Herbs recognition based on chemical properties using machine learning algorithm

Nur Fadzilah Mohamad Radzi, Azura Che Soh#, Asnor Juraiza Ishak, Mohd Khair Hassan

Control System and Signal Processing Research Group (CSSP), Department of Electrical & Electronic Engineering,
Faculty of Engineering, Universiti Putra Malaysia, 43400, Serdang, Selangor, MALAYSIA
#Corresponding author. E-Mail: azuracs@upm.edu.my; Tel: +603-97696322; Fax: +603-89466327.

ABSTRACT For decades, the headspace Gas Chromatography Mass Spectrometry (GCMS) technique has been employed to analyse Volatile Organic Compounds (VOCs), extracting chromatographic signals and identifying chemical components. In practical scenarios, identifying major chemical compounds has been a useful approach for herb experts to recognize and differentiate species. However, this process has been manual and lacked an automated herb recognition system that incorporates GCMS technology. To address this gap, a GCMS herb recognition system has been proposed, integrating the GCMS system with a pattern recognition approach. Innovatively, a new feature extraction method using the Weighted Histogram Analysis Method (WHAM) has been introduced. This method employs a reweighting technique that utilizes the peak area and peak height of VOCs to generate a unique pattern for each herb species. A comparison of classification performance between systems with WHAM shows that the Support Vector Machine (SVM) method achieves a higher percentage of accuracy, ranging from 92.32% to 95.67%, compared to without WHAM, which achieves an accuracy ranging from 57.43% to 62.11%. This method has demonstrated promising results in identifying herb species, and the classification method based on machine learning algorithms has proven successful in recognizing and distinguishing herb species.

KEYWORDS: Herbs recognition; Chemical properties; Classification; Machine learning algorithm; Feature extraction
Received 26 July 2023 **Revised** 11 August 2023 **Accepted** 18 August 20232 **In press** 4 September 2023 **Online** 29 September 2023

Original Article

INTRODUCTION

Plants generate aromas that consist of chemical substances known as phytochemicals. The presence of phytochemicals assists researchers in distinguishing herb organs (or parts of herbs) based on the pattern of compounds present (Kumar, 2020). Several conventional instruments have been developed to investigate these chemical compounds, which fall into three categories: volatile, semi-volatile, and non-volatile. These categories correspond to varying levels of volatility, with some gases being strong and others weak. This study focuses on an instrument that employs a chemical approach technique known as Gas Chromatography Mass Spectrometry (GCMS). Many herbs, being aromatic plants, possess a high volatile content and a high concentration of molecules (Ahmad *et al.*, 2014; Zouaoui *et al.*, 2019).

Gas Chromatography Mass Spectrometry (GCMS) is an analytical and measurement instrument widely used for identifying unknown samples by analyzing and quantifying chemical compounds. The GCMS system comprises two main components: a GCMS experimental machine and a computer that controls the operation of the GCMS system. Generally, herb species are identified by studying the quantitative results of the concentrations and compositions of the chemical constituents among species (Ichim & Booker, 2021; Wang *et al.*, 2018). The identification process is carried out manually due to the need for an in-depth analysis of VOCs. Furthermore, the determination of chemical compound names relies on the herb database library installed in the computer. Different installed libraries may result in varying sets of chemical compound names, necessitating the use of multiple libraries to obtain robust results. One disadvantage of GCMS is its lack of pattern recognition ability, which poses a challenge for botanists in manually identifying herb species. In this research, a GCMS herb recognition system is proposed by combining GCMS with a pattern recognition system to automate the herb identification process.

METHODOLOGY

Data Collection

Eight types of herb species from two different families were selected as samples for this investigation. Experts in this field, botanists at the Institute of Bioscience (IBS), Universiti Putra Malaysia (UPM), provided consultation before the selection of the herb samples.

Headspace GCMS Experiments

This experiment was conducted at IBS using the GCMS-QP2010 under the supervision of an expert science officer. The chromatography signal output was obtained upon the successful completion of the experiment. Each peak in the output corresponds to specific VOCs predicted at unique retention times, based on the libraries installed in the GCMS system.

