International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Extraction of Features from the GC-MS Chromatogram Unstructured Data using Multi-Class and Multi-Label Classification for the Injection of Preprocessed Dataset into Machine Learning Algorithms applicable for E-Nose

Sasedharen Chinnathambi[1], Gopinath Ganapathy[2]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: Gas Chromatography-Mass Spectrometry (GC-MS) is a powerful tool for analyzing complex chemical mixtures, particularly for characterizing chemical compositions. Our paper examines the chemical compositions of Indian Jasminum Sambac, Rosa Damascena, and Human Urine using GC-MS analysis. In the realm of Electronic Noses (E-Noses), which mimic the olfactory capabilities of living organisms, GC-MS data provide a valuable source of chemical information. However, the raw data generated by GC-MS can be complex and unstructured, posing challenges for effective integration with machine learning (ML) algorithms in E-Nose applications. This research focuses on crucial aspects of feature extraction, multi-class and multi-label classification, and proposes a machine learning algorithm for characterizing chemical compounds and their influence on odor classification. Exploratory Data Analysis (EDA) techniques are used to select important variables and explore the potential for discrimination. Linear interpolation enhances the integration of GC-MS data into ML algorithms for E-Nose applications. This research aims to leverage advanced machine learning techniques, specifically employing multi-output classifiers with various base classifiers (e.g., Random Forest, Decision Tree), for multi-level compound classification in Gas Chromatography-Mass Spectrometry (GC-MS) datasets associated with jasmine, rose, and urine extracts. This work paves the way for automated and efficient compound recognition in complex aromatic profiles.

Keywords: *Machine Learning, Random Forest, E-Nose, Feature Extraction, Support Vector Machine, Decision Tree*

1. Introduction

GC-MS stands out as an ideal tool for detecting unknown substances or contaminants, including trace elements, in samples. It facilitates the identification of chemical compounds and the quantitative analysis of floral extracts and human urine. GC-MS is particularly effective in measuring numerous organic pollutants present in complex food and environmental samples. Long-standing challenges in end-to-end smell communication include the intricate nature of smell, the inherent unpredictability of airflows, and the complexities of managing timing and intensity. For generations, researchers across diverse disciplines have sought a classification system that defines a cognitive space and facilitates objective discussions about odors. A significant gap exists in the availability of tools that can accurately compare and characterize odors while predicting their degrees of similarity. This gap, compounded by the intricate nature of existing designs and limited accessibility, has hindered progress in the field of Digital Smell Technology [1].

Our previous research explored the top-down approach to digital smell technology, aiming to capture, classify, transmit, and reproduce smell over the internet [2]. It utilized Solid Phase Extraction (SPE) and natural drying methods for analyzing samples with GC-MS. The analysis determined the molecular mass, parts per million, and peak area of the chemical compounds in the flowers. We used a SHIMADZU's GCMS-QP2010 SE instrument with a capillary column and helium as the carrier gas. The injector temperature was set to 200°C, and the oven temperature increased from 40°C to 200°C.

Our current study focuses on using machine learning algorithms to identify and classify samples based on chromatogram data. The goal is to build a model that accurately classifies samples using peak intensity information. However, correlations between chromatogram peaks can compromise its accuracy, leading to larger gaps between known and unknown peaks. Therefore, preprocessing the data is usually necessary after creating a data table containing peak information for machine learning applications with chromatogram data.

In machine learning, "generalization" refers to an algorithm's ability to perform well on diverse inputs. Ensuring an algorithm generalizes effectively is crucial during model development and performance assessment. This study evaluates the suitability of Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) algorithms for analyzing GC-MS datasets. It aims to

[1] *Sasedharen Chinnathambi, Research Scholar, School of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli, TamilNadu, India. sasedharentc@gmail.com*

[2] *Gopinath Ganapathy, Professor, Department of Computer Science, Bharathidasan University, Tiruchirappalli, TamilNadu, India. gganapathy@gmail.com*

classify and evaluate the quality of chemical compositions, offering a high-throughput alternative to traditional, time-consuming methods for compound identification and assessment.

Preprocessing aims to: (1) reduce the number of unknown features and (2) address missing values. Additionally, our research investigates the impact of chromatogram peak correlations, feature selection (narrowing down known features), and their impact on model performance. To achieve these objectives, this research employs linear interpolation techniques to interpolate between discrete data points in the GC-MS chromatography profiles, enabling a more comprehensive analysis. The raw GC-MS data is preprocessed using linear interpolation to transform it into a more structured and informative dataset. The interpolated data is then fed into machine learning algorithms to accurately identify and classify odors and volatile compounds.

The key objectives of this research are:

1. Streamline the integration of GC-MS data with machine learning algorithms for E-Nose applications.
2. Connect discrete data points in the GC-MS chromatography profiles.
3. Enhance the detection and classification performance of E-Noses.
4. Provide a framework for leveraging GC-MS data in real-time or near-real-time odor recognition systems.

The effectiveness of the proposed method is demonstrated through extensive experiments and comparisons with existing techniques, highlighting its potential to advance the field of E-Nose technology.

2. Existing Studies

Our prior research [1] aimed to use gas chromatography-mass spectrometry (GC-MS) data, exploratory data analysis techniques, and machine learning (ML) to distinguish and classify three types of samples: floral extracts, human urine, and human breath. By integrating modern classification algorithms, this system can categorize future samples by recognizing established patterns. Solid Phase Extraction for Chemical Characterization, machine learning, and evolutionary computational methods [1] are used to enhance visualization, reduce dimensionality, and highlight the impact of each variable on the final outcome.

