A Study on Image Recognition of Chinese Herbal Medicine Based on Probabilistic Convolution Voting Model

Yiheng Yan[1,*], Ziting Gao[2]

{yanyih@kean.edu[1], gaoziting@bupt.edu.cn[2]}

College of Science, Mathematics and Technology, Kean University, Union, New Jersey, 07083, US[1]
International School, Beijing University of Post and Telecommunications, Beijing, 100876, China[2]
*corresponding author

Abstract. This study investigates the application of deep learning in the field of Chinese herbal medicine (CHM) image recognition, proposing a novel multi-model ensemble approach called the Probabilistic Convolution Voting Model (PCVM) to address the limitations of current single-model systems. Traditional Chinese Medicine (TCM), with its roots in ancient Daoist practices, has gained global relevance, particularly during the COVID-19 pandemic, for its efficacy and minimal side effects. However, CHM identification faces challenges such as inefficiency and error-proneness in traditional identification methods. We leverage advances in convolutional neural networks (CNNs) and Vision Transformers (ViTs) to enhance image recognition tasks essential for CHM identification. The proposed PCVM integrates multiple pre-trained models to improve accuracy and robustness, focusing particularly on enhancing recognition capabilities in visually complex categories. The ensemble method uses a novel convolutional and voting layer system to reduce misclassification and optimize decision-making. Comprehensive experiments demonstrate that PCVM significantly outperforms traditional methods, particularly in challenging categories, making it a promising solution not only for CHM but potentially for other high-precision medical image recognition tasks such as cancer detection. Future work will focus on further optimizing and expanding the PCVM framework to enhance computational efficiency and applicability in complex image recognition scenarios.

Keywords: Ensemble Learning, Chinese Herbal Medicine, Deep Learning, Image Recognition, Probabilistic Convolution Voting Model, Convolutional Neural Networks, Vision Transformers.

1 Introduction

Traditional Chinese medicine (TCM) is one of the world's oldest healthcare systems, with its development spanning over 3000 years,[1] tracing its origins back to ancient Daoist practices in China.[2] Primarily utilized in Southeast Asia, TCM is famous for treating a variety of diseases with its efficacy and minimal side effects. The recent Coronavirus disease 2019 (COVID-19) pandemic highlighted the global relevance of CHM, especially its effectiveness in treating COVID-19 pneumonia when integrated with Western medical practices.[3][4]

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354590

Despite its numerous benefits, Chinese herbal medicine (CHM) faces challenges, particularly in identification, quality control, and standardization.[5][6] Traditional identification techniques depend on expert knowledge and visual inspection, which are inefficient and prone to errors. Advances in deep learning have introduced new methods to address these challenges[7]. Technologies such as convolutional neural networks (CNNs)[8] and Vision Transformers (ViTs) have revolutionized image recognition tasks[9][10], showing potential in the crucial processes of automatic feature extraction and classification needed for CHM identification[11][12][13].

However, despite the high accuracy of models like CNNs and ViTs, their performance in individual classes can be unsatisfactory,[14] particularly in the field of CHM image recognition,[15] where accurate identification of each herb is crucial as misidentification can lead to inadequate therapeutic outcomes or even risks of poisoning.

To fill this gap, this study aims to further explore the application of deep learning in the field of CHM image recognition, improving performance on disadvantaged classes while maintaining overall accuracy. We propose a multi-model ensemble approach aimed at enhancing the precision and robustness of these systems.[14] Our method involves constructing a comprehensive dataset, selecting robust pre-trained models, and introducing a novel multi-model ensemble technique based on a Probabilistic Convolution Voting Model (PCVM). This approach not only aims to enhance accuracy but also includes specialized assessment methods tailored to the nuances of PCVM, enabling comprehensive analysis of its effectiveness compared to traditional single-model systems. During the model training phase, we utilize cross-entropy loss functions and the Adam optimizer[16], combined with a stepwise learning rate adjustment strategy to optimize performance. Additionally, employing mixed-precision training methods reduces memory requirements and accelerates training speed. To thoroughly assess model performance, we rely not only on traditional metrics like accuracy, recall, and F1-score but also develop a new evaluation method specifically for the PCVM model to highlight its performance improvements in disadvantaged categories. Our main contributions are summarized as follows, proposing a judgment model based on multiple computer vision models called the Probabilistic Convolution Voting Model.

