Research Progress on Plant Image Classification Method Based on Convolutional Neural Networks

Mingliang Ge[1], Wei Wang[2,*], Jun Li[1], Junpeng Pei[1], Yousong Wang[1]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China
[2]PLA Naval Medical Center, Naval Medical University, Shanghai, China
[*]Corresponding author: wwang_fd@fudan.edu.cn

Abstract: *Plants are the most diverse organisms on Earth and play an irreplaceable role in maintaining ecological balance. Traditional plant identification relies on experience-based classification, which can lead to subjectivity and human error, resulting in instances where identification is either impossible or incorrect. Deep learning enables the automatic recognition and classification of plants by training neural network models, with convolutional neural networks (CNNs) being a significant technology in deep learning that demonstrates strong advantages in enhancing model performance and accuracy. This paper presents an overview of the development history of CNN models, reviews recent research utilizing these models for plant classification, and discusses future trends in the application of CNNs for plant classification.*

Keywords: *Plant Identification, Deep Learning, Convolutional Neural Networks, Image Classification*

1. Introduction

Plants are the most diverse group of organisms on Earth, widely distributed across terrestrial environments, and they play a central role in ecosystems, material cycles, and energy flow. Their contribution to maintaining balance in the natural world is irreplaceable. Understanding the diversity, characteristics, and functions of plants not only satisfies people's desire to connect with nature and protect the environment but also aids in improving ecosystems and advancing agricultural development. However, due to the vast number and complexity of plant species on Earth, identification typically relies on the overall morphology of the plant or the observation of organs such as flowers, leaves, and fruits. Traditional plant identification depends on the expertise of botanists, who use their senses of sight, smell, and taste to determine species. However, the limited range of plants they can recognize, lack of experience, or the difficulty in distinguishing between species may introduce subjectivity and human error, leading to misidentification. This traditional method, based on human knowledge and expertise, has limitations in both accuracy and efficiency. Therefore, applying modern technological methods to the scientific classification of plants has become increasingly important.

Artificial intelligence (AI) has made significant strides over the past two decades. The rapid development of AI has given rise to a popular topic known as deep learning. Deep learning techniques have achieved remarkable results in fields such as image classification, image segmentation, speech recognition, and natural language processing. By training neural network models, deep learning can automatically recognize and classify plants. Convolutional neural networks (CNNs), a key technology in deep learning, have demonstrated strong advantages in enhancing model performance and accuracy. Image-based approaches are considered the most suitable methods for plant classification. Therefore, applying these advanced deep learning techniques to plant classification can not only improve the accuracy of classification but also save time and labor, increasing efficiency.

2. Image classification method based on convolutional neural networks

With the continuous advancement of artificial intelligence, deep learning technologies have also been rapidly evolving. Deep learning-based image classification methods achieve precise classification by leveraging neural networks to learn and extract features and structures from images. The primary mechanism for this classification is the convolutional neural network (CNN). Some of the commonly used CNN architectures in the field of deep learning include LeNet, AlexNet, GoogleNet, VGG, ResNet, DenseNet, MobileNet, and EfficientNet.

2.1. LetNet

LeNet[1] is a classic convolutional neural network (CNN) architecture proposed by Yann LeCun and colleagues in 1998, originally designed for handwritten digit recognition tasks, where it achieved outstanding results, particularly on the MNIST dataset. The basic structure of LeNet includes multiple convolutional layers, pooling (subsampling) layers, and fully connected layers. As illustrated in Figure 1, LeNet's typical architecture consists of an input layer that accepts grayscale images of 32x32 pixels, followed by a convolutional layer that uses multiple convolutional kernels to extract features and generate several feature maps. These feature maps are then reduced in size through average pooling in the subsequent pooling layer, thereby decreasing computational complexity. Another convolutional layer is used to extract higher-level features, followed by another pooling layer that further reduces the feature maps' dimensions. The resulting 2D feature maps are flattened into 1D vectors and passed through a fully connected layer for classification, with the final output typically produced by a softmax layer.