This study [1] highlights a gap in existing research regarding the chemical profiles of Indian Jasmine, Rose Flowers, and Human Urine, and their parts per million (ppm), for odor classification using an analytical approach with an E-nose. The authors stress the need for further exploration, highlighting the limitations of prior research in understanding the chemical composition of these flowers and its relevance to odor classification. With further refinement and dataset expansion, this approach holds promise for sensory laboratories in aroma analysis.

Kexin Bi et al. [3] present a new method for evaluating food flavors using machine learning models and GC-MS data analysis. This method involves creating distinctive "fingerprint templates" from GC-MS data, generating individual sample "fingerprint images," and using machine learning, specifically CNN, to predict olfactometry results. Their research includes a case study on peanut oil, demonstrating a model accuracy of around 93%. The paper mentions the potential for structure optimization and dataset expansion but doesn't provide specific details or strategies.

Xiaqiong Fan et al. [4] propose a study that introduces DeepResolution2 for GC-MS data analysis. This method uses deep neural networks to segment the data profile, estimate the number of components in each segment, and predict the elution region for each component. These models enhance the multivariate curve resolution process. The proposed approach still has room for improvement, especially when dealing with peak saturation.

Fawzan Sigma Aurum et al. [5] introduced a method focusing on the fragrance of coffee, which is affected by volatile compounds (VCs). This method investigates the use of untargeted SPME-GC/MS to generate a VC fingerprint capable of predicting the origin of Indonesian coffee. Multiple machine learning (ML) models were compared to establish the most accurate origin prediction. Random Forest (RF) and Partial Least Squares Discriminant Analysis (PLS-DA) models demonstrated high accuracies, reaching 97% and 95.2%, respectively. However, the non-targeted GC/MS metabolite profiling may not capture all metabolites, potentially missing important compounds. The study also lacked sensory evaluation or cupping scores to correlate the metabolite profiles with coffee quality.

Nico Borgsmüller et al. [6] presented a machine learning approach for classifying data from GC-MS instruments. This approach holds promise for refining compound identification and characterization, providing insights for diverse applications. However, models developed on specific datasets might not generalize well to new or unseen data, limiting their practical applicability.

Kristian Pastor et al. [7] introduce a method for categorizing gluten and non-gluten cereal flours according to their botanical sources. This method generated distinct patterns for each flour class. An automated machine learning framework was used for classification, with a basic logistic classifier emerging as the most recommended choice after 10-fold cross-validation. The developed model achieved an 85.71% accuracy. The paper doesn't discuss the potential impact of variations in flour composition or processing methods on the model's accuracy.

Sastia Prama Putri et al. [8] analyzed 64 compounds in 16 green and roasted coffee beans from various Indonesian species and regions. Using Principal Component Analysis (PCA), the study revealed distinct separations among samples based on roasting methods and species. This analysis was done using gas chromatography/mass spectrometry (GC/MS), which may not capture all metabolites.

Kristian Pastor et al. [9] propose a method using GC/MS and chemometric analysis to differentiate flour samples from different wheat, hazelnut, and walnut genotypes. They apply unsupervised techniques like PCA, heat mapping, HCA, and PCoA to identify crucial factors for distinguishing flour origins. The SVM classification exhibited high performance. However, the study doesn't consider the potential impact of other factors like environmental conditions.

Kristian Pastor et al. propose a novel, unified method utilizing chemometric analysis of GC-MS data to differentiate experimental flour samples based on their botanical origins [10]. This approach addresses a previously unexplored gap, potentially facilitating the identification and tracking of cereal and pseudocereal varieties. The study demonstrates the use of pattern recognition tools like cluster analysis and principal component analysis to reveal patterns and distinctions within the samples. However, the potential impact of environmental factors on the results is not addressed, despite samples originating from the same experimental field.

3. Methodology

3.1. GC-MS Compound Analysis

In Gas Chromatography-Mass Spectrometry (GC-MS) analysis, retention time (R.Time), peak area, peak height, and base peak provide valuable information about a compound's characteristics. However, these parameters are not universal for the same compound across different samples, such as Jasmine, Rose, and Human Urine. Variations can occur due to changes in column temperature, pressure, or other experimental conditions. Consequently, the exact R.Time for a compound may differ between various GC-MS runs or instruments.

Peak area and peak height relate to a compound's concentration in the sample. These values depend on factors such as the compound's concentration itself, detector sensitivity, and sample preparation methods. For example, our previous research utilized Solid-Phase Extraction and Natural Drying (Air-Drying). The same compound in different samples or under different experimental conditions can exhibit varying peak areas and heights. Additionally, a single compound may exhibit different base peaks in different samples.

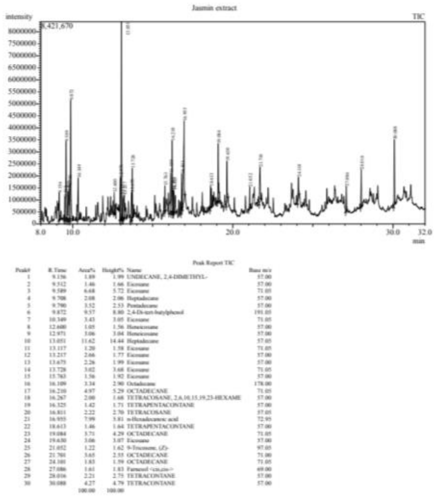
3.2. Limitations of GC-MS Sample

It's important to note that the GC-MS report provided is specific to our Indian Jasminum Sambac, Rosa Damascena, and Human Urine samples (Figures 1, 2, and 3). While retention time (R.Time), peak area, peak height, and base peak offer valuable insights for compound identification, they are not universally consistent for the same compound across different samples or GC-MS runs. Each time we run samples, even the same sample, through GC-MS, these values can vary due to factors like temperature and other spectral conditions. This limitation is a key focus of our research. We aim to use machine learning algorithms to address this inconsistency, ensuring reliable outputs from GC-MS data across different runs. By consolidating and structuring the data, we hope to improve the accuracy and consistency of chemical compound identification.

