engineering proceedings

MDPI

Proceedings

Classification of Teas Using Different Feature Extraction Methods from Signals of A Lab-Made Electronic Nose †

Irari Jiménez-López, Jeniffer Molina and Juan Manuel Gutiérrez *

Bioelectronics Section. Department of Electrical Engineering CINVESTAV-IPN; irari.jimenezl@cinvestav.mx; jeniffer.molinaq@cinvestav.mx

* Correspondence: mgutierrez@cinvestav.mx

† Presented at the 2nd International Electronic Conference on Chemical Sensors and Analytical Chemistry, online, 16-30 September.

Abstract: Tea and herbal infusions are the most consumed non-alcoholic beverages worldwide and possess bioactive components with multiple health benefits. They are categorized into different classes that depend on the elaboration process, origin, and components. Commonly, analytical methods are employed to classify tea according to its chemical composition by liquid and gas chromatography-mass spectrometry, among others. Novel methods, such as electronic noses (e-noses), effectively provide real-time and objective monitoring of odors for extended periods of time. This work aimed to classify 8 different types of tea (green, white, black, spearmint, mint, hibiscus, lemongrass, chamomile) using two feature extraction methods and two pattern recognition analyses that were compared. A total of 34 tea samples were analyzed by e-nose consisting of an olfactometer as a sample handling system, seven chemo-resistive gas sensors, and a 12-bit analog-to-digital converter. Tea samples were measured 10 times to ensure repeatability, resulting in a database of 340 tea measures with 2499 samples each per sensor. Data were pre-processed using Principal Component Analysis (PCA) and Parallel Factor Analysis (PARAFAC). The information extracted was classified by Artificial Neural Network (ANN) and k-nearest neighbor (k–NN). The best architecture in ANN and distance in k-NN were demonstrated by 10 k-fold cross-validation. The classification rate was 93% in ANN and PCA, 73% in ANN and PARAFAC, 94% in k-NN and PCA, and 84% in k-NN and PARAFAC. This demonstrates that conventional PCA is better than complex PARAFAC. Our findings not only contribute to the field of tea and herbal infusions classification but also underscore the potential of e-nose systems for discriminating between diverse tea types and herbal infusions based on their odor profiles.

Keywords: tea; e-nose; PCA; PARAFAC; ANN; k-NN

Citation: Jiménez-López, I.; Molina, J.; Gutiérrez, J.M. Classification of Teas Using Different Feature Extraction Methods from Signals of A Lab-Made Electronic Nose. **2023**, *4*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Received: date
Accepted: date
Published: date

1. Introduction

Tea is the world's most consumed aromatic, non-alcoholic beverage after water. It is common to refer to tea and herbal infusions as equal, yet the term tea focuses on the Camellia sinensis. Teas are classified into different groups depending on their manufacturing process. On the other hand, herbal teas or infusions are made with fruits, flowers, and leaves of a variety of plants [1]. They possess multiple human health functions, like antioxidation, anti-inflammation, and immune regulation, among others [2]. There are biochemical components responsible for the color, taste, and aroma of tea; those related to the aroma are named volatile organic compounds (VOCs) [3].

Commonly, analytical methods, like GC-MS and FT-IR spectrometry are employed in the industry to classify products according to their chemical composition [4,5]. A new concept of analytical methods that emerged in the last years is known as electronic noses (e-noses), which identify odors by detecting the "fingerprint" of a chemical compound.

E-noses are a new concept in analytical procedures that have arisen recently. They identify scents by detecting the "fingerprint" of a chemical component. These systems are often composed of a gas-sensor array to detect odors and a processing data tool to analyze the information [6]. Currently, e-noses consider one subsystem related to the sampling handling to deliver odors to the sensor array. Today, e-noses are employed in several applications including medicine, healthcare, food, and beverages [7].

The data processing stage is one of the most important components in e-nose development since it helps to generate a coherent and useful response. Often, this data processing and modeling are based on the use of artificial ANN, k-NN, support vector machine (SVM), and random forest (RF), to name a few. Nevertheless, the signals from e-noses are characterized by their high dimensionality and non-stationary regions that demand feature extraction methods focused on reducing data. For this purpose, data could undergo mathematical transformations such as PCA, PARAFAC [8], and multi-way analysis, to mention the most widespread.

