Odour-Profile Classification of Gelam, Acacia and Tualang Honey based on K-Nearest Neighbors Technique

Nurdiyana Zahed[1], Muhammad Sharfi Najib[2]
Faculty of Electrical and Electronic Engineering
University Malaysia Pahang
Pahang, Malaysia
nurdiyanazahed92@gmail.com[1], sharfi@ump.edu.my[2]

Saıful Nizam Tajuddin[3]
Faculty of Industrial Science and Technology
University Malaysia Pahang
Pahang, Malaysia
saifulnizam@ump.edu.my

Abstract— Recently, there has been growing interest in using agriculture food such as honey in food, beverage, pharmaceutical and medical industries. Specific honey type has their own usage and benefit. However, it is quite challenging task to classify different types of honey by simply using our naked eye.The purpose of this study is to apply an electronic nose (E-nose) as an instrument to produce odor profile pattern for Gelam, Acacia and Tualang honey which are the common honey in Malaysia. E-nose can produce signal for odor measurement in form of numeric resistance. Its measurement can pre-processed using normalization for standardized scale of unique features. Mean features is one of the unique features which extracted from the pre-processed data and statistical tool using boxplot representing the data pattern according to three types of honey (Gelam, Acacia and Tualang). Mean features that have been extracted were employed into K-Nearest Neighbors classifier as an input features. KNN performance have been evaluated using several splitting ratio. The results have shown that 100% rate of accuracy, sensitivity and specificity of classification from KNN using weigh (k=1), ratio 90:10 and Euclidean distance. It has been proven that the ability of KNN classifier as intelligent classification can be employed to classify different honey types from E-nose measured data.

Keywords—*Honey; Electronic Nose; Mean feature; Intelligent classification; K-Nearest Neighbors*

1. INTRODUCTION

Honey is a natural food that can be used as food additive, medicinal food and food preservative that comes with yellow color, sticky and have sweet taste. It is collected from exudate of trees and nectar of blossoms [1]. Honey can be differentiated within their types with where the honey is collected (botanical origin) which also sometimes influences honey quality, market price and honey appearance [2]. Honey comes with a unique of compound structure. Large amount of compounds are present in sample of honey that have many advantages to human health. Honey is rich with its nutrients and there are many researches have been embarked using honey as a resources for medicinal used. In Malaysia, there are three most common honey types which are Gelam, Acacia and Tualang honey[3]. Gelam honey which is smooth, strong penetrating odour, 99% soluble in warm water and amber liquid appearance[4] is collected from floral source which is *Melaleuca spp* (Gelam tree) that produce from monofloral *Apis mellifera*. Acacia honey or also called as *Robinia pseudoacacia*[5] is one type of honey that have milder taste as compared to others, transparent to light yellow color and not crystallized. Tualang honey is produced by *Apis Dorsata,* bee that produce their hives on Tualang tree (*Koompassia Excelsa*) [6] that is collected from in Rain forest of Peninsular Malaysia. Actually, it is quite a challenge job to classify honey within their group since its look quite similar in color.

Since honey have various types and high price value, there are several method for detection among honey whether to classify among their types or to identify pure honey and adulteration honey. Detection technique can be separated into three categories, chemical, image and electronic. E-nose analysis is one of the common methods for detection in food industry and precise in detection of honey. E-nose is a system mimicking human olfactory system by evaluate chemical profile of complex compound. Its function is depends to array of sensors with overlapping sensitivity [7]. It gives a signal reading and show the pattern recognition via the software in computer when it is connected using universal serial bus (USB) cable. Unlike others analytical

instrument, E-nose identifies mixture of organic samples without having to identified individual mixture present in the samples [8].

Data measurement from E-nose undergoes signal processing method. In signal processing, it is initialized with signal from E-nose, continue with normalization for pre-processing data and continue to feature extraction. In feature extraction, the mean features are used to compare samples and verify by statistical toot using boxplot. In a pre-processing technique, normalization is one of a vital step to increase accuracy in classification performance [9]. This technique generally functions to accommodate multiple range and unit of values multidimensional data by scaling and translating while the dimension have zero mean and unit variance [10]. Boxplot is one of statistical tool to express the specific characteristics of data presented from a group of datasets [11]. This technique is introduced by J.W. Tukey in year 1977[12].

Intelligent classification of honey using E-nose has already done by Zakaria et al using Probabilistic Neural Network method to classify 18 different samples of honey and the result of classification is 92.59%[13]. On the other hand, classification technique from honey detection using e-nose also applied using Artificial Neural Network by Simona Benedettiti using sensors reading as the input data of the system and performance classification is about 83.5%[14]. In order to increase the performance of honey classification from e-nose measurement, technique of intelligent classification using K-Nearest neighbors is proposed. As fast as literature is concerned, KNN has already been researched for honey classification from chemical data perspective, and the results of accuracy rate obtained have not exceeded 100% for all the samples.

