Pioneering AI in Chemical Data: New Frontline With GC-MS Generation

Namkyung Yoon and Hwangnam Kim*
School of Electrical Engineering Korea University Seoul, 02841, Korea
Email: nkyoon93, hnkim@korea.ac.kr

***Abstract*—The accurate detection and analysis of chemicals have become increasingly important for security and environmental monitoring with the integration of artificial intelligence (AI) methods gaining traction. However, the scarcity of certain chemicals poses significant challenges to the AI learning process. This paper presents a comprehensive AI approach and strategic direction for generating synthetic gas chromatography-mass spectrometry (GC-MS) data for such limited-availability chemicals. We conduct exploratory data analysis (EDA) on GC-MS data and apply advanced AI-driven generative algorithms, with a focus on Variational Autoencoder (VAE) and Generative Adversarial Network (GAN), acknowledging the challenges faced by current AI technologies in learning from chemical data. Additionally, we introduce a secondary contribution by developing custom Python-based tools for 3D visualization of GC-MS data, enhancing intuitive understanding and analysis precision. Our findings offer new possibilities and directions for the expansive application of AI in chemical analysis.**

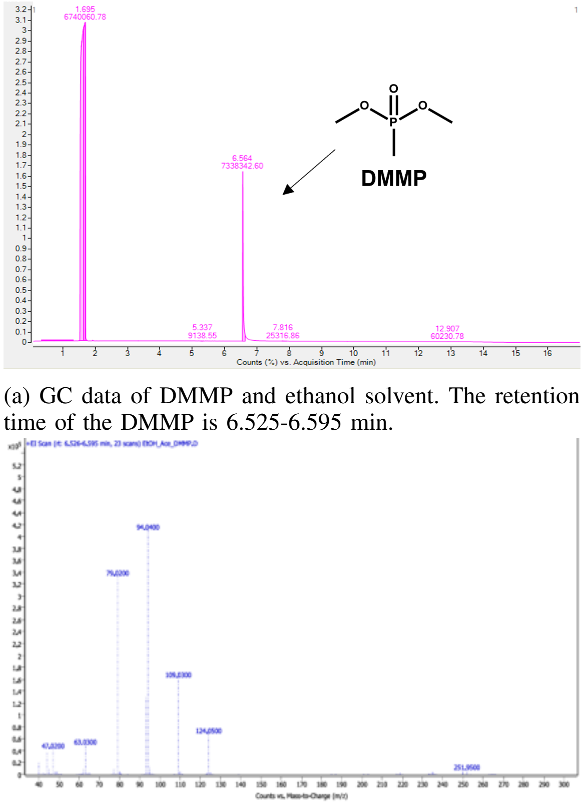
***Index Terms*—Data generation, Deep learning, Chemical data, Generative model**

I. Introduction

The application of chemical fields for tasks such as detection and analysis of substances in artificial intelligence (AI) has achieved many conveniences and developments [1]. In particular, gas chromatography mass spectrometry (GC-MS) has strong quantification and sensitivity to identification and analysis, and AI is also studied for GC-MS data analysis [2, 3]. However, comprehensive training datasets, which are essential for effective application of AI methodologies for GC-MS, can be challenging if they are data from limited chemicals.

AI for capabilities such as pattern recognition and predictive analysis requires rich and diverse dataset-based learning for guaranteed performance. Data generation can improve the quantity and quality of training data, thereby developing more accurate and generalizable AI models [4]. In view of this, we aim to explore and evaluate AI-driven data generation techniques to enrich GC-MS datasets for data with limited acquisition. In addition, exploratory data analysis (EDA) is conducted to analyze the complexity and characteristics of actual experimental GC-MS data.

In addition, this paper focuses on the use of Variational Autoencoder (VAE) and generative adversarial network (GAN), which have shown the possibility of generating one-dimensional time series data similar to GC-MS data based on previous studies [5].



(a) GC data of DMMP and ethanol solvent. The retention time of the DMMP is 6.525-6.595 min.

(b) MS data of DMMP and ethanol solvent at 6.526-6.595 min.

Fig. 1: GC-MS experimental data measured by mixing 2-CEES with ethanol solvent.

