Front. Agr. Sci. Eng. 2025, 12(2): 173–189

RESEARCH ARTICLE

An improved multiscale fusion dense network with efficient multiscale attention mechanism for apple leaf disease identification

Dandan DAI, Hui LIU (✉)

Institute of Artificial Intelligence & Robotics (IAIR), Key Laboratory of Traffic Safety on Track of Ministry of Education, School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China.

KEYWORDS

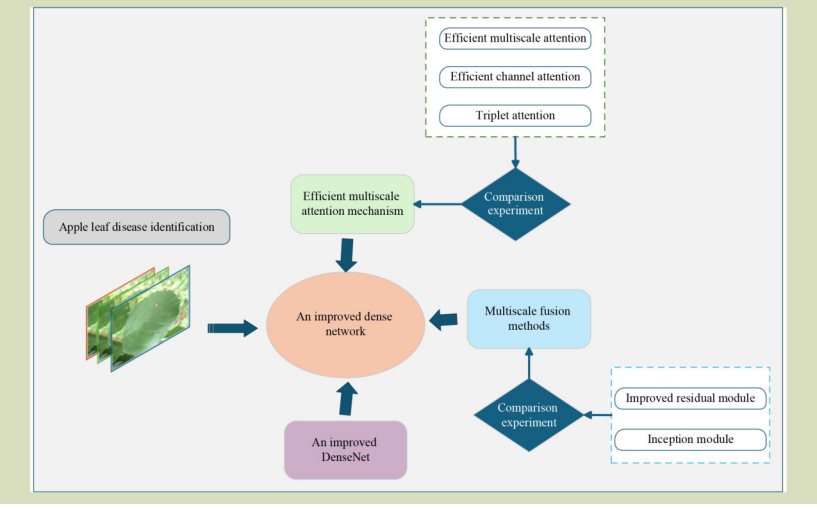
Incept_EMA_DenseNet, multi-scale fusion module, efficient multiscale attention mechanism, apple leaf disease identification

HIGHLIGHTS

- A multiscale fusion dense network with the EMA mechanism is proposed for apple leaf disease identification.
- Replace the shallow feature extraction layer with a multiscale fusion and compare the performance of two different multiscale methods
- Integrate the EMA mechanism into models based on the comparison among three types of attention mechanisms.
- An improved DenseNet is proposed based on DenseNet_121, reducing half of the parameters.

Received July 27, 2024;
Accepted September 15, 2024.

GRAPHICAL ABSTRACT



ABSTRACT

With the development of smart agriculture, accurately identifying crop diseases through visual recognition techniques instead of by eye has been a significant challenge. This study focused on apple leaf disease, which is closely related to the final yield of apples. A multiscale fusion dense network combined with an efficient multiscale attention (EMA) mechanism called Incept_EMA_DenseNet was developed to better identify eight complex apple leaf disease images. Incept_EMA_DenseNet consists of three crucial parts: the inception module, which substituted the convolution layer with multiscale fusion methods in the shallow feature extraction layer; the EMA mechanism, which is used for obtaining appropriate weights of different dense blocks; and the improved DenseNet based on DenseNet_121. Specifically, to find appropriate multiscale fusion methods, the residual module and inception module were compared to determine the performance of each technique, and Incept_EMA_DenseNet achieved an accuracy of 95.38%. Second, this work used three attention mechanisms, and the efficient multiscale attention mechanism obtained the best performance. Third, the convolution layers and bottlenecks were modified without performance degradation, reducing half of

the computational load compared with the original models. Incept_EMA_DenseNet, as proposed in this paper, has an accuracy of 96.76%, being 2.93%, 3.44%, and 4.16% better than Resnet50, DenseNet_121 and GoogLeNet, respectively, proved to be reliable and beneficial, and can effectively and conveniently assist apple growers with leaf disease identification in the field.

)

1 Introduction

Agriculture is a critical sector, supporting about 866 million people worldwide, which constitutes over 25% of the global labor force. However, there has been a troubling 16%-decline in the agricultural workforce since 2000 and led to increasing instances of abandoned farmland. This shift raises a critical question of how land use can be optimized while minimizing labor inputs. In response, automated agricultural machinery has emerged. Nevertheless, with the technological revolution and industrial transformation, current mechanized agriculture cannot satisfy the development of modern society[1], which means smart agriculture is a significant topic.

With the implementation of industry 4.0 technologies in agriculture, more researchers are advocating for the use of machine vision technology to address various challenges in the sector, including fruits and plant counting, crop growth and health monitoring, and especially disease and pest assessment[2]. In the past, pest and disease assessment relied on visual inspections by highly experienced operators, which consumed much time and was difficult for farmers to afford. In contrast, with the advent of image recognition technology, these challenges can now be addressed more efficiently and effectively.

Currently, three main image recognition technologies are used for crop disease assessment. The first one is pattern recognition, which combines image processing with pattern recognition to extract different parameters of crops, such as color and shape. After screening and optimization of parameters, some methods, such as linear classifiers and Bayesian decision theory, can be used for identification and classification[3]. The second method is machine learning, which focuses on designing and analyzing automatic learning algorithms. For example, this research considered wheat as the research object. It uses the ReliefF algorithm, and a spatial and temporal adaptive reflectance fusion model (SMLST) to fuse input image data, constructing an SMLST-support vector machine (SVM) model, which achieved an accuracy of 81.2%[4]. In addition, comparing the two operational process-based and machine-learning methods when studying rice blast disease, Nettleton et al.[5] found that process-based models can be substituted for machine learning, as they have better adaptability to new data sets. The third method is deep learning, which extracts complex features from many input pictures depending on deep neural networks, demonstrating exceptional accuracy. For example, Nirmal used Gaussian filtering, a preprocessing technology, and a convolution neural network model named DACCNN to obtain a new loss function, which achieved 98% accuracy[6].

Of these three methods, earlier digital image processing techniques require individuals with high levels of expertise to identify and evaluate detecting consequences, and machine learning has great challenges in the feature extraction segment, particularly when facing limited training samples[7]. In contrast, deep learning can significantly reduce the workload and effectively implement feature extraction to deal with large data sets, providing users with simpler and more understandable methods and outcomes.

Apple is one of the most popular fruits worldwide, with about 5 Mha planted globally, 83 Mt of fruit was produced in 2017[8]. However, leaf disease is a serious problem affecting apple yield, with reliable identification of the cause needing experience and expertise, so how to use visual intelligence to identify the cause of diseases affecting leaves is a crucial question. Depending on the visual technique, farmers can use simple and convenient photo recognition instead of paying to consult experts.

