Mathematical Biosciences and Engineering

MBE, 16(6): 6874–6891.
DOI: 10.3934/mbe.2019344
Received: 22 May 2019
Accepted: 07 July 2019
Published: 29 July 2019

http://www.aimspress.com/journal/MBE

Research article

Fast honey classification using infrared spectrum and machine learning

Hung-Yu Chien[1,*], An-Tong Shih[1], Bo-Shuen Yang[1] and Vincent K. S. Hsiao[2]

[1] Department of Information Management, National ChiNan University, Taiwan, R.O.C.
[2] Department of Applied Materials and Optoelectronic Engineering, National ChiNan University, Taiwan, R.O.C.

* **Correspondence:** Email: hychien@ncnu.edu.tw; Tel: +886492910960.

Abstract: Honey has been one previous natural food in human history. However, as the supply cannot satisfy the market demand, many incidents of adulterated and fraudulent honey have been reported. In Taiwan, some common adulterated honey and fraudulent honey incidents include (1) mixing honey with fructose, (2) importing cheap honey abroad but labeling them as domestic honey, and (3) labeling cheaper honey (for example, nectar and lychee honey) as high-price honey (for example, longan honey). It is very difficult for consumers to tell the genuineness of the labeling of honey. To protect consumers and honest honey producers, we aim at exploring and developing an efficient and convenient technology that can effectively classify honey. We analyze the infrared spectra of honey samples and apply machine learning technologies to classify honey. The experimental results confirm that this technology can effectively distinguish several main honey types in Taiwan. This technology has the advantages of non-destruction, immediacy, and low manpower. It can serve as an effective tool to fast screen honey products.

Keywords: honey; infra-red; spectrum; machine learning; adulteration; fraud

1. Introduction

Honey was called a holy product in ancient times, and it has been highly valued even then. Even though people learn better skills to keep bee and produce honey, the supply still falls far behind the market demand. The price is rising every year in Taiwan, and the supply is very unstable, due to the climate change, the environments, the virus, the pathogen, the over-usage of pesticide and herbicide,

and so on [1,2]. Similar challenges can be observed globally [3,4].

In Taiwan, several popular honey includes longan honey, lychee honey, wildflower honey, etc. People prefer longan honey to other honey, because of the taste, the flavor and the claimed nutritional values [5]. People also believe domestic honey quality is better than some other imported honey. These factors cause the prices of honey are quite different. The price of longan honey is always higher than other kinds of honey. The prices of domestic honey are higher than that of some imported honey. The big price gap between different kinds of honey allure some sellers fraudulently label low-price honey as higher-price honey [6]. For example, lychee honey and wild flower honeys are labeled as longan honey, imported honey products are labeled as domestic honey, and fruticose-mixed products are labeled as pure honey. These problems of fraudulent and adulterated honey not only affect the rights of consumers but also the interests of honest honey producers. Therefore, classifying the types and the sources of honey is an important challenge for the Taiwan honey market and industry. The global honey industry has similar issues [7]. However, because some honey types have similar colors and the skills of mixing honey have been evolving, it is very difficult for users to verify the correctness of the labelling. Therefore, we aim at exploring technologies that can efficiently and effectively classifying honey.

The rest of this paper is organized as follows. Section 2 discusses the related work and technologies. Section 3 introduces the FieldSpec4 spectroradiometer, and describes the data preprocessing technologies we apply. Our system design for the honey classification is presented in Section 4, which include the process, the design of our experiments, and the descriptions of the samples. Section 5 discusses the spectra analysis. Section 6 discusses the results of applying machine-learning classification on the honey spectra. Section 7 states our conclusions and future work.

2. Related work

There are some technologies available to classify honey, but most of these methods are destructive, and too costly in terms of man-power, money, and time. Here, destruction means that the conventional technologies would damage or compromise the honey samples. Both SNIF-NMR and EA-IRMS [8] use chemical titration and then apply the instrument to check whether honey is doped with C3/C4 isotope saccharides [9–11] and excessive animal medication. Plants like peanuts, tobacco, soybeans, rice, etc, are called “C3” due to the three-carbon compound (3-Phosphoglyceric acid, or 3-PGA) produced by the CO2 fixation mechanism in these plants. While plants like corn and sugar cane are called “C4” as they have developed the C4 carbon fixation pathway to conserve water loss, thus are more prevalent in hot, sunny, and dry climates. The isotope analysis can be used to analyze the ratios of different isotopes in a sample material [11]. The difference between the two methods is that EA-IRMS cannot detect C3 plant sugar. The strength of these two methods is that they can accurately tell whether the samples contain C3 or C4 isotope saccharides, but the weaknesses include the complicated and time-consuming process, the difficulty for farmers to acquire these technologies, and its destruction on the samples. Inductively coupled plasma mass spectrometer (ICP-MS) and atomic absorption spectroscopy (AAS) [10] are two methods to analyze the metal residues in honey. The metal trace elements in each honey samples vary, depending on their locations and species. So it can help to identify the possible locations of honey samples. Disadvantages of these methods, include (1) difficulty for farmers to acquire the technologies, (2) difficulty in identifying the C3/C4

sugars ingredients, and (3) difficulty of accessing the skillful specialists. The official method of Association of Official Agricultural Chemists (AOAC) [12,13] has been widely accepted by several international organizations to distinguish high fructose corn syrup and honey, but high fructose corn syrup cannot be correctly identified by conventional methods. AOAC can accurately determine the purity of honey in terms of C3 and C4 sugars adulteration; but it is very cumbersome, time-consuming and difficult for operators to apply than SNIF-NMR and EA-IRMS do. The operators should have well knowledge of chemical properties. ATAGO REPO series honey refractive polarimeter [14] is used to detect the sucrose ingredients, which will be completely converted to monosaccharides (glucose and fructose) gradually [15]. The advantage of ATAGO REPO is that the general public can purchase the devices and no special knowledge is required. The process just requires 3 ml of honey sample. The price is between 45 K–125 K New Taiwan dollar (NT). The disadvantage of ATAGO REPO is that it cannot identify other ingredients, if a sample is mixed with other cheaper ingredients.

Most of conventional technologies for food chemical testing or fruit sweetness identification are usually destructive: the process needs to destroy/compromise samples. For example, to test the sweetness of a group of watermelons, the samples must be randomly selected, the samples are cut open, and the juice is placed on a sweetness meter for testing. This process destroys the samples, and the tested results do not always hold for the rest of the samples. Some common weaknesses of the above technologies include (1) destruction or compromise of samples, (2) time-consuming, (3) requirement of skillful specialists, and (4) its limitations of very-narrow-specified ingredients.