LeNet's design principles laid the foundation for modern convolutional neural networks and inspired the development of later deep learning architectures. However, one of the main limitations of LeNet is its relatively simple structure and low parameter count, making it suitable only for small datasets and simple tasks. When handling modern large-scale, complex tasks, LeNet's performance is insufficient, as it is not well-suited for high-resolution images or large datasets.

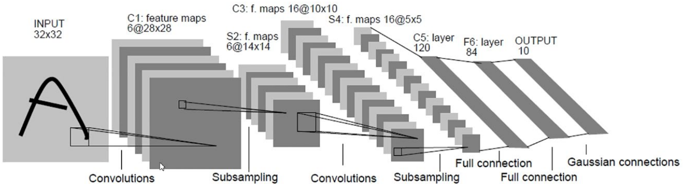


Figure 1: The structure of the LeNet.

2.2. AlexNet

AlexNet[2], introduced by Alex Krizhevsky and colleagues in 2012, marked a revolutionary achievement in the ImageNet image classification challenge, significantly advancing the field of deep learning. The architecture comprises eight layers of neural networks, as illustrated in Figure 2. AlexNet leveraged the ReLU activation function to accelerate training and used dropout to mitigate overfitting. It also incorporated data augmentation and GPU parallelism to handle large-scale image datasets. Key innovations of AlexNet include larger convolutional kernels, a deeper network structure, and local response normalization (LRN) between pooling layers. However, AlexNet has several limitations, including a large number of parameters, particularly in the fully connected layers, leading to high memory and computational demands. Additionally, for modern tasks, its structure is considered too wide and shallow, lacking the efficiency and scalability found in more recent architectures.

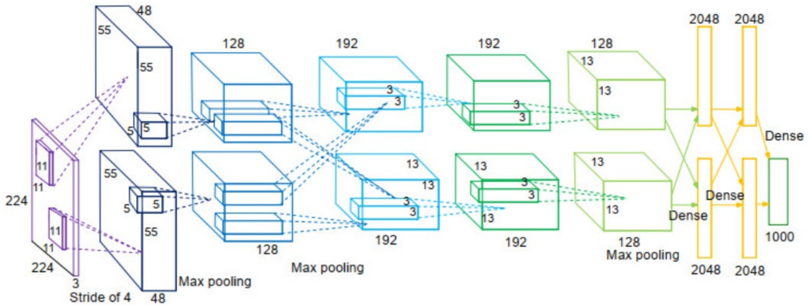
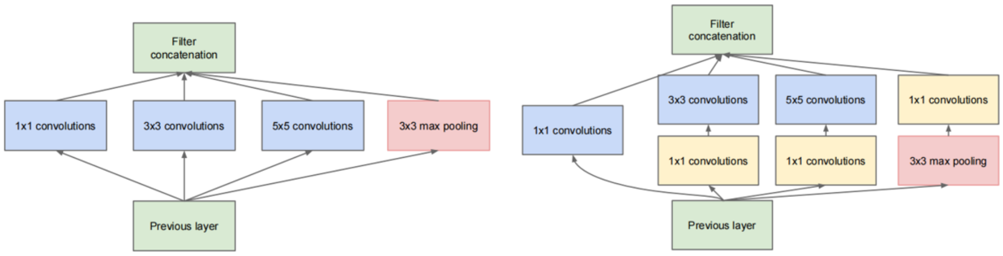


Figure 2: The structure of the AlexNet.

2.3. GoogleNet

GoogLeNet[3], a deep convolutional neural network architecture proposed by Szegedy et al. in 2014, achieved remarkable success in the ImageNet competition, thanks to its innovative Inception module. The core idea of GoogLeNet lies in the Inception module, as illustrated in Figure 3, which [3]performs multiple convolution and pooling operations at different scales within the same layer to capture features at various resolutions, thereby enhancing the network's representational capacity. This approach avoids the computational complexity associated with simply increasing the network's depth while significantly reducing the number of parameters.GoogLeNet employs a 22-layer deep architecture and replaces traditional fully connected layers with global average pooling, further reducing the risk of overfitting and decreasing the number of parameters. Additionally, it introduces auxiliary classifiers to assist with gradient backpropagation during training. This design significantly improves image classification performance while maintaining computational efficiency, marking a key advancement in the field of deep learning. However, GoogLeNet has its drawbacks. The Inception module is relatively complex, making it difficult to implement and fine-tune. Although the number of parameters is reduced, the computational cost remains high, and the network's deep structure can still lead to vanishing gradient problems.



