GRAPE LEAF VARIETY RECOGNITION BASED ON THE AF-SWIN TRANSFORMER MODEL
/
基于 AF-Swin Transformer 模型的葡萄叶片品种识别

Changmei LIANG[1)], Jiaxiong GUAN[1)], Tongtong GAO[1)], Juxia LI *[1)], Yanwen LI[1)], Qifeng ZHAO[2)], Pengfei WEN[3)], Zhifeng BI[4)], Fumin MA[5)]

[1)] College of Information Science and Engineering, Shanxi Agricultural University, Jinzhong, Shanxi/ China
[2)] Shanxi Academy of Agricultural Sciences Polomogy Institute, Jinzhong, Shanxi/ China
[3)] College of Horticulture, Shanxi Agricultural University, Jinzhong, Shanxi/ China
[4)] Department of Basic Sciences, Shanxi Agricultural University, Jinzhong, Shanxi/ China
[5)] College of Energy and Power Engineering, Lanzhou University of Technology, Lanzhou, Gansu/ China
Tel: 15803446486; E-mail: lijxsn@126.com
Corresponding author: Juxia Li
DOI: https://doi.org/10.35633/inmateh-75-92

Keywords: *Grape leaves; variety recognition; Swin Transformer; Focal Loss.*

ABSTRACT
Aiming at the problem of differentiated cultivation strategies for different grape varieties, the AF-Swin Transformer model is proposed in this study. Firstly, Focal Loss is used to effectively tackle data imbalance in grape leaves. Secondly, the AdamW optimizer is selected to better control model complexity and improve generalization. The results show that the training accuracy of AF-Swin Transformer model is 7.87 percentage points higher than that of the original Swin Transformer model. Precision and recall improved by 4.4 and 4.3 percentage points, respectively. This study enables accurate automated variety monitoring within vineyard cultivation systems, assisting growers in implementing targeted cultivation strategies.

摘要
针对不同葡萄品种栽培策略存在差异化问题，本研究提出了 AF-Swin Transformer 模型。首先，引入 Focal Loss，有效应对葡萄叶片数据不平衡，其次，选用 AdamW 优化器，更好地控制模型复杂度并提高泛化能力。结果表明，AF-Swin Transformer 模型的训练集准确比原始 Swin Transformer 模型提高了 7.87 个百分点；精准率和召回率分别提高了 4.4 和 4.3 个百分点。本研究能够在葡萄园中种植系统中实现准确的自动化品种监测，帮助种植者实施有针对性的种植策略。

INTRODUCTION
Effective recognition of grape leaf varieties can assist grape growers in managing their crops more conveniently and making precise decisions (*Cecotti et al., 2020*). During the development and maturation of grapes, they are susceptible to various diseases. Understanding the susceptibility of specific leaf varieties to certain diseases and their effective identification will enhance targeted prevention and treatment of grape diseases (*Pereira et al., 2019*). Traditional methods for identifying grape varieties primarily depend on manual observation. While this method is straightforward, it is often affected by subjective factors and environmental changes, resulting in insufficient accuracy and stability in identification. Therefore, achieving automated identification of grape varieties through leaf image analysis will provide growers with more convenient management tools, helping them make more precise decisions and ultimately enhance the efficiency and competitiveness of the entire grape industry.

In recent years, the rapid development of deep learning technologies and computer vision has brought new solutions for plant leaf recognition (*Patricio et al., 2018*). Convolutional Neural Networks (CNNs) such as AlexNet (*Ni et al., 2021)*, MobileNet *(Zou et al., 2024)*, and ResNet *(Yang et al., 2023)* can automatically extract image features (*Pushpanathan et al., 2021*) and have demonstrated superior performance in the classification tasks of different leaf varieties from the same plant. Deep learning has particularly become an effective tool for leaf feature extraction and variety recognition.

---

[1] *Changmei Liang, Prof.; Jiaxiong Guan, M.Sc.Stu.; Tongtong Gao, Juxia Li, Prof. Ph.D. Eng.; Yanwen Li, Lecturer M.S. Eng.; Qifeng Zhao,Res.; Pengfei Wen, Prof.;Zhifeng Bi,Lec.; Fumin Ma, M.S. Stud. Eng.*

Yin et al. employed the GoogLeNet model to recognize leaf images of 11 camellia plant varieties, achieving an overall accuracy of 94.1% (*Yin et al., 2023)*. Lin et al. used ResNet50 for variety recognition of 9 types of Wuyishan Fujian tea leaves, with an accuracy rate reaching 96.04% (*Lin et al., 2021*). Sun et al. conducted research on the recognition of southern medicinal leaf varieties in complex backgrounds using an improved EfficientNetv2 model, achieving an accuracy rate of 99.12% (*Sun et al., 2023*). Chen et al. utilized a Multi-Attention Fusion Convolutional Neural Network (MAFNet) for recognizing apple leaf images, with the model achieving an accuracy rate of 98.14% (*Chen et al., 2022*). Tavakoli et al. applied Convolutional Neural Networks (CNNs) for the variety recognition of 12 types of legumes, where the model showed good recognition performance on a dataset of legume leaf back images, achieving an accuracy rate of 95.86% (*Tavakoli et al., 2021*). Dong et al. used an improved RegNet model to recognize varieties of 118 mature camellia leaves that grew under natural light conditions and achieved an overall accuracy of 93.7% (*Dong et al., 2024*). Su et al. used an improved ResNet50 to recognize datasets of 12 types of wine grape leaf images collected at different growth stages, achieving an accuracy of 88.75% (*Su et al., 2021*). Zhang et al. improved the VOLO-D1 model by integrating the YOLO object detection mechanism and proposed the YOLO-VOLO-LS method, which significantly enhanced the variety identification accuracy of lettuce at the early SP growth stage, achieving a test accuracy of 93.452% (*Zhang et al., 2022)*. Islam et al. applied transfer learning based on the YOLO model to successfully recognize and localize Bangladeshi plant leaves, attaining a classification accuracy of 96% (*Islam et al., 2019*). Das et al. employed the YOLOv7 model to improve the identification and localization of medicinal plant leaves in complex environments, providing technical support for automated recognition in the herbal medicine industry (*Das et al., 2024*). Sennan et al. proposed a convolutional neural network (CNN) for spinach classification, achieving a classification accuracy of 97.5% on a dataset comprising four leaf categories (*Sennan et al., 2022*). Kaur et al. enhanced DenseNet-121 for grapevine variety identification, reaching 96% classification accuracy on high-resolution images of five grape leaf types (*Kaur et al., 2024)*. Maulana et al. conducted a comparative analysis of various CNN models for grapevine leaf classification, with DenseNet and MobileNetV2 both achieving 99% accuracy, thereby improving classification precision and model robustness *(Maulana et al., 2024)*.

