IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received January 19, 2021, accepted February 12, 2021, date of publication February 22, 2021, date of current version March 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3060822

Feature Extraction Technique Using Weighted Histogram Analysis Method (WHAM) for Herbs Discrimination Based on Gas Chromatography Signal

NUR FADZILAH MOHD RADZI, AZURA CHE SOH, (Senior Member, IEEE), ASNOR JURAIZA ISHAK, AND MOHD KHAIR HASSAN

Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia

Corresponding author: Azura Che Soh (azuracs@upm.edu.my)

This work was supported in part by the Universiti Putra Malaysia Grant Scheme, and in part by the Fundamental Research Grant Scheme through the Ministry of Education Malaysia.

ABSTRACT Herbs discrimination by investigating volatile compound using Gas Chromatography Mass Spectrometry (GCMS) is a common method adopted by botanists and scientists. Based on this common method, usually botanists and scientists would only focus on the major volatile compound in order to determine the species of the herbs. However, it is difficult to differentiate the herbs species of the same family group based on the pattern of chromatography signal since they may have almost similar physical features, characteristics, and aroma. In this case, the minor volatile compound needs to be considered in the herbs discrimination analysis. This study proposes the adoption of a Weighted Histogram Analysis Method (WHAM) that utilizes a combination histogram between two single feature histograms of peak area and peak height data in order to extract the new features based on minor and major volatile compound data (chemical properties) derived from chromatography signal patterns. From the results, it is found that WHAM technique results in better discrimination and classification between herbs species in same family group compared to the results without application of WHAM technique for feature extraction. The improvement in reducing the overlap between herbs group clustering can result in better classification as it will increase the classification accuracy.

INDEX TERMS Weighted histogram analysis method, gas chromatography signal, herbs discrimination, herbs classification, volatile organic compounds, feature extraction.

I. INTRODUCTION

Herbs are among the plant species which has emerged to become an important ingredient in the production of food, medicine, flavourings, health products, and perfume. The number of unknown plant species existing on earth are still high due to limited number of experts and resources on herbs. Botanists and forest rangers are usually experts in recognizing, identifying, and characterizing the plant species. Use of the sensory systems of smell and taste are two examples of the practical application of traditional herbs identification analysis. Such method is subjective and inaccurate, given that it is influenced by many factors such as physical fitness, mental health, fatigue, and other body conditions [1]–[3].

Each herb has its own unique characteristics which adds to the difficulty in studying and identifying them. Many engineering researchers have investigated the plant species on leaf part based on the physical appearances of leaves such as their texture, color, hardness, odor and taste. Various methods for feature extraction and techniques for herbs classification are proposed based on the image of the leaf as shown in Table 1 [4]–[15].

Many methods mimicking human sensory systems are invented such as electronic nose for smelling the released odor, electronic tongue for tasting the five basic types of taste (sour, sweet, bitter, salty, and umami), and camera for capturing the image of the leaf [1], [16]–[18]. In principle,

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Magno.

TABLE 1. Herbs recognition based on physical properties.

Refs.	Physical Properties	Method Applied/Proposed
[4]	Morphology Venation	Marginal ultimate venation pattern Divergence angles
[5]	Texture Shape Color	Kernel-based PSO Fuzzy Relevance Vector Machine (FRVM)
[6]	Texture Shape	Gabor filter and Gray Level co-occurrence matrix Curvelet transform coefficient and invariant moments
[7]	Shape Color	K-Nearest Neighbour Classifier
[8]	Texture	Overlapping leaves RankRLS learning algorithm
[9]	Color Texture	SVM Image Segmentation Digital wavelet transform
[10]	Texture	Gabor Wavelet Gradient Field Distribution
[11]	Texture	Gray level co-occurrence matrix (GLCM) Backpropogation multi-layer perceptron
[12]	Shape Venation	2D moment invariants Wavelet statistical features Self-Organizing Feature Map (SOM)
[13]	Shape	Zernike Moment Invariant Legendre Moment Invariant Tchebichef Moment Invariant General Regression Neural Network
[14]	Color	Automatic lesion segmentation Superpixel segmentation Random forest classifier
[15]	Color	Linear Discriminant Analysis (LDA) Color transformation

the invention of electronic system devices consists of several sensors array. Suitable sensors are selected based on the common chemical compound of the sample. The application of gas chromatography mass spectrometry (GCMS), high-performance liquid chromatography (HPLC), thin-layer chromatography (TLC), and high-speed counter current chromatography (HSCCC) are several chromatography methods used to identify the chemical compounds of plant species.

Apart from utilising electronic devices, botanists, scientists, and forest rangers differentiate the herbs species based on their chemical properties by using chromatography methods to analyse the pattern of chromatography signals (peaks, bands, etc.) of herbs [1]. GCMS is widely used to determine the volatile organic compounds (VOCs) and it is suitable for aromatic herbs sample [1], [19]. It is used to separate the chemical mixture, where each chromatography peak signal represents an individual volatile compound. Generally, priority is given to the major chemical compound in order to differentiate the herbs species. A critical issue is the fact that ignoring the minor signal will cause loss of information. Besides, as mentioned earlier, it is even difficult to recognize the herbs species when they are from the same family group since they have almost similar physical appearances and aroma characteristics. The similarity of aroma indicates the similar pattern of major signal of chromatography. Therefore, investigation on distribution patterns of chromatography signal without neglecting the minor signal using one of statistical techniques is a relatively new approach in herbs recognition system.

In 1989, Ferrenberg and Swendsen introduced a multiple histogram technique [20]. Later in the year 1992-1996, an extension of the multiple histogram technique was proposed by Kumar *et al.* [21]–[23], and this technique was known as Weighted Histogram Analysis Method (WHAM). WHAM presented an interesting approach in statistical technique, where the theory behind it is to apply it for weighting multiple single features histogram. It was discovered that the multiple histogram weighting technique gives the advantage of being able to extract all data information at once and reducing the dimensionality of data features [21], [24]. This research aims to explore the advantages of applying WHAM for feature extraction in herbs recognition system. The implementation of WHAM for feature extraction in herbs recognition system may influence group discrimination and classification accuracy.

This paper constructs a herbs recognition system using raw data gas chromatography signal, followed by signal pre-processing, feature selection, feature extraction using WHAM, and then herbs species discrimination which is performed by Principal Component Analysis (PCA), and finally investigates the accuracy of the classification results. The research mainly focuses on herbs species from the same family group since they may have almost similar physical appearances, characteristics, and aroma. The potential to discriminate herbs species by applying WHAM for feature extraction into herbs recognition system needs to be investigated. Results of herbs species discrimination will be discussed by looking at the PCA graph results with WHAM implementation and comparing it with the PCA graph results without WHAM. Next, the classification accuracy between the two are examined using the kernel support vector machine (SVM) method and k-Nearest Neighbors algorithm (k-NN).

II. RESEARCH MOTIVATION

There are different kinds of plant species and it has been a subject of interest to identify their species. The current practise to identify and distinguish each species is heavily dependent on botanists and scientists. The botanists have to go to the field for the identification process then to confirm the species. The scientists need to run the experiments in the laboratory to identify the chemical compounds of the plant species. This is inefficient and a waste of resources in terms of time and money.

The plant species is characterized according to their physical and chemical criteria. Each plant has its own unique characteristics and one of them is the leaf characteristic.

Based on the discussion with the botanist from Institute of Bioscience (IBS), Universiti Putra Malaysia (UPM), the critical parameter that they need to explore is when the herbs under the same family has a high possibility of having the same physical appearance with almost the same characteristic and aromas. It is difficult for botanists to recognize herbs simply based on physical properties.

Another method to differentiate the herbs species based on chemical properties is by using chromatography methods that produces the pattern of chromatography signals of herbs. Usually, this method requires various experimental exercises and the scientists will analyse the pattern of these chromatography signals. Based the GCMS experiment, the results show similar pattern of major signals of chromatography herbs under the same family. This research emphasizes on the formulation of a new algorithm using WHAM technique to distinguish distinctive chemical property patterns for herbs.

The new algorithm using the WHAM technique will extract the new features based on minor and major volatile compound data of the chemical properties. The new formulation algorithm gives a new unique pattern of herbs species and new database based on chemical properties has been developed. The idea is to use the technology in an existing chromatography method that has been improved and to incorporate it with the new database for different type of plant species. The innovation of this research could benefit especially the researchers, to identify the plant species without referring to the botanists and forest rangers for the learning and training process before they become expert in that field.