Signal Pre-processing based on Chemical Properties Data

Signal pre-processing included signal normalization, which encompassed tasks such as signal peak alignment and filtering. Fast Fourier Transform Cross Correlation (FFTCC) is one of the techniques employed in the signal alignment process. This process is used to determine whether to incorporate missing VOC peaks or eliminate unwanted VOC peaks, with the aim of obtaining the most accurate results (Zheng *et al.*, 2013)

$$(GC_{ref} * GC_{al})[n] = \sum_{m=-\infty}^{\infty} {GC_{ref}}^{*}[m] \cdot GC_{al}[n+m] \tag{1}$$

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 $GC_{ref}[m]$.

The purpose of applying mean filtering was to create a single smooth signal by reducing the intensity variations in the signal. To achieve this, the gas abundance was determined as an average derived from a series of repeated experiments carried out over the retention time. After aligning the signals, the moving average was calculated using Equation 2. In the equation, GC represents the gas abundance value, i is the number of signal samples at a specific retention time, rt and n represents the total number of chromatographic signal samples,

$$GC_{mean} = \frac{1}{n}\sum_{i=1}^{n} GC_i = \frac{1}{n}[GC_1 + GC_2 + \cdots + GC_n] \tag{2}$$

Feature Selection and Feature Extraction Based on Chemical Properties Data

In this study, a correlation between two individual histograms of peak area and peak height was proposed by applying the concept of the Weighted Histogram Analysis Method (WHAM). To extract the new features, WHAM was employed to bin the data and generate histograms (Kumar *et al.*, 1992). The new features were obtained by identifying the mid-point of the histogram correlation peak, which would serve as an input for the subsequent stage of the herb recognition system. The area-height weighted histogram is defined in Equation 3

$$P(x) = \frac{\sum_{i=1}^{N} n_i\,(x)}{\sum_{i=1}^{N} N_i\, e^{\left(\frac{F_i - U_{bias,i}(x)}{k_B T}\right)}} \tag{3}$$

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 and defined in Equation 4 ,N = number of simulations, and $n_i(x)$ = number of counts in histogram bin associated with x, U_{bias} .

$$F_i = -k_B T\ ln \left\{ \sum_{x_{bias}} P(x)\, e^{\left(-\frac{U_{bias,i}(x)}{k_B T}\right)} \right\} \qquad (4)$$

Discriminant Analysis Based on Chemical Properties Data

Principal Component Analysis (PCA) is a technique used to classify data into distinct groups by reducing the dimensionality of the data through linear transformation. It focuses on capturing the highest variance initially and gradually considers lower variances while preserving most of the information (Rana *et al.*, 2021). In this research, PCA was employed to investigate the correlation between the variables of WHAM and to comprehend the distribution of specific variables where the principal axes of the data were widely dispersed or scattered. For this study, we chose to project the data onto the first (PCA$_1$) and second (PCA$_2$) principal components. The principal components are defined as follows:

$$y = \omega^T x_{mp} \qquad (5)$$

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

Classification Based on Chemical Properties

The classification of the herbs based on GCMS data was accomplished using a machine learning method called support vector machine (SVM). The goal of SVM is to find the hyperplane that best separates the data into distinct classes, maximizing the margin between them (Huang *et al.*, 2018; Wang *et al.*, 2020; Yan and Jia, 2018). A larger margin between these hyperplanes leads to improved generalization performance for the classifier, as it helps to reduce overfitting and enhances the model's ability to accurately classify new, unseen data. To achieve this, the SVM algorithm undergoes multiple iterations of weight updates, aiming to find the optimal hyperplane that best separates the data. The final separation achieved minimizes the cost function. The structure of the hypothesis in Equation 6 and the cost function, denoted as in Equation 7, were utilized to train the SVM model.