Although the deep learning models mentioned above have achieved certain results in crop leaf variety recognition, research on grape leaf variety identification remains limited due to imbalanced sample sizes arising from varying rarity and collection difficulties, and the widely used YOLO model exacerbates this by requiring labor-intensive data annotation. This study focuses on mature grape leaves that have newly sprouted for 30 to 60 days in spring and addresses this issue by proposing the AF-Swin Transformer model. We utilize the AdamW optimizer, which incorporates weight decay. AdamW applies the weight decay term independently during parameter updates, separating it from the learning rate adjustments, which allows for a more precise implementation of weight decay. This approach effectively manages model complexity, reduces overfitting risk, and enhances stability. Moreover, the Focal Loss function is introduced to tackle the sample imbalance problem. Focal Loss mitigates the loss gradients of easily recognizable samples by introducing a modulation factor, encouraging the model to focus more on difficult-to-recognize samples. This enhances the model's ability to recognize rare varieties and improves its learning capacity for hard-to-identify samples, enabling the model to better recognize subtle differences and ultimately increase overall accuracy.

Therefore, the AF-Swin Transformer model proposed in this paper exhibits greater robustness and accuracy in recognizing grape leaf varieties, providing a scientific foundation for their identification.

MATERIALS AND METHODS

Sample dataset

All the leaf samples in this study were collected at the Fruit Tree Research Institute in Taigu District, Jinzhong City, Shanxi Province. A total of 5,516 images were collected using a Huawei Mate 40 smartphone, taken from various angles and time periods in a natural environment. The distribution of the number of samples for each variety is shown in Figure 1.

The leaves were collected from the upper-middle part of the grapevine branches, and they were healthy, mature leaves that had newly sprouted for 30 to 60 days in spring. At this stage, the leaves exhibit standard morphology, with clear leaf lobes, leaf shape, and visible venation structures. The image resolution is 3024x4032, the image format is JPEG, and the color mode is RGB.

Figure 2 displays sample images of 26 different grape varieties.