(a) Inception module, naïve version

(b) Inception module with dimensionality reduction

Figure 3: The structure of the GoogleNet.

2.4. Vgg

The VGG[4] (Visual Geometry Group) network is a deep convolutional neural network architecture proposed by the Visual Geometry Group at the University of Oxford in 2014, with its name derived from the research group. VGG gained widespread attention for its simple yet effective network design. The key feature of VGG is the use of stacked 3x3 small convolutional kernels to increase the network depth, rather than relying on larger convolutional kernels, as illustrated in Figure 4. The VGG network has several variants, with the most famous being VGG16 and VGG19, which contain 16 and 19 layers, respectively. The design principle of VGG is to maintain a consistent structure of convolutional and pooling layers, while improving image classification accuracy by increasing network depth. Although VGG has a large number of parameters and a high computational complexity, it performed exceptionally well in the ImageNet competition, becoming a classic architecture in deep learning. It has since been widely used for tasks such as transfer learning and feature extraction. However, VGG also has its drawbacks. The network's parameter count is extremely large, particularly in the fully connected layers, resulting in high memory and computational demands. Moreover, the network's computational efficiency is relatively low. While VGG offers high accuracy, its computational complexity and training time make it cumbersome compared to more modern architectures.

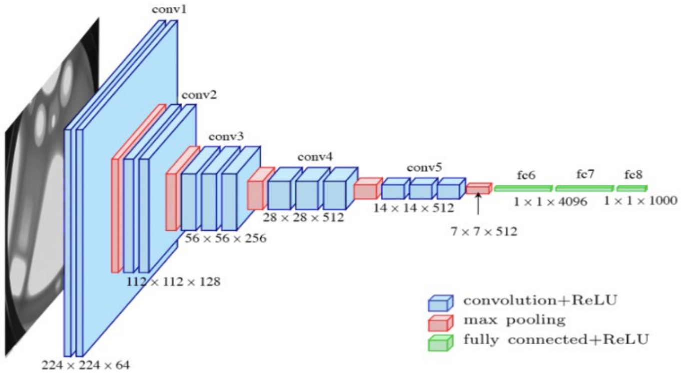


Figure 4: The structure of the Vgg.

2.5. ResNet

ResNet[5] (Residual Network) is a deep convolutional neural network architecture proposed by He K. et al. in 2015, designed to address the problems of vanishing gradients and model degradation that occur as network depth increases. The key innovation of ResNet lies in the introduction of residual blocks, which use skip connections. These connections bypass certain convolutional layers and add the input directly to the output, enabling residual learning. This structure allows the network to learn the residuals between the input and output, rather than directly learning complex mappings, ensuring that very deep networks can be trained effectively. ResNet's depth can reach 50, 101, or even 152 layers, significantly enhancing training stability while maintaining high accuracy. It achieved outstanding results in the ImageNet competition and has since become the foundation for many subsequent deep learning model designs. The architecture of the residual block is illustrated in Figure 5. While the residual blocks in ResNet mitigate the issue of vanishing gradients, increasing network depth also leads to a significant rise in computational complexity and memory requirements. Moreover, the residual connections may have limited effectiveness for certain tasks, and the model structure can be somewhat rigid.

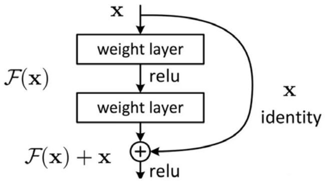

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 5: The structure of the Residual module.