3.3. The Actual GC-MS Data and Adopted Methodology (Merging Datasets)

Our initial dataset (Figures 1, 2, and 3) poses challenges for applying multi-class and multi-label classification algorithms to chemical compound identification. Key limitations include:

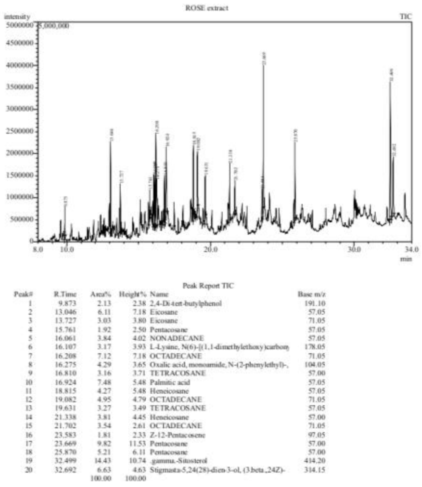
- Limited Sample Size: With only 30, 20, and 16 records respectively, the dataset is too small for robust machine learning.
- Imbalanced Classes: The non-uniform distribution of classes, with some having significantly fewer samples, creates bias and hinders accurate classification.
- Unstructured Nature: The lack of predefined patterns or organization further complicates meaningful information extraction.
- Dataset Specificity: Combining datasets from different samples without carefully considering the variability in compound identification parameters can introduce inaccuracies.



Peak Report TIC

Peak#	R.Time	Area%	Height%	Name	Base m/z
1	9.156	1.89	1.99	UNDECANE, 2,4-DIMETHYL-	57.00
2	9.512	1.46	1.66	Eicosane	57.00
3	9.589	6.68	5.72	Eicosane	71.05
4	9.708	2.08	2.06	Heptadecane	57.00
5	9.790	3.52	2.53	Pentadecane	57.00
6	9.872	9.57	8.80	2,4-Di-tert-butylphenol	191.05
7	10.349	3.43	3.05	Eicosane	71.05
8	12.600	1.05	1.56	Heneicosane	57.00
9	12.971	3.06	3.04	Heneicosane	57.00
10	13.051	11.62	14.44	Heptadecane	57.05
11	13.117	1.20	1.58	Eicosane	71.05
12	13.217	2.66	1.77	Eicosane	57.00
13	13.675	2.26	1.99	Eicosane	57.00
14	13.728	3.02	3.68	Eicosane	71.05
15	15.763	1.56	1.92	Eicosane	57.00
16	16.109	3.34	2.90	Octadecane	178.00
17	16.210	4.97	5.29	OCTADECANE	71.05
18	16.267	2.00	1.68	TETRACOSANE, 2,6,10,15,19,23-HEXAME	57.00
19	16.325	1.42	1.71	TETRAPENTACONTANE	57.00
20	16.811	2.22	2.70	TETRACOSANE	57.05
21	16.955	7.99	3.81	n-Hexadecanoic acid	72.95
22	18.613	1.46	1.64	TETRAPENTACONTANE	57.00
23	19.084	3.71	4.29	OCTADECANE	71.05
24	19.630	3.06	3.07	Eicosane	57.00
25	21.052	1.22	1.62	9-Tricosene, (Z)-	97.05
26	21.701	3.65	2.55	OCTADECANE	71.00
27	24.101	1.83	1.59	OCTADECANE	71.05
28	27.086	1.61	1.83	Farnesol <cis,cis->	69.00
29	28.016	2.21	2.75	TETRACONTANE	57.00
30	30.088	4.27	4.79	TETRACONTANE	57.00
		100.00	100.00		