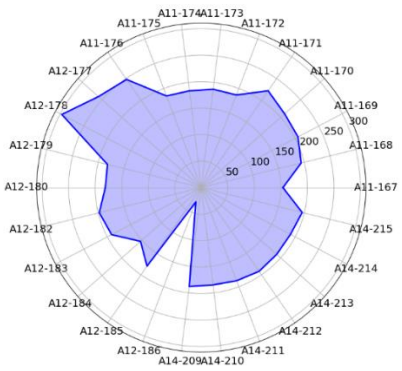


Fig. 1 - Distribution of the number of grape leaf varieties

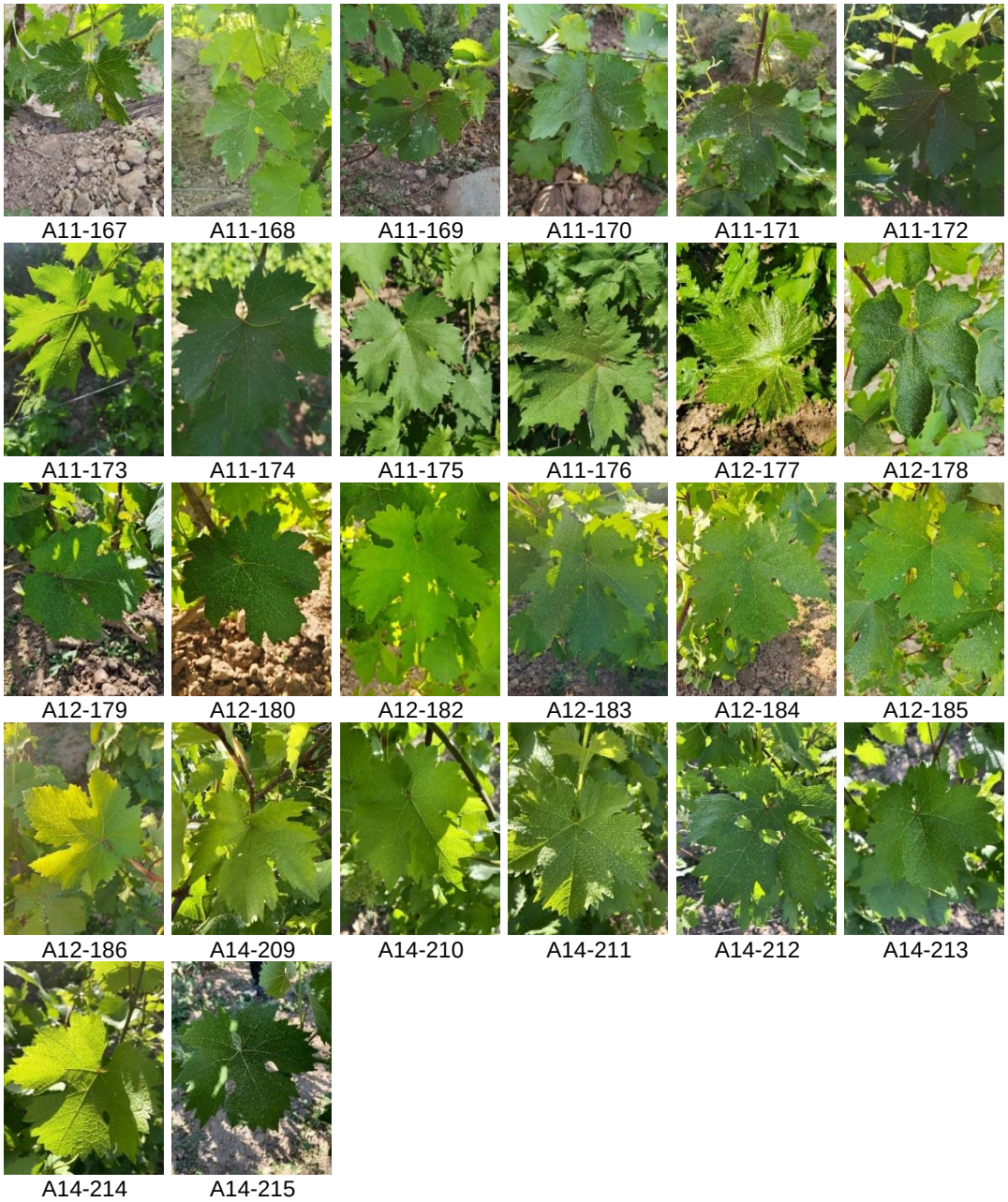


Fig. 2 - Examples of grape leaf samples from different varieties

Data augmentation

To enhance the recognition capability of the network model, four data augmentation methods—random rotation, flipping, brightness adjustment, and adding Gaussian noise—were randomly combined and artificially expanded during training to create a training dataset. The number of images for each variety after data augmentation is shown in Table 1.

This image augmentation technique is versatile and computationally efficient, effectively training deep learning models. The dataset was divided into training, validation, and testing sets in an 8:1:1 ratio. A schematic diagram of the grape leaf data augmentation (Lacate) is illustrated in figure 3. Figure 3(a) shows the original image, (b) displays the flipped image, (c) demonstrates the addition of Gaussian blur, (d) depicts an increase in brightness, (e) shows the rotated image, and (f) illustrates a decrease in brightness.

(a)Original Image | (b) Flip | (c) Gaussian Blur | (d) Increase Brightness | (e)Rotate | (f) Decrease Brightness

Fig. 3 - Data augmentation diagram (Lacate)

Table 1

Image dataset of grape leaf varieties

Variety No.	Species	Total quantity	Training set	Validation set	Test set
A11-167	Suhaike	300	240	30	30
A11-168	Baolgal	380	304	38	38
A11-169	Australia Non-	410	328	41	41
A11-170	Baiyou Malake	420	336	42	42
A11-171	Bigqi Husa	440	352	44	44
A11-172	October	370	296	37	37
A11-173	Baisha Ani	370	296	37	37
A11-174	Calaido	360	288	36	36
A11-175	Saingiovese	370	296	37	37
A11-176	Delguri Mike	490	392	49	49
A12-177	Yiqikema	510	408	51	51
A12-178	Dalbash	590	472	59	59
A12-179	Aliwalne	360	288	36	36
A12-180	Bayangxilie	360	288		36
A12-182	Kalas Rose	390	312	39	39
A12-183	Victory	380	304	38	38
A12-184	Baiwujium	300	240	30	30
A12-185	Heisther	360	288	36	36
A12-186	Kalas	50	40	5	5
A14-209	Lacete	370	296	37	37
A14-210	Aibutri	370	296	37	37
A14-211	Su-38	370	296	37	37
A14-212	Shalele	380	304	38	38
A14-213	Baikakuer	380	304	38	38
A14-214	Dashlei	380	304	38	38
A14-215	Shabash	390	312	39	39

Improved swin transformer model

The Swin Transformer model addresses image tasks through a hierarchical design by dividing the input image into multiple windows, treating the elements within each window as independent tokens, and performing linear embedding to create initial feature representations. This mechanism not only enhances the model's capacity to process large images but also effectively avoids the computational and memory limitations that traditional Transformers face when dealing with high-dimensional inputs.

In several stages, the Swin Transformer blocks progressively extract features and adjusts spatial resolution, with each stage further integrating information through a "Patch Merging" operation, which reduces computational load and increases the number of channels. The key aspect of the Swin Transformer is the combination of a sliding window multi-head self-attention mechanism (SW-MSA) and a multi-layer perceptron (MLP), which enhances the model's ability to learn local and global features. Additionally, layer normalization (LN) operations ensure the model's stability and training effectiveness.

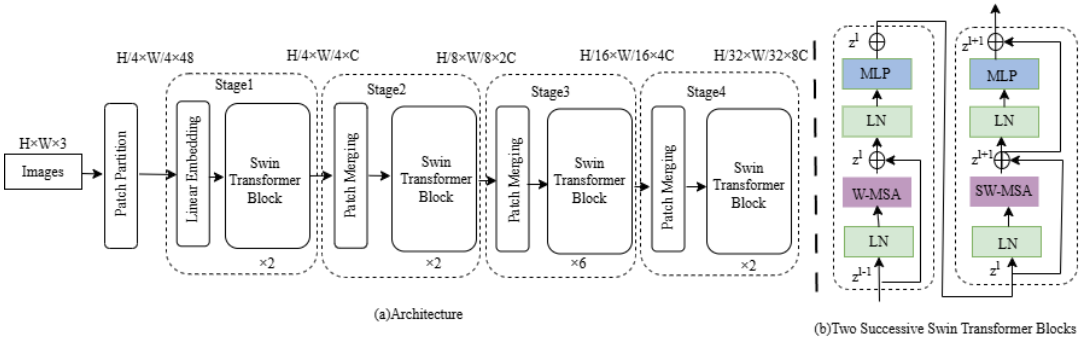


Fig. 4 - Structure diagram of the Swin Transformer model

Due to the imbalance of sample sizes among different grape leaf varieties, this study introduces the Focal Loss function. Focal Loss assesses the difficulty of each sample based on the model's predicted probabilities. It dynamically adjusts the model's focus by reducing attention to easily distinguishable samples during training, allowing the model to concentrate more on harder-to-distinguish samples. Unlike the cross-entropy loss function, Focal Loss introduces a tunable parameter that adjusts the model's focus between easily recognizable and difficult-to-recognize samples. When the value of γ is low, the model pays more attention to easily recognizable samples; when γ is high, the model focuses more on difficult-to-recognize samples. The formula for Focal Loss is as follows:

$$FocalLoss = -\alpha_t \left(1 - p_t\right)^{\gamma} log\left(p_t\right)$$

In the equation, the difficulty of recognition is reflected by p_t. When p_t is larger, it indicates a higher confidence level in identification, suggesting that the sample is easier to distinguish. Conversely, when p_t is smaller, it indicates a lower confidence level in identification, suggesting that the sample is more difficult to distinguish.

