Efficient Recognition of Easily-confused Chinese Herbal Slices Images Using Enhanced ResNeSt

Qi Zhang[1], Jinfeng Ou[1†], and Huaying Zhou[1*]
[1] School of Medical Information and Engineering, Guangdong Pharmaceutical University, Guangzhou,510006, China
[e-mail: jackie.zhang@gdpu.edu.cn, 1911065513@qq.com, huayingzhou@gdpu.edu.cn]
[*]Corresponding author: Huaying Zhou
† Co-first author

Received February 14, 2024; revised April 29, 2024; revised June 3, 2024; accepted July 11, 2024; published August 31, 2024

Abstract

Chinese herbal slices (CHS) automated recognition based on computer vision plays a critical role in the practical application of intelligent Chinese medicine. Due to the complexity and similarity of herbal images, identifying Chinese herbal slices is still a challenging task. Especially, easily-confused CHS have higher inter-class and intra-class complexity and similarity issues, the existing deep learning models are less adaptable to identify them efficiently. To comprehensively address these problems, a novel tiny easily-confused CHS dataset has been built firstly, which includes six pairs of twelve categories with about 2395 samples. Furthermore, we propose a ResNeSt-CHS model that combines multilevel perception fusion (MPF) and perceptive sparse fusion (PSF) blocks for efficiently recognizing easily-confused CHS images. To verify the superiority of the ResNeSt-CHS and the effectiveness of our dataset, experiments have been employed, validating that the ResNeSt-CHS is optimal for easily-confused CHS recognition, with 2.1% improvement of the original ResNeSt model. Additionally, the results indicate that ResNeSt-CHS is applied on a relatively small-scale dataset yet high accuracy. This model has obtained state-of-the-art easily-confused CHS classification performance, with accuracy of 90.8%, far beyond other models (EfficientNet, Transformer, and ResNeSt, etc) in terms of evaluation criteria.

Keywords: ResNeSt, multilevel perception fusion (MPF), perceptive sparse fusion (PSF), easily-confused CHS

1. Introduction

Chinese herbal medicine (CHM) classification is an important research task in intelligent medicine, which has been applied widely in the fields of smart medicinal material sorting and medicinal material recommendation [1]. Chinese herbal slices (CHS), is an important part of traditional Chinese medicine [2], which can be directly used in clinical after special concocted processes from Chinese herbs. The traditional manual identification method mainly relies on human experience to determine the quality of Chinese herbals based on their shape, color, gas, taste, and other characteristics [3]. Apparently, this method has strong subjectivity and high labor costs, making it hard to achieve rapid detection in large quantities [4]. As a result, there have been many excellent studies on machine learning and deep learning methods for CHM automated recognition in recent years. According to preliminary statistics, there are about 1200 types of commonly used CHMs [5]. Due to a large number of Chinese herbals type, the identification becomes quite complicated. Therefore, many researchers try to explore how to connect the relationship between computer science and Traditional Chinese Medicine for a long time [6].

With the rapid development of digital image processing and machine vision, the recognition effect, to some extent, can be improved via computer image processing and pattern recognition technology [7]. Specifically, machine leaning mainly depends on extracting single image features of CHS for analysis and recognition, including size, color, edge features, morphology and texture [8-10]. These methods rely on less robust handcrafted features and are not feasible for classification on large-scale datasets [1]. That is to say, shallow models extract feature information directly from image pixels without high-level semantics, resulting in poor outcomes. Therefore, with the development of artificial intelligence technology, CHS image identification based on deep learning has become a development trend and has a good application prospect [11].

In this paper, we established a small dataset for easily-confused CHS. Furthermore, we have introduced a novel ResNeSt-CHS model, which builds upon the ResNeSt architecture and incorporates MPF and PSF blocks. The main contributions of this paper can be summarized as follows:

(1) Inspired by recent research, we proposed ResNeSt-CHS, which is an enhanced version of the standard ResNeSt model.

(2) We presented the MPF block which is responsible for capture global channel and spatial features, and the PSF block which is used for channel screening of the region with the largest receptive field by combining global channel features.

(3) We examined six pairs of easily-confused CHS with highly similar appearances, and observe a significant classification improvement in using ResNeSt-CHS over ResNeSt.

The remainder of this paper is organized into seven sections. Section 2 describes the related literature of deep learning research on CHS recognition. Section 3 provides a detailed introduction to easily-confused CHS and the motivations for this study. Section 4 presents the proposed method, including MPF and PSF design. Then the dataset is described in Section 5, while the experimental results are illustrated and discussed in Section 6. Finally, the conclusion of this paper is summarized.

2. Related Work

Convolutional neural networks (CNN) have enjoyed a great success in large-scale image and video recognition [12] in the past decades due to their ability of extracting deep discriminative

features. AlexNet [13] showed the good results by constructing multi-layer architecture. VGG [12] stands for Visual Geometry Group which is standard deep CNN architecture with stacked building blocks. ResNet [14] employed residual connections to CNN to alleviate the gradient vanish problem. Xception [15] obtained better generalization by combining a linear stack of depthwise separable convolution layers with residual connections. In addition, using transfer learning methods to transfer pre-trained CNN models to tasks such as image classification [16], object detection [17], and image segmentation significantly improved the performance of these tasks and achieved good results. For CHM classification, there exists many deep learning models, such as CNN [18-19], AlexNet [20], VGG16 [21], DenseNet [22], GoogLeNet [23-24], EfficientNet [25], ResNet50 [26], and more. However, the limited receptive field of small convolutional kernels makes it difficult to obtain global information, withholding the networks of high performance on challenging classification tasks [27]. Thus, more and more researchers focus on attention mechanism [28] fuse into the deep learning model in order to extract high-level semantics more effectively.