With the rapid development of deep learning, convolutional neural networks (CNN) have achieved prominent results in image classification and recognition, which has also been widely applied to crop diseases. Iftikhar et al.[9] used a fine-tuned CNN model called ECNN, which normalized the pixel values of the input images on a 0 to 1 scale. Then, varying convolution kernels and max pooling were applied, followed by dropout layers to reduce the number of neurons. They further generalized this approach to incorporate models such as

ResNet50, AlexNet, and MobileNet V2 and achieved an accuracy of 98.17% in identifying fungi species. Also, an enhanced fine-grained robust deep CNN model was used by Thaseentaj and Ilango[10] that obtained an accuracy of 94.73%, 9.23% higher than the standard CNN model. In addition to optimizing the architecture of the CNN model for higher efficiency, researchers have undertaken numerous efforts to minimize the number of model parameters. Fu et al.[11] developed a lightweight CNN using depthwise separable convolutions (DSC) that used dilated convolutions (DC) to enlarge the receptive field of the model. Compared to ResNet50, Inception-v3 and MSR-3DCNN, their DSC-DC model performed better, achieving an accuracy of 99%.

Researchers have attempted to adjust each model segment to improve their precision and robustness to obtain better performance. For example, Singh et al.[12] introduced a data augmentation technique known as LeafyGan, which includes two stages with the initial stage a precise localization of leaves by using a segmentation model and the subsequent stage an examination of normal leaf images to generate diverse leaf image types. The enhanced data sets were used for training the MobileViT model, achieving an accuracy of 99.92% on the PlantVillage data sets, representing an improvement of 3.16% compared to earlier methods. Khan et al.[13] proposed a contrast stretching hybrid method that uses a top-bottom hat filter to modify the intensity values of the image, enhancing both local and global contrast. In addition to data argumentation, researchers have also replaced components of established network structure, such as classifying modules, to improve model efficiency. Nagachandrika et al.[14] developed a novel deep learning model called MFF-ADNet, which uses a highly adaptable CNN-AM tuned by the EGOA model to acquire features for classification, resulting in an improvement of over 4% compared to other classification models. Chen et al.[15] replaced the softmax layer with an SVM classifier for output categorization, which achieved an accuracy of 94.76% in detecting cucumber downy mildew.

In addition, researchers also integrated two to three typical blocks of standard network models to compensate for their limitations, capitalizing on their respective strengths. Radočaj et al.[16] developed the IncMB module based on an inception module, which incorporates the Mish activation function and batch normalization layer, achieving 97.78% accuracy in detecting tomato leaf diseases. Also, Liu et al.[17] integrated a multiscale module into EfficientNet, enhancing the its performance. A multiscale residual convolution module was used by Wen et al.[18] to replace the original 7×7 convolution layer in the ResNet model to create a multiscale convolution parallel structure, which achieved 98.72% accuracy in mulberry leaf disease detection. In addition, Zeng et al.[19] presented a multiscale convolution module, GMDC, that uses dilated convolutions of varying kernel sizes to extract feature information across multiple scales. In numerous research studies, substituting an initial single convolution layer with a multiscale convolution layer has proved to be an effective method, usually showing better accuracy and adaptability.

Apart from focusing on the feature extraction layer to get more picture details, researchers have also put considerable effort into various attention mechanisms to obtain more useful information and minimize data loss. Zhou et al.[20] incorporated the SIMA attention mechanism into the lightweight neural network model ShuffleNet v2 and replaced the convolution layer with max pooling in down-sampling, achieving 98.4% accuracy in maize leaf disease identification. Wang et al.[21] proposed a global attention mechanism (GAM), primarily consisting of channel and spatial attention mechanisms. By incorporating this GAM into the Spatial Pyramid Pooling–Fast (SPPF) module and the sampling process of YOLOv5, they achieved an accuracy of 91.4%, marking a significant 2.93% improvement. A new attention-enhanced model based on DenseNet_121 was proposed by Liu et al.[22] that used an algorithm to extract features according to regions of interest, achieving 96% accuracy in rice leaf disease identification.

In addition to assigning weights to different data, the attention mechanism can be combined with multiscale convolution layers to better analyze information. For example, the attention mechanism CAFF fused multiscale features[19]. Li et al.[23] proposed the mixed channel spatial attention mechanism, which incorporates a hybrid channel attention mechanism including the spatial attention mechanism and channel attention mechanism, and a multiscale feature fusion module into the ConvNeXt network, which achieved a 5.68% improvement in accuracy compared to the original model.

Therefore, this paper proposes a multiscale fusion dense network with an efficient multiscale attention mechanism (EMA) named Incept_EMA_DenseNet to precisely identify apple leaf diseases. Its framework consists of three parts: the shallow feature extraction using a multiscale fusion module, the DenseNet containing bottleneck and transition layer, which are improvements based on DenseNet_121, and a hybrid attention mechanism combined with channel attention mechanism and spatial attention mechanism. The inception module is used to extract shallow feature information, which can better extract features of different scales by focusing on

global and local information. The EMA mechanism is used to obtain different weights of the feature information of dense blocks, consisting of multiple bottlenecks, which has proved extraordinary in apple leaf classification.

The key contributions of this work are:

• A new apple leaf disease data set was created comprising 15,295 pictures, including nine types of apple leaves (healthy and 8 kinds of disease), from which a multiscale fusion dense network with the EMA mechanism was proposed having 96.76% accuracy when training for apple leaf disease identification.

• Replace the shallow feature extraction layer with a multiscale fusion module and analyze the improvement of multiscale methods, comparing the performance of two different multiscale methods Retention mechanism, which shows a significant performance compared with standard neural networks such as ResNet50, DenseNet_121, GoogLeNet and inception module.

• Integrate the EMA mechanism into models and use the attention mechanism to obtain the weights of the feature information of dense blocks. Three types of attention mechanisms are used in comparison experiments to determine the best and most appropriate attention mechanism.

• An improved DenseNet is proposed based on DenseNet_121 in apple leaf classification, which uses preactivation methods in the bottleneck and reduces half of the parameters without performance degradation.

