IDENTIFICATION OF CITRUS GREENING (HLB)
USING A VIS-NIR SPECTROSCOPY TECHNIQUE
A. R. Mishra, D. Karimi, R. Ehsani, W. S. Lee

ABSTRACT. Citrus greening, also known as Huanglongbing or HLB, is a major threat to the U.S. citrus industry. Currently, scouting and visual inspection are used for screening infected trees. However, this is a time-consuming and expensive method for HLB disease detection. Moreover, as it is subjective, the current method exhibits high detection error rates. The objective of this study was to investigate the potential of visible and near-infrared (VIS-NIR) spectroscopy for identifying HLB-infected citrus trees. The spectral data from infected and healthy orange trees of the Valencia variety were collected from four different orchards in Florida. Two different spectroradiometers with a spectral range of 350 to 2500 nm were used to collect the canopy reflectance spectral data. Three classification techniques were used to classify the data: k-nearest neighbors (KNN), logistic regression (LR), and support vector machines (SVM). Analysis showed that using only one canopy reflectance observation per tree was inadequate. None of the classification methods was successful in discriminating healthy trees from HLB-infected trees because of the large variability in the canopy reflectance spectral data. When five spectra from the same tree were used for classification, the SVM and weighted KNN methods classified spectra with 3.0% and 6.5% error rates, respectively. The results from this study indicate that canopy VIS-NIR spectral reflectance data can be used to detect HLB-infected citrus trees; however, high classification accuracy (>90%) requires multiple measurements from a single tree.

Keywords. Support vector machines, k-Nearest neighbors, Logistic regression, Citrus greening.

C

citrus greening, also known as Huanglongbing (HLB) (Candidatus Liberibacter asiaticus), is a systemic bacterial disease transmitted by Asian citrus psyllids (Diaphorina citri Kuwayama). The disease is causing substantial economic losses to the citrus industry by shortening the life span of infected trees. In the infected orchards, trees decline and the productive duration of fruit bearing is reduced. As no cure has been reported for citrus greening so far, removing the infected trees can reduce the spread of HLB. The elimination and removal of infected trees due to citrus canker (Xanthomonas citri subsp. citri) and HLB contributed to a gross loss of 8,061 ha (19,918 acres) in Florida (NASS, 2009). Muraro (2007) reported that the total costs of growing citrus fruit increased from $4,096 to $5,643 per ha due to HLB.

The first visible sign of citrus greening may either be leaf yellowing or mottling, which is not uniform and can affect one branch without having any effect on another (Schneider, 1968). Figure 1 shows HLB-infected and healthy leaves. Yellow, angular blotching has been considered a symptom specific to HLB disease and consists of blotches of yellow on dark greenish-gray leaves. HLB symptoms appear on leafy shoots as they grow from branches and after leaves have matured normally. By the time these symptoms are apparent, a tree can already be severely affected. Takushi et al. (2007) reported that the starch content of HLB-infected leaves can be 20 times higher than that of healthy leaves. Etxeberria et al. (2009) studied the anatomical distribution of abnormally high levels of starch in HLB-infected Valencia orange trees. They reported phloem collapse in HLB-infected leaves in addition to starch accumulation. They found multiple starch grains per chloroplast in HLB-infected leaf palisade cells, whereas healthy leaf chloroplasts had a small number of lipid inclusions and occasional smaller starch grains. They further reported that HLB-infected leaves have a corky texture due to the thicker photosynthetic cell walls.