Since attention mechanisms can effectively integrate local and global features [29] and are widely used in clinical images classification tasks [30-32]. Xu et al. [33] presented a new Attentional Pyramid Networks with both competitive attention and spatial collaborative attention which obtain the more efficient fused Chinese herbal images features. Miao et al. [34] improved ConvNeXt [35] network for feature extraction by adding a stacked ACMix [36] to ensure the adequacy of low-dimensional feature extraction in the Chinese medicine images. However, very few studies have attempted to focus on easily-confused CHS classification. Most of the studies related to Chinese herbal image identification have concentrated on various categories with their own created the small-size dataset.

Our proposed method is based on the ResNeSt [37], which has undergone three stages of evolution. Firstly in 2016, He et al. [14] adopted the skip connections in ResNet to solve the problem of vanishing gradient in deep neural networks. Then, Xie et al. [38] presented a variant of ResNet, that is named ResNeXt. ResNeXt retains the original ResNet architecture and builds with increasing the cardinality to improve model accuracy [39-40]. Lastly, Zhang et al. designed ResNeSt [37], that is a modularized architecture, applying the channel-wise attention on different network branches to capture cross-feature interactions and learning diverse representations. Wang et al. [41] proposed a combined channel attention and spatial attention module network (CCSM-Net) for efficiently recognizing CHS with 2D images. Tan et al. [42] mentioned a hybrid architecture based on the improved ResNet and Transformer for classifying five easily confused rhizomatic TCM decoction pieces.

3. Motivations and deep considerations

3.1 Appearance similarity in CHS images

Although the methods presented earlier applied well on various categories of CHS classification, there is no evidence that they also have promising results on easily-confused CHS due to their highly similarity. In a narrow sense, Chinese herbs can be divided into root and rhizome, stem and wood, bark, leaf, flower, fruit and seed, and whole herb, and so on [43]. Generally speaking, different categories of CHS have great differences in appearance. That is to say, some different Chinese herbals categories are easy to be classified by the global color and shape features.

However, some different varieties of CHS, known as easily-confused CHS, have similar textures and appearances. For example, the image of Radix trichosanthis and Angelica dahurica, as depicted in **Fig. 1**, is sourced from the Chinese medicinal images database of

Hong Kong Baptist University. It is unable to ensure the accuracy and stability of identifying CHS based on people's subjective judgement and personal experiences. Simultaneously inter-class and intra-class similarity will also cause some interference to the recognition results by using machine learning models. Although the development of deep learning has greatly improved the classification of CHS, it is not easy to distinguish from each other by appearance features, which will cause some interference to the recognition results [41]. Therefore, extracting easily-confused CHS features quickly and reach a high classification rate is a more challenging task as it solves the problem of unclear visual feature differences.

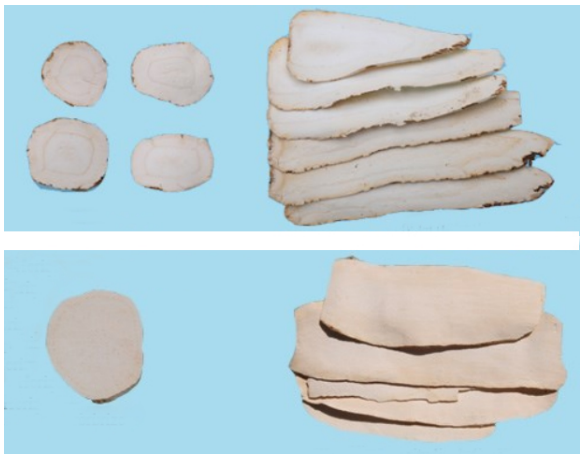

Fig. 1. Example of easily-confused CHS.
Angelica dahurica(top) and Radix trichosanthis(bottom)

3.2 Scopes for improvement

Recently, ResNet and ResNet like models have been successfully used in medical image classification [44-45]. ResNeSt is also from ResNet, which combines the channel-wise attention strategy with multi-path network layout to capture cross-feature interactions and learn diverse representations. Additionally, the model has achieved better speed-accuracy trade-offs than state-of-the-art CNN models on ImageNet dataset. Especially for small-scale datasets, the limited samples often cannot continuously improve the effectiveness of model training as the model depth grows. Due to the residual block explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions, it is suitable for small-scale datasets. However, relying solely on residual structures cannot fully solve the problem of accurately distinguishing the features between easily confused varieties, particularly in the intricate task of classifying herbal slices. The ResNeSt model adopts a split attention mechanism, which can achieve the distribution of feature attention within different feature map groups, thereby promoting the model's understanding of complex relationships. Consequently, to address the issues of high similarity and irregular shape in CHS, we propose our learning model named ResNeSt-CHS, to capture interdependencies of the CHS feature map.

4. Proposed method

Now we describe the overall architecture of our ResNeSt-CHS model. As illustrated in **Fig. 2**, we embed multilevel perception fusion (MPF) and perceptive sparse fusion (PSF) blocks to draw global context over local features, thus obtaining better feature representations for easily-

confused CHS images. The model architecture is set up as follows:

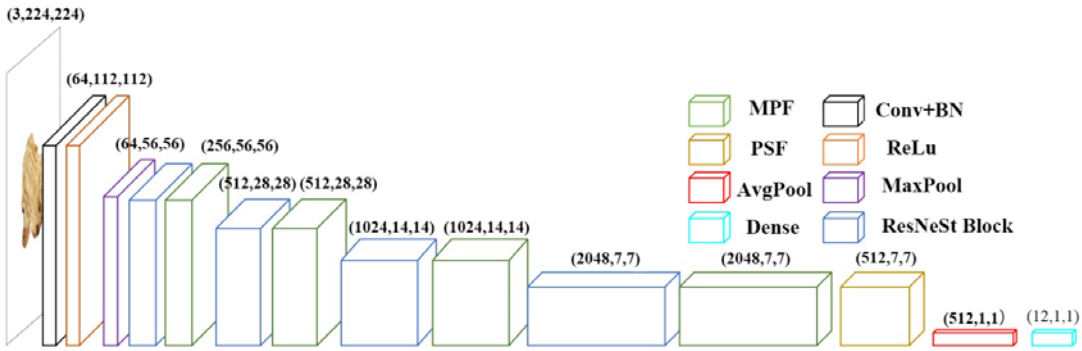

Fig. 2. The overview of ResNeSt-CHS architecture

4.1 Multilevel perception fusion block

Due to the high similarity in color and shape of the CHS, it is difficult to extract the most prominent features to distinguish easily-confused CHS. Therefore, we adopt an average concept in MPF block in order to focus more on global features information than salient features. Hence, the MPF block is a multi-channel block and is used when ResNeSt block execution ends, as shown in **Fig. 3**. Specifically, the input feature map performs channel AvgPool and global AvgPool operations to get the global spatial and channel information, and then transform and compress the dimension of the input feature map. Executing convolution to simultaneously extract local features on both spatial and channel dimensions in parallel, so as to improve the model's ability to obtain local details of CHS. A set of transformations of the right branch in the figure, which are shown as:

$$Z_c = O_c \sigma(f(\mathrm{Re}lu(f(\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} O_c(i,j))))) \tag{1}$$

where $Z_c \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$, $O_c(i,j)$ is the result of the output of previous ResNeSt block for c-th channel, $\sigma(x)$ is Sigmoid operation, and $f(x)$ is 3×3 convolution. In addition, the left branch transformations are computed as:

$$Z_{(h,w)} = O_{(h,w)} \sigma(f(\frac{1}{C}\sum_{i=1}^{C} O_{(h,w)}(i))) \tag{2}$$

where $Z_{(h,w)} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$, C denotes the size of channel dimensions, $O_{(h,w)}(i)$ is the output of (h,w) through previous ResNeSt block, $\sigma(x)$ is Sigmoid operation, and $f(x)$ is 3×3 convolution.

In ResNeSt style, although Split-Attention block has improved in capturing cross-feature interactions by integrating the channel-wise attention to multi-path network, it just concatenated the k cardinal group representations. When k increases, it may lead to redundancies cardinality groups and extract invalid feature information. While the MPF block is able to minimize the impact of redundant information on the model, in order to capture cross-channel feature correlations.

4.2 Perceptive sparse fusion block

Prior work has proved that a network should also have a large receptive field along with inductive bias to capture abundant contextual information [46]. However, the receptive field

may contain irrelevant and useless information with very sparse input [47]. Consequently, we design PSF block to reduce redundant information and strengthen salient features after completing the last MPF block calculation. In general, the input feature map undergoes a set of transformations in the PSF block, which are divided into two steps, as depicted in **Fig. 4**. First step, we apply global average pooling to obtain channel-wise texture information. After that, two consecutive 3×3 convolution to capture effective feature information from multi-channels. Then, we score and sort them through using sigmoid operation. Finally, we concatenate the high score channels to learning correlated feature representations from larger receptive fields. Next, in the second step, the input feature map is executed by a 3×3 convolution and then fused with the output of the first step to avoid data overfitting. Formally, we present aggregated transformations as:

$$L_c = f(Q_c) + g(\sigma(f(\mathrm{Re}lu(f(\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} Q_c(i,j)))))) \tag{3}$$

where $L_c \in R^{H \times W}$, $O_c(i,j)$ is the output result of the c-th channel after the last MPF block, $f(x)$ is 3×3 convolution operation, $\sigma(x)$ is Sigmoid operation, and $g(x)$ refers to sort and concatenate channels function.

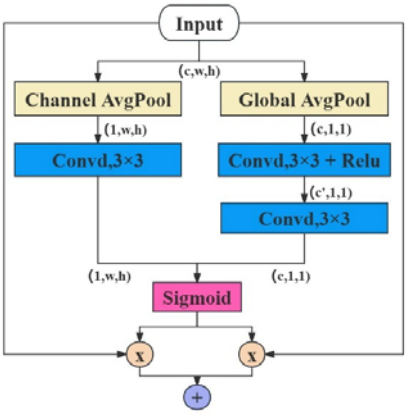

Fig. 3. The structure of the MPF

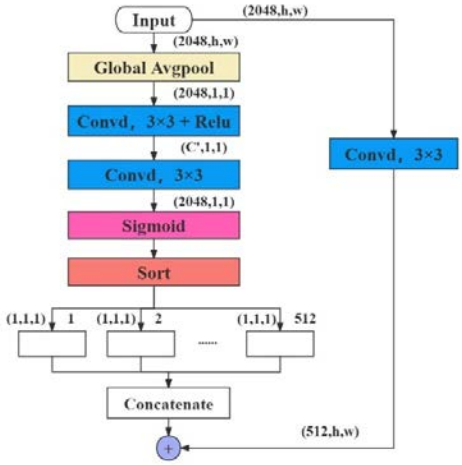

Fig. 4. The structure of the PSF

5. Experiments

5.1 Data collection and augmentation

To the best of our knowledge, there is still no available public CHS image dataset at present. Therefore, this paper starts with constructing an easily-confused CHS image dataset. We select six pairs of CHS types are very easily confused, which considering in this study. They are Atractylodes macrocephala and Rhizoma atractylodis (1st pair), Ginseng and American ginseng (2nd pair), Astragalus membranaceus and Radix isatidis (3rd pair), Angelica dahurica and Radix trichosanthis (4th pair), Polygonatum odoratum and Anemarrhena asphodeloides (5th pair), Angelica pubescens and Angelica sinensis (6th pair), as shown in **Fig. 5**. It should be noted that not only do the two CHS in the same pair look similar in appearance, texture and shape, but also the CHS in different pairs are very similar, such as the first and sixth pairs.