2 Materials and methods

2.1 Data collection and preprocessing

The data used was obtained and selected from two publicly available data sets[24,25], including eight types of apple leaf diseases, Alternaria leaf spot, brown spot, frogeye leaf spot, gray spot, mosaic, powdery mildew, rust and scab, and healthy leaves as a comparator, as shown in Fig. 1. These pictures were taken with two kinds of background, outdoor scenes with complex environmental interference and simple laboratory environment, which can help demonstrate better

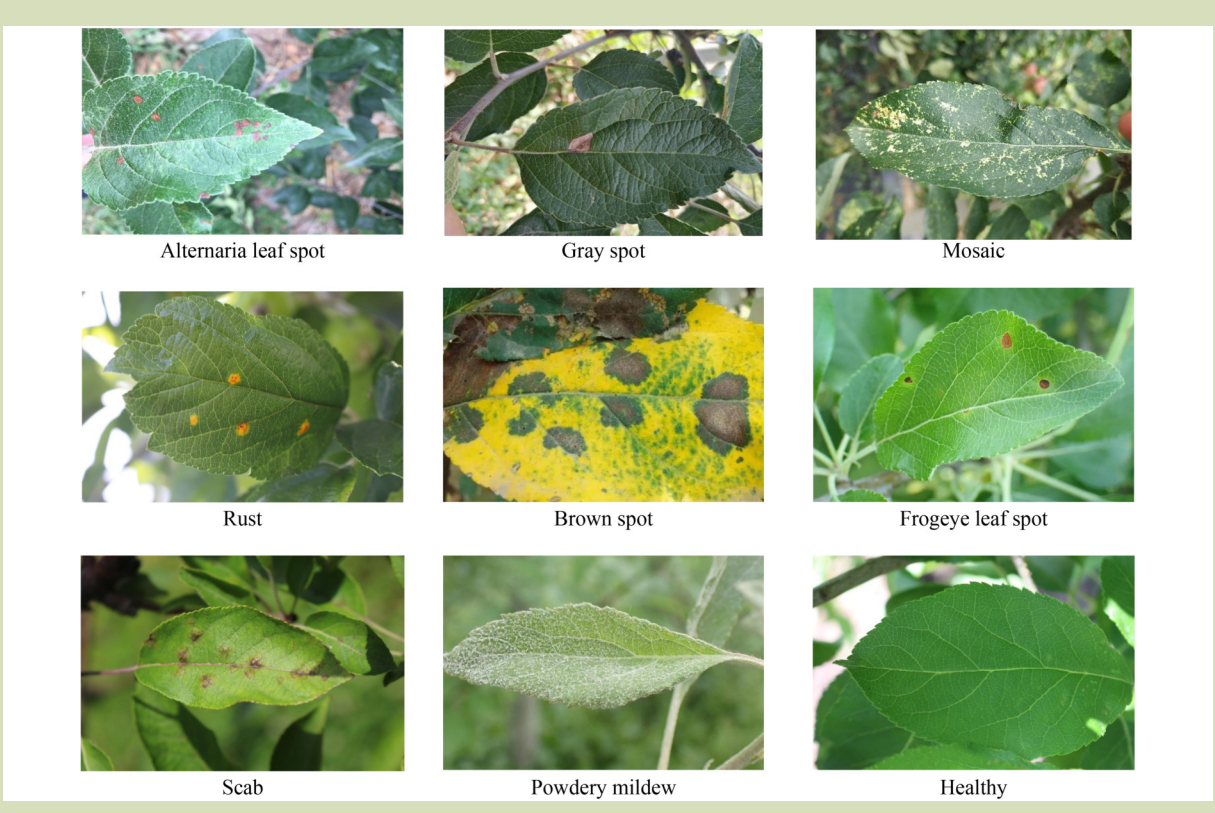


Fig. 1 Apple leaf containing eight common apple leaf diseases and a healthy leave as a comparator.

generalization performance in practical application scenarios. However, the data on four types of leaves, Alternaria leaf spot, brown spot, gray spot and mosaic, are relatively scarce, which most probably affects the learning and generalization abilities of the model, so the data sets needs to be expanded. Hu et al.[26] explored many image argumentation techniques such as data cropping, rotation and mirroring to expand maize leaf disease images. Cai et al.[27] reported that reducing the brightness and darkness of images can improve model performance. However, changes in brightness and darkness can negatively impact the distinction of similar apple leaf disease. This study mainly used rotation and mirroring to expand the image data set. The number of images in each category before and after expansion is shown in Table 1. Then, the images were interpolated to 224 × 224 pixels, and the data was divided into a training, validation and test sets at a ratio of 7:2:1 for the convenience of model training. According to the comparative experimental results between the augmented and small data sets, the accuracy rates of DenseNet_121, ResNet50, and GoogLeNet on the original data set were 91.93%, 90.13%, and 88.86%, respectively. By comparison, the accuracy of the expanded data set was 93.32%, 93.83%, and 92.60%, representing an increase of 1.39%, 3.7%, and 3.74%, respectively, which shows that the expanded data set could provide better performance.

2.2 Incept_EMA_DenseNet model structure

An enhanced convolution neural network for apple leaf disease classification was developed, which consists of three parts: inception module, EMA mechanism, and DenseNet model. The structure of the model is shown in Fig. 2. This model is an improvement of DenseNet_121, which retains the basic framework in which bottleneck and transition operate alternately to ensure model simplicity and explicitness. The bottleneck developed batch normalization technique compared with AlexNet and GoogLeNet, which ensured that the input distribution remains relatively stable across layers, accelerating the convergence speed of network training process by normalizing the input of each layer. The convolution layers with a kernel size of 1 × 1 are also used before the 3 × 3 convolution layer in the bottleneck, which adjusted the number of channels, effectively decreasing the number of output feature maps and model parameters, and the computational complexity. The crucial characteristic of DenseNet is the connectivity mechanism, where the output of each bottleneck layer is concatenated with the input along the channel dimension. In other words, the input to each layer is the concatenated feature maps from all preceding layers, leading to a total of $L(L+1)/2$ direct connections compared to L connections in standard networks, where L represents the number of layers in the network.

Although both ResNet and DenseNet use special connectivity mechanisms to alleviate the problems of gradient vanishing or exploding encountered during the training of deep neural networks, the different connection methods determine their distinct functions and characteristics. Residual connections are mainly used in ResNet, which directly adds the input to the output of the subsequent layer through skip connections, expressed as $y = F(x) + x$, where $F(x)$ represents the output of x after undergoing convolutional operations. This type of connection facilitates the propagation of gradients within deep networks, addressing the issue of gradient vanishing. Distinguished from ResNet, DenseNet expands the width of the network, enhancing the efficiency of information flow and gradient propagation enabling feature reuse and parameter efficiency. In this work, DenseNet_121 is used as a deep network architecture, which is unnecessary for this network structure, and more attention should be paid to the relation

Table 1 Apple leaf disease data sets

Leaf type	Original data	Small data set	Augmented data set
Alternaria leaf spot	417	513	1251
Brown spot	411	501	1233
Frogeye leaf spot	3181	515	3181
Gray spot	339	519	1017
Healthy	1231	516	1231
Mosaic	371	501	1113
Powdery mildew	1182	515	1182
Rust	2753	510	2755
Scab	5410	515	5410

