Genetic Algorithm for Feature and Latent Variable Selection for Nutrient Assessment in Horticultural Products

Demelza Robinson*, Qi Chen*, Bing Xue*, Daniel Killeen†, Sara Fraser-Miller‡, Keith C Gordon‡, Indrawati Oey§, and Mengjie Zhang*

*School of Engineering and Computer Science,
Victoria University of Wellington, PO Box 600, Wellington, New Zealand
†The New Zealand Institute for Plant & Food Research Limited
‡Department of Chemistry, University of Otago, Dunedin, New Zealand
§Department of Food Science, University of Otago, Dunedin, New Zealand
{Robinsdeme, Qi.Chen, Bing.Xue, Mengjie.Zhang}@ecs.vuw.ac.nz
daniel.killeen@plantandfood.co.nz, {sara.miller, keith.gordon, indrawati.oey}@otago.ac.nz

***Abstract*—Vibrational spectroscopy can be used for rapid determination of chemical quality markers in horticultural produce to improve quality control, optimize harvest times and maximize profits. Most commonly, spectral data are calibrated against chemical reference data (acquired using traditional, slower analytical methods) using partial least squares regression (PLSR). However, predictive performance of PLSR can be limited with the small number of instances, high dimensionality and collinearity of spectroscopic data. Here, a new genetic algorithm (GA) for PLSR feature and latent variable selection is proposed to predict concentrations of 18 important bioactive components across three New Zealand horticultural products from infrared, near-infrared and Raman spectral data sets. GA-enhanced PLSR models are produced from each spectroscopic data set individually. Models generated using the GA-enhanced PLSR method have notably better generalization and are less complex than the standard PLSR method. Fusing the data from the three individual sources substantially further enhances the prediction performance.**

***Index Terms*—Genetic Algorithm, Feature Selection, Partial Least Squares Regression, Vibrational Spectroscopy**

I. Introduction

Raw horticultural products vary widely in composition due mainly to factors such as genetics, cultivation region and storage conditions [1]. Therefore, quantifying the nutrient content of these products is important for quality control. Furthermore, with consumers becoming more health-conscious, the willingness to purchase nutrient-dense products increases accordingly. Meanwhile, many bioactive components affect taste, which can also influence individuals' consumption. Non-destructive, rapid assessment of these bioactive components is essential to provide the nutritional information of New Zealand horticultural products to consumers and thus maximise the profits. Such an approach is particularly desirable as current conventional wet chemistry methods, including high-performance liquid chromatography (HPLC), colorimetric assays and pH differential methods can be time-consuming, expensive and destructive [2]. These factors can preclude the application of said wet chemistry techniques in tasks requiring multiple measurements. In a production line environment, conventional methods are limited to assessing a subset of the material in a off-line manner, which can impede accuracy due to the time required to perform the analysis [3]. It is also difficult to make real time decisions to adjust processing conditions to improve product quality.

Vibrational spectroscopic techniques have been applied in previous studies attempting to quantify the phytochemical composition of New Zealand horticultural products [4], [5]. These techniques with their fingerprinting capabilities enable rapid, high-throughput and non-destructive nutrient quantification [6]. Such factors enable vibrational spectroscopic methods to be implemented on the production line. This permits the analysis of larger amounts of product and subsequently allows more accurate results. This is of particular relevance considering that storage conditions can introduce bias into measurements acquired by wet chemistry methods as they cannot be conducted in an on-line manner. Relating the vibrational spectroscopic data to the reference data of the phytochemical composition obtained through conventional wet chemistry methods can be done through the construction of a regression model. Such a model takes spectra with a range of wavelengths as the input variables, and the phytochemical components as the output/target variables. Previous studies have utilised partial least squares regression (PLSR) to achieve this [4], [5]. PLSR not only relates the spectral and reference data to each other through a linear multivariate model, but it also offers insight into both input and target variables' correlation through the generation of latent variables. Furthermore, PLSR is able to model data that is noisy, redundant and collinear; characteristics which are frequently observed in spectroscopy data [7].

Unfortunately, when the sample size is very small, it remains challenging to construct a PLSR model with high predictive accuracy. This issue is compounded when the data has severe multicollinearity or significant noise [8]. Previous stud-

ies [4], [5] found that the application of PLSR could indeed be restricted when data possessed these qualities. It is therefore desirable to assist PLSR to construct high-performing models, by manipulating the data such that the most useful information can be extracted from it. Feature selection [9] is a key method in data manipulation, which increases the data quality and reduces the dimensionality by selecting only the most useful/relevant features for modeling. However, feature selection is a challenging task due mainly to the large search space. Evolutionary computation (EC) techniques have been used for the task of feature selection [10], [11]. EC techniques, such as Genetic Algorithms (GAs) and Particle Swarm Optimisation (PSO), have previously been successfully applied numerous times to feature selection, outperforming full spectrum PLSR models [12], [13]. However, in these works, feature selection is considered based on the training performance of the regression models only, without considering any mechanism to address the potential limitation of over-fitting. It is sensible to enhance the prediction performance of PLSR via utilising a more sophisticated mechanism for evaluating the individuals in EC for feature selection while controlling the regression model complexity.

A. Goals

This study builds upon previous research which utilised both vibrational spectroscopy and PLSR to predict nutritional profiles of New Zealand horticultural products such as apricots, hops and plums. To overcome the aforementioned limitations of such spectral data, it is necessary to maximise the amount of useful information available in the data while maintaining a relatively low-complexity of model. This can be achieved through developing a new GA for feature selection and latent variable selection in PLSR.

In this work, a GA is utilised for the task of feature selection in favour of PSO. PSO may have limited performance in binary applications, as a result of its individuals' velocity and momentum values being derived in the continuous space [14]. These values must then be converted into the binary space, typically through a sigmoid function, which might reduce performance [15]. Conversely, a GA evolves individuals in the binary space, thus no conversions from continuous space are necessary and no information is lost.

