# Honey Dataset Standard Using Hyperspectral Imaging for Machine Learning Problems

Ary Noviyanto* and Waleed H. Abdulla†
Department of Electrical and Computer Engineering,
The University of Auckland
Auckland, New Zealand 1142
Email: *anov403@aucklanduni.ac.nz, †w.abdulla@auckland.ac.nz

***Abstract*—Hyperspectral imaging has been rarely investigated for honey analyses, on the contrary to the optical spectroscopy which is widely investigated. The essential missing component to kick start this research is a standard honey hyperspectral images, called hypercubes, dataset. This paper proposes a systematic procedure for the preparation of the first honey hypercube dataset using hyperspectral imaging. Moreover, a scalable and flexible dataset module is introduced to ease the interaction between raw hypercube data and machine learning software. The developed dataset greatly benefits researchers to progress on the research of honey analysis including constituents prediction and types classification using hyperspectral imaging and machine learning.**

## I. Introduction

Honey is a big commodity in the market. It has a total world production of 1.5 millions tonnes in 2004 and over 2 millions tonnes in 2013 [1]. The prediction of honey quality becomes very important to governments, industries and customers since some honey brands are far superior than the others; like mānuka honey from New Zealand which contains high antibacterial activity produced from New Zealand teatree (*Leptospermum scoparium*) [2]. This New Zealand honey has the highest price compared to the other honey brands [1].

Several approaches have been implemented for honey analysis, mostly chemical and physical approaches. Optical spectrum based honey analysis is an alternative way that provides non-contact, non-invasive, fast, and fully automatic methods. This research usually employs spectroscopy technology for predicting honey types or constituent concentration [3]. The other technology that has not been deeply explored is hyperspectral imaging. Hyperspectral imaging combined with machine learning methods is very promising for honey analysis, however, there is almost no peer-reviewed research article discussing this approach. Only one research article reported using hyperspectral imaging for adulteration detection by fructose-glucose mixture solutions [4].

Hyperspectral imaging is a study of objects in spectrum fields creating fingerprints of particular objects [5], [6]. Unlike the spectroscopy technology which captures spectral information from one location, hyperpsectral imaging technology can capture spectral information from a particular spatial region. In consequence, a hyperspectral image or *hypercube* is a cube with two spatial and spectral dimension.

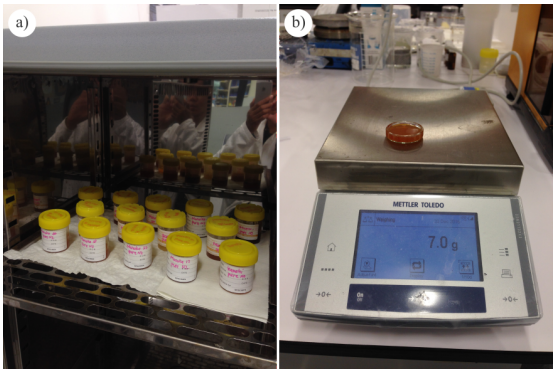



Fig. 1. (a) Honey samples are placed in an oven for the liquefying process. (b) Liquefied honey is placed in a glass container and its weight is measured.

This research proposes a standard dataset development for honey hypercubes which is essential to support research utilizing hyperspectral imaging and machine learning for honey analysis. The analyses include constituents prediction (e.g. *Maltose, Sucrose, Fructose, Glucose*, antioxidant compounds, electrical conductivity prediction, etc.) and types classification (e.g. authenticity or adulteration detection, brand identification, geographical and botanical origin classification). The standard development consists of three parts: sample preparation, hypercube acquisition and data handling. Based on our exploration, this kind of standards is still not available anywhere else. The developed database and the data handler module will be available on the University of Auckland Server.

## II. Sample Preparation

Honey samples need to be prepared carefully under identical conditions and treated in a similar way to ensure consistency of their spectra in the data acquisition phase. The defined procedure is considered to be simpler compared to the conventional honey analysis using chemical and physical methods.

