Extraction of HUA Metabolic Features Based on Deep Learning

吴昭烨 30920241154578 信院班, 张传喜 23020241154470 ai 班, 王竹 23020241154449 信院班, 陈经安 30920241154562 信院班, 纪俊煜 23020241157994 信院

班

***Abstract*— Hyperuricemia (HUA) is a metabolic disorder caused by purine metabolism dysfunction. In recent years, with the improvement of living standards and changes in lifestyle in China, the prevalence of HUA has significantly increased and is showing a trend toward younger age groups. However, it remains unclear which HUA patients will develop gout or GN, and the mechanisms underlying its occurrence and progression are yet to be fully explored.**

In this study, we designed a CNN network and an LSTM model, focusing on the scientific problem of ”constructing a risk prediction model for gout based on metabolomics and genomics.” We propose a deep learning-based method for the extraction of metabolomic and genomic features of HUA and its complications, as well as for risk prediction. This approach incorporates multi-level feature analysis and aims to deeply investigate key biomarkers.

***Index Terms*— deep learning,CNN,Lstm,HUA,LC-MS**

I. Introduction

Metabolomics analyzes low-molecular-weight metabolites in cells, tissues, or biological fluids, enabling the detection of subtle changes in metabolic pathways and aiding in the identification of biomarkers associated with pathological conditions. Metabolomics experiments involve targeted and untargeted approaches. Untargeted metabolomics utilizes various analytical techniques such as liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectrometry (GC-MS) to analyze a wide range of chemical metabolite classes simultaneously.

In the study of gout, metabolomics has been preliminarily applied to identify metabolites and metabolic pathways associated with elevated serum uric acid (sUA) levels or gout attacks by comparing metabolite profiles of gout patients and healthy controls. The metabolic mechanisms associated with HUA and its gout complications are complex and high-dimensional. Effectively extracting and selecting key features related to disease progression, and constructing a machine learning model to accurately predict the onset of HUA and its transition to gout and GN, remains a critical challenge.

Genome-wide association studies (GWAS) have reported several gout-associated genes, including those related to renal and intestinal urate transport proteins. Current research on asymptomatic HUA and gout has largely focused on serum metabolites, with limited systematic multi-omics analysis of metabolites in urine, feces, and genomics.

Several studies have used metabolomics techniques to explore the metabolic characteristics of HUA patients and applied machine learning algorithms to identify biomarkers. For example, Shen et al. used high-resolution mass spectrometry to reveal significant differences in the metabolic characteristics of HUA and gout, particularly in the arginine metabolism pathway. Using machine learning algorithms such as random forests, support vector machines, and logistic regression, they identified 13 potential biomarkers that could effectively distinguish between gout, HUA, and normal uric acid levels. Another study investigated metabolic pathways associated with HUA and gout, finding that compared to healthy individuals, patients exhibited significant dysregulation in various pathways, especially those related to amino acid and purine metabolism.

Machine learning has been widely applied to metabolomics data processing and analysis. For instance, support vector machine (SVM) and random forest (RF) algorithms have been used to identify four sepsis biomarkers. Future studies are expected to integrate multi-omics data (e.g., genomics, transcriptomics, and metabolomics) to comprehensively analyze the metabolic networks and molecular mechanisms of gout and HUA. Giuseppe' s team found that in multi-omics regression tasks, multimodal regularized linear models demonstrated competitiveness and interpretability compared to data-hungry neural network approaches when learning from experimental and model-generated omics data. Machine learning-based multimodal biomarkers have been applied for early detection and prognosis prediction of HUA. The team developed and validated a stacked multimodal ML model trained on genetic and clinical data, which synthesized in silico quantitative markers for HUA. This model shows potential for timely HUA identification and personalized risk stratification for gout and metabolism-related outcomes.

Deep learning is an essential research method in metabolomics. Mayank Baranwal' s team proposed a deep learning architecture for metabolic pathway prediction, employing a hybrid machine learning approach that combines graph convolutional networks to extract molecular shape features as inputs for a random forest classifier. Compared to previous machine learning methods for this problem, their framework directly extracts relevant shape features from input SMILES representations—standardized notations of molecular chemical structures. Their method correctly predicted the corresponding metabolic pathway categories of 95.16% of tested compounds. Additionally, their framework achieved a prediction accuracy of 97.61% for the multi-label task of classifying compounds with mixed pathway memberships. Yongjie Deng' s team published an end-to-end deep learning method for mass spectrometry data analysis, capable of uncovering disease-specific metabolic features. This interpretable deep learning-based method performs end-to-end analysis of raw metabolic signals, delivering highly accurate and reliable outputs.