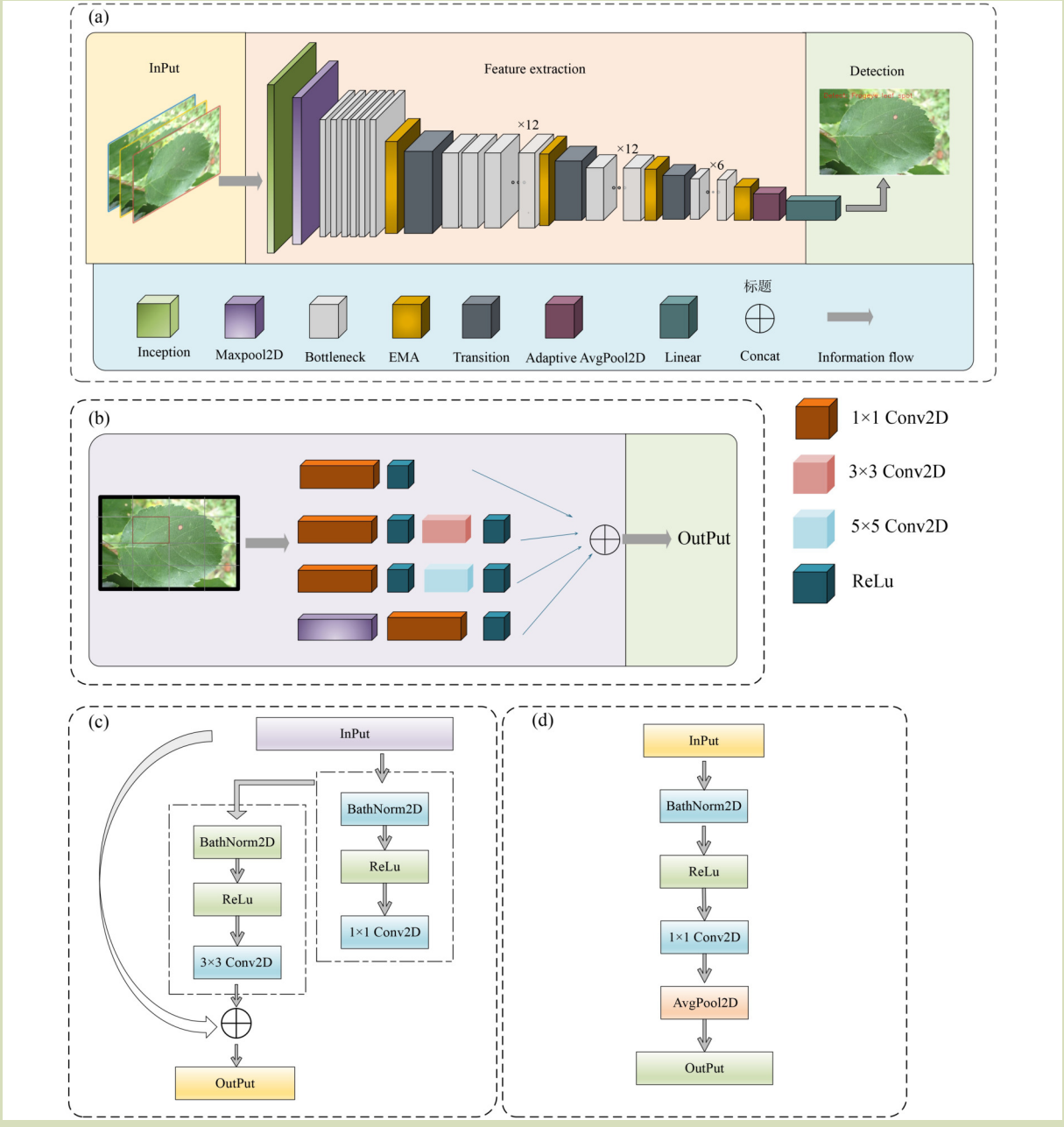


Fig. 2 Structure of the Incept_EMA_DenseNet model. (a) The structure of whole model, (b) inception module, (c) bottleneck module, and (d) transition layer.

among different channels. Also, ResNet requires more computational power and consumes more time for the training process, while DenseNet_121 can be trained faster with only a relatively small difference in accuracy.

The DenseNet_121 model has too many convolution layers, which results in a huge computation load and might lead to an overfitting problem. So, this model was reduced to 22 bottlenecks with half the number of parameters with minimal accuracy loss, and the quantity of the convolution layers corresponding to each specific location compared to DenseNet_121 are given in Table 2. The √ and × means whether the models have each module and the number represents the modules' quantity of the models. In addition, to assist in mitigating the gradient vanishing or exploding problem, preactivation, which refers to a series of operations or

transformations done before the activation function of a neuron used in the bottleneck section, and the original structure is changed to BatchNorm2D, ReLu and convolution layer as shown in Fig. 2(c).

2.3 Multiscale fusion methods inception module

The receptive field is important because it determines the pixel on a feature map in a certain layer that can be mapped back to a region of a certain size on the input image. Normally, an overly small receptive field can be incapable of capturing sufficient global information, ultimately impacting the scope for model generalization. Therefore, increasing the receptive field of the entire model while preserving model performance is crucial. Additionally, it is important to note that the receptive field is directly influenced by the kernel size of the convolutional layers. Smaller receptive fields capture fine-grained features, such as edges and textures, while larger receptive fields encompass more comprehensive information, like object shapes and positions within the input image. So, multiple small convolution kernels are used to increase the depth of the network and capture more detailed information from the image, which is extensively used as a well-regarded Visual Geometry Group network architecture, encouraging its use in research. For example, a Yang et al.[28] recommended using a 3 × 3 kernel instead of the original 7 × 7 kernel for shallow networks, as the latter can effectively introduce nonlinearity and facilitate the ability of the model to extract more intricate details from images, which means a broader receptive field affords greater access to informational extraction.

In addition to integrating a small receptive field into the model, there are many methods to enhance the receptive field without neglecting detailed information, such as dilated convolution[29]. Also, Liu et al.[30] proposed MCDCNet, which substitutes deformed convolution for standard convolution, and the adaptability of the model for geometric variations can be significantly improved by introducing learnable offsets. Also, a multiscale fusion module is also integrated into CNNs, as it captures details and characteristics of images across different scales, which can eliminate redundant information and extract more useful features, comprehensively improving the accuracy, generalization ability and robustness of the model. For example, Tian et al. [31] used concatenated inception modules and improved residual modules and integrated them into the deep layers of the DenseNet model, which achieved an accuracy of 94.31% in identifying apple leaf disease. Inspired by the multifeature fusion module, this work evaluated two multifeature fusion modules, the inception module and the improved residual module shown in Fig. 3, both of which replace the first 7 × 7 convolution layer of the DenseNet model to achieve improved feature extraction ability for apple leaf disease classification.

