Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.elsevier.com/locate/saa

Fourier transform infrared spectrum pre-processing technique selection for detecting PYLCV-infected chilli plants

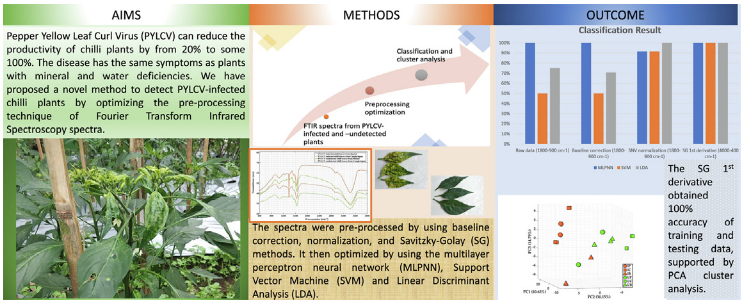
Dyah K. Agustika [a,b,*], Ixora Mercuriani [c], Chandra W. Purnomo [d], Sedyo Hartono [e], Kuwat Triyana [f], Doina D. Iliescu [a], Mark S. Leeson [a,*]

[a] School of Engineering, University of Warwick, Coventry CV4 7AL, UK
[b] Department of Physics Education, Universitas Negeri Yogyakarta, Yogyakarta, 55281 Indonesia
[c] Department of Biology Education, Universitas Negeri Yogyakarta, Yogyakarta, 55281 Indonesia
[d] Department of Chemical Engineering, Universitas Gadjah Mada, Sekip Utara Yogyakarta, 55281 Indonesia
[e] Department of Plant Protection, Faculty of Agriculture, Universitas Gadjah Mada. Jl, Flora 1, Bulaksumur, Sleman 55281, Yogyakarta
[f] Department of Physics, Universitas Gadjah Mada, Sekip Utara Yogyakarta, 55281 Indonesia

HIGHLIGHTS

- Pepper Yellow Leaf Curl Virus (PYLCV) can reduce the productivity of chilli plants by between 20% and 100%.
- The disease exhibits the same symptoms as plants with mineral and water deficiencies so is often not correctly detected.
- We propose optimized Fourier Transform Infrared Spectroscopy spectra pre-processing to detect PYLCV-infected chilli plants.
- We choose the method from denoising, normalizing and baseline correction that delivers the highest classification accuracy.
- Savitzky-Golay 1st derivative pre-processing was the best method, enabling subsequent classification accuracy of up to 100%.

GRAPHICAL ABSTRACT


AIMS
METHODS
OUTCOME
Classification Result
Classification and cluster analysis
Preprocessing optimization
FTIR spectra from PYLCV-infected and –undetected plants
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Raw data (1800-900 cm-1)
Baseline correction (1800-900 cm-1)
SNV normalization (1800-900 cm-1)
SG 1st derivative (4000-400 cm-1)
MLPNN
SVM
LDA
PC1 (40.63%)
PC2 (30.15%)
PC3 (14.79%)

Pepper Yellow Leaf Curl Virus (PYLCV) can reduce the productivity of chilli plants by from 20% to some 100%. The disease has the same symptoms as plants with mineral and water deficiencies. We have proposed a novel method to detect PYLCV-infected chilli plants by optimizing the pre-processing technique of Fourier Transform Infrared Spectroscopy spectra.

The spectra were pre-processed by using baseline correction, normalization, and Savitzky-Golay (SG) methods. It then optimized by using the multilayer perceptron neural network (MLPNN), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA).

The SG 1st derivative obtained 100% accuracy of training and testing data, supported by PCA cluster analysis.

ARTICLE INFO

Article history:
Received 20 December 2021
Received in revised form 12 April 2022
Accepted 29 April 2022
Available online 4 May 2022

Keywords:
Fourier transform infrared spectroscopy
Plant disease detection
Pre-processing
Classification

ABSTRACT

Pre-processing is a crucial step in analyzing spectra from Fourier transform infrared (FTIR) spectroscopy because it can reduce unwanted noise and enhance system performance. Here, we present the results of pre-processing technique optimization to facilitate the detection of pepper yellow leaf curl virus (PYLCV)-infected chilli plants using FTIR spectroscopy. Optimization of a range of pre-processing techniques was undertaken, namely baseline correction, normalization (standard normal variate, vector, and min–max), and de-noising (Savitzky-Golay (SG) smoothing, 1st and 2 derivatives). The pre-processing was applied to the mid-infrared spectral range (4000 – 400 cm^{-1}) and the biofingerprint region (1800 – 900 cm^{-1}) then the discrete wavelet transform (DWT) was used for dimension reduction. The pre-processed data were then used as an input for classification using a multilayer perceptron neural network, a support vector machine, and linear discriminant analysis. The pre-processing method with the highest classification model accuracy was selected for the further use in the processing. It was seen that only the SG 1st derivative method applied to both wavenumber ranges could produce 100% accuracy. This result was supported

[*] Corresponding authors.
E-mail addresses: Dyah.Agustika@warwick.ac.uk (D.K. Agustika), Mark.Leeson@warwick.ac.uk (M.S. Leeson).

https://doi.org/10.1016/j.saa.2022.121339
1386-1425/© 2022 Published by Elsevier B.V.

by principal component analysis clustering. Thus, we have demonstrated that by using the right pre-processing technique, classification success can be increased, and the process simplified by optimization and minimization of the technique used.

1. Introduction

Various plant viruses threaten worldwide food security because they can cause significant yield losses [1,2]. The global annual economic losses caused by such viruses alone has reached more than USD 30 billion [3]. The spread of a virus is very difficult to control in the field because it can only grow and reproduce inside the host cell

therefore, curative treatment cannot be performed. Virus attacks must therefore be addressed by applying reliable early diagnostic tests so that the infected plants can be rapidly eradicated and spread of the virus minimized [3]. Spectroscopic techniques have been developed as a plant disease detection system because they are faster, less expensive and more and accurate than serological, biomarker, molecular, or imaging techniques [4,5].

Plant diseases can cause changes in plant physiological development and transpiration rate, change plant tissue colour, and alter leaf shape [6]. These modifications affect the optical characteristics of plant tissues [7]. Hence, physiological changes in diseased plants can be detected by spectroscopy [8], which is thus a powerful tool for chemical analysis, characterization and identification of plant samples and biomaterials [6,9,10]. The technique presents an inexpensive method and does not require solvent extraction for sample preparation, thus reducing analysis time. Spectroscopy generates biochemical compound information from the samples and hence allows identification of major functional groups [9–11]. It has been developed in recent years as a plant disease detection tool with notable successes in detecting soil-borne fungi [12], *Geotrichum candidum* infection in tomato fruit [13], *Magnaporthe oryzae* infection in rice [14], and Huanglongbing in citrus leaves [15] in addition to the turnip yellow mosaic virus [16] and wheat streak mosaic virus [17].

Spectroscopy works by exposing a sample to polychromatic light in the infrared region, which causes molecular vibration as a result of the chemical composition of the sample and is then excited to a higher energy level due to the absorption of chemical bonds [18]. Then, backscattered light with a certain intensity becomes an indicator of the state of the plant [5]. In Fourier transform infrared (FTIR) spectroscopy, the raw interferogram data are converted into an energy transmission or absorption spectrum via the Fourier transform [18]. Important information about the character of the sample can be obtained by analyzing the differences in the area, bandwidth and intensity of the inteferogram [19]. The FTIR spectrum is large and has a complex data set so requires multivariate data processing, dealing with more than one variable at once to uncover relationships between the variables. The range of multivariate analysis methods includes classification models and cluster analysis [20]. By using such models, data patterns can be elucidated and used for future prediction [21,22].

Classification and clustering models work optimally when the FTIR spectrum is pre-processed, so that important information from the data can be separated from unwanted noise [23]. The purpose of pre-processing is thus to minimize noise and correct issues related to spectral data acquisition for multivariate analysis accuracy and improve data interpretability [20,24]. Noise can come from the scattering of light from the irradiated particles, an effect that usually appears in solid samples [23]. In addition, the diverse nature of the samples can make the identification of components in biological instances difficult. Furthermore, bio-systems consist of complex biomolecules so that under different pathological conditions, the differences from one sample to another are very small and difficult to observe in the raw spectrum [21]. These problems can be overcome by pre-processing [25]**.** The technique employs baseline spectral correction, de-noising, normalization and other manipulations [20,23,26]. In addition, the large volume of spectral data should also be reduced so that they can be easily analyzed by pattern recognition techniques. The data reduction techniques commonly used are wavenumber selection in spectral analysis by reducing the data to only include the most informative spectra [20] or use of principal component analysis (PCA), the discrete wavelet transform (DWT) and the fast Fourier transform (FFT) [27].