Raw liquid chromatography-mass spectrometry (LC-MS) is one of the most widely used analytical platforms in metabolomics. One critical challenge is processing raw signals, as LC-MS data typically consist of thousands of raw MS spectra, each with sequential numbers increasing with retention time (RT). These data include thousands of signals (features), making manual processing impractical. The typical LC-MS data processing workflow involves:

(1)Detecting regions of interest (ROI), (2)Detecting chromatographic peaks and integrating them, (3)Matching peaks across all samples in a batch (grouping), and (4)Annotating corresponding adducts and fragment ions to group peaks belonging to the same metabolite. XCMS and MZmine 2 are commonly used platforms for LC-MS data processing, but they often yield a high number of false-positive results. To address this, we applied a deep learning approach using classifier and segmentation models.

For ROI detection, we retained the classic centWave ROI method. Subsequently, we used a neural network (NN) to classify ROI regions, enabling effective peak detection.

II. Method

For ROI data, we should classify it into three categories: 1. Noise group 2. Peak-containing group, which includes one or more peaks 3. Peak-like group, representing uncertain peaks

We plan to use various neural networks to study the standard metabolites of gout. Considering the limitations of the dataset, we will start with the simplest CNN and progressively experiment with different neural network architectures.

A. Using convolutional neural networks to identify metabolites

Using convolutional neural networks (CNNs) to identify signature metabolites allows raw LC-MS metabolomics data to be directly used as input. The conventional stepwise methods for metabolomics analysis may result in significant loss of metabolic signals. By employing CNNs, the traditional processes of peak extraction and identification can be bypassed, improving both the efficiency and accuracy of data analysis.

We propose building an integrated end-to-end deep learning model that includes multiple CNNs as feature extractors to capture metabolomic signals associated with hyperuricemia (HUA).

First, we designed a classifier to categorize the types of ROI regions. This ROI classifier classifies the input ROI regions into one of the types described above. In our test dataset, the classification accuracy of the classifier reached 82%. It is worth noting that this accuracy is not low for our current design of ROI classification. If the classification is simplified to distinguishing between noise group and possible peak-containing group, the accuracy improves to 97%. However, the separation of confirmed peak-containing and possible peak-containing groups, while less accurate, is necessary.