Experimental environment

The operating system was 64-bit Windows 10, using an Intel(R) Core(TM) i7-14650HX CPU@2.20 GHz processor with 32 GB of memory. The graphics card model was the NVIDIA GeForce RTX 4060. All CNN models are developed based on the PyTorch framework, with Python 3.8 used as the programming language for implementing network model training and testing. The training parameters included an initial learning rate of 0.00001, a stochastic gradient descent (SGD) optimizer, a weight decay coefficient of 0.05, and a batch size of 4. A cosine learning rate scheduler was utilized for a total of 50 epochs, ensuring a smooth adjustment of the learning rate to avoid sudden changes during training. After each iteration, the trained model was saved in a folder, and the training logs were recorded.

Model evaluation metric

To comprehensively evaluate model performance, several commonly used classification metrics were introduced, including Confusion Matrix, Accuracy, Precision, Recall, F1-score, ROC Curve, and AUC. The ROC Curve, plotted with the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis, represents each point corresponding to a potential classification threshold. The optimal classification threshold can be determined by selecting the point on the ROC Curve that is closest to the top left corner. Furthermore, the model's performance can be assessed using the Area Under the Curve (AUC), where a larger AUC indicates better performance. The formulas for each metric are as follows:

$$ConfusionMatrix = \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \quad (1)$$

$$Presicion = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

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

$$F1 - score = \frac{2 \cdot Presicion \cdot Recall}{Presicion + Rcall} \tag{5}$$

In the equation, TP represents the number of samples predicted as positive that are actually positive; FP represents the number of samples predicted as positive, but are actually negative; FN represents the number of samples predicted as negative, but are actually positive; TN represents the number of samples predicted as negative that are actually negative.

RESULTS AND ANALYSIS

Model training

This study evaluated the performance of four convolutional neural network models: Swin Transformer, MobileNetV2, MobileNetV3, and ViT for grape leaf recognition. The model training accuracy curves and loss change curves are shown in figures 5 and 6; the testing accuracy results are presented in Table 2.

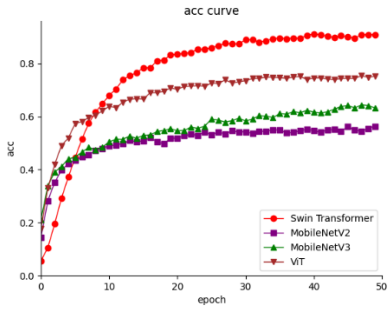


Fig. 5 - Training accuracy change curve

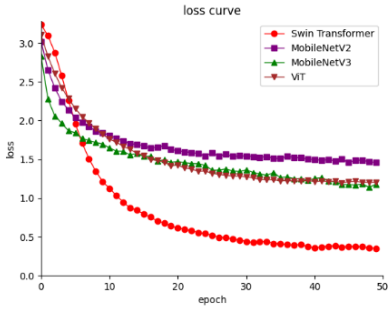


Fig. 6 - Training loss change curve

From the accuracy change curves, it can be observed that the Swin Transformer model achieves a faster improvement in accuracy, ultimately reaching approximately 90.85%. In comparison, MobileNetV2 and MobileNetV3 show relatively small changes in overall accuracy, with final accuracies of only 56.39% and 63.20%, respectively. Additionally, the Swin Transformer converges more quickly and stably than ViT. The loss change curve shows that the Swin Transformer model experiences a rapid decrease in loss during the first 30 training epochs and eventually stabilizes. In contrast, MobileNetV2, MobileNetV3, and ViT exhibit minimal loss reduction throughout the training process. Therefore, the Swin Transformer outperforms the other models in both training accuracy and loss, demonstrating strong robustness and performance advantages.

After model training, performance evaluation was conducted using a test dataset collected from real orchard environments, which included various natural conditions such as strong front lighting, backlighting, and different levels of occlusion.

Table 2

Recognition model testing accuracy

Class	Model1	Model2	Model3	Model4	Class	Model1	Model2	Model3	Model4
A11-167	0.967	0.867	0.967	0.5	A12-180	0.806	0.806	0.75	0.556
A11-168	0.947	0.842	0.868	0.711	A12-182	0.923	0.872	0.897	0.795
A11-169	0.927	0.878	0.878	0.488	A12-183	0.947	0.895	0.895	0.395
A11-170	0.929	0.786	0.976	0.667	A12-184	0.867	0.867	0.833	0.667
A11-171	0.932	0.614	0.932	0.227	A12-185	0.972	0.778	0.917	0.333
A11-172	0.946	0.919	0.838	0.432	A12-186	1.0	1.0	0.8	0.0
A11-173	0.919	0.622	0.973	0.162	A14-209	1.0	0.784	0.919	0.514
A11-174	0.972	0.75	0.917	0.306	A14-210	0.919	0.541	0.811	0.243

Class	Model1	Model2	Model3	Model4	Class	Model1	Model2	Model3	Model4
A11-175	1.0	0.865	0.973	0.324	A14-211	0.838	0.865	0.919	0.568
A11-176	0.837	0.857	0.878	0.571	A14-212	0.921	0.947	0.895	0.868
A12-177	1.0	0.922	1.0	0.627	A14-213	0.789	0.474	0.947	0.289
A12-178	0.966	0.864	0.966	0.847	A14-214	1.0	0.921	0.974	0.263
A12-179	1.0	0.833	0.861	0.194	A14-215	1.0	1.0	0.974	0.769

Note: In this paper, Model1 represents Swin Transformer, Model2 represents MobileNetV2, Model3 represents MobileNetV3, and Model4 represents Vision Transformer (ViT).

The experimental results show that the accuracy of the ViT and MobileNetV2 models is only 50.05% and 82.19%, respectively. The accuracy of the MobileNetV3 model is 90.61%, while the Swin Transformer model achieved 93.55%, the highest among the four models. It exceeds the accuracies of MobileNetV2, MobileNetV3, and ViT by 11.3, 2.94, and 43.50 percentage points, respectively.