The choice of pre-processing method greatly affects the interpretability and accuracy of the classification model [22]. Several spectrum pre-processing methods for spectroscopy are often combined to get the optimal results [20]. Liaghat *et al.* [28] used FTIR spectroscopy to detect oil palm basal stem rot. The spectra were pre-processed using baseline correction, normalization, and Savitzky-Golay (SG) smoothing to obtain the first and second derivatives. After that, data reduction was achieved by using PCA, following which the results were then classified using several multivariate classification algorithms. Meanwhile, Salman *et al.* [10] investigated *Colletotrichum*, *Fusarium oxysporum*, *Rhizoctonia* and *Verticillium* fungi attacking plants using FTIR spectroscopy. They used the baseline correction and pre-processing normalization technique and applied Ward's algorithm [14] for clustering. For detecting the four fungi, they had to select a specific wavelength region for sample clustering. In addition, Sankaran *et al.* [25] used a pre-processing combination of visible-near infrared spectra data for the detection of Huanglongbing in citrus orchards. The spectra were normalized, then the spectral values were averaged every 25 nm to reduce the dimensions. The spectra were then pre-processed again by using first derivatives and second derivatives. Three datasets of first, second derivatives and combined datasets were then reduced using PCA. The results of pre-processing were classified using quadratic discriminant analysis (QDA), linear discriminant analysis, soft independent modelling of classification analogies (SIMCA), and k-nearest neighbour.

Although pre-processing can be achieved by combining various methods, Gerretzen *et al* [29] suggested that the best approach is to use the simplest and the least complicated method. Hence the pre-processing selection process becomes fast and unbiased [30]. Therefore, the main objective of this study is to minimize the number of pre-processing methods used to process spectral data of Pepper Yellow Leaf Curl Virus (PYLCV)-infected chilli plants and PYLCV-undetected plants. PYLCV is the main virus that attacks chilli plants in Indonesia, and this virus can reduce the productivity of chilli plants by from 20% to some 100% [31]. This virus causes curling and interveinal yellowing of the chilli leaves, and the plants become stunted [32–34]. However, lack of water or minerals also causes yellowing leaves and stunted growth [35,36]. Moreover, changes in water content can also cause the leaves to curl [37]. Hence, the disease is often not detected, leading to catastrophic mishandling of the disease. Therefore, the best possible inexpensive and reliable PYLCV-infected plant detection method is needed. Until now, there have been no research reports that have investigated the detection of PYLCV by using FTIR spectroscopy. Thus, we have proposed a novel method to achieve this by optimizing the pre-processing technique to simplify the process and shorten the analysis time. In addition, to reduce the spectral dimension, the DWT was used to reduce the dimension of the data. Optimization of the pre-processing technique was achieved by choosing one method from denoising, normalizing and baseline correction, delivering classification models with the highest accuracy value. The classification models used were a supervised support vector machine (SVM), an artificial neural network (ANN) and linear discriminant analysis (LDA). The unsupervised clustering method, PCA, was also applied to observe the separation between the samples. In this way, a suitable pre-processing method was sought to provide optimal classification model results.

2. Materials and methods

2.1. Sample preparation

The samples for the experiment were Capsicum Annuum L chilli plants taken from three commercial plantations in different regions in Indonesia, namely Bantul, D.I. Yogyakarta, Sleman, D.I Yogyakarta, and Purworejo, Central Java. The distance between commercial plantations in Cangkringan and Bantul was approximately 39 km, between Bantul to Purworejo was around 64 km, and between the plantations in Purworejo and Bantul was approximately 38 km. The difference in the location of the sample origin was intended to test whether FTIR spectroscopy can still detect PYLCV attacks on chilli plants even though the plants come from different regions. Eight samples were taken from one plantation area, four from plants with no symptoms of PYLCV attack, and four from those that indicated infection by PYLCV. The infection status of the plants was confirmed by polymerase chain reaction (PCR) testing by using pepper yellow leaf curl Indonesian virus species-specific primers [31]. Samples were taken from two different trees from each type of PCR PYLCV-undetected plant and PYLCV-infected plant in one area (eight samples came from four trees two samples were taken from the same tree for repetition). Hence, the total sample size for the three areas was 8 × 3 samples (24 samples). Each sample was named based on the infection status and the sample origin for example, infected leaves from Cangkringan had the name character IC, while the healthy ones were UC. For one sampling process, three leaves were taken. Samples from Cangkringan are depicted in Fig. 1. The leaves in Fig. 1 (a) and (b) were taken from the same tree, while (c) was from the same tree as (d), (e) the same as (f), and finally (g) the same tree as (h). The fresh sample was then added to 100 mg Potassium bromide (KBr) and crushed using an agate mortar to make it fine and well mixed. After that, the sample was put in the sample holder, compressed into a pellet and then tested with the FTIR spectrometer.

2.2. Fourier transform infrared (FTIR) spectroscopy

The FTIR spectroscopy used the Thermo Scientific Nicolet iS10 spectrometer with a beam splitter (KBr/Ge mid-infrared). This had a Deuterated TriGlycine Sulfate (DTGS) detector and was equipped with Smart Omni Transmission accessories. During the reading process, the room temperature was 25 °C, and the humidity was 75%. The spectral resolution of the measurements recorded was 8 cm^{-1} and this produced a data spacing of 0.964 cm^{-1}, so that 3736 data points were generated. However, the first and last points were of consistently zero amplitude, hence were eliminated to leave 3734 data points for further analysis. Measurements were made in the transmittance mode with the transmittance spectrum of airborne KBr used as the background. For each sample, 32 scans were performed to provide the sample transmittance spectrum in the wavenumber range 4000 – 400 cm^{-1} for pre-processing, feature extraction and multivariate analysis techniques.

2.3. Data analysis

The FTIR spectra were then analyzed, as shown in Fig. 2. First, the spectra of raw data in the wavenumber range of 4000 – 400 cm^{-1} were sent to the classification model establish the classification performance without pre-processing. Afterwards, the spectra were pre-processed by using baseline correction, normalization, and denoising methods. These approaches were then optimized by using the multilayer perceptron neural network (MLPNN), SVM and LDA classification models. In the next step, the data were cut into the biofingerprint region, 1800 – 900 cm^{-1}, and the spectra passed through the same process as the 4000 – 400 cm^{-1} spectra. The spectra selection in the biofingerprint area was intended to remove the CO_2 and water absorption spectra and also eliminated any effects of baseline distortion that may exist [38]. In addition, the biofingerprint region contained absorption bands caused by axial symmetric deformation of carbonyl such as ketones, aldehydes and esters, and also axial and angular symmetric deformation of alcohols and esters [10]. In our previous study [31], by using GC–MS to detect PYLCV-infected plants and PYLCV-undetected plants, there were differences in the composition of ketones, aldehydes, alcohols and esters. Therefore, the biofingerpint region's spectra were an area of interest in finding the differences between PYLCV-undetected and PYLCV-infected chilli plants. The pre-processing method with the highest accuracy was then analyzed by PCA to prove that the methods were also successful in clustering analysis. The flowchart of the spectral analysis and data processing is depicted in Fig. 2.

3. Theory

3.1. Pre-processing

Pre-processing of raw data helps remove noise or unwanted signals, improves the accuracy of chemometric analysis in the next step, improves data interpretation capabilities, and increases the desired signal information [38,39]. A properly applied pre-processing method can affect the system classification results, so in this research, several methods were applied. The pre-processing techniques that were used were baseline correction, normalization (standard normal variate, vector, and min–max), and de-noising with SG smoothing, and 1st and 2nd derivatives. The pre-processed data were then compared by applying them to the classification models. Each of the methods was optimized. For example, normalization had three methods and that with the highest classification result was chosen for comparison with other methods. The pre-processing method selection was made by choosing the best method of baseline correction, normalization and de-noising that delivered the highest accuracy of classification models.

3.2. Baseline correction

The baseline correction technique used was obtained from the baseline correction tool from Thermo Scientific™ OMNIC™ FTIR Software. The baseline-corrected spectra were then analyzed by using pre-processing techniques. For optimization, the baseline-corrected of the raw data in the 4000 – 400 cm^{-1} wavenumber spectral range were compared to the baseline-corrected spectra in the biofingerprint region.