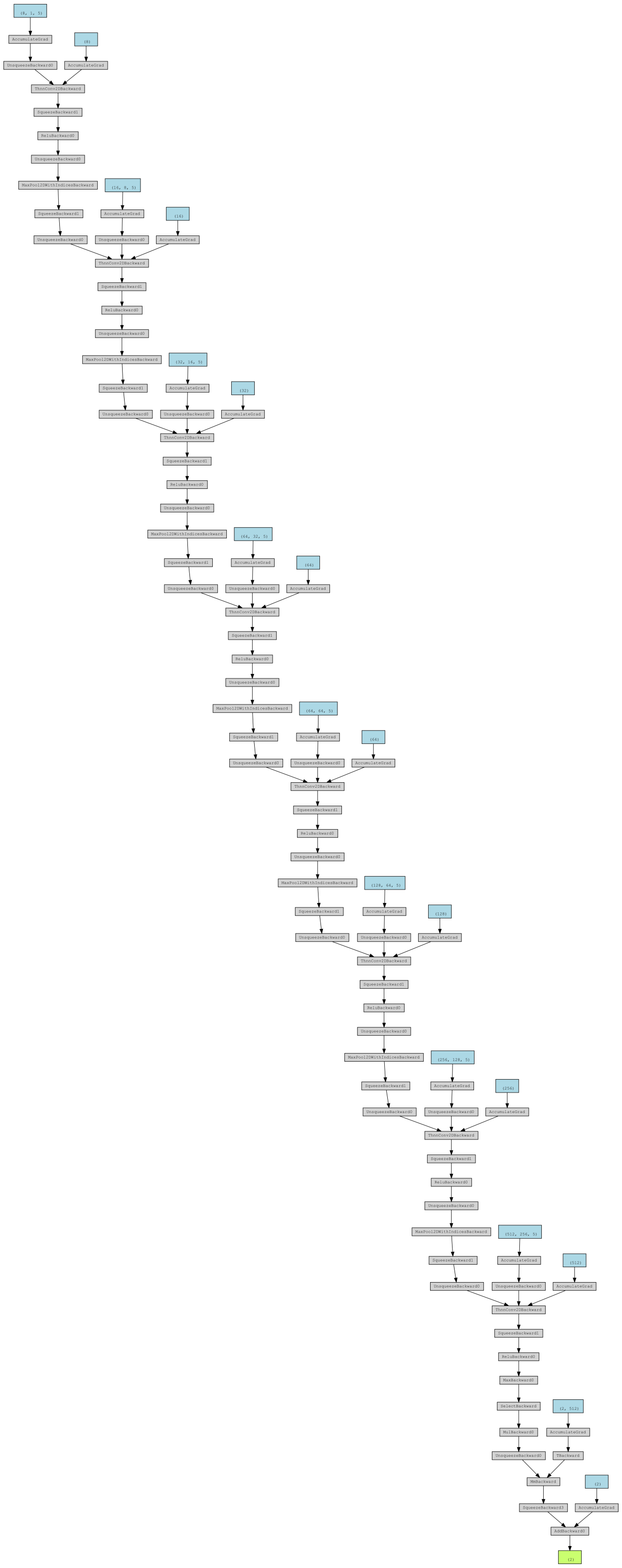


Fig. 1. CNN-based classifier architecture

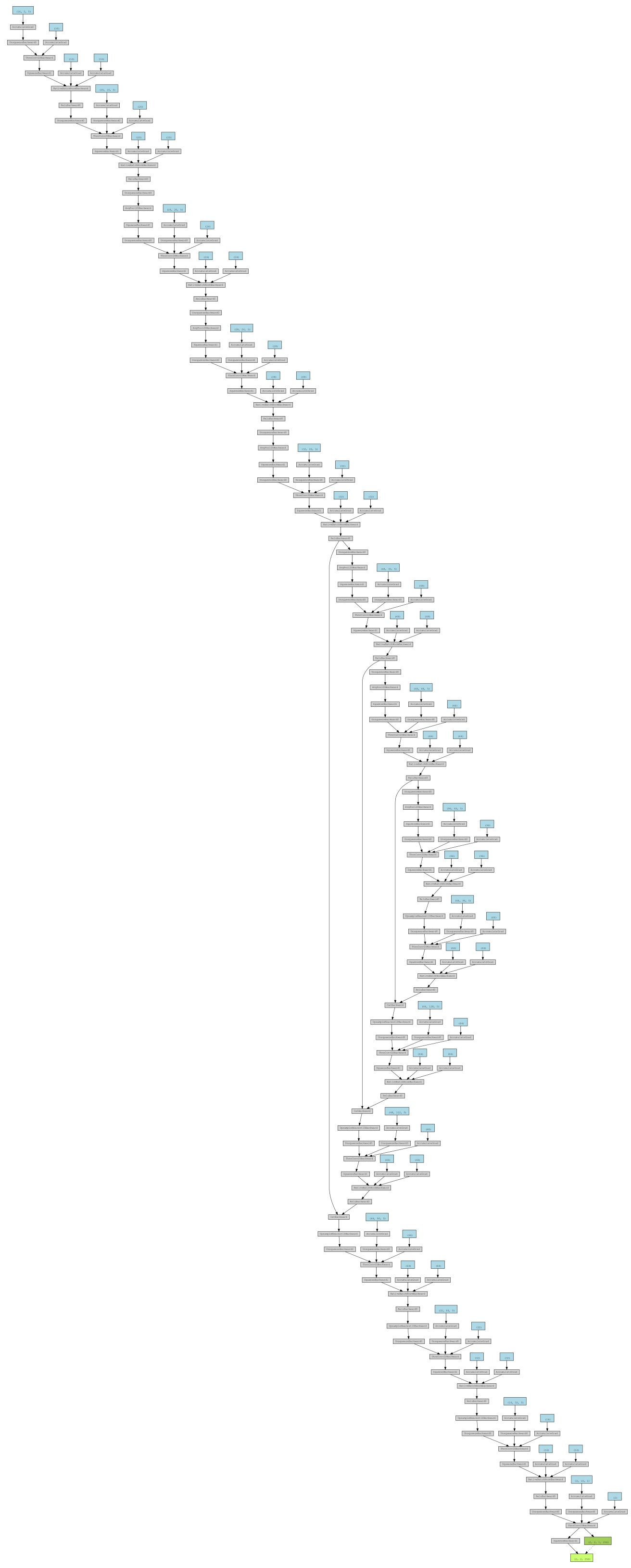


Fig. 2. CNN-based segmentator architecture

To identify peak regions, we treat the segmentation of peaks as a straightforward segmentation problem.

We designed a segmentation model to segment peak regions within the ROI. For this task, we employed a typical encoder-decoder structure commonly used in image segmentation tasks. This structure extracts features and restores the spatial structure of the image through convolution, pooling, and upsampling operations. The encoder extracts features, while the decoder maps these features back to high-resolution space.

