Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Hyperspectral imaging for identification of Zebra Chip disease in potatoes

Abhimanyu Singh Garhwal [a,*], *Reddy R. Pullanagari* [b], *Mo Li* [b], *Marlon M. Reis* [c], *Richard Archer* [a,d]

[a] School of Food and Advanced Technology, Massey University, Palmerston North 4410, New Zealand
[b] Massey AgriFood Digital Lab, Massey University, Palmerston North 4410, New Zealand
[c] AgResearch Ltd., Palmerston North Campus, 4410 New Zealand
[d] The Riddet Institute, Massey University, New Zealand

ARTICLE INFO

Article history:
Received 14 March 2020
Received in revised form
28 June 2020
Accepted 6 July 2020

Keywords:
Solanum tuberosum
Candidatus Liberibacter *solanacearum*
Spectroscopy
Partial Least Squares Discriminant Analysis (PLS-DA)
Variable importance in projection (VIP)

A Zebra Chip (ZC) disease detection system was developed based on hyperspectral imaging (HSI) to minimise economic losses in the New Zealand potato chip industry. Current detection methods for other than heavily diseased tubers require peeling or cutting of potato tubers. A rapid and non-destructive grading method would be ideal to remove ZC diseased potatoes at line before processing. The spectral signatures from a large population (n = 3352) of commercially sourced potatoes were collected using HSI in the spectral range of 550 nm–1700 nm. Spectral signatures of each potato (i.e. 1767 ZC infected and 1585 healthy potatoes) were extracted by segmentation and morphological operations. A calibration dataset (80% of the total population was randomly selected), with and without pre-processing, was used for modelling using the partial least squares discriminant analysis (PLS-DA). The model performance shows 92% accuracy for ZC potato identification on validation data (20% of total population). Waveband optimisation by variable importance in projection (VIP) method revealed 34 wavebands sensitive to ZC diseased potatoes. This optimum set of wavebands allowed ZC identification with 89% accuracy. The experiments demonstrate the potential of HSI for identification of ZC infected potatoes in whole tuber before processing. Efficient removal of diseased tubers would reduce processing losses and provide a potential opportunity to access export markets for intact tubers.

1. Introduction

The potato *(Solanum tuberosum)* industry contributes to the New Zealand economy an approximate one billion New Zealand dollars annually (Anonymous, 2019). Potato is the largest vegetable crop in New Zealand. The potato chip (note: Throughout this paper, the terminology used is in US English i.e. potato chip (US English) which is same as potato crisp (UK English)) industry is economically impacted by a disease named Zebra Chip (ZC). The ZC disease is caused by the bacterium *'Candidatus* Liberibacter solanacearum *(Lso)* or *Candidatus* Liberibacter psyllaurous *'* and causes economic loss to the New Zealand economy. First discovered in Mexico in 1994 it has spread to New Zealand and many other countries (Munyaneza, 2012). It causes a decrease in crop yield because of quick death of the plant and/

* *Corresponding author.*
E-mail address: A.S.Garhwal@massey.ac.nz (A.S. Garhwal).
https://doi.org/10.1016/j.biosystemseng.2020.07.005
1537-5110/

or impaired growth (Buchman, Fisher, Sengoda, & Munyaneza, 2012). Furthermore, ZC disease dramatically increases the level of reducing sugars (glucose and fructose) in potatoes (Kumar, Knowles, & Knowles, 2015). This results in dark brown colour strips in potato chips during frying because of the chemical reaction between reducing sugars and amino acids known as the Maillard reaction (Dyer et al., 1991). Potato chips with dark stripes are unacceptable to customers and consequently cause significant economic loss to the potato chip industry. This has increased the requirement for a technique for segregating ZC diseased potatoes from healthy ones. Traditional destructive methods like manual inspection of cut potatoes for disease detection have been widely used in the industry.

Standard lab-based methods are also available to detect ZC infection in potatoes including polymerase chain reaction (PCR) and LC-MS analysis for phenolic compounds (Levy, Ravindran, Gross, Tamborindeguy, & Pierson, 2011; Navarre, Shakya, Holden, & Crosslin, 2009) etc. Detection of certain phenolics by LC-MS would provide greater confirmation but is not definitive of ZC disease. Confirmation of the presence of DNA from the infective agent by PCR is definitive. However, there is no standard approach based on non-invasive assessment.

Hyperspectral imaging (HSI) is a fast, non-destructive and non-contact method that has been recognised for grading food commodities (Feng & Sun, 2012). The biggest advantage of HSI is that it captures both spatial and spectral information simultaneously (Elmasry, Kamruzzaman, Sun, & Allen, 2012). Physical properties like colour change of potato tubers during processing (Amjad, Crichton, Munir, Hensel, & Sturm, 2018), and textural attributes of potato tubers such as hardness, cohesiveness, gumminess upon eating were predicted by employing HSI (Su, Bakalis, & Sun, 2018). Other researchers have attempted to estimate chemical constituents including starch and cellulose (Su & Sun, 2017), and sugars such as glucose, fructose and sucrose (Ayvaz, Santos, Moyseenko, Kleinhenz, & Rodriguez-Saona, 2015) using HSI. Potatoes with external defects such as mechanical damage, common scab etc. (Riza, Suzuki, Ogawa, & Kondo, 2017) can be segregated by using HSI. Furthermore, HSI has been used to identify internal defects such as bruised tubers (López-Maestresalas et al., 2016), and hollow heart disease (Huang et al., 2015). Zhao et al. (2018) concluded from their study that ZC infection was not detected by infrared and near infrared spectra in pre-storage and post-storage potatoes. It indicates that a model fitted with reflectance measurements at four wave bands (i.e. 580, 582, 680 and 720-nm) showed a high level of accuracy for classification of infected and uninfected tubers, when infections occurred earlier in the season. However, 43% false classification also occurred, which reflected the inaccuracy of the model in classifying tubers infected later in the season. Another study has been reported focussed on identifying ZC disease in potatoes by using near infrared spectroscopy (NIRS) (Liang et al., 2018). Garhwal et al. (2020) did some preliminary spectral separation work for ZC potato identification and for detecting tampered potatoes (mimicking ZC disease) by using HSI (Garhwal, Pullanagari, Li, Archer, & Reis, 2019).

Our study investigates the use of HSI applied to detect ZC disease in a large population of 'Russet Burbank' commercial potatoes. 'Russet Burbank', is a popular and common potato cultivar. It has few eyes and the skin colour is dark brown (Bethke et al., 2014). It is used for potato chips, french fries, baking and mashing (Schlosser, 2001) and is a widely grown potato variety in the South Island of New Zealand. The objectives of this study were:

- To test the potential of HSI for segregating ZC infected potatoes.
- To identify the most important spectral bands associated with ZC disease.

The novelty of this research lies within (1) analysing visible region spectral data captured by HSI sensors to detect ZC diseased potatoes that results in development of cheap hyperspectral cameras for industrial scale ZC detection (2) the use of a very big dataset of hyperspectral images of 3352 potatoes for experimentation capturing the wide range of variability observed (3) includes hyperspectral images having additional spatial information in visible and near infrared (NIR) wavebands from 550 nm to 1700 nm for ZC detection (4) estimation of variable importance in projection (VIP) scores for selecting important wavebands for ZC identification in potatoes. To the best of our knowledge these have not been ventured for ZC potatoes grading.