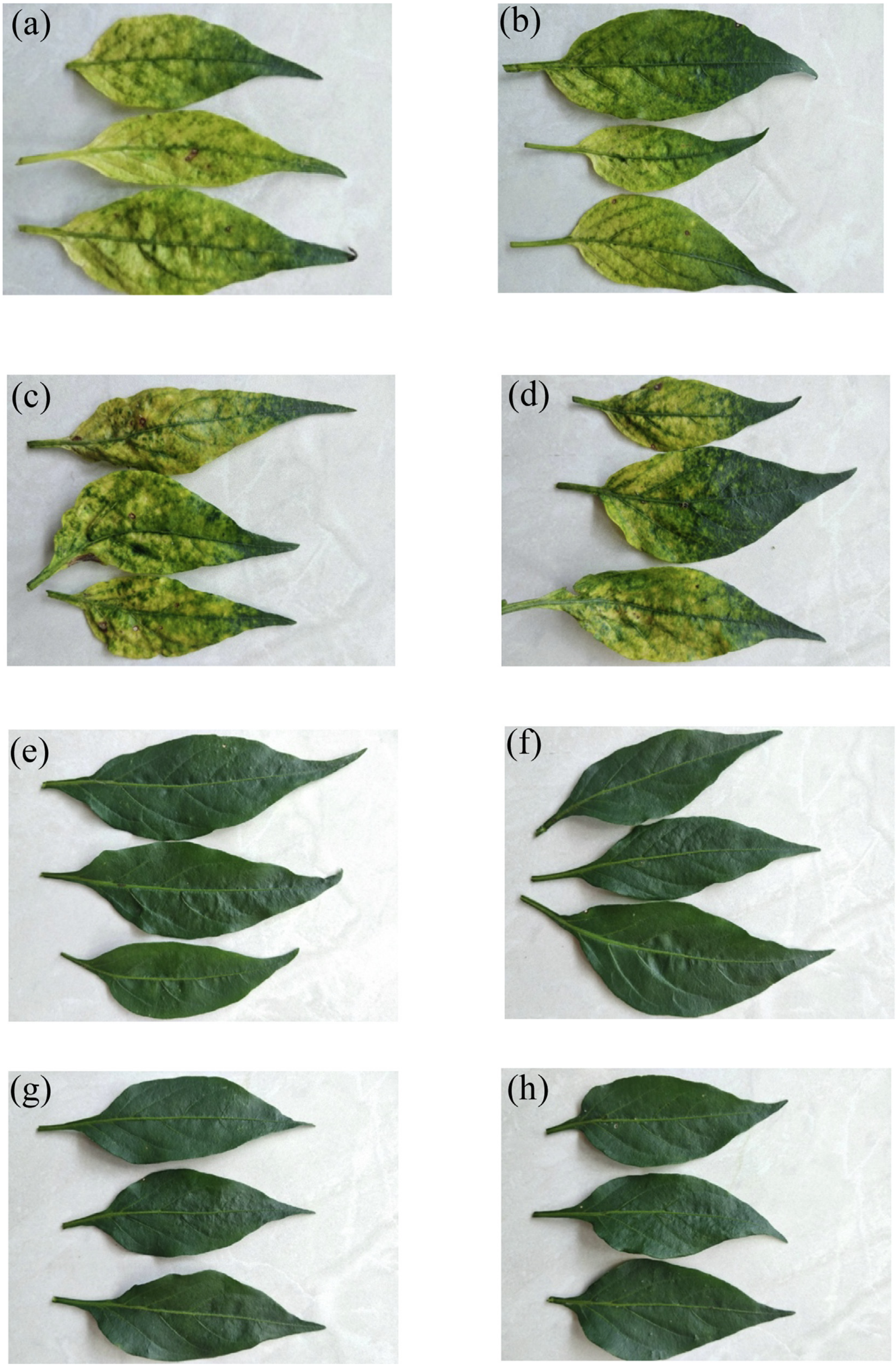


Fig. 1. Leaf samples from Cangkringan chilli plantation (a) – (b) PYLCV-infected leaves taken from infected tree 1. (c) – (d) PYLCV-infected leaves taken from infected tree 2 (e) – (f) leaves taken from PYLCV-undetected tree 1. (g) – (h) leaves taken from PYLCV-undetected tree 2.

3.3. Normalization

For normalization, three techniques were used, namely standard normal variate (SNV), vector and min–max normalization. The SNV normalization value (x_{SNV}) was obtained by subtracting for each spectral intensity value by subtracting its mean and then dividing by the standard deviation [38,40],

$$x_{SNV} = \frac{x_i - \bar{x}}{\sigma} \tag{1}$$

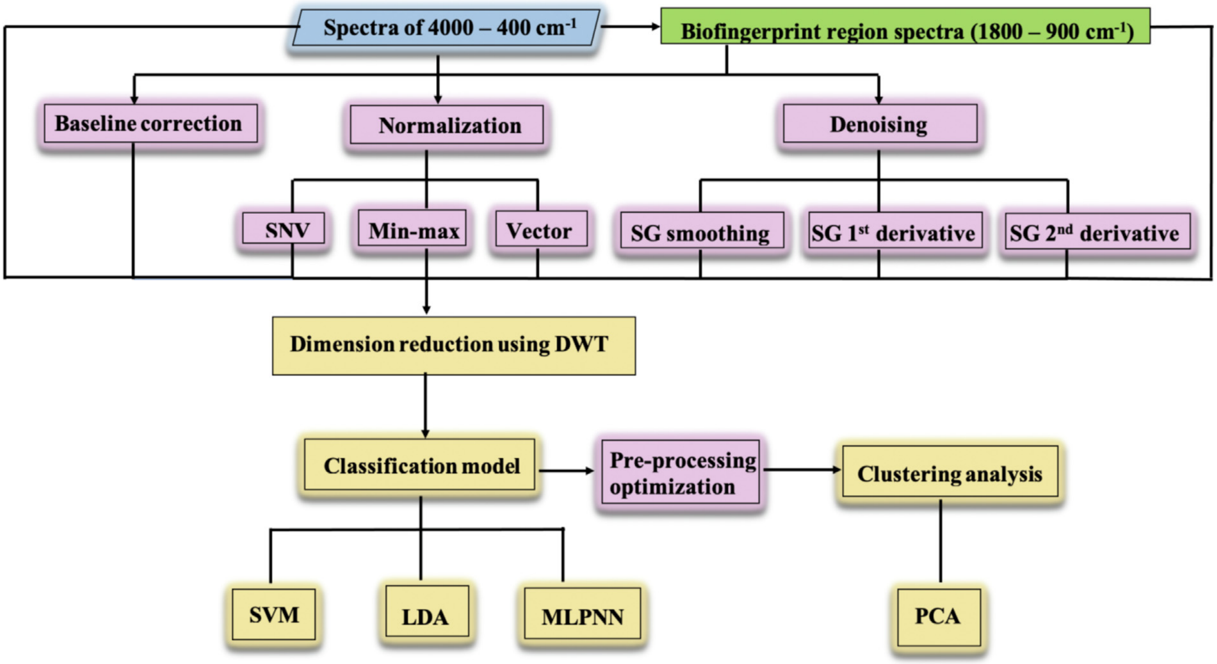


Fig. 2. Flowchart of the spectral analysis and data processing.

where x_i was the spectral data point with $i = 1, 2, \ldots, j$, $\bar{x}$ was the mean spectral value, and σ was the standard deviation of the spectral data value. Meanwhile, vector normalization (x_{vec}) was obtained from equation (2) below [38].

$$x_{vec} = \frac{x_i}{\sqrt{\sum_{i=1}^{N} x_i^2}} \qquad (2)$$

For min–max normalization, the maximum (x_{max}) and minimum (x_{min}) of the spectral intensities were first found, and then normalization values were calculated by using equation (3) [38].

$$x_{min-max} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad (3)$$

3.4. Savitzky-Golay (SG) smoothing, first and second derivative

In the SG method, signal smoothing was achieved by a least squares polynomial fit to the set of input data points in a window of $\pm M$ data points. Here, we took $M = 10$, giving a window length of $2M + 1 = 21$ and a fourth order polynomial was selected to obtain the least squared fit polynomial coefficients. Further analysis used the polynomial value at the window midpoint [41], which moved repeatedly by one data point to deliver the filter output the consecutive points. The SG filter process by fitting polynomials to data points was thus the same as convolving samples in a window with a fixed impulse response [41,42]. In addition, the first and second derivatives, obtained by differentiating the smoothed SG spectral curves, were also used as pre-processing techniques. The smoothed SG, first and second derivatives were applied to the spectral ranges of 4000 – 400 cm^{-1} and 1800 – 900 cm^{-1}, and then optimized by selecting the method that generated the highest classification model accuracy.

3.5. Dimension reduction

Since raw data in the spectral range 4000 – 400 cm^{-1} generated 3734 data points, this relatively large number could complicate the pattern recognition process. Therefore, the dimension of the data needed to be reduced by using DWT multiresolution analysis. In the analysis, the sensor response signal $x(t)$ can be described as a linear combination of scaling and base functions,

$$x(t) = \sum_{k} a_{j0}(k)\varphi_{j0,k}(t) + \sum_{k}\sum_{j=j0}^{N-1} d_j(k)\psi_{j,k}(t)(4)$$

a_{j0} is the approximation coefficient of the scaling basis vector $\varphi_{j0,k}(t)$, which represents the approximate original signal $x(t)$ and d_j is the detail coefficient of the wavelet vector base $\psi_{j,k}(t)$ for the DWT [43]. The approximation component represents the original signal, while the detail component is what is lost by high-frequency filtering. The former can be passed to the next filter to obtain more higher-level components. In the process there will be a decrease in the sampling rate because some components (samples) of the signal are removed, known as *downsampling*. The number of samples at the output of this process is a fraction of the number of samples input. In the context of signal compression, a subsignal (signal approximation in subspace) is used to represent the original signal [44].

In the analysis, level 5 Haar (db1), daubechies2 (db2), daubechies4 (db4), and symlet2 (sym2) wavelets were used to provide pairs of scaling and wavelet functions. Each approximation and detail decomposition coefficients were summed to reconstruct the FTIR spectra. Reconstruction results generally enable mean squared error (MSE) calculations between original and reconstructed signals to be made [45,46]. In this study, the spectra came from 24 samples, then to find the best wavelet, the average MSE value of each wavelet was calculated and then compared to determine the most suitable wavelet for the FTIR spectra.