2.6. MobileNet

MobileNet[6] is a lightweight convolutional neural network architecture proposed by Howard A. G. et al. in 2017, specifically designed for environments with limited computational resources, such as mobile devices and embedded systems. The key innovation of MobileNet is the introduction of depthwise separable convolutions, which decompose the standard convolution operation into two steps: depthwise convolutions and pointwise convolutions. This structure, as shown in Figure 6, significantly reduces computational costs and the number of parameters while maintaining high accuracy. This design dramatically lowers the model's complexity, enabling real-time inference on mobile devices with low latency and power consumption. MobileNet features adjustable width and resolution scaling parameters, allowing a trade-off between accuracy and efficiency, making it a popular neural network architecture for mobile computing, real-time applications, and edge computing. However, despite its suitability for mobile devices, MobileNet's performance can fall short on high-precision tasks compared to larger networks, particularly in scenarios where more computational resources are available. Additionally, the feature-capturing capability of depthwise separable convolutions may be weaker, limiting its performance on more complex tasks.

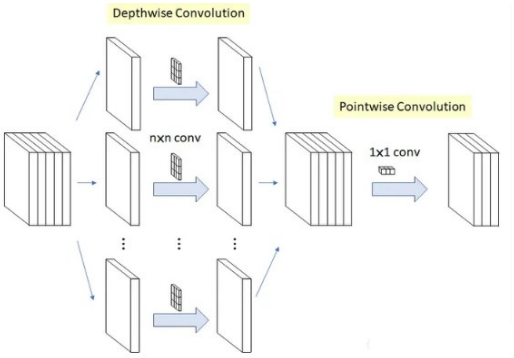


Figure 6: The structure of the Depth-separable convolution.

2.7. DenseNet

DenseNet[7] (Densely Connected Convolutional Networks) is a deep convolutional neural network architecture introduced by researchers from the University of Washington and Facebook AI Research in 2017. Its core innovation is the use of dense connections, where each layer is directly linked to all subsequent layers, maximizing feature reuse and reducing redundant computation. The structure is shown in Figure 7, the structure mitigates the vanishing gradient problem, decreases parameter count, and enhances feature propagation. While DenseNet excels in image classification tasks, particularly with limited data, it also has drawbacks. The dense connections can lead to increased memory usage and computational overhead, especially with high-resolution images, and can complicate scaling to very large tasks.

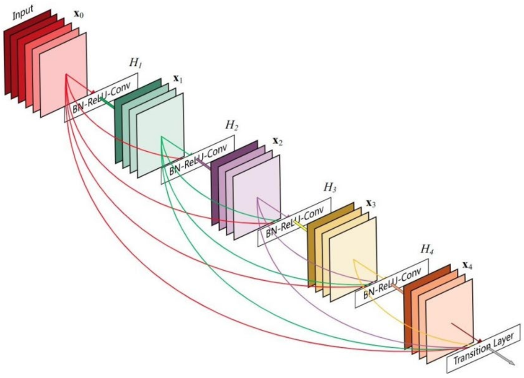


Figure 7: The structure of the DenseNet.

2.8. EfficientNet

EfficientNet[8] is a high-performance convolutional neural network architecture proposed by Google in 2019, designed to balance model accuracy and efficiency through a comprehensive scaling approach. The core idea of EfficientNet is to simultaneously adjust the network's depth, width, and resolution, rather than modifying any single dimension in isolation, achieving better performance under constrained computational resources. The structure of this comprehensive scaling is illustrated in Figure 8. EfficientNet is based on a convolution module known as MBConv, which combines MobileNet's depthwise separable convolutions with SENet's channel attention mechanism, further enhancing computational efficiency. Compared to traditional networks, EfficientNet achieves higher accuracy with the same computational resources, making it a powerful model for tasks such as image classification, object detection, and semantic segmentation. However, despite its efficient performance through compound scaling, the design of EfficientNet relies on complex hyperparameter tuning, making it less flexible in adapting to new tasks. Additionally, there may be scalability limitations for extremely large tasks, and the complexity of the MBConv module increases the difficulty of model implementation.