Skip connections allow low-level features to propagate directly to subsequent layers, improving prediction accuracy.

In our test dataset, the model achieved an Intersection over Union (IoU) score of 0.82, approaching high-quality predictions. Considering that the model was trained on data with only 256-point dimensions, it demonstrates substantial potential for further improvement.

B. Using convolutional + LSTM neural networks to identify metabolites

Reviewing our previous work, we used convolutional neural networks to achieve peak detection, but this modeling approach does not fully match the characteristics of LC-MS data.

For time-series data, the core idea of the convolutional operation is to extract local patterns, such as the shape features of peaks, through a fixed window (i.e., convolutional kernel). Convolutional operations rely on a fixed-size window (receptive field) and can only extract features near local time points. Peaks in LC-MS often exhibit global dependencies, such as the start and end positions of a peak being far apart, which a simple CNN cannot capture through a single convolutional layer. This means that $\mathbf{y}_t$ can only include information from $\{x_{t-k}, \dots, x_t, \dots, x_{t+k}\}$, losing signal context beyond the receptive field.

Furthermore, convolutional operations are insensitive to the order of input data and cannot distinguish the ”early” or ”late” characteristics of a signal. For peak detection, the rise, peak, and fall of a signal follow a sequential order, but a simple CNN cannot directly model this directionality in time series. Since convolutional operations are symmetric operations within the window, $y_t = \sum_{i=-k}^{k} w_i x_{t+i}$, a symmetric weight distribution (e.g., $w_i = w_{-i}$) cannot distinguish temporal order information. Additionally, the boundaries and shapes of peak signals may depend on the global context. For example, a weak peak may only be identifiable against the background of the entire signal sequence. A simple CNN can only expand the receptive field by stacking more convolutional layers, which significantly increases the number of parameters and risks overfitting.

When considering noise, the situation becomes worse. CNNs treat signals in each local window equally, but in peak detection tasks, noise signals may be locally very strong. A CNN without a memory mechanism cannot ignore irrelevant noise signals from the past and is prone to misidentifying high-noise points as peaks. If the noise n_t in $\mathbf{x}_{t-k:t+k}$ is strong:

$$y_t = \mathbf{w}^\top (\mathbf{s}_{t-k:t+k} + \mathbf{n}_{t-k:t+k}) + b$$

where: The noise term $\mathbf{n}_{t-k:t+k}$ may dominate the output y_t, leading to misdetection.

Considering the characteristics of peak signals, peak signals s_t typically exhibit the following temporal dependency features: Local continuity (the value of s_t changes smoothly and continuously over time t). Global correlation (the shape of a peak is related to the changes in the preceding and following signals. For example, the rise and fall phases of a peak are correlated).

Therefore, the additional introduction of an LSTM would be an effective move, as LC-MS-generated data can be represented as a time series $\mathbf{x} = \{x_t\}_{t=1}^{T}$, where: $x_t \in \mathbb{R}$: Represents the intensity signal at time t. T: The total length of the time series.

In a noisy background, the signal can be expressed as:

$$x_t = s_t + n_t$$

where: s_t: Target signal (e.g., peak signal). n_t: Noise, usually assumed to be zero-mean Gaussian noise $n_t \sim \mathcal{N}(0, \sigma^2)$.

The goal of peak detection is to identify the regions of the peak signal s_t and annotate the start and end positions (i.e., boundaries) of the peaks.

LSTM can learn the dynamic change rules $f(\cdot)$ of the time-series signal s_t through its memory state $\mathbf{c}_t$ and hidden state $\mathbf{h}_t$: For local signal changes: LSTM can learn short-term dependencies, such as the rising and falling trends of a peak, through the hidden state $\mathbf{h}_t$. For global background modeling: The memory state $\mathbf{c}_t$ retains long-term dependency information and can capture the differences between peak boundaries and background signals.

The output of the LSTM can ultimately be expressed as:

$$\hat{s}_t = \mathbf{W}_y \mathbf{h}_t + \mathbf{b}_y$$

where $\hat{s}_t$ is the predicted value of the target signal s_t.

Through its input gate and forget gate mechanism, LSTM dynamically controls the influence of the input signal x_t on the memory state $\mathbf{c}_t$: When noise is significant, the forget gate reduces the weight of the noise signal, suppressing its interference with $\mathbf{c}_t$. This selective memory mechanism makes LSTM robust to noise.