The GA is also used to select the optimal number of latent variables to be included in the PLSR model, as a means of chasing compact but accurate projection and restricting model complexity. Models in which a single spectroscopy technique is used will be compared against models that utilise a variety of spectroscopy techniques, known as data fusion. Specific goals in this paper will include investigating the followings:

1) whether the performance of the new GA-enhanced PLSR model can outperform that of the original PLSR model. Such a comparison will reveal whether or not feature and latent variable selection can improve the predictive power of the PLSR model.
2) whether a data fusion model can further improve the performance relative to models from individual spectroscopy techniques. In such the data fusion model, all three spectroscopic techniques can be used simultaneously to predict a given bioactive component. It is hypothesised that doing so will improve the accuracy over any model constructed from a single spectroscopic technique.

B. Organisation

The remainder of this paper is organised as follows. Section II provides background information. Section III presents the proposed GA approach for feature selection and latent variable selection in PLSR. Section IV describes experiment design, while Section V presents the experiment results and corresponding discussions. Section IV provides the conclusions and future work.

II. Background

A. Vibrational Spectroscopy

Three particular techniques, all of which are firmly established in the area of food analysis, are analysed in the current study. These include Fourier transform (FT) Raman spectroscopy [16], FT-infrared (IR) spectroscopy [17] and near infrared (NIR) spectroscopy [18]. There exist fundamental differences between these spectroscopy techniques such that they can offer complementary information [19]. All spectra correspond to the infrared range of the spectrum and reflect changes in molecular vibrations. The spectra provide structural information that includes characteristic vibrations known as the fingerprint region as well as information on the functional groups [20].

The mechanism underlying infrared spectroscopy involves applying an infrared light source to a sample and detecting the wavelengths of light that are absorbed by fundamental molecular vibrations. Raman spectroscopy generates similar data relating to molecular vibrations, but uses a more complicated experiment set up involving single wavelength sample irradiation and the detection of inelastically scattered light. [21]. Mid-infrared (4,000-400 cm^{-1}) is particularly useful for identifying functional groups. Near-infrared (10,000-4,000 cm^{-1}) overtones and combinations vibrations [22], and can typically penetrate a sample further than MIR [23]. Raman (4000-15 cm^{-1}) also probes fundamental vibrations of functional groups, but often detects complementary information to infrared spectroscopy.

The spectral data obtained using these techniques is used as *features*, whereby each feature represents a spectral intensity at a given wavenumber. The number of features corresponds to the number of wavenumbers.

B. Spectral Pre-processing Methods

The manually pre-processed data was acquired using different pre-processing methods for each spectroscopic technique. Spectra were pre-processed using the software Unscrambler X 10.3 (CAMO, Oslo, Norway), with the exception of concave rubberband correction on Raman spectra that was carried out using OPUS 7.5 (Bruker Optics). Concave rubberband

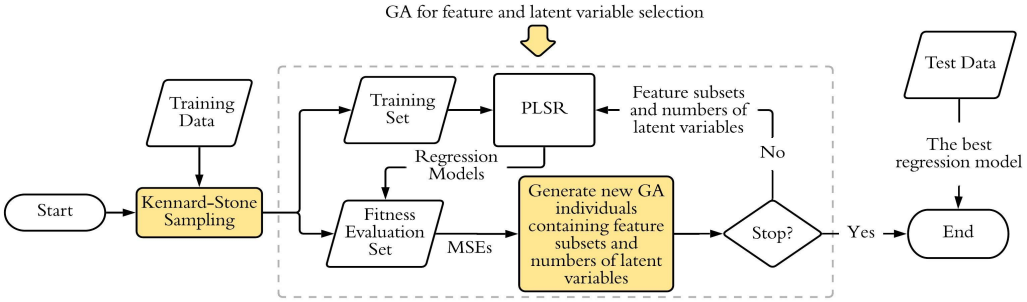

Fig. 1: The Overall Framework of the GA-enhanced PLSR Method.

correction is necessary to adjust the baseline of the spectral data. FT-Raman spectra also underwent the standard normal variate (SNV) transformation over the spectral region 1,800-300cm^{-1} which is used for scatter correction. FT-IR spectra also had SNV pre-processing applied over the spectral region 1,800-630cm^{-1}. The FT-NIR data underwent second derivative (Savitzky–Golay, second-order polynomial, 5-point window) followed by SNV over the spectral region 1670–920nm.

C. Partial Least Squares Regression (PLSR)

Partial Least Squares Regression (PLSR) aims to find a linear regression model that transforms the input features/variables (spectral data) and the target variables (reference data) into *latent variables*. This is achieved through searching for the multidimensional direction of the input features/variable space that maximises the multidimensional variance direction in the target variables. The resultant latent variables model the covariance of the input features/variables and are regressed against the target variables [24]. Given the features/input variables X, and the target variable(s) Y, the PLSR models are typically represented as:

$$X = TP^T + E_1$$
$$Y = UQ^T + E_2$$

where T and U are projections of X and projections of Y, respectively. P and Q are orthogonal loading matrices; and E_1 and E_2 are the error terms, which are assumed to be random normal variables.

D. Genetic Algorithms

Genetic algorithms are an evolutionary computation technique created by John Holland and his collaborators in the 1960s [25]. A GA imitates the natural selection process wherein a population of candidate solutions known as individuals is evolved to solve an optimisation problem. Each individual carries a chromosome is a candidate solution to the target task. The chromosomes can be mutated and altered through processes such as crossover and mutation. The chromosome itself is typically represented in binary, as a string of 0s and 1s. Each individual's chromosome is randomly initialised. The chromosome values will be modified in an iterative process with each iteration referred to as a generation. At each generation, the fitness of each individual is evaluated.

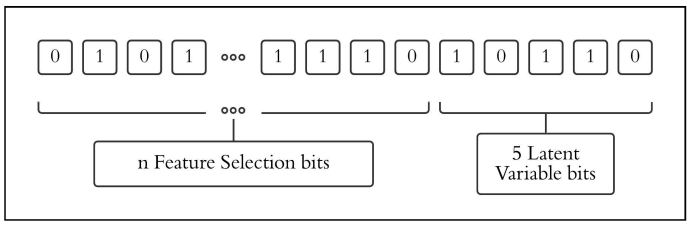

Fig. 2: The chromosome in the new GA method.

This fitness function should reflect how well the target task is being addressed by each individual. Individuals are selected from the current population based on a particular selection method with a preference to better fitted individuals. Their chromosomes are modified via recombination or crossover, and mutation operations and then added to the next generation. The algorithm will continue until a termination criterion is met, typically a maximal number of generations or a sufficiently high fitness level.