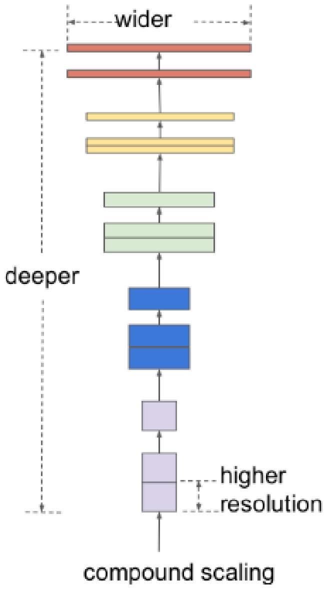


Figure 8: The structure of the EfficientNet.

3. Plant image classification method based on convolutional neural networks

Liu et al[9]. proposed a new method for plant leaf classification using a ten-layer CNN. To improve classification accuracy, they performed data augmentation on the leaf sample images, thereby expanding the database. Experiments on the Flavia dataset, which contains 4,800 leaf images across 32 species, yielded an overall accuracy of 87.92%.

Huang et al[10]. proposed an improved AlexNet model to enhance feature extraction by inserting an additional convolutional layer after the fourth layer of AlexNet to filter out more effective features. Results from their custom dataset showed that the combination of data augmentation and the improved AlexNet model achieved an accuracy of 94% on a traditional Chinese medicine image dataset.

Wagle et al[11]. employed an AlexNet-based network to classify plant leaf diseases, using the network to distinguish between healthy and diseased plants. They trained the CNN with a mixed combination of healthy and diseased leaf data. Utilizing transfer learning with a pre-trained AlexNet network, they trained the network on various amounts of data, achieving an accuracy of 91.15% on the PlantVillage dataset.

Fu et al[12]. used GoogLeNet to classify six types of fruits and vegetables, optimizing GoogLeNet to improve training speed and recognition accuracy. By reducing the number of convolutional kernels and adjusting the Inception structure, they decreased the number of parameters by nearly 48%. They also introduced a new activation function, Swish, and a DropBlock layer between convolutional layers, resulting in a classification accuracy of 98.88%.

Siddharth et al[13]. used VGG19 to classify colored images of Swedish tree leaves. The VGG-19 classifier utilized predefined hidden layers, such as convolutional layers, max pooling layers, and fully connected layers, to capture the features of the leaves, ultimately using a softmax layer to generate feature representations for all plant categories. The dataset included 15 tree species, achieving an accuracy of 99.70%.

Campos et al[14]. proposed a simplified model called N-VGG. N-VGG reduces overfitting observed in VGG16 by using the fewest possible trainable parameters. This model replaces the flatten layer in the VGG architecture with a global average pooling layer to reduce the size of the feature vector. Additionally, it eliminated one of the fully connected layers and introduced a new hyperparameter N to indicate the number of nodes in the remaining layers.

Vaidehi et al[15]. employed the ResNet50 architecture using images of six categories of Ayurvedic plants with an ECOC framework. The support vector machine classifier used a one-vs-all encoding design. They extracted image features using a pre-trained CNN and classified the data based on these extracted features. The class labels were generated through grouping, achieving an accuracy of 93%.

Ali et al[16]. proposed a more efficient method for leaf classification using transfer learning to identify plants, first learning useful leaf features directly from the input data representation using a pre-trained deep neural network model. They then employed a logistic regression classifier for leaf classification. Testing on the public datasets Flavia and Leafsnap resulted in accuracies of 99.6% and 90.54%, respectively.

Lasya et al[17]. introduced two new hybrid models utilizing a combination of deep learning architectures for accurate plant species classification. The first hybrid model combined InceptionNetV3 and MobileNet, achieving an accuracy of 88.11%. The second hybrid model integrated AlexNet and MobileNet, attaining an accuracy of 93.86%, outperforming the first model. Both hybrid models utilized deep learning classifiers, highlighting the effectiveness of hybrid models in enhancing the accuracy of plant species classification and advancing plant identification.

Wu[18] proposed a leaf recognition algorithm based on an attention mechanism and dense connections to address the diversity of leaf morphology. They first implemented DenseNet for cross-layer learning, effectively enhancing the network's generalization ability to intra-class variance while improving its ability to learn discriminative features such as leaf texture. Additionally, the attention mechanism was employed to further enhance the network's capacity for learning distinguishing features of plant leaves. Experimental results indicated an accuracy of 97.7%.

