World Academy of Science, Engineering and Technology
International Journal of Agricultural and Biosystems Engineering
Vol:14, No:10, 2020

Multivariate Analysis of Spectroscopic Data for Agriculture Applications

Asmaa M. Hussein, Amr Wassal, Ahmed Farouk Al-Sadek, A. F. Abd El-Rahman

Abstract**—**In this study, a multivariate analysis of potato spectroscopic data was presented to detect the presence of brown rot disease or not. Near-Infrared (NIR) spectroscopy (1,350-2,500 nm) combined with multivariate analysis was used as a rapid, non-destructive technique for the detection of brown rot disease in potatoes. Spectral measurements were performed in 565 samples, which were chosen randomly at the infection place in the potato slice. In this study, 254 infected and 311 uninfected (brown rot-free) samples were analyzed using different advanced statistical analysis techniques. The discrimination performance of different multivariate analysis techniques, including classification, pre-processing, and dimension reduction, were compared. Applying a random forest algorithm classifier with different pre-processing techniques to raw spectra had the best performance as the total classification accuracy of 98.7% was achieved in discriminating infected potatoes from control.

Keywords**—**Brown rot disease, NIR spectroscopy, potato, random forest.

I. Introduction

POTATOS are a popular tuber that is consumed directly and contribute to the basic diet of many people around the world. Potato (*Solanum tuberosum* L.) is a globally important crop plant producing high yields of nutritionally valuable food in the form of tubers [1]. The total world potato production was about 388,191,000 tons in 2017. Egypt is the top largest potato producer in Africa and one of the top 20 producers of potato worldwide. Egypt's potato production increased from 1,637,810 tons in 1990 to 3,659,280 tons in 2009 and reached 4,325,480 tons in 2017 [2].

Potato is one of the most important vegetable crops in Egypt for local consumption and one of the largest agricultural export crops to the European countries. The consumption (Crop Equivalent) rate of potatoes (fresh and processed) increased from 21.25 kg/capita/year in 1990 to 34.62 kg/ capita/year in 2011. In 2004, Egyptian exports totaled more than 380,000 tons of fresh potatoes and 18,000 tons of frozen potato products, mostly to markets in Europe and these exports are increasing. The amount of seeds needed for planting potato in Egypt was about 516,000 tons in 2013 [2].

Potato brown rot is an economically important disease in tropical, subtropical and temperate regions of the world. The disease is caused by the bacterium *Ralstonia solanacearum* race 3 biovar 2. The disease is generally referred to as 'bacterial wilt', but in potato it is also called 'brown rot' because tuber vascular tissue is usually a distinct grayish brown, and the discoloration may expand into the pith or cortex [3]. Quarantine restrictions on potato brown rot imposed by the European Union (EU) represent the biggest challenge to Egypt's potato exports to the EU countries. Therefore, the Central Administration for Plant Quarantine (CAPQ) of the Ministry of Agriculture, Egypt, delimit pest free areas (PFAs) that have not been shown to be infected with *Ralstonia solanacearum*, for potato cultivation for export [4].

The potato brown rot project in Egypt was set up to test imported potato seed and the exported potato for brown rot. The methods of sampling, extraction and testing used in Egypt are based on those prescribed in the European protocols.

Near Infrared Spectroscopy (NIRS) is one of the most advanced techniques for non-destructive quality control of agricultural and food products [5]. Over the past few decades, it has been used effectively for the quantitative analysis of many agricultural and food products [6], [7]. NIR can also be used for qualitative analysis, however, with the goal of classifying samples based on their spectral characteristics rather than estimating the components present in them [8]. The NIR spectroscopy advance method was evaluated using the portable systems and on-line grading/sorting machines to determine the internal qualities and external deficiencies of the potato tubers [9]. Therefore, the huge demands in yield production and measurements of food quality contribute to the expansion of feasible and real-time machine systems. Replacing the conventional method is important because the process is destructive, time-consuming, arduous and manual [10]. NIR spectroscopy techniques were therefore introduced in order to overcome the limitation of the conventional methods without compromising the physicochemical properties of food and agricultural products. Therefore, the purpose of this study was to investigate the practicality of using NIR spectroscopy together with advanced statistical analysis for rapid, non-destructive detection of potato tubers infected with potato brown rot. This could be carried out directly by correlating spectral features with potato brown rot infection.