2. Material and methods

2.1. Sample preparation

A total of 3352 'Russet Burbank' commercially grown potato tubers were harvested from the Canterbury region in the South Island of New Zealand in April 2019, for inspection in this research. To acquire an optimal number of diseased potatoes, a highly ZC infected field area was identified by an agronomist based on external symptoms of the plants. Tubers were collected from the pre-selected farm and sent to a commercial packhouse for washing and drying. These potatoes were then transported to Massey University, Palmerston North, New Zealand.

While the PCR is a standard method to detect ZC disease, the application of this method for the number of tubers investigated in this study was found to be prohibitive. Alternatively, we used a combination of approaches: 1) selected tubers from harvest known to have a high likelihood of infection; 2) inspected each individual potato; and 3) confirmed the likelihood of disease by frying a sample of potatoes.

Potatoes were peeled to identify signs for the ZC disease. Visual inspection by cutting potatoes was performed to identify ZC infection presence. Those displaying ZC disease signs were assigned to the infected group. In addition to this procedure, a subgroup of those classified as infected where fried to confirm the validity of our approach. Since the visual inspection method is subjected to bias, deep frying can support consistent performance in the validation.

2.2. Hyperspectral imaging set up

The HSI setup comprised a moving conveyor belt, two illumination sources, one VIS-NIR hyperspectral pushbroom camera (Headwall Photonics, Massachusetts, USA) and a computer.

The images were acquired using Hyperspec® III software (Headwall Photonics, Massachusetts, USA). The conveyor belt was driven at 32 mm s^{-1}. Two illuminator reflectance lamps (70 W, Malvern Panalytical Ltd, Malvern,UK), set at an angle of 45°, were employed to illuminate the moving tubers. The reference tile used (Zenith Polymer® Diffusers, SpherOptics GmbH, Germany) was stationed at the middle of the camera focus where the intensity of light approximated 85% of saturation level (11,000 au (arbitrary unit meaning no unit)). The system captures 235 wavebands in the spectral range from 550 nm to 1701 nm with a spectral resolution of 5 nm.

Batches of eight washed and dried potatoes with their respective labels were arranged on rectangular black trays for hyperspectral image collection. These trays were placed on the conveyor to pass under the camera. As shown in Fig. 1, the workflow includes four steps to identify ZC diseased potatoes using HSI. These four steps are:

Step 1. *Hyperspectral Image acquisition*: - A total of 419 hyperspectral images were collected using HSI system. Each hyperspectral image contained eight potatoes, a random mix of healthy and diseased.

Step 2. *Visual assessment of potatoes:* - Even mildly infected ZC potato had prominent tiny brown dots easily visible to the naked eye. Additional verification was performed by frying slices from a randomly selected subset of potatoes. This confirmed that potatoes judged not infected by manual inspection were not subsequently revealed to be infected via the characteristic zebra stripe discolouration induced by frying.

Immediately after the hyperspectral images were collected, the potatoes were peeled by trained staff to identify symptoms of ZC disease. If symptoms of ZC were found, the tuber was labelled as "diseased" otherwise as "healthy". From this visual inspection, 1767 potatoes tubers were assessed as diseased, and the remaining 1585 tubers as healthy. Out of 3352 potato tubers, 364 tubers were fried to validate the ZC disease severity. In this case, tubers were sliced into 5–10 mm thick slices and deep fried in canola oil for 2 min at a temperature of 100 °C.

Step 3. *Image processing:* - Collected hyperspectral images underwent a series of image processing steps to extract spectral signatures of individual potatoes as shown in Fig. 1.

Step 4. *Spectral Analysis:* - Extracted spectral data underwent a sequence of pre-process steps before development of a classification model. Waveband selection was then performed to obtain an optimum model.

2.3. Image processing

Image processing was the third step for ZC detection system as shown in Fig. 1. The total processing time for image processing was 27.11 s i.e. 1.4 s for calibration, 1.13 s for segmentation, 1.58 s for individual potato extraction and 23 s for region of interest extraction. That is not fast enough for commercial grading of potatoes and potentially affected the final throughput. Our aim is to provide a proof of concept for potato ZC detection by hyperspectral imaging. To decrease the image processing time is out of the scope of the paper that can be mitigated to meet industrial processing speed, by using parallel processing software and a dedicated hardware such as FPGA.

Image calibration was performed with a dark reference (D) collected with the lens of the camera covered to detect any effect of dark current on the camera sensor and a white reference (W) collected from a standard white plate (Zenith Polymer® Diffusers, SpherOptics GmbH, Germany). Raw intensity values of hyperspectral images were converted to reflectance values using the following Eq. (1):

$$R = \frac{I_{Tuber\ sample} - I_D}{I_W - I_D} \quad (1)$$

where, I_D corresponds to the intensity value for the dark reference, I_W is the intensity value captured on the white reference tile, $I_{Tuber\ sample}$ represents the intensity value of the tuber sample and R is the absolute reflectance.

2.3.1. Segmentation

An in house customised code was developed in Python for automatic morphological operations following thresholding for segmentation. In the first part of segmentation the background from hyperspectral images was removed using a threshold value of 0.4 at 671 nm. In the second stage, unwanted background strips were removed using a threshold value of 0.7 at 558 nm, and the third threshold value of 0.6 employed at 1396 nm to remove the labels from the image. Following this process, morphological operations including erosion, dilation, opening and closing of an image were performed to remove leftover spatial noise in the image to obtain the final segmented image separating tubers from the background.

2.3.2. Individual tuber extraction

Due to variability in the size and shape of tubers there was variability in the borders of the images due to variation in the illumination in these regions. To reduce this negative effect, preliminary analysis investigated the sizes of region of interest (ROI) and found that 20 × 20 ROI at the centre provided consistent information across the samples. The size of one pixel was 0.52 mm by 0.52 mm for each captured tuber image by HSI. The dimensions of the ROI used for spectral extraction was a 20 by 20 square. The actual physical size of the ROI was obtained by multiplying size of one pixel to the dimension of ROI ((20 × .52 mm) × (20 × .52 mm)) which result in 108.16 mm^2. Considering, these facts a 20 × 20 ROI was used to extract information from centroid of each potato to calculate a mean spectrum for the same tuber.

2.4. Multivariate data analysis

The last and fourth step for ZC detection system is named as spectral analysis in Fig. 1. The spectral data from hyperspectral images were processed and analysed using multivariate statistical methods in Spyder (The Scientific Python Development Environment, MIT, Cambridge, Massachusetts, USA) and RStudio (Version 1.1.463, Boston, MA, USA).

In this study we assume that the effect of ZC infection could affect reflectance by changes in absorption and scattering properties of the potatoes. Thus, we have considered both reflectance and absorbance when investigating the best pre-processing method.