Mathematically, the memory state is updated as:

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\cdot)$$

where $\mathbf{f}_t$ and $\mathbf{i}_t$ are dynamic adjustment coefficients, ensuring that only significant peak signals are remembered.

When the signal s_t contains multiple overlapping peaks:

$$s_t = \sum_{k=1}^{K} s_t^{(k)}$$

where: $s_t^{(k)}$: Signal of the k-th peak. K: Number of overlapping peaks.

LSTM can learn the features $f^{(k)}(\cdot)$ of each peak through its hidden state $\mathbf{h}_t$ and memory state $\mathbf{c}_t$, thereby demixing the signals.

Specifically: $\mathbf{h}_t$: Stores the comprehensive information at time point t. $\mathbf{c}_t$: Accumulates and separates the long-term features of multiple peaks.

The final output extracts the predictions of all peaks through $\mathbf{W}_y \mathbf{h}_t$, and post-processing steps separate the overlapping signals.

Based on this principle, we added LSTM layers to both the classifier and the segmenter, and the model architectures are shown in Figures 3 and 4.

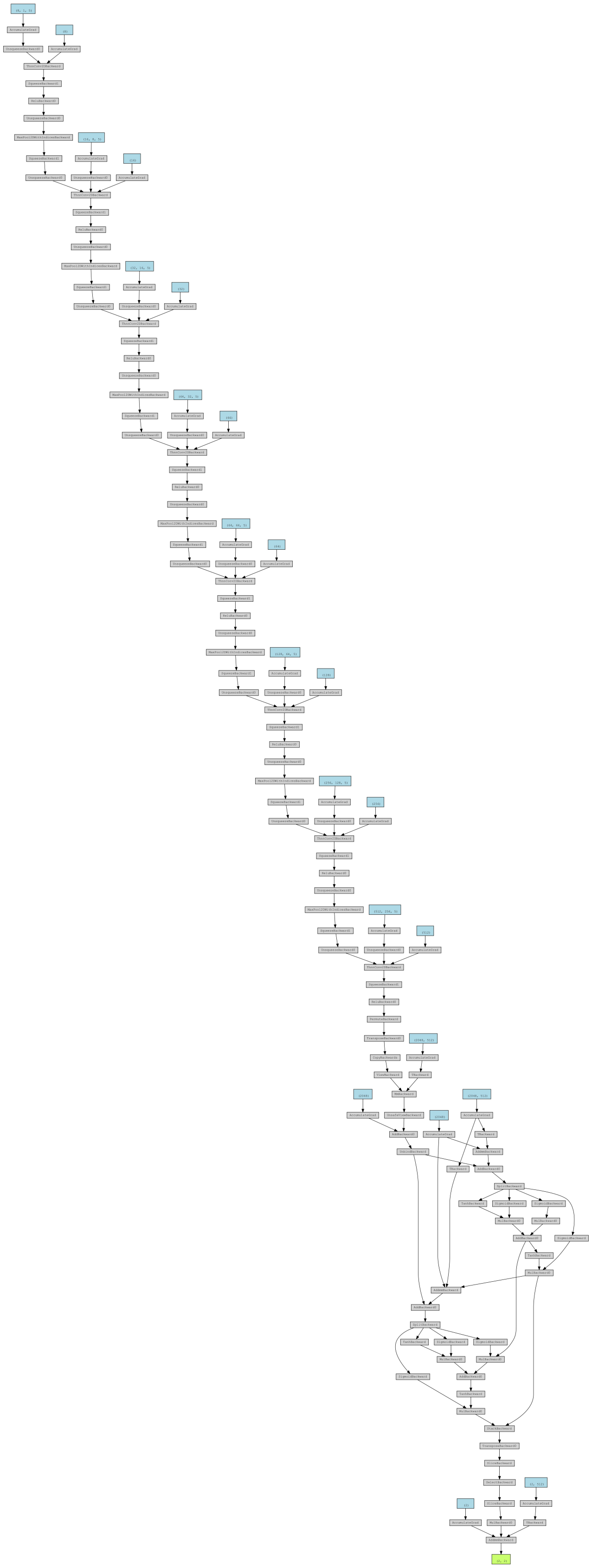
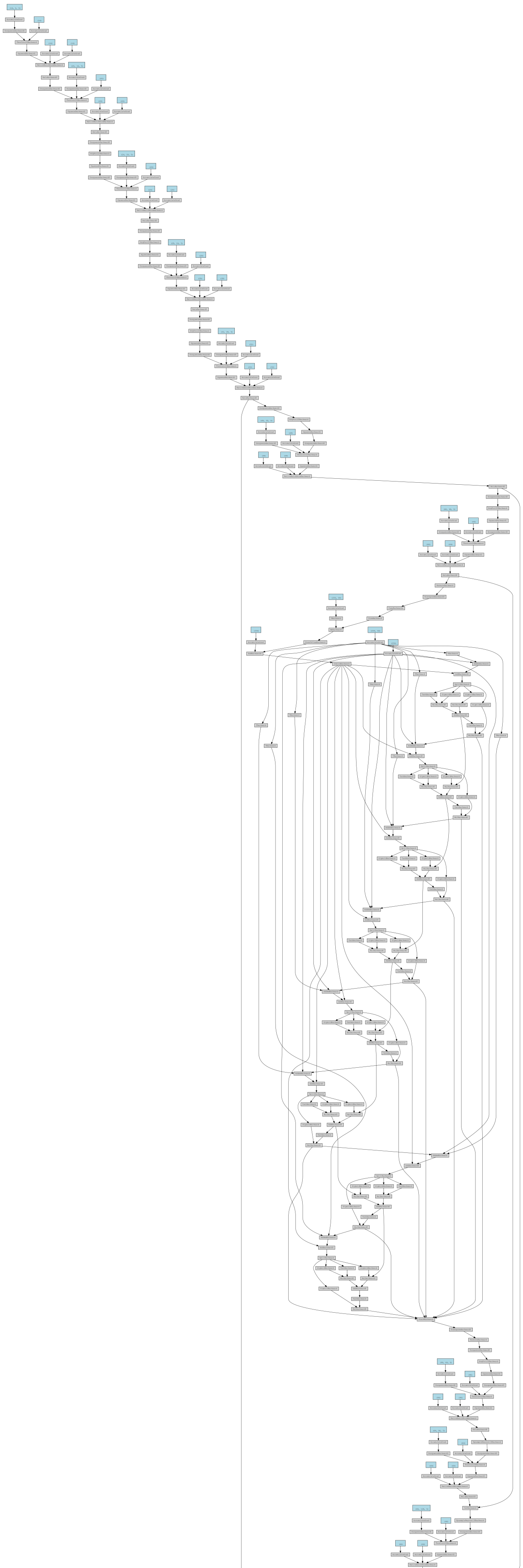