3.6. Classification models

3.6.1. SVM

These are kernel-based learning models [47,48] that employ a decision plane, known as a *hyperplane*, to separate classes between samples**.** In the SVM process, the data obtained from the measurement results are divided into training data and test data. The former set comprises pairs $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{j}$

$i = 1, 2, \cdots, j$ for inputs $\mathbf{x}_i$ with sample class labels y_i to.

$$\text{Minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{j}\xi_i \tag{5}$$

$$\text{Subject to} x_i(< \mathbf{w}, \mathbf{y_i} >) + b \geq 1 - \xi_i, \xi_i \geq 0$$

C is the cost, which is the regularization parameter and is the trade-off between margin maximization (achieved by minimizing $\|\mathbf{w}\|^2/2$) and constraint relaxation. The optimal hyperplane can be found by using a Lagrange multiplier with the Karush-Kuhn-Tucker (KKT) conditions, to produce classification of the test dataset when,

$$f(x) = \text{sign}(< \boldsymbol{w}^*.\mathbf{x} > + b^*) = \text{sign}\left(\sum_{i=1}^{n}\alpha_i^* y_i(\mathbf{x}_i.\mathbf{x}) + b^*\right) \tag{6}$$

where α_i^* is the KKT multiplier. The inner product in the equation (6) can be replaced by the kernel function $K(\mathbf{x}_i, \mathbf{x})$, which is a transformation function that maps inputs x_i and × to some feature space. Here, we used Radial Basis Function (RBF) kernels for the SVM analysis.

$$K(\mathbf{x}_i, \mathbf{x}) = \exp\left(\frac{-\|\mathbf{x}_i, \mathbf{x}\|^2}{2\sigma^2}\right) = \exp\left(-\gamma\|\mathbf{x}_i, \mathbf{x}\|^2\right) \tag{7}$$

where σ is the standard deviation of the Gaussian distribution, with $\gamma = 1/2\sigma^2$. Kernel function performance may result in overfitting, reducing SVM accuracy. This was avoided by using K-fold cross-validation on a grid search method [49] to optimize the hyperparameters C in equation (5) and γ in equation (7).

3.6.2. MLPNN

ANNs are supervised learning classification techniques that can approximate functions after a proper training process and one of the most frequently used ANN architectures is the MLPNN. This model consists of an input layer, where input data are received, a hidden layer, which oversees recognizing system patterns, and an output layer generated from the previous neuron processing layer. An output target is previously determined and the system compares the outcome of the learning process with this [50]. One of the algorithms in MLPNN that is often used is backpropagation (BP), where the training data provide learning rules so that the weights are adjusted to make the output **y** closer to the target **t**. The training data formed in the matrix x are the input to the neurons in the input layer. Furthermore, each node in a layer, both input and hidden layers, is connected to each node in the next layer with a certain weight. [51].

In this research, the MLPNN used consisted of an input layer, one hidden layer with ten nodes, and an output layer with one node. The training process for each data set was carried out 10 times, then the weights of the network with the smallest MSE were saved for the testing process.

3.6.3. LDA

This classification technique separates sample classes by using the their respective variances [52]. In LDA, the data are transformed to a low-dimensional space by maximizing the variance between classes and minimizing the variance within the same class [53], resulting in dimensionality reduction. Here, the data from the FTIR results formed a matrix **x** that could be divided into two classes (PYLCV-undetected and -infected plants), each class had mean of μ_i, while the total mean of all classes was μ. The average distance between μ_i and μ was calculated to obtain the variance between classes,. Meanwhile, the variance in each class, was obtained from the difference between the mean and element of that class. Data transformation into new dimensions was carried out based on the Fisher criterion [54,55].

The calculation of classification models accuracy of MLPNN, SVM and LDA was carried out on the training, testing, and overall data to determine the performance of the classification models [56],

$$\%\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}} \times 100\% \tag{8}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

3.6.4. Principal component analysis (PCA)

PCA is an unsupervised learning technique that reduces the dimensions of the data, without losing important information, it may even find hidden information in the data, and find data patterns by extracting important information from within them [57]. PCA transforms data by reducing redundancy so it can be more easily interpreted. The matrix data input has a covariance matrix that can be manipulated. The variance of each measurement variable is depicted on the diagonal component, and this component describes the characterization of the data. Meanwhile, the off-diagonal component describes data redundancy. Matrix manipulation is achieved by finding the eigenvalues and eigenvectors of covariance matrix. The eigenvectors, sorted from highest to lowest, represent the variability of the system. The component PC1 contains the maximum data variance or the largest sample dispersion [58].

4. Results and discussion

4.1. FTIR spectra

PYLCV-infected and -undetected plants exhibit different physiological parameters, and in this research, we utilised transmission mode FTIR spectroscopy to investigate the differences between these plant conditions. This approach generated 3734 points for one spectrum over the wavenumber range 4000–400 cm^{-1}. There were 24 spectra comprising 12 of PYLCV-infected plants and 12 from PYLCV-undetected plants. As a representative, four spectra of PYLCV-infected and -undetected chilly plants from Bantul and Cangkringan are plotted in Fig. 3.

As can be seen in Fig. 3, the FTIR spectroscopy for the PYLCV-infected and -undetected chilli plants produced spectral curves that were quite similar, differing only in a sharp downward peak at 1385 cm^{-1} for PYLCV-infected plants. This wavenumber corresponds to C-H bending in the alkane functional group [59]. In our previous research [31], alkane, which is part of hydrocarbon, was one of the functional groups with the highest variability that distinguished the PYLCH-infected and the PYLCV-undetected samples. However from GC–MS analysis, the compound group that had the largest contribution in discriminating the samples is the primary amine, and not the alkane. The different result in this work was caused by the sampling procedure. In our previous research, Gas Chromatograph-Mass Spectrometry detected the VOCs emitted by the plants when fresh leaf samples were taken directly from the trees without any sample preparation. Meanwhile, in the FTIR spectroscopy here, the leaf samples were processed into pellets by using KBr.

The peaks at 1385 cm^{-1} for the infected samples from Bantul and Cangkringan had different amplitudes. Meanwhile, other peak differences are not large and thus the spectra from the PYLCV-infected and -undetected plants alone are not sufficient to demon-

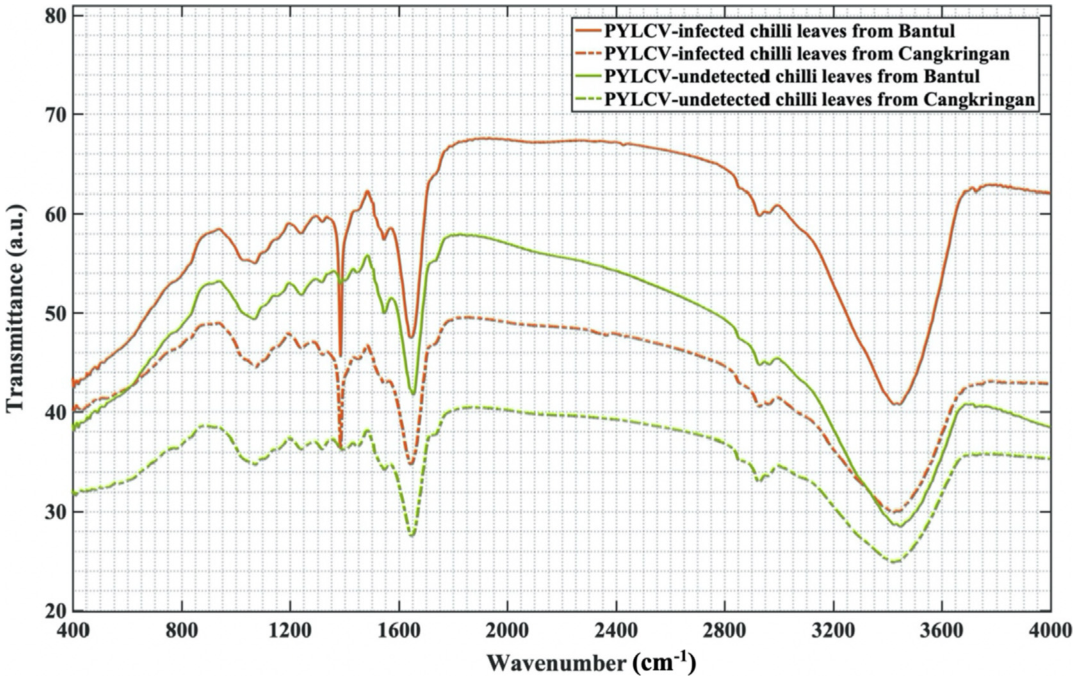


Fig. 3. Averaged spectra of PYLCV-infected and -undetected chilli plants.

strate that FTIR spectroscopy can distinguish between the plants. Moreover, PYLCV-infected samples from Cangkringan and PYLCV-undetected samples from Bantul overlap and partiall coincide in the ranges 800–400 cm^{-1} and 3400–3200 cm^{-1}, making spectra from the infected and PYLCV-undetected samples hard to distinguish. In addition, each spectrum consists of 3734 wavenumber variables, and there are 24 spectra of PYLCV-infected and -undetected samples. The large amount of data makes patterns from raw spectra difficult to identify and use to predict new test samples. Therefore, multivariate analysis is needed to handle large data sets, obtain complete information from spectra and model spectral patterns. This makes the system more reliable for future tests and we thus applied the steps shown in Fig. 2 to both raw and pre-processed data. There were three pre-processing methods used for the analysis baseline correction, normalization, and denoising. The pre-processing method was optimized by comparing the accuracy results obtained by the various classification models (SVM, LDA, and MLPNN) following their application. The pre-processing system with the highest classification model accuracy was then validated by using principal component analysis (PCA).

4.2. Raw data classification results

The size of the raw spectral dataset was too large for the classification models hence, dimension reduction was carried out using a DWT. This was applied to the full raw data spectral range from 4000 to 400 cm^{-1} utilising level five of the orthogonal wavelets db1, db2, db4 and sym2. The results of the DWT analysis appear in Table 1, where it may be seen that wavelet db1 produced the smallest MSE value, and this shows that the signal reconstruction with db1 is the most similar to the original signal compared to the other five wavelets. Therefore, this was used for the next stage of the processing.