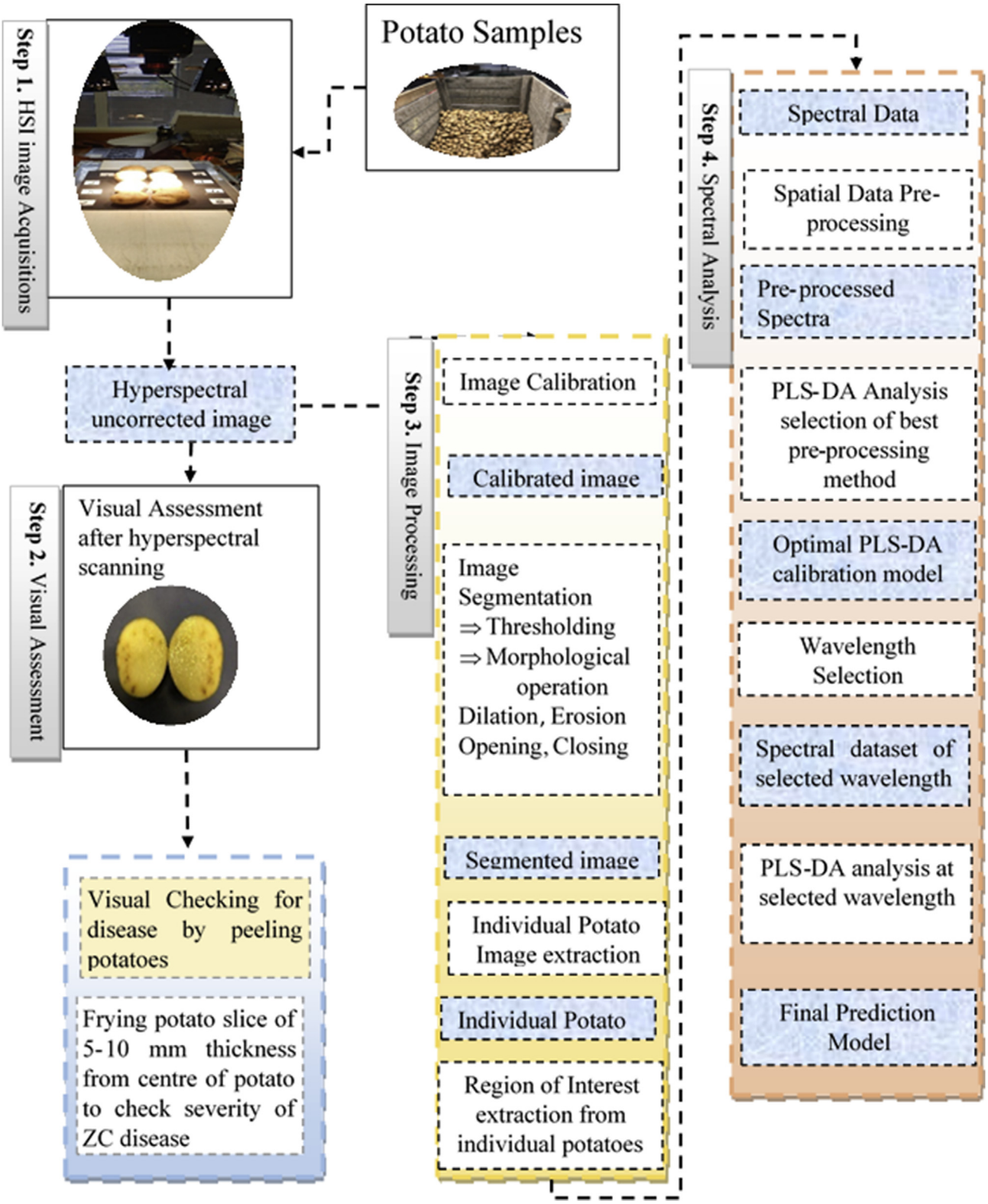

Fig. 1 – Flowchart for potato ZC disease identification using hyperspectral imaging.

Reflectance values were restricted to the spectral range between 600 nm and 1650 nm by removing 10 wavebands from both ends (i.e. from 550 nm to 595 nm, and 1655 nm–1700 nm) due to low signal to noise ratio in those extreme regions. Log transformation (LT) was performed on the reflectance spectra to convert them into absorbance values. Then standard normal variate method (SNV) was applied to reduce the effect of light scattering on log transformed data (Rinnan, Berg, & Engelsen, 2009). Derivative transformation with a window size of 11, polynomial order of

3 and a first or second derivative, was applied to the SNV spectral data (Savitzky & Golay, 1964). This operation removes the baseline offsets (Burger & Geladi, 2007). Different combinations of these pre-processing methods were evaluated for selection of the best for optimal model development.

Partial least squares discriminant analysis (PLS-DA) was used to develop models for segregating diseased and healthy tubers (Want & Masson, 2011). PLS-DA was performed by finding a link between two matrices X (containing the spectra) and Y (representing the classes of diseased and healthy). PLS generates orthogonal linear combinations of original spectra that contain maximum covariance between X and Y variables known as Latent Variables (LVs) (Höskuldsson, 1988; Wold, Sjöström, & Eriksson, 2001). Each column of the matrix Y represented one class (i.e. one column for diseased and another column for healthy class). Once the model was developed and applied to a new spectrum it generated one prediction for each class. These predictions were then compared to thresholds identified during development of the model and provided discrimination among classes to identify which class the new spectrum was allocated to (Brereton & Lloyd, 2014).

2.4.1. *Waveband selection using Variable Importance in Projection (VIP) scores of PLS-DA*

Variable importance in projection (VIP) was used to identify the most important wavebands for the classification of tubers (Wold, Sjöström, & Eriksson, 1998) (L. Eriksson, Johansson, Kettaneh-Wold, & Wold, 2001). Wavebands having VIP score values greater than one are considered as very important variables in classification (Mehmood, Martens, Sæbø, Warringer, & Snipen, 2011). The same strategy was used for selecting significant wavebands. A PLS-DA model's performance was evaluated on these selected wavebands, for identification of ZC infected potatoes.

Spectral data processed with different combinations of wavebands and models were developed for wavebands selection based on VIP scores.

(i) Three sets of wavebands were selected based on three models developed with the 235 wavebands as follows: -
   a. Without pre-processing spectral data of 235 wavebands
   b. With pre-processing of spectral data with LT, SNV and first derivative transform (FDT).
   c. With pre-processing of spectral data with LT, SNV and second derivative transform (SDT).

(ii) Another three sets of wavebands were selected on the basis of models developed with the reduced 215 wavebands spectral data as follows: -
   a. Without pre-processed spectral data of 215 wavebands
   b. With pre-processing of spectral data with LT, SNV and FDT.
   c. With pre-processing of spectral data with LT, SNV and SDT.

(iii) Common waveband selection: From the above six sets of selected wavebands, common wavebands were selected that fell in at least four sets of selected wavebands.

(iv) Below 1000 nm: Wavebands were selected for the range below 1000 nm to test the applicability of low-cost hyperspectral cameras such as the Specim FX 10 (Specim, Finland), which are readily available and may be suited to the industrial environment, but are limited wavebands between 400 and 1000 nm.

2.5. *Model validation*

The total set of tuber samples was divided randomly into training (80%) and test sets (20%) for supervised classification (PLS-DA). A training dataset was used for model development. The model performance was evaluated on the test data set to measure the capability of identification of ZC infected potatoes. The number of latent variables for PLS-DA models was selected using four-fold cross validation (Westerhuis et al., 2008).

The metrics used for measuring the developed model efficiency were accuracy, sensitivity and specificity (Parikh, Mathai, Parikh, Chandra Sekhar, & Thomas, 2008).

2.5.1. *Accuracy*

It determines the difference between true and calculated values. The proportion of potato samples classified correctly was determined by accuracy.