III. THEORY AND METHODS

A. EXPERIMENT

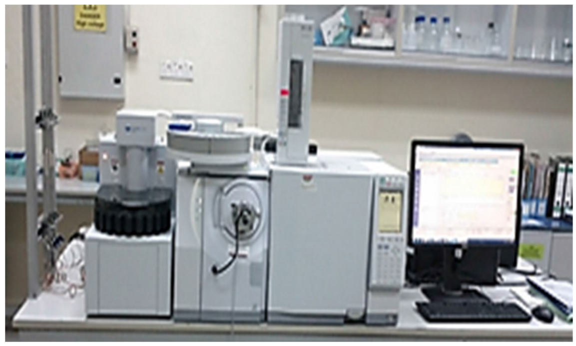
The chromatography signal was obtained using the headspace experiment of GCMS-QP2010 from SHIMADZU brand, located at Institute of Bioscience (IBS), Universiti Putra Malaysia (UPM). This model was equipped with three commercial mass spectral libraries database: Nist11, Flavors and Fragrances of Natural and Synthetic Compounds (FFNSC), and Wiley. The experiment was conducted by the expert science officer from IBS. As listed out in Table 2, eight aromatic herbs from two family groups, namely Lauraceae and Myrtaceae, are the selected samples used for the purpose of this investigation.

All the leaf samples were plucked in the morning, between 8:30am to 9:30am, in order to ensure that they were in the condition of maximum freshness. The samples were plucked from the botanical garden, which is also located in IBS. The samples were collected under the supervision of an expert botanist from IBS who validated the herbs species.

The experiment procedure for every sample started with slicing one gram of fresh leaves and placing the sliced leaves into a 10mL headspace vial. The operational condition of GCMS-QP2010 was equipped with a split injector at a temperature of 250°C, 1μL of injection volume in the split mode ratio 10:1. Helium was used as carrier gas at a constant pressure of 37.1kPa, 32.4cm/s linear velocity, and interface temperature of 300°C. MS ionization mode was set as follows: electron ionization; detector voltage at 0.87kV; acquisition mass range at 40-400u; scan speed 10000u/s; scan interval at 0.05s (20Hz); solvent delay at 5min. The experiment equipment is shown in Fig 1. Fig 2 represents a gas chromatographic signal for one sample test. Every peak represents a prediction of specific Volatile Organic Compounds (VOCs) at a distinct retention time based on the available libraries.

TABLE 2. Lists of sample herbs from lauraceae and myrtaceae species.

Family Name	Herb Name	Code Name	Country of Origin
Lauraceae	Cinnamomum Iners	L1	Botanical Garden, Institute of Biodiversity, Universiti Putra Malaysia (UPM), Malaysia
	Cinnamomum Verum	L2	
	Cinnamomum Porrectum	L3	
	Litsea Elliptica	L4	
Myrtaceae	Syzygium Aromaticum	M1	
	Syzygium Polyanthum	M2	
	Melaleuca Alternifolia	M3	
	Rhodomyrtus Tomentosa	M4	



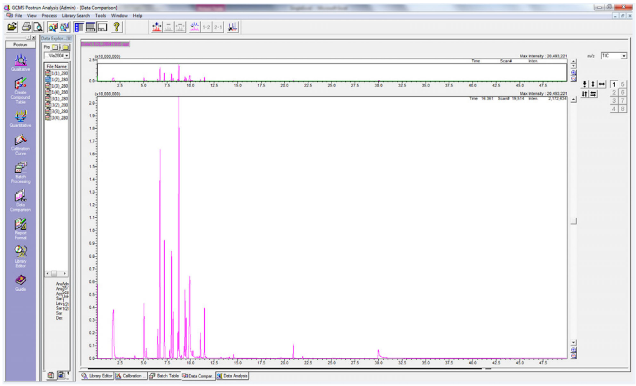
(a)

(b)

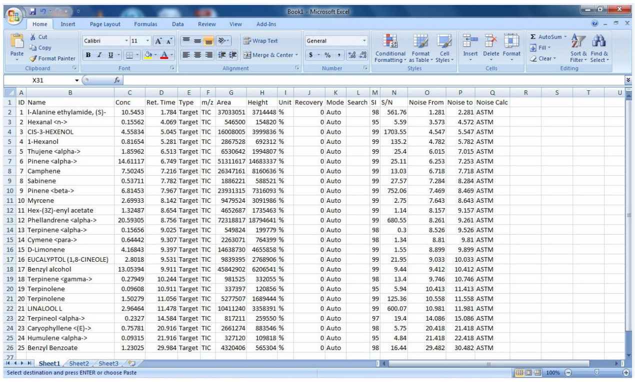
FIGURE 1. Headspace GCMS experiment setup and conducted in Botany Science Laboratory at Institute of Bioscience (IBS), Universiti Putra Malaysia (UPM), (a) GCMS-QP2010 equipment and (b) fresh leave samples.

B. SIGNAL PRE-PROCESSING

Pre-processing involves signal normalization to obtain the optimum precise results. This may include signal peak



(a)



ID	Name	Conc	Ret. Time	Type	m/z	Area	Height	Unit	Recovery	Mode	Search	SI	S/N	Noise From	Noise to	Noise Calc
1	l-Alanine ethylamide, (S)-	10.5453	1.784	Target	TIC	37033051	3714448	%	0	Auto		98	561.76	1.281	2.281	ASTM
2	Hexanal <n->	0.15562	4.069	Target	TIC	546500	154820	%	0	Auto		95	5.59	3.573	4.572	ASTM
3	CIS-3-HEXENOL	4.55834	5.045	Target	TIC	16008005	3999836	%	0	Auto		99	1703.55	4.547	5.547	ASTM
4	1-Hexanol	0.81654	5.281	Target	TIC	2867528	692312	%	0	Auto		99	135.2	4.782	5.782	ASTM
5	Thujene <alpha->	1.85962	6.513	Target	TIC	6530642	1994807	%	0	Auto		99	25.4	6.015	7.015	ASTM
6	Pinene <alpha->	14.61117	6.749	Target	TIC	51311617	14683337	%	0	Auto		99	25.11	6.253	7.253	ASTM
7	Camphene	7.50245	7.216	Target	TIC	26347161	8160636	%	0	Auto		99	13.03	6.718	7.718	ASTM
8	Sabinene	0.53711	7.782	Target	TIC	1886221	588521	%	0	Auto		99	27.57	7.284	8.284	ASTM
9	Pinene <beta->	6.81453	7.967	Target	TIC	23931315	7316093	%	0	Auto		99	752.06	7.469	8.469	ASTM
10	Myrcene	2.69933	8.142	Target	TIC	9479524	3091986	%	0	Auto		99	2.75	7.643	8.643	ASTM
11	Hex-(3Z)-enyl acetate	1.32487	8.654	Target	TIC	4652687	1735463	%	0	Auto		99	1.14	8.157	9.157	ASTM
12	Phellandrene <alpha->	20.59305	8.756	Target	TIC	72318817	18794641	%	0	Auto		99	680.55	8.261	9.261	ASTM
13	Terpinene <alpha->	0.15656	9.025	Target	TIC	549824	199779	%	0	Auto		98	0.3	8.526	9.526	ASTM
14	Cymene <para->	0.64442	9.307	Target	TIC	2263071	764399	%	0	Auto		98	1.34	8.81	9.81	ASTM
15	D-Limonene	4.16843	9.397	Target	TIC	14638730	4655858	%	0	Auto		99	1.55	8.899	9.899	ASTM
16	EUCALYPTOL (1,8-CINEOLE)	2.8018	9.531	Target	TIC	9839395	2768906	%	0	Auto		99	21.95	9.033	10.033	ASTM
17	Benzyl alcohol	13.05394	9.911	Target	TIC	45842902	6206541	%	0	Auto		99	9.44	9.412	10.412	ASTM
18	Terpinene <gamma->	0.27949	10.244	Target	TIC	981525	332055	%	0	Auto		98	13.4	9.746	10.746	ASTM
19	Terpinolene	0.09608	10.911	Target	TIC	337397	120856	%	0	Auto		95	5.94	10.413	11.413	ASTM
20	Terpinolene	1.50279	11.056	Target	TIC	5277507	1689444	%	0	Auto		99	125.36	10.558	11.558	ASTM
21	LINALOOL L	2.96464	11.478	Target	TIC	10411240	3358391	%	0	Auto		99	600.07	10.981	11.981	ASTM
22	Terpineol <alpha->	0.2327	14.584	Target	TIC	817211	259550	%	0	Auto		97	19.4	14.086	15.086	ASTM
23	Caryophyllene <(E)->	0.75781	20.916	Target	TIC	2661274	883546	%	0	Auto		98	5.75	20.418	21.418	ASTM
24	Humulene <alpha->	0.09315	21.916	Target	TIC	327120	109818	%	0	Auto		95	4.84	21.418	22.418	ASTM
25	Benzyl Benzoate	1.23025	29.984	Target	TIC	4320406	565304	%	0	Auto		98	16.44	29.482	30.482	ASTM