Table 1
MSE comparison of wavelets employed.

Type of wavelet	MSE
db1	$(2.06 \pm 0.80) \times 10^{-28}$
db2	$(1.13 \pm 0.80) \times 10^{-25}$
db4	$(1.23 \pm 0.63) \times 10^{-24}$
sym2	$(1.13 \pm 0.08) \times 10^{-25}$

A level five DWT reduced the full spectrum of 3734 data points to 117

in the biofingerprint area, 934 datapoints were reduced to 30 points. The reduced the spectra were then processed by SVM, LDA, and MLPNN, with 60% of the data points employed for training and 40% for testing, split using the Kennard-Stone algorithm [49].

First, accuracy was calculated using equation (8) The analysis was completed for full range of the raw data, then the process was repeated for the biofingerprint region and the results are shown in Fig. 4.

From Fig. 4, it can be seen that the highest total accuracy for the full range of raw data was achieved with MLPNN and LDA, returning overall accuracies of 91.7%, whereas SVM only delivered 50% accuracy. Meanwhile, using the biofingerprint spectrum, MLPNN reached an accuracy of 100% but LDA and SVM only achieved 75% and 50%, respectively. The effect of using the 1800–

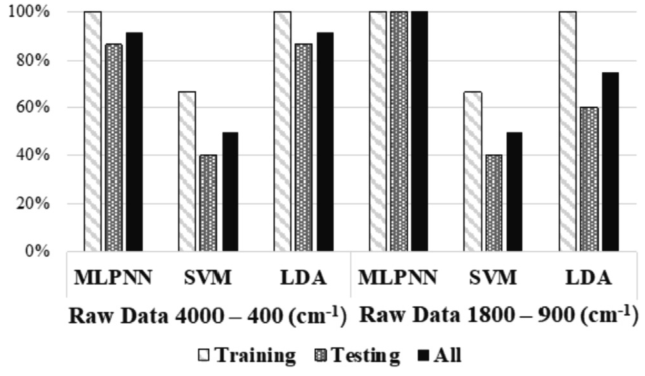


Fig. 4. Raw data accuracy results of MLPNN, SVM, and LDA classification models.

900 cm^{-1} spectral range was that the MLPNN achieved 100% accuracy, so was the best method in the raw data group.

4.3. Baseline correction Pre-processing classification results

Baseline correction aims to overcome baseline drift caused by temperature, light, humidity, and other environmental factors that affect the precision of spectrometer optical components making it difficult to determine the peak and the peak area of the spectra [60 61]. By using Thermo Scientific™ OMNIC™ FTIR Software the raw data were pre-processed to get baseline corrected spectra. These were then reduced using level five of the db1 DWT in the spectral ranges of 4000 – 400 cm^{-1} and 1800 – 900 cm^{-1}. Both pre-processed spectra were then classified by using MLPNN, SVM, and LDA. The classification results of the two resulting spectra are shown in Fig. 5.

The baseline correction method for the biofingerprint area reached 100% accuracy using MLPNN for both testing and training data, while the full spectrum reached 91.7% accuracy, with the SVM accuracies for both spectra being the same. Meanwhile, using LDA on the baseline corrected 1800 – 900 cm^{-1} region gave better results than for the whole spectral range. Overall, baseline corrected spectra in the biofingerprint region had a better classification model accuracy than the full spectral range of 4000 – 400 cm^{-1}.

4.4. Normalization pre-processing classification results

In the next step, the raw data in the 4000 – 400 cm^{-1} and 1800 – 900 cm^{-1} wavenumber spectra ranges were pre-processed by using normalization techniques. Normalization aims to compensate for the effects of scaling and spectrum shifts due to scattering, the effect of fluctuations in source power, and differences in sample characteristics such as differences in particle size [38]. The techniques used for pre-processing in this research were SNV, vector, and min–max methods. These were applied to the spectral data then the results were fed into the DWT and the classification models. The classification analysis results for normalized spectra range of 4000 – 400 cm^{-1} are depicted in Fig. 6. Next, the normalized spectra were truncated to the biofingerprint area, and the classification models applied, with results as shown in Fig. 7.

Overall, the SNV normalization pre-processing method had the highest accuracy classification results compared to vector and min–max normalization, both for the full spectral range and the truncated version. Accuracy of 100% was achieved by SNV for the former and by LDA for the latter. However, the overall accuracy of the classification model for SNV pre-processing in the 1800 – 900 cm^{-1} range was the highest compared to other pre-processing models. Min-max normalization gave the second-best accuracy result, with values more than 80% for the MLPNN, SVM, and LDA models for both wavenumber ranges Meanwhile, the accuracy of vector normalization was the lowest, even for the two spectra in the SVM model

training data accuracy did not reach 100%. As SNV normalization at for 1800 – 900 cm^{-1} had the highest classification model accuracy value, it was thus considered the best model in the normalization group.

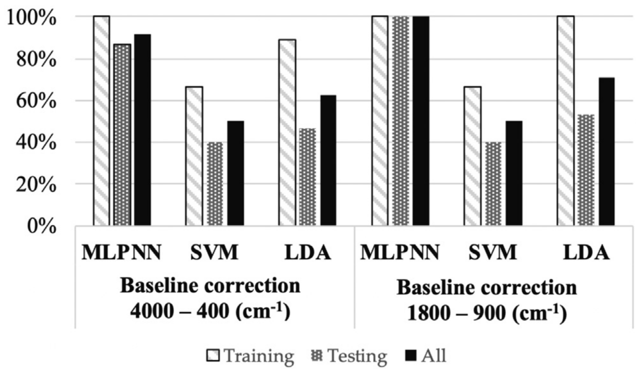


Fig. 5. Baseline corrected spectral accuracy results for MLPNN, SVM, and LDA classification models.

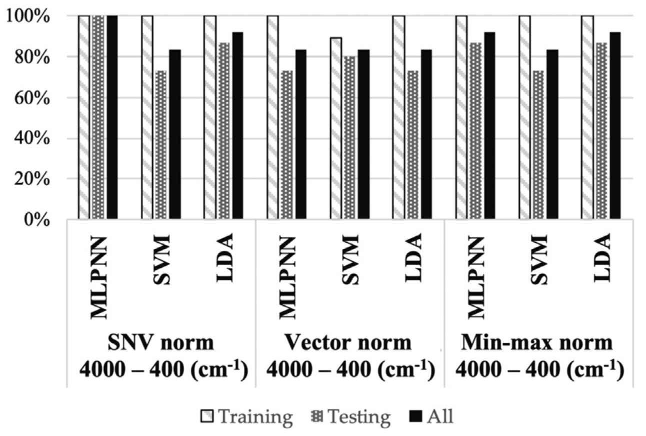


Fig. 6. Normalized spectra (4000 – 400 cm^{-1} region) accuracy results of MLPNN, SVM, and LDA classification models.

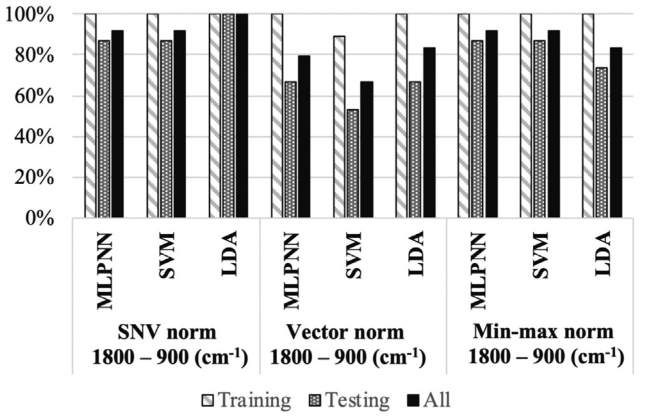


Fig. 7. Normalized spectra (biofingerprint region) accuracy results of MLPNN, SVM, and LDA classification models.

4.5. Classification results for SG analysis

The SG pre-processing began with the SG smoothing of raw data, and then generation of the first and second derivatives of the smoothing curve. These three options were then applied to MLPNN, SVM, and LDA, and the results can be seen in the Fig. 8. The SG pre-processed spectra data were then again cut in the biofingerprint area and the classification models applied with accuracy results depicted in Fig. 9.

From Figs. 8 and 9, it can be seen that the SG 1st derivative of the 4000–400 cm^{-1} spectral range and biofingerprint region reached the highest classification accuracy for MLPNN, SVM, and LDA. Meanwhile, 100% SVM total accuracy was also reached by the SG 2nd derivative pre-processed spectra of both wavenumber ranges, although the two other classification models did not show the same results. Moreover, the smoothing curve in the 4000 –

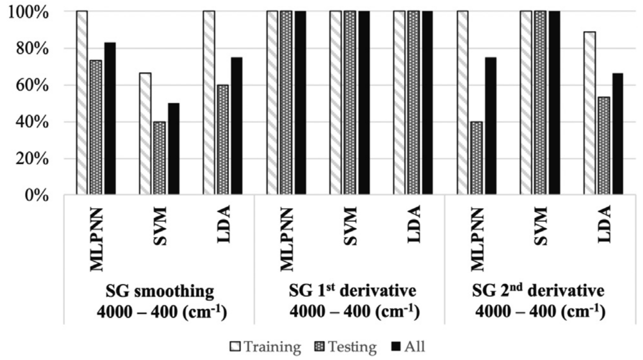


Fig. 8. De-noising spectra by using SG technique accuracy results of MLPNN, SVM, and LDA classification models in the spectral range of 4000 – 400 cm^{-1}.