This work implements a processing strategy that employs PCA and PARAFAC techniques to extract data features from an e-nose that analyzes tea samples and compares their relevance using two recognition models based on ANN and k-NN.

2. Materials and Methods

2.1. Instrumentation and Data Collection

The tea database was obtained using a lab-made e-nose consisting of an olfactometer that controlled the odor stimuli and injected the sample's VOCs into a chamber containing seven metal oxide sensors (MOXs) from MQ-series. These sensors detect various gases, including carbon monoxide, liquefied petroleum gas, natural gas, alcohol, benzene, methane, and hydrogen. The voltage values were obtained by an acquisition board with a 12-bit analog-to-digital converter at a 5 Hz sampling frequency and a Raspberry Pi 3+B single-board computer [9]. The set of samples was formed by 34 unblended tea samples from commercial brands. Teas were categorized into eight classes: green, white, black, spearmint, mint, hibiscus, lemongrass, and chamomile. Each tea sample (4.5 g) was placed in the e-nose platform for analysis. Then, odor stimuli started and lasted 500 s distributed as follows: 5 s for rest, 35 s for odors stimulation, and 460 s for relaxation. As a result, 2499 samples were collected for each tea and sensor.

Each tea was sampled and recorded 10 times to analyze the experiment's repeatability. After every experiment, the sensor chamber was cleaned ten times with pure air. Finally, the database was shaped as a tridimensional matrix of 2499 samples, 340 records (34 teas x 10 repetitions), and 7 sensors.

2.2. Data Feature Extraction and Modeling

E-nose data were analyzed using PCA and PARAFAC methods to reduce dimensionality and extract relevant features designed to improve the classification task. Whereas PCA finds the linear correlation between the original data variables to produce new uncorrelated linear combinations of these variables using an orthogonal transformation [10], PARAFAC is a multi-way data decomposition method closely related to PCA applied to higher-order arrays [11].

One of the goals of these techniques is to determine the representative number of components that better represent the original data. For PCA, significant principal components (PCs) could be chosen considering the accumulative variance since the algorithm typically orders them according to the most relevant variance; in this way, the first PCs usually represent the maximum variation present in the original variables. On the other hand, PARAFAC assumes the existence of a triple path in the data and finds an unique solution so that the components can be rearranged and scaled arbitrarily [11]; for the selection of the optimal components, a diagnostic test is usually based on a core consistency diagnostic (CORCONDIA) [12].

Two different classification models were used after the feature extraction stage, allowing to identify patterns in the data. The first was ANN, which is based on the supervised learning approach. [13]. Its optimization was done using a standard trial-and-error process, where several parameters are fine-tuned to find the best configuration to achieve the performance. The second was k-NN, a popular supervised model that finds a group of *k* objects in the training set that are near to the test object. k-NN orders the information by computing distances between feature values [14].

In this way, four combinations were performed: PCA–ANN, PCA– k-NN, PARAFAC–ANN, and PARAFAC–k-NN. A k-fold cross-validation (k=10) was carried out to determine these classification models' classification capability to compare their performance. Considering that the tea classes did not have the same number of samples; each class was split separately to ensure that the folds contained at least one sample of each type. The training matrix was built with 306 observations, while the test matrix included 34 teas. The training data were normalized in the interval of [0,1], and the maximum and minimum values obtained were used to normalize test data.

Finally, for each case, a confusion matrix was calculated to determine performance metrics: accuracy, precision, recall or sensitivity, and specificity [15].

3. Results and Discussion

The data processing was performed on an AMD Ryzen computer. Different algorithms were written by the authors in MATLAB (Math Work, Natick, MA, version R2022b) using three different toolboxes for the routines: Machine Learning Toolbox v12.0, and Deep Learning Toolbox v14.1. In addition, Eigenvector PLS_toolbox v7.8 was used to calculate PARAFAC components.

3.1. PCA Results

Data were organized in a two-dimensional array for PCA, with the rows denoting teas and the columns denoting measurements for each sensor. In this way, PCA was applied to a matrix of dimensions of 340 x 17493. Figure 1 shows a PCA plot with the first three PCs, representing an accumulated explained variance of ca. 94.6%. As observed, different clusters are partially identified as the eight tea classes measured by MOXs. Considering that the first three components failed to achieve the recommended 95% of the accumulated explained variance [16], a total of four PCs (ca. 96.8%), were used to feed the classification models.