Additionally, we address another subtle but important challenge in this field: visualization of GC-MS data. We develop a tool that enables 3D visualization of GC-MS data in a more intuitive and insightful way of mass spectrometry peaks and their respective retention times. We evaluate the performance of existing models through these qualitative visualization evaluations and quantitative evaluations, and discuss appropriate improvements to the characteristics of GC-MS identified through EDA.

The remainder of this paper is organized as follows: Section

This work was supported by Korea Research Institute of defense Technology planning and advancement (KRIT) grant funded by Defense Acquisition Program Administration(DAPA) (KRIT-CT-21-034).

*Corresponding author

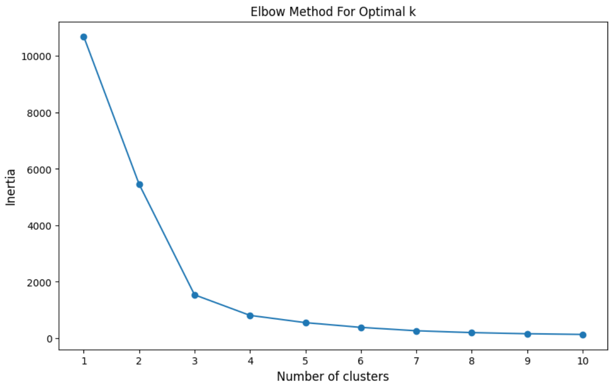


Fig. 2: Using the elbow method to find the optimal number of k-mean clusters.

2 provides a comprehensive literature review pertinent to our study. Section 3 describes the data preparation process and our exploratory data analysis methodology for GC-MS data. Section 4 presents the metrics for evaluating the quality of AI-generated data and includes a performance evaluation using 3D visualization tools. Finally, Section 5 concludes the paper by discussing the implications and future directions for the application of AI in stone chemistry.

II. Preliminary

A. GC-MS Data Analysis for CWAs

Gas chromatography-mass spectrometry (GC-MS) has played a major role in chemical warfare agent (CWA) analysis due to its unparalleled ability to separate complex mixtures and identify individual substances. Many studies have utilized GC-MS for qualitative and quantitative analysis of CWA using high resolution and sensitivity. For example, studies have shown that GC-MS can effectively identify CWA in environmental samples at trace levels that are essential for early detection and threat mitigation [6].

However, despite its capabilities, the use of GC-MS in CWA detection faces challenges such as the diversity of agents, the potential presence of interfering substances, and the need for rapid and accurate identification under varied conditions. To address these challenges, recent studies have focused on enhancing data processing and analysis methods, incorporating advanced algorithms for peak detection, deconvolution, and substance identification. These efforts aim to improve the reliability and efficiency of CWA detection, particularly in scenarios where rapid decision-making is critical [7, 8].

In this paper, we use experimental GC-MS data of Dimethyl methylphosphonate (DMMP), a simulant of a nerve agent, as shown in Fig.1. In addition, the characteristic retention time peaks in GC and MS data significantly differ in scale and magnitude compared to other peaks as shown in Fig.1a and Fig.1b. Moreover, the MS corresponding to the large peak in GC presents visual analysis challenges as shown in Fig.1a. We use models and preprocessing methods described later to demonstrate how these features work on generative AI.