Model parameter selection

Choose the Swin Transformer as the recognition model to test the impact of different training parameters on recognition performance. Keeping other parameters constant, four experimental sets (T1-T4) are established, with a batch size fixed at 4. The selected optimizers are SGD and AdamW, with initial learning rates of 0.0001 and 0.00001, respectively. The results of the experiments using different parameter selections are presented in Table 3.

Table 3

Performance of Swin Transformer under Different Parameters

Test	Optimizer	Initial Learning Rate	Bach Size	Accuracy
T1	SGD	0.0001	4	90.85%
T2	SGD	0.00001	4	42.42%
T3	AdamW	0.0001	4	94.21%
T4	AdamW	0.00001	4	95.46%

The results indicate that, with the batch size held constant, the AdamW optimizer demonstrated better performance compared to SGD, particularly at the lower learning rate (0.00001), where model T4 achieved the highest accuracy of 95.46%. For models using the SGD optimizer, a higher learning rate (0.0001) contributed to improved model performance, as seen in model T1 with an accuracy of 90.85%, whereas the accuracy of model T2 with a lower learning rate significantly dropped to 42.42%. This suggests that for SGD, an appropriate increase in the learning rate may help enhance the model's training effectiveness. Overall, AdamW demonstrates greater robustness at low learning rates, likely due to its internal mechanisms, including momentum and adaptive learning rate characteristics. Therefore, it is recommended to prioritize the AdamW optimizer for similar tasks and adjust the learning rate appropriately to find the optimal configuration.

Based on this analysis, the Swin Transformer model demonstrates the best recognition performance with the AdamW optimizer, a batch size of 4, and an initial learning rate of 0.00001.

Improvement Of Loss Function

In this study, the loss function of the Swin Transformer model was improved by replacing the original standard loss function with the Focal Loss function. To validate the effectiveness of Focal Loss, comparative experiments were conducted with Cross-Entropy Loss and Label Smoothing Loss. Focal Loss increases the loss gradient for hard-to-classify samples, enhancing the model's learning ability for these challenging samples and effectively addressing class imbalance.

Table 4

Comparison experiment of different loss functions

Test	Loss	Accuracy	Precision	Recall
T5	Cross-Entropy Loss	93.55%	93.70%	93.60%
T6	Label Smoothing Loss	95.87%	95.87%	95.80%
T7	Focal Loss	98.64%	98.64%	97.90%

From Table 4, it can be observed that Focal Loss demonstrates the best performance in grape leaf variety recognition. In terms of accuracy, the model using Focal Loss achieved an accuracy of 98.72%, which is an improvement of 3.26 percentage points over the accuracy using Cross-Entropy Loss and 2.85 percentage points over Label Smoothing Loss.

In terms of precision, Focal Loss improved by 2.19 percentage points compared to Label Smoothing Loss and by 4.40 percentage points compared to Cross-Entropy Loss. Regarding recall, Focal Loss enhanced performance by 2.10 percentage points over Label Smoothing Loss and by 4.30 percentage points over Cross-Entropy Loss. This improvement can be attributed to Focal Loss's ability to dynamically adjust the loss gradient based on sample difficulty, enabling the model to focus more on samples prone to errors or from rare categories during training. In contrast, Label Smoothing Loss primarily enhances generalization by reducing the model's overconfidence in certain categories, but it does not address class imbalance directly like Focal Loss. Label Smoothing Loss may result in insufficient learning of easier categories, potentially impacting overall performance. In contrast, Focal Loss ensures balanced learning across all categories by focusing on hard-to-classify samples. This attention mechanism enhances the model's learning for difficult samples, thereby improving overall accuracy.

Confusion Matrix

To evaluate the performance of the improved AF-Swin Transformer network model, a test set that was not used in the training or validation phases was employed. The confusion matrix generated is shown in figure 7, where the shading indicates the magnitude of the values; lighter colors represent smaller values, while darker colors represent larger values. The horizontal axis represents the true labels, while the vertical axis represents the predicted labels. The diagonal elements indicate the number of correctly identified samples.

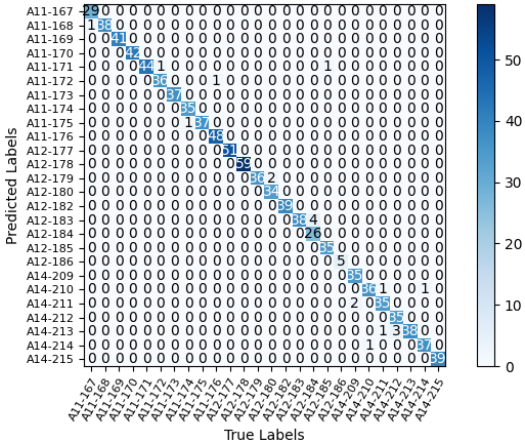
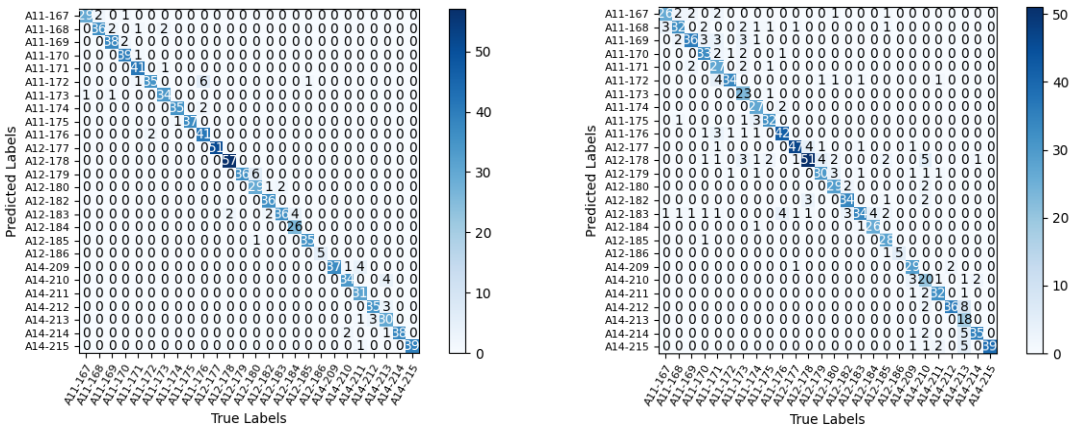
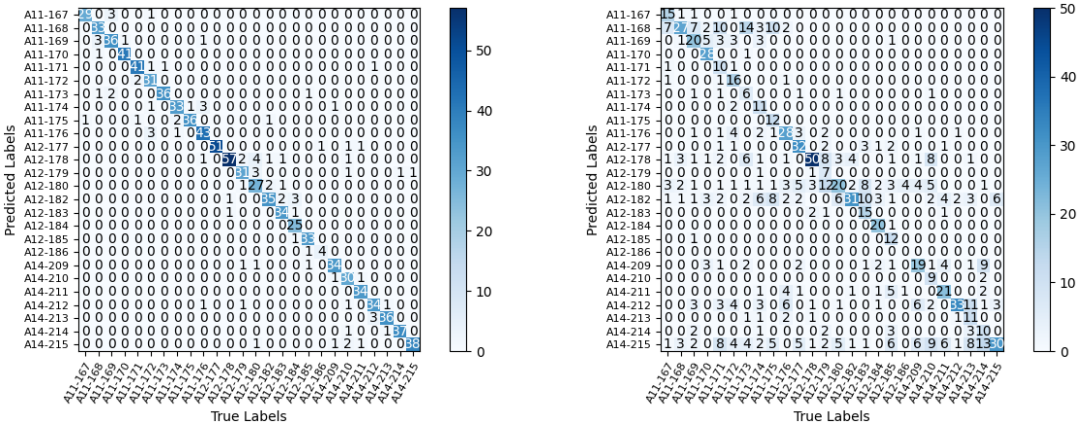


