Received: 10 June 2024 | Revised: 1 August 2024 | Accepted: 4 August 2024

DOI: 10.1002/pca.3437

RESEARCH ARTICLE

Phytochemical Analysis WILEY

HerbMet: Enhancing metabolomics data analysis for accurate identification of Chinese herbal medicines using deep learning

Yuyang Sha[1] | Meiting Jiang[2] | Gang Luo[1] | Weiyu Meng[1] | Xiaobing Zhai[1] | Hongxin Pan[1] | Junrong Li[1] | Yan Yan[3] | Yongkang Qiao[4] | Wenzhi Yang[2] | Kefeng Li[1]

[1]Center for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Macau, China

[2]National Key Laboratory of Chinese Medicine Modernization, State Key Laboratory of Component-based Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin, China

[3]Guangdong-Hong Kong-Macao University Joint Laboratory of Interventional Medicine, Center of Molecular Imaging, The Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, China

[4]Centre for Biological Science and Technology, Key Laboratory of Cell Proliferation and Regulation Biology of Ministry of Education, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai, China

Correspondence

Yongkang Qiao, Centre for Biological Science and Technology, Key Laboratory of Cell Proliferation and Regulation Biology of Ministry of Education, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519000, China.
Email: ykqiao@bnu.edu.cn

Wenzhi Yang, National Key Laboratory of Chinese Medicine Modernization, State Key Laboratory of Component-based Chinese Medicine, Tianjin University of Traditional Chinese Medicine, 10 Poyanghu Road, Tianjin 301617, China.
Email: wzyang0504@tjutcm.edu.cn.

Kefeng Li, Center for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Macau SR 999708, China.
Email: kefengl@mpu.edu.mo

Funding information

Macao Polytechnic University, Grant/Award Number: RP/FCA-14/2023; Science and Technology Development Funds (FDCT) of Macao, Grant/Award Number: 0033/2023/RIB2; Tianjin Outstanding Youth Fund, Grant/Award Number: 23JCJQJC00030

Abstract

Introduction: Chinese herbal medicines have been utilized for thousands of years to prevent and treat diseases. Accurate identification is crucial since their medicinal effects vary between species and varieties. Metabolomics is a promising approach to distinguish herbs. However, current metabolomics data analysis and modeling in Chinese herbal medicines are limited by small sample sizes, high dimensionality, and overfitting.

Objectives: This study aims to use metabolomics data to develop HerbMet, a high-performance artificial intelligence system for accurately identifying Chinese herbal medicines, particularly those from different species of the same genus.

Methods: We propose HerbMet, an AI-based system for accurately identifying Chinese herbal medicines. HerbMet employs a 1D-ResNet architecture to extract discriminative features from input samples and uses a multilayer perceptron for classification. Additionally, we design the double dropout regularization module to alleviate overfitting and improve model's performance.

Results: Compared to 10 commonly used machine learning and deep learning methods, HerbMet achieves superior accuracy and robustness, with an accuracy of 0.9571 and an F1-score of 0.9542 for distinguishing seven similar *Panax ginseng* species. After feature selection by 25 different feature ranking techniques in combination with prior knowledge, we obtained 100% accuracy and an F1-score for discriminating *P. ginseng* species. Furthermore, HerbMet exhibits acceptable inference speed and computational costs compared to existing approaches on both CPU and GPU.

Yuyang Sha, Meiting Jiang, and Gang Luo contributed equally to this work.

Conclusions: HerbMet surpasses existing solutions for identifying Chinese herbal medicines species. It is simple to use in real-world scenarios, eliminating the need for feature ranking and selection in classical machine learning-based methods.

KEYWORDS
Chinese herbal medicines, deep learning, *Gleditsia sinensis*, metabolomics, *Panax ginseng*

1 | INTRODUCTION

Chinese herbal medicines have a thousand-year-long history of clinical use in treating various diseases.[1] The most commonly used Chinese herbal medicines include *Panax ginseng*, *Gleditsia sinensis*, and *Akebiae Caulis*. Chinese herbal medicines have varying species and compositions. Using *P. ginseng* as an example, we illustrate the challenges and difficulties of identifying Chinese herbal medicines. *Panax* species (Araliaceae), such as *P. ginseng*, *P. quinquefolius*, and *P. notoginseng*, are famous worldwide for their remarkable tonifying effects and extensively consumed as health-care products, functional foods, and cosmetics.[2] Ginsenosides are the primary bioactive components of *P. ginseng* and have been shown to process anti-tumor, anti-inflammation, antioxidant, and anti-fatigue properties.[3] The different *Panax* species closely resemble each other in terms of microscopic features and chemical composition. However, the quality and effectiveness of the same *Panax* species may vary depending on several factors, such as plant growth conditions and processing procedures. Since the medicinal ingredients between various *Panax* species are similar, it is difficult to discriminate the different varieties. Therefore, designed a high-performance and robustness method for accurate identifying Chinese herbal medicines is crucial for their proper use.

Identifying the species of Chinese herbal medicines is difficult because they usually contain complex and similar bioactive components.[4] The traditional methods for classifying Chinese herbal medicines, as recorded in the Chinese Pharmacopoeia, are primarily based on characteristics, shape, microscope examination, and physicochemical properties. However, these traditional methods must face numerous challenges, such as complex sample pretreatment, susceptibility to environmental influences, and difficulties in rapid identification.[5] With the assistance of advanced omics technologies, such as liquid chromatography–mass spectrometry (LC-MS) and gas chromatography–mass spectrometry (GC-MS), several datasets for metabolomics of Chinese herbal medicines are available. Consequently, data-driven approaches for species identification have gradually become predominant in this field. Some studies[6–8] report that the LC-MS is one of the most analytical techniques for separating and characterizing multicomponent Chinese herbal medicines. The untargeted metabolomics approach is commonly used in research to comprehensively analyze all measurable analytes in Chinese herbal medicines.[9] However, the untargeted metabolomics approach pays more attention to obtaining non-biased data, which neither provides high-quality nor highly relevant datasets. Targeted metabolomics can provide sensitivity and specificity data for known compounds in the provided samples.[10] Recently, the pseudo-targeted metabolomics strategy, which integrates the advantages of both untargeted and targeted methods, has been widely used for metabolomics differential analysis, especially in identifying Chinese herbal medicines.[11] Considering the intricate nature of bioactive components in Chinese herbal medicines, gathering a substantial amount of metabolomics data for model training proves costly. Consequently, many Chinese herbal metabolomics datasets encounter various obstacles, including high data dimensionality and small sample sizes, which impede the development of effective identification methods.