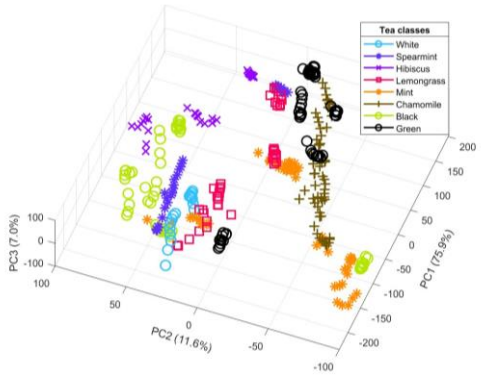
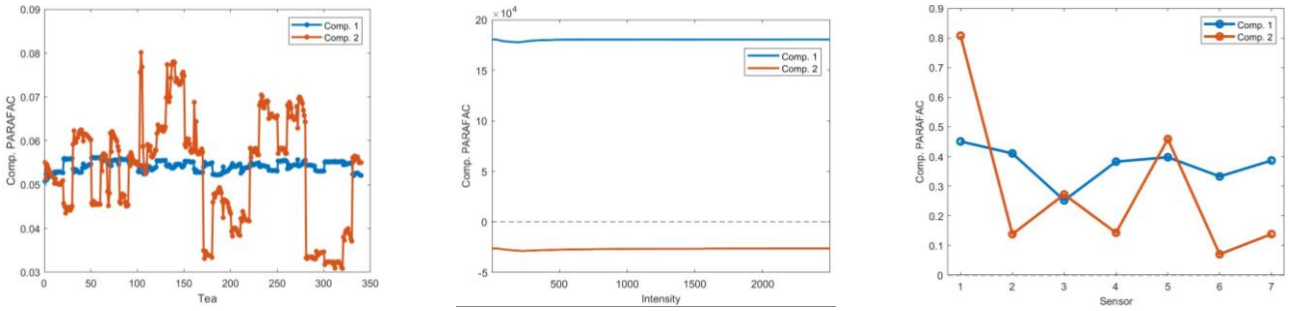


Figure 1. PCA score plot of the three first components from eight tea classes.

3.2. PARAFAC Results.

PARAFAC analysis was performed in the formed tridimensional matrix described in section 2.1. To choose the appropriate number of components, a CORCONDIA evaluation was done achieving a core consistency value of 99.2%. This result was close to the 100% described in the literature [18]. Figure 2 shows the PARAFAC components of each loading

matrix. The tea loadings represent tea variability; the intensity matrix shows changes in voltage values; finally, the sensor loadings describe the responses of each sensor. As can be seen, in order of importance, the obtained loading can be listed as follows: 1. Loadings for sensor, 2. Loadings for tea and 3. Loadings for intensities. The loadings for the sensor have the highest values and indicate which of the sensors contributes the most to detect the teas. Sensors 1 and 5 (MQ-7 and MQ-9) for component 2 provide the most, according to Figure 2c. Then, the loadings for tea indicate which teas are dominant in each component. In this case, the teas with the highest loadings for component 2 are the most influential. Therefore, sensors 1 and 5 contribute to detect such teas. The loadings for intensities are less representative. In this case, component 1 is significantly predominant over component 2, and most of the intensities contribute equally.



(**a**) (**b**) (**c**)

Figure 2. PARAFAC results loadings for (a) tea, (b) intensities, and (c) sensor.

PCA and PARAFAC data features were employed to feed pattern recognition models. Either the ANN architectures or the parameters of k-NN were selected from initial proposals and tuned by an iterative process. Optimized models were verified through the 10 k-fold cross-validation technique.

Final ANN architectures were composed of 6 layers (4 x 35 x 40 x 14 x 5 x 1) using PCs and (2 x 30 x 40 x 20 x 5 x 1) considering PARAFAC components. The first layer corresponded to input data, and the last to tea classes. Both architectures employed *logsig*, *tansig*, *logsig*, *logsig*, *tansig*, and *purelin* activations functions, respectively. Both models were adjusted by applying a resilient backpropagation training algorithm and defining proper learning rate and error values. In this way, such values for PCA-ANN were 0.002 and 0.02, while for PARAFAC–ANN were 0.09 and 0.09, respectively.