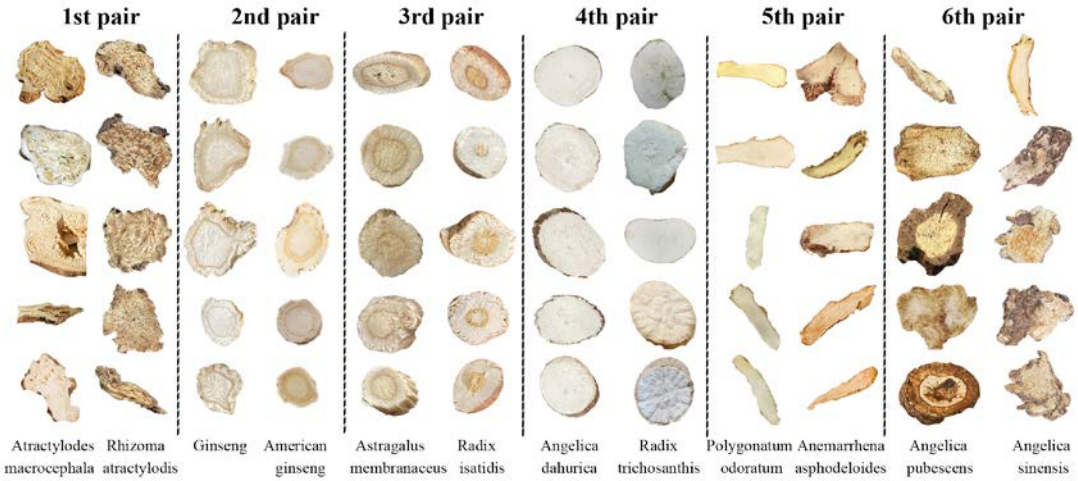

Fig. 5. Six pairs of easily-confused CHS

All samples were purchased from legitimate pharmacy and not deliberately selected with obvious features. Some of the samples are completely CHS, while others are fragmented herbal slices for data diversity. The tiny dataset consists of single-slice images, totally 2395 images with 6 easily-confused CHS pairs as shown in **Table 1**. Single-slice image can avoid the interference of overlapped CHS on feature extraction and improve the quality of training images. In order to ensure the authenticity and accuracy of the shooting color, we selected the Canon camera and took it under indoor natural light conditions. Besides, considering the issue of data balance, the quantity of each category is approximately 200 image samples.

Table 1. The number of each CHS

	CHS Name	Number
1st Pair	Atractylodes macrocephala	177
	Rhizoma atractylodis	180
2nd Pair	Ginseng	250
	American ginseng	197
3rd Pair	Astragalus membranaceus	216
	Radix isatidis	204
4th Pair	Angelica dahurica	287
	Radix trichosanthis	162
5th Pair	Polygonatum odoratum	150
	Anemarrhena asphodeloides	169
6th Pair	Angelica pubescens	200
	Angelica sinensis	203
Total		2395

We employed a specific data partitioning method, where the majority of the data was designated as the training subset to enhance the model's capability in capturing complex relationships and patterns. After that, an independent validation subset was reserved for hyperparameter tuning and model selection, aiming to mitigate issues such as overfitting. Finally, we established a test subset of equal size to the validation subset, with the objective

of evaluating the model's generalization performance on novel data, thereby ensuring a high degree of reliability and robustness in the model's outputs. As a result, randomly we select samples for training, validation, and testing in a ratio of 8:1:1. The size of each image is 256×256. To increase the generalization ability of the model, we apply some data augmentation techniques to expand twice the dataset, such as brightness adjustment and flipping.

5.2 Experimental setup

Our model is constructed using the PyTorch framework and the Python version 3.8. The system environment of platform is Linux with 12 core CPU, 86GB RAM, NVIDIA GeForce RTX 3090 GPU, and 24GB GPU RAM. We employ the optimizer SGD with 0.9 Nesterov momentum and the initial learning rate is 0.0005. To ensure the best performance of our model, we adopt warmup strategy [48] that gradually ramps up the learning rate from a small to a large value. This ramp avoids a sudden increase in learning rate, allowing healthy convergence at the start of training. Following prior research, we train our models starting from a learning rate of $\eta_n = \frac{1}{W*B}\eta_{base} + \eta_{n-1}$ ($\eta_0 = 0$), where B is the mini-batch size, W is the amount of warmup iteration, and η_{base} as the base learning rate. In this case, we set B = 14, W=10, and $\eta_{base} = 0.0005$. After the warmup, the learning rate decreases by 80% at the 15-th and 50-th epoch.

5.3 Evaluation metric

The metrics we apply to evaluate our models are accuracy, precision, recall and F-score. Among them, accuracy measures the proportion of correct predicted samples among the total number of predictions. Precision represents the proportion of the predicted samples that are correctly predicted. Recall denotes the proportion of true positives that were correctly classified. F-score is the harmonic mean of the precision and recall. We calculate mentioned metrics by following formulas:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5}$$

$$Re\,call = \frac{TP}{TP+FN} \tag{6}$$

$$F-Score = \frac{2\,Pr\,ecision*Re\,call}{Pr\,ecision+Re\,call} \tag{7}$$

Where TP (True Positive) refers to the count of samples that are actually positive and accurately predicted as positive by the model, TN (True Negative) represents the number of instances that are truly negative and correctly identified as negative by the model, FP (False Positive) corresponds to the number of cases that are actually negative but erroneously labeled as positive by the model, FN (False Negative) indicates the number of samples that are actually positive but wrongly classified as negative by the model.

6. Results and discussions

6.1 Recognition result of the proposed method

Experiments are implemented to verify the effectiveness of the proposed MPF and PSF blocks. We have tested ResNeSt50 with MPF, ResNeSt50 with PSF, ResNeSt-CHS, and compared them with the original ResNeSt50. To intuitively compare the effectiveness of different methods, **Fig. 6** illustrates the accuracy value of the validation dataset with 65 epochs.