The honey samples are placed in closed containers and heated in an oven overnight as depicted in Fig. 1.a. The temperature of the oven is set to be $40°C$ maximum to dissolve crystals without changing its characteristic because of overheating [7]. This treatment also helps to ease pouring the honey into sample containers and getting a flat surface.

ISBN 978-0-9928626-7-1 © EURASIP 2017

Ideally, all samples need to be in a same thickness so that the light travels through all samples equally. The most practical way to get a reasonable same thickness is by measuring the weight of the honey using a scale as depicted in Fig. 1.b. The honey weight and the container shape determine the thickness. For example, seven grams of honey in a cylindrical container with 3.5 cm in diameter will have around 0.5 cm thickness. Although, the density of each honey sample could be different because of water content mainly [8], in a relatively small amount of honey, the same weight produces reasonably the same thickness. It can be seen from TABLE I where the standard deviation of the thickness is very small: 0.0062 cm. The other important thing is to pour the honey carefully into the container to minimize bubbles since they will cause inconsistent information retrieval.

TABLE I
THICKNESS COMPARISON OF HONEY WITH DIFFERENT WATER CONTENT ON 3.5-CM-DIAMETER CONTAINERS.

| Water Content* (%) | Density* (g/cc) | Volume per 7 g (cc) | Thickness (cm) |
|---|---|---|---|
| 13 | 1.4457 | 4.8419 | 0.5033 |
| 14 | 1.4404 | 4.8598 | 0.5051 |
| 15 | 1.435 | 4.8780 | 0.5070 |
| 16 | 1.4295 | 4.8968 | 0.5090 |
| 17 | 1.4237 | 4.9168 | 0.5110 |
| 18 | 1.4171 | 4.9397 | 0.5134 |
| 19 | 1.4101 | 4.9642 | 0.5160 |
| 20 | 1.4027 | 4.9904 | 0.5187 |
| 21 | 1.395 | 5.0179 | 0.5216 |
| | **mean** | **4.9228** | **0.5117** |
| | **std** | **0.0600** | **0.0062** |

*) The densities of honey on different water contents are according to [9].

## III. HYPERCUBE ACQUISITION

A systematic acquisition stage is needed to get accurate and consistent data. The acquisition stage comprises three main parts: a samples platform, a lighting system, and an imager. In this research, the lighting system is configured to suit the most common sensing modes: reflectance and transmittance, where both are applicable since honey is neither 100% transparent nor 100% opaque substance. In the reflectance sensing mode, where the light source and the imager located on the same side, the light reflects from the sample to the imager, while in the transmittance sensing mode, where the light source and the imager located on opposite sides of the sample, the light travels through the sample to the imager [10]. The illustrations of the reflectance and transmittance sensing modes accompanied with their corresponding hypercubes in RGB version are shown in Fig. 2.

The choice and configuration of the lighting system determines the quality of the hypercubes. As far as the lighting source is concerned, halogen lamps are the most practical choice for visible to near infrared spectral bands. Also, the halogen bulbs are arranged to produce maximum homogeneous illumination since the hyperspectral imager captures spatial information.