1. The PCVM significantly enhances recognition accuracy in visually complex and similar categories by integrating the strengths of ResNet and three ViT models. Particularly in categories where individual models are less effective, this enhancement has been empirically validated through specific experimental results.
2. We have designed a unique convolutional and voting layer system. This system effectively integrates information from different models, reduces misclassification, and optimizes the decision-making process through probabilistic methods.
3. The design of PCVM is not only applicable to Chinese herbal image recognition; its principles and architecture can also be extended to other high-precision and robust medical and biological image recognition tasks, such as cancer detection and pathological image analysis.

2 Literature Review

In recent studies, Vision Transformers (ViT) have been compared with traditional Convolutional Neural Networks (CNN) across multiple image processing tasks. This literature

review synthesizes numerous comparisons of ViT and CNN in image classification [17]. Research [18] indicates that ViT performs better than CNN when processing natural or noisy images and that combining these two architectures can significantly enhance model accuracy. Additionally, studies [19] and [20] have found that ViT, due to its self-attention mechanism, handles full-image information more robustly in digital holography and medical image analysis (such as chest X-rays), showing accuracy and robustness that are superior to or at least equivalent to CNN. Despite ViT's advantages with smaller datasets,

Convolutional Neural Networks (CNN) are a core technology in deep learning, where Deep Convolutional Neural Networks (DCNN) stack more convolutional layers compared to basic CNNs, extracting increasingly complex and abstract features such as shapes, colors, and textures, which are crucial for identifying traditional Chinese medicine. This makes DCNN particularly suited for handling images of Chinese herbs with complex textures and shapes. In the evolution of deep convolutional neural networks, Krizhevsky, LeCun, and others have established a series of milestones for image classification [21][22][23]. The classic ResNet architecture explicitly fits these stacked layers to residual mappings rather than the underlying mappings [24]. Proposed by Kaiming, ResNet's primary aim is to reduce computational overhead during network training and to address issues related to gradient diminishing or exploding, which can degrade performance as network depth increases [9]. The architecture employs stacked non-linear layers to accommodate skip connections, establishing identity mappings. This ensures that deeper layers perform as effectively as shallower networks. ResNet architecture introduces residual connections, allowing gradients to flow directly through an alternative bypass route [25]. ResNet-50 has proven to be a versatile and effective tool for medical image analysis, as evidenced by the following studies. Chhabra and Kumar [5] developed a model to detect pneumonia in patients, making the detection process faster and more accurate. ResNet-50, combined with transfer learning, is used to classify images from two different categories (normal and pneumonia), then evaluated using a confusion matrix, showing an accuracy of up to 94.

The Transformer model was initially designed for Natural Language Processing (NLP) tasks, where its self-attention mechanism effectively extracts intrinsic attributes from text [26]. With the evolution of model architectures, ViT (Vision Transformer) [12] first demonstrated the feasibility of using a pure Transformer architecture for computer vision tasks. This pioneering research not only showcased the potential application of Transformer in image classification tasks but also extended to object detection [27], semantic segmentation [28], image generation [29], and various other visual tasks [30]. Nonetheless, studies have found that Transformer models in the visual domain often overlook local feature details when applying self-attention mechanisms. This oversight can make it challenging for the model to differentiate between the target object and the background, especially when the target's local features are not prominent [31].

Both DCNN and ViT have shown good results in image recognition, yet current research mostly focuses only on the models' overall performance, neglecting the importance of addressing underperforming categories, especially in the medical field where focusing solely on overall performance is not a sufficient standard.