Accurate identification of Chinese medicinal materials, especially different types of the same genus, is an important prerequisite for research and practical application.[12] Due to the significant strength of omics-related technologies, many advanced identification methods have adopted data-driven modeling schemes to complete analysis tasks.[13] These methods can be broadly categorized into machine learning-based and deep learning-based solutions. Establishing accurate identification models for Chinese herbal medicines using machine learning techniques and metabolomics data is a meaningful research direction that has gained considerable attention. For instance, Zhan et al.[14] proposed a novel classification system using support vector machine (SVM) and a self-assembled electronic nose framework for Chinese herbal medicines. This method was evaluated on 12 categories of herbal medicines and achieved promising results. Wang et al.[15] applied the self-organization map to classify mixtures of Chinese herbal medicines, obtaining better accuracy results for 59 types of herbal medicines. Wang et al.[16] used a combination of principal component analysis (PCA) and back-propagation artificial neural network (BP-ANN) to create a classification model for identifying three types of Chinese herbal medicines. The results showed that this combined method outperformed SVM and linear discriminant analysis (LDA). In addition, Ji et al.[17] developed a classification model using five classical machine learning methods, including ERT, XGBoost, and MLogit. Their findings indicated that tree-based methods generally perform better in identifying Chinese herbal medicines. In addition, certain studies have attempted to incorporate advanced machine learning techniques like RF, CatBoost, and KNN to classify Chinese herbal medicines. Nevertheless, most machine learning methods need to conduct feature ranking before producing final prediction results. This process may lose some important information, potentially negatively impacting model performance.

In recent years, deep learning methods have become the primary approach in computer vision,[18,19] natural language processing,[20] and data mining.[21] Research showed that deep neural networks have been successfully used to identify Chinese herbal medicines, producing promising results. For example, Liu et al.[22] developed an automatic classification framework for Chinese herbal medicines using deep neural networks. The method demonstrated promising performance but was limited to processing image samples and not sequence data. Chen et al.[23] introduced the S-TextBLCNN model, which included several Bi-LSTM modules and achieved superior performance in Chinese herbal medicine classification. This project also offered a new perspective on the field by suggesting using NLP-related technologies in building classification models. Compared with classical identification methods, deep learning-based approaches can automatically identify the relationship between different features and have demonstrated significant advantages in high-dimensional omics data over machine learning methods.[24] Most deep learning models can be trained in an end-to-end manner, and we can obtain complex models in a simple way using raw input samples without any manual feature extraction. However, deep learning models require plenty of data to train effectively. Insufficient training samples may lead to severe overfitting, seriously affecting the algorithm's performance. Although these current classification methods can achieve high performance in classification, numerous challenges still need to be tackled. Remarkably, most methods only categorize herbal medicines, such as *P. ginseng* and *G. sinensis*. However, it is challenging to differentiate species within the same genus, such as *P. ginseng* and *P. quinquefolius*. *Besides*, these approaches also have limitations when applied to the analysis of Chinese herbal medicines, including model overfitting, complex interpretation, and poor reproducibility.

To address the above issues, we introduce HerbMet, a novel AI-based system for accurately identifying Chinese herbal medicines using deep learning and metabolomics data. Inspired by successful computer vision[25,26] and natural language processing[20,27] architectures, we design a 1D ResNet-like architecture to extract distinctive representations from input samples effectively. Subsequently, these features are mapped to generate prediction results via a multilayer perceptron (MLP). We also propose a double dropout regularization module (DDR) to mitigate overfitting. To demonstrate the advantages of the proposed model, we conduct several experiments between HerbMet and several widely used machine learning and deep learning algorithms on two metabolomics datasets. These datasets consist of seven types of *P. ginseng* and three varieties of *G. sinensis*. We also utilize 25 feature ranking techniques to select the most distinguishing features and use them to build a more accurate model. By integrating AI and metabolomics, HerbMet has the potential to revolutionize quality control, pharmacological research, and clinical practice in the field of Chinese herbal medicines, ultimately benefiting patients and advancing the field of herbal medicine research.

2 | MATERIALS AND METHODS

2.1 | Data collection

We utilize two metabolomics datasets on the roots of seven *P. ginseng* species[28] and the seeds of three *G. sinensis* varieties[29] to develop and evaluate our AI models. For detailed information on the protocols for metabolite extraction and analysis using LC-Q-TOF, please refer to the original cited publications. The *P. ginseng* dataset comprises 70 samples, including 7 species with 10 samples per species. Each sample involves 253-dimensional features, which can be represented as a 70×253 matrix. The *G. sinensis* dataset contains 45 samples divided into 3 categories, with 15 samples for each type. The feature dimension of the *G. sinensis* dataset is much higher than that of the *P. ginseng* dataset, with each sample containing 2,867-dimensional features. We define a $45 \times 2{,}867$ matrix to represent the collected *G. sinensis* dataset. The details of collected datasets are presented in Tables S1 and S2, and the raw metabolomics data can be found in Data S1 and S2.

2.2 | Data processing

Some studies[30,31] have demonstrated that data processing technologies can have a significant impact on model performance. To achieve an accurate and robust model, we have incorporated multiple data processing methods to address null values and outliers in the metabolomics dataset. First, we use the chained equations algorithm to fill in missing values. Second, we apply the Box-Cox algorithm to harmonize the data and reduce variations in distribution between different institutions. Following that, we standardize the data by using the min-max method. Finally, we address data imbalance with the adaptive synthetic sampling method while maintaining a specific balancing ratio.

2.3 | Proposed method

In this paper, we present HerbMet, an AI-based system designed to accurately identify Chinese herbal medicines with deep neural network and metabolomics data. The details of proposed system are shown in Figure 1, which comprises data acquisition, data preprocessing, data generation, and model analysis.

2.3.1 | Main architecture

In this study, we consider the task of Chinese herbal medicines identification as supervised learning problems. Therefore, we can use $D = \{(x_i, y_i)\}_{i=1}^{n}$ to represent the input dataset, where $x_i = \left(x_i^{num}, x_i^{cat}\right) \in \mathbb{X}$ defines numerical x_{ij}^{num} and categorical x_{ij}^{cat} features of an object and $y \in \mathbb{Y}$ denotes the corresponding object label.

Artificial intelligence has revolutionized numerous fields, including facial recognition, data mining, and machine translation. In the task of Chinese herbal medicines identification, deep learning-based methods have gained significant attention and are widely used.[32] Unlike traditional machine learning algorithms like SVM and RF, deep learning methods can automatically extract distinctive features from input samples and use them to develop an efficient classification or regression model. Deep neural networks can leverage large amounts of training data and advanced hardware devices, providing significant