2.5.2. *Sensitivity*

Binary classification performance was estimated by sensitivity. It is expressed as the fraction of the actual positives that are correctly identified (total positive classified/total actual positive in the population). In this paper, binary classification was determined in terms of diseased (positive) and healthy (negative) potatoes. Our aim was to identify diseased potatoes from a group of potatoes. Hence, sensitivity measures the part of diseased potato samples correctly classified as diseased potatoes.

2.5.3. *Specificity*

Another measurement for binary classification performance was the fraction of negatives identified (total negatives classified/total actual negative in the population). The aim of the specificity in this paper was to tell the fraction of healthy potato samples correctly classified as healthy potatoes.

The values of sensitivity and specificity lie in between 0 and 1. Sensitivity and specificity values nearer to 1 indicate good performance of model.

3. Results and discussions

3.1. *Mean raw and second derivative spectra*

The mean spectra of 3352 (1589 healthy and 1763 diseased) potatoes were obtained and plotted (Fig. 2(a)) to inspect any probable variation among healthy and ZC infected potatoes. Additionally, analysis was performed to detect difference between processed spectral data (Fig. 2(b)) and raw spectral data (Fig. 2(a)) of healthy and ZC infected potatoes.

The average spectrum of each group (1589 healthy and 1763 diseased) was obtained and are presented in Fig. 2(a). The solid line shows the average spectrum corresponding to diseased potato and the dotted line corresponds to healthy

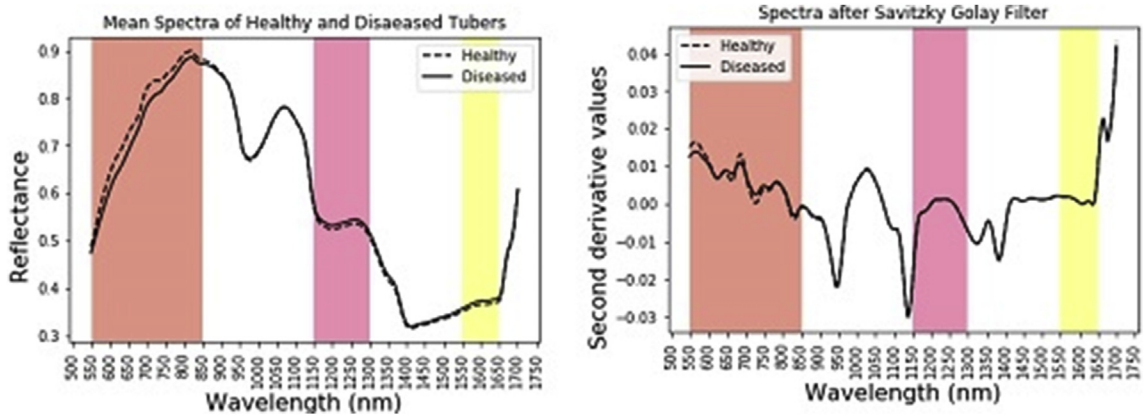

Fig. 2 – Mean Spectra of 1763 diseased (solid line) and 1589 healthy tubers (dotted line) (a) Raw Mean Spectra (b) Second derivative of Raw Mean Spectra.

potato spectra. From Fig. 2(a) most of the raw spectra of healthy and diseased potatoes overlap except for the wavebands between 550 nm and 850 nm, 1150 nm and 1300 nm; 1550 nm and 1650 nm. Second derivative was applied to these spectra to reduce effect of light scattering (by eliminating linear baseline shifts) and to enhance the effect of spectral bands originally observed as shoulders in the reflectance spectra. The difference in spectra from 550 nm to 850 nm remains after applying second derivative. The difference in spectra for the other spectral regions vanishes after taking the second derivative as shown in Fig. 2(b) suggesting that difference observed Fig. 2(a) could be associated with difference in scattering properties between the two groups.

3.2. *Deep frying validation*

Figure 3 illustrates different levels of infection. The highest level of ZC infection is represented by Fig. 3(a), where dark brown strips due to phenolic oxidation and excessive Maillard browning can be seen all over the slice. Figure 3(b) has a mild level of infection which is verified by the dark brown pigments only at the corner of the slice. This dark brown colouring makes chips highly unacceptable to consumers. Figure 3(c) represents a healthy potato slice with no browning at all.

The results show that all samples classified as infected and fried displayed the signs of the disease thus validating the procedure. Several levels of infection (Fig. 3) were observed from low to high and during model testing all diseased were labelled. Deep frying provided additional validation of the manual identification of diseased potatoes, but no attempt was made to measure the infection severity. To study the different level of infections (severity) is out of the scope of this paper.

3.3. *Model performance*

A series of spectral pre-processing combinations were evaluated (data not shown) and the best combination was utilised for spectral data analysis.

The best combination found first used log transformation (LT) of reflectance. After that Standard Normal Variate (SNV) and then second derivative (SDT) or first derivative (FDT) and mean centring (MC) were effective. This pre-processing combination was performed uniformly and in the same sequence each time.

Extracted spectral data with 235 wavebands were utilised to develop three PLS-DA models. The spectral data were pre-processed with LT+SNV+SDT+MC and LT+SNV+FDT+ MC respectively. Eleven latent variables were selected based on cross validation to develop the final PLS-DA model having cumulative explained variance of 99.98%. Maximum accuracy was achieved by using LT+SNV+SDT+MC pre-processed spectral data i.e. around 91% with highest specificity (91%) and sensitivity (91%) during calibration.

The results for the test dataset, Table 1, show that without any pre-processing of spectral data, 87% accuracy was observed, and it was increased between 3% and 5% after pre-processing. Also, when the combination of pre-processing

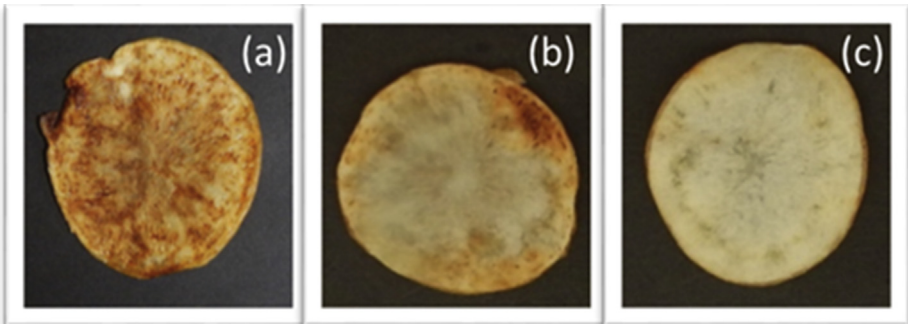

Fig. 3 – Slices with and without ZC disease (a) for severely diseased tuber (b) for mildly diseased tuber and (c) for healthy tuber.

Table 1 – Model performance comparison in ZC diseased potato identification using different pre-processing technique combinations with different number of wavebands.

Pre-Processing	Number of Wavebands	PLS-DA on Full Spectrum with 11 latent components					
		Cross Validated Calibration			Test		
		A	Sen	Spe	A	Sen	Spe
No Pre-processing	235	89	89	89	87	87	87
LT+SNV+FDT+MC	235	90	90	90	90	90	90
LT+SNV+SDT +MC	235	91	91	91	92	92	92

A: Accuracy, Sen: Sensitivity, Spe: Specificity, FDT: Savitzky Golay with first derivative, SDT: Savitzky Golay with second derivative, SNV: for standard normal variate, LT: Log Transformation, the numbers are % values.