On the other hand, spectra of samples can be easily collected through infra-red spectrum, multi-spectrum, and hyperspectral instruments [16,17]. The process does not destroy or compromise the samples [18], and it is fast and easy to train the operators. Each ingredient inside a sample would generate different reflection on different spectrum, and these data (or curves) can be used to identify the ingredients. When there are many different ingredients contained in a food sample (like meat, fruits, oils, proteins), the accumulated effects on the reflections are quite complex [19,20]. Luckily, as more samples are collected, one can still differentiate the differences and Machine Learning (ML) technologies can be applied to learn the differences.

Various ML technologies have been successfully applied in several fields like image identification, voice identification, natural language interpretation and translation, etc. Its application in food safety challenges has drawn some attention from the academia and industry [21,22]. A MultiLayer Perceptron (MLP) [23] is a class of feedforward artificial neural network. A MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Its multiple-layer architecture and non-linear activation make MLP good at distinguishing data that is not linearly separable. A Convolutional Neural Network (CNN) [24] is a regularized version of multilayer perceptron. CNNs take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. They are commonly used in analyzing visual imagery. Both MLP and CNN are feedforward neural networks. Unlike feedforward neural networks like MLP and CNN, a recurrent neural network (RNN) [25] is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence, and RNNs can use their internal state (memory) to process sequences of inputs. This allows RNNs to exhibit temporal dynamic behavior, and makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. Support-vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and

regression analysis. SVMs can efficiently perform both a non-probabilistic binary linear classification and a non-linear multi-class classification. The Principal Component Analysis (PCA) [26] is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. PCA is mostly used as a tool in exploratory data analysis and for making predictive models. Due to the properties of the scanned spectra of the honey samples, we would apply MLP, SVM and PCA in classify honey samples. This paper describes our experiments on analyzing honey samples' infra-red spectra and the classification of these samples using MLP, SVM and PCA respectively.

3. Preliminaries

We introduce the instrument and the pre-processing technologies applied. The instrument we use is the FieldSpec4 spectroradiometer from American ASD Company (now Malvern Panalytical company) [27]. FieldSpec4 spectrum ranges from 350 nm to 2500 nm. It uses graded index InGaAs photodiode SWIR detectors [16]. It provides 3 nm spectral resolution in the VNIR (350 nm–1000 nm) range and 10 nm in the SWIR (1001 nm–2500 nm) range. FieldSpec Dual collection software automates and synchronizes white reference and target radiance measurements, eliminating the need for manual white reference measurements. The light source we use is a 75 watt quartz-tungsten-halogen light.

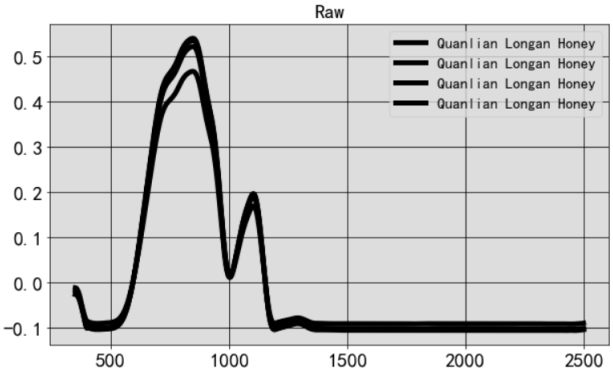

Figure 1. Four Spectrograms of the same longan honey.

The raw data from FieldSpec4 need to be further processed in order to solve the scattering phenomenon [28], which mainly occurs in the operation and are inevitable during the experimental operation. Ahmed and Akinbode et al. [18] applied the technologies of the first-order, the second-order differential, normalization, Standard Normal deViate (SNV) correction, Multiplicative Signal Correction (MSC), and median center to solve the scattering problems of their meat samples. We apply the first-order differentials, SNV [29], and MSC [29]. The scattering phenomenon was shown in Figure 1. We can see that although the same sample was scanned several times, there are some differences among the spectrograms, which is the scattering phenomenon caused by white noise. Figure 1 shows the raw spectra, Figure 2 shows that after MSC processing, and Figure 3 shows the results after SNV processing [29]. We can see that both SNV and MSC can eliminate the scattering noises. The vertical axis unit of the spectra charts below represents the ratio of the

reflected wave to the incident wave power.

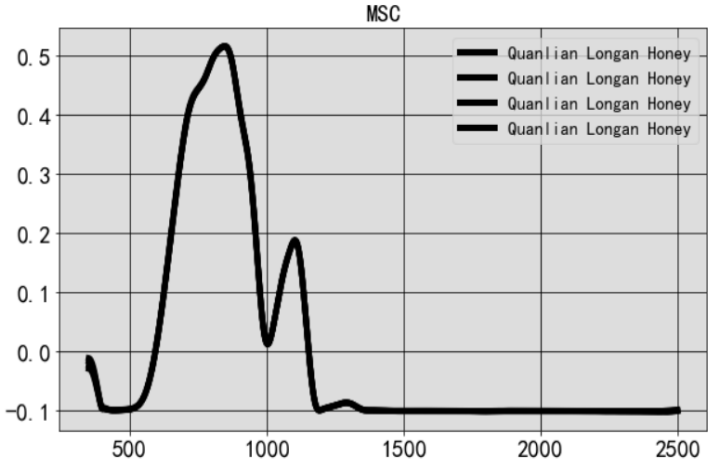

Figure 2. Four Spectrograms of the same Longan Honey after MSC processing.

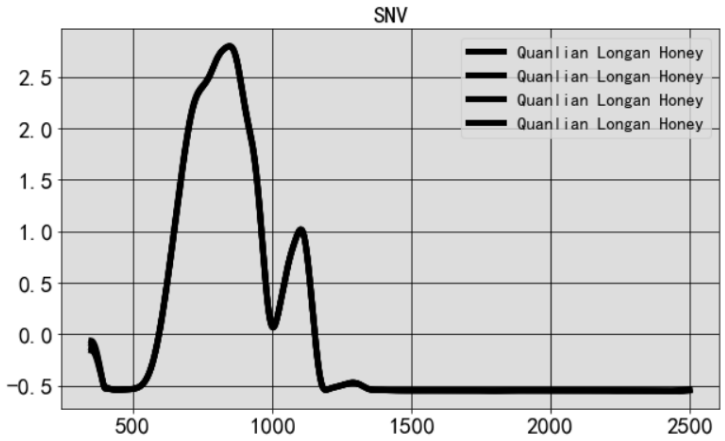

Figure 3. Four Spectrograms of the same Longan Honey after SNV processing.