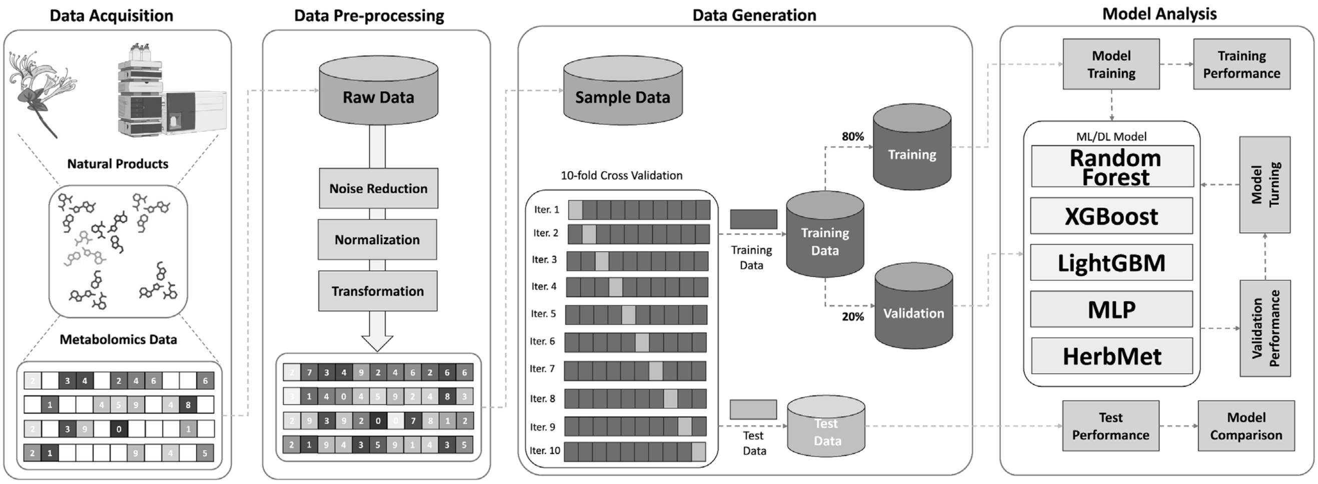

FIGURE 1 The overall study flow for this work. The metabolomic analysis of the rootsfor seven species of *Panax ginseng* species and the seeds of 3 *Gleditsia sinensis* herbalmedicines were described as in the data source of the method section. The metabolomics peak area data from the chromatographic features were pre-processed by noise reduction, normalization, and transformation. The dataset was then randomly separated into training (80%), validation (10%), and test (10%) sets. The performance of our developed HerbMet model was compared against 10 commonly used machine learning and deep learning models.

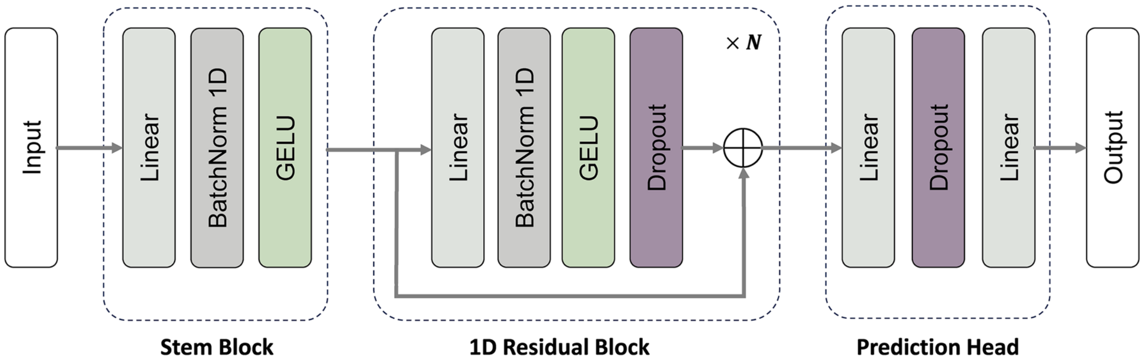

FIGURE 2 Architecture of the proposed 1D ResNet-like structure. The structure contains three main components: the stem block as the starting point, followed by new designed repeating 1D residual blocks (×*N*), and the prediction head as the final output. [Colour figure can be viewed at wileyonlinelibrary.com]

advantages over machine learning systems in model performance, training protocols, and inference speed.

We aim to develop an efficient system for identifying various species of Chinese herbal medicines via metabolomics data. Usually, most Chinese herbal medicines metabolomics datasets contain three characteristics: high dimensionality, small sample size, and similar biomarkers. Therefore, these may be challenging for existing methods to produce promising results. In order to overcome these challenges, we propose HerbMet, which is designed to learn semantic knowledge from these complex metabolomics datasets. The proposed system consists of two parties: 1D ResNet-like structure and DDR module.

Previous studies[33,34] have indicated that many deep learning-based sequence data analysis solutions are built with MLP. These methods usually perform well in speed and feature representation learning. However, MLP-based approaches may struggle when dealing with complex datasets. Inspired by several works[25,26] in computer vision, we design a 1D ResNet-like structure for analyzing 1D sequence data using Linear Layer. The proposed structure can be described as Equation (1), mainly composed of ResBlock, Linear Layer, and GeLU.[35] Figure 2 shows the detailed components of our introduced 1D ResNet-like structure. Compared with the standard ResNet, we replace Conv2D with Linear Layer. The simplified main building block is beneficial for optimization and creating a clear path from input to output.

$$\begin{aligned} ResNet(x) &= Prediction(ResBlock(\ldots(ResBlock(Linear(x))))) \\ ResBlock(x) &= x + Drop(Linear(Drop(GeLU(Linear(BatchNorm(x)))))) \\ Prediction(x) &= Linear(GeLU(BatchNorm(x))) \end{aligned} \tag{1}$$

2.3.2 | Double dropout regularization module

It is well known that deep learning models are susceptible to overfitting, especially when dealing with high feature dimensionality and small sample sizes in many Chinese herbal medicines metabolomics datasets.[36] Therefore, we develop a simple and effective regularization method, named DDR, based on standard dropout technology[37,38] to address this issue. During the model training phase, the standard dropout function randomly drops some units in each neural network layer, which helps prevent overfitting and optimize model performance. This technique does not affect the model's inference speed and computational cost. However, standard dropout may lead to discrepancies between model training and testing, which can have a negative impact on the model's performance. In the task of identifying Chinese herbal medicines, slight differences can significantly impact the model's performance. Therefore, we use DDR to boost the model accuracy, by minimizing the negative log-likelihood loss function. The details of DDR are illustrated in Figure 3.

In the training stage, we feed the input sample x_i into the proposed DDR. Unlike the normal training protocol, x_i would go through the forward pass of the neural network twice. Therefore, we can obtain two distributions of the model prediction for the input sample of x_i, which can be defined as $f_1(y_i|x_i)$ and $f_2(y_i|x_i)$. Due to the characteristics of dropout, the distributions of $f_1(y_i|x_i)$ and $f_2(y_i|x_i)$ for the same input sample (x_i, y_i) are different. Therefore, our proposed DDR attempts to regularize the model predictions by minimizing the bidirectional Kullback–Leibler (KL) divergence between these two predicted distributions for the same sample during the training stage. The $\mathcal{L}_{DDR}$ is defined as follows:

$$\mathcal{L}_{DDR} = \frac{1}{2}\left(D_{kl}\left(f_1(y_i|x_i)\middle\|f_2(y_i|x_i)\right) + D_{kl}\left(f_2(y_i|x_i)\middle\|f_1(y_i|x_i)\right)\right), \tag{2}$$

where the D_{kl} denotes the loss function of Kullback–Leibler (KL) divergence.

It is important to note that the proposed DDR is only utilized during the model training phase. Therefore, it does not have any impact on the inference speed or computational costs.