LT+SNV+SDT+MC was used the maximum accuracy of 92% was obtained for the test dataset, and highest specificity and sensitivity of 92% (Table 1).

Furthermore, no considerable difference was observed in accuracy, sensitivity and specificity after keeping the same pre-processing except using first derivative rather than second derivative. A similar procedure was performed after removing the first and last 10 wavebands from the 235 wavebands i.e. 215 wavebands were used to develop three more PLS-DA models. The results were similar to the results obtained with 235 wavebands, so these results are not reported in this paper (See Appendix A: Tables 4 and 5 and represented in Appendix Table 4).

3.4. *Selecting important bands for discriminating potatoes*

VIP scores indicate the importance of each waveband in differentiating healthy and diseased tubers.

Three different sets of wavebands were selected based on VIP score value equal to or exceeding one. These sets of wavebands were extracted from three PLS-DA models developed above using the spectral data of all 235 wavebands. A model developed using no pre-processed spectral data had 46 wavebands with VIP scores values greater than or equal to one (Fig. 4). Similarly, the models developed by using pre-processed spectral data with LT+SNV+SDT+ MC and LT+SNV+FDT+ MC had 77 and 72 wavebands, respectively.

The classification performances of selected wavebands was checked by developing another three calibration models using PLS-DA with and without pre-processing the spectral data. Only best results were reported in Table 2, i.e. 46 wavebands achieve 91% classification accuracy with no pre-processing of spectral data. Similarly, the best results were reported in the case of 77 and 72 selected wavebands.

The same model development and waveband selection process was performed on the truncated 215 waveband spectral data. The number of wavebands selected were 47, 61 and 59. The results for ZC potatoes identification were similar to the results of 235 wavebands selected wavebands spectral data, so these results are not reported in this paper (See Appendix A: Table 5) and revealed in Appendix Table 5.

For further optimisation common wavebands were selected from 235 wavebands spectral data (i.e. 46, 77 and 72 wavebands) and 215 wavebands spectral data (47, 61 and 59 wavebands). That resulted in 40 common wavebands of which 25 lie in the visible region and 15 in the near infrared region (NIR). Different combinations of spectral data pre-processing were employed on these 40 waveband subsets of spectral data for model development and only the best results were reported. PLS-DA models were developed and tested on ZC potato identification. The accuracy for identification of ZC diseased potatoes was 89% (Table 3), similar to the accuracy of the earlier model with 46, 77 and 72 waveband subsets of spectral data. It appears that 40 wavebands were sufficient to identify the ZC defected potatoes without compromising accuracy.

A final selection was performed by choosing spectral data from 40 wavebands by removing 6 selected wavebands above 1000 nm i.e. 1346 nm, 1351 nm, 1356 nm, 1371 nm, 1376 nm, 1381 nm to test the potential suitability of less capable and less expensive hyperspectral cameras. That gave 34 wavebands below 1000 nm where 25 lie in the visible region (i.e. 607 nm, 611mn, 616 nm, 626 nm, 636 nm, 652 nm, and from 661 nm to 750 nm with a spectral resolution of 5 nm) and remaining 9 lie in the NIR region (i.e. 755 nm, 760 nm, 765 nm, 770 nm, 775 nm, 805 nm, 810 nm, 815 nm, 883 nm). PLS-DA applied to spectral data of the 34 wavebands gave a model of almost similar accuracy to the full 40 wavebands (Table 3).

The above analysis represents that brown colour of the ZC infection areas/dots is one of the most important factor in identifying the ZC infected potatoes. This is because of the fact of polyphenolic oxidation results in decolouration of healthy potato tissues developing brown spots/areas.

These results indicated that the 34 wavebands were the main contributor in discrimination between healthy and infected potatoes (Fig. 5).

3.5. *Evaluating the ZC disease detection system*

The results from this study show that HSI has potential to discriminate ZC disease infected potatoes from healthy. Chemical changes in ZC diseased potatoes appear to have resulted in noticeable differences in reflectance values in the visible (550 nm–750 nm) and near infrared regions (750 nm–1400 nm) (Fig. 5). These spectral ranges have also shown to be related to browning in bruised potatoes (Ye, Sun, Tan, Che, & Yang, 2018). This could lead to confounding effects between browning and ZC disease. This could result in bruised potatoes being rejected as diseased potatoes, which is not a problem as bruising is also classified as a quality defect.

Liang et al. (2018) investigated the use of NIRS for non-destructive detection of ZC diseased potatoes. In their study the spectral range utilised was between 900 nm and 2600 nm and it was reported an accuracy of 97%, which 5%–7% higher than current study. The difference in performance could be associated with size of the population investigated, where

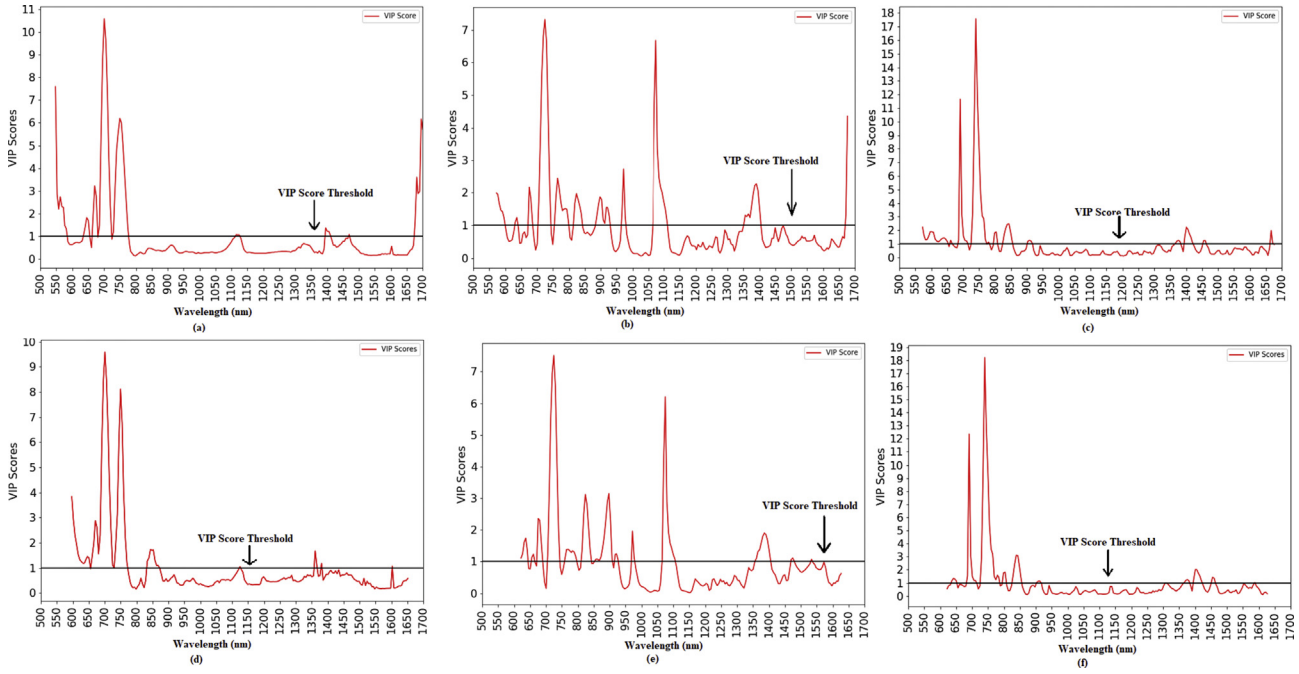

Fig. 4 – VIP Scores of PLS-DA model developed with and without pre-processing of spectral data (a) 235 wavebands without pre-processing (b) 235 wavebands with first derivative(c) 235 wavebands with second derivative (d) 215 wavebands without pre-processing (e) 215 wavebands with first derivative (f) 215 wavebands with second derivative.