4. The design of our honey classification process

The goals of this study include (1) investigating the effectiveness of various pre-processing technologies on honey samples, (2) classifying honey samples based on processed spectra, and (3) evaluating ML technologies on honey classification. To achieve the goals, we design and conduct several experiments, which can be divided into two phases. In Phase 1, we design a general process for identifying which pre-processing technology can effectively differentiate the honey while eliminating the noises. In Phase 2, we investigate how effectively Machine Learning (ML) technologies classify the honey.

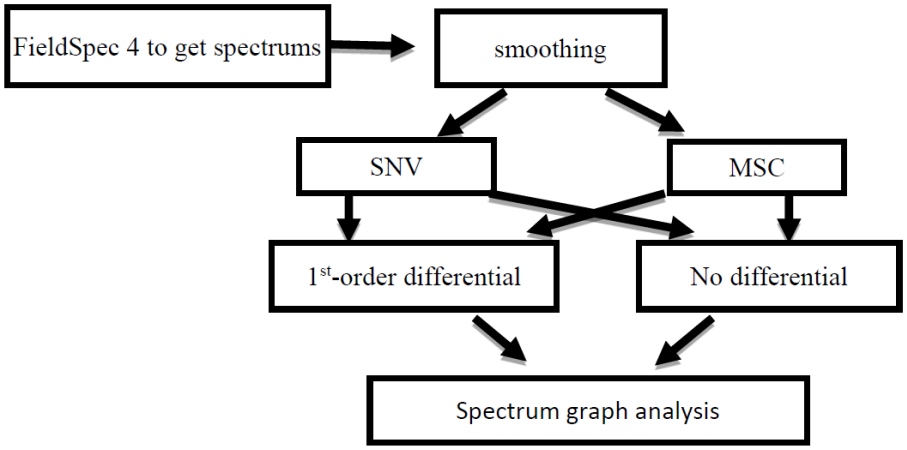

Figure 4. Phase 1 process: investigating various pre-processing technologies.

Figure 4 depicts our Phase-1-process: evaluating pre-processing technologies on honey classification. In the first step, we get the spectra of samples from FieldSpec 4, and apply the savitzky_golay smoothing algorithm [30] with the window size of 81 and a polynomial order of 3. Table 1 summarizes the acronyms used in this article.

Table 1. The acronyms used in this article.

terms	acronyms	terms	acronyms
Machine Learning	ML	FrucTose	FT
MultiLayer Perceptron	MLP	Principal Component Analysis	PCA
Convolutional Neural Network	CNN	Support-Vector Machine	SVM
Recurrent Neural Network	RNN	Mixed Longan Honey	MLGH
standard normal variate	SNV	Domestic LonGan Honey	DLGH
multiplicative scatter correction	MSC	Imported LonGan Honey	ILGH
First LonGan Honey	1LGH	Citrus Honey	CH
Second LonGan Honey	2LGH	Wild Flowers Honey	WFH
General LonGan Honey	GLGH	LyChee Honey	LCH
Pure LonGan Honey	PLGH		

Several pre-processing technologies are evaluated, which include smoothed data with the 1st order differential (referred as smooth-1st-order-differential for short), smoothed data with SNV (smooth-SNV), smoothed data with MSC (smooth-MSC), smoothed data with SNV and the 1st order differential (smooth-SNV-1st-order-differential), and smoothed data with MSC and the 1st order differential (smooth-SNV-1st-order-differential). Totally, we have 8 kinds of spectra for each sample: raw spectrogram, 1st order differential spectrogram, smoothed spectrogram, smooth-1st-order-differential spectrogram, smooth-SNV spectrogram, smooth-SNV-1st-order-differential spectrogram, the smooth-MSC spectrogram, and the smooth–MSC-1st-order-differential spectrogram.

Based on the 8 kinds of spectra, we investigate several questions: (1) does the smoothing processing would lose important information, (2) do SNV and MSC processes could solve the scattering phenomenon on honey samples, (3) how does the 1st-order differential operation impact

the information embedded in the raw spectrum, and (4) does it amplify the differences between different kinds of honey samples by applying the 1st-order differential after SNV/MSC operations? These questions will be investigated through the following honey classification experiments.

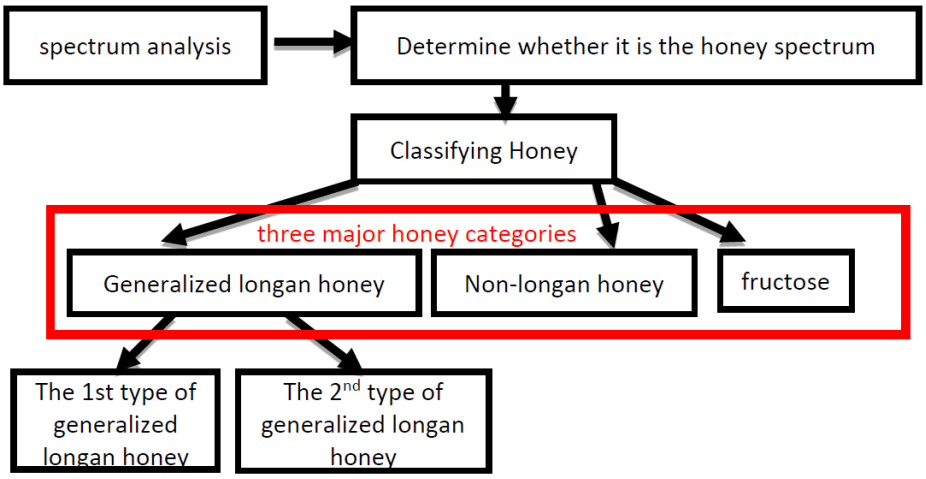

Figure 5. Honey classification experiments and the relation.

Figure 5 shows several honey classification experiments and the relation among these experiments. The 1st experiment is to classify samples into three major honey categories: (1) Generalized LonGan Honey (GLGH) which include pure longan honey and longan honey mixed with any other kinds of honey, (2) Non-LonGan Honey (NLGH), and (3) FrucTose(FT). The main goal of this experiment is to distinguish longan honey from others.

The 2nd experiment is to classify samples into four categories: (1) the 1st-type of LonGan Honey (1GLH), the 2nd type of LonGan Honey (2LGH), NLGH, and FT. The reason for further classifying longan honey is because the different locations of longan honey productions might have some differences in their spectra. The 2LGH includes the samples labeled with "winery longan honey", "neighbor's longan honey", and "Zhong Liao longan honey", and the rest of the GLGH samples are classified as the 1LGH.

