Journal of Theoretical and Applied Information Technology
30th June 2024. Vol.102. No. 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP LEARNING FOR UNTANGLING THE CHEMISTRY OF SCENT: A NOVEL APPROACH TO ODOUR CLASSIFICATION USING GC-MS DATA

SASEDHAREN CHINNATHAMBI [1], GOPINATH GANAPATHY [2]

[1] Research Scholar, School of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli, Tamilnadu, India.

[2] Professor, Department of Computer Science, Bharathidasan University, Tiruchirappalli, Tamilnadu, India.

E-mail: [1]sasedharentc@gmail.com, [2]gganapathy@gmail.com

ABSTRACT

The study evaluates the efficacy of employing a deep-learning neural network, for classifying the samples from the Gas Chromatography-Mass Spectrometry(GC-MS) chromatogram dataset. By utilizing a deep learning neural network, this paper endeavours to classify chemical compositions, facilitating high-throughput compound identification and assessment. The unstructured nature of the data, variability in compound identification parameters, and the need to consider experimental conditions pose additional challenges in accurate classification. To overcome these challenges, the research implements data preprocessing techniques such as linear interpolation to bridge gaps in chromatography profiles, thereby transforming the raw dataset into a structured and informative dataset. The objectives include streamlining the integration of GC-MS data into deep learning models, improving the detection and classification of odours, and providing a framework for real-time odour recognition systems. This study delves into the intricacies of GC-MS data analysis within the realm of olfactory classification, with particular attention to the fragrances of Jasminum Sambac, Rosa Damascena, and Human Urine, encompassing both pleasant and unpleasant scents. Through exploratory data analysis, crucial variables are identified, and a novel deep-learning neural network is proposed for characterizing chemical compounds and their impact on odour classification. By pioneering the application of supervised deep learning directly on raw GC-MS datasets, this study achieves remarkable accuracy in classifying floral and human urine samples. Linear interpolation emerges as a key technique for seamless data integration and augmentation, offering valuable insights into the aromatic profiles of culturally and economically significant flowers.

Keywords: *Deep Learning, Interpolation, Neural Network, GC-MS, Feature Extraction, Feature Classification*

1. INTRODUCTION

The rationale behind why flower samples are being compared to human urine is to analyse the chemical composition of pleasant and unpleasant smells. The two floral plants Jasminum Sambac and Rosa Damascena chemical composition were taken for the study of chemical compound characterization since these flowers have attained commercial significance in India. The choice of these flowers as the subject for chemical compound characterization carries several rationales: Jasminum Sambac and Rosa Damascena possess distinct and captivating aroma profiles characterized by a rich combination of volatile organic compounds (VOCs). Understanding the chemical composition of these aromas is essential for quality control, product development, and fragrance formulation. Analyzing the chemical composition of these extracts using advanced analytical techniques like GC-MS allows for the identification and quantification of individual aroma compounds, providing valuable insights into the complexity of their aromatic chemistry. In addition to qualitative identification, GC-MS can also provide quantitative information about the concentration of each compound present in the given extract. This quantitative data can elucidate the relative abundance of key aroma compounds and their contribution to the overall aromatic profile.

This paper has several key objectives:

i. Seamless Integration: We aim to seamlessly integrate GC-MS data with deep learning

models for efficient characterization of chemical compositions.

ii. Connecting the Dots: We address the challenge of discrete data points in GC-MS profiles by employing linear interpolation, enabling a more holistic analysis.
iii. Enhanced Odour Detection: Our approach aims to significantly improve the detection and classification performance of odours present in the samples.
iv. Real-Time Odour Recognition: We envision a future where this framework can be leveraged to develop real-time or near-real-time odour recognition systems.

This work represents a pioneering effort in applying supervised deep learning directly to raw GC-MS datasets for the classification of floral and human urine samples with exceptional accuracy. The use of linear interpolation plays a crucial role in data augmentation, significantly improving the performance of the deep learning model, especially when dealing with limited datasets. By unravelling the intricate relationship between chemical compounds and perceived odour, this research offers valuable insights into the aromatic profiles of these culturally and economically significant samples, paving the way for advancements in odour science and related fields.

2. EXISTING STUDIES

Our previous research, directed towards identifying and quantifying volatile compounds through GC-MS analysis, has the potential to support the advancement of odour classification[1], which seeks to detect and categorize various odours. The accurate measurement of parts per million (PPM) of each chemical compound is crucial for the classification of odour. This study further highlights the importance of considering factors such as sensory evaluation and processing methodology differences when analysing the effects of different identification processes on the chemical compounds of Jasminum Sambac and Rosa Damascena[2].

According to *Anne Bech Risum* and *Rasmus Bro*, the use of deep learning to evaluate peaks[3], to automate the analysis of gas chromatographic data, specifically in identifying whether each resolved component represents a peak suitable for integration. The deep network is trained to classify four different classes, shows high accuracy in classifying peaks, with few misclassification. Misclassification can occur due to the variability in peak shape, the choice of appropriate classifiers. 'Training the model on more diverse data may further improve its performance'. *Dmitriy D. Matyushin et.al.,* introduces the usage of deep learning ranking for small molecules identification using low-resolution electron ionization mass spectrometry by using a deep neural network[4] to reduce the probability of wrong answers in library search procedures. Paper contributes the deep learning ranking model outperforms other approaches, reducing the fraction of wrong answers (at rank-1) by 9-23% depending on the dataset used. The study tested the model using spectra from the Golm Metabolome Database, Human Metabolome Database, and FiehnLib, demonstrating its applicability for small molecule identification in metabolomics. 'Filtering and selecting spectra for accurate library searches is an ongoing challenge that requires further improvements'.