III. The Proposed Method

As mentioned in the previous section that the vibrational spectroscopic data usually has a small number of samples, while the features are highly collinear. To maximise the amount of useful information while maintaining a relatively low-complexity of partial least squares (PLSR) model in vibrational spectroscopic data for predicting nutritional profiles, this work proposes a new method for feature selection and latent variable selection based on a GA. The overall framework of the GA-enhanced PLSR method is shown in Fig. 1. The proposed GA method has been utilised for selecting the features to feed to the learning process in PLSR while it also selects the number of latent variables used during this process. This is fulfilled by a new representation and a new fitness evaluation method. The details of the method are presented in the following sections.

A. Representation in the New GA based Method

As shown in Fig. 2, the chromosomes in the new GA based method are represented by a string consist of two parts. The first n bits of binary values are reserved for feature selection, and n is the number of features. The following l bits represents a binary value corresponding to the number of latent variables

in the PLSR model. The value of l depends on the maximal number of latent values used in the PLSR model. In this work, the maximal number of latent variable is set to 32, thus the value of l is set to $l = 5$ since $2^5 = 32$.

B. The Fitness Evaluation Method

1) Kennard-Stone Sampling to generate the fitness evaluation data set for the GA individuals: The vibrational spectroscopic data usually has a small number of instances. As a result, it is probable that any model constructed from the training data may not elucidate all information that is present in the unseen data. Evaluating the fitness of the GA individuals directly from the training performance of the PLSR models can potentially introduce a lower bias error but a higher variance error, which indicates a poor generalisation on the unseen data. To avoid this limitation, in this work, the training data is split into the training set and the fitness evaluation set. The training set is used for constructing the PLSR model, while the fitness evaluation set, consists of a portion of the training data which never expose to the training of the PLSR models, is therefore used as the fitness evaluation set for GA individuals. In other words, each GA individual is decoded to a set of selected features and latent variables. The training set will be transformed according to the decoded features/variables. PLSR will be trained on the transformed training set. The performance of the trained PLSR model is evaluated on the fitness evaluation set, and the obtained performance is used as the fitness value of the GA individual.

Kennard-Stone sampling [26], which allows to select samples with a uniform distribution over the feature space, is used to divide the original training data into a training set for learning PLSR models and a fitness evaluation set for GA individuals. Kennard-Stone operates by iteratively finding the most separated points in the training data. The metric for separation is often Euclidean distance. This approach aims to allocate representative and diverse samples into the training set, with the remaining samples used in the fitness evaluation set. The chance of missing valuable information, which is needed for the model to generalise to unseen data, should therefore theoretically be reduced.

2) The Fitness Function: The fitness function of the new GA method has three components that it attempts to minimise the mean squared error (MSE) of the PLSR model, the number of selected features and the number of latent variables in PLSR. The former one is included as a metric for the regression performance. The latter two are included so as to reduce the complexity of PLSR models. The fitness function is shown as follows:

$$fitness = w_1 \times MSE + w_2 \times \#Features + (1 - w_1 - w_2) \times \#LVs \quad (1)$$

where MSE is the mean square error of the regression model on the fitness evaluation set, #Features is the number of features, which is to be minimised. In such a way features that do not offer much useful information should hypothetically be excluded from the model. #LVs refers to the number of Latent variables, which is minimised so that the PLSR model will use a small number of latent variables, or components, that will further reduce the model complexity of PLSR models. The two parameters w_1, w_2 in the fitness function are set to 0.7, 0.2, respectively. These values are set via trail and error with a bias to a larger value of w_1 because the regression performance is the most important component. The weights affect the influence of each component on the overall fitness of the individual. w_1 is weighted more highly as MSE is considered the most significant to the overall fitness. Specifically, there is no sense in minimising the number of features or latent variables at the expense of constructing an extremely low accuracy model. The fitness function therefore is a means of balancing the accuracy and complexity of the model.

In most cases, it is argued that a univariate PLSR model, with one target variable, yields better predictive accuracy than a multivariate PLSR model with multiple target variables [27]. As such, to gain optimal predictive performance, univariate PLSR models, which are also known as PLS1, will be constructed for each individual target variable [28].

IV. Experiment Design

To examine the performance of the proposed GA-enhanced PLSR method, various experiments have been conducted using three vibrational spectroscopic data sets.

A. The Benchmark Technique

The original PLSR technique is used as the benchmark technique in the experiments. To allow for fair comparisons, 50 runs using the same set of random seeds are undertaken for both the GA-enhanced PLSR approach and the original PLSR approach. In original PLSR, 10 latent variables are used to construct PLSR models for comparisons, which was found to yield the best overall performance on test data in our preliminary experiments.

B. Data Sets and Reference Data

Table I shows the three data sets used in experiments. All spectroscopic data sets were acquired at Otago University, New Zealand. Reference data were generated by Otago University Food Science Department for the *Plums* and *Apricot* data sets using the methods described in McIntyre et al. [5] Reference data was generated by The New Zealand Institute for Plant and Food Research for the *Hops* data as described in Killeen et al [4]. Each data set was pre-processed such that noise was removed. Additionally, areas of the spectra containing suspected redundant or unimportant information were removed.

In each data set, three different Vibrational Spectroscopy techniques including Fourier transform (FT) Raman spectroscopy (Raman), FT-infrared (IR) spectroscopy and near infrared (NIR) spectroscopy have been utilised to obtain three sets of spectral data. Reference data, which measures the bioactive component, was collected using wet chemistry methods. The bioactive component(s), which are the target

TABLE I: Data Sets

	Spectroscopy Technique	# Features	# Instances
Apricots	IR	637	99
	NIR	122	
	Raman	2440	
Hops	IR	679	139
	NIR	429	
	Raman	1524	
Plums	IR	520	216
	NIR	122	
	Raman	1659	

TABLE II: Bioative Components in Each Data Set

	Bioactive Components **(Target Variables)**
Apricots	Vitamin C, Phenolics, Carotenoids
Hops	Alpha Acids (HPLC), Alpha Acids (UV), Beta Acids (HPLC), Beta Acids (UV), Total Acids (HPLC), Total Acids (UV), Cohumulone, Colupulone, Humulone, Lupulone, Xanthohumo
Plums	Anthocyanin, Antioxidants, Phenolics, Vitamin C

variable(s) that should be predicted in the regression tasks in each data set, are summaries in Table II. This work treats the bioactive components in each data set independently, e.g. for each data set, the number of PLSR models that are needed be constructed is equal to the types of bioactive components.