II. Materials and Methods

In this section, the materials used in the study and also the computational methods for classifying are presented. As shown in Fig. 1, the activities started by collecting the potato

Asmaa M. Hussein and Ahmed Farouk Al-Sadek are with the Climate Change Information Center, Renewable Energy, and Expert Systems, Agricultural Research Center, Giza, Egypt (e-mail: asmaama.ahmed2020@gmail.com, afsadek@gmail.coml).

Amr Wassal is with the Computer Engineering Department, Cairo University, Cairo 12613, Egypt (e-mail: wassal@eng.cu.edu.eg).

A.F.Abd El-Rahman is with Bacterial Diseases Research Department, Plant Pathology Research Institute, Agricultural Research Center, Giza, Egypt (e-mail: aabdelrahman2012@gmail.com).

samples then applying the spectral radiation and collecting its results on data files that were introduced to the classification model. The model is illustrated in Fig. 1, starting by outlier removal then normalizing the data using different algorithms. Then, the model applied the dimensionality reduction on the spectroscopic data to get a smaller number of features. Lastly, the finalized data were introduced to different machine learning classifiers to get the final results and select the best one. In the following subsections, the detailed results for each of those model parts will be presented.

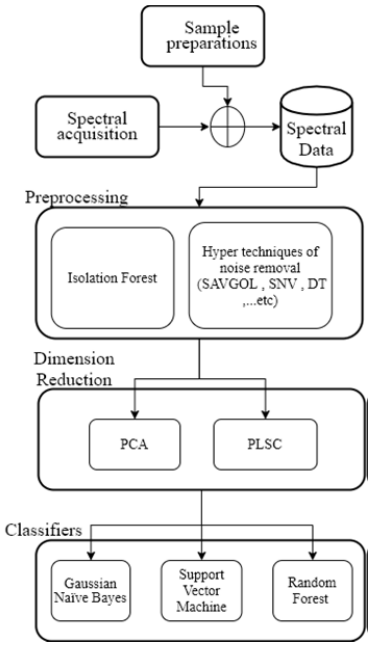

Fig. 1 System architecture

A. Sample Preparations

Healthy potato tubers and potato tubers showing brown color and necrosis in the vascular tissues (symptoms of brown rot) of Spunta variety were kindly obtained from the Potato Brown Rot Project (PBRP, ARC, Egypt). Immunofluorescence (IF) test was performed as described in EPPO Standard PM 7/97 and EPPO Standard PM 7/21 (2) to all potato tubers to confirm whether or not the tubers were infected with brown rot disease [11], [12]. Each tuber was considered as a separate sample. Potatoes were cut width-wise into slices as shown in Fig. 2.

Fig. 2 Potatoes cut width wise in to slices

B. Spectral Acquisition

The mean diffuse reflectance spectra in the long-wavelength NIR regions (1,350-2,500 nm), were measured in this study at different random locations of each whole, potato slice using a spectrophotometer (NeoSpectra-Micro – SWS62231). Spectral Sensor is linked to a personal computer running the Spectro Most micro version 1.0 software. The sensors are based on Fourier Transform Infrared (FT-IR) technology, which is a standard technique used in laboratory-based spectrometers that offers a wide spectral range for the best qualification and quantification of materials [13]. To obtain consistent measurements, each slice was placed such that the light beam struck random infection places as shown in Fig. 3 (a). Spectral measurements were performed in 565 samples which were chosen randomly at the infection place in the slice, see Fig. 3 (b). Spectral measurements were performed at the same places in the uninfected potato slices, see Fig. 3 (c). The analyzed samples used in this study were 254 infected samples with Potato Brown Rot infection and 311 uninfected samples.