2.3.3 | Loss function

Focal loss[39] is effective in classification tasks, especially in imbalanced scenarios. Therefore, we use focal loss instead of the CE loss during model training. The focal loss $\mathcal{L}_{FL}$ can be defined as follows:

$$\mathcal{L}_{FL} = -a_t(1-p_t)^{\gamma}\log(p_t), \tag{3}$$

where a_t is the balance parameter. And the p_t can be calculated as

$$p_t = \begin{cases} p & \mathit{if}\ y=1 \\ 1-p & \mathit{other} \end{cases}. \tag{4}$$

The $p \in [0,1]$ represents the model's predicted probability for the input sample.

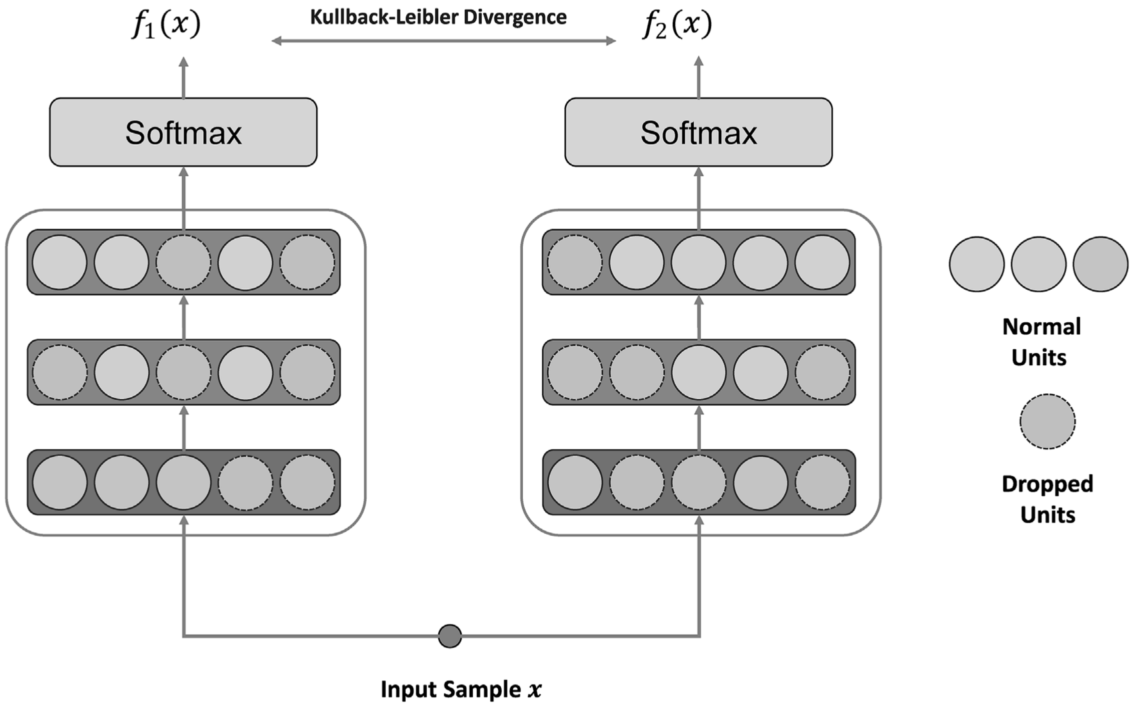

FIGURE 3 Structure of the proposed double dropout regularization (DDR) module for mitigating model overfitting. The DDR module consists of one dropout layers, which is applied twice during the training process. As the input sample passes through the neural network, it undergoes two forward passes, each with a different configuration of dropped units in the dropout layer. This results in two distinct distributions of model predictions for the input sample. The DDR aims to minimize the bidirectional Kullback–Leibler divergence between these two predicted distributions to effectively reduce overfitting and enhance model robustness.

The final training objective is to minimize the $\mathcal{L}$ for input samples. It can be defined as follows:

$$\mathcal{L}=\mathcal{L}_{FL}+\gamma\mathcal{L}_{DDR}, \tag{5}$$

where the γ is the coefficient weight to control the importance of $\mathcal{L}_{DDR}$.

2.4 | Implementation details

We use the Adam optimizer and the cosine learning rate scheduler during the training stages. Our proposed HerbMet is trained from scratch on one Nvidia 4090 GPU with a batch size of 64 and 60 epochs. The initial learning rate is set to 0.0003, and we use momentum of 0.9 along with weight decay. To reduce the impact of partition randomness on the obtained results, we adopt tenfold cross-validation to verify the prediction performance of the proposed model. All deep learning-based approaches are built with PyTorch 1.11 and Python 3.10.

2.5 | Evaluation metric

We evaluate model performance using mainstream metrics like accuracy, precision, recall, and F1 score. Definitions of these metrics are provided in Equations (6)–(9).

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN}, \tag{6}$$

$$Precision=\frac{TP}{TN+FP}, \tag{7}$$

$$Recall=\frac{TP}{TP+FN}, \tag{8}$$

$$F1-Score=\frac{2\times Precision\times Recall}{Precision+Recall}. \tag{9}$$

The abbreviation *TP* stands for "true positive," which means that the sample is in the positive category and has been correctly classified. On the other hand, *FN*, or "false negative," refers to the sample being in the positive category but is predicted to be negative. Similarly, *TN* and *FP* represent "true negatives" and "false positives," respectively. The definitions of *TN* and *FP* are similar to *TP* and *FN*, as mentioned above.

3 | RESULTS AND DISCUSSION

3.1 | Main results

In this section, we compare HerbMet with several classic machine learning-based approaches, such as SVM, Random Forest (RF), XGBoost,[40] LightGBM,[41] and CatBoost.[42] We build an MLP-based method to serve as the benchmark model for the deep learning approach. The LSTM and Conv1D algorithms are also used as comparison methods for identifying Chinese herbal medicines. Deep learning-based methods have significant advantages in feature selection compared with machine learning approaches. To ensure a fair comparison, we design two experiments to evaluate the model performance. First, all methods are trained and evaluated on the raw dataset. Second, we conduct experiments on feature-selected datasets formed by the top 20 features. It is important to mention that the top 20 features are identified through various feature selection methods and practical experience. The comparison results for the raw metabolomics datasets of *P. ginseng* and *G. sinensis* are presented in Tables 1 and 2, respectively. According to the results, HerbMet demonstrates better performance in identifying Chinese herbal medicines compared to existing machine learning and deep learning methods. For instance, when comparing HerbMet with Decision Tree on the raw *P. ginseng* dataset, HerbMet achieves a relative improvement of 25.62% and 17.88% on accuracy and F1 score, respectively. Furthermore, as depicted in Table 2, HerbMet can achieve 100% prediction accuracy on the raw *G. sinensis* dataset, indicating a significant improvement over existing methods.

Tables S3 and S4 illustrate the experimental results of HerbMet and other comparison methods on the feature-selected datasets of *P. ginseng* and *G. sinensis*. It is clear that our proposed HerbMet still outperforms the existing methods in predicted accuracy. For instance, HerbMet demonstrates a relative improvement over Random Forest with 16.67% and 13.30% in accuracy and F1 score, respectively. When we compare Tables 1 and S3, we can find that the performance of all methods has been improved significantly in the feature-selected dataset. Classic machine learning-based methods have been considerably optimized, while the performance of deep learning-based approaches minimal changes between the raw dataset and the feature-selected one. The results also imply that our proposed HerbMet can omit feature ranking, simplifying the operation process for Chinese herbal medicines identification.