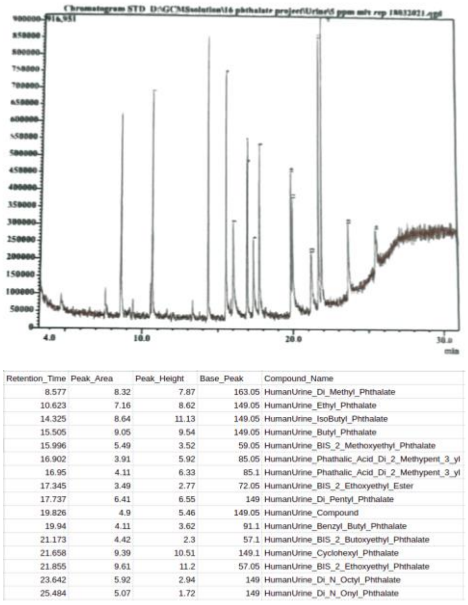
Fig. 1 – GC-MS Report – Jasmine Extract



Peak Report TIC

Peak#	R.Time	Area%	Height%	Name	Base m/z
1	9.873	2.13	2.38	2,4-Di-tert-butylphenol	191.10
2	13.046	6.11	7.18	Eicosane	57.05
3	13.727	3.03	3.80	Eicosane	71.05
4	15.761	1.92	2.50	Pentacosane	57.05
5	16.061	3.84	4.02	NONADECANE	57.05
6	16.107	3.17	3.93	L-Lysine, N(6)-[(1,1-dimethylethoxy)carbony	178.05
7	16.208	7.12	7.18	OCTADECANE	71.05
8	16.275	4.29	3.65	Oxalic acid, monoamide, N-(2-phenylethyl)-,	104.05
9	16.810	3.16	3.71	TETRACOSANE	57.00
10	16.924	7.48	5.48	Palmitic acid	57.05
11	18.815	4.27	5.48	Heneicosane	57.05
12	19.082	4.95	4.79	OCTADECANE	71.05
13	19.631	3.27	3.49	TETRACOSANE	57.05
14	21.338	3.81	4.45	Heneicosane	57.00
15	21.702	3.54	2.61	OCTADECANE	71.05
16	23.583	1.81	2.33	Z-12-Pentacosene	97.05
17	23.669	9.82	11.53	Pentacosane	57.00
18	25.870	5.21	6.11	Pentacosane	57.00
19	32.499	14.43	10.74	.gamma.-Sitosterol	414.20
20	32.692	6.63	4.63	Stigmasta-5,24(28)-dien-3-ol, (3.beta.,24Z)-	314.15
		100.00	100.00		

Fig. 2 – GC-MS Report – Rose Extract



Retention_Time	Peak_Area	Peak_Height	Base_Peak	Compound_Name
8.577	8.32	7.87	163.05	HumanUrine_Di_Methyl_Phthalate
10.623	7.16	8.62	149.05	HumanUrine_Ethyl_Phthalate
14.325	8.64	11.13	149.05	HumanUrine_IsoButyl_Phthalate
15.505	9.05	9.54	149.05	HumanUrine_Butyl_Phthalate
15.996	5.49	3.52	59.05	HumanUrine_BIS_2_Methoxyethyl_Phthalate
16.902	3.91	5.92	85.05	HumanUrine_Phathalic_Acid_Di_2_Methypent_3_yl
16.95	4.11	6.33	85.1	HumanUrine_Phathalic_Acid_Di_2_Methypent_3_yl
17.345	3.49	2.77	72.05	HumanUrine_BIS_2_Ethoxyethyl_Ester
17.737	6.41	6.55	149	HumanUrine_Di_Pentyl_Phthalate
19.826	4.9	5.46	149.05	HumanUrine_Compound
19.94	4.11	3.62	91.1	HumanUrine_Benzyl_Butyl_Phthalate
21.173	4.42	2.3	57.1	HumanUrine_BIS_2_Butoxyethyl_Phthalate
21.658	9.39	10.51	149.1	HumanUrine_Cyclohexyl_Phthalate
21.855	9.61	11.2	57.05	HumanUrine_BIS_2_Ethoxyethyl_Phthalate
23.642	5.92	2.94	149	HumanUrine_Di_N_Octyl_Phthalate
25.484	5.07	1.72	149	HumanUrine_Di_N_Onyl_Phthalate

Fig. 3 – GC-MS Report – Human Urine

GC-MS data, including retention time, peak area, peak height, and base peak values, is intrinsically sample-specific. These values can vary due to the sample matrix, experimental conditions, and instrument settings. This is particularly important when dealing with different samples like jasmine and rose, where direct concatenation of datasets can lead to misleading associations between compounds and samples due to inherent variations. Before processing any GC-MS data, unstructured data must be transformed into a structured format suitable for machine learning algorithms. While adding features can be a common approach to increase accuracy, it's crucial to carefully consider the added features and their relevance to the specific problem. For clustering similar features, feature engineering might be more appropriate than simply adding them all. Additionally, standardization or normalization of features should be considered to address potential biases arising from variables with larger scales.

3.4. Exploratory Data Analysis (EDA)

Our initial exploration of the GC-MS data involved various descriptive statistics (Mean, Median, Standard Deviation, etc.) to understand the data distribution and variability. These statistics revealed minimal deviations, with all values within a 10% range. Additionally, we employed data visualization techniques such as:

- Bar charts: To illustrate the distribution of different features within the dataset.
- Histograms: To visually compare the distribution of individual features.
- Density plots: To visualize the probability density of features, particularly focusing on symmetry for Peak Area, Peak Height, and Base Peak.
- Box plots: To identify potential outliers in Peak Area, Peak Height, and Base Peak during feature engineering. Notably, box plots ensured values remained within the expected range (9.00 - 30.00 mins) based on the GC-MS chromatogram peaks.

Furthermore, correlation analysis provided insights into potential relationships between variables. This analysis can be helpful in understanding how adding or removing features might impact the model's predictive power and accuracy.

3.5. The Rationale for Interpolation and Deviation in Datasets

The EDA in our process provided insights into the data, aiding in understanding its structure and the distribution of variables. Initially, our raw merged dataset contained 66 records. Since the dataset is unstructured and skewed, the entire dataset has been scaled using linear interpolation for multi-class and multi-label classification. When dealing with unstructured data, it is not advisable to directly process a raw dataset; instead, interpolation should be employed. Linear interpolation on three columns (Peak_Area, Peak_Height, and Base_Peak) based on the 'Retention_Time' index. Given two points (x_1, y_1) and (x_2, y_2), the linearly interpolated value y at a point x between x_1 and x_2 is calculated as:

$$y = y_1 + ((x-x_1)*(y_2-y_1)) / (x_2-x_1) \qquad (1)$$

Here, linear interpolation on specified columns based on the 'Retention_Time' index, filling in missing values by estimating intermediate values through a linear equation between adjacent data points. The initial computation of the Mean for Peak_Area, Peak_Height, and Base_Peak from a raw sample was 4.5, 4.5, and 93.3, respectively. After Linear Interpolation, an imbalance emerged in the dataset. To address this, we further scaled the dataset values between 9.00 Mins to 30 Mins based on Peaks in the GC-MS chromatogram. This adjustment aimed to handle gaps within the merged interpolated dataset. Expanding the original dataset from 66 records to 1458 records through interpolation resulted in mean deviations to 4.7, 4.8, 103.9, respectively. Despite linear interpolation, the data exhibited bias, leading to oversampling (5568 Records) of features to eliminate dataset deviations. Label Encoding was performed for non-numerical variables before oversampling to simplify complexities.