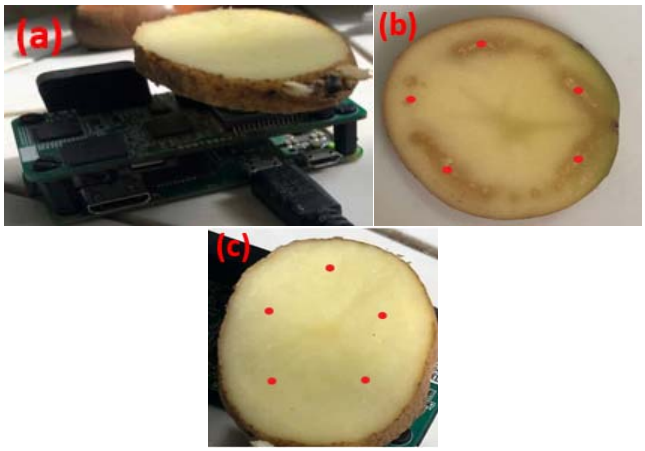

Fig. 3 (a) Potato slice placed on a spectral sensor such that the light beam struck random places, (b) Light beam struck random infection places (red spots) and (c) Light beam struck the same places in the uninfected potato (red spots)

C. Isolation Forest

Outliers are, in general, less frequent than normal observations and differ in terms of values from them (they lie further away from the normal observations in the feature space). The Isolation Forest algorithm isolates observations by selecting a feature randomly, and then randomly selecting a split value between the selected feature's maximum and minimum values. The way the algorithm creates the separation is through first building isolation trees, or random decision trees. Then the score was calculated to isolate the observation as the length of the path. Therefore, outliers should be identified closer to the root of the tree (shorter average path length, i.e., the number of edges of the tree from the root to the terminal node) with fewer splits needed, see Fig. 4 [14]. Therefore, Isolation Forest analysis was performed to select suitable experimental data for model construction and to detect and eliminate the outliers. All spectral data were processed and analyzed using statistical software (Anaconda Navigator1.9.6).

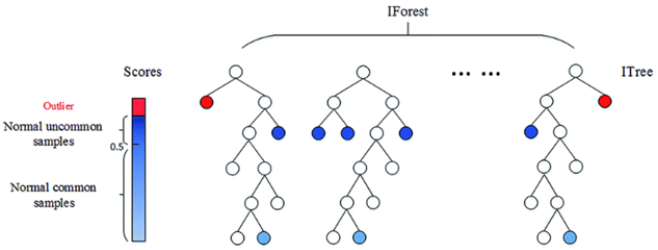

Fig. 4 Isolation Forest

D. Data Preprocessing

Pre-processing or pre-treatment of spectral data is often required to reduce/remove noise or undesirable background information and boost robustness in the subsequent analysis of data such as exploratory analysis, calibration, and classification modeling [15]. With a well-designed pre-processing step, the performance of the model can be greatly improved. Therefore, in this study, the NIR raw spectra obtained after outlier exclusion were treated by different data preprocessing techniques, such as Savitzky – Golay filter (SAVGOL) that was used to remove noise from the spectral data. The first derivative transforms were useful for eliminating baseline offset variations within a set of spectra, while the second derivative could help separate overlapping peaks and sharpen spectral features [16].

To eliminate the effects of baseline variation, light scattering, and path length differences, it was necessary to perform mathematical pretreatments including multiplicative scatter correction (MSC), standard normal variate (SNV), D-trend (DT) [17]. Data normalization is useful for the purpose of classification. There are so many ways to classify but all can vary greatly from one to another. The normalization technique is required to make them closer in order to maintain a large variety of classification [18]. Smoothing will eliminate the high-frequency noise.

The spectra were also treated using D-trend plus Normal plus Smooth, Smooth plus SAVGOL, Normal plus SAVGOL, Normal plus Smooth, Normal plus Smooth plus SAVGOL.

E. Dimension Reduction

Dimension reduction is mapping data to a lower-dimensional space to discard uninformative variance in the data, or to detect a subspace where the data lives. Dimension reduction method is for visualizing data and extracting key low-dimensional features (for example, an object's two-dimensional orientation from its high dimensional representation of images) [19].

In this study, we used Principle Component Analysis (PCA) and Partial Least Squares Canonical (PLSC) dimension reduction techniques. PLSC implements the two blocks canonical PLS of the original Wolf algorithm [20], [21]. PCA selects input data's maximum variance estimate, placing an orthonormality limit on the projection vectors. PCA works on the assumption that the high variance estimates include the information relevant to the learning task at hand [22].