3.2 | Effectiveness of DDR module

In this paper, we introduce a regularization function named DDR, which is primarily used to reduce overfitting and enhance model performance. To explore the effectiveness of the proposed module, we conduct several ablation experiments on the collected dataset. The results on the raw *P. ginseng* dataset are presented in Table 3. As the results show, our proposed DDR can optimize the model performance with a significant margin. For instance, with the help of DDR, HerbMet can achieve 10.48% and 10.15% relative improvements in accuracy and F1 score, respectively. We also carry out similar experiments on the *G. sinensis* dataset, which is shown in Table S5. The scale of most Chinese herbal medicine metabolomics datasets is small due to high collection costs and complex processing. Therefore, numerous analysis models in this field often struggle with overfitting due to

TABLE 1 Performance comparison of HerbMet and 10 machine learning and deep learning methods for distinguishing seven *Panax ginseng* species using all collected metabolomics features

	All collected metabolomics features from *P. ginseng*			
Machine learning/deep learning models	**Accuracy**	**F1 score**	**Precision**	**Recall**
Logistics regression	0.8571 ± 0.02	0.8603 ± 0.03	0.8571 ± 0.02	0.9238 ± 0.03
Random forest	0.8571 ± 0.01	0.8095 ± 0.02	0.8333 ± 0.03	0.875 ± 0.02
KNN	0.8095 ± 0.03	0.8095 ± 0.04	0.8857 ± 0.02	0.8857 ± 0.01
Decision tree	0.7619 ± 0.02	0.8095 ± 0.01	0.8095 ± 0.02	0.8428 ± 0.02
SVM	0.8571 ± 0.03	0.8989 ± 0.03	0.9285 ± 0.02	0.9166 ± 0.01
XGBoost	0.8095 ± 0.01	0.8102 ± 0.02	0.8333 ± 0.03	0.8433 ± 0.02
LightGBM	0.8095 ± 0.02	0.8272 ± 0.01	0.8333 ± 0.03	0.8714 ± 0.03
CatBoost	0.9047 ± 0.02	0.8214 ± 0.01	0.8571 ± 0.04	0.7999 ± 0.01
MLP	0.8857 ± 0.02	0.8837 ± 0.03	0.9381 ± 0.03	0.8856 ± 0.04
Conv1D	0.8571 ± 0.01	0.8810 ± 0.02	0.8571 ± 0.02	0.8524 ± 0.03
LSTM	0.9286 ± 0.03	0.9524 ± 0.04	0.9286 ± 0.03	0.9238 ± 0.02
HerbMet	**0.9571** ± 0.01	**0.9542** ± 0.03	**0.9714** ± 0.01	**0.9571** ± 0.02

Note: Higher values indicate better performance. Bold denotes the best performance. Data are represented as mean ± SD.

TABLE 2 Performance comparison of HerbMet and 10 machine learning and deep learning methods for distinguishing three *Gleditsia sinensis* species using all collected metabolomics features

	All collected metabolomics features from G. *sinensis*			
Machine learning/deep learning models	**Accuracy**	**F1 score**	**Precision**	**Recall**
Logistics regression	0.7778 ± 0.02	0.8055 ± 0.01	0.8667 ± 0.01	0.8333 ± 0.01
Random forest	0.7778 ± 0.01	0.7222 ± 0.01	0.7778 ± 0.03	0.8333 ± 0.02
KNN	0.6667 ± 0.01	0.5757 ± 0.02	0.5238 ± 0.02	0.6667 ± 0.03
Decision tree	0.7778 ± 0.02	0.6000 ± 0.02	0.5556 ± 0.01	0.6667 ± 0.01
SVM	0.8889 ± 0.03	0.8666 ± 0.03	0.8889 ± 0.01	0.8889 ± 0.03
XGBoost	0.8889 ± 0.01	0.8222 ± 0.02	0.8889 ± 0.01	0.8333 ± 0.01
LightGBM	0.7778 ± 0.02	0.7500 ± 0.01	0.7500 ± 0.01	0.7500 ± 0.01
CatBoost	0.7778 ± 0.01	0.8056 ± 0.02	0.8333 ± 0.01	0.8667 ± 0.02
MLP	0.8889 ± 0.01	0.8667 ± 0.01	0.8889 ± 0.02	0.8889 ± 0.02
Conv1D	0.8889 ± 0.01	0.8889 ± 0.02	0.8889 ± 0.02	0.8889 ± 0.01
LSTM	**1.0** ± 0.02	**1.0** ± 0.01	**1.0** ± 0.02	**1.0** ± 0.02
HerbMet	**1.0** ± 0.01	**1.0** ± 0.01	**1.0** ± 0.01	**1.0** ± 0.01

Note: Higher values indicate better performance. Bold denotes the best performance. Data are represented as mean ± SD.

TABLE 3 Performance of the double dropout regularization (DDR) module in HerbMet for distinguishing seven *Panax ginseng* species using all collected metabolomics features or the top 20 selected features

Dataset	Type	Accuracy	F1 score	Precision	Recall
All metabolomics features	*wo.* DDR	0.8571	0.8571	0.9286	0.8571
	***w.* DDR**	**0.9571**	**0.9542**	**0.9714**	**0.9571**
Top 20 selected features	*wo.* DDR	0.9286	0.9238	0.9524	0.9286
	***w.* DDR**	**1.0**	**1.0**	**1.0**	**1.0**

Note: The 20 features were selected based on 25 feature ranking methods in combination with prior knowledge. These features include C207, C202, C75, C32, C6, C10, C71, C50, C36, C138, C110, C93, C189, C187, C186, C46, C111, C201, C102, and C190 (see Data S1 for details about these compounds). Abbreviations: *w.* DDR, HerbMet with DDR; *wo.* DDR, HerbMet without DDR.

limited sample size. Experimental results demonstrate that the proposed DDR can tackle this issue effectively. In addition, DDR is only used in the training phase, so it does not impact the model's operating efficiency and inference speed.

3.3 | Effectiveness of backbone architecture

Generally, we often prioritize the model's accuracy over algorithm efficiency and inference speed. Nevertheless, algorithm efficiency is critical when implementing models in real-world projects. Deep learning models with substantial parameters and high computational costs typically deliver superior performance, but they are inefficient and demand advanced hardware for deployment. Therefore, we conduct experiments to analyze the algorithm's operating efficiency and inference speed, aiming to demonstrate that our proposed method is suitable for real-world scenarios. We compared the proposed 1D ResNet-like structure with classic deep learning-based solutions, such as VGG[43] and ResNet.[26] The results are presented in Table 4, which shows that the proposed backbone architecture offers significant advantages in terms of model parameters and computational costs.