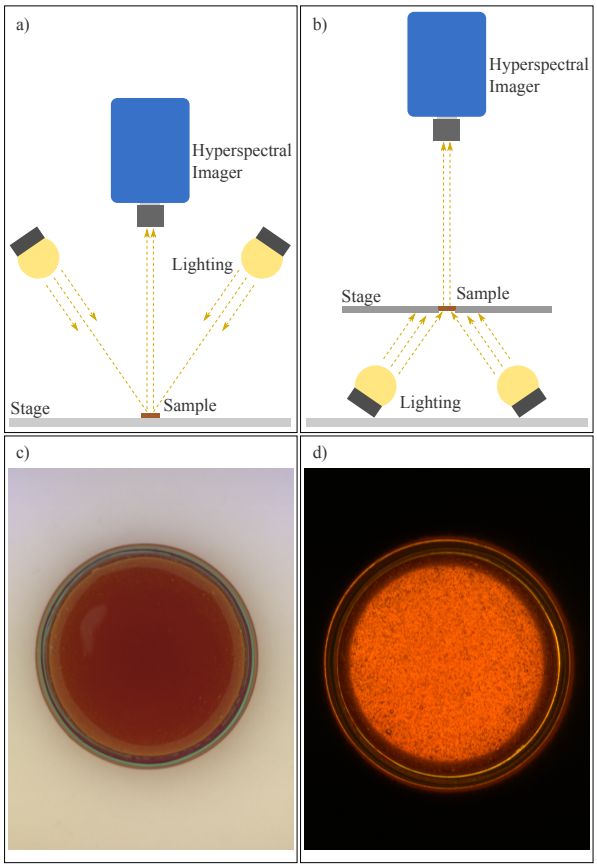



Fig. 2. (a) The reflectance sensing mode. (b) The transmittance sensing mode. (c) and (d) The corresponding RGB versions of hypercubes respectively.

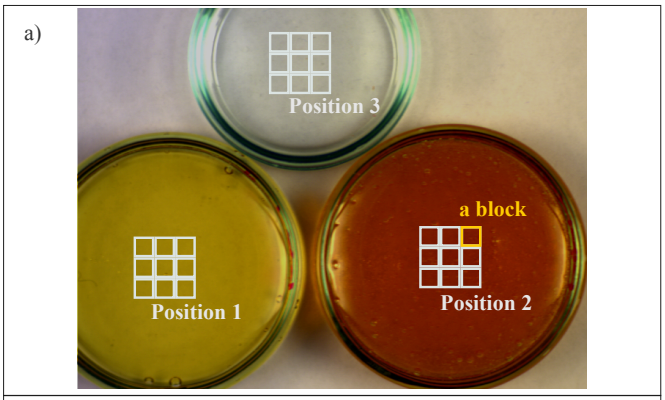
## IV. HYPERCUBE METADATA

Metadata is needed to give information about honey characteristics for each hypercube. The metadata will be provided in a separate file. To map the information to corresponding hypercubes, a primary key is required in the metadata. The best practice primary key is a combination of the filename and its corresponding position since one hypercube file can possibly contain more than one sample as shown in Fig. 3.a. The corresponding metadata is shown in Fig. 3.b where each row represents a particular honey characteristic for a specific position in the hypercube file.

## V. DATA HANDLING

A data handler is needed to process and manipulate the acquired hypercubes for further processing. Two main tasks need to be considered. The first task concerns the hypercube database in the server-side, where the hypercubes are collected and grouped together in a server; it enables remote access by many researchers. The second task concerns constructing module in the client sides to process the hypercubes conforming individual needs.

ISBN 978-0-9928626-7-1 © EURASIP 2017




b)

| Filename | Position | Type | No | Batch | Temp | Sensing | ... |
|---|---|---|---|---|---|---|---|
| 2015-12-20-hypercube-1.cube | 1 | Rata | 1 | 1 | 20.6 | Reflectance | ... |
| 2015-12-20-hypercube-1.cube | 2 | Manuka | 1 | 1 | 20.6 | Reflectance | ... |
| 2015-12-20-hypercube-1.cube | 3 | Empty | 1 | 1 | 20.6 | Reflectance | ... |
| ... | ... | ... | ... | ... | ... | ... | ... |

Fig. 3. (a) A hypercube contains more than one sample. Each sample is segmented as a region which has specific information in the metadata file. (b) the corresponding metadata.