Each sample has been scanned from time to time to see how it would change as we store it for a 2-month period. The spectrum of a sample is also measured several times when we scan it using the FieldSpec 4 instrument. We apply several pre-processing operations on the raw spectra, and there are totally 104 spectrum data. 75% of the data are used as training data, and 25% of them are used for testing.

5. Honey spectrum analysis

Each honey sample is denoted by their product name. Four scanned values of each sample has been recorded during a 2-month period. A sequence number (or date) is annexed to the name to differentiate each scanning. For example, Nongqi Shi citrus honey 1, Nongqi Shi citrus 2, and so on. The samples marked with "Li Lunde" is to highlight the samples harvested from Lunde Li's (or his relatives') bee farms, which have won several honey competitions in Taiwan.

Figure 6 and Figure 7 respectively depict the spectra of pure longan honey after SNV or MSC

pre-processing. We can easily tell that these pure longan honey samples share very similar spectrum charts, but there are still some differences in some channels (the reflections on different wavelength). These pure longan honeys samples can be roughly divided into two groups: the first group includes Li Lunde, Thailand Longan, Chiang Mai honey, pure honey; the second group includes winery longan honey and Zhong Liao longan honey.

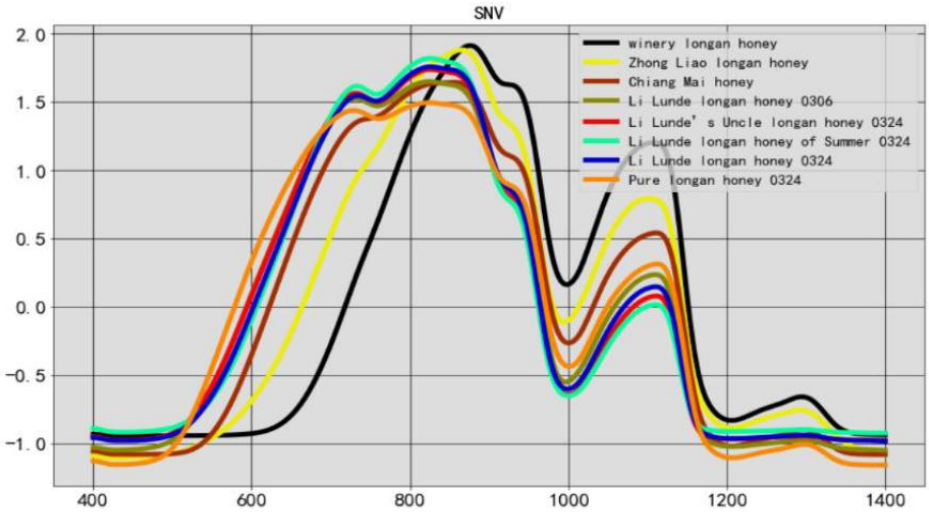

Figure 6. Spectra of pure longan honey after SNV processing.

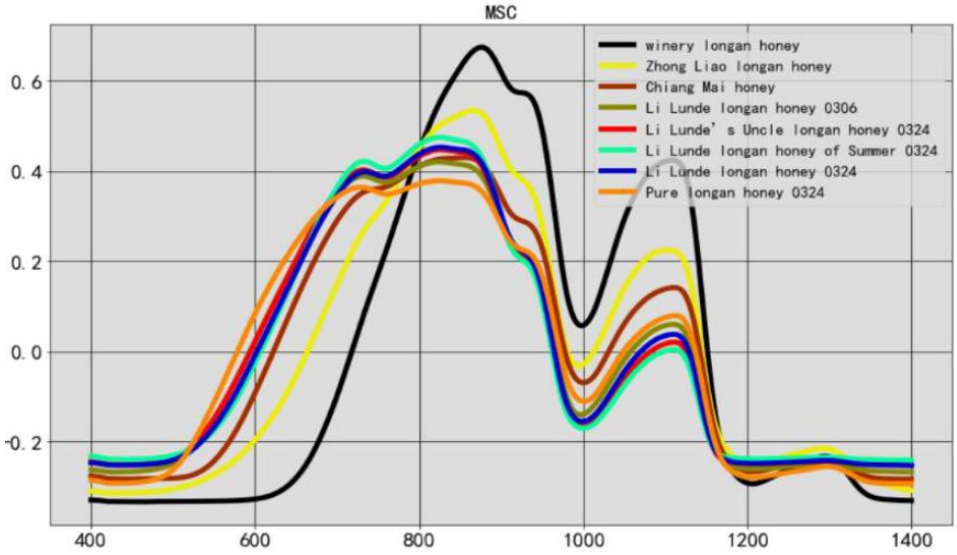

Figure 7. Spectra of pure longan honey after MSC processing.

Figure 8 and Figure 9 respectively depict the charts of the pure longan honey after SNV-1st-differential operation and after MSC-1st- differential operation. We can see that the 1st-order-differential operation does amplify the differences of the samples at some channels. For example, the Pure labeled "honey0324" stands out from the group of Lunde LI's honeys.

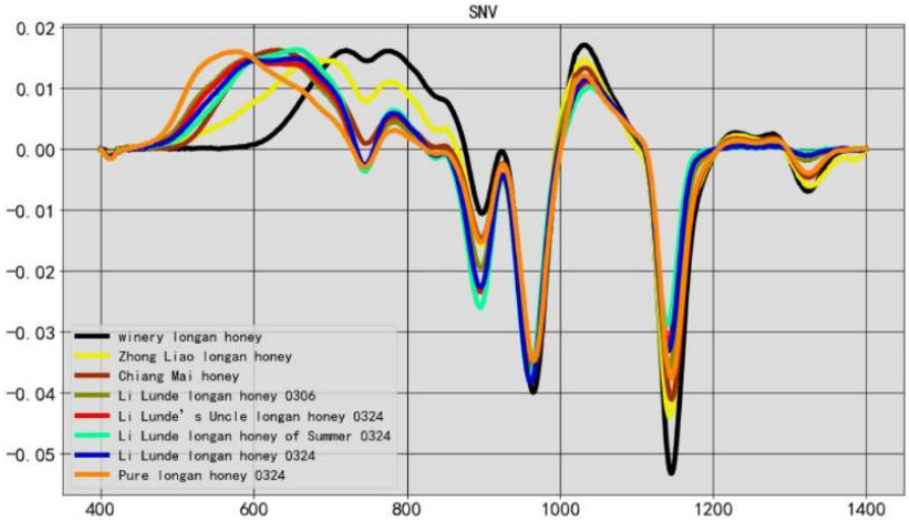

Figure 8. Charts of pure longan honey after SNV-1st-order-derivative-differential operation.