The inception module contains four paths, and each path generates different channel information, which can extract more diverse features and describe the input data more

Table 2 The structure comparison between DenseNet_121 and improved DenseNet

Layer	Output size	DenseNet_121	Improved DenseNet
Shallow feature extraction	112 × 112	7 × 7 Convolution layer	Inception module
Maxpool2D	56 × 56	√	√
Dense block1 (bottleneck)	56 × 56	6	6
Attention mechanism	56 × 56	×	√
Transition layer1	28 × 28	√	√
Dense block2 (bottleneck)	28 × 28	12	12
Attention mechanism	28 × 28	×	√
Transition layer2	14 × 14	√	√
Dense block3 (bottleneck)	14 × 14	24	12
Attention mechanism	14 × 14	×	√
Transition layer3	7 × 7	√	√
Dense block4 (bottleneck)	7 × 7	16	6
Attention mechanism	7 × 7	×	√
Adaptive Avgpool2D	1 × 1	√	√

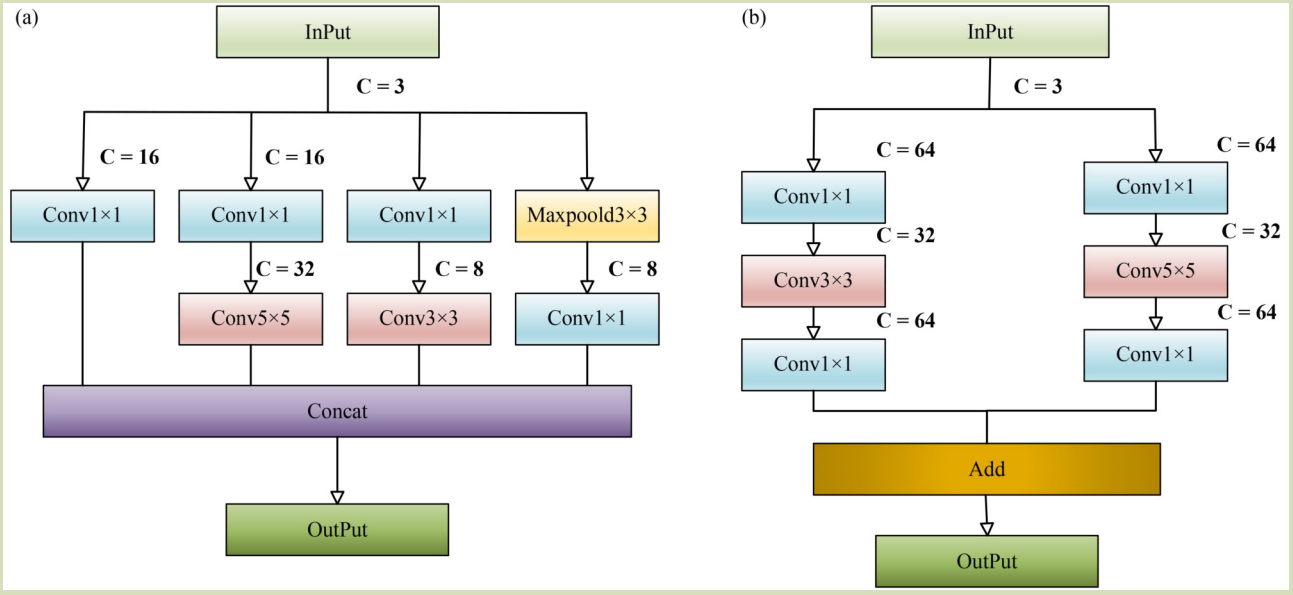


Fig. 3 Structural comparison of the inception and improved residual module. (a) Inception module and (b) mproved residual module.

comprehensively. The inception module uses three types of convolutional layers, where different kernel sizes correspond to distinct channel weights. The fourth path involves a max pooling layer, which reduces the dimensionality of feature maps and lowers the computational complexity of the subsequent layers. Then, the concatenate methods merge four paths in the channel dimension compared to the residual module that uses add methods to aggregate the element values at corresponding positions. The improved residual module is adjusted based on the long-established residual module, which splits a single feature extraction path into two, and each path has 64 channels, effectively preventing network degradation and minimizing the risk of overfitting problems.

2.4 Efficient multiscale attention mechanism

The attention mechanism is beneficial for capturing semantic associations among target features, thereby facilitating the model to comprehend and classify data more comprehensively. Wang et al.[32] integrated squeeze-and-excitation (SE) attention mechanism, which includes three steps into DenseNet, squeeze (global information embedding), excitation (adaptive recalibration) and weighting, enabling the model to learn which channels are more important for the task and adjust the feature maps accordingly. Although the dimensionality reduction operation in the SE attention mechanism can improve computational efficiency, it results in the loss of partial channel information, which may dismiss crucial nuanced differences for the final task. Therefore, efficient channel attention (ECA) is used to solve the loss information problem[33], in which a one-dimensional convolution generator substitutes the two fully connected layers in the SEA module to avoid dimensionality reduction without degrading the performance of the attention module in other aspects.

The ECA mechanism involves three key steps: global average pooling, sigmoid activation, and elementwise multiplication. First, a global average pooling operation is applied to each channel of the input feature map, calculating the mean value for each channel. Next, attention weights for each channel are generated by using a sigmoid function to the channel averages obtained in the previous step. Finally, elementwise multiplication is performed between each channel of the original feature map and its corresponding attention weight, resulting in a weighted feature map. The relevant equations are:

$$X = \frac{1}{c}\sum_{i=1}^{C} x_i \tag{1}$$

$$X = \sigma\left(\frac{1}{c}\sum_{i=1}^{C} X_i\right) \tag{2}$$

$$X = X \odot \sigma\left(\frac{1}{c}\sum_{i=1}^{C} X_i\right) \tag{3}$$

Apart from the channel attention mechanism, EMA is also important for interchannel dependencies and the encoding of global information. Hao et al.[34] combined EMA into YOLOv5, which reorganizes the channel and batch dimension, leverages cross-dimensional interactions to capture pixel-level relationships, and then encodes global information in parallel

branches to recalibrate the channel weights.