According to *Jesse Read* and *Fernando Perez-Cruz*, traditional multi-label classification methods often do not explicitly model dependencies between labels, leading to limited predictive performance. The authors empirically evaluate their deep network and show that it outperforms several competitive methods from the literature in multi-label classification. The study suggests that 'focusing on feature modelling'[5], rather than solely on modelling dependencies between output labels, can lead to significant improvements in multi-label classification. *Vishakha Pareek and Santanu Chaudhury,* proposes two deep learning-based architectures for gas identification and quantification. These architectures automatically *tune hyper-parameters for optimal performance* [6, 11]. The performance of traditional pattern recognition methods[7] for gas identification and quantification depends heavily on feature engineering and hyper-parameter selection, while the proposed deep learning-based methods overcome these limitations.

Mike Li et. al., introduces a deep learning model for alignment of peaks in GC-MS data[8], which is more adept at handling complex and fuzzy data sets. Testing the model on various GC-MS data sets of different complexities and analyzing the alignment results quantitatively. The model showed very good performance, with an AUC of approximately 1 for simple data sets and an AUC of approximately 0.85 for very complex data sets. Missing additional evaluation metrics such as precision, recall, or F1 score could provide a more comprehensive assessment of its alignment accuracy. According to *Xiaqiong Fan et.al.,* a fully automated approach, using tensor-based modeling[9] and deep learning assistance, to convert gas chromatography-mass spectrometry (GC-MS) data into peak tables without user interactions,

addressing the issues of software uncertainty and reproducibility. The automated approach provides improvements over current analysis methods in terms of analysis time and reproducibility. The proposed automated approach still has room for improvement, especially when the data collinearity is broken, such as in the case of peak saturation.

Yan Huang et.al., proposed a multi-task deep neural network architecture for multi-label learning, which transforms multi-label learning into multiple binary-class classification tasks. The effectiveness of using deep neural networks[10] for multi-label learning and shows the potential for improving performance in tasks such as image annotation. The multi-label classifier of MT-DNN compares the outputs of different nodes corresponding to different labels to determine the labels for the instance.

2.1 Limitations

This analysis reveals several opportunities to enhance the application of deep learning in chromatography and mass spectrometry.

Data Considerations:

- **Training Data Diversity:** Studies highlight the need for broader training datasets to improve model performance in tasks like peak classification and alignment[3]. This will lead to more robust and generalizable models.
- **Feature Engineering Focus:** Shifting focus towards feature modeling, beyond just output label dependencies, holds promise for significant improvements, especially in multi-label classification tasks[5].

Evaluation and Refinement:

- **Comprehensive Evaluation Metrics:** While existing studies report metrics like AUC, incorporating additional metrics like precision, recall, and F1 score will provide a more holistic assessment of model accuracy[8].
- **Complex Data Handling:** Addressing challenges associated with complex or non-ideal data, such as peak saturation and data collinearity[9], is crucial to ensuring the effectiveness of deep learning methods in these scenarios.

Automation and Reproducibility:

- **Enhanced Automation:** While progress has been made towards automating tasks like peak table generation, further advancements are necessary to streamline workflows, improve reproducibility, and address software uncertainty[9].

Overall, the gaps in the existing studies seem to revolve around the need for more diverse and comprehensive training data, improved evaluation metrics, better handling of complex or non-ideal data scenarios, exploration of feature modeling techniques, and further advancements in automation and reproducibility for various analytical tasks in the field of chromatography and mass spectrometry.

2.2 Problem Statement

- The research paper aims to address the challenges associated with classifying samples from Gas Chromatography-Mass Spectrometry (GC-MS) datasets, particularly focusing on Challenges include the limitations of the initial GC-MS datasets due to their size and structure, leading to biased classification results and difficulties in extracting meaningful information with the help of Deep Learning.
- The study also highlights the importance of addressing imbalanced datasets and the impact of ensemble techniques and hyper-parameter tuning on model performance

3. HARNESSING DEEP LEARNING FOR ENHANCED ODOUR IDENTIFICATION AND CLASSIFICATION

3.1 Sample Preparation and GC-MS Analysis

The analysis begins with solvent extraction using hexane to isolate aromatic compounds from the sample. Following extraction, solid-phase extraction (SPE) further purifies and concentrates the extracted compounds[2, 12]. During SPE, the target molecules are retained on a solid phase while unwanted substances are washed away. Finally, the samples are prepared for GC-MS analysis. Helium serves as the carrier gas, and the injector and oven temperatures are carefully optimized to achieve optimal separation and detection of the compounds. The gas chromatograph (GC) separates the compounds based on their volatility and affinity for the stationary phase within the column. The mass spectrometer (MS) then identifies and quantifies the separated compounds by analyzing their mass-to-charge ratio.