C. Data Fusion

In this work, we also would like to investigate the performance of PLSR methods to predict a given bioactive component when all the three spectroscopic techniques are used concurrently, whereas previously only one spectroscopic technique would be used each time. This setting will take advantage of the fact that all three aforementioned spectroscopy techniques provide complementary information where each technique targets different regions of the spectrum. Implementing this data fusion in conjunction with feature selection is anticipated to yield improved predictive power, as this fused model hypothetically includes the greatest amount of useful information in the features it is constructed from.

D. Parameter Settings

In the experiments, all instances in each data set are randomly divided into two sets. As illustrated in Fig. 3, 80% will form the training set and the remaining 20% will from the test set. 50 random splits are used for the 50 runs of experiments in each method, one split every run. This enables different instances to be assigned to the training and test sets each time in the splitting process.

For the GA-enhanced PLSR method, as is also shown in Fig. 3, the training set is further divided into two sets. 75% of the training data will be the training set and the rest 25% will be the fitness evaluation set for GA. This division is carried out using Kennard-Stone sampling as described in Section III-B1.

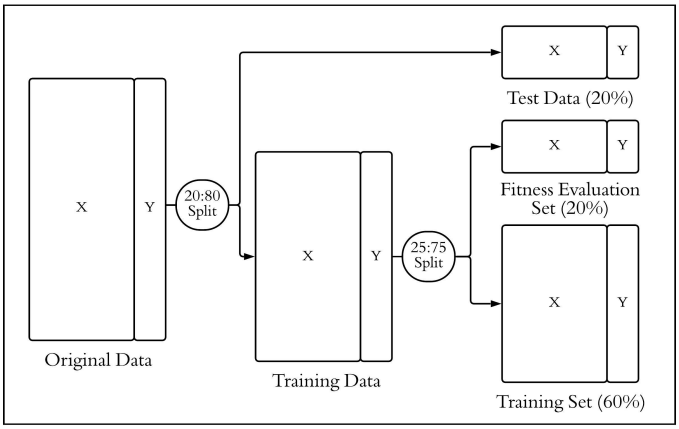

Fig. 3: Illustrates how the original data is divided in to the test, fitness evaluation and training sets with a 20:20:60 split.

In the GA approach, the population size $P = 5$ and the maximum number of generations $G = 5$ are used. Tournament selection with a tournament size of 3 is used. The crossover rate and the mutation rate are set to 0.5 and 0.2, respectively. These parameter settings were found to be the best combination in several trails of preliminary experiments.

V. Results and Discussions

The experiment results of the two PLSR methods on the three data sets are presented and analysed in this section. Tables III, IV and V show the performance of the two PLSR methods on each spectroscopic technique respectively. Table VI shows their performance on the fused data sets. For an easier comparison, following previous related work [4], [5], the R-squared (denoted as R^2) and the MSE of the PLSR models on the training sets and the test sets are both reported. For these two performance metrics, the larger R^2 the better, while the smaller the MSE the better. The Mann Whitney U test [29], which is a non-parametric statistical significance test, with a significant level of 0.05, has been used for test the difference in terms of both R^2s and MSEs between the two methods in each data set in pairs. The significantly better results are shown in bold in the tables.

A. Comparisons between GA-enhanced PLSR and PLSR on Each Spectroscopic Technique

On the training set, as shown in Tables III, IV and V, the GA-enhanced PLSR has worse training performance than the original PLSR in many cases of the three data sets. Among the three spectroscopic techniques, the patterns are slightly different. When using IR spectroscopy, the new PLSR method has a notably better training performance on assessing the components of Carotenoids and Phenolics in the Apricots data set and the Anthocyanin and the Antioxidants in the Plums data set, which are all significant. The GA-enhanced PLSR has slightly but not significantly worse training performance than PLSR on the six Hops cases.

TABLE III: IR Models

Data Set	BioActive Component	Training				Test			
		PLSR		GA-enhanced PLSR		PLSR		GA-enhanced PLSR	
		R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Apricots	Carotenoids	0.78±0.03	14.51±1.37	**0.90±0.23**	**9.11±20.79**	-0.50±1.24	67.45±27.59	**0.96±0.13**	**0.36±1.24**
	Phenolics	0.70±0.03	176.95±16.46	**0.95±0.06**	**31.52±33.42**	0.15±0.34	471.80±152.90	**1.00±0.00**	**0.01±0.05**
	Vitamin C	0.84±0.02	1.23±0.11	0.73±0.23	2.05±1.80	0.24±0.48	4.51±1.70	**0.94±0.10**	**0.25±0.41**
Hops	Alpha Acids (HPLC)	0.96 ±0.00	0.61±0.05	0.94±0.07	1.00±1.08	0.89 ±0.04	1.63±0.56	**0.96±0.07**	**0.53±1.04**
	Alpha Acids (UV)	0.96 ±0.00	0.73±0.06	0.95±0.05	1.05 ±0.91	0.89 ±0.04	1.90±0.55	**0.98±0.03**	**0.04±0.69**
	Beta Acids (HPLC)	**0.89 ±0.01**	**0.19±0.02**	0.83 ±0.12	0.31 ±0.23	0.67 ±0.13	0.55±0.18	**0.89±0.12**	**0.15±0.19**
	Beta Acids (UV)	0.89 ±0.01	0.25±0.02	0.87 ±0.10	0.33 ±0.26	0.68 ±0.14	0.70±0.25	**0.98±0.03**	**0.11±0.15**
	Total Acids (HPLC)	0.95 ±0.00	1.10±0.08	0.93 ±0.07	1.81 ±1.65	0.87 ±0.06	2.93±1.01	**0.96±0.07**	**0.88±1.63**
	Total Acids (UV)	0.95 ±0.00	1.30±0.09	0.95 ±0.05	1.59 ±1.36	0.87 ±0.05	3.36±1.06	**0.98±0.04**	**0.56±1.25**
	Cohumulone	**0.97 ±0.00**	**0.05±0.00**	0.92 ±0.09	0.13 ±0.14	0.92 ±0.03	0.11±0.04	**0.93±0.11**	**0.08±0.13**
	Colupulone	**0.93 ±0.01**	**0.05±0.00**	0.89 ±0.08	0.09 ±0.07	0.80 ±0.08	0.15±0.05	**0.87±0.15**	**0.09±0.10**
	Humulone	**0.95 ±0.00**	**0.39±0.03**	0.90 ±0.09	0.83 ±0.77	0.87 ±0.06	1.05±0.38	**0.96±0.08**	**0.45±0.84**
	Lupulone	0.86 ±0.01	0.06±0.00	0.74 ±0.23	0.10 ±0.09	0.53 ±0.22	0.16±0.06	**0.91±0.14**	**0.02±0.04**
	Xanthohumol	**0.94 ±0.01**	**0.00±0.00**	0.90 ±0.08	0.01 ±0.00	0.86 ±0.07	0.01±0.00	**0.93±0.10**	**0.00±0.00**
Plums	Anthocyanin	0.96 ±0.00	29.76±1.67	**0.99 ±0.01**	**5.80±3.52**	0.90 ±0.04	67.42±25.97	**1.00±0.00**	**0.01±0.09**
	Antioxidants	0.61 ±0.02	15.61±1.46	**0.90±0.14**	**2.29±3.07**	0.32 ±0.25	23.17±6.51	**0.93±0.19**	**6.43±17.06**
	Phenolics	**0.86 ±0.01**	**0.00±0.00**	0.63±0.19	0.01±0.00	0.72 ±0.11	0.01±0.00	**0.96±0.07**	**0.00±0.00**
	Vitamin C	**0.92 ±0.00**	**0.13±0.01**	0.79±0.12	0.19±0.11	0.84 ±0.12	0.25±0.20	0.95±0.08	**0.01±0.02**