Concretely, combining ResNeSt50 with MPF and PSF obtains 1.7%, 1.3% and 2.1% higher accuracy than ResNeSt50 with MPF, ResNeSt50 with PSF and ResNeSt50, respectively. And it demonstrates that our model can extract features quickly even with a limited amount of data, the recognition rate tends to stabilize after the 30-th epoch.

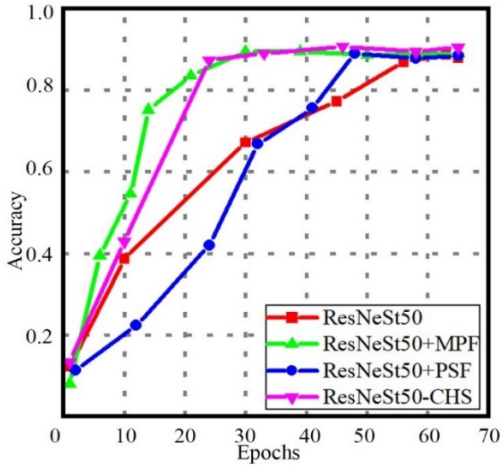

Fig. 6. Performance comparison of different methods

It is also worth pointing out that the recognition rate of second pair is very high, reaching 100% accuracy, while the first and sixth pair precision is relatively low, as shown in **Table 2**. For 1st and 6th pair, the accuracy of using ResNeSt50 to identify Atractylodes macrocephala and Angelica sinensis is only 55.6% and 55%, while our model is 72.3% and 75%, respectively. Taking Atractylodes macrocephala as an example, there are six sample recognition errors by ResNeSt50, of which three are mistakenly identified as Rhizoma atractylodis and the other three are misclassified as Angelica sinensis. This indicates that due to highly similar texture, it is extremely difficult to distinguish and learn the easily-confused CHS pairs, especially from the same pair. Obviously, compared with ResNeSt, our model is more effective in capturing abundant feature correlations and improving classification accuracy.

Table 2. The recognition rates of six CHS pairs

		ResNeSt50	**ResNeSt-CHS**
1st Pair	Atractylodes macrocephala	**0.556**	**0.723**
	Rhizoma atractylodis	0.944	0.889
2nd Pair	Ginseng	1.0	1.0
	American ginseng	1.0	1.0
3rd Pair	Astragalus membranaceus	1.0	0.952
	Radix isatidis	0.95	1.0
4th Pair	Angelica dahurica	1.0	1.0
	Radix trichosanthis	0.75	0.812
5th Pair	Polygonatum odoratum	1.0	0.933
	Anemarrhena asphodeloides	0.941	0.941
6th Pair	Angelica pubescens	0.85	0.8
	Angelica sinensis	**0.55**	**0.75**

ResNeSt has been recently the promising recognition model. To validate that ResNeSt-CHS is superior to the ResNeSt50 on the easily-confused CHS dataset, we design the experiments from two aspects: (1) compared them with same ResNeSt hyperparameters. (2) compared them with different ResNeSt hyperparameters.

When setting radix = 2 and cardinality=4, the comparison results are indicated in **Table 3**. It can be shown that ResNeSt50 with MPF performs better than ResNeSt50 in obtaining global context information. And ResNeSt50 with PSF achieves further optimization. Our model ResNeSt-CHS integrates MPF and PSF blocks, achieving the optimal result and outperforming ResNeSt50 by 2.1% of accuracy, 0.6% of precision, 0.13% of recall, and 1.6% of the F-score, respectively. It reveals that our method is conducive for getting global context information and achieving better identification result with the same size of hyperparameters.

Table 3. Evaluation results of models

Model	Param(M)	Accuracy	Precision	Recall	F-score
ResNeSt50	93.2	0.887	0.896	0.886	0.88
ResNeSt50+MPF	118.3	0.891	0.9	0.891	0.886
ResNeSt50+PSF	121.5	0.895	0.894	0.891	0.887
ResNeSt-CHS	146.6	**0.908**	0.902	0.899	0.896

With different value of radix and cardinality, ResNeSt-CHS has been significantly improved compared to ResNeSt. Our model has better fitting and stability for different parameters as shown in **Table 4**. Notably, we can see that ResNeSt-CHS50 and ResNeSt-CHS101 outperform their ResNeSt counterparts under the same value of radix and cardinality. When radix=4 and cardinality=2, ResNeSt-CHS50 performs best, which indicates that inner-class differences can be recognized in most easily-confused CHS categories. It can be observed that our model is beneficial to acquire the global context information and get a better generalization ability.

Table 4. Evaluation results of models with different R and K

Model	Param(M)	R	K	Accuracy	Precision	Recall	F-score
ResNeSt50	25.4	1	2	0.866	0.87	0.866	0.858
ResNeSt50	29.8	1	4	0.849	0.85	0.849	0.838
ResNeSt50	48.0	2	2	0.861	0.882	0.871	0.869
ResNeSt50	56.9	2	4	0.874	0.88	0.87	0.869
ResNeSt50	93.2	4	2	0.887	0.896	0.886	0.88
ResNeSt50	110.9	4	4	0.878	0.88	0.878	0.872
ResNeSt101	46.1	1	2	0.857	0.861	0.857	0.852
ResNeSt101	89.4	2	2	0.87	0.876	0.87	0.868
ResNeSt101	175.9	4	2	0.887	0.891	0.887	0.882
ResNeSt-CHS50	78.8	1	2	**0.878**	0.882	0.878	0.872
ResNeSt-CHS50	101.4	2	2	**0.874**	0.886	0.887	0.878
ResNeSt-CHS50	146.6	4	2	**0.908**	0.902	0.899	0.896
ResNeSt-CHS50	164.3	4	4	**0.891**	0.902	0.891	0.889