3.2 Understanding Variability in GC-MS Compound Analysis

While Gas Chromatography - Mass Spectrometry (GC-MS) offers a wealth of information about a compound's identity and relative

abundance, the results are not absolute and can vary depending on several factors. One source of variability is a compound's retention time (R.Time), which reflects the time it takes to travel through the GC column. Changes in column temperature, pressure, or other experimental conditions can influence retention time. Consequently, the exact retention time for the same compound may differ across GC-MS runs or instruments.

Peak area and peak height are two other parameters susceptible to variation based on the sample and experimental conditions. Peak area signifies the total amount of a compound present, while peak height represents the signal intensity at the compound's elution time. Factors like the compound's concentration, detector sensitivity, and sample preparation method can affect these parameters. For example, a compound might exhibit varying peak areas and heights in samples with different concentrations or prepared using different methods.

In some cases, a compound may even exhibit different base peaks in different samples. The base peak is the most prominent in the compound's mass spectrum. Occasionally, it can vary depending on the sample matrix or the ionization conditions used in the mass spectrometer. These variations in GC-MS results emphasize the importance of considering the experimental context and sample conditions when interpreting data. Although GC-MS is a valuable tool for compound analysis, it's crucial to recognize that the obtained results are not always absolute and require interpretation in light of the specific analysis conditions.

3.3 Limitations of GC-MS Sample Analysis

The GC-MS report generated is specific to the analyzed samples: Indian Jasmine (Jasminum Sambac), Damask Rose (Rosa Damascena), and Human Urine (Figures 1, 2, and 3). Retention time (R.Time), peak area, peak height, and base peak values obtained for these samples are unique to them. While valuable for compound identification, these parameters can exhibit inconsistencies for the same compound across different samples or GC-MS runs. This variability arises because each GC-MS run can introduce slight variations in R.Time, peak area, peak height, and base peak for a specific compound. Factors like temperature, pressure, and other experimental conditions influence this variance. To address this inconsistency, our research employs a deep learning algorithm. This algorithm consolidates data from multiple GC-MS runs of the same sample and aims to rectify inconsistencies in the generated reports. By combining and analysing data from various sources, we strive to achieve consistent output and enhance the accuracy of chemical compound identification.

3.4 Challenges of Applying Deep Learning to GC-MS Data and Our Proposed Solutions

The initial GC-MS dataset for Jasmin Sambac, Rosa Damascena, and Human Urine presents limitations for deep learning applications due to their size and structure (Figures 1, 2, and 3). With limited set of records and the imbalanced nature of the data, where some classes have significantly fewer samples, can lead to biased classification results if left unaddressed. An additional challenge lies in the unstructured nature of the data. Lacking predefined patterns or organization makes it difficult to extract meaningful information. Simply combining these datasets without considering the inherent variability in compound identification parameters can introduce errors. Retention time, peak area, peak height, and base peak values are all sample-specific in GC-MS, influenced by factors like the sample matrix, experimental conditions, and instrument settings. Concatenating the dataset without accounting for these variations can lead to incorrect associations between compounds and samples. To overcome these challenges, we have implemented the following steps:

3.4.1 feature engineering

We have enriched the merged dataset with additional features that group similar features within GC-MS chromatograms. This effectively increases the data available for analysis, potentially improving classification accuracy.

3.4.2 data structuring

The unstructured data undergoes conversion to a structured format before processing. This may involve feature engineering and standardizing or normalizing features using TensorFlow.keras for data preprocessing, then shuffled to ensure the data is not skewed towards variables with larger scales. After Normalization, the dataset gets converted into a tensorflow framework. These steps aim to enhance the quality and usability of the GC-MS data, ultimately enabling more accurate and reliable compound identification using deep learning techniques.

4. EXPLORATORY DATA ANALYSIS (EDA)

To deepen our comprehension of the GC-MS dataset, we utilized Exploratory Data Analysis (EDA) techniques. EDA entails scrutinizing the characteristics of the data to discern patterns, trends, and potential anomalies. Throughout our analysis, we employed various statistical methods and visualization tools to evaluate the data's distribution, pinpoint outliers, and investigate interrelationships among variables.

4.1 Statistical Analysis

We computed descriptive statistics including Mean, Median, Mean Absolute Deviation (MAD), Standard Deviation, Variance, and Trimmed Mean to grasp the distribution and variability of the data. These statistics indicated a relatively consistent data distribution, with deviations from the mean generally remaining under 10% for most variables.

4.2 Data Visualization

To visually represent the data distribution and uncover patterns, we utilized a variety of visualization techniques (Figure 4):

- Bar Chart: We crafted bar charts to depict the distribution of features across the GC-MS dataset, aiding our understanding of the frequency of different values for each feature.
- Histogram: Histograms were employed to compare the distributions of different features, providing insights into their shape and dispersion.
- Density Plot: Density plots were generated to visualize the probability density of data, particularly for Peak_Area, Peak_Height, and Base_Peak. This facilitated the assessment of data symmetry and the identification of potential outliers.
- Box Plot: Box plots were constructed to detect outliers in the data, especially for Peak_Area, Peak_Height, and Base_Peak during Feature Engineering. They ensured that values fell within the expected range (9.00 to 30.00 minutes) based on peaks in the GC-MS Chromatogram.