Accurate diagnosis of HLB is essential before applying control strategies like tree removal. HLB diagnosis is difficult to make based on field observations, as the symptoms resemble nutrient deficiencies such as iron or zinc deficiency (Etxeberria et al., 2007). Electron microscopy and bioassay can be used to diagnose HLB, but these methods are time-consuming and impractical (Chung and Bransky, 2005). Molecular methods such as real-time polymerase chain reaction (PCR) based assays are used to detect the presence of HLB. However, identification of suspected plants by foliar and fruit symptoms is required by a trained scouting crew prior to real-time PCR assays. Because the current HLB detection methods are expensive, time-
correlation coefficients to characterize infestation severity. The authors used linear correlation intensity analysis and determined the coefficient of determination ($R^2$) of a ratio coefficient ($r$) at 426 nm was the highest ($r = 0.878$), and the coefficient of determination ($R^2$) of a ratio between non-symptomatic HLB-infected trees and healthy trees. Delalieux et al. (2007) used hyperspectral imaging to detect HLB by applying spectral angle mapping (SAM) and spectral feature fitting (SFF) methods. They reported that it was difficult to obtain good results with SAM and SFF because of the positioning errors in GPS ground truthing and aerial imaging, and the spectral similarity between non-symptomatic HLB-infected trees and healthy trees. Delalieux et al. (2007) used hyperspectral imaging and parametric approaches such as logistic regression, partial least squares, discriminant analysis, and tree-based modeling to detect biotic stress (apple scab; *Venturia inaequalis* (Cooke) G. Winter) in apple trees.

A method for timely detection of HLB in the field can assist growers to better manage and control the disease, resulting in significant production and economic benefits. The long-term goal of this study is to develop a ground-based method to detect HLB at the early stages of development in citrus orchards. The specific objective of this study was to investigate the possibility of identifying HLB-infected trees using VIS-NIR spectroscopy and to compare three classification techniques to analyze the spectral data.

**MATERIALS AND METHODS**

**FIELD EXPERIMENTS**

A total of 1,239 spectra were collected from 135 (80 HLB, 55 healthy) Valencia orange trees. Table 1 shows detailed information on the data used in this study. All the healthy and HLB-infected trees in this study were 15 to 20 years old. Spectral data were collected using two portable...
spectroradiometers. The first series of data were collected with a FieldSpec 3 spectroradiometer (Analytical Spectral Devices - ASD, Boulder, Colo.). Reflectance data from 350 to 2,500 nm were collected using the spectroradiometer and transmitted wirelessly to a laptop computer. The equipment was set up to use an average of ten scans to represent a single observation. Bare fiber optic cable with 25° field of view (FOV) was used for data collection. The ASD FieldSpec spectroradiometer has a circular field of view. The integration time was optimized using the optimization option within the software. Optimization values depend on the response to light in a particular spectral region. The bare fiber optic cable was placed at a distance of approximately 50 to 80 cm from the tree canopy, where the canopy was in the range of FOV of the sensor. During data collection, the area scanned by the spectroradiometer was approximately 385 to 988 cm².

The second series of data were collected using a SVC HR-1024 portable spectroradiometer (Spectra Vista Corp., Poughkeepsie, N.Y.). The spectral range of the SVC is 350 to 2,500 nm with spectral resolutions of 3.5 nm (at 350 to 1,000 nm), 9.5 nm (at 1,000-1,850 nm), and 6.5 nm (at 1,850 to 2,500 nm), which was similar to the ASD FieldSpec spectroradiometer. The SVC HR-1024 spectroradiometer communicated with a handheld PDA through a wireless connection. All the data were collected with 4° FOV. The FOV of the SVC spectroradiometer was rectangular, which covered an area of $4.8 \times 9.4$ cm from a distance of 1 m; the area was approximately 45 cm². The SVC spectroradiometer had a minimum integration time of 1 ms. Under field conditions, the scan time was set at 4 s. Laboratory tests were conducted with various objects to compare the spectral signatures of the ASD FieldSpec and SVC HR-1024 spectroradiometers. Similar results were obtained from both spectroradiometers.

Dark current measurements were automatically taken immediately prior to the reference or target scans by both the ASD FieldSpec and SVC spectroradiometers. The FOV for both sensors would only cover a small section of a branch. The spectral data were collected from sunlit canopies between 11:00 a.m. to 2:00 p.m. with no additional artificial light. Spectral reflectance was collected from tree canopies from sunlit sides of the tree canopy to maximize the incident light intensity. In the case of HLB-infected trees, the reflectance spectral data were collected only from branches with symptomatic leaves. These symptomatic leaves were marked by trained scouts and confirmed by PCR test before data collection. Reference readings of a white panel were collected every 10 min to reduce errors in reflectance caused by changes in atmospheric conditions.