Table 2 – Model performance comparison in ZC diseased potato identification using selected wavebands.

Pre-Processing	Number of Wavebands	PLS-DA on selected wavebands with 11 latent components					
		Cross Validated Calibration			Test		
		A	Sen	Spe	A	Sen	Spe
LT	46	90	90	90	91	91	91
LT	77	89	89	89	89	89	89
LT	72	89	89	89	90	90	90

A: Accuracy, Sen: Sensitivity, Spe: Specificity, LT: Log Transformation, the numbers are % values.

Table 3 – Model performance comparison in ZC diseased potato identification using optimised wavebands.

Pre-Processing	Number of Wavebands	PLS-DA on Full Spectrum with 11 latent components					
		Cross Validated Calibration			Test		
		A	Sen	Spe	A	Sen	Spe
LT	40	89	89	89	89	89	89
LT	34	89	88	88	88	88	88

A: Accuracy, Sen: Sensitivity, Spe: Specificity, LT: Log Transformation, the numbers are % values.

Liang et al. (2018) used 363 tubers (175 infected and 188 non-infected) in comparison with the current study, which utilised 3352 (1589 Healthy and 1763 diseased). The difference in performance could also be associated to the validation approach, where Liang et al. reports the accuracy based on cross validation results while the current study is based on validation data set (30% of that was not used in the development of the calibration). Results of Liang et al. shows important wavebands lies in between 1250 and 1350 nm, around 1400 nm, around 1700 nm, 1800–1870 and 2050–2150 nm with the greatest contribution with the band between 1800 nm and 1870 nm. The current study identified 25 important wavebands in the visible region whereas 6 in between 607 and 652 nm, and 19 wavebands from 661 nm to 750 nm) and 15 in the NIR region i.e. 5 in between 755 nm and 775 nm, 4 in between 805 and 883 nm and 6 wavebands in between 1346 nm and 1381 nm. The highest contribution was observed for visible region wavebands. Altogether the current study and Liang et al. (2018) demonstrate the value of VIS-NIR for detection of ZC diseased potatoes.

Most of the important wavebands for ZC identification were found in the visible region ranging from 600 nm to

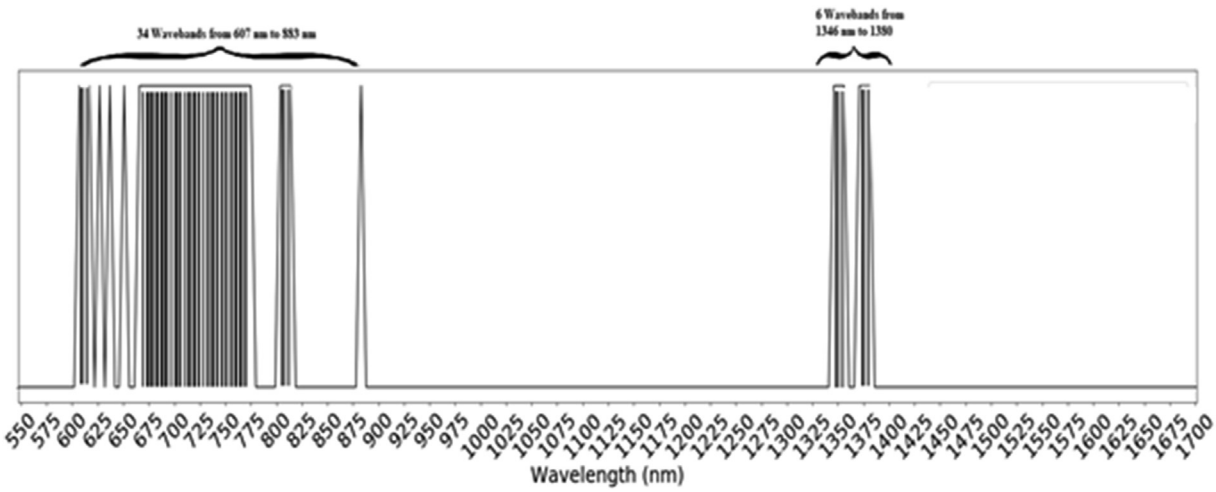

Fig. 5 – Important wavebands for identification of ZC.

750 nm (i.e. 24 wavebands) (Fig. 5). This region is attributed to colour changes caused by phenolic oxidation (Xing, Bravo, Moshou, Ramon, & De Baerdemaeker, 2006). Usually, the ZC disease causes brown dots that can have strong absorption characteristics in the visible region (Ye et al., 2018). The wavebands from 750 nm to 885 nm represent the third overtone of N–H and C–H vibrations that could be associated with phenolic compounds (Porteous, Muir, & Wastie, 1981).

The bacterium 'Candidatus Liberibacter solanacearum' decreases specific gravity by reducing the amount of water in tubers (AUSVEG, 2017). As ZC infection increases, we can expect higher reflectance values in the water absorption region (between 1350 nm and 1400 nm) of the spectrum. The other seven wavebands are between 1300 nm and 1400 nm (Fig. 5), which includes the starch relevant wavebands at 1380 nm, and may reflect changes in starch composition due to ZC infection.

Selecting the optimal number of wavebands (34 wavebands) that are sensitive to ZC disease could improve acquisition speed and reduce computational time without compromising the performance of the predictions.

The highest accuracy for ZC infected tubers detection observed in this study is 92%. The fact that we scanned potatoes only from one side and ZC infection might be possible on the other side of potato may contribute to inaccuracy. This can be mitigated by taking multiple images of the same tuber while rotating on the moving conveyor belt.

This research proved that HSI has potential for rapid and non-destructive identification of Zebra disease in potatoes. As we used a wide range potatoes in the model development, the model is simplified and can be deployable in an industrial environment for automatic screening. However, real-time processing tools are required to process the data.

As this research has been conducted on specific variety of potato, further research is required to test the performance of model on different varieties from wider potato growing regions.

This work is based on the assumption that visual inspection of peeled potatoes enabled identification of signs of the ZC infected potatoes. This allowed to use a large population of samples for the study. A subset of the potatoes identified as diseased was fried and confirmed that our approach for visual assessment of ZC signs was robust. A limitation of our approach is the possibility of having infected potatoes that did not show visual signs of the disease, and therefore were assigned to the healthy potatoes group. However, our results show similar sensitivity and specificity and therefore similar classification performance for both groups. Considering that two groups are well balanced in the number of samples, this suggests that the likelihood of having incorrectly assigned potatoes to the heathy groups is low.