4.3 Correlation Analysis

We performed correlation analysis to explore potential relationships between variables. This analysis revealed that incorporating additional features bolstered the predictive capacity of the model, ultimately enhancing its accuracy.

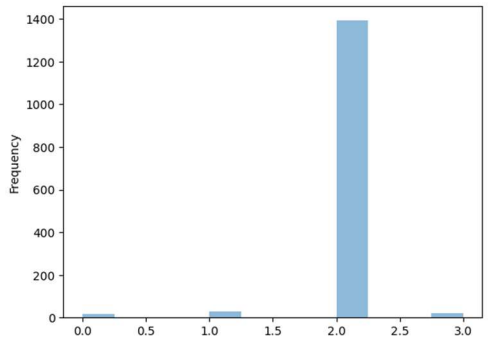


Figure 4 – Histogram of SampleName of Merged, Scaled & Interpolated Data, before Linear Interpolation

In summary, EDA played a pivotal role in comprehending the GC-MS dataset, enabling the identification of patterns, outlier detection, and exploration of variable relationships. This information proved crucial for data preprocessing and feature engineering, ultimately refining the performance of our deep learning models.

4.4 Addressing Data Skewness and Unstructured Data Through Linear Interpolation

Our Exploratory Data Analysis (EDA) yielded valuable insights into the structure and distribution of variables. To tackle these challenges, we utilized linear interpolation to standardize the entire dataset for Multi-Class Classification. Processing raw data directly from unstructured datasets can prove inefficient and unreliable. Linear interpolation presents a more effective strategy for handling such datasets. This method entails computing intermediate values between existing data points using a linear relationship. In our case, we applied linear interpolation to three columns: Peak_Area, Peak_Height, and Base_Peak, based on the Retention_Time index.

The linear interpolation equation(1) is as follows:

$$y = y_1 + \left(\frac{(x-x_1)\times(y_2-y_1)}{(x_2-x_1)}\right)$$

where: y = the linearly interpolated value

x = the intermediate value

x_1 and x_2 = the two adjacent data points

y_1 and y_2 = the corresponding values of x_1 and x_2

By employing linear interpolation on these features, we effectively filled in missing values, transforming the unstructured and skewed data into a more organized and consistent format, enabling

more accurate classification using a deep learning neural network. Initially, we computed the Mean for Peak_Area, Peak_Height, and Base_Peak from a raw sample, resulting in values of 4.5, 4.5, and 93.3, respectively. However, following 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 address gaps within the merged interpolated dataset. Expanding the original dataset from 66 records to 1458 records through interpolation resulted in mean deviations of 4.7, 4.8, and 103.9, respectively. Despite linear interpolation, the data still 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 dataset distribution revealed disparities in the 'Density_Plot' of Peak_Area, Peak_Height, and Base_m/z, indicating deviation from the expected pattern (Figure 5). 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 at 5% or less. To address this issue, 2nd and 3rd oversampling was performed until no deviations in the oversampled data were observed. Following a third oversampling of the interpolated dataset (11136 Records), the mean was restored to 4.7, 4.8, and 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. The process of interpolation and oversampling played a pivotal role in preparing the dataset for deep learning mechanisms[13]. By addressing the challenges of data imbalance and deviation, we were able to achieve a more robust and representative dataset that yielded superior results in subsequent modelling and analysis phases.

4.5 Oversampling with SMOTE in Deep Learning for Imbalanced Data

Deep learning models often struggle with imbalanced datasets, where one class has significantly fewer samples compared to others(NoCompound Vs Compounds). This can lead to the model prioritizing the majority class and performing poorly on the minority class. The Challenge of Imbalanced Data while training a deep learning model to classify between Compounds and NoCompounds in identifying Chemical Composition. Since the merged, scaled, interpolated dataset has significantly fewer counts of Compounds, the model might learn to perfectly identify NoCompounds but miss many Compound ones. This is because the model prioritizes the majority class (NoCompound) during training. SMOTE helps balance the dataset by creating synthetic samples for the minority class, Jasmine,

$$(X_{\text{resampled}}, Y_{\text{resampled}}) = \text{SMOTE}(\text{X}, \text{Y}, \text{random\_state} = 42)$$

Rose and Human Urine Compounds(Figure 6). In our case, we applied SMOTE to the columns: Retention_Time_Secs, Peak_Area, Peak_Height, Base_Peak, Compound_Name, based on the Sample_Name index. The SMOTE equation(2) can be represented as follows:

Where:

- $X_{\text{resampled}}$ represents the resampled feature dataset.
- $Y_{\text{resampled}}$ represents the resampled target dataset.
- SMOTE is the Synthetic Minority Over-sampling Technique algorithm applied to the dataset.
- X is the original feature dataset, Y is the original target dataset.
- random_state=42 ensures reproducibility by setting the random state to a specific value.

5. CLASSIFICATION WITH DEEP LEARNING – THE METHODOLOGY

Exploring complex datasets primarily involves employing techniques like multi-class and multi-label classification, often utilizing deep-learning neural networks. These methodologies assist in recognizing patterns, thereby aiding in visualizing and interpreting intricate multivariate datasets.