Figure 10 and Figure 11 respectively depict the charts of all samples' spectra after SNV-1st-order- differential and after MSC-1st-order- differential. Based on these charts, the samples could be roughly divided into four group. The first group is the 1LGH group (some longan honey samples from specified locations); the 2nd group is NLGH group which contains samples from wildflower honey, and citrus honey, and lychee honey; the third group is the FT group (fructose); the fourth group is the 2LGH group which includes winery longan honey, neighbor's longan honey, and Zhong Liao longan honey.

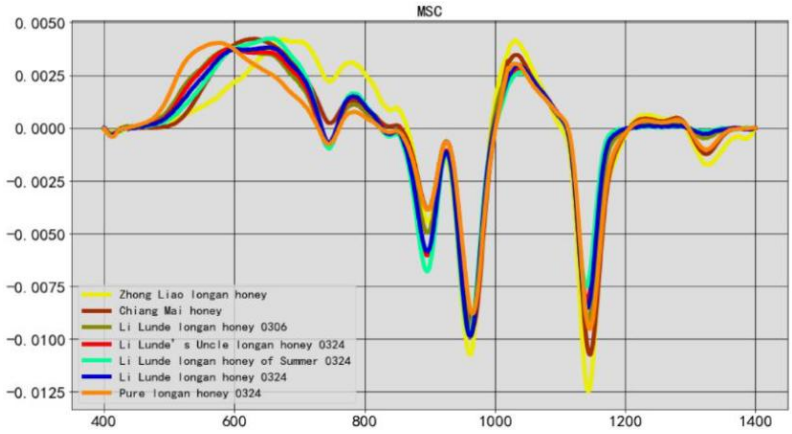

Figure 9. Charts of pure longan honey after MSC-1st-order- differential operation.

In the 1LGH group, the spectrograms of Li Lunde's pure longan honey and Li Lunde longan honey mixed with other honey (flowers, lychee) sample are very similar. According to producer's explanations, the blossom period of longgan flowers last several weeks, and bees harvest both longan flowers and other flowers both in the beginning and in the end of that period. Therefore, the producer Li Lunde would not label these honey as "pure" longan honeys, but other producers would still label

them as “pure” longan honey. In our classification, we all label them as Generalized LongGan Honey (GLGH) in this stage. From the charts of Figure 10 and 11, we can notice that there are apparently two different types of longan honey, the red-color type and the black-color type. We speculate that there are several possible causes that result in the phenomenon. One possible cause is the overlap of the logan blossom period and other flower blossom period. Bees forage honey from longan flowers as well as other flowers both at the beginning period and at the will-end period of longan blossom period. Most bee keepers still label their honey products as “pure” logan honey even though their harvested logan honey contain other honey types like lychee honey and wild-flower honey. The second possible cause might be the small cultivated lands in Taiwan. The size of each cultivated land in Taiwan is very small: the size of a farm is usually only one to several “fen”, where one “fen” equals 969 square meter. But, the foraging radius of a bee can range from 2~7 kilo meters. The third possible cause might be the variance of lonagn nectar in different locations. To precisely identifying and quantifying the ingredients of honey, we think some conventional time-consuming technologies like melissopalinological analysis applied on pollens are necessary [31].

There is an interesting phenomenon. After about 800nm, most of the spectra have overlapped together, which is supposed to be the spectrum channel of fructose [32].

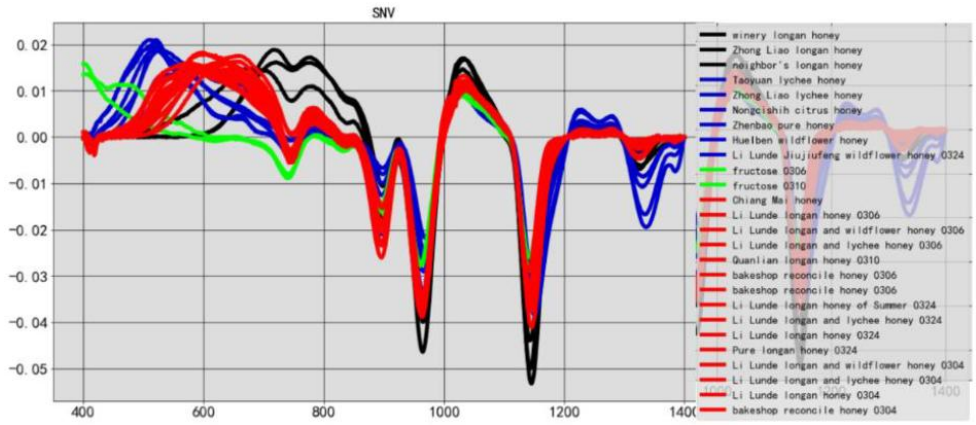

Figure 10. Charts of all-samples with SNV-1st-order- differential.

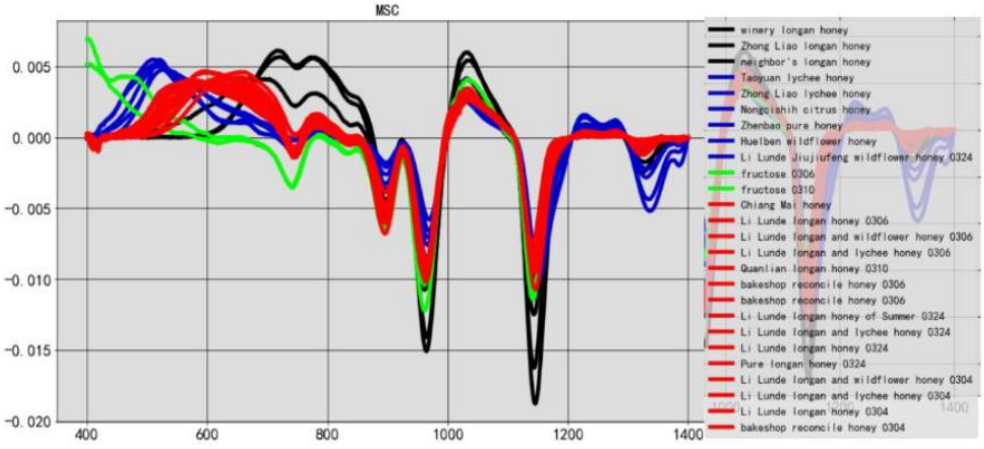

Figure 11. Charts of all-samples with MSC-1st-order- differential.

6. Honey classification using MLP, SVM, and PCA

We respectively discuss our honey classification experiments using the MultiLayer Perceptron (MLP) in TensorFlow package [33], SVM [34], and PCA [35].