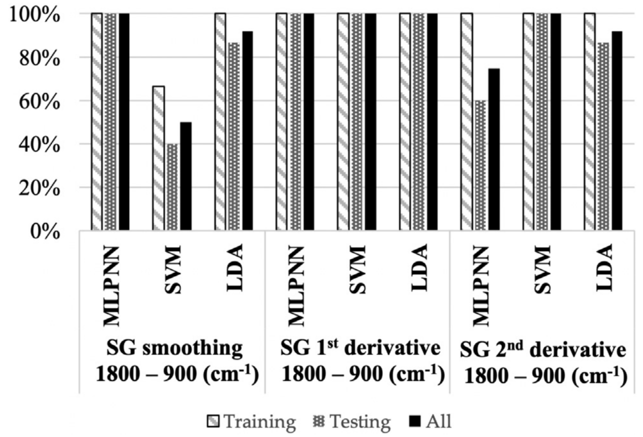


Fig. 9. De-noising spectra by using SG technique accuracy results of MLPNN, SVM, and LDA classification models in the spectral range 1800 – 900 cm^{-1}.

400 cm^{-1} range shows the lowest accuracy results of the classification models compared to the other SG methods.

SG smoothing is used to smooth the signal by attenuating high-frequency noise, hence the waveform peak shape, amplitude and width of the desired signal will be maintained [62]. Meanwhile, the SG derivative is used as a band-pass filter which can also reduce signals at low frequencies [63,64]. The better classification results from the SG derivative implies that the noise in the spectra occurs not only at high frequencies but also at low frequencies. Therefore, the SG 1st derivative of the 4000–400 cm^{-1} and 1800–900 cm^{-1} wavenumber spectral ranges will be used for optimization between the pre-processing methods.

4.6. PCA analysis

The sub-methods of the three methods of pre-processing that gives the highest accuracy were compared to find which procedure has the highest accuracy among the classification models. The highest ones were then processed by PCA to ascertain the degree of sample separation in the cluster analysis. The best sub-methods of each pre-processing methods can be seen in Table 2.

Table 2
Comparison of classification accuracy between the optimized pre-processing methods.

Type of pre-processing	MLPNN	SVM	LDA
Raw data (1800–900 cm^{-1})	100%	50%	75%
Baseline correction (1800–900 cm^{-1})	100%	50%	70.83%
SNV normalization (1800–900 cm^{-1})	91.67%	91.67%	100%
SG 1st derivative (4000–400 cm^{-1})	100%	100%	100%
SG 1st derivative (1800–900 cm^{-1})	100%	100%	100%

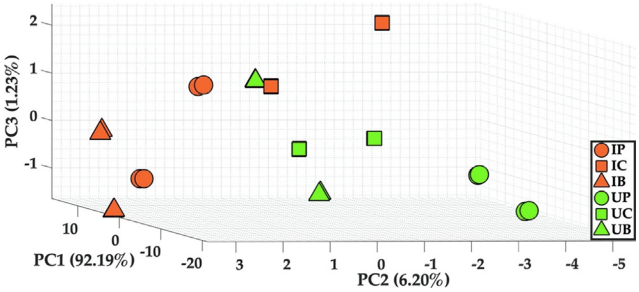


Fig. 10. PCA results of raw data in 4000 – 400 cm^{-1} spectral range.

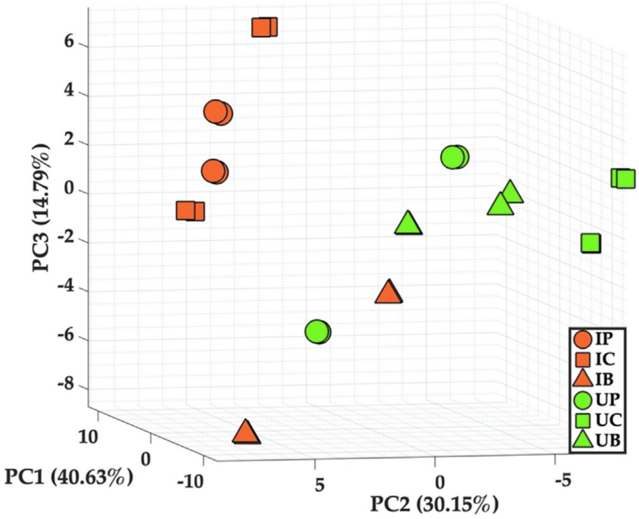


Fig. 11. PCA results of pre-processed spectra by using SG 1st derivative for the range 4000 – 400 cm^{-1}.

From Table 2, the SG 1st derivative results in both wavenumber ranges reached 100% for all the classification methods and were thus chosen for PCA analysis. The PCA clusters of the methods were then compared with the raw spectral data to ascertain the effect of the pre-processing method on the cluster analysis. The PCA result for the raw spectral data is shown in Fig. 10. In the figure, the samples that came from the same trees are overlapping. It is apparent that the PYLCV-infected and -undetected chilli plant regions are also overlapping, and the distances between the data points of the groups are too close. Therefore, the applying raw data directly to the classification models will deliver poor results.

The PCA results of the SG 1st derivative for the 4000–400 cm^{-1} and 1800–900 cm^{-1} spectral ranges are shown in Figs. 11 and 12.

It is clear from Figs. 11 and 12, that pre-processing the PYLCV-infected and -undetected plant groups delivered more distinguishable results than employing the raw data in Fig. 10, moreover SG 1st derivatives achieved the highest classification performance of 100% for all three classification models in SG analysis applied over both wavenumber ranges. In Fig. 10, the position of the PYLCV-undetected UBs are close to the infected IC ones. Moreover, the infected IC results are far from the rest of the group of infected plants. However, when pre-processed by the SG 1st derivative, as can be seen in Fig. 11, the infected plants of IB appear to be in the PYLCV-undetected plants' cluster region, but there is no overlap between the samples. Therefore, with the slack variable and flexible RBF kernel of SVM, the classification of PYLCV-undetected and diseased plants for training and testing data was still 100% successful, and this was reinforced with the MLPNN and LDA results. On the other hand, pre-processing with the SG 1st derivative in the spectra range of 1800–900 cm^{-1} gave better clustering results as can be seen in the Fig. 12, where groups of PYLCV-infected and -undetected plants were clearly clustered and could be distinguished by a linear plane.

Overall, SG 1st derivative pre-processing was the best method when compared to the other techniques investigated. Using it, the MLPNN, SVM, and LDA classification results were 100% in training and test data for both wavenumber ranges. Meanwhile, the clustering analysis of the 1st derivative SG spectra in the range 1800 – 900 cm^{-1} shows a clearer separation than the same technique at the 4000 – 400 cm^{-1} range. However, the MLPNN, SVM, LDA can still provide 100% classification results for the full spectral range. Although baseline correction and normalization can overcome the spectrum shifts and scaling problems, the best classification accuracy was achieved when the SG 1st derivative was applied. This proves that the system will have a better result when what is effectively a band-pass filter is used to reduce high and low frequency noise.

For comparison, we also combined several pre-processing systems with the best classification accuracy of each method (baseline correction, SNV normalization, and SG 1st derivative) and applied them to both the 4000–400 cm^{-1} spectral range and the biofingerprint area, as shown in Table 3.

From Table 3, the three classification models have 100% accuracy for recognizing training and testing data if baseline correction, SNV normalization, and SG1st derivative are combined together and applied to the biofingerprint range. The second-best accuracy of all the models was obtained by the combination of baseline correction and SG 1st derivative, followed by baseline correction combined with SNV normalization both also applied to the biofingerprint range. Meanwhile, the best classification accuracy for the full spectra range was achieved by the combination of baseline correction with the SG 1st derivative and SNV normalization with the SG 1st derivative. However, these cannot provide 100% accuracy for all three classification models.

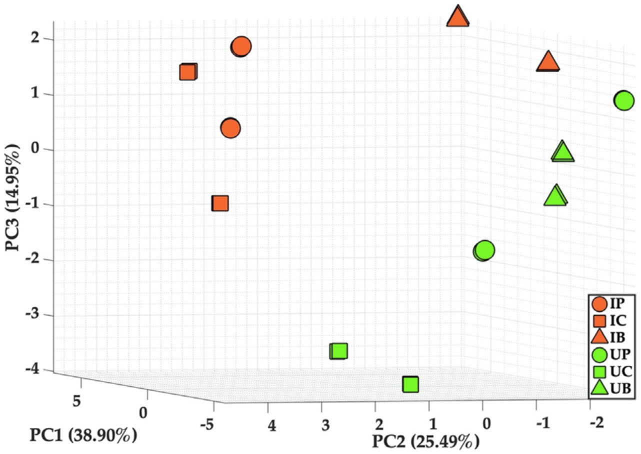


Fig. 12. PCA results of pre-processed spectra by using SG 1st derivative for the range 1800 – 900 cm$^{-1.}$

For further analysis of distinguishing PYLCV-infected and -undetected chilli plants by using FTIR spectroscopy, we suggest using spectra in the wavenumber range of 4000–400 cm^{-1} with the SG 1st derivative because, in this range, all the information and phenomena that occurred in the signal are included.