TABLE IV: NIR Models

Data Set	Bioactive Component	Training				Test			
		PLSR		GA-enhanced PLSR		PLSR		GA-enhanced PLSR	
		R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Apricots	Carotenoids	**0.93±0.01**	**4.32±0.73**	0.77±0.10	16.19±7.72	-1.30±1.50	110.46±29.33	**1.00±0.00**	**0.01±0.04**
	Phenolics	**0.95±0.01**	**28.16±6.68**	0.83±0.13	109.38±81.38	-0.53±0.57	838.76±221.85	**0.99±0.03**	**5.22±12.53**
	Vitamin C	**0.97±0.01**	**0.25±0.04**	0.80±0.08	1.24±0.53	-0.10±0.53	6.76±2.20	**1.00±0.01**	**0.05±0.01**
Hops	Alpha Acids (HPLC)	**0.99±0.00**	**0.13±0.02**	0.91 ±0.11	1.39 ±1.63	0.74±0.08	2.94±1.16	**0.96±0.07**	**0.68±1.28**
	Alpha Acids (UV)	**0.99±0.00**	**0.15±0.03**	0.94±0.05	1.05±0.90	0.75±0.07	4.41±1.19	**0.98±0.02**	**0.35±0.49**
	Beta Acids (HPLC)	**0.97±0.00**	**0.05±0.01**	0.84 ±0.16	0.28 ±0.28	-0.12±0.70	1.85±1.23	**0.93±0.13**	**0.10±0.19**
	Beta Acids (UV)	**0.98±0.01**	**0.06±0.01**	0.83 ±0.17	0.41 ±0.41	0.09±0.64	2.06±1.52	**0.92±0.14**	**0.13±0.22**
	Total Acids (HPLC)	**0.99±0.00**	**0.17±0.03**	0.93 ±0.09	1.63 ±2.16	0.77±0.09	5.14±1.77	**0.97±0.06**	**0.66±1.61**
	Total Acids (UV)	**0.99±0.00**	**0.16±0.03**	0.96 ±0.05	1.12 ±1.33	0.82±0.07	4.92±1.80	**0.99±0.02**	**0.32±0.72**
	Cohumulone	**0.99±0.00**	**0.02±0.00**	0.87 ±0.17	0.17 ±0.23	0.53±0.17	0.60±0.17	**0.90±0.17**	**0.12±0.19**
	Colupulone	**0.98±0.00**	**0.02±0.00**	0.79 ±0.19	0.15 ±0.14	-0.08±0.59	0.78±0.41	**0.78±0.27**	**0.15±0.18**
	Humulone	**0.99±0.00**	**0.10±0.02**	0.79 ±0.19	1.43 ±1.63	0.58±0.14	3.40±1.23	**0.90±0.15**	**1.02±1.57**
	Lupulone	**0.96±0.01**	**0.02±0.00**	0.70 ±0.30	0.13 ±0.12	-0.47±0.75	0.50±0.26	**0.89±0.18**	**0.03±0.05**
	Xanthohumol	**0.98±0.00**	**0.00±0.00**	0.84 ±0.17	0.01 ±0.01	0.37±0.27	0.03±0.01	**0.91±0.14**	**0.00±0.00**
Plums	Anthocyanin	**0.97±0.00**	**20.59±1.66**	0.96 ±0.01	22.83±5.44	0.85±0.04	97.65±28.14	**0.99±0.01**	**2.02±2.21**
	Antioxidants	**0.86±0.02**	**5.63±0.52**	0.82±0.05	3.41±0.92	0.28±0.38	22.72±5.23	**0.97±0.03**	**2.93±2.93**
	Phenolics	**0.92±0.01**	**0.00±0.00**	0.68±0.15	0.01±0.00	0.67±0.10	0.01±0.00	**0.92±0.12**	**0.00±0.00**
	Vitamin C	**0.93±0.01**	**0.11±0.01**	0.73±0.15	0.23±0.13	0.75±0.08	0.39±0.10	**0.87±0.16**	**0.04±0.05**

When using the other spectroscopic methods, i.e. NIR and Raman, the GA-enhanced PLSR has a worse training performance than PLSR on all the three data sets, most of which are significant. The disadvantage of the new GA-enhanced PLSR on the training performance is actually not surprising, since the PLSR model in the GA-enhanced method is trained on a portion of the training set only. Moreover, the new GA method controls the model complexity in PLSR via feature and latent variable selection. These additional objectives in the fitness function are potentially conflicting with minimising the model training error only, particularly when the regression model has an over-fitting trend.