A 10-fold cross-validation procedure was performed to validate the classification capability of the models. Each k-fold uses the same fit criterion, where the weights and biases were initialized randomly before the training process. Each result was saved and averaged, and class metrics were calculated. Table 1 shows the mean confusion matrix and performance metrics results for PCA–ANN, and PARAFAC–ANN.

Table 1. Confusion matrix and classification rate of PCA and PARAFAC using ANN.

	Classes	**White**		**Spearmint**		**Hibiscus**		**Lemongrass**		**Mint**		**Chamomile**		**Black**		**Green**		**Class. rate**	
		FE_1*	**FE_2****	**FE_1**	**FE_2**	**FE_1**	**FE_2**	**FE_1**	**FE_2**	**FE_1**	**FE_2**	**FE_1**	**FE_2**	**FE_1**	**FE_2**	**FE_1**	**FE_2**	**FE_1**	**FE_2**
True class	**White**	**100**	**90**	0	10	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.99
	Spearmint	0	0	**90**	**86.6**	10	3.3	0	10	0	0	0	0	0	0	0	0	0.99	0.96
	Hibiscus	0	0	3.3	6.6	**86.6**	**60**	6.6	6.6	0	20	0	6.6	3.3	0	0	0	0.97	0.93
	Lemongrass	0	0	0	5	7.5	12.5	**85**	**57.5**	7.5	17.5	0	7.5	0	0	0	0	0.97	0.87
	Mint	0	0	0	0	0	2	2	20	**94**	**58**	4	16	0	4	0	0	0.98	0.85
	Chamomile	0	0	0	0	0	0	0	5	0	8.3	**93.3**	**75**	5	8.3	1.6	3.3	0.98	0.88
	Black	0	0	0	0	0	0	0	2	0	8	0	12	**100**	**74**	0	4	0.98	0.91
	Green	0	0	0	0	0	0	0	0	0	0	0	1.6	5	8.3	**95**	**90**	0.99	0.96
																	Total	0.98	0.92
	Predicted class																		

FE_1* = PCA. FE_2** = PARAFAC.

Lastly, k-NN modeling was implemented using euclidean distance and defining three neighbors to classify different types of tea. As in the case of ANN models, a k-fold cross-validation was carried out to reveal the variability of model classification against different data set sequences. Table 2 shows the mean confusion matrix and performance metrics results for PCA–k-NN and PARAFAC–k-NN.

Table 2. Confusion matrix and classification rate of PCA and PARAFAC using k-NN.

	Classes	White		Spearmint		Hibiscus		Lemongrass		Mint		Chamomile		Black		Green		Class. rate	
		FE_1*	FE_2**	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2	FE_1	FE_2
True class	White	**100**	**95**	0	0	0	0	0	5	0	0	0	0	0	0	0	0	1.00	0.99
	Spearmint	0	0	**100**	**90**	0	0	0	10	0	0	0	0	0	0	0	0	0.99	0.98
	Hibiscus	0	0	3.3	0	**90**	**73.3**	0	3.3	0	13.3	0	6.6	6.6	3.3	0	0	0.99	0.97
	Lemongrass	0	2.5	7.5	7.5	0	0	**80**	**77.5**	12.5	10	0	0	0	0	0	2.5	0.97	0.94
	Mint	0	0	2	0	0	2	2	6	**92**	**78**	0	0	4	14	0	0	0.97	0.91
	Chamomile	0	0	0	0	0	0	0	3.3	0	0	**91.6**	**86.6**	0	0	8.3	10	0.98	0.95
	Black	0	0	0	0	0	0	0	0	0	16	0	0	**100**	**84**	0	0	0.99	0.95
	Green	0	0	0	0	0	0	0	0	0	0	0	8.3	0	0	**100**	**91.6**	0.98	0.96
																	Total	0.98	0.96
		Predicted class																	

FE_1* = PCA. FE_2** = PARAFAC.