Both spectroradiometers provided relative reflectance values of the tree canopy based on the reference value. The data analysis described in this article was performed in Matlab (Mathworks, Inc., Natick, Mass.).

### SPECTRAL PRETREATMENT AND FEATURE SELECTION

Multiplicative scatter correction (MSC) was performed on the NIR portion of the spectra (Martens and Naes, 2001). Then the local mean and the first and second derivatives were calculated at 25 nm intervals from 375 to 1,325 nm, from 1,500 to 1,750 nm, and from 2,050 to 2,300 nm (removing the noisy regions). At each of these 61 points, the local mean was computed by averaging the spectral reflectance values for that wavelength and its four neighbors. For example, the local mean at 375 nm was computed by averaging from 373 to 377 nm. In order to avoid noise amplification, which results from differentiation, the Savitzky-Golay method (Orfanidis, 1996) was used to compute the first and second derivatives. A quadratic polynomial and a window size of 21 were used. Since three values were computed for each of the 61 points mentioned above (i.e., the mean and the first and second derivatives), each spectrum was represented by a feature vector of 183 elements. Because the number of features was significantly high and many of these features could be correlated, principal component analysis (PCA) was performed to reduce the number of features (fig. 2). The 183 spectral elements were reduced to 25 principal components (PCs) that accounted for more than 90% of the variance within the data. From figure 2, it can be seen that the first 25 PCs contribute about 90% of the variance. A preliminary analysis was performed to see whether increasing the number of PCs would improve the classification results. No significant improvements were obtained by increasing the number of PCs beyond 25. Figure 3 shows a plot of PCA loadings for the first two principal components.
The vectors of the principal components were used as the input to the classification algorithms. The output classes obtained from the classification models were the “diseased” and “healthy” trees. For each classification method, 75% of the data were used for training, while 25% of the data were used for testing the accuracy of the classifier.

**Weighted k-Nearest Neighbors (KNN)**

Weighted $k$-nearest neighbors (KNN) is an instance-based classification algorithm. This type of algorithm does not develop an explicit function or model for predicting the target classes. Instead, all the training samples are stored, and computation is delayed until a new unknown sample must be classified. For a new sample, the Euclidian distance between its feature vector and the feature vectors of the training samples is computed. The computed distances are then used to find the $k$ training samples that are closest to the unknown sample, and a prediction is made based on these nearest neighbors. For a classification problem, this can be done simply by a majority vote. A more sophisticated approach, used in this study, was to give a different weight to the contribution of each of the $k$-nearest neighbors in the prediction. Commonly, a weight that is inversely proportional to the square of the Euclidian distance is used (Mitchell, 1997):

$$
\hat{f}(x_{\text{unknown}}) = \arg\max_{v \in V} \sum_{i=1}^{K} w_i \delta(v, f(x_i))
$$

where

$$
w_i = \frac{1}{\sum_{j=1}^{n} \left(x_{\text{unknown}}^j - x_i^j\right)^2}
$$
In equations 1 and 2, \( \hat{f} \) indicates the predicted class for the new unknown feature vector \( x_{\text{unknown}} \). \( x_{i,j} \) denotes the \( j \)th element of the \( i \)th feature vector, \( w_i \) are the weights, and the function \( \delta \) is defined as follows:

\[
\delta(a,b) = \begin{cases} 
1 & \text{if } a = b \\
0 & \text{otherwise}
\end{cases}
\]  (3)

A \( k \) of 12 was used in this study based on preliminary analysis of the data.