Kumar et al[19]. presented a non-averaged DenseNet-169 (NADenseNet-169) CNN architecture, which reduced performance (lowering accuracy and increasing computation time) by inserting a 7 × 7 Global Average Pooling (GAP) layer into the architecture. The NADenseNet-169 model was tested on a non-augmented dataset, significantly decreasing the number of mispredictions for real-time images. Consequently, the NADenseNet-169 model trained on the Leaf-12 dataset is well-suited for real-time plant species recognition systems.

Arun et al[20]. classified 11 different plant leaf categories using the EfficientNetB5 model. This method alleviated the reliance on the physical features of plant morphology, eliminating the need for preprocessing of leaf images. The features were extracted and classified by the pre-trained network, achieving an accuracy of 98.43% on the test set.

4. Common datasets and evaluation indicators

4.1. Common datasets

This section introduces public data sets commonly used in current plant taxonomic research, as shown in Table 1.

Table 1: Common datasets.

Datasets	Classes	Num	Location	Source link
Flavia	32	1907	Leaf	https://flavia.sourceforge.net/
Swedish leaf	15	1125	Leaf	https://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/
D_leaf	43	1290	Leaf	https://figshare.com/articles/dataset/D-Leaf_Dataset/5732955
Leafsnap	185	30866	Leaf	https://leafsnap.com/dataset/
Oxford Flower17	17	1362	Flower	https://www.robots.ox.ac.uk/vgg/data/flowers/
Oxford Flower102	102	8189	Flower	https://www.robots.ox.ac.uk/vgg/data/flowers/
MEW2012	153	9745	Herb	http://zoi.utia.cas.cz/node/662
ImageCLEF2011	71	5436	Herb	https://www.imageclef.org/2011/plants
ImageCLEF2013	250	26077	Herb	https://www.imageclef.org/2013/plant
PlantCLEF2014	500	60000	Herb	https://www.imageclef.org/2014/lifeclef/plant
PlantCLEF2015	1000	113205	Herb	https://www.imageclef.org/lifeclef/2015/plant
LifeCLEF2020	1000	60000	Herb	https://www.imageclef.org/PlantCLEF2020
LifeCLEF2021	1000	60000	Herb	https://www.imageclef.org/PlantCLEF2021
PlantCLEF2022	80000	4000000	Herb	https://www.imageclef.org/PlantCLEF2022
PlantCLEF2023	80000	4000000	Herb	https://www.imageclef.org/PlantCLEF2023

4.2. Evaluation indicators

The metrics are based on the confusion matrix, which contains information about true and false classified pixels relative to the ground truth of the classification. This includes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP refers to the number of samples correctly predicted as the positive class, while TN denotes the number of samples accurately predicted as the negative class. Conversely, FP represents the number of samples incorrectly predicted as the positive class, and FN indicates the number of samples incorrectly predicted as the negative class. Evaluation metrics such as classification accuracy, precision, recall, and F1 score can be computed based on the confusion matrix.

4.2.1. Confusion Matrix

The classification results are presented in matrix form, showcasing the correspondence between predicted outcomes and actual results across different categories. This format is suitable for in-depth analysis of the model's performance across various classes, providing a comprehensive perspective for evaluating model effectiveness. Particularly in scenarios where the dataset is imbalanced, it can assist in identifying categories where the model performs poorly. The formulas for the evaluation indicators such as accuracy, accuracy, recall and F1 score are shown below.

$$A_{\text{ccuracy}} = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{1}$$

$$P_{\text{recision}} = \frac{T_P}{T_P+F_P} \tag{2}$$

$$R_{\text{ecall}} = \frac{T_P}{T_P+F_N} \tag{3}$$

$$F_1 = \frac{2\times P_{\text{recision}} \times R_{\text{ecall}}}{P_{\text{recision}} + R_{\text{ecall}}} \tag{4}$$

5. Conclusion

In the field of plant classification, tasks based on CNNs still have vast development potential. First, improving datasets is crucial; existing datasets need to be optimized and innovated for specific plant classification tasks, while also considering variations in plants across different environments, seasons, and perspectives. These are pressing challenges that need to be addressed for practical recognition. Secondly, interdisciplinary technological integration will create new opportunities for plant classification. For instance, the combination of Transformer architectures with CNNs has demonstrated positive effects on enhancing model performance, which could yield unexpected results when applied to the field of plant classification. Finally, innovations in new models and modules will inject fresh vitality into this area.