In the aforementioned review, comparisons were observed between Vision Transformers (ViT) and Convolutional Neural Networks (CNN) across multiple image processing tasks, with an

emphasis on their respective advantages. Existing studies primarily focus on the performance of models on large datasets, with insufficient consideration for their application on specific categories or smaller datasets. Additionally, the limitations of single models when dealing with complex or similar images have not been adequately addressed. To counter these deficiencies, this study introduces the innovative Probabilistic Convolution Voting Model (PCVM), which combines the strengths of ResNet and ViT, and significantly enhances the model's performance and accuracy through specific convolutional and voting layer designs. The PCVM integrates the strengths of multiple models, not only extending the recognition capabilities of a single model in specific categories but also reducing classification errors through its unique voting mechanism. This is particularly crucial in the recognition of medical and herbal images, where images often appear similar yet belong to critically distinct categories. By employing probabilistic convolution processing, the PCVM optimizes the recognition process, ensuring precision in final decisions. Moreover, the model's collective decision-making mechanism leverages the collective intelligence of multiple models, enhancing the robustness and adaptability of the recognition system. In summary, by addressing the shortcomings of existing research innovatively, the PCVM model not only improves the accuracy of specific image recognition but also provides an effective new approach to handling complex image recognition challenges. These improvements and enhancements significantly advance the development of image-processing technologies and pave new paths for future research.

3 Methodology

3.1 Data Set Description

Our dataset includes a total of 266,767 images of herbal medicines in 163 categories from [32], with a balanced number of images for each category of herbal medicine. Images of herbs that did not meet the criteria, such as being blurred or mislabelled, were removed. First, we used random sampling to select 10,000 images from a total of 266,767 images as an independent test set for the final model evaluation to ensure the accuracy of the model evaluation. The remaining 256,767 images were divided into a training set and a validation set, respectively, using a ratio of 4:1. This stratification ensured that both the training and validation sets represented the characteristics of the complete dataset, thus maintaining the integrity of the model evaluation process.

In the data preprocessing phase, we defined specific transformation operations for the different models. For the test and training sets, we used the following preprocessing steps: for the ResNet50-based model, the images were first resized to 256 pixels, then centrally cropped to 224 pixels, converted to a tensor, and normalised with a normalised mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225]. For the Vision Transformer-based model, we uniformly resized the images to 224x224, converted the greyscale images to pseudo-RGB format, and then performed the same tensor quantisation and normalisation. The average size of the images is 299x299 pixels, which helps to maintain the accuracy of the image quality and details.

3.2 Pretrained Model

For multi-class image classification of herbal pictures, four pre-trained models known for their proficiency in handling image recognition were selected. The selected models are ResNet50, three Vision Transformer[33][34][35]. The pre-trained ResNet-50 model, a widely used convolutional neural network for image recognition tasks[36], is used in the code. Replacing the last fully-connected layer with migration learning to adapt to a new classification task can take advantage of the knowledge learned by the ResNet50 pre-trained model on large datasets, thus accelerating the learning process and improving the performance of the new task. Since the results of fine-tuning the training of fully- connected classification layers are not satisfactory, training all layers using the training set fine-tuning resulted in a significant improvement in accuracy, precision, recall score and f1 score. This experiment uses RTX*4090*5 GPU to accelerate the computation and DataParallel for parallel training of the model, which helps to improve the training efficiency. Cross-entropy loss function and Adam optimiser are used. Cross-entropy is a popular loss function for multi-class classification problems, while the Adam optimiser is respected for its adaptive learning rate adjustment. A stepped learning rate tuning strategy (halved every 5 cycles) is used, which helps to refine the model parameters in the later stages of training, potentially avoiding overfitting and improving generalisation.

For Vision Transformer fine-tuning training, we use DataLoader to load data in parallel to batch process images during training and testing. Speed up training and allocate GPU resources wisely. Dynamically adjust the output dimension of the classifier layer based on the number of categories in the dataset. Use Adam optimiser with learning rate set to 1e-3. Use GradScaler and auto-cast for mixed-precision training to reduce memory consumption while speeding up training. Training Cycle: The model is trained in multiple iterations (set to 50 rounds), and the performance of the model is evaluated by calculating the cross-entropy loss in each round.