$$h_\theta(x_{pca1}) = \frac{1}{1 + e^{-\theta^T x_{pca1}}} \qquad (6)$$

$$\sigma = \min_\theta C \sum_{i=1}^{n} \left[y_i \text{cost}_1\left(\theta^T x_{pca1_i}\right) + (1 - y_i)\text{cost}_0\left(\theta^T x_{pca1_i}\right) \right] + \frac{1}{2}\sum_{j=1}^{d} \theta_j^2 \qquad (7)$$

where the maximum margin given;

$$\min_\theta \frac{1}{2}\sum_{j=1}^{d} \theta_j^2 \quad ; \begin{cases} \theta^T x_{pca1_i} \geq 1 & if \quad y_i = 1 \\ \theta^T x_{pca1_i} \leq -1 & if \quad y_i = -1 \end{cases} \qquad (8)$$

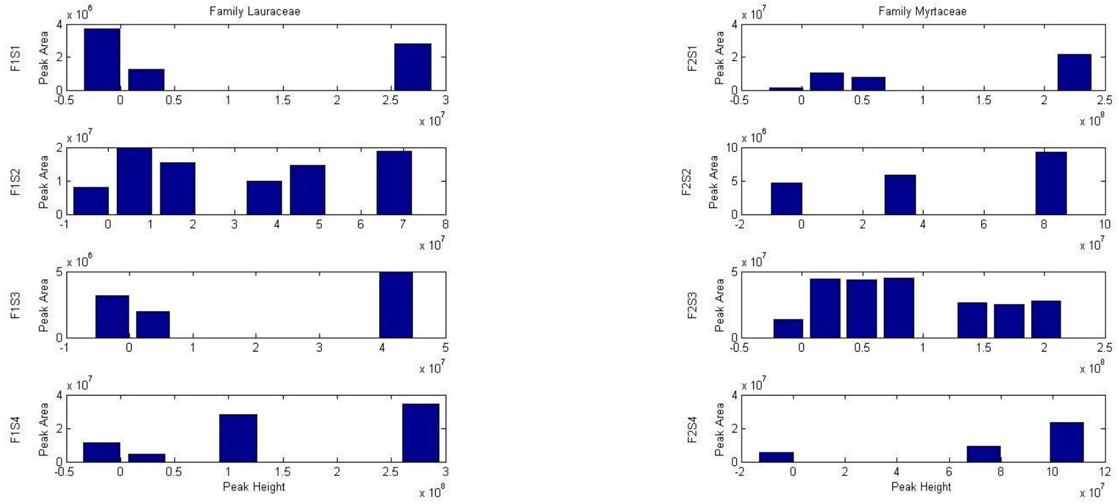
RESULT AND DISCUSSION

The study of VOCs distribution patterns involved not only analyzing a single histogram but also employing a WHAM to gain additional insights, such as understanding the correlation between the peak area and height of VOCs. It was necessary to ensure a strong relationship between the two single histograms before applying WHAM. The analysis revealed a positive correlation coefficient between the peak area and peak height, with the degrees of correlation presented in Table 1.

Table 1. The correlation between peak area and peak height of each species.

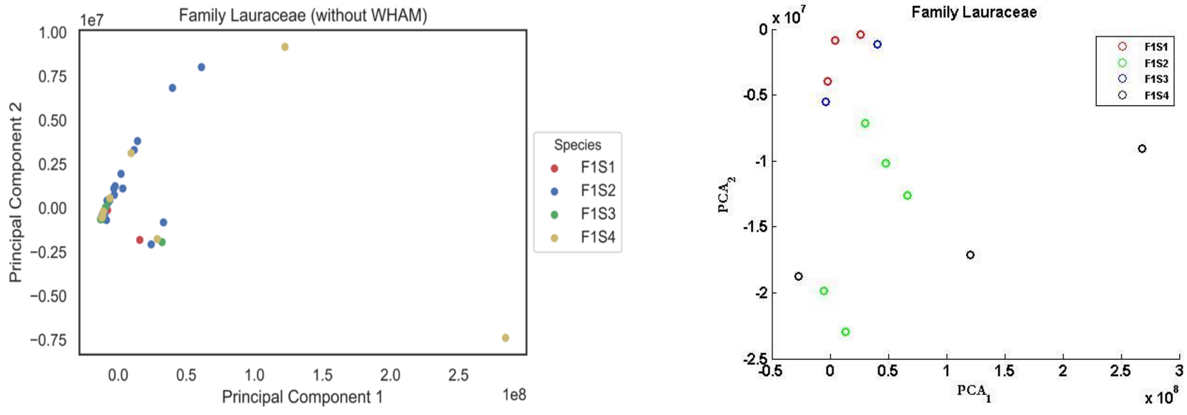
Group Species	Herbs Species	Degree of Correlation
Lauraceae	Medang Teja(F1S1)	0.9582
Lauraceae	Kayu Manis (F1S2)	0.9340
Lauraceae	Medang Wangi (F1S3)	0.9615
Lauraceae	Medang Kesing (F1S4)	0.9495
Myrtaceae	Cengkih (F2S1)	0.9479
Myrtaceae	Daun Salam(F2S2)	0.9685
Myrtaceae	Gelam Wangi (F2S3)	0.9233
Myrtaceae	Kemunting (F2S4)	0.9476