Visualization of the datasets distribution revealed differences in the 'Density_Plot' of Peak_Area, Peak_Height, and Base_m/z, indicating a deviation from the expected pattern(Figure 4). Similar challenges were encountered during Downsampling. With more oversampling, values approached those of the raw dataset. Although there isn't a fixed rule, the general aim was to keep deviations 5% or less. To solve this problem, 2nd and 3rd oversampling was done until the no deviations of the oversample data is observed. Following a third oversampling of the interpolated dataset (11136 Records), the mean was restored to 4.7, 4.8, 96.4, representing almost less than 10% deviation and closely resembling the original raw dataset. The density graph also exhibited striking similarities, with minimal deviations in the mean among the raw, interpolated, and oversampled datasets.

4. Classification with Machine Learning

Complex datasets are primarily explored using multi-class and multi-label classification techniques like Decision Trees(DT), Support Vector Machine(SVM), Random Forest(RF), Gradient Boosting, and more. These techniques help identify patterns, enabling visualization and interpretation of complex multivariate datasets[11]. They aim to capture as much variance from the original data as possible, facilitating the identification of major patterns and potential causal factors. Machine Learning algorithms, specifically Random Forests(RF), offer significant advantages in analyzing extensive data arrays, particularly when the number of variables exceeds the sample count. RF functions as a bootstrapping classification tool, creating multiple decision trees that each use randomly selected input variables for predictions or class allocation. This approach effectively combats over fitting and does not require scaling before analysis[12]. Due to Random Forest superior performance, the algorithm has been applied across various domains, ranging from gene selection studies for diagnostic purposes[13], monitoring grasslands[14], classification of milk based on animal species[15], intra-regional classification of grape seeds[16], to predict liquid chromatographic retention times[17].

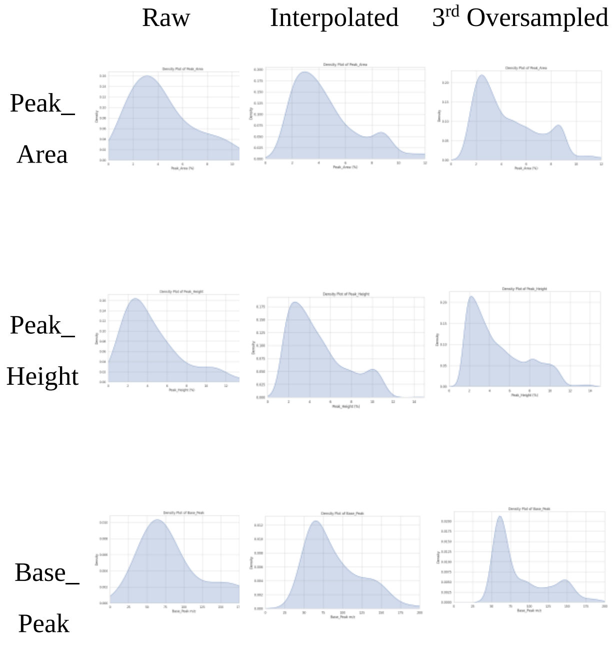


Fig. 4 – Visualization of Data, Density Plot after Third Oversampling

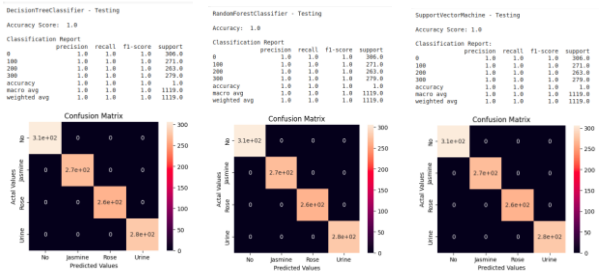
4.1. Testing Phase

4.1.1. Multi-Class Classification

During the testing phase, the Decision Tree Classifier, Random Forest, and Support Vector Machine were utilized to predict outcomes in a test set. The algorithm was trained using features and hyper-parameters were fine-tuned for handling individual samples, focusing on parameters like Regularization (specifically L2 regularization), Penalty type, Cross-validation strategy (using KFold with n_splits=5), and Scoring metrics. Notably, when validated with Merged Samples, the Decision Tree Classifier, RF, and SVM exhibited 100% accuracy[Figure. 5].