TABLE 4 Comparison of model parameters and computational cost for different deep learning algorithms

Models	Parameters (M)	MACs (M)
ResNet-50	25.557	33.977
MobileNetv2	3.505	4.176
ShuffleNetv2	2.279	2.497
EfficientNet-B0	5.289	5.994
MLP	0.683	0.228
HerbMet	1.102	0.337

Abbreviation: MAC, multiply-accumulate operations.

Based on the findings in Section 3.2 and Table 4, the proposed HerbMet demonstrates the ability to achieve high accuracy in classifying Chinese herbal medicines. Furthermore, it offers significant advantages in model operation efficiency and computational costs. This method could potentially serve as a primary approach to further the development of Chinese herbal medicines analysis.

3.4 | Evaluation of inference speed

In practical scenarios, it is crucial to consider the inference speed of the algorithm. Therefore, we compare the running speed of classic machine learning methods and deep learning methods on CPU (Intel Xeon Gold 6338 @ 2.00GHz) and GPU (Nvidia GTX 4090). The experiment results are presented in Table 5. Specifically, most classic machine learning methods (such as SVM and Random Forests) are challenging to run on advanced hardware such as GPU or NPU. Therefore, we only verify the machine learning algorithm on the CPU. Meanwhile, the proposed MLP and HerbMet are evaluated using CPU and GPU. We hope that the proposed HerbMet can be used in most practical scenarios, so we select an entry-level GPU device for model training and testing instead of employing expensive professional devices, such as Nvidia A100 or H100. We perform 10,000 tests for each algorithm and calculate the average time for the final result. Initially, we use a single sample as an input to evaluate the model processing speed. We observe that the proposed HerbMet performs faster inference than Random Forest and CatBoost. With the assistance of GPU, HerbMet's inference speed can be further improved. Fortunately, modern computing devices can handle complex parallel computing tasks efficiently. Thus, we conduct extensive experiments to verify the parallel computing efficiency of the proposed method. All algorithms are required to predict 1,000 samples simultaneously. Table 5 shows that the proposed HerbMet takes only 0.435 ms to yield the final results, which is 136X faster than KNN. Although our

TABLE 5 Comparison of inference speed between different machine learning and deep learning methods on CPU and GPU

	Batch size = 1		Batch size = 1,000	
	CPU (ms)	GPU (ms)	CPU (ms)	GPU (ms)
Logistics regression	0.089	-	7.470	-
Random forest	3.072	-	12.659	-
KNN	1.190	-	59.249	-
Decision tree	0.096	-	5.722	-
SVM	0.121	-	15.439	-
XGBoost	0.383	-	6.869	-
LightGBM	0.903	-	7.658	-
CatBoost	1.410	-	52.270	-
MLP	0.161	0.096	2.100	0.138
HerbMet	0.620	0.251	3.594	0.435

Note: The CPU used for testing was Intel Xeon Gold 6338 @ 2.00GHz, and the GPU was Nvidia GTX 4090. All of the results were mean values after running 10,000 times. Batch size = 1 defines the model using one sample to get the final result. Batch size = 1,000 represents the model predicting 1,000 results simultaneously.

method is based on a deep neural network, its inference speed surpasses most classic machine learning methods. The advantages of HerbMet should become more pronounced when deployed on a more powerful GPU or NPU device.

3.5 | Discussion

In recent years, AI-related technologies have gained significant attention. Significant developments in large language and multimodal vision models have recently revolutionized our knowledge of deep learning. Advanced AI technology has found widespread application in various fundamental science and interdisciplinary fields, such as bioinformatics, computational chemistry, and artificial intelligence drug discovery. In this article, we discuss the use of advanced artificial intelligence techniques for identifying Chinese herbal medicines. Our approach involves taking metabolomics data as input and analyzing it using a deep neural network model. This allows us to quickly and accurately identify the species of Chinese herbal medicines. Unlike classic machine learning solutions, our approach eliminates the need for complex pre-processing steps, such as feature ranking. The proposed HerbMet relies on deep neural networks to extract discriminate representations directly from complex data, optimizing the algorithm's performance and simplifying the analysis steps. Although HerbMet shows promising results on two Chinese herbal medicines metabolomics datasets, multiple factors may impact the model's performance. Therefore, we conducted extensive experiments from various perspectives to comprehensively analyze the model structure and feature selection.

Various studies[37,44] have highlighted that overfitting can seriously impact the performance of deep learning models. The metabolomics datasets for Chinese herbal medicine typically have small sample sizes because of high collection costs and complex processing steps. As a result, model overfitting is a more challenging problem in Chinese herbal medicines analysis. In order to address the problem, we develop the DDR module as a regularization function to enhance the algorithm's performance. We conduct a series of experiments on the raw *P. ginseng* dataset to compare the effectiveness of the DDR module and other regularization methods such as standard Dropout, Drop Connect,[45] and Ada-Dropout.[44] Results are shown in Table 6. Notably, the baseline model does not use any regularization method in the model training stage. Our proposed DDR module can lead to a significant improvement in model performance compared to other regularization techniques. The results show that DDR has demonstrated a relative improvement of 11.67% in accuracy and 11.71% in F1 score compared to the baseline model.

Based on the results from Section 3.1, it is evident that most methods perform better when using the feature-selected dataset rather than the raw one. Can we identify a more representative set of features and use them to create a powerful classification model for Chinese herbal medicines analysis? To answer this question, we design comprehensive experiments to investigate these features. In this part, we utilize 25 feature selection models to determine the top 20 most important features for *P. ginseng* and *G. sinensis* datasets. Details of the feature selection methods can be found in Table S7. The top 20 important features of the *P. ginseng* dataset are depicted in Figure S1A. Moreover, experiments are conducted to access the contribution of these ranked features to the final results. Table 7 reports the comparison results of HerbMet in the *P. ginseng* dataset. Specifically, "Feature Top 1" refers to the model performance using the top 1 feature, whereas "Feature Top 1∼2" elucidates the prediction outcome of HerbMet based on the top 1 and 2 features. From the results, we can see that our proposed HerbMet is able to achieve 100% accuracy in classifying the species using only the top 14 features for the *P. ginseng* identification task. A similar experiment is carried out on the *G. sinensis* dataset, resulting in a 100% prediction accuracy based on the top 4 features. The results are presented in Figure S1B and Table S6. Tables S8 and S9 display the top 20 ranked features of the *P. ginseng* and *G. sinensis* datasets by different feature selection methods. Based on the findings, we can create a more efficient classification model by incorporating representative features. The proposed HerbMet model can accurately identify Chinese herbal medicines with a limited number of selected features. Consequently, the measurement of a large amount of metabolomics data in practical settings can be avoided, thereby simplifying the process of classifying Chinese herbal medicines. Furthermore, this approach can facilitate the development of AI-related techniques in this field.