4. Conclusions

A ZC disease detection system based on HSI was proposed. The potential of HSI was evident in this study for rapid identification of ZC diseased potato tubers in a non-destructive manner. PLS-DA was useful in modelling full spectral data (i.e. 235 wavebands) and truncated spectra (i.e. 215 wavebands) with (first and second derivative) and without preprocessing. The 92% detection accuracy for ZC diseased potatoes confirmed the capability of HSI. Important wavebands were successfully selected based on VIP scores of PLS-DA models. ZC detection accuracy of 90% and 89% were achieved using 40 and 34 selected wavebands respectively.

This study shows that diffuse reflectance measure with external illumination and with the line scanning detection successfully discriminated healthy and diseased potatoes. Even better performance should be possible by making better use of spatial data which might better recognise uneven ZC distribution especially in lightly infected tubers. This approach allows that further enhancement is achieved with deep learning (such as 3D CNN) based methods to take in consideration the spatial distribution within the scanned potato, which is focus of future research. Overall, this research demonstrates the potential of HSI as a fast ZC detection tool for whole washed tubers in the potato chip industry.

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.

Acknowledgements

Authors acknowledge their funding bodies: FIET (Food Industry Enabling Technologies) Programme funding from New Zealand Ministry of Business, Innovation & Employment (MBIE), and the Potato NZ, New Zealand.

Appendix A. Tables

Table 4 – Model performance comparison in ZC diseased potato identification using different pre-processing technique combinations with different number of wavebands.

Pre-Processing	Number of Wavebands	PLS-DA on Full Spectrum with 11 latent components					
		Cross Validated Calibration			Test		
		A	Sen	Spe	A	Sen	Spe
LT	235	89	89	89	87	87	87
LT+SNV+FDT+ MC	235	90	90	90	90	90	90
LT+SNV+SDT+MC	235	91	91	91	92	92	92
Selected wavebands based on VIP scores of PLS-DA model without pre-processed spectral data of 235 wavebands							
LT	46	90	90	90	91	91	91
LT+SNV+FDT+MC	46	89	89	89	90	90	90
LT+SNV+SDT+MC	46	89	89	89	91	91	91
Selected wavebands based on VIP scores of PLS-DA model with first derivative pre-processed spectral data of 235 wavebands							
LT	77	89	89	89	89	89	89
LT+SNV+FDT+MC	77	89	89	89	88	88	88
LT+SNV+SDT+MC	77	89	89	89	89	89	89
Selected wavebands based on VIP scores of PLS-DA model with second derivative pre-processed spectral data of 235 wavebands							
LT	72	89	89	89	90	90	90
LT+SNV+FDT+MC	72	89	89	89	89	89	89
LT+SNV+SDT+MC	72	89	89	89	89	89	89

A: Accuracy, Sen: Sensitivity, Spe: Specificity, FDT: Savitzky Golay with first derivative, SDT: Savitzky Golay with second derivative, SNV: for standard normal variate), the numbers are % values.

Table 5 – Model performance comparison in ZC diseased potato identification using different pre-processing technique combinations with different number of wavebands.

Pre-Processing	Number of Wavebands	PLS-DA on selected wavebands with 11 latent components					
		Cross Validated Calibration			Test		
Trimmed spectral data i.e. 215 wavebands							
LT	215	90	90	90	91	91	91
LT+SNV+FDT+MC	215	90	90	90	90	90	90
LT+SNV+SDT+MC	215	91	91	91	92	92	92
Selected wavebands based on VIP scores of PLS-DA model without pre-processed spectral data of 215 wavebands							
LT	47	89	89	89	60	59	59
LT+SNV+FDT+MC	47	89	89	89	81	81	81
LT+SNV+SDT+MC	47	89	89	89	86	86	86
Selected wavebands based on VIP scores of PLS-DA model with first derivative pre-processed spectral data of 215 wavebands							
LT	61	90	90	90	60	58	58
LT+SNV+FDT+MC	61	89	89	89	80	80	80
LT+SNV+SDT+MC	61	89	89	89	68	68	68
Selected wavebands based on VIP scores of PLS-DA model with second derivative pre-processed spectral data of 215 wavebands							
LT	59	89	89	89	89	89	89
LT+SNV+FDT+MC	59	88	88	88	89	89	89
LT+SNV+SDT+MC	59	88	88	88	87	87	87
From the above six sets of selected wavebands, common wavebands were selected that fell in at least four sets of selected wavebands							
LT	40	89	89	89	89	89	89
LT+SNV+FDT+MC	40	89	89	89	88	88	88
LT+SNV+SDT+MC	40	88	88	88	88	88	88
From above 40 wavebands, 34 wavebands selected that fall below 100 nm							
LT	34	89	89	89	88	88	88
LT+SNV+FDT+MC	34	86	86	86	83	83	83
LT+SNV+SDT+MC	34	87	87	87	87	87	87

A: Accuracy, Sen: Sensitivity, Spe: Specificity, FDT: Savitzky Golay with first derivative, SDT: Savitzky Golay with second derivative, SNV: for standard normal variate), the numbers are % values.

REFERENCES

Amjad, W., Crichton, S. O. J., Munir, A., Hensel, O., & Sturm, B. (2018). Hyperspectral imaging for the determination of potato slice moisture content and chromaticity during the convective hot air drying process. *Biosystems Engineering, 166*, 170–183. https://doi.org/10.1016/j.biosystemseng.2017.12.001

Anonymous. (2019). *Potatoes now worth one billion dollars to NZ economy*. Retrieved from https://potatoesnz.co.nz/news-info/industry-news/potatoes-now-worth-one-billion-dollars-to-nz-economy/.

AUSVEG. (2017). *Tomato-potato psyllid and zebra chip information sheet. 6*. Retrieved from AUSVEG website https://ausveg.com.au/app/uploads/2017/05/TPP-and-Zebra-Chip-Information-Sheet.pdf.

Ayvaz, H., Santos, A. M., Moyseenko, J., Kleinhenz, M., & Rodriguez-Saona, L. E. (2015). Application of a portable infrared instrument for simultaneous analysis of sugars, asparagine and glutamine levels in raw potato tubers. *Plant Foods for Human Nutrition, 70*(2), 215–220. https://doi.org/10.1007/s11130-015-0484-7

Bethke, P. C., Nassar, A. M. K., Kubow, S., Leclerc, Y. N., Li, X.-Q., Haroon, M., et al. (2014). History and origin of Russet Burbank (netted gem) a sport of Burbank. *American Journal of Potato Research, 91*(6), 594–609. https://doi.org/10.1007/s12230-014-9397-5

Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: Taking the magic away. *Journal of Chemometrics, 28*(4), 213–225. https://doi.org/10.1002/cem.2609

Buchman, J. L., Fisher, T. W., Sengoda, V. G., & Munyaneza, J. E. (2012). Zebra Chip progression: From inoculation of potato plants with liberibacter to development of disease symptoms in tubers. *American Journal of Potato Research, 89*(2), 159–168. https://doi.org/10.1007/s12230-012-9238-3

Burger, J., & Geladi, P. (2007). Spectral pre-treatments of hyperspectral near Infrared images: Analysis of diffuse reflectance scattering. *Journal of Near Infrared Spectroscopy, 15*(1), 29–37. https://doi.org/10.1225/jnirs.717