(b)

FIGURE 2. Chromatographic signal graph (Gas abundance (mAU) versus Retention time (minutes)) for one herb sample (a) detection of VOCs from GC/MS data file SGIMADZU software and (b) converted VOCs result to Microsoft Excel.

alignment and filtering. Generally, several data will be collected by performing the same experiment several times on one herb species to achieve robust results. Unfortunately, some of the chromatography peaks are subject to missing value and time delays caused by external factors during the conduct of the experiment. Consequently, this may cause poor discrimination and difficulty in classifying the herbs species. The purpose of signal alignment is to decide whether to include the missing VOCs peak or remove the unwanted VOCs peak. Certain data may or may not be useful. There are two categories of approaches for signal alignment, which are feature-based and profile-based [25]–[27]. Feature based approach has been selected to be applied in this study where the chromatography signal is aligned to peak matching. A chromatographic signal contains a huge number of peaks. It is necessary to extract the desired information.

Fast Fourier Transform Cross Correlation (FFTCC) is one of the methods that have been used in signal alignment application. Cross correlation applies a time-lag technique to one of two similarity measurement signals, where the correlation between two series is estimated. It is used to find the position where two signals match [28]. In this study, we employed the FFTCC proposed by Zheng *et al.* [28] to do the cross-correlation of two discrete signals in signal alignment for matching the peak signals. The chromatographic signal is gas abundance, versus retention time, as shown in Fig 2. Thus, the standard equation of cross-correlation for two discrete chromatographic signals $GC_{ref}\ (rt)$ and $GA_{al}\ (rt)$ of a real variable is defined in the Eq. (1).

$$\left(GC_{ref} * GC_{al}\right)[n] = \left\{\sum\nolimits_{m=-\infty}^{\infty} GC_{ref}{}^{*}\ [m]\right\} GC_{al}\ [n+m] \tag{1}$$

where; GC_{ref} is the reference chromatographic signal, GC_{al} is the chromatographic signal to be aligned, $GC_{ref} * GC_{al}$ is the cross-correlation values for all the variables, and $GC^{*}_{ref}\ [m]$ is the conjugate of $GC_{ref}\ [m]$. By using the FFTCC method, cross-correlation is given as notation cc in Eq. (2). Then, the forward and reverse Discrete Fourier Transform (DFT) are defined in Eq. (3). Given the discrete Fourier transformed data, RT in wavelength domain and complex number, N of rt data $(rt_0, rt_1, \ldots, rt_{n-1})$.

$$cc = real\left(F^{-1}\left\{GC_{ref} \cdot GC^{*}_{al}\right\}\right) \tag{2}$$

$$RT_k = \sum\nolimits_{n=0}^{N-1} rt_n e^{-i2\pi\left(\frac{k}{N}\right)n} \quad k = 0, \ldots, N-1 \tag{3a}$$

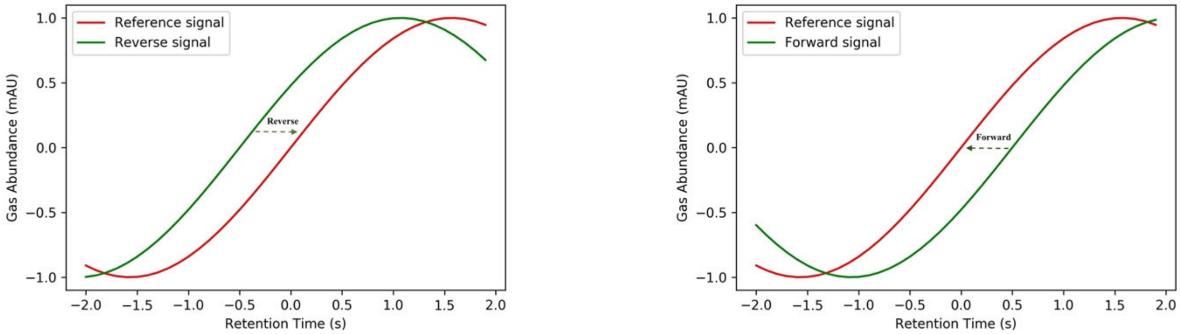
$$RT_n = \frac{1}{N}\sum\nolimits_{k=0}^{N-1} RT_k e^{+i2\pi\left(\frac{k}{N}\right)n} \quad k = 0, \ldots, N-1 \tag{3b}$$

As shown in Fig 3, DFT is used in the calculation of cross-correlation for shifting purposes. Forease of understanding, let the red line serve as our reference chromatographic signal, GC_{ref} and the other two lines (green and blue) represent the chromatographic signal which needs to be aligned, GC_{al}. Forward DFT, Eq. (3a) will be activated when the signal GC_{al} comes after the reference signal GC_{ref}. Meanwhile, reverse DFT, Eq. (3b) will be activated when the signal GC_{al} comes before the reference signal GC_{ref}. When DFT is activated, signal GC_{al} will be shifted along the rt-axis for a certain data-shift determined by the cross-correlation until it is successfully aligned to signal GC_{ref}, where at this time the value of cc reached its maximum value. It is noted that GC_{ref} and GC_{al} are DFT and inverse DFT of function $GC_{ref}(rt)$ and $GC_{al}(rt)$, respectively. GC^{*}_{al} is the conjugate of GC_{al}.

The idea of applying the mean filtering is to smooth out the several signals into one smooth signal and to reduce the amount of intensity. The average of the gas abundance is taken from n number of repeating experiments over the retention time series after the alignment signals. The equation of moving average was calculated according to Eq. (4), where GC is gas abundance value, i is the number of signal sample at retention time, rt, and n is the total of chromatographic signal samples.

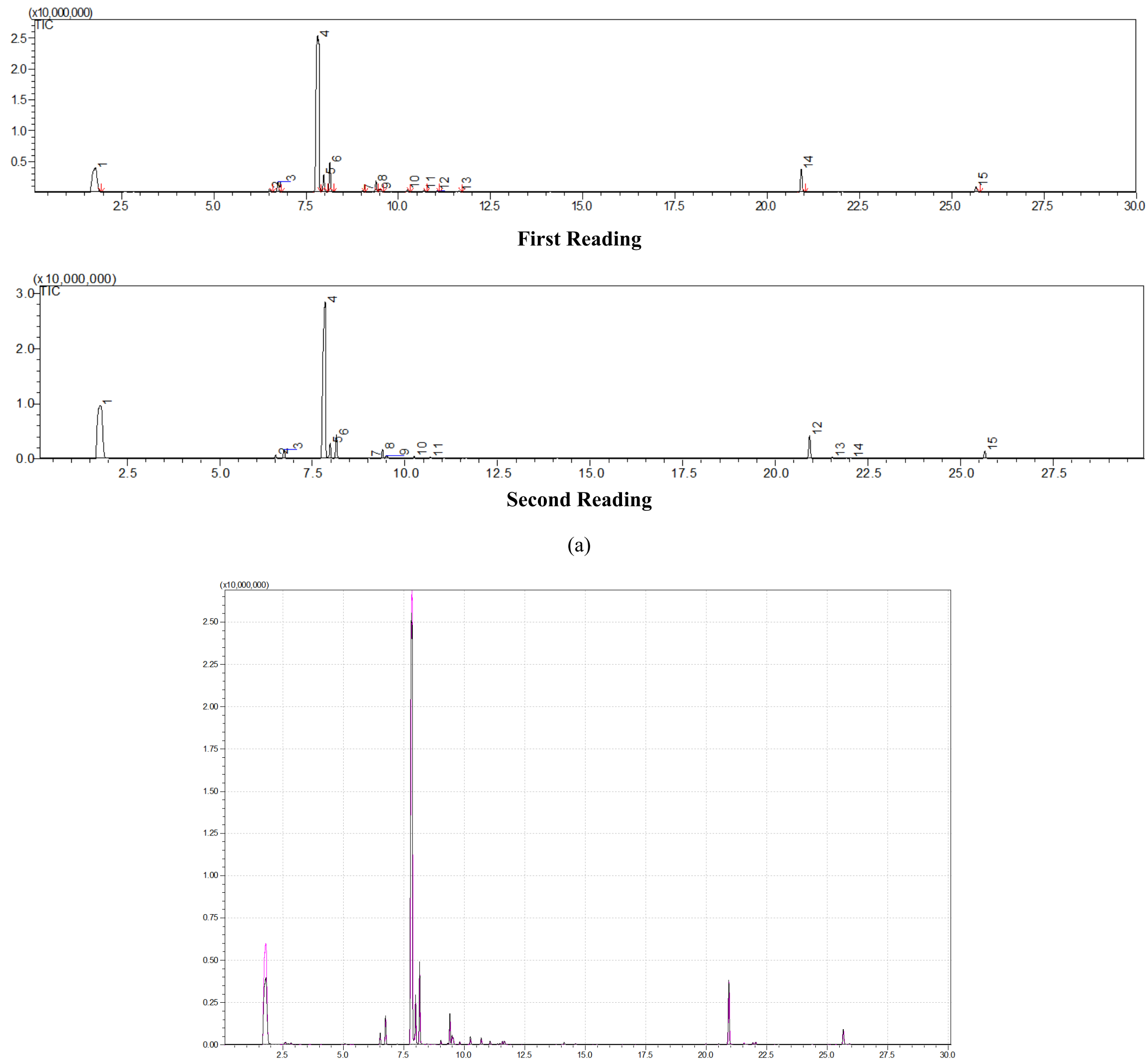
$$GC_{mean} = \frac{1}{n}\sum\nolimits_{i=1}^{n} GC_i = \frac{1}{n}\left[GC_1 + GC_2 + \ldots + GC_n\right] \tag{4}$$