5. Conclusions

A novel method to detect PYLCV-infected chilli plants by using FTIR spectroscopy has been developed and the best pre-processing method to improve system classification results has been investigated. We optimized the pre-processing methods and then retained only one for the next analysis step to simplify and shorten the process. Optimization was undertaken for baseline correction, normalization, and de-noising methods with the signal processing applied to both the full spectrum of raw data (4000 – 400 cm^{-1}) and the biofingerprint region (1800 – 900 cm^{-1}). Considerable reduction in the spectral dimension was needed and achieved via the DWT. The pre-processed techniques were then compared by applying their outputs to MLPNN, SVM, and LDA. As a result, the SG 1st derivative applied to both spectral ranges obtained 100% accuracy of training and testing data for the three methods, supported by PCA cluster analysis. However, the analysis should be performed on the full wavenumber spectrum of 4000 – 400 cm^{-1} because this contains all the information and phenomena in the signal without the inevitable loss by truncating the spectral range. By selecting the proper pre-processing technique, the analysis of FTIR spectra data may be simplified by employing just one pre-processing method and still generate high accuracy classification results.

CRediT authorship contribution statement

Dyah K. Agustika: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Visualization, Funding acquisition. **Ixora Mercuriani:** Validation, Resources, Investigation. **Chandra W. Purnomo:** Validation. **Sedyo Hartono:** Conceptualization, Validation. **Kuwat Triyana:** Validation, Resources. **Doina D. Iliescu:** Methodology, Supervision. **Mark S. Leeson:** Conceptualization, Formal analysis, Investigation, Data curation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 3
Comparison of classification accuracy between the combination of pre-processing methods.

Combination of pre-processing method	ANN	SVM	LDA
SNV + SG 1st derivative (4000–400 cm^{-1})	100%	83.33%	79.17%
Baseline correction + SNV (4000–400 cm^{-1})	100%	66.67%	83.33%
Baseline correction + SNV + SG 1st derivative (4000–400 cm^{-1})	100%	62.50%	91.67%
Baseline correction + SG 1st derivative (4000–400 cm^{-1})	100%	66.67%	62.50%
Baseline correction + SNV + SG 1st derivative (1800–900 cm^{-1})	100%	100%	100%
Baseline correction + SG 1st derivative (1800–900 cm^{-1})	100%	91.67%	100%
Baseline correction + SNV (1800–900 cm^{-1})	100%	83.33%	91.67%

Acknowledgements

This research was funded by the Indonesia Endowment Fund for Education (LPDP), Ministry of Finance, Republic of Indonesia, grant number S-2507/LPDP.4/2019.

References

[1] R.A.C. Jones, Global plant virus disease pandemics and epidemics, Plants 10 (2) (2021) 233, https://doi.org/10.3390/plants10020233.

[2] D.S. Gandolfo, H. Mortimer, J.W. Woodhall, N. Boonham, Fourier transform infra-red spectroscopy using an attenuated total reflection probe to distinguish between Japanese larch, pine and citrus plants in healthy and diseased states, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 163 (2016) 181–188, https://doi.org/10.1016/j.saa.2016.03.022.

[3] V. Nicaise, Crop immunity against viruses Outcomes and future challenges, Front. Plant Sci. 5 (2014) 1–18, https://doi.org/10.3389/fpls.2014.00660.

[4] A.Y. Khaled, S. Abd Aziz, S.K. Bejo, N.M. Nawi, I.A. Seman, D.I. Onwude, Early detection of diseases in plant tissue using spectroscopy–applications and limitations, Appl. Spectrosc. Rev. 53 (1) (2018) 36–64, https://doi.org/10.1080/05704928.2017.1352510.

[5] M. I. S. Mohd Hilmi Tan, M. F. Jamlos, A. F. Omar, F. Dzaharudin, S. Chalermwisutkul, and P. Akkaraekthalin, "Ganoderma boninense disease detection by near-infrared spectroscopy classification A review," *Sensors*, vol. 21, no. 9, p. 3052, 2021, 10.3390/s21093052.

[6] A.-K. Mahlein, Plant Disease Detection by Imaging Sensors – Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping, Plant Dis. 100 (2) (2016) 241–251, https://doi.org/10.1094/PDIS-03-15-0340-FE.

[7] D.G. Ivanova, B.R. Singh, Nondestructive FTIR monitoring of leaf senescence and elicitin-induced changes in plant leaves, Biopolym. - Biospectroscopy Sect. 72 (2) (2003) 79–85, https://doi.org/10.1002/bip.10297.

[8] C. Zhang, X. Feng, J. Wang, F. Liu, Y. He, W. Zhou, Mid-infrared spectroscopy combined with chemometrics to detect Sclerotinia stem rot on oilseed rape (Brassica napus L.) leaves, Plant Methods 13 (1) (2017) 39, https://doi.org/10.1186/s13007-017-0190-6.

[9] T. Durak and J. Depciuch, "Effect of plant sample preparation and measuring methods on ATR-FTIR spectra results," *Environ. Exp. Bot.*, vol. 169, no. October 2019, p. 103915, 2020, 10.1016/j.envexpbot.2019.103915.

[10] J.V. Link, A.L.G. Lemes, I. Marquetti, M.B. dos Santos Scholz, E. Bona, Geographical and genotypic classification of arabica coffee using Fourier transform infrared spectroscopy and radial-basis function networks, Chemom. Intell. Lab. Syst. 135 (2014) 150–156, https://doi.org/10.1016/j.chemolab.2014.04.008.

[11] J. Johnson, J. Mani, N. Ashwath, M. Naiker, Potential for Fourier transform infrared (FTIR) spectroscopy toward predicting antioxidant and phenolic contents in powdered plant matrices, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 233 (2020), https://doi.org/10.1016/j.saa.2020.118228 118228.

[12] A. Salman, L. Tsror, A. Pomerantz, R. Moreh, S. Mordechai, M. Huleihel, FTIR spectroscopy for detection and identification of fungal phytopathogenes, Spectroscopy 24 (2010) 261–267, https://doi.org/10.3233/SPE-2010-0448.

[13] P. Skolik, M.R. McAinsh, F.L. Martin, ATR-FTIR spectroscopy non-destructively detects damage-induced sour rot infection in whole tomato fruit, Planta 249 (3) (2019) 925–939, https://doi.org/10.1007/s00425-018-3060-1.

[14] L. Gaoqiang, D. Changwen, M. Fei, S. Yazhen, Z. Jianmin, Responses of leaf cuticles to rice blast Detection and identification using depth-profiling Fourier transform mid-infrared photoacoustic spectroscopy, Plant Dis. 104 (3) (2020) 847–852, https://doi.org/10.1094/PDIS-05-19-1004-RE.

[15] S.A. Hawkins, B. Park, G.H. Poole, T. Gottwald, W.R. Windham, K.C. Lawrence, Detection of citrus huanglongbing by fourier transform infrared-attenuated total reflection spectroscopy, Appl. Spectrosc. 64 (1) (2010) 100–103, https://doi.org/10.1366/000370210790572043.

[16] S. Kim, S. Lee, H.-Y. Chi, M.-K. Kim, J.-S. Kim, S.-H. Lee, H. Chung, Feasibility study for detection of Turnip yellow mosaic virus (TYMV) infection of Chinese cabbage plants using Raman spectroscopy, Plant Pathol. J. 29 (1) (2013) 105–109.

[17] C. Farber, R. Bryan, L. Paetzold, C. Rush, D. Kurouski, Non-Invasive Characterization of Single-, Double- and Triple-Viral Diseases of Wheat With a Hand-Held Raman Spectrometer, Front. Plant Sci. 11 (2020), https://doi.org/10.3389/fpls.2020.01300.

[18] H.J. Butler, M.R. McAinsh, S. Adams, F.L. Martin, Application of vibrational spectroscopy techniques to non-destructively monitor plant health and development, Anal. Methods 7 (10) (2015) 4059–4070, https://doi.org/10.1039/c5ay00377f.

[19] G. Cakmak-Arslan, Monitoring of Hazelnut oil quality during thermal processing in comparison with extra virgin olive oil by using ATR-FTIR spectroscopy combined with chemometrics, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 266 (2022), https://doi.org/10.1016/j.saa.2021.120461 120461.

[20] M.J. Baker et al., Using Fourier transform IR spectroscopy to analyze biological materials, Nat. Protoc. 9 (8) (2014) 1771–1791, https://doi.org/10.1038/nprot.2014.110.

[21] L. Bai, Y. Liu, Classification of FTIR cancer data using wavelets and fuzzy C-means clustering, Wavelet Appl. Ind. Process. III 6001 (2005) 60010B, https://doi.org/10.1117/12.629946.

[22] L.C. Lee, C.Y. Liong, A.A. Jemain, A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum, Chemom. Intell. Lab. Syst. 163 (2017) 64–75, https://doi.org/10.1016/j.chemolab.2017.02.008.

[23] Å. Rinnan, Pre-processing in vibrational spectroscopy-when, why and how, Anal. Methods 6 (18) (2014) 7124–7129, https://doi.org/10.1039/c3ay42270d.

[24] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC - Trends Anal. Chem. 28 (10) (2009) 1201–1222, https://doi.org/10.1016/j.trac.2009.07.007.

[25] S. Sankaran, A. Mishra, J.M. Maja, R. Ehsani, Visible-near infrared spectroscopy for detection of Huanglongbing in citrus orchards, Comput. Electron. Agric. 77 (2) (2011) 127–134, https://doi.org/10.1016/j.compag.2011.03.004.