ResNeSt-CHS101	99.5	1	2	**0.87**	0.867	0.866	0.861
ResNeSt-CHS101	142.8	2	2	**0.887**	0.888	0.887	0.883
ResNeSt-CHS101	229.3	4	2	**0.903**	0.906	0.903	0.902

6.2 Comparison with different models

In order to assess the effectiveness of ResNeSt-CHS, lightweight CNN, and Transformer models are compared in parameter sizes, accuracy, precision, recall and F-score, including EfficientNet, DenseNet, GoogLeNet, Vision Transformer, and Swin Transformer. The experimental results are listed in **Table 5**. It can be seen that the lightweight CNN models have few parameters and a fast-training speed, but the model's ability to extract features is not strong. And the Transformer models are mainly applied to large datasets, resulting in unsatisfactory training results on our dataset, indicating significant limitations in practical application. The result with ResNeSt-CHS50 is the best without a remarkably increased in parameter sizes. ResNeSt-CHS50 outperforms 4.7%, 17.3% and 11.8% accuracy than GoogLeNet, Vision Transformer (ViT) and Swin Transformer, respectively. Accordingly, it is certificated that our model is beneficial to easily-confused CHS classification effectively.

Table 5. Comparison of different models

Model	Image Size	Param(M)	Accuracy	Precision	Recall	F-score
EfficientNet-B0	244×244	4.02	0.777	0.770	0.777	0.768
DenseNet121	244×244	6.7	0.845	0.856	0.845	0.843
GoogLeNet	244×244	12.0	0.861	0.861	0.861	0.854
ViT_b_16	244×244	85.8	0.706	0.699	0.706	0.698
ViT_b_32	244×244	87.5	0.693	0.702	0.693	0.693
ViT_l_16	244×244	303.3	0.735	0.737	0.735	0.724
Swin Transformer_t	244×244	27.5	0.786	0.78	0.786	0.779
Swin Transformer_s	244×244	48.8	0.782	0.796	0.782	0.778
Swin Transformer_b	244×244	86.8	0.79	0.788	0.79	0.782
ResNeSt-CHS50-4-2(ours)	256×256	146.6	**0.908**	0.902	0.899	0.896

6.3 Comparison with datasets of different sizes

To evaluate the effectiveness of the ResNeSt-CHS model across different magnitudes of training samples, we systematically compared its accuracy, precision, recall, and F1-score under conditions of no data augmentation, one-fold augmentation, two-fold augmentation, and three-fold augmentation. With radix parameter set at 4, cardinality at 2, and a batch size of 14 as default settings, the resultant performances are summarized in **Table 6**, demonstrating that our improved model consistently outperforms the original model across all ranges of training sample quantities. As the degree of data augmentation increases, the ResNeSt-CHS model exhibits a progressive enhancement in its overall learning capability. Specifically, it is noteworthy that while data augmentation techniques do not alter the intrinsic nature of images,

they enhance the model's comprehension and adaptability by applying a series of transformations such as rotations and scaling. However, as the number of training samples grows, signs of overfitting become increasingly evident in the model, manifested by a decline in various performance metrics following an initial improvement.

Table 6. Comparison of different sizes of datasets

Model	Number of training dataset	Accuracy	Precision	Recall	F-score
ResNeSt50	1919	0.832	0.832	0.832	0.827
ResNeSt50+MPF	1919	0.84	0.843	0.84	0.836
ResNeSt50+PSF	1919	0.845	0.842	0.845	0.838
ResNeSt-CHS	1919	**0.861**	0.862	0.857	0.854
ResNeSt50	3838	0.857	0.854	0.857	0.851
ResNeSt50+MPF	3838	0.861	0.864	0.861	0.852
ResNeSt50+PSF	3838	0.866	0.873	0.866	0.865
ResNeSt-CHS	3838	**0.874**	0.875	0.874	0.871
ResNeSt50	5757	0.887	0.896	0.886	0.88
ResNeSt50+MPF	5757	0.891	0.9	0.891	0.886
ResNeSt50+PSF	5757	0.895	0.894	0.891	0.887
ResNeSt-CHS	5757	**0.908**	0.902	0.899	0.896
ResNeSt50	7676	0.878	0.881	0.878	0.876
ResNeSt50+MPF	7676	0.882	0.883	0.882	0.88
ResNeSt50+PSF	7676	0.887	0.888	0.887	0.884
ResNeSt-CHS	7676	**0.895**	0.9	0.895	0.892

7. Conclusion

Automated recognition of CHS utilizing computer vision technology holds a pivotal position in the practical implementation of intelligent Chinese medicine. Although traditional classification algorithms perform well in CHS with significant differences in appearance features, the classification performance of deep learning methods is significantly higher than that of traditional classification algorithms when faced with similar shaped herbal slices. For the challenge of distinguishing similar shaped slices, we are currently facing issues such as the need for a large amount of data in deep learning and the lack of image datasets for slices of herbs. This paper establishes a tiny dataset for easily-confused CHS. Then, we have presented a novel ResNeSt-CHS model based on ResNeSt, which adds MPF and PSF blocks. This model achieves promising easily-confused CHS classification results with better outcome and efficiency. In addition, numerous experiments have been carried out to verify ResNeSt-CHS is the optimal model for easily-confused CHS classification, even in a small-scale dataset. Future work will explore a larger easily-confused CHS dataset with more categories and more accuracy method for identifying CHS.

Acknowledgement

This work is supported by Guangdong Administration of Traditional Chinese Medicine, China (No.20221221 and No.20231221); the College Student Innovation and Entrepreneurship Training Program of Guangdong Province (No. 202310573014) and Special Fund for Science and Technology Innovation Strategy of Guangdong Province ("Climbing Program")(No. pdjh2023b0273).