These two processes help to reduce the processing time for herbs discrimination. Fig 4 represents the example of alignment results from two chromatography signals.



(a) (b)

FIGURE 3. **Alignment between two signals and reference signal (a) reverse signal and (b) forward signal.**



(a)

(b)

FIGURE 4. **Example of result comparison before and after signal alignment for one herb species (a) two chromatographic signals need to be aligned reverse signal and (b) signal alignment between 2 signals.**

C. FEATURE SELECTION AND FEATURE EXTRACTION

The dataset for each peak area and height can be generated by the GCMS data file software and saved in excel file as shown in Fig 2(b). Generally, the volatile organic compounds (VOCs) obtained from the GCMS experiment provide too much information, making it difficult to process the

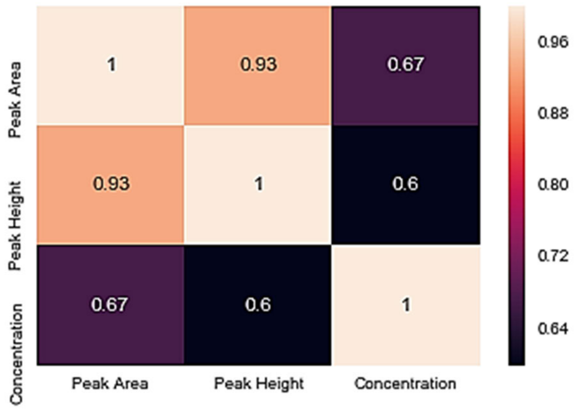


FIGURE 5. **Correlation between features from family Lauraceae.**

discrimination data. Besides, it is dependent on the available library in the GCMS to determine the compound for each peak. The histogram concept approach was applied to the raw data to assist in investigating the pattern of distribution data graphically. Histogram is focused on the frequency of data distribution of only one feature. In our case, we try to investigate the correlation histogram where the combination histogram between two single features, the peak area and the peak height, is performed. Therefore, WHAM is applied.

WHAM is one of the methods for reweighting. Using this technique to extract the new features, WHAM is used to put the data into bins in order to generate histogram. WHAM is a technique that allows better estimates to be obtained by combining several single histograms of frequency distribution as a weighted sum over the data extracted from all the single histograms and determining the functional form of weight factors that minimizes the statistical error [20]–[21]. The purpose of translating the several single histograms to WHAM histogram is to investigate the correlation between the peak area and the peak height. This is referred to as histogram correlation. Extracting new features basically refers to the mid-point of histogram correlation peak which will become the input for the next stage of herbs recognition system. The area-height weighted histogram is defined in Eq. (5).

$$P(x) = \frac{\sum_{i=1}^{N} n_i(x)}{\sum_{i=1}^{N} N_i e^{\left(\frac{F_i - U_{bias,i}(x)}{k_B T}\right)}} \tag{5}$$

$$F_i = -k_B T ln ln \left\{ \sum_{x_{bias}} P(x) e^{\left(-\frac{U_{bias,i}(x)}{k_B T}\right)} \right\} \tag{6}$$

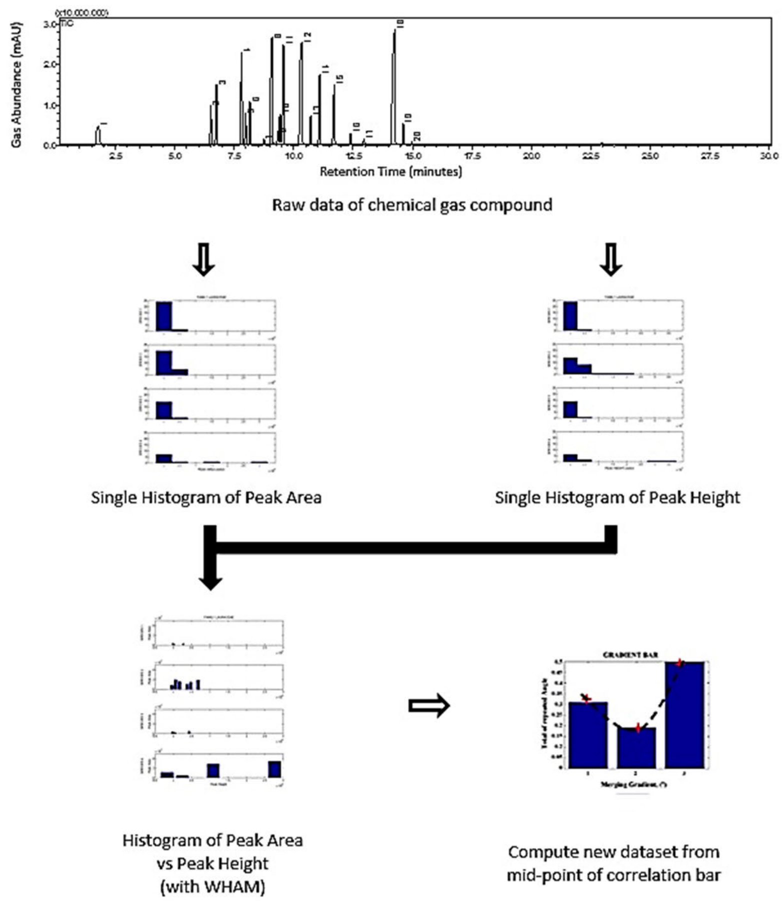


FIGURE 6. **Feature extraction process to determine the correlation between two features.**

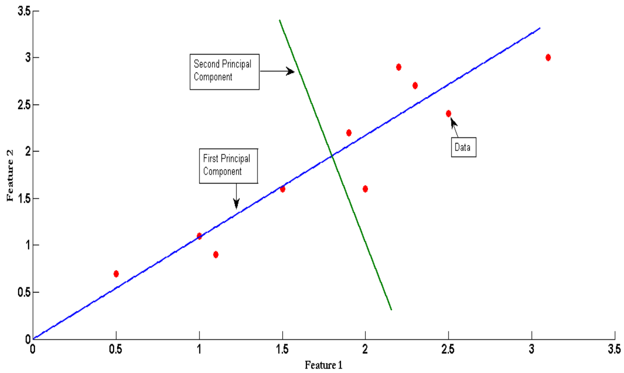


FIGURE 7. Data is projected to maximum-to-lower variance.

where; $P(x)$ = best estimate of unbiased probability distribution F_i, $P(x)$ are unknowns, F_i = biasing potential and free energy shift from simulation i, N = number of simulations, and $n_i(x)$ = number of counts in histogram bin associated with x, U_{bias}.

For feature selection, peak area and peak height of chromatographic signal are two informative features to be used in this herbs recognition system as they demonstrate the highest correlation, 0.93 as shown in Fig 5.

The feature extraction process of transforming the chromatographic signal to a single histogram of peak area and height, and subsequently to a histogram correlation as shown in Fig.6.

D. DISCRIMINANT ANALYSIS

Principal component analysis (PCA) is one of the techniques used to discriminate data into group classes by reducing the dimension data using linear transformation concept from highest variance (first principal component) to lower variance while retaining most of the information as shown in Fig. 7 [29].