Logistic Regression (LR)

Logistic regression provides a model for the probability of occurrence of an event by fitting the data to a logistic curve. As shown in figure 4, this curve maps the entire real axis onto the interval \([0,1]\), making it ideal for modeling the probability of an event. When used in classification, the variable \( z \) is usually defined as a linear combination of the features (Witten and Frank, 2005):

\[
g_\theta(x) = f(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}
\]  (4)

where \( \theta \) is the parameter vector, \( x \) is the feature vector, and \( f(t) = 1 \) if \( t \geq 0 \) and \( f(t) = -1 \) if \( t < 0 \). A batch gradient descent approach was used for training the model, as shown in the following pseudo code:

\[
\text{Do until convergence } \{
\text{for all } j
\quad \theta_j = \theta_j + \varepsilon \sum_{i=1}^{N} x_{i,j} \left( y_i - g_\theta \left( x_i \right) \right)
\}
\]

where \( N \) is the number of features, \( \varepsilon \) is the learning rate, and \( y \) is the true label (either 1 or 0). A sufficiently small \( \varepsilon \) will ensure that the global optimum will be achieved, although the computation time will increase by decreasing the value of \( \varepsilon \). In this study, a value of \( \varepsilon = 0.001 \) was used.

Support Vector Machines (SVM)

Support vector machines (SVM) is one of the most successful machine learning algorithms that is widely used in various fields (Byvatov and Schneider, 2003; Rumpf et al., 2010). The basic SVM solves a classification problem with only two target classes. However, it can be generalized to solve problems that involve more than two classes. For linearly separable data, such as the data that are shown in figure 5a, one can draw many different hyperplanes that can separate the data. The goal in SVM is to find a hyperplane that separates the data with the largest possible margin, as shown in figure 5b. SVM uses numeric labels 1 and -1 to identify the two classes. In this article, the following notation is used to show the SVM classifier (Webb, 2002):

\[
y = g_w(x) = f \left( w^T x + w_0 \right)
\]  (5)

The parameters \( w \) and \( w_0 \) define the decision boundary (i.e., the separating hyperplane). The intercept \( w_0 \) is a scalar, while \( w \) and the feature vector \( x \) are \( n \)-dimensional vectors. In equation 5, \( f(t) = 1 \) if \( t \geq 0 \) and \( f(t) = -1 \) if \( t < 0 \). The equation defining the separating hyperplane is as follows:

\[
w^T x + w_0 = 0
\]  (6)

The geometric margin of the \( i \)th example from the separating hyperplane can be calculated using the following equation:

\[
D_i = y_i \left( \frac{w^T x_i + w_0}{\|w\|} \right)
\]  (7)

Here, a positive \( D_i \) would mean that \( y_i \) and \( w^T x_i + b \) are of
the same sign, or equivalently, the $i$th example is correctly classified. The geometric margin of the classifier with respect to the set of all training samples, $\{(x_{i}^{1}, y_{i}), \ldots, (x_{i}^{N}, y_{i})\}$, is defined as the minimum of the geometric margins for each sample:

$$D = \min_{i=1:N} D^{i}$$

Therefore, the goal of SVM is to find a separating hyperplane that maximizes $D$. In mathematical terms, SVM will pose the following optimization problem:

$$\begin{align*}
\max_{w, w_{0}} D & \\
\text{such that: } & y^{i} \left( w^{T} x^{i} + w_{0} \right) \geq D, & i = 1, \ldots, N \\
& \|w\| = 1
\end{align*}$$

The second constraint, $\|w\| = 1$, is a non-convex constraint, and this optimization problem cannot be directly solved. It can be shown (Gunn, 1998) that this optimization problem can be replaced by the following equivalent problem:

$$\begin{align*}
\min_{w, w_{0}} & \frac{1}{2} \|w\|^{2} \\
\text{such that: } & y^{i} \left( w^{T} x^{i} + w_{0} \right) \geq 1, & i = 1, \ldots, N
\end{align*}$$

In equation 10, $y^{i} \left( w^{T} x^{i} + w_{0} \right)$ is called the functional margin. The difference between the functional margin and the geometric margin is that multiplying $w$ and $w_{0}$ by a constant will not change the geometric margin but will change the functional margin. The functional and geometric margins are related as follows:

$$D^{i} = d^{i} / \|w\|$$

Equation 10 means that the optimal margin classifier sought by SVM can be found by minimizing the norm of $w$, under the constraint that the functional margin for all examples is at least equal to 1. It is easy to see that the minimum margin always belongs to the points that are on the edge of the widest strip that separates the data, as shown in figure 5b. Only for these points $y^{i} \left( w^{T} x^{i} + w_{0} \right) = 1$; for the remaining points, $y^{i} \left( w^{T} x^{i} + w_{0} \right) > 1$. These points are called the “support vectors” and usually comprise a very small fraction of the total number of points, resulting in a huge reduction in the computational cost.