6.1. Honey classification using MLP

In the MLP experiments, we use three hidden layers, and we label the samples, according to the experiment goals. The parameters used the MLP experiment are summarized in Table 2, where the notation "1LGH*60" means that there are 60 samples belonging to the 1LGH group, ReLu is one of the activation function [31], and softmax is used in the output layer.

Table 2. The parameters and the numbers of samples in our MLP experiments

	4-category lab.	3-category lab.
number of epochs, ratio of validation data	5, 25%	5, 25%
Sample Type * quantity	1LGH*60, 2LGH*12, FT*8, Other Honey*24	GLGH*72, FT*8, Other Honey*24
NoN of input layer	2151	2151
NoN of 1st hidden layer (ReLu)	228	208
Non of 2nd hidden layer (ReLu)	152	146
Non of 3rd hidden layer (ReLu)	76	63
NoN of output layer	4	3

NoN: Number of Nodes. The algorithm ReLu [23] is used in the hidden layers. Softwax [23] is used in the output layer.

Figure 12 shows the flow chart for applying supervised learning in our honey spectrum classification. The detailed steps are described as follows.

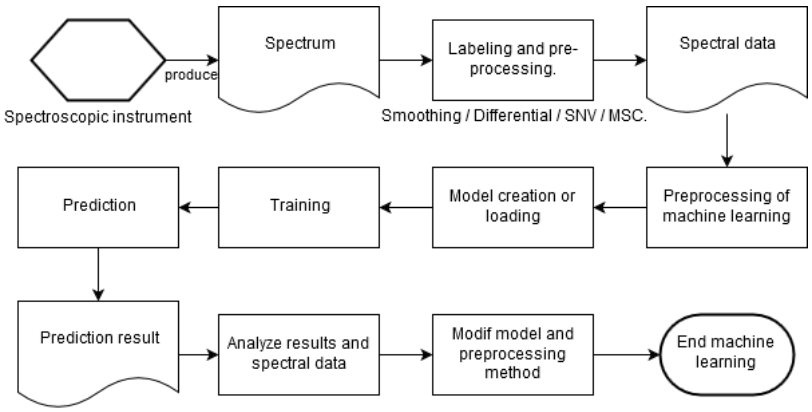

Figure 12. The flow chart of applying ML in honey spectrum classification.

Step 1: The experimenter operates the spectroradiometer to scan the samples and obtain the spectra.

Step 2: Create Label data from the samples. Preprocess the spectra. For example: smoothing, differentiation, and performing SNV/MSC to eliminate the scattering noises.

Step 3: Preprocessing for machine learning.

Step 4: Load or create a model.

Step 5: Model training.

Step 6: The model is used to predict the labels of the spectra.

Step 7: Analyze the results of the predictions and spectra, and investigate the reasons for the success/error of the predictions.

Based on the results from the MLP-based classification experiments, we have the following observations.

The 1st observation: This experiment is to observe the accuracy and prediction results of both the 4-category and the 3-category classifications, using only smoothed data. Table 3 shows the results. The 4-category experiment has the highest model accuracy.

Table 3. The results of MLP experiments and the 1st observation.

Pre-processing	Classification	Model accuracy	# of correct predictions	# of wrong predictions
Smooth	4-category classification	0.980	25	1
Smooth	3-category classification	0.923	23	3

2nd observation of MLP experiments: This experiment is to investigate the 4-category MLP classifications using different preprocessed data. Table 4 shows the results. The SNV-based/MSC-based MLP classifications have the highest model accuracy.

Table 4. MLP classifications using different preprocessed data.

Processing	Classification	Model accuracy	Correct prediction	Wrong prediction
Raw data	4-category classification	0.961	24	2
Data smoothing	4-category classification	0.932	22	4
First-order differential	4-category classification	0.971	24	2
Smooth and first-order differential	4-category classification	0.961	24	2
SNV	4-category classification	0.980	25	1
SNV and first-order differential	4-category classification	0.980	25	1
MSC	4-category classification	0.980	25	1
MSC and first-order differential	4-category classification	0.980	25	1

3rd observation of MLP experiments: This experiment is to verify whether the SNV-1st-order-differential-preprocessed-data would have better results for all the MLP-based classification goals. The model was retrained five times. Table 5 shows the results. It does improve all the accuracy for several classification goals.

Table 5. MLP classifications using SNV-1st-order-derective-differential-preprocessed data.

Preprocessing	Classification	Highest accuracy	Minimum accuracy	Average accuracy
SNV and first-order differential	4-category classifications	0.980	0.95	0.969
SNV and first-order differential	3-category classifications	0.974	0.94	0.954

In a short summary, we have the following results. The results from both the preprocessed spectrum analysis and the MLP classifications show that the SNV-1st-order-differnetial preprocessing has the best accuracy. The 2nd winner is the MSC-1st-order-differential preprocessing, with a 0.03 less accuracy, compared to the SNV-1st-order-differnetial preprocessing. Among several classification goals, the 4-category classification with accuracy 0.9897 has the best result, and the 3-category classification with accuracy 0.9743 comes the second; it shows that the results from the 4-categry classification and the 3-category classification, in spectrum analysis and in MLP experiments, are satisfactory.

6.2. Honey classification using SVM

Here, we are interested in two questions for applying SVM classification: (1) which preprocessing procedure can provide better classification results? (2) how well SVM does on classifying the honey samples? Table 6 summarizes the results of the 4-category classification experiment for the 1st question. We apply SVM on 8 versions of the pre-processed spectra (raw/smoothed/1st-order differential/ smoothed+1st-order differential/ SNV/SNV + 1st-order differential/MSC/MSC + 1st-order differential). Here, we concentrate on the spectra ranging from 400–800 nm, based on the observations from the previous spectra analysis. The test data are to be classified as one of the four groups (C1–C4), where C1 denotes 1LGH, C2 denotes 2LGH, C3 denotes the FT group, and C4 denotes the other honey group.

Table 6. SVM-based honey classification using different preprocessing procedures

	Raw				smoothed				1st-order diff.				Smoothed + 1st order diff.				SNV				SNV+ 1st order diff.				MSC				MSC + 1st order diff.			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
pre.	.94	1	1	1	.94	1	1	1	.58	0	0	0	.58	0	0	0	.94	1	1	1	.58	0	0	0	1	1	1	1	.58	0	0	0
re	1	1	1	.83	1	1	1	.83	1	0	0	0	1	0	0	0	1	1	1	.83	1	0	0	0	1	1	1	1	1	0	0	0
f1	.97	1	1	.91	.97	1	1	.91	.73	0	0	0	.73	0	0	0	.97	1	1	.91	.73	0	0	0	1	1	1	1	.73	0	0	0
sup	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6

pre.: precision; re: re-call; f1: f1-score; sup: support. C1 is the 1LGH group; C2 is the 2LGH group; C3 s the FT group; C4 is the other honey group.