Although no groundbreaking architectures akin to VGG and ResNet have emerged in recent years, technological advancements are expected to lead to the development of more efficient and accurate models.

This paper provides an in-depth exploration of the classic models, fundamental principles, and advantages and disadvantages of deep learning-based CNNs. It also reviews recent studies utilizing CNNs for plant classification, evaluating the performance of various models across different classification tasks, and highlights several models that have achieved remarkably high accuracy, providing references for further enhancing plant classification performance. Additionally, common publicly available plant datasets are summarized, along with the various metrics used to evaluate model performance. Finally, the paper discusses future trends in CNN-based plant classification.

References

[1] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-324.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. Advances in neural information processing systems, 2012, 25.

[3] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2015.

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:14091556, 2014.

[5] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016.

[6] Howard A G. Mobilenets: Efficient convolutional neural networks for mobile vision applications [J]. arXiv preprint arXiv:170404861, 2017.

[7] Huang G, Liu Z, Van, et al. Densely connected convolutional networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017.

[8] Tan M. Efficientnet: Rethinking model scaling for convolutional neural networks [J]. arXiv preprint arXiv:190511946, 2019.

[9] Liu J, Yang S, Cheng Y, et al. Plant leaf classification based on deep learning[C]. Proceedings of the 2018 Chinese Automation Congress (CAC), F, 2018. IEEE.

[10] Huang F, Yu L, Shen T, et al. Chinese herbal medicine leaves classification based on improved AlexNet convolutional neural network[C]. Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), F, 2019. IEEE.

[11] Wagle S A. Comparison of Plant Leaf Classification Using Modified AlexNet and Support Vector Machine [J]. Traitement du Signal, 2021, 38(1).

[12] Yuesheng F, Jian S, Fuxiang X, et al. Circular fruit and vegetable classification based on optimized GoogLeNet [J]. IEEE Access, 2021, 9: 113599-611.

[13] Siddharth T, Kirar B S, Agrawal D K. Plant species classification using transfer learning by pretrained classifier VGG-19 [J]. arXiv preprint arXiv:220903076, 2022.

[14] Campos-Leal J A, Yee-Rendón A, Vega-López I F. Simplifying vgg-16 for plant species identification [J]. IEEE Latin America Transactions, 2022, 20(11): 2330-8.

[15] Vaidehi M V, Vinod M V. ResNet based classification in CNN for ayurvedic plant categorization using deep learning [J]. Design Engineering, 2021: 1507-16.

[16] Beikmohammadi A, Faez K. Leaf classification for plant recognition with deep transfer learning[C]. Proceedings of the 2018 4th Iranian Conference on Signal Processing and Intelligent Systems, F, 2018.

[17] Lasya. S , Jyothsna. S , Pushpa B R .Optimized Plant Species Classification through MobileNet-Enhanced Hybrid Models[C].2024 5th International Conference for Emerging Technology (INCET).[2024-10-30].DOI:10.1109/INCET61516.2024.10593020.

[18] Wu H, Shi Z, Huang H, et al. Automatic Leaf Recognition Based on Attention DenseNet[C]. proceedings of the Cognitive Systems and Signal Processing: 5th International Conference, ICCSIP 2020, Zhuhai, China, December 25–27, 2020, Revised Selected Papers 5, F, 2021. Springer.

[19] Sathiesh Kumar V, Anubha Pearline S. Real-Time Plant Species Recognition Using Non-averaged DenseNet-169 Deep Learning Paradigm[C]. proceedings of the International Conference on Computer Vision and Image Processing, F, 2022. Springer.

[20] Arun Y, Viknesh G. Leaf classification for plant recognition using EfficientNet architecture[C]. proceedings of the 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC), F, 2022. IEEE.