Although the optimization problem in equation 10 can be solved by commercially available quadratic programming code, an easier equivalent problem can be found by using the method of Lagrange multipliers (Strang, 1991). Using this method, the following dual optimization problem will be obtained:

$$\begin{align*}
\max_{\alpha} & \sum_{i=1}^{N} \alpha_{i} - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_{i} \alpha_{j} y^{i} y^{j} \left( x^{i} \cdot x^{j} \right) \\
\text{such that: } & \sum_{i=1}^{N} \alpha_{i} y^{i} = 0 \\
& \alpha_{i} \geq 0, & i = 1, \ldots, N
\end{align*}$$

In equation 12, $\alpha_{i}$ are the Lagrange multipliers, and $(x^{i}, y^{i})$ indicates the inner product between the $i$th and $j$th feature vectors. The interesting point is that all of the Lagrange multipliers are zero except for the ones that correspond to the support vectors. Once the above optimization problem is solved, $w$ and $w_{0}$ can be found using the following equations:

$$\begin{align*}
w & = \sum_{i=1}^{N} \alpha_{i} y^{i} x^{i} \\
w_{0} & = -\max_{i,j=1} \frac{w^{T} x^{i} + \min_{i,j=1} w^{T} x^{j}}{2}
\end{align*}$$

It is important to note that the new optimization problem in equation 12 is only in terms of the inner products of the feature vectors. Moreover, with the new definition of $w$ in terms of Lagrange multipliers (eq. 13), for a new example $x$, the prediction of the SVM classifier can be written as follows:

$$y = g_{w}(x) = f \left( \sum_{i=1}^{N} \alpha_{i} y^{i} \left( x \cdot x^{i} \right) + w_{0} \right)$$

In other words, the entire algorithm can be written in terms of the inner products of the feature vectors.

In many practical applications, the data are not linearly separable, or even if they are, the optimal margin classifier described so far may not be the best choice because it can be very sensitive to outliers. Therefore, a modified version of the SVM problem is defined as follows (Webb, 2002):

$$\begin{align*}
\min_{w, w_{0}} & \frac{1}{2} \|w\|^{2} + C \sum_{i=1}^{N} \xi_{i} \\
\text{such that: } & y^{i} \left( w^{T} x^{i} + w_{0} \right) \geq 1 - \xi_{i}, & i = 1, \ldots, N \\
& \xi_{i} \geq 0, & i = 1, \ldots, N
\end{align*}$$

With this definition, some of the examples are allowed to fall on the wrong side of the separating hyperplane. However, for each such example, a cost ($\xi$) will be considered. The parameter $C$ determines the importance that we attach to these errors. Once again, the method of Lagrange multipliers can convert this problem into an easier one. The resulting problem is as follows:
The parameters $C$ and $\gamma$ in the problem are unknown. Therefore, the first step in solving the problem is to find the optimum values for these parameters. The grid search procedure suggested by Hsu et al. (2008) was used for this purpose. First, the approximate values of the optimum $C$ and $\gamma$ were found on a coarse grid with $C = 2^{-8}, 2^{-4}, \ldots, 2^{30}$ and $\gamma = 2^{-14}, 2^{-10}, \ldots, 2^{24}$. After finding the approximate values for the best $C$ and $\gamma$, a finer grid was defined to search for more accurate values of the optimum $C$ and $\gamma$.