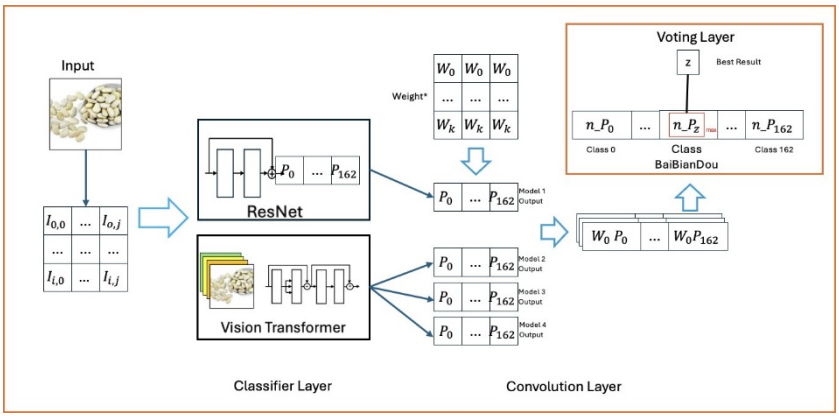

Fig. 1. Examples of Probabilistic Convolution Voting Model(PCVM) architecture diagram.

3.3 PCVM

A judgment method based on multiple computer vision models is called the Probabilistic convolution voting model (PCVM). In this case, PCVM includes ResNet and three ViT models, which perform well in herbal medicine image recognition.

PCVM consists of three main parts: classifiers layer, convolution layer, and voting layer.

The Classifiers layer is a multi-visual model container. This layer can incorporate multiple models based on different algorithms that are computed separately. This approach can fully utilise the strengths of each model in its unique domain. In this study ResNet and the three ViT models have more prominent performance in different classes of the dataset, respectively. They constitute the first layer of PCVM.

Consider that these four models have misclassification flaws in some of the categories, despite their obvious advantages. PCVM focuses more on optimising the shortcomings of each model and avoiding the bucket effect. The Convolution layer is an important component to achieve this feature. Typically, the probability of a model's final outcome on an image of an uncertain category is not significantly more salient than the rest of the category probabilities. Therefore, when multiple models are considered together, the model that has a significant advantage in this image category can dominate the judgement of the PCVM, thus improving the lower bound of the overall model and reducing the frequency of the bucket effect. Convolution gets its name from stacking probabilities, a computational approach as shown in figure 1.

When performing model inference, this study used a voting mechanism to combine the predictions of different models to improve the final classification accuracy. For each test image, all models make predictions independently, and then the majority voting method is used to select the final classification result. In categorization, if a category is supported by the majority of models, that category is selected as the final prediction. In addition, the prediction probabilities of each model were collected for further evaluating the model performance and performing more complex analyses.

Through these methods, this study successfully integrated the global feature extraction capability of the ViT model and the local feature extraction capability of the CNN model, resulting in a significant performance improvement in the task of herbal medicine image recognition.

3.4 Evaluation

For the pre-trained models, we use the traditional f1-scores evaluation method. The metrics are mathematically defined as follows:

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

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

$$\text{Recall}=\frac{TP}{TP+FN}$$

$$\text{Top-10 accuracy}=\frac{Number\ of\ correct\ predictions\ in\ top\ 10}{Total\ number\ of\ samples}$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. In this study, there is no obvious precision recall tendency, so the weights are taken as 1:1. F1scores can reflect the model's ability on this image classification. For PCVM, a two-track evaluation method is used, and f1-scores is referenced as a traditional evaluation method, and a new evaluation method applicable to PCVM is proposed. As PCVM

focuses on improving the lower bounds of all models and maintaining a high level of category judgement. f1-scores is hardly able to improve on the overall performance of the already highly dominant category judgement, and is unable to reflect the advantages of PCVM on items with inferior category judgement. The new evaluation algorithm focuses more on the improvement in each model's ability to make disadvantageous category judgements.

This evaluation approach is more reflective of the strengths of our model.

4 Experiment and Result

The use of the PCVM based on the pretrained models ResNet-50 and three Vision Transformers, for classifying images of Chinese herbal medicines has achieved significant improvements in performance metrics for disadvantaged categories.