On the more important test set performance, the pattern is considerably different from that on the training set. The GA-enhanced PLSR offers notable improvements on all three test sets for all of the 18 target variables /bioactive components consistently. These are shown with much higher R^2 values and notably lower MSEs in all the cases. In almost all the comparisons on the test sets, the GA-enhanced PLSR has significantly better results on both R^2s and MSEs than original PLSR using the three spectroscopic technique except for Vitamin C on Plums. Compared with the original PLSR, which has a relatively low R^2 in many cases, the GA-enhanced PLSR has an R^2 around 1 in many cases. This indicates that the GA-enhanced PLSR can explain a high proportion (around 100% in many cases) of the variance for the target variable. Note that in some cases, original PLSR has a negative value of R^2. This means that it has a worse generalisation performance than the simple regression model that takes the average target value as the prediction value. The superior performance of the GA-enhanced PLSR over original PLSR confirms the positive effect of the new GA method on enhance the generalisation of PLSR via feature and latent variable selection.

Comparing the results using the three spectroscopy techniques, Raman appears to be the most promising technique for quantifying bioactive content on the three examined data

TABLE V: Raman Models

Data Set	Model	Training				Test			
		PLSR		GA-enhanced PLSR		PLSR		GA-enhanced PLSR	
		R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Apricots	Carotenoids	1.00±0.00	0.00±0.00	1.00±0.02	0.35±1.13	0.13±0.71	41.71±19.80	**1.00±0.00**	**0.00±0.01**
	Phenolics	**1.00±0.00**	0.04±0.01	0.97±0.11	19.89±69.89	0.15±0.27	480.69±125.34	**0.99±0.05**	**3.27±19.74**
	Vitamin C	**1.00±0.00**	0.00±0.00	0.80±0.08	1.24±0.53	0.04±0.64	5.83±2.64	**1.00±0.01**	**0.01±0.05**
Hops	Alpha Acids (HPLC)	**0.99±0.00**	**0.21±0.03**	0.95 ±0.05	0.64 ±0.71	0.74±0.08	3.94±1.16	**0.98±0.04**	**0.32±0.62**
	Alpha Acids (UV)	**0.99±0.00**	**0.24±0.04**	0.95±0.06	0.86 ±0.95	0.73±0.08	4.74±1.16	**0.97±0.05**	**0.55±0.99**
	Beta Acids (HPLC)	**0.98±0.00**	**0.04±0.01**	0.82 ±0.18	0.29 ±0.29	0.48±0.18	0.85±0.22	**0.89±0.19**	**0.15±0.26**
	Beta Acids (UV)	**0.98±0.00**	**0.04±0.01**	0.87 ±0.12	0.29 ±0.27	0.57±0.15	2.06±1.52	**0.93±0.14**	**0.12±0.23**
	Total Acids (HPLC)	**0.99±0.00**	**0.32±0.05**	0.93 ±0.07	1.58 ±1.47	0.68±0.10	7.05±1.48	**0.97±0.05**	**0.84±1.25**
	Total Acids (UV)	**0.99±0.00**	**0.37±0.06**	0.94 ±0.05	1.57 ±1.24	0.72±0.09	7.46±1.74	**0.97±0.04**	**0.95±1.31**
	Cohumulone	**0.98±0.00**	**0.02±0.00**	0.86 ±0.15	0.18 ±0.19	0.66±0.08	0.43±0.09	**0.88±0.15**	**0.13±0.18**
	Colupulone	**0.98±0.00**	**0.02±0.00**	0.79±0.16	0.14±0.11	0.50±0.18	0.36±0.10	**0.80±0.25**	**0.13±0.17**
	Humulone	**0.98±0.00**	0.16±0.03	0.93 ±0.08	0.49 ±0.57	0.62±0.11	3.09±0.77	**0.97±0.05**	**1.02±1.57**
	Lupulone	**0.97±0.01**	**0.01±0.00**	0.74 ±0.25	0.10±0.10	0.27±0.26	0.25±0.07	**0.90±0.16**	**0.03±0.04**
	Xanthohumol	**0.99±0.00**	**0.00±0.00**	0.92 ±0.09	0.00 ±0.00	0.76±0.08	0.01±0.00	**0.93±0.11**	**0.00±0.00**
Plums	Anthocyanin	**1.00±0.00**	0.45±0.07	0.99 ±0.012	5.27±10.30	0.84±0.05	100.87±24.30	**1.00±0.00**	**0.12±0.50**
	Antioxidants	**1.00±0.00**	**0.14±0.03**	0.89±0.14	2.09±2.73	-0.12±0.43	37.71±9.88	**0.97±0.07**	**2.59±6.27**
	Phenolics	**1.00±0.00**	**0.00±0.00**	0.84±0.23	0.00±0.00	0.19±0.22	0.02±0.00	**0.98±0.06**	**0.00±0.00**
	Vitamin C	1.00±0.00	0.00±0.00	0.88±0.17	0.11±0.16	0.57±0.09	0.68±0.11	**0.96±0.08**	**0.01±0.02**