Fig. 7 - Confusion Matrix of AF-Swin Transformer Model

From figure 7, it can be observed that varieties A11-170, A11-171, and A11-172 exhibited no misidentification, indicating that their features are distinct. However, varieties A11-172 and A12-185 were misidentified as A11-171. This suggests that the recognition features for varieties A11-172 and A12-185 do not differ significantly from those of other varieties, making them susceptible to interference. This may also be influenced by shooting angles and lighting conditions.



(a) Swin Transformer (b) MobileNetV2



(c) MobileNetV3 (d) ViT

Fig. 8 - Confusion Matrix of Different Models

By comparing figures 7 and 8, which show the confusion matrices of the AF-Swin Transformer, Swin Transformer, MobileNetV2, MobileNetV3, and ViT models, it can be observed that the AF-Swin Transformer model correctly identifies varieties A11-169, A11-173, and A11-178, while the other models exhibit misidentifications for these varieties. This is primarily because Focal Loss enhances the model's ability to learn from hard-to-identify samples, addressing the issue of class imbalance, while the weight decay characteristics of the AdamW optimizer help improve the model's generalization capability, making it more precise when dealing with varieties that have subtle differences. These improvements provide the AF-Swin Transformer with a significant advantage in processing varieties with closely similar features. The above analysis demonstrates that the proposed improved model, AF-Swin Transformer, has strong robustness in recognizing grape leaf varieties.

Grad-CAM Visual Analysis

To understand the feature learning of grapevine leaf sample, this study used the Grad-CAM algorithm to output the gradient heatmap of the weights in the final convolutional layer and visualize the network model. As shown in figure 9, areas that are redder indicate that these features play a more critical role in class orientation.

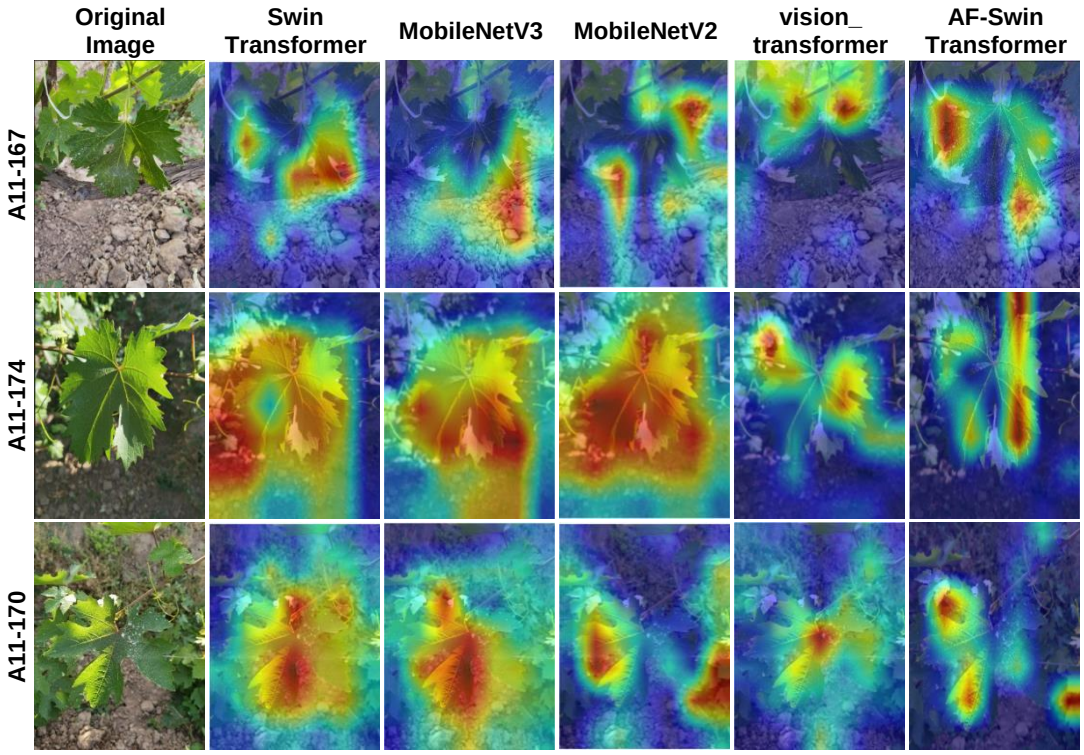


Fig. 9 - Grad-CAM Visualization of Recognition Results for Different Convolutional Neural Network Algorithms

Figure 9 compares the heatmaps of the Swin Transformer, MobileNetV2, MobileNetV3, ViT, and AF-Swin Transformer. The heatmap of the AF-Swin Transformer focuses on both the leaf veins and edges, with a broad distribution that covers a significant portion of the leaf area. In contrast, the heatmaps of MobileNetV2 and MobileNetV3 show higher activity only at the leaf edges, while the Swin Transformer and ViT heatmaps primarily focus on regions where leaf veins are located. Therefore, compared to the other four models, AF-Swin Transformer has a broader and more accurate recognition capability for grapevine leaves.