4.1.2. Multi-Label Classification

GC-MS reports may exhibit an imbalance in labels, making it crucial to address this imbalance to prevent model biases. The optimal algorithm for multi-label classification hinges on the dataset's specific characteristics and the desired trade-off between accuracy and computational efficiency. However, some algorithms tend to perform better across a broader range of datasets. In multi-label classification, unlike multi-class and multi-output classification, we predict multiple output labels, assigning as many labels as applicable to the input data (e.g., jasmine, rose, and human urine). The system can predict anywhere from no compounds to the maximum number of available compounds. Following data preprocessing, we constructed a multi-label classifier using Scikit-Learn. Within Scikit-Learn, we utilized the MultiOutputClassifier object to train the multi-label classifier model. This model's strategy involves training one classifier per label, essentially assigning a separate classifier to each label. In this system, we employed Random Forest Classifier, and MultiOutputClassifier extended it to all labels. We can further modify the model and adjust its parameters, which are passed into the MultiOutputClassifier, to suit our needs for Support Vector Machines and Decision Trees. In a multi-label classification task, the anticipated output is a set of predicted labels for each testing sample. Evaluation metrics for multi-label classification, including Hamming Loss, F1-Score, Precision, and Recall, accurately assess the model's performance.



```
DecisionTreeClassifier - Testing

Accuracy Score:  1.0

Classification Report
                precision  recall  f1-score  support
0                     1.0     1.0       1.0    306.0
100                   1.0     1.0       1.0    271.0
200                   1.0     1.0       1.0    263.0
300                   1.0     1.0       1.0    279.0
accuracy              1.0     1.0       1.0      1.0
macro avg             1.0     1.0       1.0   1119.0
weighted avg          1.0     1.0       1.0   1119.0
```

```
RandomForestClassifier - Testing

Accuracy:  1.0

Classification Report
                precision  recall  f1-score  support
0                     1.0     1.0       1.0    306.0
100                   1.0     1.0       1.0    271.0
200                   1.0     1.0       1.0    263.0
300                   1.0     1.0       1.0    279.0
accuracy              1.0     1.0       1.0      1.0
macro avg             1.0     1.0       1.0   1119.0
weighted avg          1.0     1.0       1.0   1119.0
```

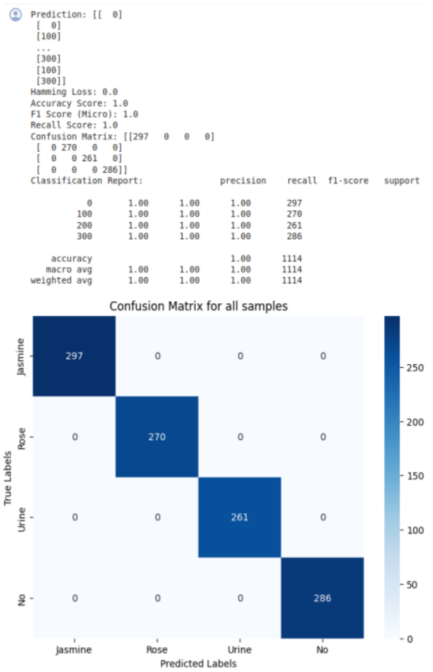
```
SupportVectorMachine - Testing

Accuracy Score: 1.0

Classification Report:
                precision  recall  f1-score  support
0                     1.0     1.0       1.0    306.0
100                   1.0     1.0       1.0    271.0
200                   1.0     1.0       1.0    263.0
300                   1.0     1.0       1.0    279.0
accuracy              1.0     1.0       1.0      1.0
macro avg             1.0     1.0       1.0   1119.0
weighted avg          1.0     1.0       1.0   1119.0
```

Fig. 5 – Testing Phase – DT, RF, SVM Algorithms

The Hamming loss metric evaluates the accuracy of a multi-label classifier, particularly in scenarios where each sample can belong to multiple classes or have multiple labels. It measures the proportion of incorrectly predicted labels for a given sample. However, achieving a perfect Hamming loss in practical scenarios with complex and unstructured datasets can be extremely challenging. In our case, Multi-Label Decision Trees, SVM, and Random Forests resulted in a Hamming Loss of '0' [Figure 6,7]. This indicates perfect performance, where all predicted labels for each sample precisely match the true labels. Similarly, the Accuracy score, F1, and Recall score results of '1' demonstrate that the model predicts the exact label combination [Figure 6,7].

4.2. Validation Phase

Transitioning to the validation phase, the RF model was applied to the merged dataset, optimizing it by evaluating preprocessed encoded data and determining the number of features used in the raw dataset. The model's performance was assessed through a cross-validation process using the sklearn library. The entire dataset was divided into training (70%), testing (20.1%), and validation (9.9%) subsets for both A and B, resulting in 3897, 1119, and 552 samples, respectively. Differences between raw and processed data (Linear Interpolated, Encoded, and Oversampled) were examined. Despite the RF parameters remaining at default settings, the RF model achieved 100% accuracy in Multi-Class classifying features as Jasmine, Rose, Human_urine, and No_Compound data (Figure 8).



```
Prediction: [[  0]
 [  0]
 [100]
 ...
 [300]
 [100]
 [300]]
Hamming Loss: 0.0
Accuracy Score: 1.0
F1 Score (Micro): 1.0
Recall Score: 1.0
Confusion Matrix: [[297   0   0   0]
 [  0 270   0   0]
 [  0   0 261   0]
 [  0   0   0 286]]
Classification Report:               precision    recall  f1-score   support

           0       1.00      1.00      1.00       297
         100       1.00      1.00      1.00       270
         200       1.00      1.00      1.00       261
         300       1.00      1.00      1.00       286

    accuracy                           1.00      1114
   macro avg       1.00      1.00      1.00      1114
weighted avg       1.00      1.00      1.00      1114
```