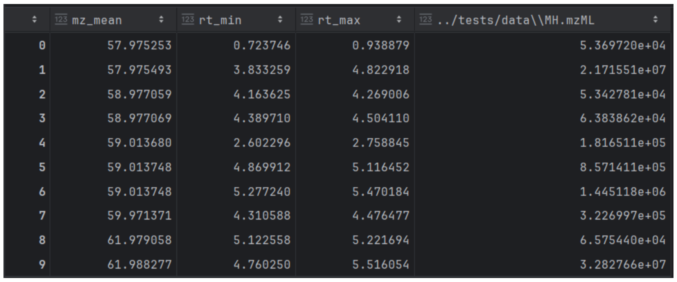
Fig. 3. CNN-based classifier architecture with lstm



The model achieved an IoU of 0.85 on the test set, and the improvement might seem modest. However, it is important to consider that the current ROI region setting is only 256, which is far from sufficient to fully leverage the capabilities of LSTM. As the ROI region selection method improves and the ROI region increases, the performance of our model will also show significant improvement.

C. Detection of HUA Characteristic Metabolites Based on Deep Learning Models

We used our trained model to perform a simple analysis on the LC-MS spectrum data of patients with HUA, as shown in Figure 5. We detected a total of 2,210 characteristic peaks, and the target regions have also been segmented, as shown in Figure 6.



	mz_mean	rt_min	rt_max	../tests/data\\MH.mzML
0	57.975253	0.723746	0.938879	5.369720e+04
1	57.975493	3.833259	4.822918	2.171551e+07
2	58.977059	4.163625	4.269006	5.342781e+04
3	58.977069	4.389710	4.504110	6.383862e+04
4	59.013680	2.602296	2.758845	1.816511e+05
5	59.013748	4.869912	5.116452	8.571411e+05
6	59.013748	5.277240	5.470184	1.445118e+06
7	59.971371	4.310588	4.476477	3.226997e+05
8	61.979058	5.122558	5.221694	6.575440e+04
9	61.988277	4.760250	5.516054	3.282766e+07