From the reported tables, it observes that the classes with the highest accuracy were white tea with 100% for PCA–ANN, 99% for PARAFAC–ANN, 100% for PCA–k-NN, and 99% for PARAFAC–k-NN; and green tea with 99% for PCA–ANN, 96% for PARAFAC–ANN, 100% for PCA–k-NN, and 96% for PARAFAC–k-NN. In comparison, black tea remained above 94% in general. The above metrics denote that the models correctly identify the different types of tea. Nevertheless, the algorithm occasionally mislabels the samples since the teas are extracted from the same plant, and more others share some VOCs. The difference between them is the manufacturing process, which provides chemical changes [17]. On the other hand, the algorithms fail to identify whether the tea is hibiscus or lemongrass, likely because they share VOCs as linalool, limonene, and hexanal, among others [18,19]. Figure 3 shows the average performance metrics for every model, exhibiting the difference between models and the achieved rates per metric. As can be observed, higher metrics were obtained using the combination PCA-ANN and PCA- k-NN; the accuracy for both pattern recognition algorithms stands above 98%, suggesting that PCA successfully extracts the most relevant information for classification in this dataset. On the other hand, the combinations with PARAFAC performed the lowest metrics, possibly because three-dimension analysis includes irrelevant information in the data rather than meaningful features, as shown in Figure 2b.

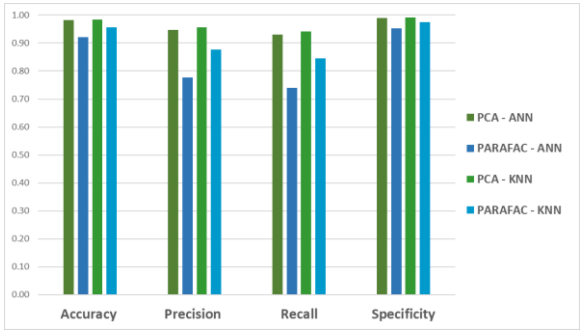


Figure 3. Metrics comparison of all the classification models.

4. Conclusions

In this work, two feature extraction methods, PCA and PARAFAC, and two classification techniques, ANN and k-NN, were compared to provide information about the processing techniques to enhance the classification accuracy of the e-nose in predicting the eight types of tea and infusions instead off the conventional works where they focus on

one type of tea (black or green tea). The obtained classification results show that feature extraction by PCA has superior metrics than using PARAFAC components; this was corroborated by the PCA-ANN combination that achieved the most remarkable accuracy. This fact is mainly related to particularities in the dataset, due to PARAFAC finding that one of the dimensions of the information (intensities) is not representative. Therefore, using only the variance as the main feature allows a better data evaluation. Although both techniques focused on discrimination tasks related to qualitative analyses, current results motivate the study of quantitative analyses of chemical species, considering the content of VOCs in tea. In this way, e-noses could be sensitive to the mixture of VOCs per tea, allowing their possible quantification from acquired MOXs signals. Combining an appropriate sensor array and a processing system, makes the e-noses competitive systems to evaluate and identify food products instead of the standard methods. The shift toward advanced sensory technologies underlines not only the importance of this research line, but also its potential for improving both qualitative and, in future works, a quantitative evaluation in the field of tea and herbal infusions classification.