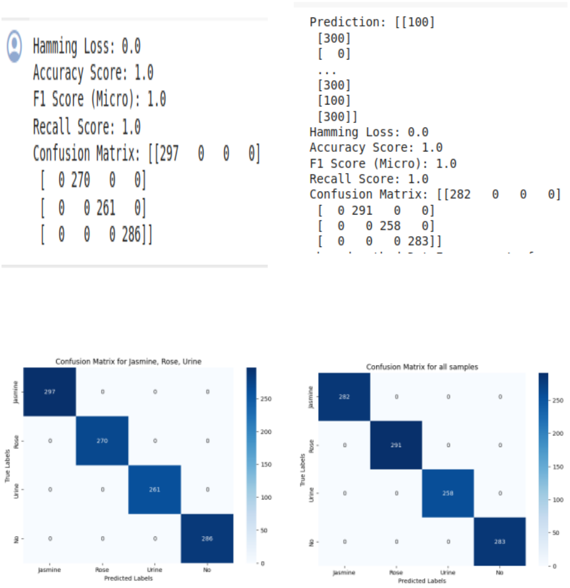
Fig. 6 – Testing Phase – Multi-Label Random Forest Classifier

Similar like testing phase, we employed Random Forest, and MultiOutputClassifier extended it to all labels. We can further modify the model and adjust its parameters, which are passed into the MultiOutputClassifier, to suit our needs for Support Vector Machine and Decision Trees. The evaluation metrics of Hamming Loss, F1-Score, Precision, and Recall are instrumental in accurately evaluating the performance of the multi-label classification model during the validation phase(Fig. 9).

5. Results

One aspect regarding imbalanced datasets is that they don’t heavily impact ensemble techniques. Instead, fine-tuning hyper-parameters and adjusting class weights to penalize misclassifying the minority class can be beneficial. The use of appropriate techniques during model training in classification algorithms may result in better model performance and accuracy. We used samples of Jasmine, Rose, and Human Urine for Gas Chromatography testing, resulting in a GC-MS Report, which was later transformed into a structured dataset. Following this, we conducted data preprocessing and exploratory data analysis, achieving 100% accuracy across all machine learning classifier algorithms. Typically, such high accuracy is attained by one classifier, but in our case, it was achieved universally across all algorithms.

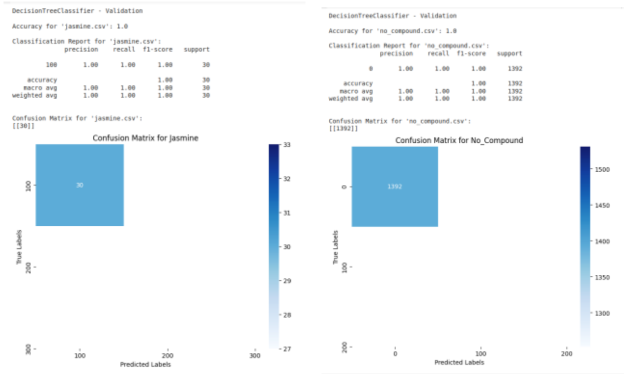
Despite this success, we opted for the Random Forest classifier as our primary choice due to its capability to determine feature importance, aiding in understanding the features' contributions to predictions. By leveraging feature engineering, encoding techniques, and comprehensive exploratory data analysis, we achieved 100% accuracy, while the Random Forest's ability to select pertinent features and reduce over fitting made it less susceptible to noise and outliers.



```
Hamming Loss: 0.0
Accuracy Score: 1.0
F1 Score (Micro): 1.0
Recall Score: 1.0
Confusion Matrix: [[297   0   0   0]
 [  0 270   0   0]
 [  0   0 261   0]
 [  0   0   0 286]]
```

```
Prediction: [[100]
 [300]
 [  0]
 ...
 [300]
 [100]
 [300]]
Hamming Loss: 0.0
Accuracy Score: 1.0
F1 Score (Micro): 1.0
Recall Score: 1.0
Confusion Matrix: [[282   0   0   0]
 [  0 291   0   0]
 [  0   0 258   0]
 [  0   0   0 283]]
```

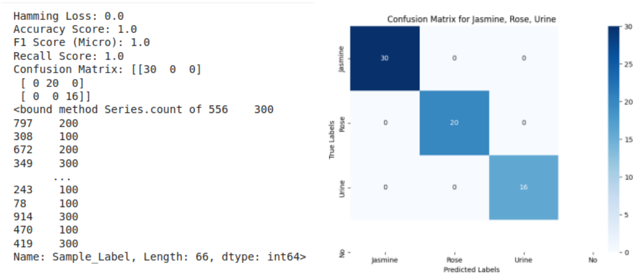
Fig. 7 – Testing Phase – Multi-Label Decision Tree and SVM



```
DecisionTreeClassifier - Validation

Accuracy for 'jasmine.csv': 1.0

Classification Report for 'jasmine.csv':
              precision    recall  f1-score   support

         100       1.00      1.00      1.00        30

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30


Confusion Matrix for 'jasmine.csv':
[[30]]
```

```
DecisionTreeClassifier - Validation

Accuracy for 'no_compound.csv': 1.0

Classification Report for 'no_compound.csv':
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1392

    accuracy                           1.00      1392
   macro avg       1.00      1.00      1.00      1392
weighted avg       1.00      1.00      1.00      1392


Confusion Matrix for 'no_compound.csv':
[[1392]]
```

Fig. 8 – Validation Phase – Multi-Class Random Forest Classifier



```
Hamming Loss: 0.0
Accuracy Score: 1.0
F1 Score (Micro): 1.0
Recall Score: 1.0
Confusion Matrix: [[30  0  0]
 [ 0 20  0]
 [ 0  0 16]]
<bound method Series.count of 556    300
797    200
308    100
672    200
349    300
      ...
243    100
78     100
914    300
470    100
419    300
Name: Sample_Label, Length: 66, dtype: int64>
```

Fig. 9 – Validation Phase, Multi-Label Random Forest Classifier

6. Conclusion

The effectiveness of EDA and other techniques may vary depending on the specific characteristics of a dataset. Both machine learning and deep learning algorithms rely on understanding the underlying data patterns. If these patterns vary, the output will also differ. This approach is crucial, especially when dealing with datasets containing significant bias, as adopting this methodology ensures dataset consistency and accurate identification of chemical compounds. Additionally, it contributes to a deeper understanding of the relationship between chemical data and odour perception, advancing the field of E-Nose technology and its potential applications in various domains. Our research overcomes all the limitations or drawbacks mentioned above, having the potential for structure optimization and dataset expansion, shows promising results for classifying odour samples based on their chemical profiles. Further it investigates the potential impact of variations in chemical composition, processing methods, environmental conditions, and other factors on the accuracy of the classification model. Understanding the relationship between chemical compounds and perceived odour is essential, as it provides insights into the aromatic profile of these culturally and economically significant flowers.