Fig. 5. CNN-based segmentator architecture with lstm

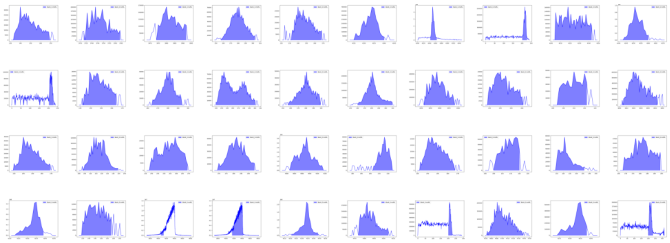


Fig. 6. CNN-based segmentator architecture with lstm

III. Discussion

Our network still has room for improvement, with the main limiting factor being the rigid division of the ROI. The ROI areas are quite small, which prevents our model from fully demonstrating its performance. Additionally, the classifier's categorization method can be optimized, as the definition of the classification criteria has led to some unclear data annotations. One potential improvement could be to treat the ROI division method as a parameter and include it in the training of the neural network, which could be a good way to enhance performance.

IV. Conclusion

We designed a deep neural network to analyze LC-MS data. Initially, we implemented this network using CNNs and then experimented with adding LSTM modules. Our approach successfully achieved ROI classification and peak segmentation. This method effectively enhances the accuracy of data analysis, providing a more reliable feature detection tool compared to traditional methods. Using this network, we can identify potential peaks from LC-MS spectra, enabling subsequent algorithms to determine the desired characteristic metabolites based on these peaks.

References

[1] 中华医学会内分泌学分会, 中国高尿酸血症与痛风诊疗指南 (2019). 中华内分泌代谢杂志, 2020. 36(1): p. 1-13.

[2] Martinon, F., et al., Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature*, 2006. 440(7081): p. 237-241.

[3] Sellmayr, M., et al., Only Hyperuricemia with Crystalluria, but not Asymptomatic Hyperuricemia, Drives Progression of Chronic Kidney Disease. *Journal of the American Society of Nephrology: JASN*, 2020. 31(12): p. 2773-2792.

[4] Brown, J. and G.K. Mallory, Renal changes in gout. *The New England Journal of Medicine*, 1950. 243(9): p. 325-329.

[5] Dalbeth, N., et al., Gout. *Lancet (London, England)*, 2021. 397(10287): p. 1843-1855.

[6] Johnson, C.H., J. Ivanisevic, and G. Siuzdak, Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews. Molecular Cell Biology*, 2016. 17(7): p. 451-459.

[7] Banimfreg, B.H., et al., Untargeted Metabolomic Plasma Profiling of Emirati Dialysis Patients with Diabetes versus Non-Diabetic: A Pilot Study. *Biomolecules*, 2022. 12(7).

[8] Ikram, M.M.M., et al., GC-MS Based Metabolite Profiling to Monitor Ripening-Specific Metabolites in Pineapple (Ananas comosus). *Metabolites*, 2020. 10(4).

[9] Shen, X., et al., Serum Metabolomics Identifies Dysregulated Pathways and Potential Metabolic Biomarkers for Hyperuricemia and Gout. *Arthritis Rheumatol*, 2021. 73(9): p. 1738-1748.

[10] Huang, Y., et al., Identification of the urine and serum metabolomics signature of gout. *Rheumatology (Oxford)*, 2020. 59(10): p. 2960-2969.

[11] Köttgen, A., et al., Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat Genet*, 2013. 45(2): p. 145-54.

[12] Matsuo, H., et al., Genome-wide association study of clinically defined gout identifies multiple risk loci and its association with clinical subtypes. *Ann Rheum Dis*, 2016. 75(4): p. 652-9.

[13] Chang, B.S., Ancient insights into uric acid metabolism in primates. *Proc Natl Acad Sci U S A*, 2014. 111(10): p. 3657-8.

[14] Dalbeth, N., et al., Urate crystal deposition in asymptomatic hyperuricaemia and symptomatic gout: a dual energy CT study. *Ann Rheum Dis*, 2015. 74(5): p. 908-11.

[15] Stewart, S., et al., Ultrasound Features of the First Metatarsophalangeal Joint in Gout and Asymptomatic Hyperuricemia: Comparison With Normouricemic Individuals. *Arthritis Care Res (Hoboken)*, 2017. 69(6): p. 875-883.

[16] Wang, P., et al., Identification of monosodium urate crystal deposits in patients with asymptomatic hyperuricemia using dual-energy CT. *RMD Open*, 2018. 4(1): p. e000593.

[17] Jung, S.W., et al., Uric acid and inflammation in kidney disease. *Am J Physiol Renal Physiol*, 2020. 318(6): p. F1327-f1340.

[18] Zhu, P., et al., Serum uric acid is associated with incident chronic kidney disease in middle-aged populations: a meta-analysis of 15 cohort studies. *PLoS One*, 2014. 9(6): p. e100801.