In the experimental results, we initially fine-tuned only the fully connected classification layer of the ResNet-50 model, which yielded high accuracy for a smaller number of categories. However, when expanded to 163 categories, the accuracy in the multi-class Chinese herbal medicine recognition task was only 0.7814, which did not meet our model selection criteria for PCVM. Subsequent comprehensive fine-tuning of all layers of the ResNet-50 model significantly increased its recognition accuracy to 0.9818. This outcome clearly demonstrates that extensive fine-tuning can greatly enhance the model's performance in complex image recognition tasks, although it exhibited a higher misclassification rate in individual categories, as shown in figure 2.

The fine-tuning approach significantly enhanced the precision, recall, and F1 score of the ResNet model in multi-class Chinese herbal medicine image recognition, highlighting an effective balance between recall and precision. The F1 score (i.e., the harmonic mean of precision and recall) was very high, reflecting the model's effectiveness in detecting true positives and maintaining accuracy across categories.

All Vision Transformer models showed similar and high levels of accuracy after fine-tuning, slightly lower than the fully fine-tuned ResNet-50, with accuracies of the three ViT models at 0.9329, 0.9373, and 0.9472 respectively. They significantly sped up the training process while maintaining high classification accuracy through data parallelism and mixed precision training. Notably, the recall for typically challenging categories improved, indicating these models' capability to minimize false negatives.

Both traditional and newly proposed methods were used to evaluate the PCVM, which was designed to improve model performance in challenging categories.

The F1 score for PCVM was slightly higher than that of individual models, mainly due to its complex method of leveraging the strengths of different models to enhance overall performance. This result highlights the advantage of integrating multiple model outputs to improve prediction accuracy. Using PCVM, the final result's F1 score reached 0.9872, an improvement of 0.0052 over the highest among the four models (ResNet) and 0.0549 over the lowest ViT, as shown in Figure 2. With limited scope for improvement in a context where accuracy is already near 1, this evaluation does not perfectly reflect the model's comparative merits.

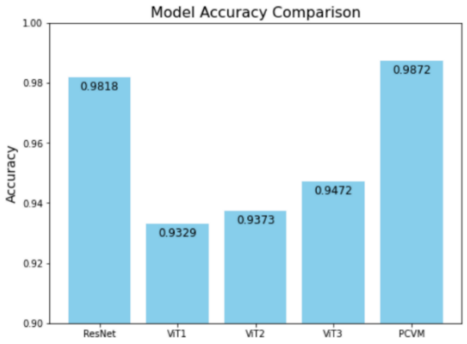

Fig. 2. 4 Models Accuracy Comparison.

Therefore, we employed a new evaluation criterion that focuses more on the specific strengths and weaknesses of each category and the overall optimization of shortcomings. The results indicated that PCVM significantly increased the accuracy and F1 score floor for these categories. This was particularly evident in herbal categories where individual models had previously struggled. Using PCVM, the category with the highest misclassification in ResNet, "malt," saw a reduction in misclassifications from 40 to 21, improving by 0.475. In Figure 3, comparing the prediction errors before and after using PCVM for each category reveals PCVM's substantial impact on enhancing recognition capabilities in various categories. Compared to ResNet, 48 categories showed improved recognition efficiency, with only three showing a decrease. The second ViT model performed the best, improving recognition in 145 categories with only one showing a decrease. In Figure 4, we observe the overall change in prediction errors, with an improvement of about 0.59-0.84 compared to individual models. This indicates PCVM's significant effect on enhancing category-specific and overall performance.

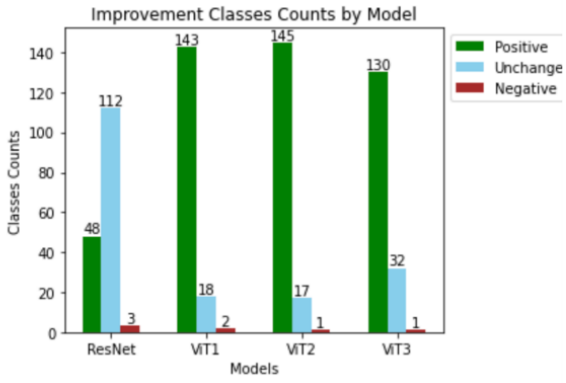

Fig. 3. Improved Class Counts by Model.