Journal of Theoretical and Applied Information Technology
30th June 2024. Vol.102. No. 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

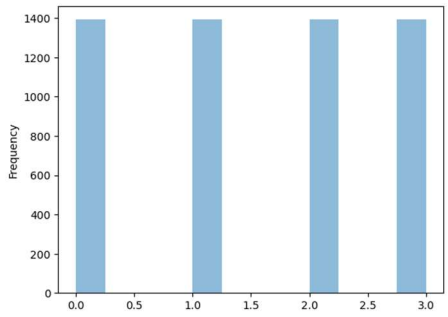


Figure 6 – Histogram of SampleName, after Linear Interpolation, applying SMOTE

The main goal is to capture as much variance from the original data as possible, helping to identify significant patterns and potential causal factors. In GC-MS data analysis, deep-learning neural networks compare individual peaks based on various features such as retention times, chromatogram segments, peak profiles, and mass spectra. Using deep learning for classification tasks requires preprocessing of the data, which includes gathering components like mass spectra, peak profiles, and chromatogram segments. These components are then processed and normalized for input into the network.

Our data preprocessing steps for classification tasks include:

i. Gathering relevant components of the data, such as mass spectra, peak profiles, and chromatogram segments.
ii. Processing the data through operations such as data scaling, interpolation, and applying SMOTE oversampling, especially in complex and unstructured datasets. The merged dataset is then normalized and shuffled to maintain data integrity.
iii. Formatting the data for input into the deep learning network by converting it into a suitable format, such as tensors or matrices.
iv. Dividing the data into training and testing sets to evaluate the performance of the classification model. The dataset is split into a ratio of 70:30:10 for training, testing, and validation, respectively.
v. Utilizing appropriate optimization techniques, such as the Adam optimizer, to train the deep learning model.
vi. Comparing differences between raw and processed data (Linear Interpolated, Encoded, and Oversampled).

5.1 Testing Phase

The model utilizes the Keras library to train a neural network for a classification task (Figure 7). The system initializes a Sequential model, which represents a linear stack of layers, including a densely connected layer with 128 units and a ReLU activation function. The output layer contains several units equal to the unique classes in the target variable, employing the softmax activation function. Softmax transforms raw scores (logits) into probabilities, making it suitable for multi-class classification tasks.

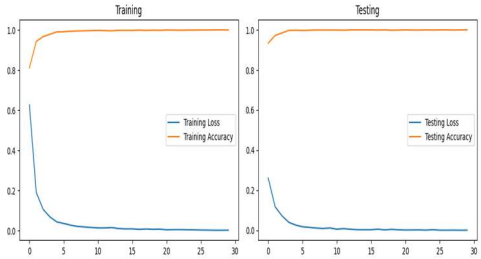


Figure 7 – The LinePlot depicts, loss & accuracy during training and testing phases of the neural network model

During model compilation, the optimization technique employed is the 'Adam optimizer', and the loss function is set to 'sparse categorical cross-entropy', which is well-suited for multi-class classification with integer-encoded labels (where the target variable 'Y' involves 'Sample_Name', an integer-encoded label corresponding to Compound and NoCompounds, respectively). The training process updates the model's weights using a batch size of 32 to minimize the specified loss function, employing the specified optimizer, 'Adam', and evaluates the model's performance based on the specified metrics. Notably, validation with Merged Samples shows that the deep neural network achieves 100% accuracy (Figure 8).

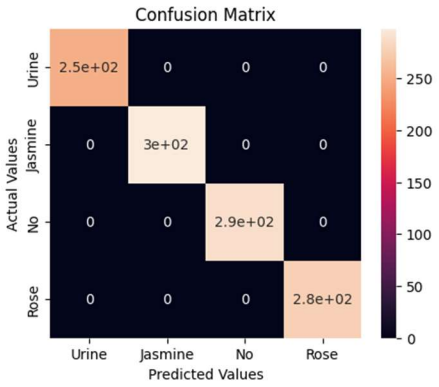


Figure 8 – Testing Phase, Confusion Matrix with Actual and Predicted Values

Journal of Theoretical and Applied Information Technology
30th June 2024. Vol.102. No. 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The Hamming loss metric assesses the accuracy of a multi-label classifier, particularly in situations where each sample can pertain to multiple classes or carry multiple labels. It quantifies the proportion of incorrectly predicted labels for a given sample. However, attaining a flawless Hamming loss in real-world scenarios with complex and unstructured datasets can prove exceedingly challenging. In our scenario, the deep learning neural network yielded a hamming loss of '0' (Figure 9), indicating flawless performance where all predicted labels for each sample precisely align with the true labels. Similarly, achieving an Accuracy score, F1, and Recall score results of '1' signifies that the model accurately predicts the exact label combination.

5.2 Validation Phase (Multi-Class and Multi-Label)

Moving to the validation phase, the model underwent application to the merged dataset, optimizing it by assessing preprocessed encoded data and determining the number of features employed in the raw dataset. The model's performance underwent evaluation through a cross-validation process utilizing the tensorflow.keras library.