The principal idea of KNN is where most frequently class level is selected as the class level for one testing sample [15]. This approach is suitable for text based problems including visual pattern recognition, in which the similarities property is compare in term of "k" nearest input between datasets with neighbourhood [16]. KNN has a huge benefit for the implementation because it is very easy to implement and computationally efficient [17] due to the less calculation, high accuracy and finer timeless compare to others machine learning algorithm [18]. Moreover, it has been proven that this approach can be very effective in time series classification problem [19]. In this approach, the new sample is compare to the training sample in the term of nearest objective function value in training space [20]. Training and testing data is declared based on splitting data process. The 10:90 training to testing data splitting is generated with one portion of first data is declared as training and the remaining portion is for testing purposed [21]. Classification problem is one of the step in KNN algorithm to measure the similarity of each cases. There are different way to calculate the similarity, for instance by using geographical graph, questionnaire, colour and so on. To make the calculation simpler, all the similarity characteristics are converted to numeric value and example of calculations are using Euclidean distance.

2. EXPERIMENTAL

2.1 Sample Preparation.

Three samples of pure honey were taken from Terengganu Honey hunter Co-operative Limited, a project that have collaboration with Terengganu Agriculture and Department. The samples were placed in room temperature condition to avoid the adulteration of the honey. All types of honey was collected from Hutan Simpan Merchang, Terengganu. Selected types of honey which are more common honey in Malaysia have been prepared which are Gelam Honey, Acacia Honey and Tualang Honey. Each types of honey was taken in a same amount of volume which is 50 ml. Then, it was placed in three labelled vials.

2.2 E-nose Setup

E-nose system was setup as Figure 1. The main component of the E-nose system is a computer to acquire and display odor measured data results, microcontroller circuit to read the data response of sample test, fan to spread the odor sample in constant flow, sensor array as the indicator for detection honey sample and last but not least is chamber house to trap odor sample to avoid mixed with surrounding odor. Firstly, the computer that has been installed with dedicated software where the result from E-nose has been displayed. USB wire was connected from computer to circuit board that contain microcontroller. A special coding software that read the output from microcontroller and display result in numeric form was employed. The fan function is to spread odor from sample, so the sensor array will be responded to give signal for the odor test. A special pipe was used to connect wiring sensors to the circuit board. Honey sample was placed under the sensor array.

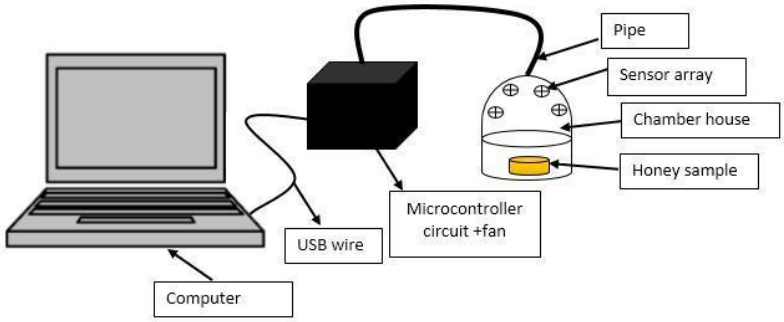

Figure 1: E-nose setup

Since the E-Nose system use non-specific sensor for detection, four gas sensors were selected in this research. The main sensor that was used which properties are conducting polymers or metal oxide sensors. Sensors array consist of four different types of sensors (S1, S2, S3 and S4).

2.3 Data Measurement

For one experiment, a sample was measured for 120s (data collection) and clean phase is 320s (E-nose neutralisation). For one sample (honey type), the experiment was repeated for five times. Then, the E-nose was changed to rest mode within 30 minutes. After 30 minutes, the E-nose was then switched to active mode. This process was repeated for the rest of the other sampled data measurements.

Table 1: Analytical condition for E-nose system

Quantity of sample in container	50 ml
Baseline phase	60 s
Measurement phase	120 s
Clean phase	320 s
Delay phase between sample test	2400 s

Table 1 shows the analytical condition for measurement using E-nose system. The total output from E-Nose is matrix numeric number with size 3000x4 that called as raw data. The generated data is from 1000x4 for Gelam test, 1000x4 for Acacia test and 100x4 for Tualang test.