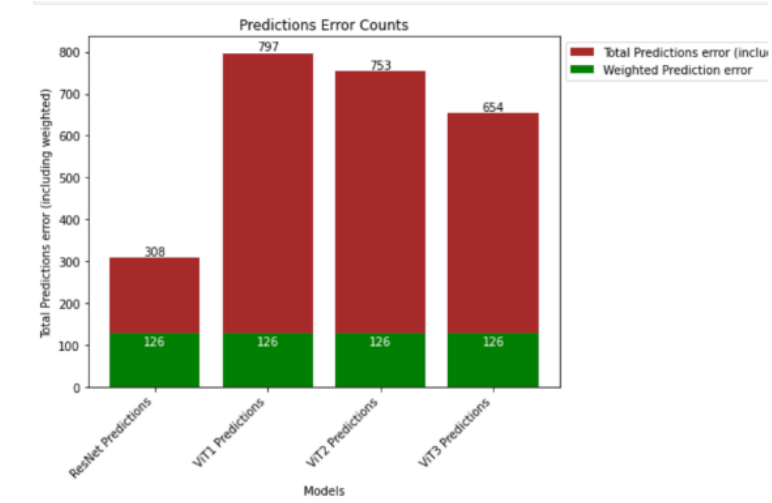

Fig. 4. Predictions Error Counts.

5 Conclusion

In conclusion, the application of advanced pretrained models alongside the PCVM has markedly enhanced the performance of Chinese herbal medicine image classification. This integration has not only improved overall accuracy but has also addressed challenges in categorizing visually similar and complex groups. By leveraging the combined strengths of ResNet and three ViTs models, the PCVM has significantly elevated recognition capabilities, particularly in categories where individual models traditionally falter. Our experimental results affirm the efficacy of this approach, demonstrating notable reductions in misclassification rates through a unique convolutional and voting layer system that synthesizes information across models and employs probabilistic methods to refine decision-making processes.

Furthermore, the architecture and principles underlying the PCVM extend beyond the realm of Chinese herbal medicine, offering promising applications in other critical areas of medical and biological image recognition, such as cancer detection and pathological image analysis. This versatility underscores the potential of PCVM to contribute broadly to high-precision and robust image-based diagnostic tools, setting a new standard in the field of medical image analysis.

Future research should focus on exploring the application of additional models within the PCVM framework. There is a need to optimize aspects of PCVM that negatively affect its performance. Furthermore, efforts should be made to enhance the computational efficiency of PCVM, as it currently demands substantial computational resources. These improvements could further refine the model's utility and applicability in complex image recognition tasks.

References

[1] Gary Nestler. "Traditional Chinese medicine". In: *Medical Clinics* 86.1 (2002), pp. 63–73.
[2] Ved P Kamboj. "Herbal medicine". In: *Current science* 78.1 (2000), pp. 35–39.