References

[1] Han, M., Zhang, J., Zeng, Y., Hao, F., & Ren, Y., "A Novel Method of Chinese Herbal Medicine Classification Based on Mutual Learning," *Mathematics*, vol.10, no.9, 2022. Article(CrossRef Link)

[2] Hu, H., & Chung, C. C., "The innovation and modernisation of 'herbal pieces' in China: System evolution and policy transitions (1950s–2010s)," *European Journal of Integrative Medicine*, vol.7, no.6, pp.645-649, 2015. Article (CrossRef Link)

[3] Xie, Z. W., ""Differentiation of symptoms and discussion of quality" in traditional experience identification of traditional Chinese medicine varieties," *Lishizhen Medicine and Materia Medica Research*, vol.5, no.3, pp.19-21, 1994. Article(CrossRefLink)

[4] Zhang, Y., Wan, H., and Tu, S. Q., "Technical review and case study on classification of Chinese herbal slices based on computer vision," *Journal of Computer Applications*, vol.42, no.10, pp.3224-3234, 2022. Article (CrossRef Link)

[5] Tang, Y., Wang, Y., Li, J., Zhang, W., Wang, L., Zhai, X., & Han, A., "Classification of Chinese Herbal Medicines by deep neural network based on orthogonal design," in *Proc. of 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pp.574-583, 2021. Article (CrossRef Link)

[6] Cai, C., Liu, S., Wang, L., Yang, B., Zhi, M., Wang, R., & He, W., "Classification of Chinese Herbal Medicine Using Combination of Broad Learning System and Convolutional Neural Network," in *Proc. of 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp.3907-3912, 2019. Article (CrossRef Link)

[7] Zhu, X., Zhu, M., & Ren, H., "Method of plant leaf recognition based on improved deep convolutional neural network," *Cognitive Systems Research*, vol.52, pp.223-233, 2018. Article (CrossRef Link)

[8] Kadir, A., Nugroho, L. E., Susanto, A., & Santosa, P. I., "Leaf Classification Using Shape, Color, and Texture Features," *International Journal of Computer Trends and Technology (IJCTT)*, vol.1, no.3, pp.225-230, 2011. Article (CrossRef Link)

[9] Dehan, L., Jia, W., Yimin, C., & Hamid, G., "Classification of Chinese Herbal medicines based on SVM," in *Proc. of 2014 International Conference on Information Science, Electronics and Electrical Engineering*, vol.1, pp.453-456, 2014. Article (CrossRef Link)

[10] Mahajan, S., Raina, A., Gao, X. Z., & Kant Pandit, A., "Plant Recognition Using Morphological Feature Extraction and Transfer Learning over SVM and AdaBoost," *Symmetry*, vol.13, no.2, 2021. Article (CrossRef Link)

[11] Xing, C., Huo, Y., Huang, X., Lu, C., Liang, Y., & Wang, A., "Research on Image Recognition Technology of Traditional Chinese Medicine Based on Deep Transfer Learning," in *Proc. of 2020 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, pp.140-146, 2020. Article (CrossRef Link)

[12] Simonyan, K., & Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. of International Conference on Learning Representations*, 2015. Article (CrossRef Link)

[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E., "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol.60, no.6, pp.84-90, 2017. Article (CrossRef Link)

[14] He, K., Zhang, X., Ren, S., & Sun, J., "Deep Residual Learning for Image Recognition," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016. Article (CrossRef Link)
[15] Chollet, F., "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1251-1258, 2017. Article (CrossRef Link)
[16] Lee, S., Choi, G., Park, H.-C., Choi, C., "Automatic Classification Service System for Citrus Pest Recognition Based on Deep Learning," *Sensors*, vol.22, no.22, 2022. Article (CrossRef Link)
[17] Dilshad, N., Khan, T., Song, J., "Efficient Deep Learning Framework for Fire Detection in Complex Surveillance Environment," *Computer Systems Science and Engineering*, vol.46, no.1, pp.749-764, 2023. Article (CrossRef Link)
[18] Wang, W., Tian, W., Liao, W., Cai, B., & Li, B., "Identifying Chinese Herbal Medicine by Image with Three Deep CNNs," in *Proc. of CCEAI '21: Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence*, pp.1-8, 2021. Article (CrossRef Link)
[19] Sun, X., & Qian, H., "Chinese Herbal Medicine Image Recognition and Retrieval by Convolutional Neural Network," *PLoS ONE*, vol.11, no.6, 2016. Article (CrossRef Link)
[20] Huang, F., Yu, L., Shen, T., & Jin, L., "Chinese Herbal Medicine Leaves Classification Based on Improved AlexNet Convolutional Neural Network," in *Proc. of 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, vol.1, pp.1006-1011, 2019. Article (CrossRef Link)
[21] Zhao, P., "Explore the identification of Chinese herbal medicine based on the VGG-16 model," in *Proc. of the 3rd International Conference on Signal Processing and Machine Learning*. vol.4, pp. 645-650, 2023. Article (CrossRef Link)
[22] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q., "Densely Connected Convolutional Networks," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.2261-2269, 2017. Article (CrossRef Link)
[23] Liu, S., Chen, W., & Dong, X., "Automatic Classification of Chinese Herbal Based on Deep Learning Method," in *Proc. of 2018 14th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)*, pp.235-238, 2018. Article (CrossRef Link)
[24] Liu, S., Chen, W., Li, Z., & Dong, X., "Chinese Herbal Classification Based on Image Segmentation and Deep Learning Methods," in *Proc. of Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Proceedings of the ICNC-FSKD 2021*, pp.267-275, 2022. Article (CrossRef Link)
[25] Hao, W., Han, M., Yang, H., Hao, F., & Li, F., "A novel Chinese herbal medicine classification approach based on EfficientNet," *Systems Science & Control Engineering*, vol.9, no.1, pp.304-313, 2021. Article (CrossRef Link)
[26] Wu, C., TAN, C. Q., Huang, Y. L. Wu, C. J., Chen, H., "Intelligent Identification of Fritillariae Cirrhosae Bulbus, Crataegi Fructus and Pinelliae Rhizoma Based on Deep Learning Algorithms," *Chinese Journal of Experimental Traditional Medical Formulae*, vol.26, no.21, pp.195-201, 2020. Article (CrossRef Link)
[27] Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., & Xu, C., "CMT: Convolutional Neural Networks Meet Vision Transformers," in *Proc. of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.12165-12175, 2022. Article (CrossRef Link)
[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., "Attention is all you need," in *Proc. of NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp.6000-6010, 2017. Article (CrossRef Link)
[29] Niu, Z., Zhong, G., & Yu, H., "A review on the attention mechanism of deep learning," *Neurocomputing*, vol.452, pp.48-62, 2021. Article (CrossRef Link)
[30] Wang, Y., Feng, Y., Zhang, L., Zhou, J. T., Liu, Y., Goh, R. S. M., & Zhen, L., "Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images," *Medical Image Analysis*, vol.81, 2022. Article (CrossRef Link)