2.4 Data Pre-processing

The measured raw data was pre-processed using normalization technique for reducing error and normalizing the range of the data. There are several equations were used for normalization application and the equation selection in this research is by using equation (1) below:

$$R' = R/Rmax \quad (1)$$

In equation 1, R' presents the each data reading in ohm unit from e-nose calibration while *Rmax* is the maximum reading in each data measurement. Then, the data was processed using clustering technique. It was used for data mining process and commonly used for statistical data analysis. The aim of this analysis is to grouping various data according to their cluster (group) by using function in MATLAB software. From the whole set of normalized data, there were matrix (3000 measurement x 4 array of sensors) data that were present. Those data was clustered according to group of sample data (1000 measurement x 4 array of sensors) matrix for each sample which is Gelam, Acacia and Tualang honey. After the datasets were grouped based on their sample, the data was minimized by clustering based on experiment for each sample since each of a sample measurement based on 5 different sets of experiment. Each experiment consist of (200 mean of experiment x 4 array of sensors) matrix data. The data was collected based on mean calculation for all the experiment clusters. The final mean data was represented (30 set of mean measurement x 4 array of sensors) data measurement which include (10x4 Gelam) odor data, (10x4 Acacia) odor data and (10x4 Tualang) odor data.

2.5 Feature Extraction

A boxplot or known as a whisker diagram is one of statistical analysis tool used in this research done using MATLAB software. Generally, it is used to present the distribution of data or full range of variation data. It summarizes data to five distribution which include minimum data, maximum data, median data, 1st quartile, range data (Q1) and 3rd quartile range data (Q3). Range between Q1 until Q3 represent the interquartile (IQR). The data that out of range will present as the outlier.

2.6 Intelligent Classification

KNN classifier is one of intelligent classification technique that can be run in MATLAB software as there are the setting algorithm for this function in software. To complete classification technique in this system, the distance of data was measure by applying selective rule. The measured data was compared between training and testing data.

This system was started with input and output assignment. In this step, input and output of the system was clearly declared. The input the mean of cluster data for each sample and output is the class for each sample. The value assign for input and output is remain same. Second step was data preparation. To find the best performance of intelligent classification using KNN, it undergoes data splitting or split sample technique. Total data is subsample to 'training' data and the remaining data is subsample to 'testing' data which is prepared accordingly to training to testing ratio. This practice approach is accepted by Cool et al, 1987 and already practice by other researcher using statistical measure 70:30, 60:40 and 50:50 [22]. In this step, the total data was splitted from ratio 10:90 until 90:10 before it was inserted into the system. The next step was assigning the training and testing prepared data. The data of train was assigned first in the system and continued with testing data. The system has automatically calculated the class of the testing data based on training data. Consequently, the confusion matrix was done to measure true and false case from result of classification. Lastly, performance measures of honey classification using KNN was evaluated.

Performance measure of KNN was measured using statistical analysis of error calculation. The error of classification result was done by applying the formula of MSE from equation 2.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Zi - \widehat{Zi})^2 \quad (2)$$

From the equation 2, Zi and $\widehat{Zi}$ represent observe and target value from ith observation and N is the total number of such values. Performance measure was calculated based on confusion matrix. Total predicted is 'target' and total true case is 'observe' value.

3. RESULT AND DISCUSSION

3.1 Data Case based on sample

Figure 2 until Figure 4 indicate the final data cases after clustering the data. The total cases for each sample is 10 cases present in line as indicate in 2D graph. Each graph represents a different pattern of data based on odor sample. The data is used as the input for intelligent classification using CBR technique.Sensor array 1 is methane CNG gas sensor (S1), sensor array 2 is carbon monoxide gas sensor (S2), sensor array 3 is alcohol sensor (S3) and last sensor is CO/combustible gas sensor (S4). From figure 2 and figure 3, the highest response of sensor to odorant sample is S4 due to its high sensitivity to propane which is one of the element compounds found in Gelam and Acacia honey. By observation in Figure 4, the most responsive sensor is S3 since there are high volatile compound are benzoic in Tualang honey.