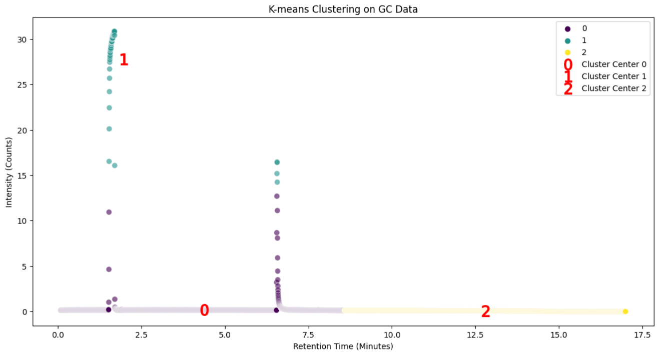
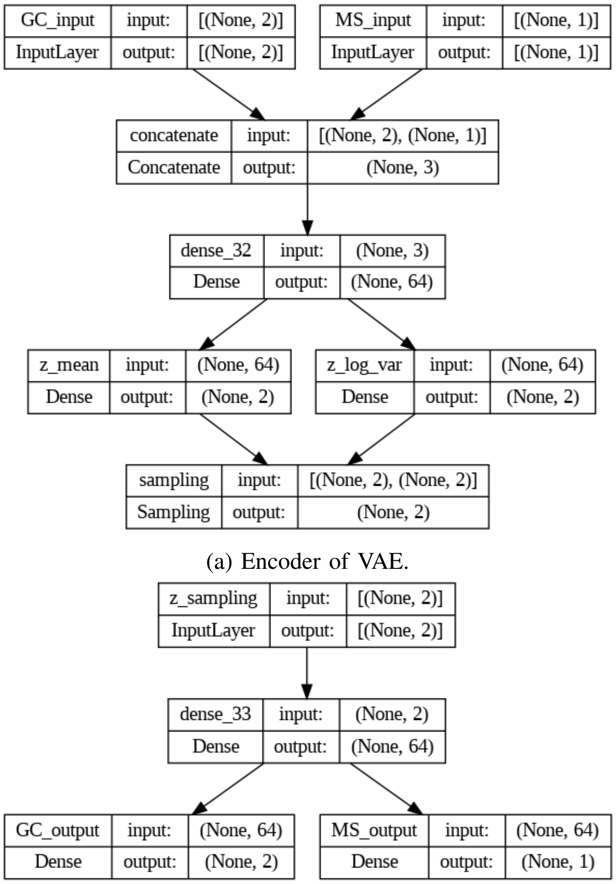


Fig. 3: Visualize K-means clustering on the original GC data feature space.



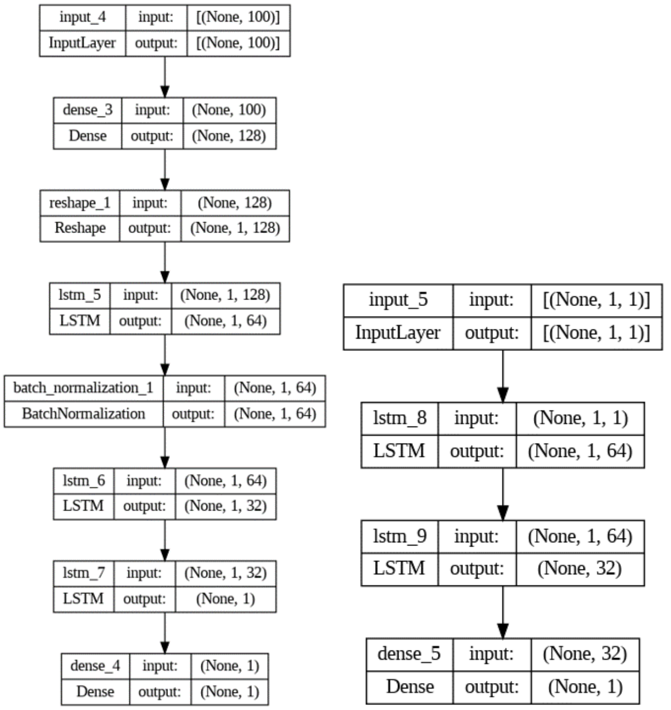
(a) Encoder of VAE.

(b) Decoder of VAE.

Fig. 4: Architecture of VAE.

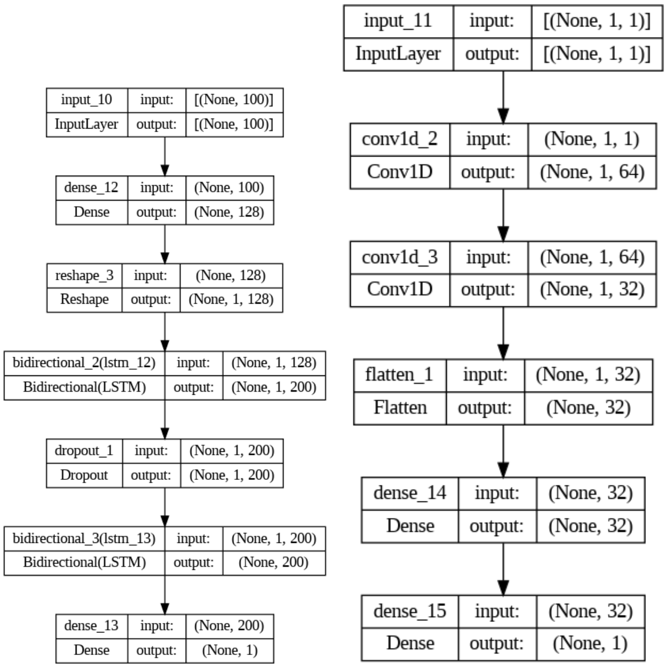
B. Generative AI Models

AI-based generation models evolve into various architectures and paradigms according to the development and purpose of technology. Since GC data is time series data, while MS data is not time-series data, it can be represented in a 1D format analogous to time series. According to previous studies, we apply VAE and GAN, which emerge as potent tools for time series data generation [5].