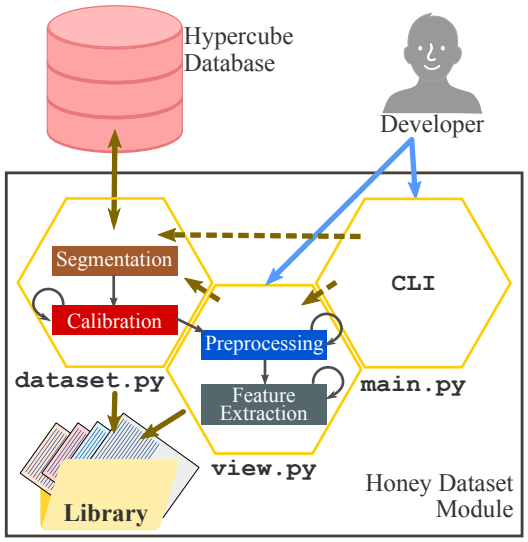



Fig. 4. The honey hypercube dataset module.

The hypercube database could be structured in a simple way. A collection of hypercubes is placed on a server grouped by folders where a metadata file is provided in each folder.

A new *honey hypercube dataset module* is developed to process the hypercubes from the database into ready-to-use data for other machine learning software. The module is constructed based on *skdata* (scikit-data) [11] written in the Python programming language. A brief overview of the module is shown in Fig. 4 where there are three core python files (`dataset.py`, `view.py` and `main.py`) to handle hypercubes.

The `dataset.py` contains a class which make connection to the hypercube database for downloading and basic processing. The basic processing includes segmentation and calibration. The segmentation methods are used to design how regions of interest are segmented. For example, in Fig. 3.a, three regions containing $3 \times 3$ grids are defined on the hypercube. Calibration methods are used to correct anomalies of the segmented hypercubes caused by temperature, spatial heterogeneity and other unknown factors. The `dataset.py` also contains a rule to define an individual spectrum. For example in Fig. 3.a, each small rectangular block is extracted into one spectrum, so that nine spectra can be extracted from each sample. Researchers can define how the hypercubes will be segmented and calibrated according to their specific problems or personal preferences.

The `view.py` contains a class executing preprocessing, feature extraction and generating training-testing sets. The preprocessing methods are intended to enhance the spatial-spectral data without changing its original space. Common preprocessing methods for spectral analysis are *Multplicative Scatter Correction* (MSC) and its variants [12], *Standard Normal Variate* (SNV) and *De-Trending* (DT) [13]. Feature extraction methods are used to form the spatial-spectral data to emphasis its characteristics for better prediction of the targeted classes. It is common to change the original space to get better features. The examples of feature extraction methods are derivative signals, spectral averaging, etc. It is also possible to define feature selection methods in the feature extraction code section where the algorithms choose some most important original bands or features. The examples of feature selection methods are entropy-based method (*info gain*) [14], correlation-based method [15], etc. The researchers can freely establish preprocessing and feature extraction methods according to their particular problems. After the final features are formed, the training and testing sets can be generated automatically according to a particular validation strategy. The common validation strategies are the percent split (holdout method) and cross-validation [16], [17]. The data structure of the training and testing sets is defined according to skdata which is a `Split` object containing pairs of `train` and `test` variables. The `train` and `test` are `Task` objects which are classified as *vector classification* containing a `x` variable as features and a `y` variable as targeted classes. For example, `Validation.splits[0].train.x` means to access the training set in the first fold which is a 2D-matrix (data $\times$ features) and `Validation.splits[0].train.y` means to get the corresponding targeted classes which can be integers, vectors or string labels.

The `main.py` is a command-line interface (CLI) entry point which can instantly be used to download, extract and preview databases through a terminal or console.

The honey hypercube dataset module is organized as in Fig. 5 containing the three core python files and a library folder (**libs**). In the library folder, prefix `sg_*`, `cl_*`, `pp_*` and `fe_*` followed by method names represent segmentation, calibration, preprocessing and feature extraction respectively. Each python file in the library folder has a standard `run` function with predefined input and output variables. The segmentation method can only be executed once but the

ISBN 978-0-9928626-7-1 © EURASIP 2017

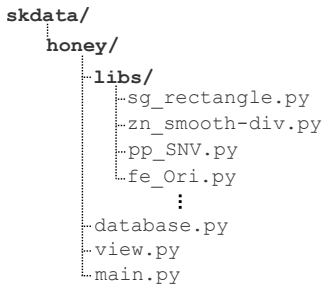



Fig. 5. The structure of the honey hypercube dataset module.