The first principal component represents the highest percentage of data transformation carried forward to the next stage. The second principal component carries the second highest data transformation, and so on. The highest of principal component percentage means the lowest of data loses during data transformation. For purposes of this paper, this technique was used to study the performance of herbs discrimination of species within the same family group. The principal component is defined as:

$$y = \omega^T x_{mp} \tag{7}$$

where; y is new projected data from highest variance to lower variance, $\omega^T = \omega_1 + \omega_2 + \cdots + \omega_n$ is eigenvector, and $x_{mp} = \left\{x_{mp}^1, x_{mp}^2, \cdots, x_{mp}^n\right\}$ is the mid-point of the correlation histogram data.

E. CLASSIFICATION

Support vector machine is a powerful classification technique based on statistical approach. It is suitable for application in cases of supervised classification. Compared to other classification methods, it is capable of achieving high classification accuracy depending on how the cost and kernel parameters are set [30]. K-fold cross validation is applied in SVM in order to obtain the optimal parameter. SVM works to simultaneously search for the maximum geometric margin and minimize classification error as shown in Fig. 8 [31]. SVM tries to find the maximum separation between two hyperplanes that separate the data. The larger the margin of these hyperplanes, the better the generalization error of the classifier. Parallel hyperplanes is described in Eq. (8) where w is width or margin, b is a constant, and $f\left(x_{pca1}\right) = 0$ is a decision boundary that completely separates the 2 classes, $f\left(x_{pca1}\right) > 0, \forall x_{pca1}$ of class red, and $f\left(x_{pca1}\right) < 0, \forall x_{pca1}$ of class green.

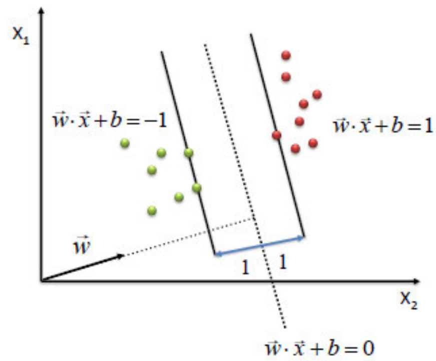


FIGURE 8. Maximum margin for two classes [31].

$$f\left(x_{pca1}\right) = wx_{pca1} + b \tag{8}$$

Data points along the hyperplanes are called Support Vectors (SV). The vector theta has to be perpendicular to decision boundary. Finding the optimal hyperplane which could best separate the data requires multiple iteration of weight, w updates in which the final separation gives the minimum cost function. The cost function in Eq. (9) is used to train the SVM giving the final equation as shown in Eq. (10).

$$h_\theta\left(x_{pca1}\right) = \frac{1}{1 + e^{-\theta^T x_{pca1}}} \tag{9}$$

$$\boxed{\begin{aligned} \min_\theta C \sum_{i=1}^{n} \Big[& y_i cost_1\left(\theta^T x_{pca1_i}\right) \\ & + (1 - y_i)\, cost_0\left(\theta^T x_{pca1_i}\right)\Big] \\ & + \frac{1}{2}\sum_{j=1}^{d} \theta_j^2 \qquad\qquad (10) \end{aligned}}$$

where the maximum margin given

$$\min_\theta \frac{1}{2}\sum_{j=1}^{d}\theta_j^2; \quad \begin{cases} \theta^T x_{pca1_i} \geq 1 & if\ y_i = 1 \\ \theta^T x_{pca1_i} \leq -1 & if\ y_i = -1 \end{cases}$$

TABLE 3. VOCs peak area and peak height from sample herb species from family lauraceae and family myrtaceae.

Species name	Time region (min)	Peak area (x 10^6mAU)	Peak height (x 10^6mAU)	Total of VOCs area (x 10^6mAU)
Family Lauraceae				
Cinnamomum Iners (L1)	0:00 – 5:00	29.741949	3.047838	46.187381
	5:00 – 10:00	14.415423	4.129691	
	10:00 – 15:00	1.410058	0.414310	
	15:00 – 20:00	0.152195	0.044640	
	20:00 – 25:00	0.119094	0.034656	
	25:00 – 30:00	0.348662	0.079755	
Cinnamomum Verum (L2)	0:00 – 5:00	37.579551	3.869268	351.180670
	5:00 – 10:00	288.467439	75.653115	
	10:00 – 15:00	17.824880	5.760296	
	15:00 – 20:00	0	0	
	20:00 – 25:00	2.988394	0.993364	
	25:00 – 30:00	4.320406	0.565304	
Cinnamomum Porrectum (L3)	0:00 – 5:00	46.241399	5.278788	62.189492
	5:00 – 10:00	15.229713	4.659738	
	10:00 – 15:00	0.248981	0.074189	
	15:00 – 20:00	0	0	
	20:00 – 25:00	0.469399	0.132451	
	25:00 – 30:00	0	0	
Litsea Elliptica (L4)	0:00 – 5:00	42.630359	4.936399	506.439362
	5:00 – 10:00	6.747486	2.157224	
	10:00 – 15:00	0	0	
	15:00 – 20:00	430.586630	63.395289	
	20:00 – 25:00	1.686776	0.457958	
	25:00 – 30:00	0	0	
Family Myrtaceae				
Syzygium Aromaticum (M1)	0:00 – 5:00	67.211317	7.544410	349.275010
	5:00 – 10:00	0	0	
	10:00 – 15:00	0	0	
	15:00 – 20:00	241.088546	22.212474	
	20:00 – 25:00	40.975147	11.383210	
	25:00 – 30:00	0	0	
Syzygium Polyanthum (M2)	0:00 – 5:00	121.970170	15.128273	137.885344
	5:00 – 10:00	8.984919	3.025867	
	10:00 – 15:00	2.171015	0.666748	
	15:00 – 20:00	4.380618	0.906117	
	20:00 – 25:00	0.378622	0.120814	
	25:00 – 30:00	0	0	
Melaleuca Alternifolia (M3)	0:00 – 5:00	43.844988	4.807609	1108.778199
	5:00 – 10:00	503.214117	121.685477	
	10:00 – 15:00	561.719094	97.833289	
	15:00 – 20:00	0	0	
	20:00 – 25:00	0	0	
	25:00 – 30:00	0	0	
Rhodomyrtus Tomentosa (M4)	0:00 – 5:00	86.950797	9.288368	216.736763
	5:00 – 10:00	126.322045	27.844246	
	10:00 – 15:00	0.756677	0.196789	
	15:00 – 20:00	0	0	
	20:00 – 25:00	2.707244	0.848694	
	25:00 – 30:00	0	0	

The success of SVM in classifying for non-linear separable data depends on the tuning of several parameters (cost parameter, C and kernel parameters (γ,d)). Grid-search method is applied in cross validation to obtain the best parameters and radial basis function (RBF) kernel. Soft margin SVM is applied to tolerate the outlier so that the constraints of the optimization problem can be solved.

Other method used for herbs classification in this research is k-Nearest Neighbor (k-NN). The k-NN classification is a non-parametric model that is described as instance-based learning in which the model is characterized by memorizing the training dataset. The algorithm is a special case of instance-based learning that is associated with zero cost during the learning process [32].

The k-NN is a supervised learning algorithm that classifies a sample by a majority vote of its neighbours. Based on the algorithm concept, the sample is allocated to the class that supported the foremost common class among its k closest neighbours. In order to determine the class, this algorithm requires training data and pre-defined k value. The

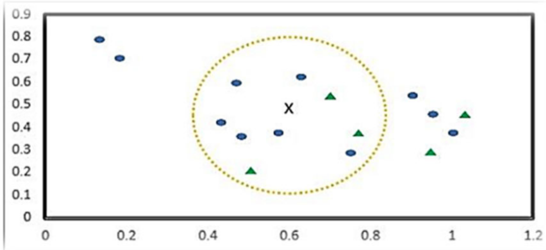


FIGURE 9. The classification of k-NN algorithm with using Euclidean distance [34].

value of k is usually a small integer with positive value. The algorithm will search through the training sample space for the k-most similar samples based on a similarity measure a distance metrics [33]. The distance metrics is one of important parameter that will also affect the performance of classification. In this study, Euclidean distance is used to find distance between a new data point and existing training dataset. Euclidean distance examines the root of square differences between coordinates of a pair of objects. For each feature x_i calculate the Euclidean distance to all other features in sample. Euclidean distance $d(x, y)$ between features x_i and y_i is calculated using the formula:

$$d(x, y) = \sqrt{\sum_{i=1} (x_i - y_i)^2} \tag{11}$$

where x_i is coordinate of the reference features and y_i is coordinate of other than the reference features.