By partitioning the reweighted potentials into bins, it became possible to determine the correlation frequency between two features using WHAM derivation. Different choices of the number of bins yielded a range of reweighted potentials. The number of bins that provides better discrimination was selected as shown in Figure 1 - 3. The application of the WHAM technique enabled the generation of graphical representations from complex chemical raw data. All the WHAM histograms were compiled into the GCMS herbs database.



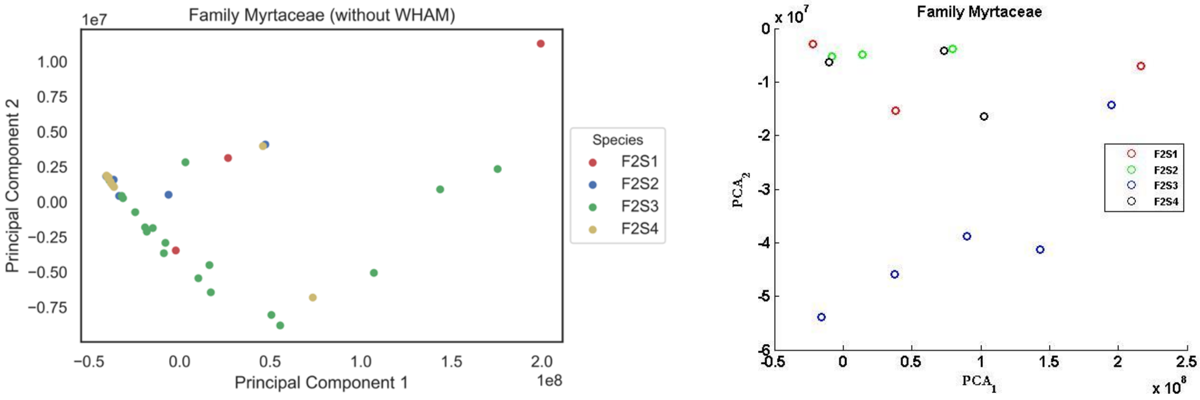
(a) (b)

Figure 1. Reweighed of the Area and Height of volatile compound into 8 bins for (a) Family *Lauraceae* and (b) Family *Myrtaceae.*



(a) (b)

Figure 2. Data distribution for Family *Lauraceae*: (a) PCA without WHAM, (b) PCA with WHAM



(a) (b)

Figure 3. Data distribution for Family *Myrtaceae*: (a) PCA without WHAM, (b) PCA with WHAM

After converting the single histograms into correlation histograms, the herb species were successfully discriminated. The midpoint of the correlation histogram peak was determined using the PCA technique, and the results of both principal components are presented in Table 2. PCA was employed to transform the data into a higher-dimensional space that captured a greater variance in the distribution. In Figure 2(a) and Figure 3(a), the graphical distribution of the projected data using PCA without WHAM is not well separated for all species. There is significant overlap observed between species within the same family group. In contrast, Figure 2(b) and Figure 3(b) exhibit reduced overlap when the WHAM technique is applied. Based on the discrimination results obtained, WHAM has proven to be effective in distinguishing between herb species belonging to different family groups. The information extracted from the midpoint of the histogram correlation between peak area and peak height played a crucial role in this discrimination. It has demonstrated superior separation results with the highest similarity signal pattern, forming distinct herb species patterns that serve as a solution for group clustering. By applying WHAM, classification performance results showed the highest accuracy for SVM, as compared to the models without WHAM as shown in Table 3. This is because data with bad separation among classes will give negative impact on the classification performance.