F. Classification

In this study, Gaussian Naïve Bayes (NB) classifier, Support Vector Machine (SVM) classifier, and Random Forest (RF) algorithm classifier were used. NB is a supervised learning algorithm that uses Bayes' rule to calculate the probability that a sample belongs to a certain class based on the features that it contains, with the naïve assumption that the features are statistically independent conditional on class membership [23]. SVM is a classifier that tries to find the optimum class hyperplane. Optimality is defined here as the maximum margin between the classes and the hyperplane described by the vectors of support [24]. RF algorithm is a method of regression or classification that uses decision trees: a sequence of rules that divide the data in a way that eliminates variance in the most optimal way.

Each tree receives a random subset of training samples and the algorithm selects a subset of variables randomly at each split in the tree [25]. These trees, which are individually relatively poor classifiers, are combined into a tree's ensemble called a RF that is used for prediction A random forest's predictive results are a summary of many individual trees ' predictive outcomes. Here, there are only two classes (control and brown rot-infected).

The accuracy can be defined as the percentage of correctly classified instances (TP + TN)/(TP + TN + FP + FN) where TP, FN, FP, and TN represent the number of true positives, false negatives, false positives, and true negatives, respectively.

III. Results and Discussion

The absorbance spectra of 311 control (brown rot-free) samples and 254 brown rot-infected samples were plotted and shown in Fig. 2 A. Isolation Forest analysis was performed to select suitable experimental data for model construction and to detect and eliminate the outliers, see Fig. 5. Various noise removal techniques mentioned in the previous section materials and methods were applied after outlier removal from the row of data. The data were mapped with different criteria from domain value to another. In other words, all of those techniques are a domain transformation as shown in Figs. 7-12.

Hyper noise removal techniques were also applied to get more suitable experimental data to be forwarded to the dimension reduction process. In this study the following hyper methods: Detrend then Normal then Smooth, Smooth then Savgol, Normal then Savgol, Normal then Smooth, Normal then Smooth then Savgol were applied as shown in Figs. 13-17. Then, the dimension reduction was applied with different techniques and a different number of components two and five. Now, the experimental data are suitable to be classified, three classifiers were applied. Different techniques to build the model were applied as described in Fig. 1.

Finally, various results with different techniques were obtained as shown in Table I. One category data were selected from the table to be evaluated with 60% training data and 40% testing data, or 70% training data and 30% testing data.

Accuracy of the classification models with different pre-processing techniques and dimension reduction methods, with 60% training data and 40% testing data is shown in Figs. 18 and 19, where (1) refers to Detrend then Normal then Smooth, (2) refers to SNV, (3) refers to Normal then SAVGOL, (4) refers to Normal then Smooth, and (5) refers to Normal then Smooth then SAVGOL. (a) refers to PLSC and (b) refers to PCA. So (1.a) is Detrend then Normal then Smooth pre-processing with PLSC dimension reduction and (1.b) is Detrend then Normal then Smooth pre-processing with PCA dimension reduction.