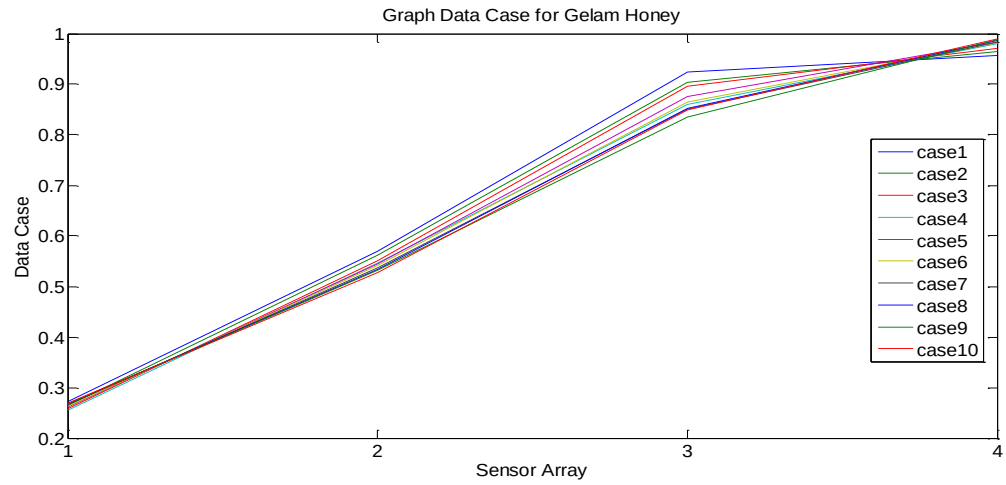

Figure 2: Graph Data Case for Gelam Honey

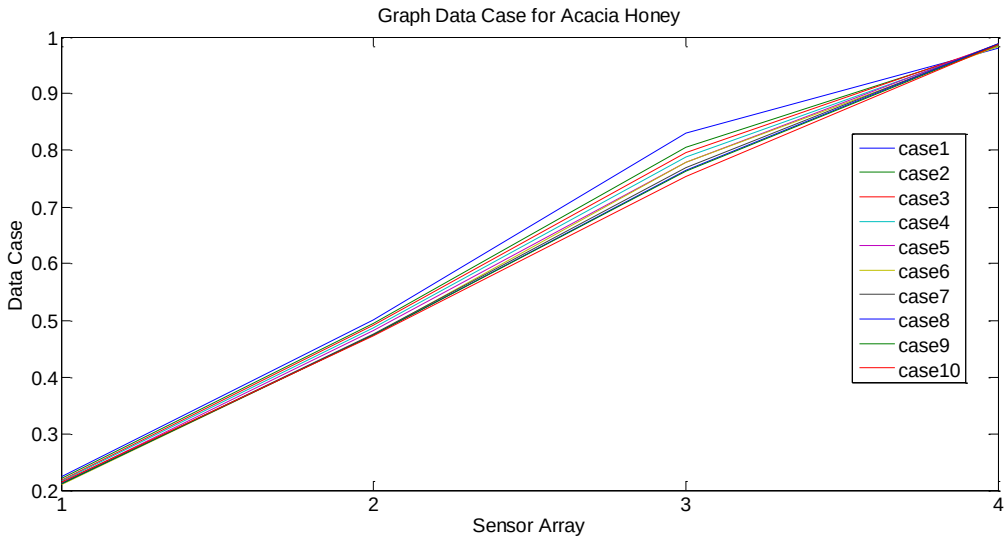

Figure 3: Graph Data Case for Acacia Honey

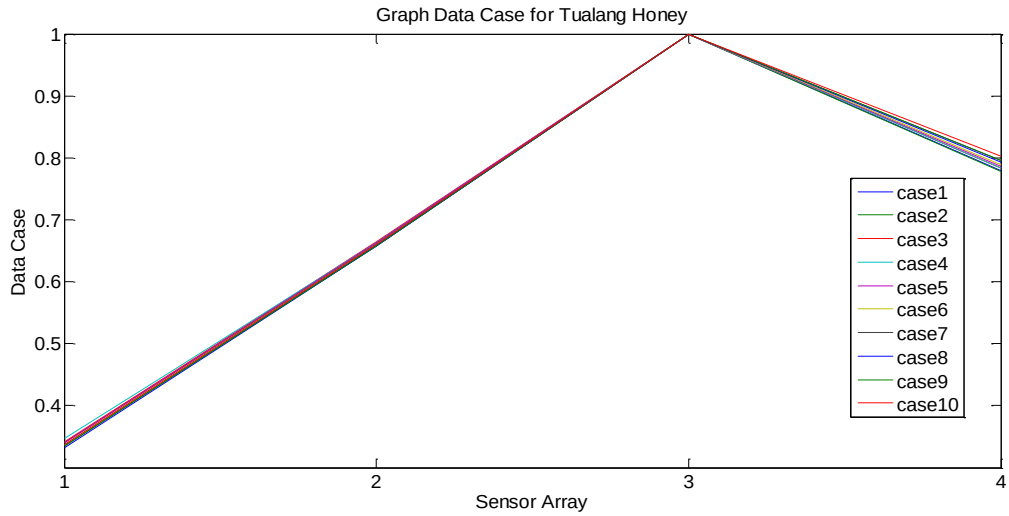

Figure 4: Graph Data Case for Tualang Honey

3.2 Statistical Analysis

In this section, the distribution data for three types of honey were represented in whole statistical tool using boxplot include minimum, maximum, median, 1st quartile, 3rd quartile, interquartile and outlier (if present). In designing this boxplot analysis using MATLAB software, the data cases for each sample with size of (10x4 matrix) data was inserted as the input data. The distribution data present in the boxplot is based on sensor array.