Fig. 9 illustrates the concept of k-NN algorithm with Euclidean distance as distance metrics is used to determine the appropriate class of the new data. The data to be classified is marked "X" and the big circle is represented by the Euclidean distance computation. Based on the Euclidean distance computation, it shown that there are two possible classes which are circle class with six instances and triangle class with three instances. From the calculation of Euclidean distance, higher value of distance will indicate better separation of group or species compare to the lower value of distance. The algorithm will classify marked "X" to the circle class as the circle class has the majority of data within the radius [34].

In conclusion, the proposed herbs recognition algorithm for this study is shown in Fig 10. The flowchart sets out the process of identifying the herbs species, from the starting point of having raw chromatography signal until classification of the herbs species is achieved.

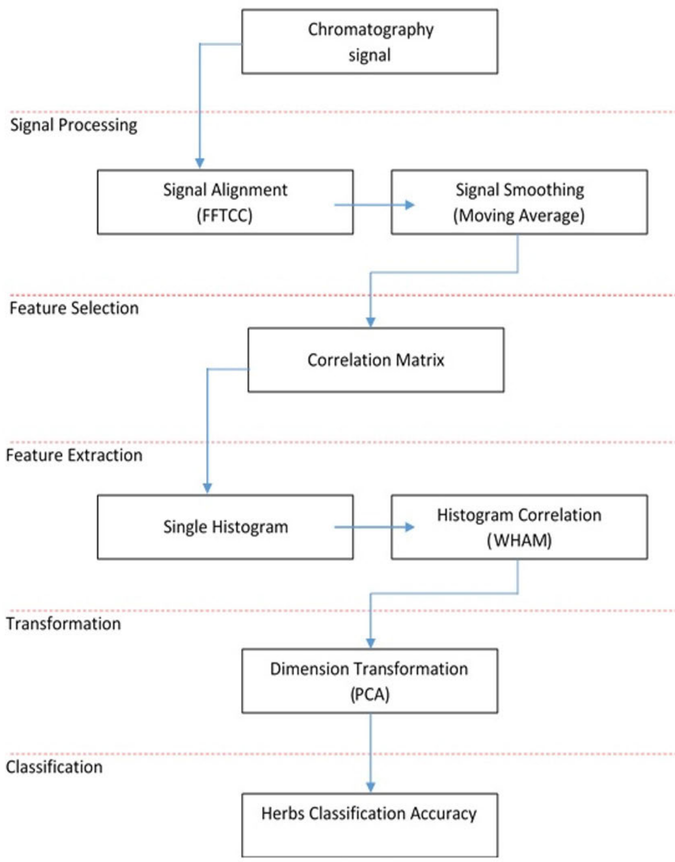


FIGURE 10. Flowchart of herbs recognition algorithm.

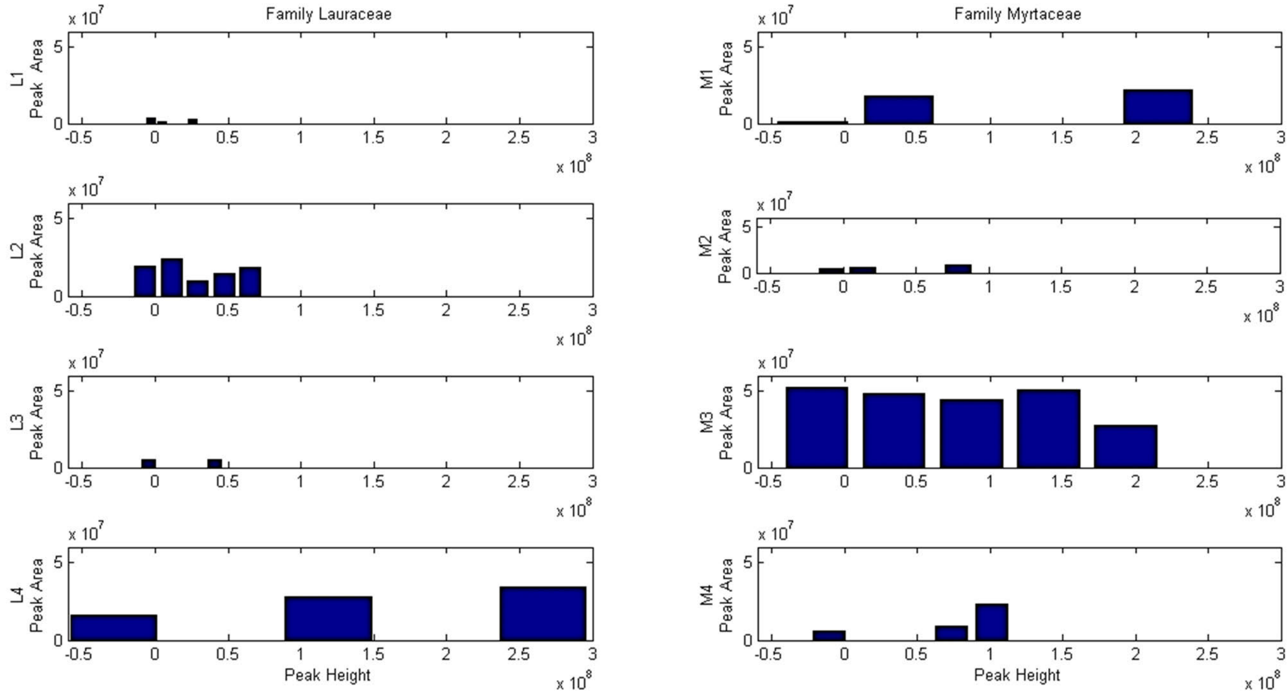


FIGURE 11. Data is projected to maximum-to-lower variance.

IV. RESULTS AND DISCUSSION

A. DATA COLLECTION

In this study, eight samples of aromatic herbs species from the two-family groups of Lauraceae and Myrtaceae were examined. The VOCs signal of each of the herb samples was collected using GCMS-Headspace experiment. Within a duration of 30 minutes, all the compounds were completely released from the samples of fresh herbs leaves. The raw data of VOCs signals for each herb are pre-processed and the results are tabulated in Table 3.

The measurement of peak area and height will be divided into six-time regions in order to investigate the distribution pattern. The correlation coefficient between the two features (area and height) shows a relationship of positive correlation as well as the degree of correlation, as tabulated in Table 4 below.

B. WEIGHTED HISTOGRAM ANALYSIS METHOD (WHAM)

Instead of studying the distribution pattern of VOCs from a single histogram, WHAM helps to gather out more information such as the correlation between features from two single histograms of VOCs' peak area and height. WHAM derivation determines the correlation of frequency between two features by assigning the reweighed potentials into bins. A different choice of number of bins leads to different reweighed potentials. The number of bins that gives better discrimination is chosen. Histogram correlation between feature peak area and peak height with 5 bins is shown in Fig10.

TABLE 4. The correlation between peak area and peak height of each species in family lauraceae and family myrtaceae.

Group Species	Code Species	Degree of Correlation
Family Lauraceae	L1	0.9582
	L2	0.9340
	L3	0.9615
	L4	0.9495
Family Myrtaceae	M1	0.9479
	M2	0.9685
	M3	0.9233
	M4	0.9476

Herbs species was discriminated applying PCA technique using the mid-point of histogram correlation peak. The first and second principal components are obtained as listed in Table 5. Fig. 11 shows the discrimination results for Family Lauraceae and Family Myrtaceae, respectively. Fig 11(a) represents the scatter plot from the original dataset, while Fig11(b) represents the PCA plotting results without WHAM, and Fig11(c) represents the PCA plotting results when WHAM is applied as feature extraction.