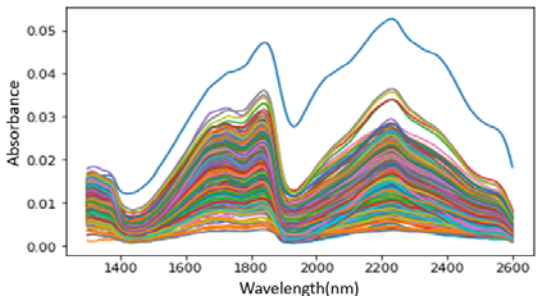

Fig. 5 Row of Data

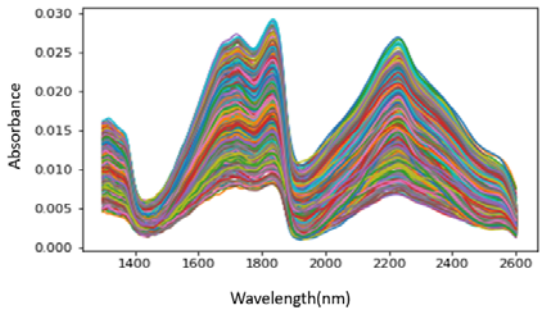

Fig. 6 Data after outlier removing

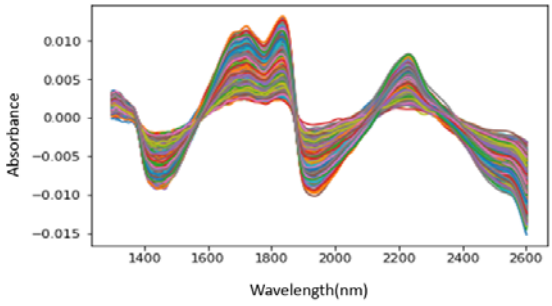

Fig. 7 NIR spectra pre-processed using the Detrend method

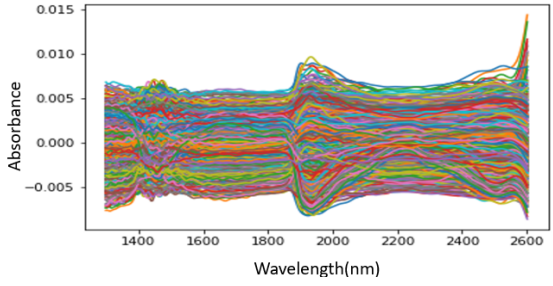

Fig. 8 NIR spectra pre-processed using the MSC method

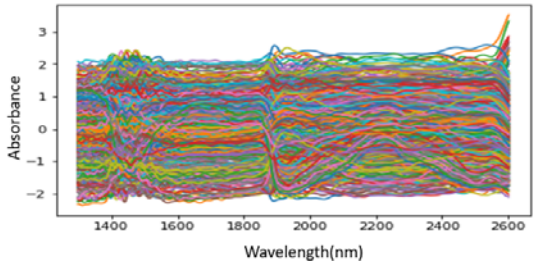

Fig. 9 NIR spectra pre-processed using the SNV method

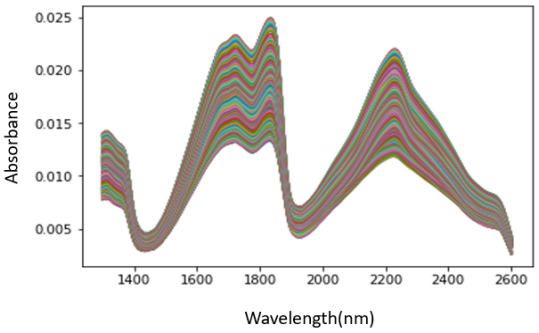

Fig. 10 NIR spectra pre-processed using the Smooth method

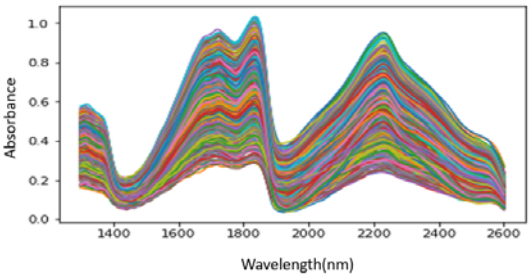

Fig. 11 NIR spectra pre-processed using the Normal method

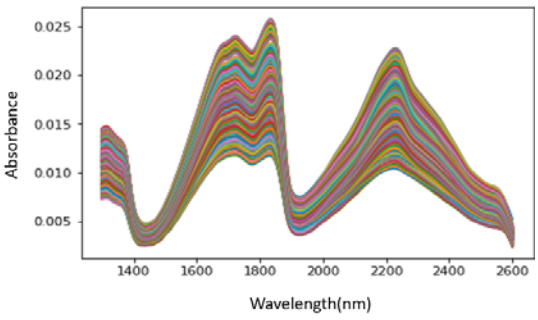

Fig. 12 NIR spectra pre-processed using the SAVGOL method

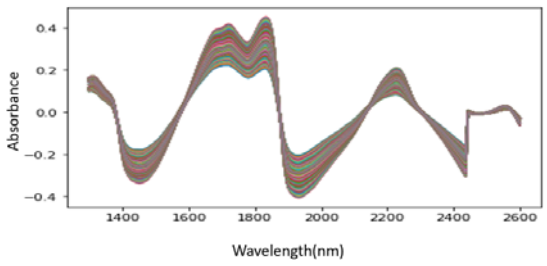

Fig. 13 NIR spectra pre-processed using the Detrend then the Normal then the Smooth method