Figure 5 indicate the boxplot for Gelam honey. For data in first feature (sensor1), it can be seen that it is the constant value for median, 1st quartile, 3rd quartile and interquartile (IQR).This trend depicts that the data is quiet similar for feature 1 and range data is lower than 0.3. Feature 3 (F3)shown high range between maximum and minimum value as compared to the others features.

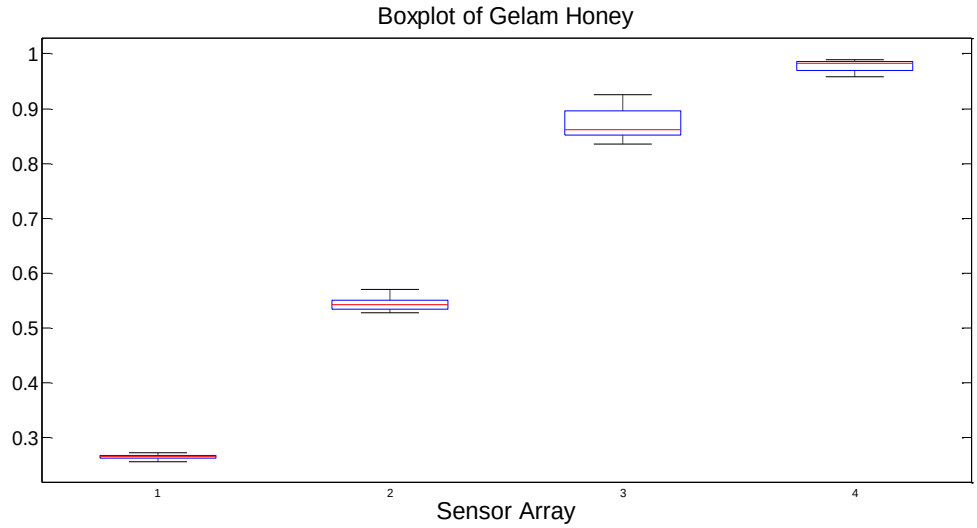

Figure 5: Boxplot of Gelam honey

Figure 6 explains the boxplot for Acacia honey. The most similar data in their feature is at Feature 4 (F4) because there are no IQR and 1st quartile, 3rd quartile, median and maximum data is at same data value. Same as Gelam case, F3 shows higher range between feature data.

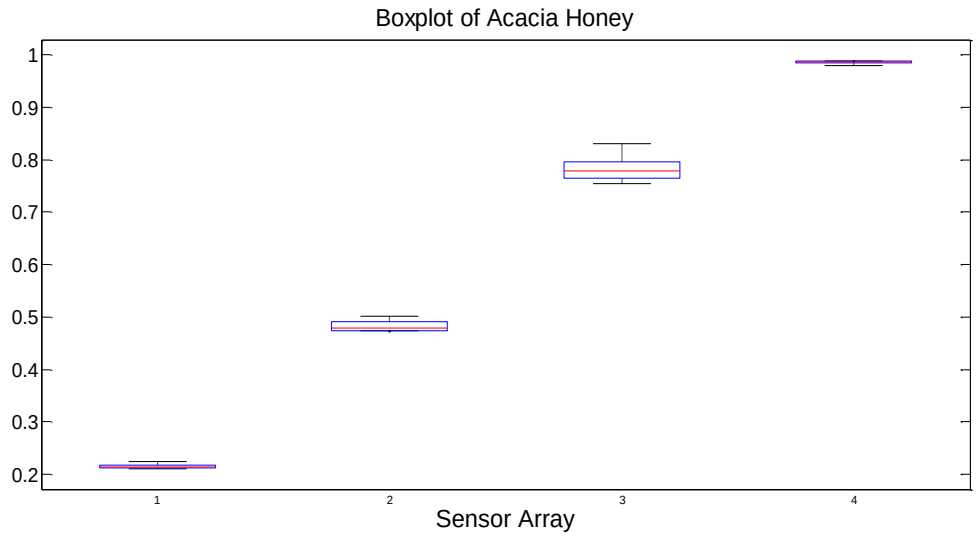

Figure 6: Boxplot of Acacia honey

Figure 7 indicates boxplot for Tualang honey. Feature 3 show only one line, means all the data in the feature is equal. In F1 and F2, there are no IQR present in feature data. F4 shows a little range of IQR.