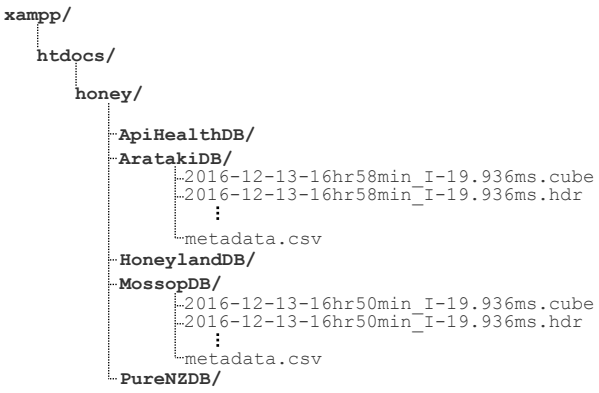



Fig. 6. A simple hypercube database in the server side. Each hypercube data (`*.cube`) is accompanied with a header file (`*.hdr`).

others (calibration, preprocessing and feature extraction) can be executed more than once sequentially. This is very useful because, for example, in the preprocessing step; SNV is usually followed by DT.

The highlights of the developed module are listed below:

- Scalability: segmentation, calibration, preprocessing and feature extraction methods can be add dynamically.
- Flexibility: it handles data from selected hypercubes and defines targeted classes through a filter mechanism.
- Buffer mode: for a limited storage computer by downloading hypercubes data per allocated memory.
- Visualization functions: Plotting of wavelength and intensity, 2D and 3D scatter graphs.
- Timer function: execution times calculation for each method.
- SHA-1 checksum: it keeps originality and avoid processing of corrupted files. SHA-1 is a standard secure hash algorithm defined by the National Institute of Standards and Technology (NIST) [18].

## VI. Hypercubes Collection and Simulation

In the preliminary development, 32 honey samples with various types obtained from five brands (*ApiHealth, Arataki Honey, Honeyland, Mossop's Honey and Pure New Zealand Honey*) in the market are collected. For each sample, seven grams of honey is placed in a circular lime glass container with 3.5 cm in diameter prior the data acquisition. The imager is from Surface Optic Corporation with code name *SOC710-VP* which is capable to capture 128 bands from 400-1,000 nm with $\pm$ 4.9 nm increment and produce 520 $\times$ 696 pixels spatial resolution with 12-bit data per pixel. One pixel in the hypercube is equal to 1 mm$^2$ in the real spatial resolution. Honey hypercudes are acquired using the imager illuminated by a homogeneous halogen lighting system. To capture variability of the samples, the hypercubes are obtained in six batches resulting 192 reflectance and 192 transmittance hypercubes in the database. The number of samples is planned to be growing from time to time.

A simulation is conducted to test the proposed honey hypercube dataset module in a local network. A local server is created using an Apache Web Server included in the XAMPP package [19]. The hypercubes are placed in folders as depicted in Fig. 6. A metadata file is placed in each folder to give information about corresponding hypercube files. In this paper, Arataki samples are used as an example to show the visualization functions for honey botanical origin classification purposes. The plot of wavelength and intensity can be depicted as in Fig. 7. The figure shows a potential segregation among the honey types. The clover honey is significantly different from mānuka honey and each mānuka variants has some differentiation visually. The module can also produce a two-dimensional scatter plot by inputing two selected bands as in Fig. 8. In addition, particular groups can be made like in Fig. 9 where the focus is to see the difference between clover and mānuka honey. The module is also able to plot an interactive 3D scatter as in Fig. 10 by inputing three selected bands. The targeted classes will follow filtering and grouping mechanisms. There are five classes in Fig. 7, 8 and 10 and two classes in Fig. 9 (the variants of mānuka are grouped together).