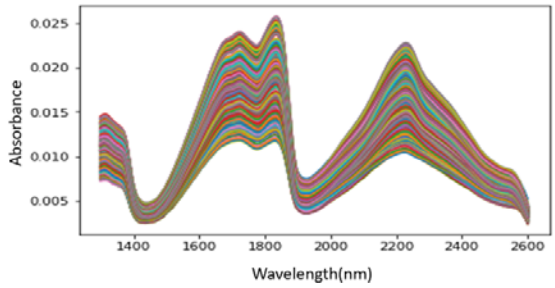

Fig. 14 NIR spectra pre-processed using the Smooth then the SAVGOL method

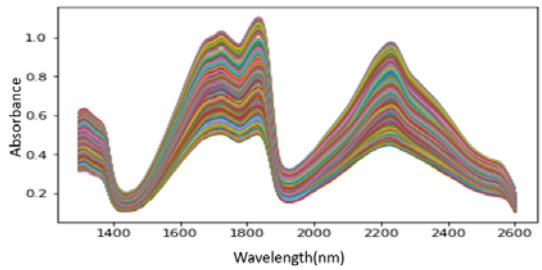

Fig. 15 NIR spectra pre-processing using the Normal then the SAVGOL method

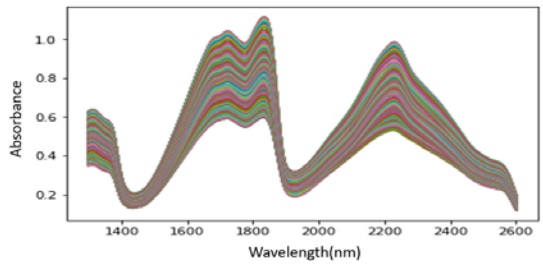

Fig. 16 NIR spectra pre-processed using the Normal then the Smooth method

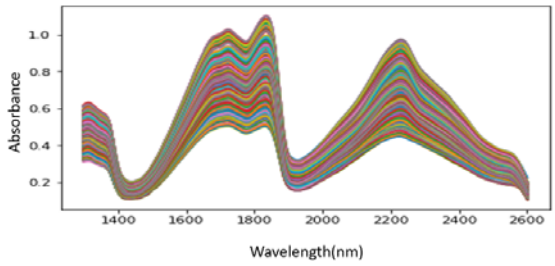

Fig. 17 NIR spectra pre-processed using the Normal then the Smooth then the SAVGOL method

TABLE I
CLASSIFICATION RESULTS BASED ON DIFFERENT CLASSIFICATION MODELS WITH DIFFERENT PREPROCESSING METHODS AND FULL WAVELENGTHS

Preprocessing	Dimension Reduction	Number of Component	Testing Percentage	Accuracy with SVM	Accuracy with GNB	Accuracy with RF
DETREND NORML SMOOTH	*PLSC*	*2*	*30%*	*96.7%*	*96.7%*	*98.5%*
			40%	*54.9%*	*54.9%*	*98.5%*
		5	*30%*	*98%*	*98%*	*98.5%*
			40%	*73.5%*	*73.5%*	*98.5%*
	PCA	*2*				
			30%	*95.4%*	*95.4%*	*98.3%*
			40%	*80.3%*	*80.3%*	*98.7%*
		5	*30%*	*61.4%*	*61.4%*	*98.3%*
			40%	*79.4%*	*79.4%*	*98.7%*
SNV	*PLSC*	*2*	*30%*	*89.5%*	*90%*	*84.9%*
			40%	*88.7%*	*85%*	*87%*
		5	*30%*	*90.8%*	*89.5%*	*88.8%*
			40%	*85.7%*	*89%*	*89.7%*
	PCA	*2*				
			30%	*85.6%*	*82.2%*	*89.5%*
			40%	*88%*	*84.3%*	*84.3%*
		5	*30%*	*91.5%*	*85.6%*	*85.6%*
			40%	*86.2%*	*88%*	*87%*
NORMAL SAVGOL	*PLSC*	*2*	*30%*	*90.8%*	*92%*	*97%*
			40%	*90%*	*87.9%*	*98.7%*
		5	*30%*	*86.2%*	*92.8%*	*97%*
			40%	*89.7%*	*90%*	*97.5%*
	PCA	*2*				
			30%	*87.5%*	*87.5%*	*98%*
			40%	*92.6%*	*92.6%*	*98.5%*
		5	*30%*	*89.5%*	*85.6%*	*98.6%*
			40%	*86.7%*	*86.7%*	*97%*
NORMAL SMOOTH	*PLSC*	*2*	*30%*	*92%*	*92%*	*98.6%*
			40%	*90%*	*92%*	*98%*
		5	*30%*	*90%*	*92%*	*98%*
			40%	*90%*	*90%*	*98%*
	PCA	*2*				
			30%	*94.7%*	*90.8%*	*98%*
			40%	*93%*	*93%*	*98.5%*
		5	*30%*	*92.7%*	*92.8%*	*98%*
			40%	*91.6%*	*93%*	*98.7%*
NORMAL SMOOTH SAVGOL	*PLSC*	*2*	*30%*	*88%*	*88%*	*98%*
			40%	*89%*	*89%*	*98%*
		5	*30%*	*88.8%*	*88.8%*	*98%*
			40%	*90%*	*90%*	*98.7%*
	PCA	*2*				
			30%	*88.8%*	*88.8%*	*96%*
			40%	*87.7%*	*87.7%*	*96.5%*
		5	*30%*	*90.8%*	*86.9%*	*97%*
			40%	*90.6%*	*90.6%*	*97%*