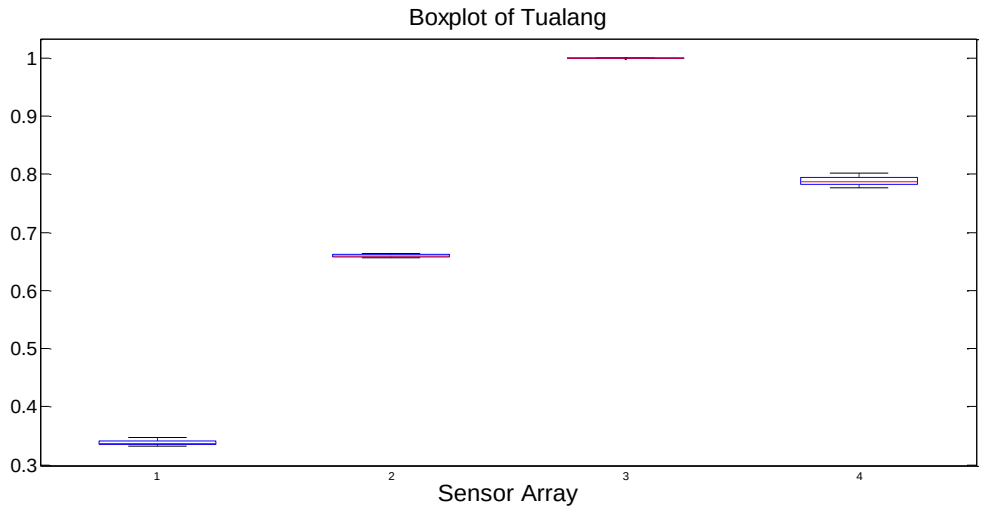

Figure 7: Boxplot of Tualang honey

3.3 K-Nearest Neighbors

Training and testing data in KNN was tested in KNN classifier by varying the rule parameters (nearest, random, consensus) using distance parameters (Euclidean, city block, cosine and correlation). The results can be observed by using weight or k=1, 2, 3.

Table 2: Summarize Mean Data as class for KNN classifier

Sample	Mean Data (Ω)
Gelam	0.6648
Acacia	0.6166
Tualang	0.6967

Table 2 indicates the mean data for each sample that was assigned as the class in KNN classifier. From all the data input, one mean data was calculated to represent the group. There are the mean data for Gelam, acacia and Tualang with the value of 0.6648, 0.6166 and 0.6967 respectively.

Table 3: Parameter Optimization KNN for K=1

Distance	Rule	Parameter Optimization (%)								
		Ratio 10:90	Ratio 20:80	Ratio 30:70	Ratio 40:60	Ratio 50:50	Ratio 60:40	Ratio 70:30	Ratio 80:20	Ratio 90:10
Euclidean	Nearest	78.33	91.67	93.81	99.17	99.33	100.00	100.00	100.00	100.00
	Random	78.70	91.46	93.81	99.17	99.33	99.58	100.00	100.00	100.00
	Consensus	78.52	91.67	93.81	99.17	99.33	99.58	100.00	100.00	100.00
City Block	Nearest	78.33	88.96	93.81	98.89	99.33	99.17	99.44	100.00	100.00
	Random	78.52	88.75	93.81	98.89	99.33	99.17	99.44	100.00	100.00
	Consensus	78.33	88.75	93.81	98.61	99.33	99.17	99.44	100.00	100.00
Cosine	Nearest	80.56	93.13	97.14	99.72	99.67	99.58	99.44	100.00	100.00
	Random	76.85	93.13	96.90	99.72	99.67	99.58	99.44	100.00	100.00
	Consensus	80.56	93.13	97.14	99.72	99.67	99.58	99.44	100.00	100.00
Correlation	Nearest	74.63	76.88	73.57	75.56	77.00	77.50	76.67	79.17	80.00
	Random	74.63	73.96	73.57	75.56	77.33	77.50	76.67	82.50	80.00
	Consensus	74.63	73.96	73.57	75.56	77.33	77.50	76.67	79.17	81.67

Table 3 shows the result of percentage similarity using KNN classifier by using K=1 and varies distance, rule and separation ratio of training to testing data from the total of 200 measurement data for three samples of honey. The input data for one observation is not overlap between training and testing data. The results from 10:90 until 90:10 training testing ratio have produced different rate of accuracy performance. It can be observed that the accuracy of lowest training ratio of 10:90 has the lowest performance. Clearly, it is shown and proved that the performance is increased by increasing the ratio of training to testing data. The rate of accuracy of 70:30 to 90:10 training testing data splitting ratio has shown consistent improvement. Based

on the table, the successful performance was selected while applying percentage of 70:30 until 90:10 training testing data splitting from input data, using distance Euclidean and varies rule nearest, random and consensus. From the result obtained for all the training testing data ratio, it can be seen that Correlation distance for all the rule have low performance as compared to other distance.