input_4 | input: | [(None, 100)]
InputLayer | output: | [(None, 100)]
dense_3 | input: | (None, 100)
Dense | output: | (None, 128)
reshape_1 | input: | (None, 128)
Reshape | output: | (None, 1, 128)
lstm_5 | input: | (None, 1, 128)
LSTM | output: | (None, 1, 64)
batch_normalization_1 | input: | (None, 1, 64)
BatchNormalization | output: | (None, 1, 64)
lstm_6 | input: | (None, 1, 64)
LSTM | output: | (None, 1, 32)
lstm_7 | input: | (None, 1, 32)
LSTM | output: | (None, 1)
dense_4 | input: | (None, 1)
Dense | output: | (None, 1)
input_5 | input: | [(None, 1, 1)]
InputLayer | output: | [(None, 1, 1)]
lstm_8 | input: | (None, 1, 1)
LSTM | output: | (None, 1, 64)
lstm_9 | input: | (None, 1, 64)
LSTM | output: | (None, 32)
dense_5 | input: | (None, 32)
Dense | output: | (None, 1)

(a) Generator of LSTM-GAN. (b) Discriminator of LSTM-GAN.

Fig. 5: Architecture of LSTM-GAN.


input_10 | input: | [(None, 100)]
InputLayer | output: | [(None, 100)]
dense_12 | input: | (None, 100)
Dense | output: | (None, 128)
reshape_3 | input: | (None, 128)
Reshape | output: | (None, 1, 128)
bidirectional_2(lstm_12) | input: | (None, 1, 128)
Bidirectional(LSTM) | output: | (None, 1, 200)
dropout_1 | input: | (None, 1, 200)
Dropout | output: | (None, 1, 200)
bidirectional_3(lstm_13) | input: | (None, 1, 200)
Bidirectional(LSTM) | output: | (None, 200)
dense_13 | input: | (None, 200)
Dense | output: | (None, 1)
input_11 | input: | [(None, 1, 1)]
InputLayer | output: | [(None, 1, 1)]
conv1d_2 | input: | (None, 1, 1)
Conv1D | output: | (None, 1, 64)
conv1d_3 | input: | (None, 1, 64)
Conv1D | output: | (None, 1, 32)
flatten_1 | input: | (None, 1, 32)
Flatten | output: | (None, 32)
dense_14 | input: | (None, 32)
Dense | output: | (None, 32)
dense_15 | input: | (None, 32)
Dense | output: | (None, 1)

(a) Generator of LSTM-CNN GAN. (b) Discriminator of LSTM-CNN GAN.

Fig. 6: Architecture of LSTM-CNN GAN.

VAE is a probabilistic model that can generate new data points by learning the distribution of existing data [9]. In the context of GC-MS, VAE can generate a data distribution that maintains the statistical properties of real samples, thereby enriching the dataset and enhancing model training of AI.

On the other hand, GAN produce high-fidelity synthetic data because they involve competitive learning in which one network generates data and the other evaluates it [10]. Recent studies have successfully applied GAN to generate realistic time series data, which allows for extensive model training and enhanced robustness against data variability and noise [11]. Both VAE and GAN have shown promise in different time series regions, suggesting potential applicability for GC-MS data generation for CWA detection.

Improving AI performance through data generation not only helps overcome data scarcity, but can also take advantage of the combined strengths of GC-MS and AI to lead the big development of analytical chemistry.