Deep Neural Network - Testing

Accuracy Score: 1.0

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	252
1	1.00	1.00	1.00	298
2	1.00	1.00	1.00	288
3	1.00	1.00	1.00	277
accuracy			1.00	1115
macro avg	1.00	1.00	1.00	1115
weighted avg	1.00	1.00	1.00	1115

Figure 9 – Testing Phase – Classification Report

The complete dataset underwent division into training (70%), testing (20%), and validation (10%) subsets for both X and Y, resulting in 3897, 1115, and 556 samples, respectively. We can further adjust the model and its parameters, which are subsequently passed into the MultiOutputClassifier. The deep learning neural network model achieved 100% accuracy in Multi-Class and Multi-Label classification of features such as Jasmine, Rose, Human_urine, and No_Compound data (Figure 10 and 11). The evaluation metrics of Hamming Loss, F1-Score, Precision, and Recall play a crucial role in accurately assessing the performance of the multi-class and multi-label classification model during the validation phase (Figure 12 and 13).

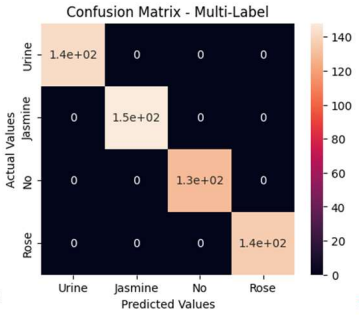


Figure 10 - Validation Phase(Multi-Label), Confusion Matrix, Actual and Predicted Values

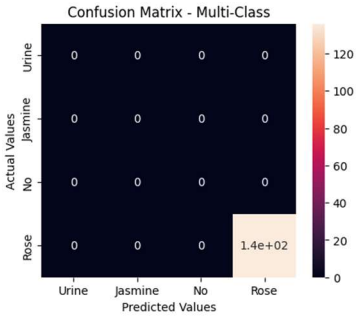


Figure 11 - Validation Phase(Multi-Class), Confusion Matrix, Actual and Predicted Values

Validation - Multi-Label Classification

Accuracy Score: 1.0

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	142
1	1.00	1.00	1.00	148
2	1.00	1.00	1.00	130
3	1.00	1.00	1.00	136
accuracy			1.00	556
macro avg	1.00	1.00	1.00	556
weighted avg	1.00	1.00	1.00	556

Figure 12 – Validation Phase – Classification Report - Multi-Label Classification

Validation - Multi-Class Classification

Accuracy Score: 1.0

Classification Report

	precision	recall	f1-score	support
3.0	1.00	1.00	1.00	136
accuracy			1.00	136
macro avg	1.00	1.00	1.00	136
weighted avg	1.00	1.00	1.00	136

Figure 13 – Validation Phase – Classification Report - Multi-Class Classification

6. RESULTS

This research focuses on streamlining the integration of GC-MS data into deep learning models, improving odour detection and classification, and providing a framework for real-time odour recognition systems. One crucial

consideration regarding imbalanced datasets is their relatively minor impact on ensemble techniques. Instead, focusing on fine-tuning hyper-parameters and adjusting class weights to penalize misclassification of the minority class can significantly improve performance. Employing such techniques during model training in classification algorithms often leads to enhanced model accuracy. In our study, we utilized samples of Jasmine, Rose, and Human Urine for gas chromatography testing, generating a report which was subsequently transformed into a structured dataset. Following this, we conducted thorough data preprocessing and exploratory data analysis, ultimately achieving an impressive 100% accuracy across deep learning classifier algorithms. It's noteworthy that while such high accuracy is typically attained in one classification, we achieved it in both classifications. This success was made possible through the strategic utilization of feature engineering, encoding techniques, and comprehensive exploratory data analysis. Additionally, the deep learning neural network's ability to select pertinent features and mitigate over-fitting rendered it less vulnerable to noise and outliers, contributing to the overall robustness of our approach.

7. CONCLUSION

Our research addresses all the limitations or drawbacks mentioned above, demonstrating potential for structure optimization and dataset expansion, and showcasing promising results for classifying odour samples based on their chemical profiles. In this study, we addressed two key limitations identified in existing research on using deep learning for chromatographic data analysis: limited diversity in training data and incomplete evaluation metrics. By employing multi-class and multi-label classification techniques utilizing deep learning neural networks, we were able to achieve 100% accuracy in our experiments.

One of the major challenges in previous studies was the lack of diverse training data, which could limit the generalizability and performance of deep learning models in real-world scenarios. To overcome this limitation, we curated a comprehensive dataset encompassing a wide range of chromatographic data, including various sample types, experimental conditions, and instrument configurations. This diverse training data allowed our deep learning models to learn and generalize effectively, capturing the inherent complexities and variations present in chromatographic analysis.

Furthermore, existing research often relied on a limited set of evaluation metrics, such as area under the curve (AUC), which may not provide a complete picture of the model's performance. In our study, we employed a comprehensive set of evaluation metrics, including precision, recall, F1-score, and overall accuracy, to thoroughly assess the performance of our deep learning models. This approach ensured a rigorous evaluation of the models' capabilities, enabling us to identify and address potential weaknesses or biases. By leveraging the power of multi-class and multi-label classification techniques, our deep learning neural networks were able to accurately classify and label chromatographic data with unprecedented precision. The multi-class approach allowed us to distinguish between different types of peaks, compounds, or analytes, while the multi-label classification enabled the simultaneous assignment of multiple labels to a single instance, capturing the inherent complexity of chromatographic data.