3.6 | Limitations and future works

The method presented in this paper exhibits superior performance compared to traditional methods. Nonetheless, it still faces many obstacles, such as structural design, feature processing, and model

TABLE 6 Comparison of the double dropout regularization module (DDR) with other common regularization algorithms in HerbMet for distinguishing seven *Panax ginseng* species using all collected metabolomics features without feature selection

	Performance metrics for distinguishing seven different *P. ginseng* species			
Dropout methods	Accuracy	F1 score	Precision	Recall
Baseline	0.8571	0.8524	0.8810	0.8571
Standard Dropout	0.9286	0.9238	0.9524	0.9286
DropConnect	0.8571	0.8571	0.8571	0.8571
Ada-Dropout	0.9286	0.9238	0.9524	0.9286
DDR	**0.9571**	**0.9542**	**0.9714**	**0.9571**

Note: Higher is better. Bold denotes the best performance.

TABLE 7 Performance comparison of metabolomic feature numbers used in HerbMet for distinguishing seven *Panax ginseng* species

	Performance metrics for distinguishing seven different *P. ginseng* species			
Number of metabolomic features used in HerbMet	Accuracy	F1 score	Precision	Recall
Top 1	0.5000	0.4646	0.4762	0.5714
Top 1∼2	0.5000	0.5190	0.5238	0.5714
Top 1∼3	0.7143	0.5857	0.5238	0.7143
Top 1∼4	0.7857	0.7129	0.7738	0.7143
Top 1∼5	0.7857	0.7143	0.7857	0.7143
Top 1∼6	0.7857	0.7177	0.7500	0.7143
Top 1∼7	0.8571	0.8000	0.7619	0.8571
Top 1∼8	0.8571	0.8000	0.7619	0.8571
Top 1∼9	0.9286	0.8286	0.8095	0.8571
Top 1∼10	0.9286	0.8286	0.8095	0.8571
Top 1∼11	0.8571	0.8367	0.8929	0.8571
Top 1∼12	0.9286	0.9238	0.9524	0.9286
Top 1∼13	0.9286	0.9238	0.9524	0.9286
Top 1∼14	0.9286	0.9238	0.9524	0.9286
Top 1∼15	1.0	1.0	1.0	1.0
Top 1∼16	1.0	1.0	1.0	1.0
Top 1∼17	1.0	1.0	1.0	1.0
Top 1∼18	1.0	1.0	1.0	1.0
Top 1∼19	1.0	1.0	1.0	1.0
Top 1∼20	1.0	1.0	1.0	1.0

Note: Top 1 defines HerbMet using only the top 1 feature to build the model. Similarly, Top 1∼10 indicates HerbMet is trained and evaluated using the top 10 features. The 20 features were selected based on 25 feature ranking methods in combination with prior knowledge. These features include C207, C202, C75, C32, C6, C10, C71, C50, C36, C138, C110, C93, C189, C187, C186, C46, C111, C201, C102, and C190 (see Data S1 for details about these compounds).

deployment. For example, even though the HerbMet shows faster inference speeds than current methods, it mostly relies on advanced hardware. Nevertheless, some particular operators might not be supported by embedded devices, which could make it more difficult to use the suggested approach in real-world projects. Furthermore, HerbMet only evaluate on two Chinese herbal medicines datasets. Consequently, we have to use additional datasets to comprehensively verify the model performance.

In light of this, we will do follow-up work from the perspectives of model design and data gathering. In terms of model design, we will continue to improve model performance and regularization capabilities by optimizing the 1D ResNet-like architecture and DDR module. For example, we may include the Cross Attention Module and Transformer Block into the backbone to increase the feature extraction capability. Regarding data collection, we plan to collect more Chinese herbal medicine metabolomics datasets and verify the generalization and robustness of HerbMet on various datasets. Furthermore, we intend to launch an online platform that offers feature selection, model training, and species identification for Chinese herbal medicines.

4 | CONCLUSION

This paper presents a novel framework, called HerbMet, which applies deep neural networks and metabolomics data to accurately identify species of Chinese herbal medicines. We evaluate the proposed method's performance on the *P. ginseng* and *G. sinensis* datasets. The results demonstrate that the HerbMet outperforms previous solutions significantly. To address the issue of model overfitting in the identification of Chinese herbal medicine species, we developed a novel regularization module named DDR, which can improve model performance without reducing inference speed. Furthermore, we employ several popular feature ranking methods to identify some important features, which is beneficial to constructing efficient classification models.

ACKNOWLEDGMENTS

This work was supported by the fund from Macao Polytechnic University (RP/FCA-14/2023), Science and Technology Development Funds (FDCT) of Macao (0033/2023/RIB2), and Tianjin Outstanding Youth Fund (23JCJQJC00030).

DATA AVAILABILITY STATEMENT

The datasets and source code used and/or analyzed during the current study are available through GitHub (https://github.com/syysha0k/HerbMet) or Zenodo (https://zenodo.org/records/13117936) for research purpose only.