ROC Curve

To evaluate the recognition performance of the models comprehensively, ROC curves were analyzed. By comparing the area under the curve (AUC), we can intuitively assess the strengths and weaknesses of different models. The ROC curve for the AF-Swin Transformer is shown in Figure 10, while Figure 11 displays the ROC curves for the various models.

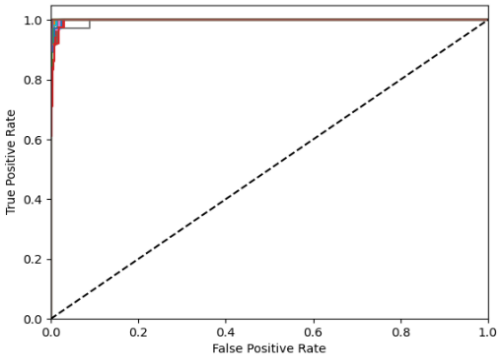
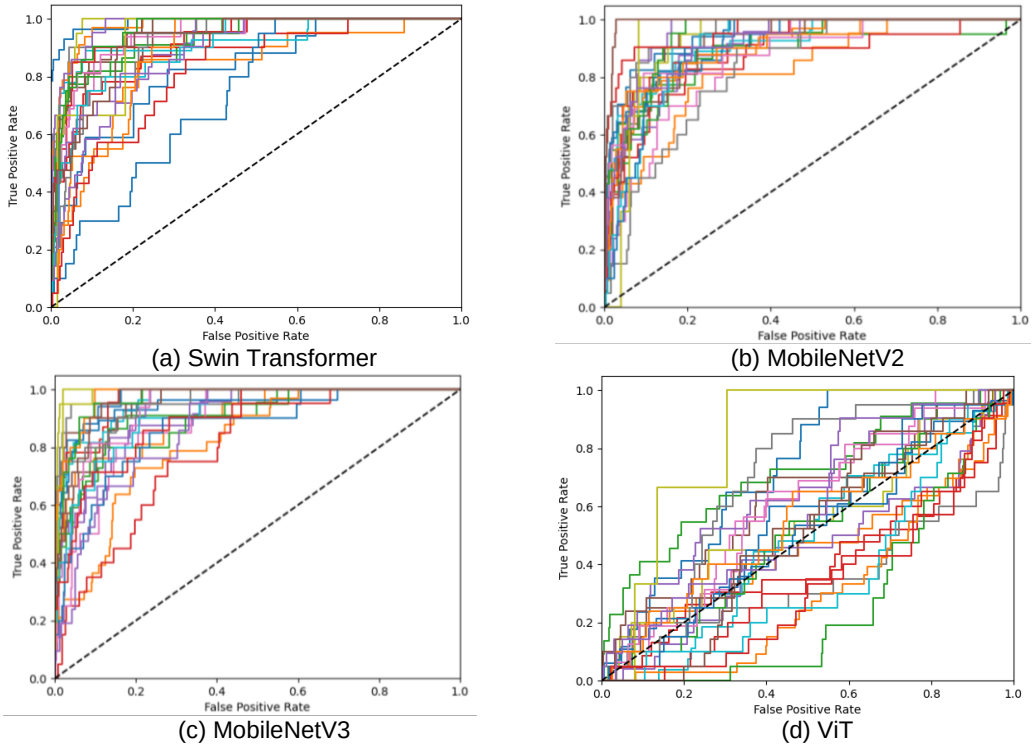


Fig. 10 - ROC Curve of AF-Swin Transformer

From figure 10, it can be seen that the AF-Swin Transformer model exhibits good distinguishing performance among different varieties.



(a) Swin Transformer (b) MobileNetV2

(c) MobileNetV3 (d) ViT

Fig. 11 - ROC Curves of Different Models

Note: A11-167 A11-168 A11-169 A11-170 A11-171 A11-172 A11-173 A11-174 A11-175 A11-176 A12-177 A12-178 A12-179 A12-180 A12-182 A12-183 A12-184 A12-185 A12-186 A14-209 A14-210 A14-211 A14-212 A14-213 A14-214 A14-215

To further comprehensively evaluate the performance of the AF-Swin Transformer model, a comparative assessment was conducted using evaluation metrics such as precision, recall, F1-score, and AUC value for all the models used. As shown in Table 5:

Table 5

Evaluation Metrics of Different Models

Models	Precision	Recall	F1 Score	AUC
Swin Transformer	0.937	0.936	0.934	0.910
MobileNetV2	0.811	0.798	0.795	0.918
MobileNetV3	0.859	0.863	0.856	0.932
ViT	0.626	0.474	0.478	0.520
AF-Swin Transformer	0.981	0.979	0.980	0.999

The table 5 shows that the AF-Swin Transformer achieves an overall precision of 0.981 in grapevine leaf variety identification, which is higher than that of the Swin Transformer, MobileNetV2, MobileNetV3, and ViT by 4.4, 17, 12.2, and 35.5 percentage points, respectively. The Recall value for the AF-Swin Transformer is 0.979, exceeding that of the Swin Transformer, MobileNetV2, MobileNetV3, and ViT by 4.3, 18.1, 11.6, and 50.5 percentage points, respectively. Additionally, the AF-Swin Transformer has the highest area under the ROC curve (AUC) among the models. The Focal Loss function reduces overfitting on simple samples compared to the cross-entropy loss function, enabling the model to concentrate on misclassified samples and effectively capture complex features and patterns. The AdamW optimizer maintains model stability during training and mitigates overfitting, particularly when handling complex datasets with subtle varietal differences. Therefore, the AF-Swin Transformer proposed in this study significantly outperforms other models in identifying grape leaf varieties and demonstrates clear advantages across various evaluation metrics.