Competing Interests

I/We certify that we have No affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. I/We have no conflicts of interest to disclose. The above information is true and correct, up to our knowledge.

References

[1] Sasedharen Chinnathambi and Gopinath Ganapathy, "Qualitative Analysis of Chemical Components of Jasminum Sambac and Rosa Damascena by Gas Chromatography-Mass Spectrometry and Its Influences on E-nose to Classify Odour", *Applied Ecology and Environmental Sciences*, 2022, Vol. 10, No. 12, 766-775, DOI:10.12691/aees-10-12-10

[2] Sasedharen Chinnathambi and Gopinath Ganapathy, "A literature review of scent technology and analysis on digital smell to capture, classify, transmit and reproduce smell over internet", *Journal of Theoretical and Applied Information Technology*, 31'st May 2023. Vol.101. No 10, ISSN: 1992-8645

[3] Kexin Bi et.al., "GC-MS Fingerprints Profiling Using Machine Learning Models for Food Flavor Prediction", *MDPI*,www.mdpi.com/journal/processes, Processes 2020, 8, 23; doi:10.3390/pr8010023[4]. Xiaqiong Fan et.al., "Fully automatic resolution of untargeted GC-MS data with deep learning assistance", *Talanta*, Volume 244, 1 July 2022, 123415

[4] Fawzan Sigma Aurum et.al., “Predicting Indonesian coffee origins using untargeted SPME − GCMS - based volatile compounds fingerprinting and machine learning approaches”, *European Food Research and Technology*, 249(8), May 2023, DOI:10.1007/s00217-023-04281-2

[5] Nico Borgsmüller et.al., “Machine learning-based classification to improve Gas Chromatography-Mass spectrometry data processing”, *European RFMF Metabomeeting 2020*, Jan 2020, Toulouse, France. 263 p., 2020

[6] Kristian Pastor et.al., “Classification of Cereal Flour by Gas Chromatography – Mass Spectrometry (GC-MS) Liposoluble Fingerprints and Automated Machine Learning”, *Analytical Letters*, Taylor & Francis Online, Volume 55, 2022 - Issue 14, Pages 2220-2226, 21 Mar 2022

[7] Sastia Prama Putri et.al., “GC/MS based metabolite profiling of Indonesian specialty coffee from different species and geographical origin”, *Metabolomics,* Vol. 15, Iss: 10, pp 126, 18 Sep 2019

[8] Kristian Pastor et.al., “A rapid dicrimination of wheat, walnut and hazelnut four samples using chemometric algorithms on GC/MS data”, *Journal of Food Measurement and Characterization*, Springer Nature 2019, 10 July 2019

[9] Kristian Pastor et.al., “Discriminating cereal and pseudocereal speciesusing binary system of GC/MS data – Pattern Recognition Approach”, *Journal of The Serbian Chemical Society (National Library of Serbia)*, Vol. 83, Iss: 3, pp 317-329, 01 Apr 2018

[10] Alban Ramette, “Multivariate analyses in microbial cology”, Federation of European Microbiological Societies, *Blackwell Publishing Ltd*., FEMS Microbiol Ecol 62, 142–160, 2007

[11] L. Tedone, A. Ghiasvand and B. Paull, “Random Forests machine learning applied to gas chromatography – Mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils”, *Talanta*, doi: 10.1016/j.talanta.2019.120471, 2019

[12] Jacobs et.al., “Genetic fingerprinting of salmon louse (Lepeophtheirus salmonis) populations in the North-East Atlantic using a random forest classification approach”, Sci. Rep. 8(1) 1203. https://doi.org/10.1038/s41598-018-19323-z, 2018

[13] Melville, A. Lucieer and J. Aryal, “Object-based random forest classification of Landsat ETM+ and WorldView-2 satellite imagery for mapping lowland native grassland communities in Tasmania”, Australia, Int. J. Appl. Earth Obs. Geoinf. 66 (2018) 46-55 https://doi.org/10.1016/j.jag.2017.11.006.

[14] Amjad et.al., “Raman spectroscopy based analysis of milk using random forest classification”, Vib. Spectrosc. 99 (2018), 124-129. https://doi.org/10.1016/j.vibspec.2018.09.003.

[15] B.V. Canizo et.al., “Intra-regional classification of grape seeds produced in Mendoza province (Argentina) by multi-elemental analysis and chemometrics tools”, Food Chem. 242 (2018) 272-278. https://doi.org/10.1016/j.foodchem.2017.09.062.

[16] F. Tian, L. Yang, F. Lv and P. Zhou, “Predicting liquid chromatographic retention times of peptides from the Drosophila melanogaster proteome by machine learning approaches”, Anal. Chim. Acta 644(1-2) (2009) 10-6. https://doi.org/10.1016/j.aca.2009.04.010.