## VII. Conclusions

In this research, the framework of the first standard honey hypercube dataset is proposed. The standard dataset is very essential for supporting the continuity of honey analysis based on hyperspectral imaging and machine learning. The sample preparation procedure and development of the acquisition stage have been discussed and explained in order to get accurate and consistent data. The proposed honey hypercube dataset module is important to progress in segmentation, calibration, preprocessing and feature extraction methods, which can be used along with the machine learning software. The proposed module is scalable, flexible and embedded with handy features including buffer mode, timer and visualization functions. The simulation shows that the honey hypercube database can be developed and used effectively.

## References

[1] FAOSTAT. (2017) Food and agriculture data. [Online]. Available: http://www.fao.org/faostat

[2] K. Allen, P. Molan, and G. Reid, "A survey of the antibacterial activity of some new zealand honeys," *Journal of pharmacy and pharmacology*, vol. 43, no. 12, pp. 817–822, 1991.

[3] A. Noviyanto, W. Abdullah, W. Yu, and Z. Salcic, "Research trends in optical spectrum for honey analysis," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 416–425.

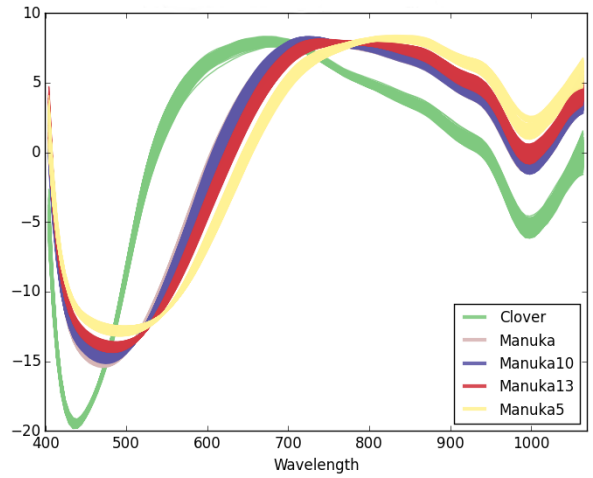



Fig. 7. The preview function on Arataki Honey containing clover, mānuka, mānuka with UMF 5+, UMF 10+, and UMF 13+.

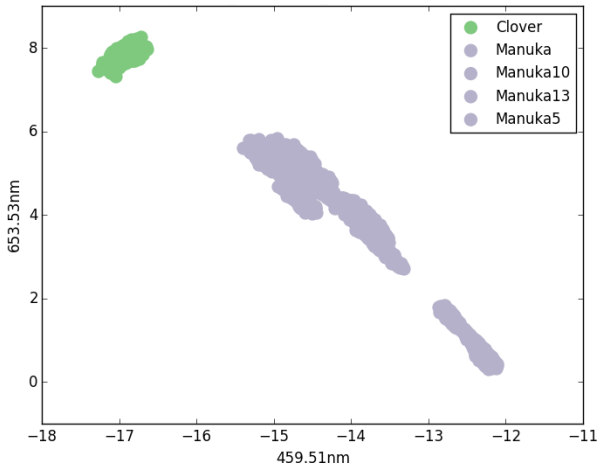



Fig. 9. The scatter plot of Arataki Honey on 459.51 and 653.53 nm grouped into two classes: clover and mānuka honey.