In the EMA mechanism, the first group is used to split the channels of the future and turns C×H×W into C//G×H×W (with // being integer division, also referred to as floor division or downward rounding division). Then, two branches are generated. The first performs a one-dimensional global pooling operation and the second extracts features through a 3 × 3 convolution. After that, the output features from both branches are then modulated through a sigmoid function (Eq. (4)), normalization operation (Eq. (5)) and softmax function (Eq. (6)), which generates C//G×H∗W and C//G×1×1 and finally, the two feature outputs are combined through a cross-dimension interaction module. The more specific information content can be illustrated as shown in Fig. 4(a), and the comparison structure between the EMA and ECA mechanisms is also shown in Fig. 4.

$$\sigma_X = \left(\frac{1}{1+e^{-x}}\right) \tag{4}$$

$$X_{\text{norm}} = \frac{x-u}{\sigma} \tag{5}$$

$$\sigma(Z)_j = \frac{e^z j}{\sum_{k=1}^{K} e^z k} \tag{6}$$

Additionally, there are also some hybrid attention mechanisms not only focusing on channel information, such as the convolutional block attention module (CBAM) proposed by Chen et al.[35], which combines spatial, channel and triplet

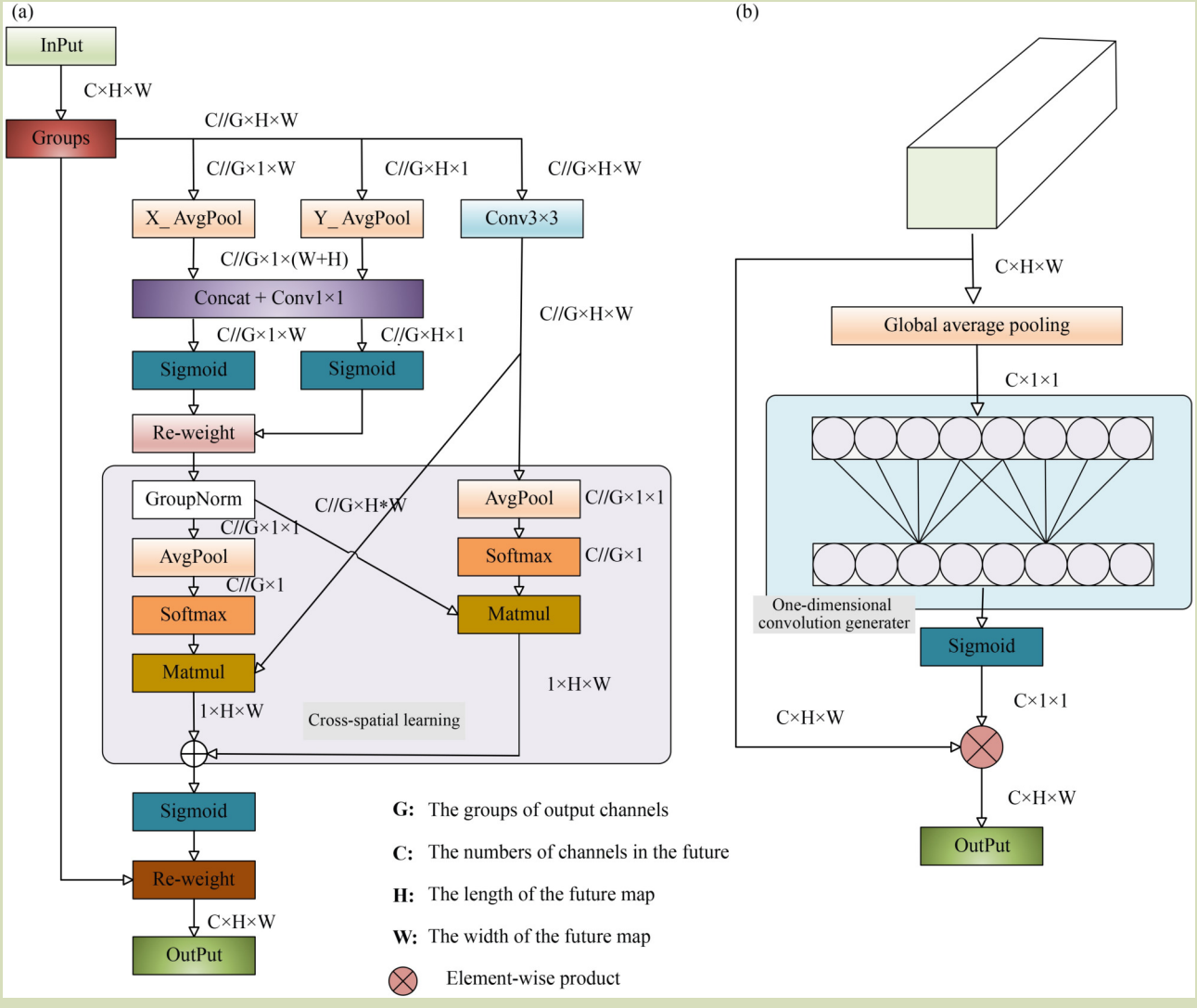


Fig. 4 Structural comparison between efficient multiscale attention (EMA) and efficient channel attention (ECA) mechanisms. (a) EMA mechanism and (b) ECA mechanism.

attention as mentioned by Hasanpour Zaryabi et al.[36] This work combined the triplet attention mechanism with the inception module to compare the other two attention mechanisms, ECA and EMA.

Table 3 shows all the parameter sizes of the model. The number of parameters are halved in improved DenseNet (abbreviated as DenseNet in the context) compared to DenseNet_121. The inception module does not increase the computational load but the residual module creates more than twice that load. Three attention mechanisms do not need many parameters to learn, but triplet attention presents the highest estimated total size among the three attention mechanisms, which is usually connected with the simplicity and efficiency of the model computational process, and the EMA mechanism is the lowest.

3 Experiments and results

In this section, the advancement of the Incept_EMA_DenseNet model is demonstrated, and some comparative experiments were be conducted to test the performance of the inception module and EMA mechanism. First, the model proposed was compared with DenseNet_121, ResNet50, and GoogLeNet to test its relative ability. Second, ablation experiments were conducted: first the inception module was compared with the residual module then comparison of ECA, EMA and triplet attention mechanisms was done based on Incept_DenseNet.

3.1 Experimental setup and evaluation metric

During training, the ratio of training data to validation data was 8:2, and cross-entropy loss and softmax activation function were used to classify the features. Adam optimization was used to update the model parameters using gradient descent. Adam uses exponential moving averages to compute the first-moment estimate, also known as the mean, and the second-moment estimate, commonly referred to as the uncentered variance of the gradients as the following equations, which are explained in Table 4.