TABLE VI: IR + NIR + Raman Model

Data Set	Bioactive Component	Training				Test			
		PLSR		GA-enhanced PLSR		PLSR		GA-enhanced PLSR	
		R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Apricots	Carotenoids	1.00±0.00	0.01±0.00	0.98±0.06	1.95±4.67	0.41±0.42	29.43±12.20	**1.00±0.01**	**0.01±0.06**
	Phenolics	**1.00±0.00**	**0.10±0.03**	0.97±0.08	17.76±50.35	0.20±0.30	442.58±127.95	**1.00±0.01**	**0.68±2.18**
	Vitamin C	**1.00±0.00**	0.00±0.00	0.94±0.16	0.44±1.14	0.33±0.43	4.08±1.67	**0.99±0.05**	**0.05±0.19**
Hops	Alpha Acids (HPLC)	0.99±0.00	0.12±0.02	0.97±0.04	0.42±0.61	0.90±0.04	1.53±0.48	**0.99 ±0.01**	**0.11 ±0.22**
	Alpha Acids (UV)	0.99±0.00	0.12±0.02	0.97±0.04	0.51±0.71	0.90±0.03	1.70±0.53	**0.99±0.01**	**0.13 ±0.23**
	Beta Acids (HPLC)	**0.98±0.00**	**0.03±0.01**	0.89±0.13	0.19±0.22	0.70±0.14	0.49±0.21	**0.93 ±0.11**	**0.10 ±0.15**
	Beta Acids (UV)	**0.99±0.00**	**0.03±0.01**	0.93±0.09	0.16±0.22	0.78±0.09	0.48±0.17	**0.97 ±0.07**	**0.06 ±0.11**
	Total Acids (HPLC)	**0.99±0.00**	**0.21±0.03**	0.97±0.05	0.79±1.09	0.88±0.05	2.65±0.96	**0.99 ±0.03**	**0.28 ±0.65**
	Total Acids (UV)	**0.99±0.00**	**0.19±0.03**	0.97±0.04	0.98±1.16	0.91±0.04	2.51±0.83	**0.99 ±0.02**	**0.38 ±0.72**
	Cohumulone	0.99±0.00	0.01±0.00	0.96±0.05	0.05±0.08	0.89±0.04	0.14±0.04	**0.97 ±0.05**	**0.03 ±0.06**
	Colupulone	**0.99±0.00**	0.01±0.00	0.95±0.06	0.04±0.05	0.81±0.08	0.14±0.06	**0.96 ±0.09**	**0.03 ±0.06**
	Humulone	**0.99±0.00**	0.08±0.01	0.96±0.07	0.35±0.58	0.87±0.05	1.05±0.39	**0.99 ±0.04**	**0.14 ±0.41**
	Lupulone	**0.97±0.00**	**0.01±0.00**	0.83±0.24	0.07±0.10	0.60±0.18	0.14±0.06	**0.95 ±0.09**	**0.01 ±0.02**
	Xanthohumol	**0.99±0.00**	**0.01±0.00**	0.96±0.05	0.00±0.00	0.87±0.06	0.01±0.00	**0.98 ±0.05**	**0.00 ±0.00**
Plums	Anthocyanin	1.00±0.00	0.68±0.15	0.99±0.01	4.01±6.93	0.89±0.02	69.11±11.25	**1.00 ±0.00**	**0.08±0.25**
	Antioxidants	**0.99±0.00**	**0.21±0.03**	0.91±0.13	1.78±2.65	0.14±0.35	28.78±7.29	**0.98±0.05**	**2.01±4.84**
	Phenolics	**1.00±0.00**	**0.01±0.00**	0.85±0.22	0.00±0.00	0.60±0.11	0.01±0.00	**0.98±0.04**	**0.00±0.00**
	Vitamin C	**1.00±0.00**	**0.01±0.00**	0.94±0.11	0.06±0.10	0.77±0.05	0.36±0.07	**0.98±0.05**	**0.01±0.02**

sets. Both of the two PLSR methods have best prediction performance when approximating the target variable using the Raman data. This is consistent with the findings in previous work [4], [5]. We also note that NIR is the least promising technique, where both the original and GA-enhanced PLSR methods have worse regression performance in most cases than when using either of the other two spectroscopy techniques.

Comparing the performance of the proposed GA-enhanced PLSR method on assessing the bioactives content on the three data sets, the improvement on the Apricots data set is greater than those obtained for the remaining two data sets. This could be attributed to the Apricots spectral data possessing a greater amount of redundancy or background noise in measurements. The feature selection approach was able to extract useful features from this spectral data, and removing irrelevant or noisy features, and the latent variable selection enabled an optimal number of latent variables to be included in the PLSR model.

B. Comparisons between GA-enhanced PLSR and PLSR on Data Fusion

The results of the two PLSR methods on the fused data of the three spectroscopy techniques are shown in Table VI. As it shows, the overall pattern found from the comparison of the two PLSR methods is similar to that obtained when using a solo spectroscopy technique. The GA-enhanced PLSR method still has a worse training performance but a much better test performance than PLSR on all the three data sets.

Both PLSR methods, when trained on fused data as opposed to using any given spectroscopy technique independently, demonstrate a marked improvement on the training and the test performance. This is due to taking advantage of the complementary nature of the three spectroscopic techniques, whereby the maximal amount of useful information is made available to the PLSR model therefore enabling increased predictive performance on the test data. The GA-enhanced PLSR achieves more than 90% predictive accuracy for all bioactive

components across all data sets. Through this approach, the best predictive accuracy on the test data is achieved.

VI. Conclusions and Future Work

This paper has explored the effect of using a GA-based approach for feature and latent variable selection for developing PLSR models that quantify bioactive components in New Zealand fruits and plants from spectral data. The experiments have confirmed that our new GA method enhanced the construction of PLSR models by selecting feature subsets and a smaller number of latent variables that offered a notably positive effort on enhancing the generalisation performance of PLSR models and reducing the model complexity. The comparisons also highlight the particular value of Raman spectroscopy in quantifying the phytochemical composition of New Zealand horticultural products.

Additionally, this study investigated the impact of data fusion, which proved to be fruitful. The complementary nature of these spectral methods, whereby each is capable of offering unique information not provided by the others, is likely what enables these improvements.

The PLSR models in this work learned from the manually pre-processed spectral data which required human domain knowledge. Future work will explore the algorithm's performance when using raw spectral data directly, by implementing an automatic selection of pre-processing methods.

Acknowledgement

This work is supported in part by the MBIE Endeavour Fund on Research Program under the contract of C11X2001 and Marsden Fund under the contract number of VUW1913 and VUW2016. We would also like to thank our project leaders Sue Marshall at Plant and Food Research and Jeremy Rooney at University of Otago for their support.

References

[1] Shewfelt, R., 1990. Sources of variation in the nutrient content of agricultural commodities from the farm to the consumer. Journal of Food Quality, 13(1), pp.37-54.

[2] Leong, S. & Oey, I., 2012. Effects of processing on anthocyanins, carotenoids and vitamin C in summer fruits and vegetables. Food Chemistry, 133(4), pp.1577-1587.

[3] Kucha, C., Liu, L., & Ngadi, M. (2018). Non-Destructive Spectroscopic Techniques and Multivariate Analysis for Assessment of Fat Quality in Pork and Pork Products: A Review. Sensors, 18(2), p.377.