[3] Lin Ang et al. “Herbal medicine for the treatment of coronavirus disease 2019 (COVID-19): a systematic review and meta-analysis of randomized controlled trials”. In: *Journal of Clinical Medicine* 9.5 (2020), p. 1583.
[4] Ming Lyu et al. “Traditional Chinese medicine in COVID-19”. In: *Acta Pharmaceutica Sinica B* 11.11 (2021), pp. 3337–3363.
[5] Yi-Zeng Liang, Peishan Xie, and Kelvin Chan. “Quality control of herbal medicines”. In: *Journal of chromatography B* 812.1-2 (2004), pp. 53–70.
[6] Shilong YANG et al. “Current Situation and Thinking on "Odor and Taste" Identification of Traditional Chinese Medicine.” In: *World Science and Technology-Modernization of Traditional Chinese Medicine* (2014), pp. 1876–1879.
[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.
[8] Neena Aloysius and M Geetha. “A review on deep convolutional neural networks”. In: *2017 international conference on communication and signal processing (ICCSP)*. IEEE. 2017, pp. 0588–0592.
[9] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
[10] Laith Alzubaidi et al. “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions”. In: *Journal of big Data* 8 (2021), pp. 1–74.
[11] Rahul Chauhan et al. “Convolutional neural network (CNN) for image detection and recognition”. In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. IEEE. 2018, pp. 278–282.
[12] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
[13] Andrés Hernández-Serna and Luz Fernanda Jiménez-Segura. “Automatic identification of species with neural networks”. In: *PeerJ* 2 (2014), e563.
[14] José Maurício, Inês Domingues, and Jorge Bernardino. “Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review”. In: *Applied Sciences* 13.9 (2023). ISSN: 2076-3417. DOI: 10.3390/app13095521. URL: https://www.mdpi.com/2076-3417/13/9/5521.
[15] Adibaru Kiflie Mulugeta, Durga Prasad Sharma, and Abebe Haile Mesfin. “Deep learning for medicinal plant species classification and recognition: a systematic review”. In: *Frontiers in Plant Science* 14 (2024), p. 1286088.
[16] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv* (2014). eprint: 1412.6980 (cs.LG).
[17] José Maurício, Inês Domingues, and Jorge Bernardino. “Comparing vision transformers and convolutional neural networks for image classification: A literature review”. In: *Applied Sciences* 13.9 (2023), p. 5521.
[18] Michal Filipiuk and Vasu Singh. “Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems.” In: *SafeAI@ AAAI*. 2022.18Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.
[19] Stéphane Cuenat and Raphaël Couturier. “Convolutional neural network (cnn) vs vision transformer (vit) for digital holography”. In: *2022 2nd International Conference on Computer, Control and Robotics (ICCCR)*. IEEE. 2022, pp. 235–240.

[20] Khushal Tyagi et al. “Detecting pneumonia using vision transformer and comparing with other techniques”. In: *2021 5th international conference on electronics, communication and aerospace technology (ICECA)*. IEEE. 2021, pp. 12–16.
[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.
[22] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833.
[23] Yann LeCun et al. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551.
[24] Fengxiang He, Tongliang Liu, and Dacheng Tao. “Why resnet works? residuals generalize”. In: *IEEE transactions on neural networks and learning systems* 31.12 (2020), pp. 5349–5362.
[25] Michal Drozdzal et al. “The importance of skip connections in biomedical image segmentation”. In: International workshop on deep learning in medical image analysis, international workshop on large-scale annotation of biomedical data and expert label synthesis. Springer. 2016, pp. 179–187.
[26] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
[27] Nicolas Carion et al. “End-to-end object detection with transformers”. In: *European conference on computer vision*. Springer. 2020, pp. 213–229.
[28] Sixiao Zheng et al. “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 6881–6890.
[29] Ziyu Wan et al. “High-fidelity pluralistic image completion with transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 4692–4701.
[30] Kai Han et al. “A survey on vision transformer”. In: *IEEE transactions on pattern analysis and machine intelligence* 45.1 (2022), pp. 87–110.
[31] Xiangzuo Huo et al. “HiFuse: Hierarchical multi-scale feature fusion network for medical image classification”. In: *Biomedical Signal Processing and Control* 87 (2024), p. 105534.
[32] Yiheng Yan. *Chinese Herbal Medicine Dataset*. https://www.kaggle.com/datasets/ yihengyan111/ chinese-herbal-medicine. 2024.
[33] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *CoRR* abs/2010.11929 (2020). arXiv: 2010.11929. URL: https://arxiv. org/abs/2010. 11929.
[34] *MS Windows NT Kernel Description*. https://huggingface.co/nateraw/vit-base- patch16-224-cifar10. Accessed: 2010-09-30.
[35] *MS Windows NT Kernel Description*. https://huggingface.co/amunchet/rorshark-vit-base. Accessed: 2010-09-30.
[36] Sheldon Mascarenhas and Mukul Agarwal. “A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification”. In: *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*. Vol. 1. IEEE. 2021, pp. 96–99.