III. Methodology

A. Exploratory Data Analysis

Before applying complex AI algorithms, we perform exploratory data analysis (EDA) on the GC-MS data to uncover underlying patterns, detect outliers, and understand the data's structure. The EDA encompasses various statistical and machine learning techniques as follows:

- **K-means clustering**: K-means clustering effectively segments data into groups based on similarities as an unsupervised learning algorithm. It identifies patterns or groups in chemical signatures, which may correspond to interactions with various types of CWA or different solvents [12]. Utilizing the elbow method as shown in Fig.2, we determine the optimal number of clusters to be $K = 3$. These clusters allow us to explore data patterns without prior labeling, which is crucial in analyzing the unstructured nature of GC-MS data.
- **Cluster Interpretation**: The clusters are identified as shown in Fig.3. Cluster 0 corresponds to baseline values in zones characterized by pronounced peaks, suggesting regions of lower chemical activity or baseline noise. Cluster 1 consists of data points with intensity values significantly above the baseline, indicative of prominent chemical peaks. Cluster 2 comprises baseline values where no significant peaks are identified, representing areas of no or minimal chemical activity.

Our proposed EDA phase lays the foundation for subsequent application of AI-driven generative models, providing an integrated understanding of the structure and peak discrimination of GC-MS data. The insights at this stage suggest a direction to move forward through the selection of AI algorithms and techniques such as appropriate preprocessing.

B. AI-based Generative Models

We apply a preprocessing technique to address the inherent scaler deviation of GC-MS data identified through EDA before models suitable for time series data augmentation based on previous studies [5]. The normalization preprocessing technique of peak to overcome the deviation of GC-MS data scale is as follows Eq. (1), and we use it for all generative models that are applied later.

$$\text{Peak\_normalized} = \frac{\text{Peak\_value} - \text{Peak\_min}}{\text{Peak\_max} - \text{Peak\_min}}. \tag{1}$$

C. Variational Autoencoders (VAE)

We utilize the preprocessing as Eq. (1) and fully connected layer to apply the VAE which is known to be effective in learning time series data to GC-MS and construct it as shown in Fig.4 [13]. The model also balances the possibility of data with the complexity of latent representation and uses the Evidence Lower Bound (ELBO), which combines reconstruction loss with Kullback-Leibler divergence as follows:

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + \text{KL}(q(z|x)||p(z)).$$

D. Generative Adversarial Network (GAN)

To learn GC-MS data effectively, we design LSTM-CNN GAN and LSTM-GAN with the same preprocessing technique as Eq. (1). For both LSTM-CNN GAN and LSTM-GAN, the following minimax loss functions of generator G and discriminator D are used for adversarial learning as follows [10]:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = \min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] \\ + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))].\end{aligned} \tag{2}$$

- **LSTM-GAN**: We design LSTM-GAN using LSTM into the GAN architecture as shown in Fig.5, particularly suitable for time-series data where understanding long-term dependencies is crucial [14]. This is essential for GC-MS data, which exhibits complex temporal dynamics. The LSTM layers in both the generator and discriminator allow the model to capture these dynamics effectively:

  $$\mathbf{h}_t, \mathbf{c}_t = \text{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}). \tag{3}$$

  In Eq. (3), $\mathbf{h}_t$ and $\mathbf{c}_t$ represent the hidden state and cell state of the LSTM at time t, respectively, and $\mathbf{x}_t$ is the input at time t. The hidden state $\mathbf{h}_{t-1}$ and cell state $\mathbf{c}_{t-1}$ are from the previous time step, enabling the network to propagate information over longer periods. The LSTM-GAN uses these states to generate new synthetic GC-MS data that mimic the long-range temporal correlations present in the experimental data.
- **LSTM-CNN GAN**: We design the LSTM-CNN GAN using a bidirectional LSTM and 1D convolutional layer that have been studied to be effective for ECG data, which is similar in shape to GC-MS data as shown in Fig.6 [15]. We extend the LSTM used in Eq. (3) to be learned in both directions, and modify the proposed LSTM-CNN GAN model to be used for learning GC-MS data. The convolution operation in the 1D CNN layer applies a filter to the input sequence to capture the local features of GC-MS data as follows:

  $$\mathbf{f}_{\text{out}}[i] = \sum_{k=0}^{K-1} \mathbf{f}_{\text{in}}[i+k] \cdot \mathbf{w}[k] + b. \tag{4}$$

  In Eq. (4), $\mathbf{f}_{\text{out}}[i]$ represents the output feature at position i, $\mathbf{f}_{\text{in}}$ is the input feature sequence, $\mathbf{w}$ denotes the weights of the convolutional filter of size K, and b is the bias term. The LSTM-CNN GAN leverages this convolutional feature extraction to analyze and generate the complex chromatographic patterns in the GC-MS data.