References

1. Ponce, M.d.V.; Cina, M.; López, C.; Cerutti, S. Polyurethane Foam as a Novel Material for Ochratoxin A Removal in Tea and Herbal Infusions–A Quantitative Approach. *Foods* **2023**, *12*, 1828, doi: 10.3390/foods12091828.
2. Tang, G.Y.; Meng, X.; Gan, R.Y.; Zhao, C.N.; Liu, Q.; Feng, Y.B.; Li, S.; Wei, X.L.; Atanasov, A.G.; Corke, H.; et al. Health Functions and Related Molecular Mechanisms of Tea Components: An Update Review. *Int. J. Mol. Sci.* **2019**, *20*, 6196, doi:10.3390/ijms20246196.
3. Liu, Y.; Guo, C.; Zang, E.; Shi, R.; Liu, Q.; Zhang, M.; Zhang, K.; Li, M. Review on herbal tea as a functional food: classification, active compounds, biological activity, and industrial status. *J. Future Foods* **2023**, *3*, 206-219, doi:10.1016/j.jfutfo.2023.02.002.
4. Chen, W.; Hu, D.; Miao, A.; Qiu, G.; Qiao, X.; Xia, H.; Ma, C. Understanding the aroma diversity of Dancong tea (Camellia sinensis) from the floral and honey odors: Relationship between volatile compounds and sensory characteristics by chemometrics. *Food Control* **2022**, *140*, 109103, doi:10.1016/j.foodcont.2022.109103.
5. Yousefbeyk, F.; Ebrahimi-Najafabadi, H.; Dabirian, S.; Salimi, S.; Baniardalani, F.; Azmian Moghadam, F.; Ghasemi, S. Phytochemical Analysis and Antioxidant Activity of Eight Cultivars of Tea (Camellia sinensis) and Rapid Discrimination with FTIR Spectroscopy and Pattern Recognition Techniques. *Pharm. Sci.* **2023**, *29*, 100-110, doi:10.34172/ps.2022.27.
6. Persaud, K.; Dodd, G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature* **1982**, *299*, 352-355, doi:10.1038/299352a0.
7. Covington, J.A.; Marco, S.; Persaud, K.C.; Schiffman, S.S.; Nagle, H.T. Artificial Olfaction in the 21st Century. *IEEE Sens. J.* **2021**, *21*, 12969-12990, doi:10.1109/JSEN.2021.3076412.
8. Padilla, M.; Montoliu, I.; Pardo, A.; Perera, A.; Marco, S. Feature extraction on three way enose signals. *Sens. Actuators B: Chem.* **2006**, *116*, 145-150, doi:10.1016/j.snb.2006.03.011.
9. Valdez, L.F.; Gutiérrez, J.M. Chocolate Classification by an Electronic Nose with Pressure Controlled Generated Stimulation. *Sensors* **2016**, *16*, 1745, doi:10.3390/s16101745.
10. Jolliffe, I.T. Introduction. In *Principal Component Analysis*, 2 ed.; Springer New York, USA, 2006; pp. 1-6.
11. Acar, E.; Yener, B. Unsupervised Multiway Data Analysis: A Literature Survey. *IEEE Trans. Knowl Data Eng.* **2009**, *21*, 6-20, doi:10.1109/TKDE.2008.112.
12. Bro, R.; Kiers, H.A.L. A new efficient method for determining the number of components in PARAFAC models. *J. Chemom.* **2003**, *17*, 274-286, doi:10.1002/cem.801.
13. Mahmood, L.; Ghommem, M.; Bahroun, Z. Smart Gas Sensors: Materials, Technologies, Practical Applications, and Use of Machine Learning – A Review. *J. Appl. Comput. Mech-* **2023**, *9*, 775-803, doi:10.22055/jacm.2023.41985.3851.
14. Kaushal, S.; Nayi, P.; Rahadian, D.; Chen, H.-H. Applications of Electronic Nose Coupled with Statistical and Intelligent Pattern Recognition Techniques for Monitoring Tea Quality: A Review. *Agriculture* **2022**, *12*, 1359 doi:10.3390/agriculture12091359.
15. Hossin, M.; M.N, S. A Review on Evaluation Metrics for Data Classification Evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 01-11, doi:10.5121/ijdkp.2015.5201.
16. Li, B.; Gu, Y. A Machine Learning Method for the Quality Detection of Base Liquor and Commercial Liquor Using Multidimensional Signals from an Electronic Nose. *Foods* **2023**, *12*, 1508, doi:10.3390/foods12071508.
17. Wong, M.; Sirisena, S.; Ng, K. Phytochemical profile of differently processed tea: A review. *J. Food Sci.* **2022**, *87*, 1925-1942, doi:https://doi.org/10.1111/1750-3841.16137.
18. Jirovetz, L.; Jaeger, W.; Remberg, G.; Espinosa-Gonzalez, J.; Morales, R.; Woidich, A.; Nikiforov, A. Analysis of the volatiles in the seed oil of Hibiscus sabdariffa (Malvaceae) by means of GC-MS and GC-FTIR. *J. Agric. Food Chem.* **1992**, *40*, 1186-1187, doi:10.1021/jf00019a021.
19. Skaria, B.P.; Joy, P.P.; Mathew, S.; Mathew, G.; Joseph, A.; Joseph, R. Major sources of aromatic oils. In *Aromatic Plants,* 1st ed.; New India Publishing Agency: New Delhi, India, 2007; pp. 100-109.