[19] Oh, T.R., et al., Hyperuricemia has increased the risk of progression of chronic kidney disease: propensity score matching analysis from the KNOW-CKD study. *Sci Rep*, 2019. 9(1): p. 6681.

[20] Piani, F. and R.J. Johnson, Does gouty nephropathy exist, and is it more common than we think? *Kidney Int*, 2021. 99(1): p. 31-33.

[21] Wang, S., et al., Research progress of risk factors and early diagnostic biomarkers of gout-induced renal injury. *Front Immunol*, 2022. 13: p. 908517.

[22] Li, H., et al., Kidney and plasma metabolomics provide insights into the molecular mechanisms of urate nephropathy in a mouse model of hyperuricemia. *Biochim Biophys Acta Mol Basis Dis*, 2022. 1868(6): p. 166374.

[23] Liu, M., et al., Synergistic effect of Aconiti Lateralis Radix Praeparata water-soluble alkaloids and Ginseng Radix et Rhizoma total ginsenosides compatibility on acute heart failure rats. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2020. 1137: p. 121935.

[24] Ohashi, Y., et al., Plasma and Urinary Metabolomic Analysis of Gout and Asymptomatic Hyperuricemia and Profiling of Potential Biomarkers: A Pilot Study. *Biomedicines*, 2024. 12(2).

[25] Han, T., et al., Temporal Relationship Between Hyperuricemia and Insulin Resistance and Its Impact on Future Risk of Hypertension. *Hypertension*, 2017. 70(4): p. 703-711.

[26] Bombelli, M., et al., Uric acid and risk of new-onset metabolic syndrome, impaired fasting glucose and diabetes mellitus in a general Italian population: data from the Pressioni Arteriose Monitorate E Loro Associazioni study. *J Hypertens*, 2018. 36(7): p. 1492-1498.

[27] Hahn, K., et al., Serum uric acid and acute kidney injury: A mini review. *J Adv Res*, 2017. 8(5): p. 529-536.

[28] Miao, H., et al., 1-Hydroxypyrene mediates renal fibrosis through aryl hydrocarbon receptor signalling pathway. *Br J Pharmacol*, 2022. 179(1): p. 103-124.

[29] Tan, Y.M., et al., Plasma Metabolome and Lipidome Associations with Type 2 Diabetes and Diabetic Nephropathy. *Metabolites*, 2021. 11(4).

[30] Dalbeth, N., L.K. Stamp, and T.R. Merriman, The genetics of gout: towards personalised medicine? *BMC Med*, 2017. 15(1): p. 108.

[31] Kawamura, Y., et al., Genome-wide association study revealed novel loci which aggravate asymptomatic hyperuricaemia into gout. *Ann Rheum Dis*, 2019. 78(10): p. 1430-1437.

[32] Phipps-Green, A.J., et al., Twenty-eight loci that influence serum urate levels: analysis of association with gout. *Ann Rheum Dis*, 2016. 75(1): p. 124-30.

[33] Nakayama, A., et al., GWAS of clinically defined gout and subtypes identifies multiple susceptibility loci that include urate transporter genes. *Ann Rheum Dis*, 2017. 76(5): p. 869-877.

[34] Li, C., et al., Genome-wide association analysis identifies three new risk loci for gout arthritis in Han Chinese. *Nat Commun*, 2015. 6: p. 7041.

[35] Sulem, P., et al., Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat Genet*, 2011. 43(11): p. 1127-30.

[36] Stiburkova, B., et al., The impact of dysfunctional variants of ABCG2 on hyperuricemia and gout in pediatric-onset patients. *Arthritis Res Ther*, 2019. 21(1): p. 77.

[37] Merriman, T. and R. Terkeltaub, PPARGC1B: insight into the expression of the gouty inflammation phenotype: PPARGC1B and gouty inflammation. *Rheumatology (Oxford)*, 2017. 56(3): p. 323-325.

[38] Chen, Y., et al., CARD8 rs2043211 polymorphism is associated with gout in a Chinese male population. *Cell Physiol Biochem*, 2015. 35(4): p. 1394-400.

[39] Rasheed, H., et al., The Toll-Like Receptor 4 (TLR4) Variant rs2149356 and Risk of Gout in European and Polynesian Sample Sets. *PLoS One*, 2016. 11(1): p. e0147939.

[40] Rasheed, H., et al., Replication of association of the apolipoprotein A1-C3-A4 gene cluster with the risk of gout. *Rheumatology (Oxford)*, 2016. 55(8): p. 1421-30.