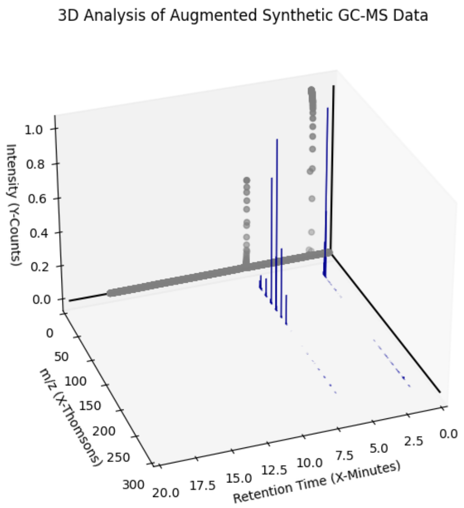


Fig. 7: Visualization of mixed GC-MS data of ethanol solvent and DMMP using 3D tool.

Parameter	Value
latent_dim	100
batch_size	32
epochs	1000

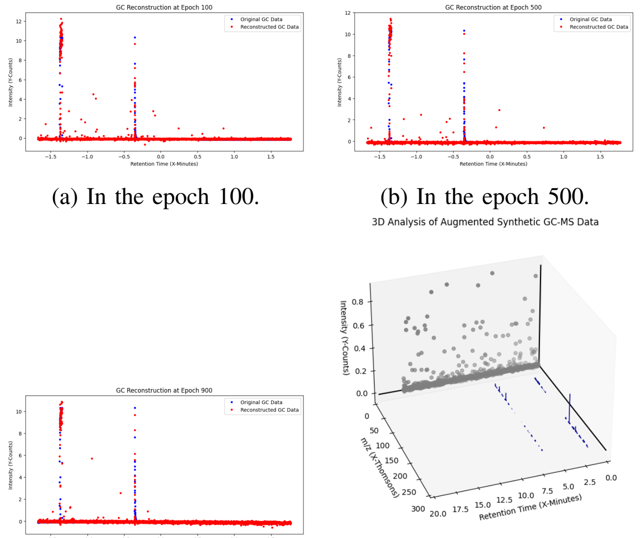
TABLE I: Parameters of AI-based generation models.

IV. Evaluation

A. GC-MS Experimental Setup

In this paper, data obtained by GC-MS analysis with an ethanol solvent are used for dimethyl methylphosphonate (DMMP), a similar agent of CWAs with limited acquisition. We conduct GC-MS experiments using standard protocols optimized for detection of CWA. We develop a customized 3D visualization tool to enhance interpretability of GC-MS data as an additional contribution to this study as shown in Fig.7. This facilitates the identification of patterns and correlations in GC-MS data that may not appear clearly in existing two-dimensional images as shown in Fig.1.

Visual qualitative evaluation with these tools is complemented by quantitative measurements such as the time it takes to identify the peak and the accuracy of the analysis, which shows a significant improvement over traditional 2D visualization methods.



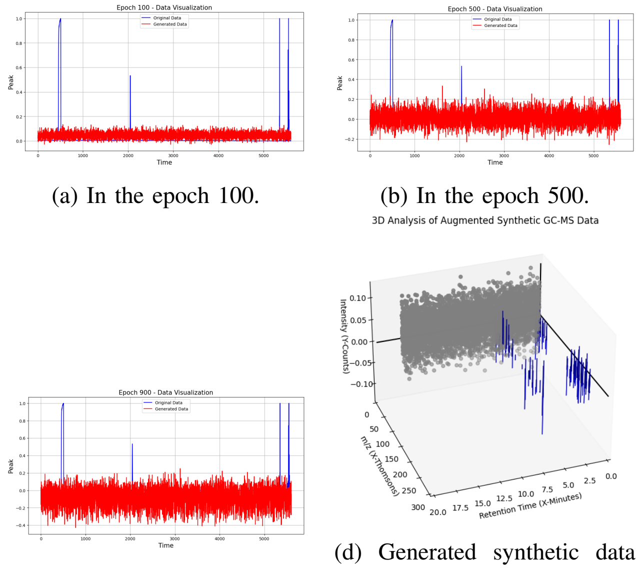
(a) In the epoch 100.