Figure 8 represents the graph of honey classification with ratio 90:10 by applying KNN method. The total of training data is 180 from each samples and testing data is 20 measurement each samples. From the graph, it can be observed that all of sample is classified correctly according to their mean data of sample.

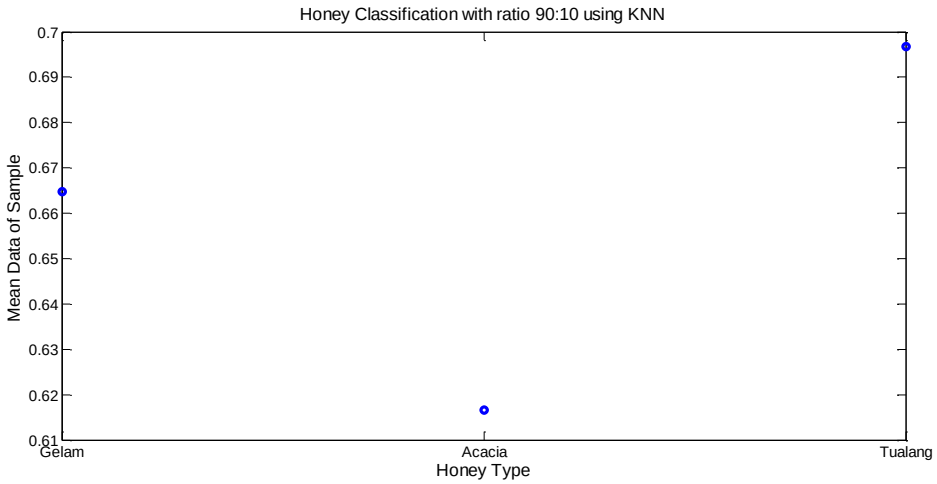

Figure 8: Graph Honey Classification with ratio 90:10 using KNN

Table 4: MSE calculation

Ratio	Mean Square Error (MSE)
90:10	0
80:20	0
70:30	0
60:40	9.68E-06
50:50	1.55E-05
40:60	1.94E-05
30:70	0.000138
20:80	0.000194
10:90	0.000499

Table 4 focuses on the error calculation using Mean Square Error (MSE) to check the performance using KNN classifier using k=1, distance=Euclidean, rule=nearest as mention the best performance to classify three types of common honey in Malaysia. Based on the result, there are zero value of MSE in ratio of 90:10, 80:20 and 70:30.The highest error can be obviously depicted in 10:90 training testing data splitting ration. Thus, the result of the classification has proven that by increasing the training to testing ratio the performance of classifier will increase.

4. CONCLUSION

This paper has successfully presented different honey odour pattern classification. Three types of pure honey, Gelam, Acacia and Tualang honey odour profiles have been classified using E-nose without going through heating process to avoid change honey chemical properties. The E-nose was able to distinguish between three types of honey by showing different data of odor profile pattern from four sensor array features. It was proven by graphical representation of statistical tool using boxplot. KNN technique was demonstrated for intelligent classification. For honey classification, KNN classifier has presented the performance of 100% rate of accuracy, sensitivity and specificity using ratio 90:10 and Euclidean distance. The contribution to knowledge of this paper is that, the KNN as the intelligent classification method for honey which produced 100% rate of accuracy .

ACKNOWLEDGMENT

We thank to Faculty of Electrical and Electronics and Faculty of Industrial Science and Technology Universiti Malaysia Pahang for provided the equipment for this project. The author acknowledges the financial scheme provided from UMP, Graduate Research Scheme(GRS).

REFERENCES

[1] S. Shafiee, S. Minaei, N. Moghaddam-Charkari, and M. Barzegar, "Honey characterization using computer vision system and artificial neural networks.," *Food Chem.*, vol. 159, pp. 143–50, Sep. 2014.

[2] S. Seisonen, E. Kivima, and K. Vene, "Characterisation of the aroma profiles of different honeys and corresponding flowers using solid-phase microextraction and gas chromatography-mass spectrometry/olfactometry.," *Food Chem.*, vol. 169, pp. 34–40, Feb. 2015.

[3] L. S. Chua, N. L. a Rahaman, N. A. Adnan, and T. T. Eddie Tan, "Antioxidant activity of three honey samples in relation with their biochemical components," *J. Anal. Methods Chem.*, vol. 2013, pp. 1–8, 2013.

[4] M. Kassim, M. Achoui, M. R. Mustafa, M. A. Mohd, and K. M. Yusoff, "Ellagic acid, phenolic acids, and flavonoids in Malaysian honey extracts demonstrate in vitro anti-inflammatory activity," *Nutr. Res.*, vol. 30, no. 9, pp. 650–659, 2010.