[4] Killeen, D., Andersen, D., Beatson, R., Gordon, K. & Perry, N., 2014. Vibrational Spectroscopy and Chemometrics for Rapid, Quantitative Analysis of Bitter Acids in Hops (Humulus lupulus). Journal of Agricultural and Food Chemistry, 62(52), pp.12521-12528.

[5] McIntyre, S., Ma, Q., Burritt, D., Oey, I., Gordon, K. & Fraser-Miller, S., 2020. Vibrational spectroscopy and chemometrics for quantifying key bioactive components of various plum cultivars grown in New Zealand. Journal of Raman Spectroscopy, 51(7), pp.1138-1152.

[6] Hackshaw, K., Miller, J., Aykas, D. & Rodriguez-Saona, L., 2020. Vibrational Spectroscopy for Identification of Metabolites in Biologic Samples. Molecules, 25(20), p.4725.

[7] Amuah, C., Teye, E., Lamptey, F., Nyandey, K., Opoku-Ansah, J. & Adueming, P., 2019. Feasibility Study of the Use of Handheld NIR Spectrometer for Simultaneous Authentication and Quantification of Quality Parameters in Intact Pineapple Fruits. Journal of Spectroscopy, 2019, pp.1-9.

[8] Spiegelman, C., McShane, M., Goetz, M., Motamedi, M., Yue, Q. & Coté, G., 1998. Theoretical Justification of Wavelength Selection in PLS Calibration: Development of a New Algorithm. Analytical Chemistry, 70(1), pp.35-44.

[9] Xue, B., Zhang, M., Browne, W. N., & Yao, X., "A Survey on Evolutionary Computation Approaches to Feature Selection," in IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, pp.606-626, Aug. 2016.

[10] Chen, Q., Zhang, M. & Xue, B., "Feature Selection to Improve Generalization of Genetic Programming for High-Dimensional Symbolic Regression," in IEEE Transactions on Evolutionary Computation, vol. 21, no. 5, pp.792-806, Oct. 2017.

[11] Al-Sahaf, H., Bi, Y., Chen, Q., Lensen, A., Mei, Y., Sun, Y., Tran, B., Xue, B. & Zhang, M., 2019. A survey on evolutionary machine learning. Journal of the Royal Society of New Zealand, 49(2), pp.205-228.

[12] Leardi, R., 2000. Application of genetic algorithm-PLS for feature selection in spectral data sets. Journal of Chemometrics, 14(5-6), pp.643-655.

[13] Xue, L., Cai, J., Li, J., & Liu, M., 2012. Application of Particle Swarm Optimization (PSO) Algorithm to Determine Dichlorvos Residue on the Surface of Navel Orange with Vis-NIR Spectroscopy. Procedia Engineering, 29, pp.4124-4128.

[14] Lee, S., Soak, S., Oh, S., Pedrycz, W., & Jeon, M., 2008. Modified binary particle swarm optimization. Progress In Natural Science, 18(9), pp.1161-1166.

[15] Nguyen, B., Xue, B., Andreae, P., & Zhang, M., 2021. A New Binary Particle Swarm Optimization Approach: Momentum and Dynamic Balance Between Exploration and Exploitation. IEEE Transactions On Cybernetics, 51(2), pp.589-603.

[16] Zhu, X., Xu, T., Lin, Q. & Duan, Y., 2014. Technical development of Raman spectroscopy: from instrumental to advanced combined technologies. Applied Spectroscopy Reviews, 49(1), pp.64-82.

[17] Paterova, A., Lung, S., Kalashnikov, D.A. & Krivitsky, L.A., 2017. Nonlinear infrared spectroscopy free from spectral selection. Scientific reports, 7(1), pp.1-8.

[18] Manley, M. & Baeten, V., 2018. Spectroscopic technique: Near infrared (NIR) spectroscopy. Modern techniques for food authentication, Academic Press, pp.51-102.

[19] Hashimoto, K., Badarla, V., Kawai, A. & Ideguchi, T., 2019. Complementary vibrational spectroscopy. Nature Communications, 10(1).

[20] Dyer, R., & Woodruff, W., 2011. Vibrational Spectroscopy. Encyclopedia Of Inorganic And Bioinorganic Chemistry.

[21] Matthäus, C., Bird, B., Miljković, M., Chernenko, T., Romeo, M., & Diem, M., 2008. Chapter 10 Infrared and Raman Microscopy in Cell Biology. Methods In Cell Biology, pp.275-308.

[22] Grabska, J., Beć, K., Kirchler, C., Ozaki, Y. & Huck, C., 2019. Distinct Difference in Sensitivity of NIR vs. IR Bands of Melamine to Inter-Molecular Interactions with Impact on Analytical Spectroscopy Explained by Anharmonic Quantum Mechanical Study. Molecules, 24(7), p.1402.

[23] Ash, C., Dubec, M., Donne, K. & Bashford, T., 2017. Effect of wavelength and beam width on penetration in light-tissue interaction using computational methods. Lasers in medical science, 32(8), pp.1909-1918.

[24] Rosipal, R., & Krämer, N., 2006. Overview and Recent Advances in Partial Least Squares. Subspace, Latent Structure And Feature Selection, pp.34-51.

[25] Yang, X., 2014. Genetic Algorithms. Nature-Inspired Optimization Algorithms, pp.77-87.

[26] Pétillot, L., Pewny, F., Wolf, M., Sanchez, C., Thomas, F., Sarrazin, J., Fauland, K., Katinger, H., Javalet, C. & Bonneville, C., 2020. Calibration transfer for bioprocess Raman monitoring using Kennard Stone piecewise direct standardization and multivariate algorithms. Engineering Reports, 2(11), p.e12230.

[27] Li, Y., 2009, September. Partial least squares regression and its modeling realization. 2009 International Conference on Management and Service Science, IEEE, pp. 1-3.

[28] Garthwaite, P., 1994. An Interpretation of Partial Least Squares. Journal Of The American Statistical Association, 89(425), 122-127.

[29] MacFarland, T.W. & Yates, J.M., 2016. Mann–whitney u test. In Introduction to nonparametric statistics for the biological sciences using R (pp. 103-132). Springer, Cham.