IV. CONCLUSION

This study shows that the use of a high-speed NIR spectrophotometer (Neo Spectra-Micro – SWS62231 – Spectral Sensor) instrument with a suitable multivariate analysis technique could be effective in detecting the potato infected with brown rot disease. The results indicate that the classification model built in this study has performed well with high accuracy in classification.

Applying RF algorithm classifier with PLSC or PCA dimension reduction method and various pre-processing techniques to raw spectra has the total classification accuracy 98.7% was achieved in discriminating infected potatoes from control. Future studies will be conducted on the use of the results in the determination of latent infection of brown rot and the distinction between brown rot and some other diseases that cause discoloration in potato tubers. The value of these conclusions can be useful in moving NIR reflection

technology from laboratory to industrial application for non-destructive, real-time, or portable potato quality measurement.

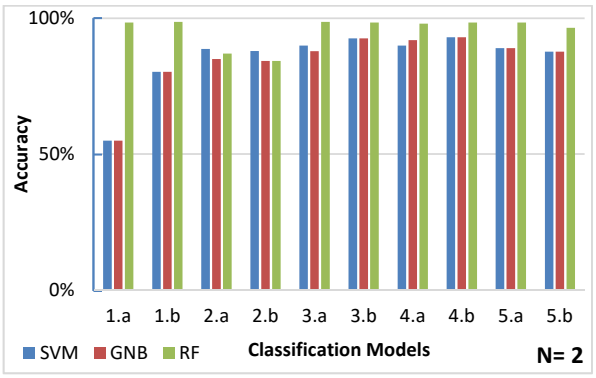

Fig. 18 Accuracy of classification models with different pre-processing techniques and dimension reduction methods with N = 2, where N is the number of components

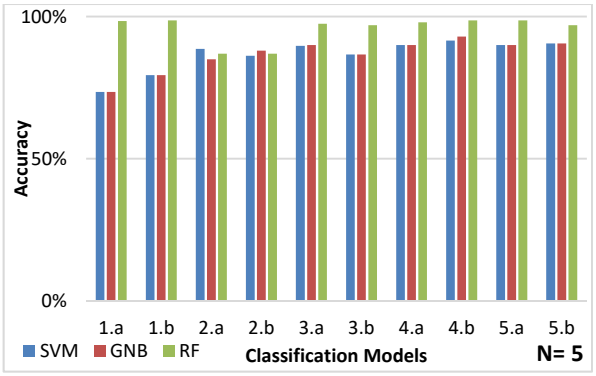

Fig. 19 Accuracy of classification models with different pre-processing techniques and dimension reduction methods with N = 5, where N is the number of components