CONCLUSIONS

This study focuses on recognizing grapevine leaf varieties and proposed the AF-Swin Transformer model, which efficiently identifies different grapevine leaf varieties despite sample imbalance conditions. The main conclusions are as follows:

(1) Compared to four other deep learning models, the AF-Swin Transformer model demonstrates better performance in grapevine leaf variety identification.

(2) To address the issue of sample imbalance among different grape leaf varieties, the Swin Transformer model's loss function was replaced with the Focal Loss function. Additionally, the AdamW optimizer was introduced to improve the model's generalization capability. The results show that the AF-Swin Transformer model demonstrates good stability in identifying grape leaf varieties.

This study identified only 26 grape leaf varieties, and future research will expand to include more varieties and further optimize the model. Additionally, exploration of data augmentation techniques and transfer learning methods will be undertaken to achieve efficient recognition of various plant leaves.

ACKNOWLEDGEMENTS

The data utilized in this study was provided by the National Grape Germplasm Resource Center (Pomology Institute, Shanxi Academy of Agricultural Sciences) and funded through Shanxi Agricultural University's Technology Innovation Improvement Project (CXGC2023046) and Youth Science & Technology Innovation Project (No.2019024).

REFERENCES

[1] Cecotti H., Rivera A., Farhadloo M. (2020). Grape detection with Convolutional Neural Networks. *Expert Systems with Applications,* 159:113588. DOI:10. 1016/j.eswa.2020.113588.

[2] Chen J., Han J., Liu C., Wang Y., Shen H., Li L. (2022). A deep learning method for the classification of apple varieties via leaf images from different growth periods in natural environment. *Symmetry*, 14(8):1-14.

[3] Das S., Chatterjee M., Stephen R., Singh A. K., Siddique A. (2024). Unveiling the Potential of YOLO v7 in the Herbal Medicine Industry: A Comparative Examination of YOLO Models for Medicinal Leaf Recognition. *International Journal of Engineering Research & Technology (IJERT)*, Vol.13, Issue 11, Paper ID: IJERTV13IS110019

[4] Dong Z., Yang F., Du J., (2024), Identification of varieties in Camellia oleifera leaf based on deep learning technology. *Industrial Crops and Products*, 216: 118635.

[5] Islam M.K., Habiba S., Ahsa S.M.M. (2019). Bangladeshi plant leaf classification and recognition using YOLO neural network. *2nd International Conference on Innovation in Engineering and Technology (ICIET),* pp. 1-5. IEEE.

[6] Kaur A. (2024). Leaf Detectives: A Deep Learning Approach to Grapevine Varietal Identification with DenseNet-121. *In 2024 Global Conference on Communications and Information Technologies (GCCIT),* pp. 1-5. IEEE.

[7] Lin Lihui, Wei Yi, Pan Junhong (2021). Classification Method of Wuyi Rock Tea Leaves Based on Convolutional Neural Networks (基于卷积神经网络的武夷岩茶叶片分类方法). *Journal of Ningde Normal University*: *Natural Science Edition*, 33(4): 7.

[8] Maulana F.A., Kertarajasa K., Yasa Y.S., Sari S.A., Sulistiyo M.D. (2024). Grapevine Leaves Classification Using Various CNN Model. *In 2024 11th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE),* pp. 224-229, IEEE.

[9] Ni Jiangong, Yang Haoyan, Li Juan, Han Zhongzhi. (2021), Identification of Peanut Pod Varieties Based on Improved AlexNet (基于改进型 AlexNet 的花生荚果品种识别). *Journal of Peanut Science*, 050(004): 14-22.

[10] Pereira C.S., Morais R., Reis M.J.C.S. (2019). Deep learning techniques for grape plant species identification in natural images. *Sensors*, 19(22): 4850.

[11] Patricio I.R.R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and Electronics in Agriculture*, 153.

[12] Pushpanathan K., Hanafi M., Mashohor S.I.W.F.F. (2021). Machine learning in medicinal plants recognition: a review. Artificial Intelligence Review: *An International Science and Engineering Journal*, 54(1):305-327.

[13] Sennan S., Pandey D., Alotaibi Y., Alghamdi S. (2022). A Novel Convolutional Neural Networks Based Spinach Classification and Recognition System. *Computers, Materials & Continua*, 73(1).

[14] Sun D.Z., Liu J.Y., Ding Z. (2023). A Multi-Variety Classification Method for Southern Medicinal Plant Leaves Based on the Improved EfficientNetv2 Model (基于改进 EfficientNetv2 模型的多品种南药叶片分类方法). *Journal of Huazhong Agricultural University*, 42(1): 258-267.

[15] Su B.F., Shen L., Chen S. (2021). Multi-Feature Classification Method for Grape Varieties Based on Attention Mechanism (基于注意力机制的葡萄品种多特征分类方法). *Journal of Agricultural Machinery*, (011): 052.

[16] Tavakoli H., Alirezazadeh P., Hedayatipour A. (2021). Leaf image-based classification of some common bean cultivars using discriminative convolutional neural networks. *Computers and electronics in agriculture*, 181: 105935.

[17] Yang J., Run P., Zhang Y.Y. (2023). Research on Visual Recognition Methods of Chinese Herbal Medicine Plants Based on Deep Learning (基于深度学习的中草药植物视觉识别方法研究). *Chinese Journal of Stereology and Image Analysis*, 28(2): 203-211.

[18] Yin X., Ji Y., Zhang R. (2023). Research on recognition of Camellia oleifera leaf varieties based on deep learning[J]. *Journal of Nanjing Forestry University*, 47(3): 29.

[19] Zhang P., Li D. (2022), YOLO-VOLO-LS: a novel method for variety identification of early lettuce seedlings. *Frontiers in plant science*, 13, 806878.

[20] Zou Wei, Yue Yanbin, Feng Enying. (2024). Identification of Anthracnose Disease in Chili Fruits Based on MobileNet V2 and Its Application (基于 MobileNet V2 的辣椒果实炭疽病识别及其应用). *Guizhou Agricultural Sciences*, 52(09): 125-132.