C. CLASSIFICATION OF HERBS

The classification accuracy is discussed to investigate the outcome of applying WHAM as feature extraction. The efficacy of two different classification methods Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN) have been

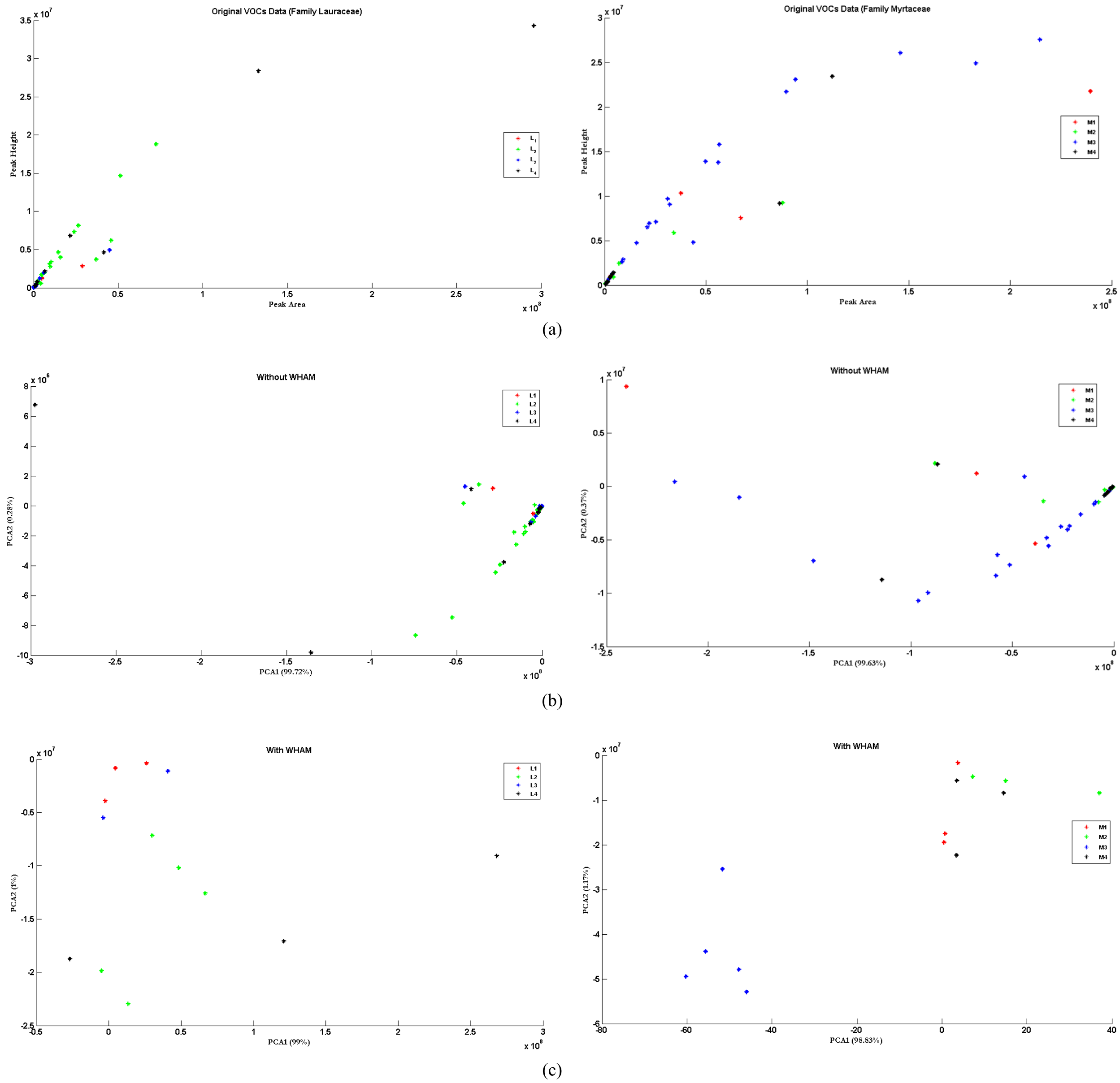


FIGURE 12. Discrimination results between herbs species using PCA (a) original dataset, (b) PCA without WHAM, and (c) PCA with WHAM.

TABLE 5. VOCs peak area and peak height from family lauraceae and family myrtaceae.

Group Species		1st Principle Component PCA1 (%)	2nd Principle Component PCA2 (%)
Family Lauraceae	With WHAM	99.00	1.00
	Without WHAM	99.72	0.28
Family Myrtaceae	With WHAM	98.83	1.17
	Without WHAM	99.63	0.37

compared. The accuracy results without WHAM and with WHAM are set out in Table 6. Improvement of classification performance is achieved when WHAM is applied as feature extraction for both Family Lauraceae and Family Myrtaceae. Based on the results, the WHAM technique shows improvement by reducing the overlap or redundancy signal between the herbs group clustering especially for the case of herb species from the same family. The enhancement in the group clustering will be improved in the classification accuracy.

In order to compare the performance of classification techniques, SVM is concluded as the better technique given the higher percentage of accuracy for range 92.32%- 95.67% compared to k-NN for 50%-75.01% percentage of accuracy for classification with WHAM technique. SVM works relatively well when there is a clear margin of separation between classes and relative memory efficiency. However, in this case

TABLE 6. Classification accuracy for family lauraceae and family myrtaceae.

Group	Family Lauraceae		Family Myrtaceae	
	Without WHAM	With WHAM	Without WHAM	With WHAM
k-NN	32.00%	50.00%	50.10%	75.01%
SVM	**57.43%**	**95.67%**	**62.11%**	**92.32%**

k-NN shows the low efficiency for herbs classification. This method depends on the selection of a "good value" for k. It is impractical for k-NN methods to assign a fixed k value to all test samples and it is also time-consuming to assign different k values to different test samples by using cross validation method.

V. CONCLUSION

In this study, eight herbs species from Family Lauraceae and Family Myrtaceae were used for herbs discrimination analysis. WHAM was adopted to investigate the correlation between features extracted from volatile compound released from the herbs leaves. The discrimination results obtained demonstrated that WHAM can be used to discriminate herbs species for both family groups by applying PCA techniques to extract the information from mid-point of histogram correlation between peak area and peak height. The problem in group clustering for the herbs species with the same family was solved using WHAM. The WHAM gives better separation result for group clustering, which has the highest similarity signal pattern and gives out a unique pattern for herbs species. The results show that the performance of classification has the highest accuracy for both SVM and k-NN by applying WHAM compared without using WHAM. However, the SVM shows better classification accuracy performance of 95.67 % (Family Lauraceae) and 92.32 % (Family Myrtaceae) compared to k-NN. As a conclusion, the formulation of this new algorithm using WHAM makes it possible to transform complicated chemical raw data into a graphical representation for a better visualization. The study also shows that the proposed technique has a good potential to improve the performance accuracy of classification for the highest similarity signal from different group classes.

ACKNOWLEDGMENT

Great appreciation goes to the Institute of Bioscience (IBS), Universiti Putra Malaysia (https://ibs.upm.edu.my) for providing all the herb samples.