Table 2. The correlation between peak area and peak height of each species.

Group Species	Feature Extraction	PCA_1 (%)	PCA_2 (%)
Lauraceae	Without WHAM	0.9582	1.00
Lauraceae	With WHAM	0.9340	0.28
Myrtaceae	Without WHAM	0.9615	1.17
Myrtaceae	With WHAM	0.9495	0.37

Table 3. Herbs classification accuracy using SVM

Feature Extraction	*Lauraceae*	*Myrtaceae*
Without WHAM	57.43%	62.11%
With WHAM	95.67%	92.32%

CONCLUSION

In conclusion, the integration of a pattern recognition system into the proposed GCMS system has facilitated the automation of herb species identification. The feature extraction method, WHAM, was introduced using the concept of histogram correlation between VOCs peak area and peak height. By combining WHAM with the SVM classification method, the system's performance was significantly enhanced. The results demonstrated that WHAM exhibited superior separation capabilities for group clustering, showing the highest similarity in signal patterns and producing distinct patterns for each herb species. This study underscores the substantial potential of the proposed technique in enhancing classification accuracy across various herb species. The ability to generate distinct and unique patterns for each species highlights the effectiveness of the approach in species differentiation. This advancement in automated herb species identification holds promise for diverse applications and may lay the groundwork for further research and developments in the future.

ACKNOWLEDGEMENTS

Special gratitude is extended to the Institute of Bioscience (IBS) at Universiti Putra Malaysia (https://ibs.upm.edu.my) for generously providing all the herb samples.

REFERENCES

[1] Ahmad, R., Baharum, S., Bunawan, H., Lee, M., Mohd Noor, N., Rohani, E. R., Ilias, N. & Zin, N. M. 2014. Volatile Profiling of Aromatic Traditional Medicinal Plant, Polygonum minus in Different Tissues and Its Biological Activities. *Molecules*, 19, 19220-19242.

[2] Huang, S., Cai, N., Pedro, P.P., Narrandes, S., Wang Y. & Xu, W. 2018. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics*, 15(1), 41-51.

[3] Ichim, M. C. & Booker, A. 2021. Chemical authentication of botanical ingredients: A review of commercial herbal products. *Frontiers in Pharmacology*, 12, 1-130.

[4] Kumar, A. 2020. Phytochemistry, pharmacological activities and uses of traditional medicinal plant Kaempferia galanga L. – An overview. *Journal of Ethnopharmacology*, 253, 112667.

[5] Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. 1992. The weighted histogram analysis method for free-energy calculations on biomolecules. *Journal of Computational Chemistry*, 13(8), 1011-1021.

[6] Rana, P., Liaw, S. Y., Lee, M. S. & Sheu, S. C. 2021. Discrimination of four Cinnamomum species with physico-functional properties and chemometric techniques: Application of PCA and MDA models. *Foods*, 10(11), 2871.

[7] Wang, Z., Chen, W., Gu, S., Wang, Y. & Wang, J. 2020. Evaluation of trunk borer infestation duration using MOS E-nose combined with different feature extraction methods and GS-SVM. *Computers and Electronics in Agriculture*, 170, 105293.

[8] Yan, X. & Jia, M. 2018. A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing. *Neurocomputing*, 313, 47-64

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

[10] Zouaoui, N., Chenchouni, H., Bouguerra, A., Massouras, T. & Barkat, M. 2019. Characterization of volatile organic compounds from six aromatic and medicinal plant species growing wild in North African drylands. *NFS Journal*, 18, 19-28.