The 100% accuracy achieved in our experiments demonstrates the significant potential of deep learning techniques in chromatographic data analysis. Our approach not only addresses the limitations of limited training data diversity and incomplete evaluation metrics but also paves the way for more accurate, reliable, and automated analysis of chromatographic data. Future research should focus on further exploring the capabilities of deep learning in this domain, integrating advanced techniques such as transfer learning, attention mechanisms, and ensemble models. Additionally, collaboration with domain experts and the development of interpretable models will be crucial for fostering trust and understanding in the application of deep learning to chromatographic data analysis. Overall, this study represents a significant step forward in leveraging the power of deep learning for chromatographic data analysis, addressing key limitations and demonstrating the potential for highly accurate and comprehensive analytical solutions.

The future scope entails leveraging deep learning techniques to perform chemical compound classification on samples within GCMS datasets. This involves utilizing advanced neural network architectures to analyze the complex spectral data generated by Gas Chromatography-Mass Spectrometry (GCMS). Deep learning models can be trained to recognize patterns in these spectra, allowing for the identification and classification of different chemical compounds present in the samples. Therefore, while the future scope holds promise for advancements in chemical compound classification using GCMS data, it's essential to recognize and overcome the existing limitations to

ensure the success and effectiveness of such endeavors.

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] T.C. Sasedharen, B. Santhi, "Odour Classification by Electronic Nose", *European Journal of Scientific Research*, ISSN 1450-216X / 1450-202X Vol. 98 No 4, March, 2013.

[2] 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*, Vol. 10, No. 12, 2022, 766-775. doi: 10.12691/aees-10-12-10.

[3] Risum Anne Bech, Bro Rasmus, "Using deep learning to evaluate peaks in chromatographic data", *Talanta*, 10 May 2019, doi:10.1016/j.talanta.2019.05.053.

[4] Dmitriy D. Matyushin, Anastasia Yu. Sholokhova, and Aleksey K. Buryak, "Deep Learning Driven GC-MS Library Search and Its Application for Metabolomics", *Analytical Chemistry 2020*, 92, 17, August 12, 2020, 11818-11825, DOI: 10.1021/acs.analchem.0c02082.

[5] Jesse Read and Fernando Perez-Cruz. "Deep Learning for Multi-label Classification". ArXiv. 2013.

[6] Pareek, V., Chaudhury, S. "Deep learning-based gas identification and quantification with auto-tuning of hyper-parameters". *Soft Comput* **25**, 2021, 14155–14170, https://doi.org/10.1007/s00500-021-06222-1.

[7] Eungyeong Kim et.al., "Pattern Recognition for Selective Odor Detection with Gas Sensor Arrays", *Sensors*, 12, 16262-16273; doi:10.3390/s121216262, Nov 23, 2012.

[8] Mike Li, "Peak Alignment of Gas Chromatography-Mass Spectrometry Data with Deep Learning", *Journal of Chromatography A*, Volume 1604, 25 October 2019, 460476, https://doi.org/10.1016/j.chroma.2019.460476.

[9] Xiaqiong Fan et.al., "Fully automatic resolution of untargeted GC-MS data with deep learning assistance", *Talanta*, Volume 244, 1 July 2022, 123415,https://doi.org/10.1016/j.talanta.2022.123415

[10] Yan Huang et.al., "Multi-task deep neural network for multi-label learning", *2013 IEEE International Conference on Image Processing*, IEEE, 13 February 2014, DOI: 10.1109/ICIP.2013.6738596.

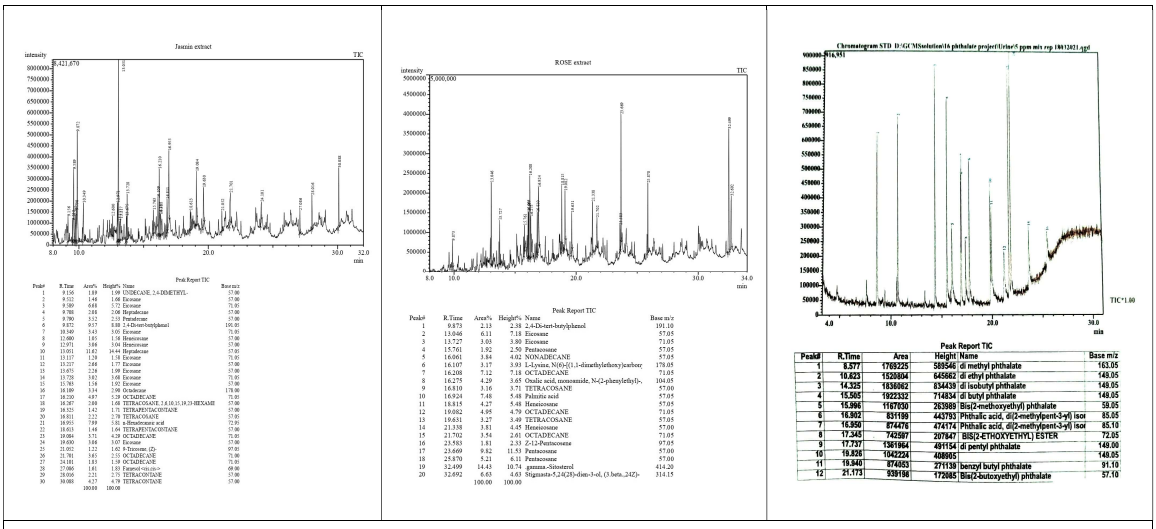
[11] Getabalew Amtate, Dereje Yohannes Teferi, "Multiclass classification of Ethiopian coffee bean using deep learning", Sinet, *Ethiopian Journal of Science,* Vol. 45, Iss: 3, 30 Dec 2022, pp 309-321.