REFERENCES

[1] T. C. Pearce, S. S. Schiffman, H. T. Nagle and J. W. Gardner, *Handbook of Machine Olfaction*. Weinheim, Germany: Wiley-VCH, 2003.
[2] M. D. Valle, "Bioinspired sensor systems," *Sensors*, vol. 11, no. 11, pp. 10180–10186, Oct. 2011.
[3] R. L. Doty, K. Saito, and S. M. Bromley, "Disorders of taste and smell," in *The Senses: A Comprehensive Reference*, vol. 4. 2010, pp. 859–887.
[4] E. F. Oliveira, D. G. Bezerra, M. L. Santos, M. H. Rezende, and J. A. M. Paula, "Leaf morphology and venation of psidium species from the Brazilian savanna," *Revista Brasileira de Farmacognosia*, vol. 27, no. 4, pp. 407–413, Jul. 2017.
[5] B. VijayaLakshmi and V. Mohan, "Kernel-based PSO and FRVM: An automatic plant leaf type detection using texture, shape, and color features," *Comput. Electron. Agricult.*, vol. 125, pp. 99–112, Jul. 2016.
[6] J. Chaki, R. Parekh, and S. Bhattacharya, "Plant leaf recognition using texture and shape features with neural classifiers," *Pattern Recognit. Lett.*, vol. 58, pp. 61–68, Jun. 2015.
[7] T. Munisami, M. Ramsurn, S. Kishnah, and S. Pudaruth, "Plant leaf recognition using shape features and colour histogram with K-nearest neighbour classifiers," *Procedia Comput. Sci.*, vol. 58, pp. 740–747, Jan. 2015.
[8] T. Pahikkala, K. Kari, H. Mattila, A. Lepistö, J. Teuhola, O. S. Nevalainen, and E. Tyystjärvi, "Classification of plant species from images of overlapping leaves," *Comput. Electron. Agricult.*, vol. 118, pp. 186–192, Oct. 2015.
[9] Q. K. Man, C. H. Zheng, X. F. Wang, and F. Y. Lin, "Recognition of plant leaves using support vector machine," *Commun. Comput. Inf. Sci.*, vol. 15, pp. 192–199, Sep. 2008.
[10] A. J. Ishak, A. Hussain, and M. M. Mustafa, "Weed image classification using Gabor wavelet and gradient field distribution," *Comput. Electron. Agricult.*, vol. 66, no. 1, pp. 53–61, Apr. 2009.
[11] G. Mukherjee, A. Chatterjee, and B. Tudu, "Study on the potential of combined GLCM features towards medicinal plant classification," in *Proc. 2nd Int. Conf. Control, Instrum., Energy Commun. (CIEC)*, Jan. 2016, pp. 98–102.
[12] L. Zhang, J. Kong, X. Zeng, and J. Ren, "Plant species identification based on neural network," in *Proc. 4th Int. Conf. Natural Comput.*, Oct. 2008, pp. 90–94.
[13] Z. Zulkifli, P. Saad, and I. A. Mohtar, "Plant leaf identification using moment invariants & general regression neural network," in *Proc. 11th Int. Conf. Hybrid Intell. Syst. (HIS)*, Dec. 2011, pp. 430–435.
[14] X. Mai and M. Q.-H. Meng, "Automatic lesion segmentation from rice leaf blast field images based on random forest," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot. (RCAR)*, Jun. 2016, pp. 255–259.
[15] N. Wu, M. Li, L. Chen, Y. Yuan, and S. Song, "A LDA-based segmentation model for classifying pixels in crop diseased images," in *Proc. 36th Chin. Control Conf. (CCC)*, Jul. 2017, pp. 11499–11505.
[16] H. Alam and S. H. Saeed, "Modern applications of electronic nose: A review," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 3, no. 1, pp. 52–63, Feb. 2013.
[17] P. B. Bhandare, N. S. Pendbhaje, and A. P. Narang, "Electronic nose: A review," *Res. Rev., J. Eng. Technol.*, vol. 2, no. 4, pp. 1–8, 2013.
[18] Y. Tahara and K. Toko, "Electronic tongues—A review," *IEEE Sensors J.*, vol. 13, no. 8, pp. 3001–3011, May 2013.
[19] B. K. Ghimire, J. H. Yoo, C. Y. Yu, and I.-M. Chung, "GC–MS analysis of volatile compounds of perilla frutescens britton var. Japonica accessions: Morphological and seasonal variability," *Asian Pacific J. Tropical Med.*, vol. 10, no. 7, pp. 643–651, Jul. 2017.
[20] A. M. Ferrenberg and R. H. Swendsen, "Optimized Monte Carlo data analysis," *Phys. Rev. Lett.*, vol. 63, no. 12, pp. 1195–1198, Sep. 1989.
[21] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," *J. Comput. Chem.*, vol. 13, no. 8, pp. 1011–1021, Oct. 1992.
[22] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "Multidimensional free-energy calculations using the weighted histogram analysis method," *J. Comput. Chem.*, vol. 16, no. 11, pp. 1339–1350, Nov. 1995.
[23] S. Kumar, P. W. Payne, and M. Vásquez, "Method for free-energy calculations using iterative techniques," *J. Comput. Chem.*, vol. 17, no. 10, pp. 1269–1275, Jul. 1996.
[24] M. Souaille and B. Roux, "Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations," *Comput. Phys. Commun.*, vol. 135, no. 1, pp. 40–57, Mar. 2001.

[25] M. D. Robinson, D. P. D. Souza, W. W. Keen, E. C. Saunders, M. J. McConville, T. P. Speed, and V. A. Likic, ''A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments,'' *BMC Bioinf.*, vol. 8, no. 419, pp. 1–14, 2007.

[26] T.-H. Tsai, M. G. Tadesse, Y. Wang, and H. W. Ressom, ''Profile-based LC-MS data alignment—A Bayesian approach,'' *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 2, pp. 494–503, Mar. 2013.

[27] J. W. H. Wong, C. Durante, and H. M. Cartwright, ''Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets,'' *Anal. Chem.*, vol. 77, no. 17, pp. 5655–5661, Sep. 2005.

[28] Y.-B. Zheng, Z.-M. Zhang, Y.-Z. Liang, D.-J. Zhan, J.-H. Huang, Y.-H. Yun, and H.-L. Xie, ''Application of fast Fourier transform cross-correlation and mass spectrometry data for accurate alignment of chromatograms,'' *J. Chromatography A*, vol. 1286, pp. 175–182, Apr. 2013.

[29] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, ''An overview of principal component analysis,'' *J. Signal Inf. Process.*, vol. 4, pp. 173–175, Aug. 2013.

[30] D. K. Srivastava and L. Bhambhu, ''Data classification using support vector machine,'' *J. Theor. Appl. Inf. Technol.*, vol. 2, no. 1, pp. 1–7, 2012.

[31] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, and R. A. Dudley, ''N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit,'' *J. Amer. Med. Inform. Assoc.*, vol. 21, no. 5, pp. 871–875, Sep. 2014.

[32] S. Raschka, *Python Machine Learning*. Birmingham, U.K.: Packt Publishing, 2015.

[33] M. Mohammadpour, M. Ghorbanian, and S. Mozaffari, ''Comparison of EEG signal features and ensemble learning methods for motor imagery classification,'' in *Proc. 8th Int. Conf. Inf. Knowl. Technol. (IKT)*, Sep. 2016, pp. 288–292.

[34] N. E. M. Isa, A. Amir, M. Z. Ilyas, and M. S. Razalli, ''The performance analysis of k-nearest neighbors (k-NN) algorithm for motor imagery classification based on EEG signal,'' in *Proc. Int. Conf. Emerg. Electron. Solutions IoT (ICEESI)*, Penang, Malaysia, 2017, pp. 1–6.

NUR FADZILAH MOHD RADZI received the B.Eng. degree in electrical engineering and the Master of Engineering Science degree in electrical engineering from the University of Malaya (UM), in 2009 and 2013, respectively. She is currently pursuing the Ph.D. degree in control engineering with the Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM). Her main research interests include control systems and signal and image pattern recognition systems.

AZURA CHE SOH (Senior Member, IEEE) received the B.Eng. degree in electronic/computer and the M.Sc. (Eng.) degree in electrical and electronic from Universiti Putra Malaysia (UPM), Serdang, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering from the Universiti Teknologi Malaysia, in 2011. She was a Researcher with the Department of Electrical and Electronics Engineering, Control System and Signal Processing (CSSP) Research Unit, UPM. She is also an Associate Researcher with the Malaysian Research Institute on Ageing (MyAgeing^{TM}). She is currently an Associate Professor with the Department of Electrical and Electronics Engineering, Faculty of Engineering, UPM. She has published 70 journals and 81 proceedings. Her research interests include intelligent control systems, control systems, simulation and system modeling, industrial process control, and fault detection and diagnosis systems. She is also supervising six M.Sc. and three Ph.D. students directly under the supervision, including 16 M.Sc. and ten Ph.D. as co-supervision postgraduate students. She is a Professional Technologist registered under the Malaysia Board of Technologists (MBOT) and a member of the International Association of Engineers (IAENG), Asian Control Association (ACA), Malaysia Society of Engineering & Technology (MySET), and Malaysian Society for Automatic Control Engineers (MACE).

ASNOR JURAIZA ISHAK received the bachelor's degree in electrical-mechatronic engineering from the University of Technology Malaysia (UTM), the M.Sc. degree in control automation system engineering from Universiti Putra Malaysia (UPM), and the Ph.D. degree in electrical, electronic, and system engineering from Universiti Kebangsaan Malaysia (UKM).

She is currently an Associate Professor with the Department of Electrical and Electronic Engineering, UPM. Her research interests include intelligent control systems, control system design, pattern recognition, image and signal processing, biomedical engineering, system modeling, rehabilitation, and assistive robotic.

MOHD KHAIR HASSAN was born in Melaka, Malaysia. He received the B.Eng. degree (Hons.) in electrical and electronic engineering from the University of Portsmouth, U.K., in 1998, the M.Eng. degree in electrical engineering from Universiti Teknologi Malaysia (UTM), Skudai, Malaysia, in 2001, and the Ph.D. degree from Universiti Putra Malaysia (UPM), Serdang, Malaysia, in 2011.

He was a Researcher with the Control System and Signal Processing (CSSP) Research Center, UPM, where he is currently an Associate Professor with the Department of Electrical and Electronic Engineering. His areas of interests include control systems, automotive control, electric vehicle, and AI applications. His focuses are on x-by-wire technology and optimal strategy for energy consumption in electric vehicle. He is a Professional Engineer registered under Board of Engineers Malaysia (BEM), a Corporate Member of the Institution of Engineers Malaysia (IEM), and a member of Society of Automotive (SAE).