### Reducing the Classification Error by Using Multiple Measurements

The goal of the present study was to evaluate classification algorithms that can be used for automatic detection of HLB-infected citrus trees in the field. The preliminary results indicated that the classification accuracy from the statistical methods used to classify HLB-infected trees were low when data with single measurements were considered for analysis (data not shown). This could be due to variability in the experimental conditions, such as the amount of sunlight, orientation of the leaves with respect to the sensor, and sensor measuring angle. Therefore, it was hypothesized that high detection accuracy can be achieved through multiple measurements on a single tree. In this study, we investigated the accuracy of the three classification methods (KNN, LR, and SVM) using spectral data from one, three, or five measurements from different canopy areas of the same tree.

When more than one measurement was presented to the classifier, a simple method based on majority voting can be used to determine the classification group. The tree is labeled an HLB-infected tree when more than half of the measurements from that tree are classified as HLB infected. An improved approach is to consider the confidence of the classifier for its prediction on each of the measurements. For example, if a classifier predicts “infected” on two out of five measurements with very high confidence and “healthy” on the three remaining measurements with very low confidence, it would be more reasonable to classify the tree as “infected.” Our analysis was based on the majority with confidence of the classifier.

For the KNN algorithm, the following equation was used to estimate the confidence of each prediction:

$$
\text{KNN confidence} = \sum wi \quad \text{(for those among the } k \text{ nearest neighbors that have the same class as the predicted class)}
$$

\[+ \sum wi \quad \text{(for all of the } k \text{ nearest neighbors)}
\]

where $w_i$ is as defined in equation 2. For the LR method, the value of the function $g_\theta(x)$ in equation 4 can be used to calculate a measure of confidence in prediction. The closer $g_\theta(x)$ is to either 0 or 1, the more confident is the prediction. Therefore, the value $|g_\theta(x) - 0.5|$ was used to evaluate the confidence of the predictions by the LR method:

$$
\text{LR confidence} = |g_\theta(x) - 0.5| \quad (21)
$$

For the SVM classifier, the value of the function $g_\alpha(x)$ in equation 14 can be used as a measure of confidence in pre-
prediction. The larger the absolute value of this function, the more confident is the prediction. Therefore, $|g_w(x)|$ was used to evaluate the confidence of predictions by the SVM method:

$$\text{SVM confidence} = |g_w(x)|$$

(22)

**RESULTS AND DISCUSSION**

Figure 6 shows an example of the collected spectra for both healthy and HLB-infected leaves. Each observation is an average of ten measurements. In general, the reflectance in the visible region (400 to 700 nm) of HLB-infected leaves was higher than that of healthy leaves. This finding is accordance with Poole et al. (2008) and supports our hypothesis that due to the chlorosis in HLB-infected trees, they absorb less visible light, resulting higher reflectance in the visible region. In the NIR region, the canopy reflectance of HLB-infected leaves in most cases was lower than that of healthy leaves.

For the SVM method, the first step was to find the optimum values for $C$ and $\gamma$. As mentioned before, this was performed by a grid search. Figure 7 shows typical contour graphs that were used to find the optimum parameter values. First, the classification error was evaluated on a coarse grid (fig. 7a) to find approximate estimates for $C$ and $\gamma$. Then a finer grid was defined (fig. 7b) and used to determine accurate values of $C$ and $\gamma$ that minimized the classification error. In the specific case shown in figure 7, the lowest classification error was approximately 18%, which was obtained for $C = 28$ and $\gamma = 34$.

Analysis showed that MSC correction of the spectra did not improve the classification accuracy. In other words, when MSC, as described earlier in the Spectral Pretreatment and Feature Selection section, was removed as a pretreatment procedure, the classification error did not improve for any of the three classification techniques. Table 2 shows the classification error for each of the three classification techniques without MSC correction. The table shows the percentage of misclassified citrus trees when one, three, or five spectral measurements from each tree were used for classification. In addition to the overall error, the table also shows false positive rate (i.e., the percentage of healthy trees that were classified as HLB infected) and the false negative rate (i.e., the percentage of HLB-infected trees that were classified as healthy).