First-moment estimate:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{7}$$

Second-moment estimate:

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{8}$$

This methodology gave the algorithm a high degree of efficiency and stability during the parameter update process. Also, the small memory requirements of Adam optimization make it an ideal choice for tackling large-scale data sets and extensive parameter space tasks. By leveraging exponential moving averages, Adam adaptively tunes the learning rate for each parameter according to its historical gradient information, thereby improving performance and robustness during the training phase.

In addition, a learning rate scheduler was used to adjust the learning rate of an optimizer, which can optimize model performance by dynamically changing the learning rate as the training progresses, enabling the model to converge faster and achieve better generalization. In this work, the learning rate was decreased in an exponential manner, which allowed faster updates to the weight parameters in a large learning rate during the initial stages of training and aided the model in making finer adjustments in a small learning rate when nearing the optimal solution as the training progressed. An initial large learning rate allowed the model to approach the optimal solution region quickly. A lower learning rate reduced oscillation of the parameters around the optimal solution, facilitating the model to converge to the optimal solution more stably.

From Eq. (9), lr is the new learning rate, lr_{initial} is the initial

Table 3 Parameter number and size for all evaluated models

Model	Total parameters	Parameter size (MB)	Estimated total size (MB)
DenseNet_121	6,971,145	26.59	392.48
DenseNet	3,300,809	12.59	312.52
Incept_DenseNet	3,304,565	12.61	333.49
Residual-DenseNet	6,203,241	23.66	435.47
Incept_ECA_DenseNet	3,310,325	12.63	343.90
Incept_Triplet_DenseNet	3,305,765	12.61	378.15
Incept_EMA_DenseNet	3,314,565	12.64	344.56

Table 4 Definition of each parameter in the first and second-moment estimate equations

Parameter	Definition
m_t	First-moment estimate (momentum) at the current time step t, an exponentially moving average of past gradients
β_1	Decay rate for the first-moment estimate controls the exponential decay rate of the first-moment estimate
m_{t-1}	First-moment estimate at the previous time is step $t-1$
g_t	Gradient at the current time step t
$1-\beta_1$	Complement of the decay rate is used to adjust the weight of the current gradient
v_t	Second-moment estimate (uncentered variance) at the current time step t, an exponentially moving average of the squares of past gradients
β_2	Decay rate for the second moment estimate, typically set to a value close to 1, smaller than β_1
v_{t-1}	Second moment estimates at the previous time step $t-1$
g_t^2	Square of the gradient at the current time step t, also expressed as $g_t \times g_t$, where $\times$represent element-wise multiplication
$1-\beta_2$	Complement of the decay rate, used to adjust the weight of the current gradient square

learning rate, and γ and epoch are the decay rate and the number of epochs, respectively. In the experiments, γ was set to 0.9 and lr_{initial} to 0.001.

$$\text{New\_}lr = \text{intial\_}lr \times \gamma^{\text{epoch}} \tag{9}$$

To evaluate the results of all experiments, some evaluation metrics were used. ACC is the most commonly proposed classification performance metric, representing the proportion of samples correctly classified by the model out of the total number of samples. Precision and recall represent the proportion of instances predicted as positive that are truly positive and the proportion of all true positive instances correctly identified, respectively. The F1 score is the harmonic mean of precision and recall, which is used to balance both and obtain a comprehensive result and the equation of ACC. The false positive rate (FPR) describes the proportion of negative samples that are incorrectly predicted as positive among all negative samples, evaluating the ability of a model to distinguish negative samples. ACC, precision, recall, FPR and F1 are defined as:

$$\text{ACC} = \frac{(\text{TP}+\text{TN})}{(\text{P}+\text{N})} \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{12}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN}+\text{FP}} \tag{13}$$

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{14}$$

A confusion matrix is a table that displays the detailed classification results of a model containing true positives, false positives, true negatives and false negatives. The confusion matrix not only offers an intuitive representation of the correspondence between model predictions and actual labels, but also provides essential data for calculating key performance metrics. This information is crucial for optimizing the model and conducting thorough error analysis. The receiver operating characteristic (ROC) curve depicts the trade-off between the true positive rate (also referred to as recall) and the FPR at various threshold settings, which is more robust compared to other evaluation metrics when dealing with imbalanced data sets. The area under the curve (AUC), which is used to quantify the performance of a classifier, stands for the area under the ROC curve. The AUC value ranges from 0 to 1, where 0.5 indicates that model performance is equivalent to random guessing, and 1 represents a perfect classifier. Therefore, the closer the AUC value is to 1, the better the model performs across different thresholds.

3.2 Incept_EMA model compared with standard convolution neural networks

First, to compare the proposed model, some standard convolution neural networks were also trained and tested to serve as the control group. In addition, as this model was an improvement based on DenseNet_121, the comparison between DenseNet_121 and Incept_EMA_DenseNet showed the improvement obtained in the multiscale fusion module and attention mechanism.

As shown in Table 5, the Incept_EMA_DenseNet model had the highest accuracy of 96.76%, being 2.93%, 4.16%, and 3.44% higher than ResNet50, GoogLeNet and DenseNet_121,

respectively. With other types of assessment metrics, including precision, recall, F1 score and FPR, Incept_EMA_DenseNet gave the best performance, being 96.34%, 96.75%, 96.50%, and 0.42%, respectively. The accuracy and loss of all evaluated models during training are shown in Fig. 5. From the validation accuracy curve, each model eventually converged, and model accuracy was ranked from highest to lowest as: Incept_EMA_DenseNet, Incept_ECA_DennseNet, Incept_Triplet_DenseNet, Incept_DenseNet, Residual_DenseNet, ResNet50, DenseNet_121, GoogLeNet.

Also, the value loss curve was used to evaluate the performance of a model on unseen data to detect potential overfitting or underfitting problems of the model. In Fig. 5, GoogLeNet shows an overall poor performance, and Residual-DenseNet produced a degree of oscillation. Other models converged upon a fixed value, and Incept_EMA_DenseNet has the lowest loss value being less than 2%.

3.3 Ablation experiments

Although the previous experiment demonstrated that the Incept_EMA_DenseNet model achieved the highest accuracy in classifying apple leaf diseases, it did not fully reveal the benefits of multiscale fusion methods and attention mechanisms. Therefore, the ablation experiments were specifically designed to focus on the independent contributions of multiscale fusion methods and attention mechanisms. In addition, the model incorporating a multiscale fusion module was compared to DenseNet_121 to highlight the significance of multiscale fusion methods. The results showed that Residual_DenseNet and Incept_DenseNet improved overall performance. As shown in Table 6, there were 2.06% and 1.7% improvements with Incept_DenseNet and Residual_DenseNet compared with DenseNet_121, and the accuracy of Incept_DenseNet was 0.36% higher than Residual-DenseNet, which also had the highest F1 score of 0.9536 and lowest FPR of 0.0063 for different multiscale fusion methods.