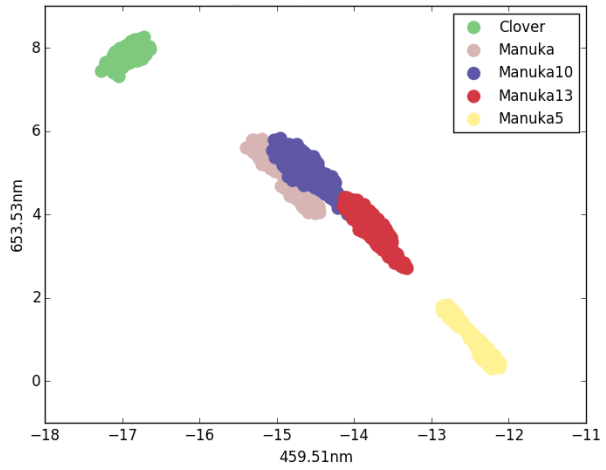



Fig. 8. The scatter plot of Arataki Honey on 459.51 and 653.53 nm.

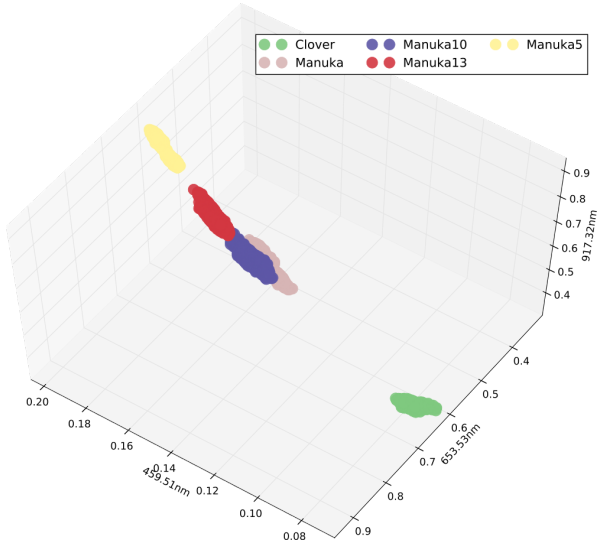



Fig. 10. The 3D scatter plot of Arataki Honey on 459.51, 653.53 nm and 917.32nm. The figure can be rotated 360°.

[4] S. Shafiee, G. Polder, S. Minaei, N. Moghadam-Charkari, S. Van Ruth, and P. M. Kuś, "Detection of honey adulteration using hyperspectral imaging," *IFAC-PapersOnLine*, vol. 49, no. 16, pp. 311–314, 2016.

[5] H. Huang, L. Liu, and M. O. Ngadi, "Recent developments in hyperspectral imaging for assessment of food quality and safety," *Sensors*, vol. 14, no. 4, pp. 7248–7276, 2014.

[6] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O. L. García-Navarrete, and J. Blasco, "Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment," *Food and Bioprocess Technology*, vol. 5, no. 4, pp. 1121–1142, 2012.

[7] K. Hamdan, "Crystallization of honey," *Bee World*, vol. 87, no. 4, pp. 71–74, 2010.

[8] R. Krell, *Value-added products from beekeeping*. Food & Agriculture Org., 1996, no. 124.

[9] J. W. White Jr, "Physical characteristics of honey," *Honey: A Comprehensive Survey. E. Crane, ed*, 1975.

[10] D. Wu and D.-W. Sun, "Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review–part i: Fundamentals," *Innovative Food Science & Emerging Technologies*, vol. 19, pp. 1–14, 2013.

[11] J. Bergstra. (2017, jan) skdata (scikit-data). [Online]. Available: https://github.com/jaberg/skdata

[12] H. Martens and E. Stark, "Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy," *Journal of pharmaceutical and biomedical analysis*, vol. 9, no. 8, pp. 625–635, 1991.

[13] R. Barnes, M. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.

[14] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 2012, vol. 454.

[15] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[16] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Stanford, CA, 1995, pp. 1137–1145.

[17] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[18] P. FIPS, "180-4–secure hash standard," *National Institute of Standards and Technology*, March 2012.

[19] A. Friends. (2017, jan) Xampp apache + mariadb + php + perl. [Online]. Available: https://www.apachefriends.org/index.html

ISBN 978-0-9928626-7-1 © EURASIP 2017