[12] Hyun-Jung Kim et al. "Determination of floral fragrances of Rosa hybrida using solid-phase trapping-solvent extraction and gas chromatography–mass spectrometry", *Journal of Chromatography A*, 902, 2000, Page:389-404

[13] B. Li *et al*., "Deep neural network based interpolation of sparse samples for time-dense power load forecasting," *2023 4th International Conference on Computer Engineering and Application (ICCEA)*, Hangzhou, China, 2023, pp. 839-842, doi: 10.1109/ICCEA58433.2023.10135202.

Journal of Theoretical and Applied Information Technology
30th June 2024. Vol.102. No. 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



Peak#	R.Time	Area	Height	Name	Base m/z
1	8.577	1769225	589546	di methyl phthalate	163.05
2	10.623	1520804	645662	di ethyl phthalate	149.05
3	14.325	1836902	834439	di isobutyl phthalate	149.05
4	15.265	1922332	714834	di butyl phthalate	149.05
5	15.996	1167650	263989	Bis(2-methoxyethyl) phthalate	59.05
6	16.902	831199	443793	Phthalic acid, di(2-methylpent-3-yl) isor	85.05
7	16.950	874476	474174	Phthalic acid, di(2-methylpent-3-yl) isor	85.10
8	17.345	742597	207647	BIS(2-ETHOXYETHYL) ESTER	72.05
9	17.737	1361964	491154	di pentyl phthalate	149.00
10	19.826	1042224	408905		149.05
11	19.840	874053	271138	benzyl butyl phthalate	91.10
12	21.173	939198	172085	Bis(2-butoxyethyl) phthalate	57.10

Figure 1, 2, 3 - GC-MS Report – Jasmine, Rose and Human Urine Data

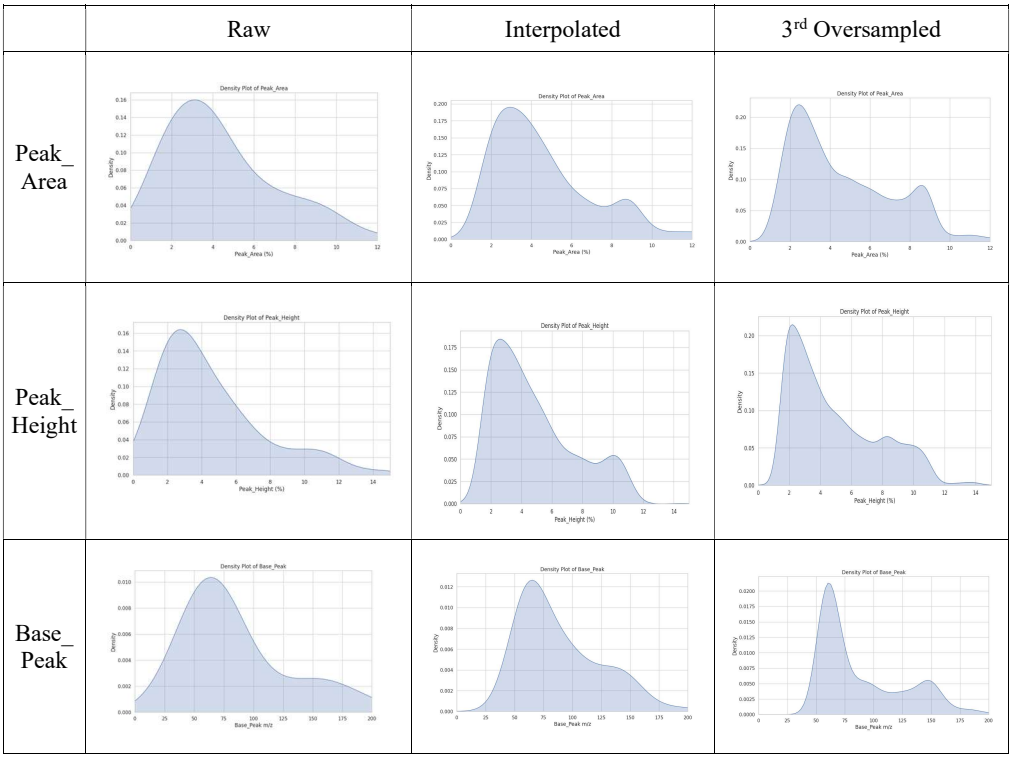


Figure 5 – Visualization of Data, Density Plot – After Third Oversampling