ORCID

Kefeng Li https://orcid.org/0000-0002-7233-4347

REFERENCES

1. Zhu Y, Ouyang Z, Du H, et al. New opportunities and challenges of natural products research: when target identification meets single-cell multiomics. *Acta Pharm Sin B*. 2022;12(11):4011-4039. doi:10.1016/j.apsb.2022.08.022
2. Li X, Liu J, Zuo TT, et al. Advances and challenges in ginseng research from 2011 to 2020: the phytochemistry, quality control, metabolism, and biosynthesis. *Nat Prod Rep*. 2022;39(4):875-909. doi:10.1039/d1np00071c
3. Yao W, Guan Y. Ginsenosides in cancer: a focus on the regulation of cell metabolism. *Biomed Pharmacother*. 2022;156:113756. doi:10.1016/j.biopha.2022.113756
4. Zhou X, Seto SW, Chang D, et al. Synergistic effects of Chinese herbal medicine: a comprehensive review of methodology and current research. *Front Pharmacol*. 2016;7:201. doi:10.3389/fphar.2016.00201
5. Shen T, Li W, Zhang X, et al. High-sensitivity determination of nutrient elements in panax notoginseng by laser-induced breakdown spectroscopy and chemometric methods. *Molecules*. 2019;24(8):1525. doi:10.3390/molecules24081525
6. Yang KY, Lin LC, Tseng TY, Wang SC, Tsai TH. Oral bioavailability of curcumin in rat and the herbal analysis from Curcuma longa by LC-MS/MS. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2007;853(1–2):183-189. doi:10.1016/j.jchromb.2007.03.010
7. Banerjee S, Kar A, Mukherjee PK, Haldar PK, Sharma N, Katiyar CK. Immunoprotective potential of ayurvedic herb Kalmegh (Andrographis paniculata) against respiratory viral infections—LC-MS/MS and network pharmacology analysis. *Phytochem Anal*. 2021;32(4):629-639. doi:10.1002/pca.3011
8. Caldeirão L, Sousa J, Nunes LCG, Godoy HT, Fernandes JO, Cunha SC. Herbs and herbal infusions: determination of natural contaminants (mycotoxins and trace elements) and evaluation of their exposure. *Food Res Int*. 2021;144:110322. doi:10.1016/j.foodres.2021.110322
9. Tang C, Li X, Wang T, et al. Characterization of metabolite landscape distinguishes medicinal fungus cordyceps sinensis and other cordyceps by UHPLC-Q Exactive HF-X untargeted metabolomics. *Molecules*. 2023;28(23):7745. doi:10.3390/molecules28237745
10. Cui S, Li K, Ang L, et al. Plasma phospholipids and sphingolipids identify stent restenosis after percutaneous coronary intervention. *JACC Cardiovasc Interv*. 2017;10(13):1307-1316. doi:10.1016/j.jcin.2017.04.007
11. Yang F, Chen B, Jiang M, et al. Integrating enhanced profiling and chemometrics to unveil the potential markers for differentiating among the leaves of *Panax ginseng*, *P. quinquefolius*, and *P. notoginseng* by ultra-high performance liquid chromatography/ion mobility-quadrupole time-of-flight mass spectrometry. *Molecules*. 2022;27(17):5549. doi:10.3390/molecules27175549
12. Chen S, Pang X, Song J, et al. A renaissance in herbal medicine identification: from morphology to DNA. *Biotechnol Adv*. 2014;32(7):1237-1244. doi:10.1016/j.biotechadv.2014.07.004
13. Muneer A, Fati SM. Efficient and automated herbs classification approach based on shape and texture features using deep learning. *IEEE Access*. 2020;8:196747-196764. doi:10.1109/ACCESS.2020.3034033
14. Zhan X, Guan X, Wu R, Wang Z, Wang Y, Li G. Discrimination between alternative herbal medicines from different categories with the electronic nose. *Sensors (Basel)*. 2018;18(9):2936. doi:10.3390/s18092936
15. Wang M, Li L, Yu C, et al. Classification of mixtures of Chinese herbal medicines based on a self-organizing map (SOM). *Mol Inform*. 2016;35(3–4):109-115. doi:10.1002/minf.201500115
16. Wang J, Liao X, Zheng P, Xue S, Peng R. Classification of Chinese herbal medicine by laser-induced breakdown spectroscopy with principal component analysis and artificial neural network. *Anal Lett*. 2018;51(4):575-586. doi:10.1080/00032719.2017.1340949
17. Ji X, Tong W, Liu Z, Shi T. Five-feature model for developing the classifier for synergistic vs. antagonistic drug combinations built by XGBoost. *Front Genet*. 2019;10:600. doi:10.3389/fgene.2019.00600
18. Sha Y, Zhai X, Li J, Meng W, Tong HH, Li K. A novel lightweight deep learning fall detection system based on global-local attention and channel feature augmentation. *Inter Nurs Res*. 2023;2(2):68-75. doi:10.1097/NR9.0000000000000026
19. Sha Y, Meng W, Zhai X, Xie C, Li K. Accurate facial landmark detector via multi-scale transformer. In: *Pattern Recognition and Computer Vision*. Springer; 2024:278-290.
20. Li J, Li D, Savarese S, Hoi S. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning. PMLR; 2023:19730-19742.
21. Sha Y, Meng W, Luo G, et al. MetDIT: transforming and analyzing clinical metabolomics data with convolutional neural networks. *Anal Chem*. 2024;10-18. doi:10.1021/acs.analchem.3c04607
22. Liu S, Chen W, Dong X. Automatic classification of Chinese herbal based on deep learning method. In: *International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*. IEEE; 2018:235-238. doi:10.1109/FSKD.2018.8687165
23. Cheng N, Chen Y, Gao W, et al. An improved deep learning model: S-TextBLCNN for traditional Chinese medicine formula classification. *Front Genet*. 2021;12:807825. doi:10.3389/fgene.2021.807825
24. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform*. 2022;23(1):bbab454. doi:10.1093/bib/bbab454
25. Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Anal Mach Intell*. 2021;43(10):3349-3364. doi:10.1109/TPAMI.2020.2983686
26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Computer Vision and Pattern Recognition*. IEEE; 2016:770-778.
27. Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto TB. Benchmarking large language models for news summarization. *Trans Assoc Comput Linguist*. 2024;12:39-57. doi:10.1162/tacl_a_00632
28. Wang X, Jiang M, Lou J, et al. Pseudotargeted metabolomics approach enabling the classification-induced ginsenoside characterization and differentiation of ginseng and its compound formulation products. *J Agric Food Chem*. 2023;71(3):1735-1747. doi:10.1021/acs.jafc.2c07664
29. Xie H, Wang H, Chen B, et al. Untargeted metabolomics analysis to unveil the chemical markers for the differentiation among three *Gleditsia sinensis*-derived herbal medicines by ultra-high performance liquid chromatography/quadrupole time-of-flight mass spectrometry. *Arab J Chem*. 2022;15(5):103762. doi:10.1016/j.arabjc.2022.103762
30. Cai G, Huang F, Gao Y, et al. Artificial intelligence-based models enabling accurate diagnosis of ovarian cancer using laboratory tests in China: a multicentre, retrospective cohort study. *Lancet Digit Health*. 2024;6(3):e176-e186. doi:10.1016/S2589-7500(23)00245-5
31. Gao Y, Zeng S, Xu X, et al. Deep learning-enabled pelvic ultrasound images for accurate diagnosis of ovarian cancer in China: a

retrospective, multicentre, diagnostic study. *Lancet Digit Health*. 2022;4(3):e179-e187. doi:10.1016/S2589-7500(21)00278-8

32. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning--based text classification: a comprehensive review. *ACM Comput Surv (CSUR)*. 2021;54(3):1-40. doi:10.1145/3439726
33. Zhou K, Yu H, Zhao WX, Wen JR. Filter-enhanced MLP is all you need for sequential recommendation. *ACM Web Conf*. 2022;2388-2399. doi:10.1145/3485447.3512111
34. Li M, Zhang Z, Zhao X, et al. AutoMLP: automated MLP for sequential recommendations. *Proc ACM Web Conf*. 2023;1190-1198. doi:10.1145/3543507.3583440
35. Hendrycks D, Gimpel K. Gaussian error linear units (gelus). arXiv preprint arXiv:160608415.
36. Sha Y. Efficient facial landmark detector by knowledge distillation. *Int Conf Automat Face Gesture Recogn*. 2021;1-8.
37. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929-1958.
38. Liang X, Wu L, Li J, et al. R-drop: regularized dropout for neural networks. *Adv Neural Inform Process Syst*. 2021;34:10890-10905.
39. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(2):318-327. doi:10.1109/TPAMI.2018.2858826
40. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. *ACM Int Conf Knowl Discov Data Mining*. 2016;785-794.
41. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inform Process Syst*. 2017;30.
42. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inform Process Syst*. 2018;31.
43. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556.
44. Ba J, Frey B. Adaptive dropout for training deep neural networks. *Adv Neural Inform Process Syst*. 2013;26.
45. Wan L, Zeiler M, Zhang S, le Cun Y, Fergus R. Regularization of neural networks using dropconnect. In: *International Conference on Machine Learning*. PMLR; 2013:1058-1066.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sha Y, Jiang M, Luo G, et al. HerbMet: Enhancing metabolomics data analysis for accurate identification of Chinese herbal medicines using deep learning. *Phytochemical Analysis*. 2024;1-12. doi:10.1002/pca.3437