[31] Shanshan, W., Tao, Z., Fei, L., ZhenPing, R., Zhen, Y., Shu, Z., & ZhiQiang, Z., "A Synergic Neural Network For Medical Image Classification Based On Attention Mechanism," in *Proc. of 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, pp.82-87, 2022. Article (CrossRef Link)
[32] Zhu, H., Wang, J., Wang, S. H., Raman, R., Górriz, J. M., & Zhang, Y. D., "An Evolutionary Attention-Based Network for Medical Image Classification," *International Journal of Neural Systems*, vol.33, no.3, 2023. Article (CrossRef Link)
[33] Xu, Y., Wen, G., Hu, Y., Luo, M., Dai, D., Zhuang, Y., & Hall, W., "Multiple attentional pyramid networks for Chinese herbal recognition," *Pattern Recognition*, vol.110, 2021. Article (CrossRef Link)
[34] Miao, J., Huang, Y., Wang, Z., Wu, Z., & Lv, J., "Image recognition of traditional Chinese medicine based on deep learning," *Frontiers in Bioengineering and Biotechnology*, vol.11, 2023. Article (CrossRef Link)
[35] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S., "A ConvNet for the 2020s," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, pp.11966-11976, 2022. Article (CrossRef Link)
[36] Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., & Huang, G., "On the Integration of Self-Attention and Convolution," in *Proc. of the 2022 IEEE/CVF conference on computer vision and pattern recognition*, pp.805-815, 2022. Article (CrossRef Link)
[37] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A., "ResNeSt: Split-Attention Networks," in *Proc. of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.2735-2745, 2022. Article (CrossRef Link)
[38] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K., "Aggregated Residual Transformations for Deep Neural Networks," in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp.5987-5995, 2017. Article (CrossRef Link)
[39] Nan, F., Zeng, Q., Xing, Y., & Qian, Y., "Single Image Super-Resolution Reconstruction based on the ResNeXt Network," *Multimedia Tools and Applications*, vol.79, pp.34459-34470, 2020. Article (CrossRef Link)
[40] Go, J. H., Jan, T., Mohanty, M., Patel, O. P., Puthal, D., & Prasad, M., "Visualization Approach for Malware Classification with ResNeXt," in *Proc. of 2020 IEEE Congress on Evolutionary Computation (CEC)*, pp.1-7, 2020. Article (CrossRef Link)
[41] Wang, J., Mo, W., Wu, Y., Xu, X., Li, Y., Ye, J., & Lai, X., "Combined Channel Attention and Spatial Attention Module Network for Chinese Herbal Slices Automated Recognition," *Frontiers in Neuroscience*, vol.16, 2022. Article (CrossRef Link)
[42] Tan, D. Q. et al., "Research on identification of confusing TCM decoction pieces by integrating of improved residual network," *China Digital Medicine*, vol.18, no.6, pp.42-50, 2023. Article (CrossRef Link)
[43] Yang, Z., Wang, X., Hong, W., Zhang, S., Yang, Y., Xia, Y., & Yang, R., "The pharmacological mechanism of Chinese herbs effective in treating advanced ovarian cancer: Integrated meta-analysis and network pharmacology analysis," *Frontiers in Pharmacology*, vol.13, 2022. Article (CrossRef Link)
[44] Sarwinda, D., Paradisa, R. H., Bustamam, A., & Anggia, P., "Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer," *Procedia Computer Science*, vol.179, pp.423-431, 2021. Article (CrossRef Link)
[45] Showkat, S., & Qureshi, S., "Efficacy of Transfer Learning-based ResNet models in Chest X-ray image classification for detecting COVID-19 Pneumonia," *Chemometrics and Intelligent Laboratory Systems*, vol.224, 2022. Article (CrossRef Link)
[46] Lou, M., Zhou, H. Y., Yang, S., Yu, Y., "TransXNet: Learning Both Global and Local Dynamics with a Dual Dynamic Token Mixer for Visual Recognition," *arXiv:2310.19380*, 2023. Article (CrossRef Link)

[47] Sun, X., Ponce, J., Wang, Y. X., "Revisiting Deformable Convolution for Depth Completion," in *Proc. of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp.1300-1306, 2023. Article (CrossRef Link)
[48] Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K., "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour," *arXiv:1706.02677*, 2017. Article (CrossRef Link)

Qi Zhang, She received M.S. degree in Hong Kong Polytechnic University in 2009. She is as a lecturer in Department of Digital Media Technology, Guangdong Pharmaceutical University. Her current research interests include machine learning and computer vision.

Jinfeng Ou, He is currently pursuing the B.S degree in School of Medical Information and Engineering, Guangdong Pharmaceutical University. His research interests focus on computer vision and deep learning.

Huaying Zhou, She received Ph.D. degree in Guangdong University of Technology in 2019, China. She is as a director in Department of Computer, Guangdong Pharmaceutical University. Her current research interests include intelligent odor identification and machine learning.