Dyer, D. G., Blackledge, J. A., Katz, B. M., Hull, C. J., Adkisson, H. D., Thorpe, S. R., et al. (1991). The Maillard reaction in vivo. *Zeitschrift für Ernährungswissenschaft, 30*(1), 29–45. https://doi.org/10.1007/bf01910730

Elmasry, G., Kamruzzaman, M., Sun, D.-W., & Allen, P. (2012). Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: A review. *Critical Reviews in Food Science and Nutrition, 52*(11), 999–1023. https://doi.org/10.1080/10408398.2010.543495

Eriksson, L., Johansson, E., Kettaneh-Wold, N., & Wold, S. (2001). Multi-and megavariate data analysis. Principles and applications. *Journal of Chemometrics, 16*, 261–262. https://doi.org/10.1002/cem.713

Feng, Y.-Z., & Sun, D.-W. (2012). Application of hyperspectral imaging in food safety inspection and control: A review. *Critical Reviews in Food Science and Nutrition, 52*(11), 1039–1058. https://doi.org/10.1080/10408398.2011.651542

Garhwal, A. S., Pullanagari, R., Li, M., Archer, R., & Reis, M. (2019). *Feasibility study of detecting phenolic compounds changes in potatoes using Hyperspectral Imaging*. Gold Coast, Australia: Paper presented at the 19th International Council of Near Infrared Spectroscopy Conference.

Garhwal, A. S., Pullanagari, R., Li, M., Bailey, D., Reis, M., & Archer, R. (2020). Spectral separations of potatoes and honey. *Food New Zealand, 19*, 9–11.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics, 2*(3), 211–228. https://doi.org/10.1002/cem.1180020306

Huang, T., Li, X. Y., Xu, M. L., Jin, R., Ku, J., Xu, S. M., et al. (2015). Non-destructive detection research for hollow heart of potato based on semi-transmission hyperspectral imaging and SVM. *Guang pu xue yu guang pu fen Xi=Guang pu, 35*(1), 198–202. https://doi.org/10.3964/j.ssn.1000-0593(2015)01-0198-05

Kumar, G. N. M., Knowles, L. O., & Knowles, N. R. (2015). Zebra Chip disease decreases tuber (Solanum tuberosum L.) protein content by attenuating protease inhibitor levels and increasing protease activities. *Planta, 242*(5), 1153–1166. https://doi.org/10.1007/s00425-015-2346-9

Levy, J., Ravindran, A., Gross, D., Tamborindeguy, C., & Pierson, E. (2011). Translocation of 'Candidatus liberibacter solanacearum', the zebra chip pathogen, in potato and tomato. *Phytopathology®, 101*(11), 1285–1291. https://doi.org/10.1094/PHYTO-04-11-0121

Liang, P.-S., Haff, R. P., Hua, S.-S. T., Munyaneza, J. E., Mustafa, T., & Sarreal, S. B. L. (2018). Nondestructive detection of zebra chip disease in potatoes using near-infrared spectroscopy. *Biosystems Engineering, 166*, 161–169. https://doi.org/10.1016/j.biosystemseng.2017.11.019

López-Maestresalas, A., Keresztes, J. C., Goodarzi, M., Arazuri, S., Jarén, C., & Saeys, W. (2016). Non-destructive detection of blackspot in potatoes by Vis-NIR and SWIR hyperspectral imaging. *Food Control, 70*, 229–241. https://doi.org/10.1016/j.foodcont.2016.06.001

Mehmood, T., Martens, H., Sæbø, S., Warringer, J., & Snipen, L. (2011). A Partial Least Squares based algorithm for parsimonious variable selection. *Algorithms for Molecular Biology, 6*(1), 27. https://doi.org/10.1186/1748-7188-6-27

Munyaneza, J. E. (2012). Zebra chip disease of potato: Biology, epidemiology, and management. *American Journal of Potato Research, 89*(5), 329–350. https://doi.org/10.1007/s12230-012-9262-3

Navarre, D. A., Shakya, R., Holden, J., & Crosslin, J. M. (2009). LC-MS analysis of penolic compounds in tubers showing Zebra Chip symptoms. *American Journal of Potato Research, 86*(2), 88–95. https://doi.org/10.1007/s12230-008-9060-0

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*(1), 45–50. https://doi.org/10.4103/0301-4738.37595

Porteous, R. L., Muir, A. Y., & Wastie, R. L. (1981). The identification of diseases and defects in potato tubers from measurements of optical spectral reflectance. *Journal of Agricultural Engineering Research, 26*(2), 151–160. https://doi.org/10.1016/0021-8634(81)90066-4

Rinnan, Å., Berg, F. V. D., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TRAC Trends in Analytical Chemistry, 28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Riza, D. F. A., Suzuki, T., Ogawa, Y., & Kondo, N. (2017). Diffuse reflectance characteristic of potato surface for external defects discrimination. *Postharvest Biology and Technology, 133*, 12–19. https://doi.org/10.1016/j.postharvbio.2017.07.006

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Schlosser, E. (2001). *Fast food nation: The dark side of the all-American meal*. Boston, New York, US: Houghton Mifflin Harcourt.

Su, W.-H., Bakalis, S., & Sun, D.-W. (2018). Fourier transform mid-infrared-attenuated total reflectance (FTMIR-ATR) microspectroscopy for determining textural property of microwave baked tuber. *Journal of Food Engineering, 218*, 1–13. https://doi.org/10.1016/j.jfoodeng.2017.08.016

Su, W.-H., & Sun, D.-W. (2017). Chemical imaging for measuring the time series variations of tuber dry matter and starch concentration. *Computers and Electronics in Agriculture, 140*, 361–373. https://doi.org/10.1016/j.compag.2017.06.013

Want, E., & Masson, P. (2011). Processing and analysis of GC/LC-MS-based metabolomics data. In T. O. Metz (Ed.), *Metabolic*

profiling: Methods and protocols (pp. 277–298). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-61737-985-7_17.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., et al. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*(1), 81–89. https://doi.org/10.1007/s11306-007-0099-6

Wold, S., Sjöström, M., & Eriksson, L. (1998). Partial least squares projections to latent structures (PLS) in Chemistry. *Encyclopedia of Computational Chemistry*. https://doi.org/10.1002/0470845015.cpa012

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109–130. https://doi.org/10.1016/s0169-7439(01)00155-1

Xing, J., Bravo, C., Moshou, D., Ramon, H., & De Baerdemaeker, J. (2006). Bruise detection on 'Golden Delicious' apples by vis/NIR spectroscopy. *Computers and Electronics in Agriculture, 52*(1), 11–20. https://doi.org/10.1016/j.compag.2006.01.006

Ye, D., Sun, L., Tan, W., Che, W., & Yang, M. (2018). Detecting and classifying minor bruised potato based on hyperspectral imaging. *Chemometrics and Intelligent Laboratory Systems, 177*, 129–139. https://doi.org/10.1016/j.chemolab.2018.04.002

Zhao, Z., Prager, S. M., Cruzado, R. K., Liang, X., Cooper, W. R., Hu, G., et al. (2018). Characterizing Zebra Chip symptom severity and identifying spectral signatures associated with 'Candidatus Liberibacter solanacearum'-infected potato tubers. *American Journal of Potato Research, 95*(5), 584–596. https://doi.org/10.1007/s12230-018-9666-9