[26] B. J. Lee *et al.*, "Discrimination and prediction of the origin of Chinese and Korean soybeans using Fourier transform infrared spectrometry (FT-IR) with multivariate statistical analysis," *PLoS One*, vol. 13, no. 4, 2018, 10.1371/journal.pone.0196315.

[27] Y. Meng, S.N. Qasem, M. Shokri, S. Shahab, Dimension reduction of machine learning-based forecasting models employing principal component analysis, Mathematics 8 (8) (2020) 1233, https://doi.org/10.3390/MATH8081233.

[28] S. Liaghat, S. Mansor, R. Ehsani, H.Z.M. Shafri, S. Meon, S. Sankaran, Mid-infrared spectroscopy for early detection of basal stem rot disease in oil palm, Comput. Electron. Agric. 101 (2014) 48–54, https://doi.org/10.1016/j.compag.2013.12.012.

[29] J. Gerretzen et al., Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments, Anal. Chem. 87 (24) (2015) 12096–12103, https://doi.org/10.1021/acs.analchem.5b02832.

[30] H.J. Butler, B.R. Smith, R. Fritzsch, P. Radhakrishnan, D.S. Palmer, M.J. Baker, Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy, Analyst 143 (24) (2018) 6121–6134, https://doi.org/10.1039/c8an01384e.

[31] D.K. Agustika et al., Gas Chromatography-Mass Spectrometry Analysis of Compounds Emitted by Pepper Yellow Leaf Curl Virus-Infected Chili Plants A Preliminary Study, Sep. MDPI 8 (5) (2021) 136, https://doi.org/10.3390/separations8090136.

[32] A. Dombrovsky, E. Glanz, M. Pearlsman, O. Lachman, Y. Antignus, Characterization of Pepper yellow leaf curl virus, a tentative new Polerovirus species causing a yellowing disease of pepper, Phytoparasitica 38 (5) (2010) 477–486, https://doi.org/10.1007/s12600-010-0120-x.

[33] C. P, S. B, and Y. S, "Begomoviruses Associated to Pepper Yellow Leaf Curl Disease in Thailand," *Open Access J. Agric. Res.*, vol. 3, no. 7, 2018, 10.23880/oajar-16000183.

[34] C. Fadhila et al., The threat of seed-transmissible pepper yellow leaf curl Indonesia virus in chili pepper, Microb. Pathog. vol. 143, no. March (2020), https://doi.org/10.1016/j.micpath.2020.104132 104132.

[35] H. J. S. Finch, A. M. Samuel, and G. P. F. Lane, "Diseases of farm crops," *Lockhart Wiseman's Crop Husb. Incl. Grassl.*, pp. 142–179, Jan. 2002, 10.1533/9781855736504.1.142.

[36] "Identifying nutritional deficiencies in backyard plants." [Online]. Available https://www.agric.wa.gov.au/identifying-nutritional-deficiencies-backyard-plants?nopaging=1. [Accessed 26-Mar-2022].

[37] J.L. Liu, P.C. Zuo, J. Sun, D. Lan, Why leaves curl with water content varied Mechanics can illustrate biology, Int. J. Mech. Eng. Educ. 43 (2) (2015) 110–117, https://doi.org/10.1177/0306419015591323.

[38] R. Gautam, S. Vanga, F. Ariese, S. Umapathy, Review of multidimensional data processing approaches for Raman and infrared spectroscopy, EPJ Techn Instrum 2 (1) (2015).

[39] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, TrAC - Trends Anal. Chem. 132 (2020), https://doi.org/10.1016/j.trac.2020.116045 116045.

[40] D. Ami, R. Posteri, P. Mereghetti, D. Porro, S.M. Doglia, P. Branduardi, Fourier transform infrared spectroscopy as a method to study lipid accumulation in oleaginous yeasts, Biotechnol. Biofuels 7 (1) (2014) 12, https://doi.org/10.1186/1754-6834-7-12.

[41] R.W. Schafer, What Is a Savitzky-Golay Filter? [Lecture Notes], IEEE Signal Process. Mag. 28 (4) (2011) 111–117, https://doi.org/10.1109/MSP.2011.941097.

[42] B. Zimmermann, A. Kohler, Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy, Appl. Spectrosc. 67 (8) (2013) 892–902, https://doi.org/10.1366/12-06723.

[43] M. Dct, Wavelet Theory and Application, Wavelet Theory Appl. (1993), https://doi.org/10.1007/978-1-4615-3260-6.

[44] E. PHAISANGITTISAGUL, "Signal Processing using Wavelets for Enhancing Electronic Nose Performance," North Carolina State University, 2007.

[45] A. Dixit, S. Majumdar, Comparative Analysis of Coiflet and Daubechies Wavelet using Global TRhreshold for Image De-Noising, Int. J. Adv. Eng. Technol. 6 (5) (2013) 2247–2252.

[46] N.A.S. Alwan, Z.M. Hussain, Image quality assessment for different wavelet compression techniques in a visual communication framework, Model. Simul. Eng. 2013 (2013), https://doi.org/10.1155/2013/818696.

[47] Q. Li, W. Wang, X. Ling, J.G. Wu, Detection of gastric cancer with fourier transform infrared spectroscopy and support vector machine classification, Biomed Res. Int. 2013 (2013), https://doi.org/10.1155/2013/942427.

[48] R. F. de Mello and M. A. Ponti, *Machine Learning A Practical Approach on the Statistical Learning Theory*. Springer International Publishing AG.

[49] C.L.M. Morais, K.M.G. Lima, M. Singh, F.L. Martin, Tutorial multivariate classification for vibrational spectroscopy in biological samples, Nat. Protoc. 15 (7) (2020) 2143–2162, https://doi.org/10.1038/s41596-020-0322-8.

[50] N.A. Almansour, H.F. Syed, N.R. Khayat, R.K. Altheeb, R.E. Juri, J. Alhiyafi, S. Alrashed, S.O. Olatunji, Neural network and support vector machine for the prediction of chronic kidney disease A comparative study, Comput. Biol. Med. 109 (2019) 101–111.

[51] T. Munakata (Ed.), Texts in Computer ScienceFundamentals of the New Artificial Intelligence, Springer London, London, 2007.

[52] D. Tafintsev, "Multivariate Classification Methods for Spectroscopic Data with, Multiple Class Structure" (2016).

[53] M. Hilario, A. Kalousis, Approaches to dimensionality reduction in proteomic biomarker studies, Brief. Bioinform. 9 (2) (2008) 102–118, https://doi.org/10.1093/bib/bbn005.

[54] A. Tharwat, T. Gaber, A. Ibrahim, A.E. Hassanien, Linear discriminant analysis A detailed tutorial, AI Commun. 30 (2017) 169–190, https://doi.org/10.3233/AIC-170729.

[55] S. Balakrishnama, A. Ganapathiraju, LINEAR DISCRIMINANT ANALYSIS - A BRIEF TUTORIAL, Department of Electrical and Computer Engineering, Mississippi State University, Institute for Signal and Information Processing, 1998.

[56] C. Chen, F. Chen, B. Yang, K. Zhang, X. Lv, C. Chen, "A novel diagnostic method FT-IR, Raman and derivative spectroscopy fusion technology for the rapid diagnosis of renal cell carcinoma serum", *Spectrochim*, Acta Part A Mol. Biomol. Spectrosc. 269 (2022), https://doi.org/10.1016/j.saa.2021.120684 120684.

[57] C. Distante, M. Leo, P. Siciliano, K.C. Persaud, On the study of feature extraction methods for an electronic nose, Sensors Actuators, B Chem. 87 (2) (2002) 274–288, https://doi.org/10.1016/S0925-4005(02)00247-2.

[58] D. K. Agustika, S. N. Hidayat, K. Triyana, D. D. Iliescu, and M. S. Leeson, "Steady-state response feature extraction optimization to enhance electronic nose performance," in *2020 7th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2020, pp. 144–149, 10.23919/EECSI50503.2020.9251887.

[59] "IR Spectrum Table & Chart." [Online]. Available https://www.sigmaaldrich.com/GB/en/technical-documents/technical-article/analytical-chemistry/photometry-and-reflectometry/ir-spectrum-table. [Accessed 27-Sep-2021].

[60] F. Zhang, X. Tang, A. Tong, B. Wang, J. Wang, Y. Lv, C. Tang, J. Wang, Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method, Spectrosc. Lett. 53 (3) (2020) 222–233.

[61] F. Zhang, X. Tang, A. Tong, B. Wang, and J. Wang, "An automatic baseline correction method based on the penalized least squares method," *Sensors (Switzerland)*, vol. 20, no. 7, 2020, 10.3390/s20072015.

[62] M. A. de Oliveira, N. V. S. Araujo, R. N. da Silva, T. I. da Silva, and J. Epaarachchi, "Use of Savitzky-Golay filter for performances improvement of SHM systems based on neural networks and distributed PZT sensors," *Sensors (Switzerland)*, vol. 18, no. 1, 2018, 10.3390/s18010152.

[63] H.L. Kennedy, Improving the frequency response of Savitzky-Golay filters via colored-noise models, Digit. Signal Process. A Rev. J. 102 (2020) 1–16, https://doi.org/10.1016/j.dsp.2020.102743.

[64] N.B. Gallagher, "Savitzky–Golay smoothing and differentiation filter", *white Pap*, Eig. Inc. (2020).