In Table 6, four metrics are specified. "Precision" is defined as the number of true positives (T_p) over the number of true positives plus the number of false positives (F_p): $P \coloneqq T_p/(T_p + F_p)$ [36]. "Recall" (R) is defined as the number of true positives (T_p) over the number of true positives plus the number of false negatives (F_n): $R \coloneqq T_p/(T_p + F_n)$. The ($f1$) score is defined as the harmonic mean of precision and recall: $f1$-score$\coloneqq 2 * (P * R)/(P + R)$. The "support" is the number of tested samples in that group.

High precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). From the table, we have two observations: (1) SVM applied on MSC-preprocessed spectra has the highest precision and the highest recall; (2) the 1st-order preprocessing on honey spectra will degrade the performance of SVM classification. Therefore, we apply SVM on the SMC-preprocessed spectra for the 3-category

classification. The results are shown in Table 7. From the table, we can see that applying SVM on the MSC-preprocessed spectra can 100% correctly classify the samples for the 3-category honey classification.

Table 7. SVM-based 3-category honey classification using MSC-preprocessed spectra.

	MSC		
group	C1	C2	C3
precision	1	1	1
recall	1	1	1
f1-score	1	1	1
support	18	2	6

C1 is the GLGH; C2 is the FT group; C3 is the other honey group.

6.3. Honey classification using PCA

Here, we are still interested in the two fundamental questions. (1) which preprocessing procedure can provide better classification results? (2) how well PCA does on classifying the honey samples? Table 8 summarizes the results of the 4-category classification experiment for the 1st question. We apply PCA on the 8 versions of the spectra. Here, we concentrate on the spectra ranging from 400–800 nm, based on the observations from the previous spectra analysis.

Table 8. PCA-based honey classification using different preprocessing procedures.

	Raw				smoothed				1st-order diff.				Smoothed + 1st order diff.				SNV				SNV+ 1st order diff.				MSC				MSC + 1st order diff.			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
pre	.94	1	1	1	.94	1	1	1	1	1	1	1	1	1	1	1	.94	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
re	1	1	1	.83	1	1	1	.83	1	1	1	1	1	1	1	1	1	1	1	.83	1	1	1	1	1	1	1	1	1	1	1	1
f1	.97	1	1	.91	.97	1	1	.91	1	1	1	1	1	1	1	1	.97	1	1	.91	1	1	1	1	1	1	1	1	1	1	1	1
sup	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6	15	3	2	6

pre.: precision; re: re-call; f1: f1-score; sup: support. C1 is the 1LGH group; C2 is the 2LGH group; C3 s the FT group; C4 is the other honey group.

Table 9. PCA-based 3-category honey classification using 1st-order-preprocessed spectra.

	1st-order diff.		
group	C1	C2	C3
precision	1	1	1
recall	1	1	1
f1-score	1	1	1
support	18	2	6

C1 is the GLGH; C2 is the FT group; C3 is the other honey group.

From the table, we have one observation: both MSC and 1st-order differential operation can provide best PCA classification results. Next, we apply PCA on the 1st-order preprocessed spectra for

the 3-category classification. Table 9 summarizes the results.

6.4. Summary of the classification experiments

Now we summarize the observations.

(1) Among the three classification technologies we applied, both SVM and PCA out-perform MLP in terms of accuracy.
(2) When we apply MLP classification, both SNV pre-processing and MSC pre-processing can improve the accuracy.
(3) When we apply SVM classification, MSC-preprocessed data can provide best accuracy, but the 1st-order differentiation would downgrade the accuracy.
(4) When we apply PCA classification, both MSC and the 1st-order differentiation can provide best classification accuracy.
(5) Applying suitable ML technologies or statistical tools like PCA on preprocessed honey spectra can effectively classify honey samples when enough authentic samples are collected. This approach is much fast, and requires low man-power than other conventional approaches.

Based on the experiments, there are some interesting open questions.

(1) Why and how does the 1st-order differentiation pre-processing significantly downgrade the classification accuracy?
(2) To identifying and quantifying the ingredients inside honey samples, some conventional time-consuming technologies like melissopalinological analysis applied on pollens are necessary, at least during the ML model training phase.

7. Conclusions

In this paper, we have investigated the effectiveness of several spectrum preprocessing technologies for classifying honey samples, and have run MLP, SVM, PCA classification experiments using the preprocessed honey spectra. Both the spectra analysis and the classification experiments provide several promising observations. (1) All three classification technologies (MLP, SVM, and PCA) can effectively perform both the 4-category honey classification and the 3-category classification. But, SVM and PCA outperform MLP in terms of prediction accuracy, when applying on our honey samples. (2) The MSC preprocessing can improve the performance of all the three classification process. Considering the merits of (1) non-destruction of samples, (2) fastness, (3) easy operation, (4) low man power, (4) no requirement of skillful operators, and (5) ML-based approach can learn and accumulate new knowledge of honey classification, this spectra-ML-based approach shows it as a very promising tool for fast and cheap honey sample screening and classification. However, to identify and quantify the ingredients of honey samples, we think collecting very large quantity of authentic samples and accompanying with some conventional time-consuming technologies like melissopalinological analysis is necessary.

During the experiments, we also found that, some classification experiments on the lychee-verse- wildflower honey classification and the domestic-verse-imported honey classification do not provide consistent results among the three classification technologies, because the number of trusted samples are not enough. It is very difficult at this stage to gather large quantity of trusted samples, as the reports say many of the products in Taiwan be fraudulent. To further quantifying the

ingredients and extend the experiments to other honey classification challenges, we plan to co-operate with the government to acquire more trusted samples, and investigate other ML technologies in the future. Another interesting open question is exploring the rationale of why the 1st-order differentiation would downgrade the accuracy of SVM classification on honey samples.

Acknowledgments

This project is partially supported by the National Science Council, Taiwan, R.O.C., under grant no. MOST 107-2218-E-260-001.