[5] L. A. Marghitas, D. S. Dezmirean, C. B. Pocol, M. Ilea, O. Bobis, and I. Gergen, "The development of a biochemical profile of acacia honey by identifying biochemical determinants of its quality," *Not. Bot. Horti Agrobot. Cluj-Napoca*, vol. 38, no. 2, pp. 84–90, 2010.

[6] K. Bashkaran, E. Zunaina, S. Bakiah, S. A. Sulaiman, K. Sirajudeen, and V. Naik, "Anti-inflammatory and antioxidant effects of Tualang honey in alkali injury on the eyes of rabbits: experimental animal study.," *BMC Complement. Altern. Med.*, vol. 11, no. 1, p. 90, 2011.

[7] E. Westenbrink, R. P. Arasaradnam, N. O'Connell, C. Bailey, C. Nwokolo, K. D. Bardhan, and J. A. Covington, "Development and application of a new electronic nose instrument for the detection of colorectal cancer.," *Biosens. Bioelectron.*, vol. 67, pp. 733–8, May 2015.

[8] A. D. Wilson and M. Baietto, "Applications and advances in electronic-nose technologies.," *Sensors (Basel).*, vol. 9, no. 7, pp. 5099–148, Jan. 2009.

[9] F. A. Halim, M. S. Najib, K. H. Ghazali, and M. F. Zahari, "Classification of Ammonia Odor-profile Using k-NN Technique," *Colloq. Robot. Unmanned Syst. Cybern. 2014(CRUSC 2014)*, vol. 2014, no. Crusc, pp. 4–8, 2014.

[10] G. Chakraborty and B. Chakraborty, "A novel normalization technique for unsupervised learning in ANN," *IEEE Trans. Neural Networks*, vol. 11, no. 1, pp. 253–257, 2000.

[11] K. Potter, "Methods for Presenting Statistical Information : The Box Plot a b c," *Vis. Large Unstructured Data Sets*, vol. 4, pp. 97–106, 2006.

[12] J. W. Tukey, "Exploratory Data Analysis," *Analysis*, vol. 2, no. 1999, p. 688, 1977.

[13] A. Zakaria, A. Y. M. Shakaff, M. J. Masnan, M. N. Ahmad, A. H. Adom, M. N. Jaafar, S. a. Ghani, A. H. Abdullah, A. H. A. Aziz, L. M. Kamarudin, N. Subari, and N. A. Fikri, "A Biomimetic Sensor for the Classification of Honeys of Different Floral Origin and the Detection of Adulteration," *Sensors*, vol. 11, no. 12, pp. 7799–7822, 2011.

[14] M. R. Branco, N. A. C. Kidd, and R. S. Pickard, "Electronic nose and neural network use for the classification of honey," *Apidologie*, vol. 37, pp. 452–461, 2006.

[15] N. A. Samsudin and A. P. Bradley, "Nearest neighbour group-based classification," *Pattern Recognit.*, vol. 43, no. 10, pp. 3458–3467, Oct. 2010.

[16] H. K. Lam, U. Ekong, H. Liu, B. Xiao, H. Araujo, S. H. Ling, and K. Y. Chan, "A study of neural-network-based classifiers for material classification," *Neurocomputing*, vol. 144, pp. 367–377, Nov. 2014.

[17] M. Devak, C. T. Dhanya, and A. K. Gosain, "Dynamic coupling of support vector machine and K-nearest neighbour for downscaling daily rainfall," *J. Hydrol.*, vol. 525, pp. 286–301, Jun. 2015.

[18] F. Li, J. Wang, B. Tang, and D. Tian, "Life grade recognition method based on supervised uncorrelated orthogonal locality preserving projection and K-nearest neighbor classifier," *Neurocomputing*, vol. 138, pp. 271–282, Aug. 2014.

[19] M. González, C. Bergmeir, I. Triguero, Y. Rodríguez, and J. M. Benítez, "On the stopping criteria for k-Nearest Neighbor in positive unlabeled time series classification problems," *Inf. Sci. (Ny).*, vol. 328, pp. 42–59, Jan. 2016.

[20] Y. Liu and F. Sun, "A fast differential evolution algorithm using k-Nearest Neighbour predictor," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4254–4258, Apr. 2011.

[21] M. Naughton, N. Stokes, and J. Carthy, "Sentence-level event classification in unstructured texts," *Inf. Retr. Boston.*, vol. 13, no. 2, pp. 132–156, 2010.

[22] B. Surendiran and A. Vadivel, "Classifying Benign and Malignant Masses Using Statistical Measures," *ICTACT J. image video Process.*, vol. 9102, no. November, pp. 319–326, 2011.