(b) In the epoch 500.

(c) In the epoch 900.

(d) Generated synthetic data using VAE.

Fig. 8: Visualization results of learning DMMP data using VAE.



(a) In the epoch 100.

(b) In the epoch 500.

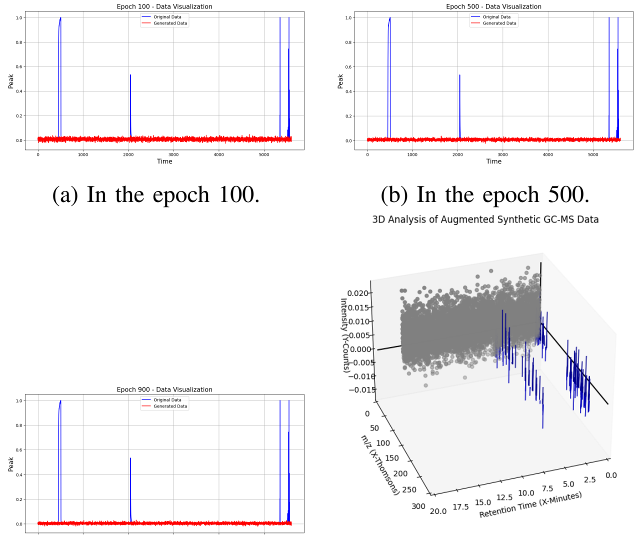
(c) In the epoch 900.

(d) Generated synthetic data using LSTM-GAN.

Fig. 9: Visualization results of learning DMMP data using LSTM-GAN.

	Data	PCC	Jacquard	PD
VAE [13]	GC	-0.0070	0.029	+63
	MS	-0.0187	0.066	+9
LSTM-GAN [14]	GC	0.0099	0.004	+27
	MS	0.0301	0.005	+10
LSTM-CNN GAN [15]	GC	0.0286	0.03	+54
	MS	0.0575	0.06	+13

TABLE II: Experimental Results of VAE, LSTM-GAN, and LSTM-CNN GAN Models



(a) In the epoch 100.

(b) In the epoch 500.

(c) In the epoch 900.

(d) Generated synthetic data using LSTM-CNN GAN.

Fig. 10: Visualization results of learning DMMP data using LSTM-CNN GAN.

B. Performance evaluations

We implement the proposed generative models to learn the dataset composed of GC and MS data using parameters as shown in Table I. Synthetic data generated through learning is compared with the original GC-MS data using the following metrics.

- **Pearson Correlation Coefficient (PCC)** The PCC measures the linear relationship between the composite dataset and the features of the original dataset. It ranges from -1 to +1, and the closer to zero, the less correlation.
- **Jaccard Similarity (Jaccard)** Jacquard similarity evaluates the similarity between two sets. This metric provides insight into how well the synthetic data captures the overall structure and pattern found in the original data.
- **Peak Difference (PD)** The peak difference evaluates the diversity of spectral peaks using local maxima by comparing adjacent values [16].

When evaluating the generated synthetic datasets for GC and MS data, PCC recorded values close to zero for all models. This represents a weak linear relationship between the generated synthetic data and the original data as shown in Fig.8d, Fig.9d, Fig.10d. An increase in the number of peaks was observed in the synthetic dataset compared to the original data for all models. This trend suggests the propensity of models to generate new chemical signal peaks. This suggests that new features can be introduced into the data synthesized by AI-based generation models to secure diversity of data different from simple modulation. Overall, low PCC and Jaccard, and the process by which models have significantly difficulty learning GC-MS data, as shown in Fig.8, Fig.9, Fig.10, indicate that the generative models studied for similar data forms are inadequate for the GC-MS field. Therefore, we propose an approach that fits the characteristics of the data, as shown in Fig.3, for the progress of GC-MS data generation beyond the limitations of these existing models. In order to properly learn GC-MS data analyzed through EDA, it is necessary to focus the input data on a specific part of the data. We implement and apply an attention mechanism layer for each model to selectively focus on important parts of GC-MS data [17]. As a result, it is confirmed that the performance is improved by 2% to 5%, but the influence of the attention mechanism on the model is insignificant. Through this evaluation, we propose that an innovative approach that goes beyond the paradigm of previously studied models is needed for successful generation of GC-MS data.