Conflict of interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. Agence France-Presse, South China Morning Post, Taiwan's beekeepers battle to cash in on rapidly growing market for pure honey, South China Morning Post, Oct 23, 2016. Available from: https://www.scmp.com/news/china/society/article/2039345/taiwanese-beekeepers-battle-cash-pure-honey-buzz.
2. TRIDGE, Taiwan Honey suppliers, wholesale prices, and market information. Available from: https://www.tridge.com/intelligences/honey/TW.
3. United States Department of Agriculture, NATIONAL HONEY REPORT, United States Department of Agriculture, April 24, 2019. Available from: https://www.ams.usda.gov/mnreports/fvmhoney.pdf.
4. F. E. Murphy, IoT Applications for Honey Bee Colony Condition: What's the Buzz All About? IEEE IoT Newsletter–(May 2019).
5. HealthyWithHoney.com, Longan Honey benefits? It's a powerful antimicrobial honey, March 15th, 2018. Available from: https://healthywithhoney.com/longan-honey-benefits-its-a-powerful-antimicrobial-honey/.
6. Taipei Times, Researchers say 75% of Taiwan's honey not pure, Sep. 14, 2013. Available from: http://www.taipeitimes.com/News/lang/archives/2013/09/14/2003572064.
7. The Economist, The scourge of honey fraud, Aug. 30, 2018, Available from: https://www.economist.com/united-states/2018/08/30/the-scourge-of-honey-fraud.
8. Eurofins Scientific, Use SNIF-NMR and EA-IRMS to check whether honey is doped with C3/C4, Available from: https://www.eurofins.tw/%E9%A3%9F%E5%93%81%E6%AA%A2%E9%A9%97/%E8%9C%82%E8%9C%9C%E6%AA%A2%E9%A9%97/..
9. J. B. Paine III, Y. B. Pithawalla and J. D. Naworal, Carbohydrate pyrolysis mechanisms from isotopic labeling: Part 3. The Pyrolysis of d-glucose: Formation of C3 and C4 carbonyl compounds and a cyclopentenedione isomer by electrocyclic fragmentation mechanisms, *J. Anal. Appl. Pyrol.*, **82** (2008), 42–69.
10. Council of Agriculture, Executive Yuan R.O.C., Origin Authentication of Taiwan Longan Honey, April 2009. Available from: https://www.coa.gov.tw/ws.php?id=19303&print=Y.

11. Wikipedia, Fractionation of carbon isotopes in oxygenic photosynthesis, Available from: https://en.wikipedia.org/wiki/Fractionation_of_carbon_isotopes_in_oxygenic_photosynthesis.
12. KKnews, The developments for the inspection methods of honey (use Aoac to distinguish high fructose corn syrup and honey), July 7, 2017. Available from: https://kknews.cc/zh-tw/news/zx6lkkp.html.
13. AOAC (Association of Official Analytical Communities) Taiwan Section, Introduction to the standards of an international organization. Available from: http://www.aoac.org.tw/.
14. KKnews, New instruments for identifying honey, May 31, 2016. Available from: https://kknews.cc/zh-tw/news/3v9y3.html.
15. S. Riddle, The chemistry of honey. "honey convert to monosaccharides, Bee Culture, July 25, 2016. Available from: https://www.beeculture.com/the-chemistry-of-honey/.
16. S. Chen, W. H. Chang and K. W. Hsieh, The study on prediction models for determination of sugar content in fruit juice, *J. Agr. Mach.*, **7** (1998), 41–60.
17. C. Pasquini, Near infrared spectroscopy: A mature analytical technique with new perspectives – A review, *Anal. Chim. Acta*, **1026** (2018), 8–36.
18. M. Ferreiro-González, E. Espada-Bellido, L. Guillén-Cueto, et al., Rapid quantification of honey adulteration by visible-near infrared spectroscopy combined with chemometrics, *Talanta*, **188** (2018), 288–292.
19. L. Pan, Q. Zhu, R. Lu, et al., Determination of sucrose content in sugar beet by portable visible and near-infrared spectroscopy, *Food Chem.*, **167** (2015), 264–271.
20. L. Pan, R. Lu, Q. Zhu, et al., Measurement of moisture, soluble solids, sucrose content and mechanical properties in sugar beet using portable visible and near-infrared spectroscopy, *Postharvest Biol. Technol.*, **102** (2015), 42–50.
21. X. Fu, X. Wang, and X. Rao, An LED-based spectrally tuneable light source for visible and near-infrared spectroscopy analysis: A case study for sugar content estimation of citrus, *Biosyst. Eng.*, **163** (2017), 87–93.
22. Y. Yang, H. Zhuang, S. C. Yoon, et al., Rapid classification of intact chicken breast fillets by predicting principal component score of quality traits with visible/near-Infrared spectroscopy, *Food Chem.*, **244** (2018), 184–189.
23. Wikipedia, Multilayer perceptron, Available from: https://en.wikipedia.org/wiki/Multilayer_perceptron.
24. Wikipedia, Convolutional neural network, Available from: https://en.wikipedia.org/ wiki/Convolutional_neural_network.
25. Wikipedia, Recurrent neural network, Available from: https://en.wikipedia.org/ wiki/Recurrent_neural_network.
26. Wikipedia, Principal component analysis. Available from: https://en.wikipedia.org/ wiki/Principal_component_analysis.
27. Malvern Panalytical Products, ASD range, Available from: https://www.malvernpanalytical.com/ en/products/product-range/asd-range.
28. A. M. C. Davies and T. Fearn, Something has happened to my data: potential problems with standard normal variate and multiplicative scatter correction spectral pre-treatments, *Spectroscopy Europe*, **21** (2009), 15–19.

29. D. Pelliccia, Instruments & Data Tools, Two scatter correction techniques for NIR spectroscopy in Python, July 21, 2018. Available from: https://www.idtools.com.au/two-scatter-correction-techniques-nir-spectroscopy-python/.
30. RIP Tutorial, Using a Savitzky–Golay filter to smooth spectrum, Available from: https://riptutorial.com/scipy/example/15878/using-a-savitzky-golay-filter.
31. L. Svecnjak, N. Biliskov, D. Bubalo, et al., Application of Infrared Spectroscopy in Honey Analysis, *Agr. Conspectus Sci.*, **76** (2011), 191–195.
32. Glorybee, HONEY FACTS & NUTRITION, Available from: https://glorybee.com/content/honey-facts-nutrition.
33. Tensorflow, Available from: https://www.tensorflow.org/.
34. U. Malik, Implementing SVM and Kernel SVM with Python's Scikit-Learn, April 17, 2018. Available from: https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/.
35. G. Seif, Principal Component Analysis: Your Tutorial and Code, Nov 9, 2018. Available from: https://towardsdatascience.com/principal-component-analysis-your-tutorial-and-code-9719d3d3f376.
36. scikit-learn, Precision-Recall, Available from: https://scikit-learn.org/stable/auto_ examples/model_ selection /plot_precision_recall.html.

AIMS Press