It can be seen from table 2 that the classification error with a single spectrum was relatively large. However, when three or five spectra from the same tree were used, the classification error decreased. The SVM method demonstrated lower classification errors than the other two methods, especially with five spectra. For the SVM and KNN methods, the false negative error rate was slightly higher than the false positive error rate, whereas for the LR technique, the reverse was true. Because the amount of data was limited, it was decided to build the classifiers using the pooled data from all seven field experiments. As previously mentioned, 75% of the entire dataset was used to build each classifier, and the rest of the data were used to evaluate the classifiers. The errors mentioned in table 2 are the average classification errors for the three classification techniques.

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Logistic Regression</th>
<th>$k$-Nearest Neighbors</th>
<th>Support Vector Machines</th>
</tr>
</thead>
<tbody>
<tr>
<td>One measurement</td>
<td>False positive (%)</td>
<td>37</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>False negative (%)</td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Overall error (%)</td>
<td>35</td>
<td>23</td>
</tr>
<tr>
<td>Three measurements</td>
<td>False positive (%)</td>
<td>25</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>False negative (%)</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Overall error (%)</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>Five measurements</td>
<td>False positive (%)</td>
<td>20</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>False negative (%)</td>
<td>17</td>
<td>7.1</td>
</tr>
<tr>
<td></td>
<td>Overall error (%)</td>
<td>19</td>
<td>6.5</td>
</tr>
</tbody>
</table>
tion errors for the testing data from all seven field experiments. The classification error can also be computed for the testing data from each field experiment separately. Our analysis showed that the classification error was very close for different field experiments. For field experiments that provided more data (i.e., 13 June 2007 and 24 March 2009), the classification error was slightly less than that of the field experiments that provided less data. We assume that this was because more spectra from the field experiments with a larger amount of data were used in building the classifier, thus biasing the classifier toward the pattern in the data from those field experiments.

The classifiers obtained by the weighted KNN, logistic regression, and SVM algorithms can be programmed on a microcontroller. The three algorithms were also very different in terms of computation time. For SVM-based classification, the most time-consuming step was to find the optimum values for \( C \) and \( \gamma \). The grid search method for optimal value selection used in this study took several hours (about 8 h) on a PC. Although there are faster methods for finding \( C \) and \( \gamma \) (Hsu et al., 2008), this step is inherently time consuming. The batch gradient descent algorithm described previously for the logistic regression method took approximately 12 s to complete on a PC with a 4.8 GHz processor and 512 MB of RAM. As mentioned before, the KNN method does not produce any model from the data. Therefore, for the KNN algorithm, there is no computation until a new spectrum is to be classified. Once the model (i.e., the classifier) is obtained, the computation time for classification of a new spectrum is very fast for SVM and logistic regression. On the same PC as mentioned above, the time to make a prediction based on three spectra from the same tree was only 0.06 ms for the logistic regression classifier, whereas for the weighted KNN and SVM methods, this time was 5.5 and 5.0 ms, respectively. Therefore, the logistic regression method was computationally much faster. However, because the KNN and SVM algorithms yielded better classification accuracy than logistic regression, they are preferable algorithms for HLB detection in citrus groves.

**CONCLUSION**

The goal of this study was to evaluate machine learning techniques for rapid detection of HLB-infected citrus trees. Canopy reflectance spectra were measured on infected and healthy trees using a spectroradiometer, and three common classification algorithms (logistic regression, \( k \)-nearest neighbor, support vector machines) were used to distinguish the infected trees from the healthy trees. The results indicated that a single measurement was insufficient for accurate detection of the infected trees with the three approaches used in this study. The classification error was between 18% and 35% using a single spectrum. However, using multiple spectral measurements from the tree canopy of a single tree, the classification accuracy increased. The SVM method showed an accuracy greater than 97% when it was provided with five spectra from different leaves on the same tree. Under real field conditions, variation of sunlight and other environmental factors can produce noise that might reduce the classification accuracy. Under these conditions, multiple measurements will be necessary to ensure acceptable classification accuracy.

**ACKNOWLEDGEMENTS**

This research was supported under Project No. 81086 of the Florida Citrus Production Research Advisory Council, Peter McClure, Chairman. The authors would like to thank Dr. Sindhuja Sankaran and Sherrie Buchanon for their assistance in this study.

**REFERENCES**