V. Conclusion

This work proposes an urgent need for generative model studies for limited data such as CWA and represents a comprehensive approach to artificial intelligence utilization, especially for generating GC-MS data. Machine learning-based EDA is conducted using a similar agent of actual experimental CWA to analyze the inherent characteristics. We also demonstrate the ability to create new synthetic instances using previously studied generative models aimed at GC-MS data and similar data, but experimentally reveal that training is challenged by complex patterns and large deviations of real GC-MS data. Ultimately, this work presents the need for a new approach tailored to GC-MS characteristics demonstrated by machine learning-based EDA for future research. It also marks an important step towards scalable integration, where AI-based generative models can be widely applied to chemical data synthesis.

References

[1] Rola Houhou and Thomas Bocklitz. Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data. *Analytical Science Advances*, 2(3-4):128–141, 2021.

[2] Nils Krone, Beverly A Hughes, Gareth G Lavery, Paul M Stewart, Wiebke Arlt, and Cedric HL Shackleton. Gas chromatography/mass spectrometry (gc/ms) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (lc/ms/ms). *The Journal of steroid biochemistry and molecular biology*, 121(3-5):496–504, 2010.

[3] Giacomo Baccolo, Beatriz Quintanilla-Casas, Stefania Vichi, Dillen Augustijn, and Rasmus Bro. From untargeted chemical profiling to peak tables–a fully automated ai driven approach to untargeted gc-ms. *TrAC Trends in Analytical Chemistry*, 145:116451, 2021.

[4] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2019.

[5] Guillermo Iglesias, Edgar Talavera, Ángel González-Prieto, Alberto Mozo, and Sandra Gómez-Canaval. Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications*, 35(14):10123–10145, 2023.

[6] Carlos A Valdez, Roald N Leif, Saphon Hok, and Bradley R Hart. Analysis of chemical warfare agents by gas chromatography-mass spectrometry: methods for their direct detection and derivatization approaches for the analysis of their degradation products. *Reviews in Analytical Chemistry*, 37(1):20170007, 2017.

[7] Xiuxia Du and Steven H Zeisel. Spectral deconvolution for gas chromatography mass spectrometry-based metabolomics: current status and future perspectives. *Computational and structural biotechnology journal*, 4(5):e201301013, 2013.

[8] Danny Yeap, Mitchell M McCartney, Maneeshin Y Rajapakse, Alexander G Fung, Nicholas J Kenyon, and Cristina E Davis. Peak detection and random forests classification software for gas chromatography/differential mobility spectrometry (gc/dms) data. *Chemometrics and Intelligent Laboratory Systems*, 203:104085, 2020.

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[11] Adam Wunderlich and Jack Sklar. Data-driven modeling of noise time series with convolutional generative adversarial networks. *Machine Learning: Science and Technology*, 4(3):035023, 2023.

[12] TR Noviandy, A Maulana, NR Sasmita, R Suhendra, GM Idroes, M Paristiowati, Z Helwani, E Yandri, S Rahimah, R Idroes, et al. The implementation of k-means clustering in kovats retention index on gas chromatography. In *IOP Conference Series: Materials Science and Engineering*, volume 1087, page 012051. IOP Publishing, 2021.

[13] Junying Li, Weijie Ren, and Min Han. Variational autoencoders based on the shift correction for imputation of specific missing in multivariate time series. *Measurement*, 186:110055, 2021.

[14] Md Abul Bashar and Richi Nayak. Algan: Time series anomaly detection with adjusted-lstm gan. *arXiv e-prints*, pages arXiv–2308, 2023.

[15] Fei Zhu, Fei Ye, Yuchen Fu, Quan Liu, and Bairong Shen. Electrocardiogram generation with a bidirectional lstm-cnn generative adversarial network. *Scientific reports*, 9(1):6734, 2019.

[16] P Virtanen, R Gommers, TE Oliphant, M Haberland, T Reddy, D Cournapeau, E Burovski, P Peterson, W Weckesser, J Bright, et al. Fundamental algorithms for scientific computing in python and scipy 1.0 contributors. scipy 1.0. *Nat. Methods*, 17:261–272, 2020.

[17] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.