REFERENCES

[1] Millam S (2007). Potato (*Solanum tuberosum* L.). Methods in Molecular Biology 344, 25–35.
[2] Potatopro, 2019. https://www.potatopro.com/world/potato-statistics
[3] Messiha NAS, van Bruggen AHC, van Diepeningen AD, de Vos OJ, Termorshuizen AJ, Tjou-Tam-Sin NNA, Janse JD (2007). Potato brown rot incidence and severity under different management and amendment regimes in different soil types. Eur J Plant Pathol 119:367–381
[4] Kabeil SS, Lashin SM, El-Masry MH, El-Saadani MA, Abd Elgawad, MM and Aboul-Einean AM (2008). Potato brown rot disease in Egypt: current status and prospects. American–Eurasian J Agric Environ Sci, 4 (1) : 44-54
[5] Magwaza LS, Opara UL, Nieuwoudt H, Cronje PJR, Saeys W and Nicolaï B, NIR spectroscopy applications for internal and external quality analysis of citrus fruit – A review. Food Bioprocess Technol 5: 425– 444 (2012).
[6] Gunasekaran S and Irudayaraj J, Nondestructive Food Evaluation. Techniques to Analyse Properties and Quality, Optical Methods: Visible NIR and FTIR Spectroscopy. Marcel Dekker, New York (2000).
[7] Nicolaï BM, Beullens K, Bobelyn E, Peirs A, Saeys W, Theron KI et al., Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest Biol Technol 46: 99–118 (2007).
[8] Lister SJ, Dhanoa MS, Stewart JL and Gill M, Classification and comparison of Gliricidia provenances using near infrared reflectance spectroscopy. Anim Feed Sci Technol 86: 221– 238 (2000).
[9] Berardinelli, A., Cevoli, C., Silaghi, F. A., Fabbri, A., Ragni, L., Giunchi, A.; and Bassi, D. (2010). FT-NIR spectroscopy for the quality characterization of apricots (*Prunus armeniaca* L.). Journal of Food Science, 75(7), 462-468.
[10] Jie, D., Xie, L., Rao, X.; and Ying, Y. (2014). Using visible and near infrared diffuse transmittance technique to predict soluble solids content of watermelon in an on-line detection system. Postharvest Biology and Technology, 90, 1-6.
[11] EPPO (2009). PM 7/97 (1): Indirect immunofluorescence test for plant pathogenic bacteria. Bulletin OEPP/EPPO Bulletin, 39: 413–416.
[12] EPPO (2018). PM 7/21 (2): *Ralstonia solanacearum*, *R. pseudosolanacearum* and *R. syzygii* (*Ralstonia solanacearum* species complex). Bulletin OEPP/EPPO Bulletin, 48: 32–63.
[13] R. Davis and L.J. Mauer. Fourier tansform infrared (FT-IR) spectroscopy: A rapid tool for detection and analysis of foodborne pathogenic bacteria. In book: Current research, technology and education topics in Applied Microbiology and Microbial Biotechnology Volume II. Publisher: Formatex Research Center. Editors: A. Mendez-Vilas
[14] Wo-Ruo Chen, Yong-Huan Yun, Ming Wen, Hong-Mei Lu, Zhi-Min Zhang* and Yi-Zeng Liang*., Representative subset selection and outlier detection via isolation forest. DOI: 10.1039/C6AY01574C (Paper) Anal. Methods, 2016, 8, 7225-7231
[15] (Reich, 2005) Reich, G., 2005. Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications. Advanced Drug Delivery Reviews 57, 1109e1143
[16] (Cen and He, 2007; Wu et al., 1995) Cen, H., He, Y., 2007. Theory & application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science & Technology 18, 72e83
[17] Dhanoa MS, Sister SJ and Barnes RJ, On the scales associated with near infrared reflectance difference spectra. Appl Spectros 49:765–772 (1995)
[18] S. Gopal Krishna Patro, Kishore Kumar sahu, March 2015, Normalization: A Preprocessing Stage, DOI: 10.17148/IARJSET.2015.2305
[19] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," Cognitive Science, vol. 9, 1985
[20] Tenenhaus, M. (1998). La regression PLS: theorie et pratique. Paris: Editions Technic.
[21] Jacob A. Wegelin. A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle, 2000.
[22] Marc Rebillat, Nazih Mechba, (2019). Principal Least Squares Canonical Correlation Analysis for damage quantification in aeronautic composite structures. Processes and Engineering in Mechanics and Materials Laboratory (PIMM, UMR CNRS 8006, Arts et Métiers ParisTech (ENSAM)), 151, Boulevard de l'Hôpital, Paris, F-75013, France.
[23] Mitchell TM (1997) Machine Learning. 1st edition. New York: McGraw-Hill
[24] Qi, X., Silvestrov, S., & Nazir, T. (2017). Data classification with support vector machine and generalized support vector machine. doi:10.1063/1.4972718
[25] Breiman, 2001 L. Breiman, Random forests, Mach. Learn., 45 (2001), pp. 5-32