Table 5 The experiment performance between Inccept_EMA_DenseNet and standard models

Model	Accuracy	Precision	Recall	F1	FPR
ResNet50	93.83	0.9303	0.9351	0.9326	0.0081
DenseNet_121	93.32	0.9269	0.9227	0.9245	0.0089
GoogLeNet	92.60	0.9302	0.9170	0.9228	0.0102
Incept_EMA_DenseNet	96.76	0.9634	0.9675	0.9650	0.0042

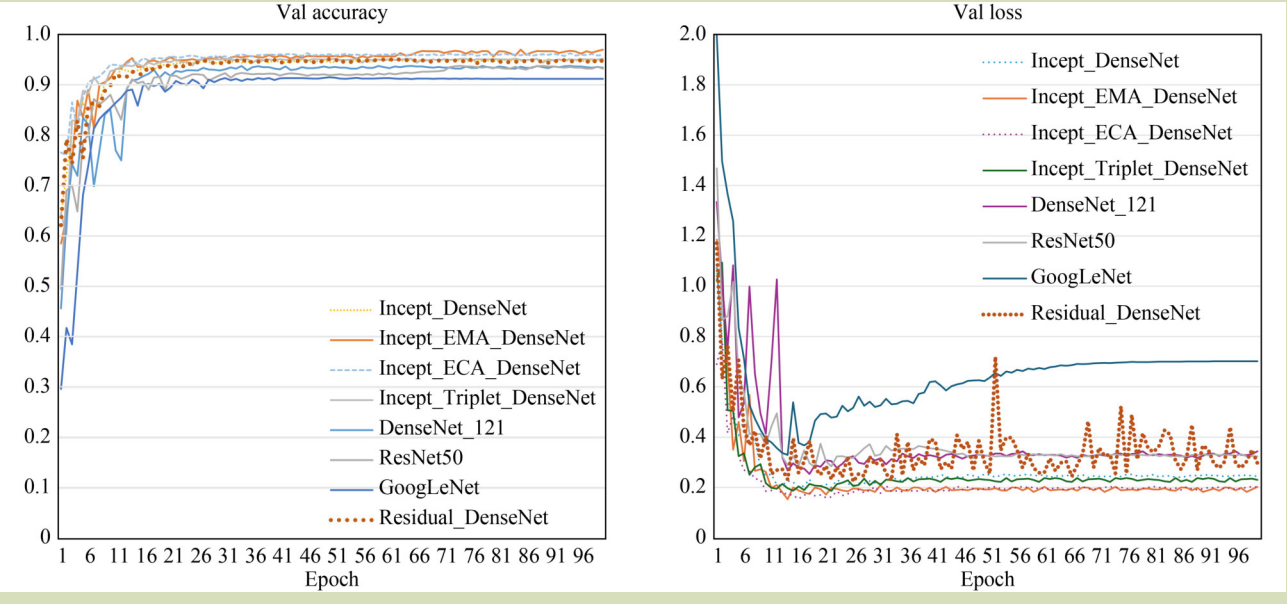


Fig. 5 Validation accuracy and loss of all evaluated models.

Table 6 The experiment performance between two multiscale fusion methods

Model	Accuracy	Precision	Recall	F1	FPR
DenseNet_121	93.32	0.9269	0.9227	0.9245	0.0089
Residual_DenseNet	95.02	0.9425	0.9547	0.9483	0.0065
Incept_DenseNet	95.38	0.9558	0.9520	0.9536	0.0063

From the validation of different multiscale fusion methods experiments, the Incept_DenseNet gave extraordinary performance, so the validation experiments for the different attention mechanisms could be executed based on the Incept_DenseNet. Therefore, different attention mechanisms were compared to determine which mechanism could provide the greatest enhancement based on Incept_DenseNet. As shown in Table 7, three attentions mechanism all enhanced the ability of the Incept_DenseNet, giving improvements in accuracy of 0.92%, 0.82%, and 1.38% with ECA, triplet attention and EMA mechanisms, respectively.

The accuracy for each model is evident in the comparison diagram of the confusion matrix (Fig. 6), and the healthy leaves had the lowest accuracy compared to the disease leaves in all evaluated models, which may be related to the limited learning image data for this leaf type. Although standard neural networks performed poorly in healthy apple leaf classification, which only reached 85%, the model combined with a multiscale fusion module and attention mechanism achieved above 90% accuracy, and the Incept_EMA_DenseNet achieved the highest accuracy of 94%. The ROC for all evaluated models is given in Fig. 7, in which the 45° line represents a random classifier. The AUC was used to quantify the performance of a classifier based on the ROC curve, revealing the model performance for the nine apple leaf types. The closer the AUC is to 1, the better the model performance, with Incept_EMA_DenseNet giving the best performance.

4 Discussion

First, the experiments showed that among standard networks, ResNet50, DenseNet_121, and GoogLeNet achieved accuracies above 90%, which means the augmented data set of apple leaf disease contains sufficient and relevant samples that accurately reflect the characteristics and variations present. In addition, after integrating a variable learning rate that decreases exponentially over time, the convergence speed of all the evaluated models was accelerated, and the models tended to achieve a stable equilibrium at about 50 epochs. Also, oscillation phenomena, which can result in strong performance on the training set but poor generalization on the test set, were significantly reduced. This allowed the models to deliver more consistent predictive performance when applied to real-world context.

Second, as the ablation experiment between the residual module and inception module showed, compared to the original 7×7 convolution layer, both the incept and residual modules enhanced the receptive field, providing different kernel sizes for shallow extraction of the images, which considerably improves the accuracy of the model without neglecting the global information. Also, this experiment demonstrated that substituting multiscale methods for a single convolution layer can enhance the ability of the model to capture diverse features. In addition, it is noteworthy that the inception module had less than half of the computational load than the residual module but it gave higher accuracy and precision, which may be attributed to its dispersion and then integration of the channels. Different channel weights for different kernel size convolution layers facilitate the model in better learning local detailed information and global information.

Third, the three attention mechanisms improved model

Table 7 Experiment performance for different attention mechanism

Model	Accuracy	Precision	Recall	F1	FPR
Incept_DenseNet	95.38	0.9558	0.9520	0.9536	0.0063
Incept_ECA_DenseNet	96.30	0.9598	0.9659	0.9626	0.0049
Incept_Triplet_DenseNet	96.20	0.9635	0.9597	0.9615	0.0052
Incept_EMA_DenseNet	96.76	0.9634	0.9675	0.9650	0.0042

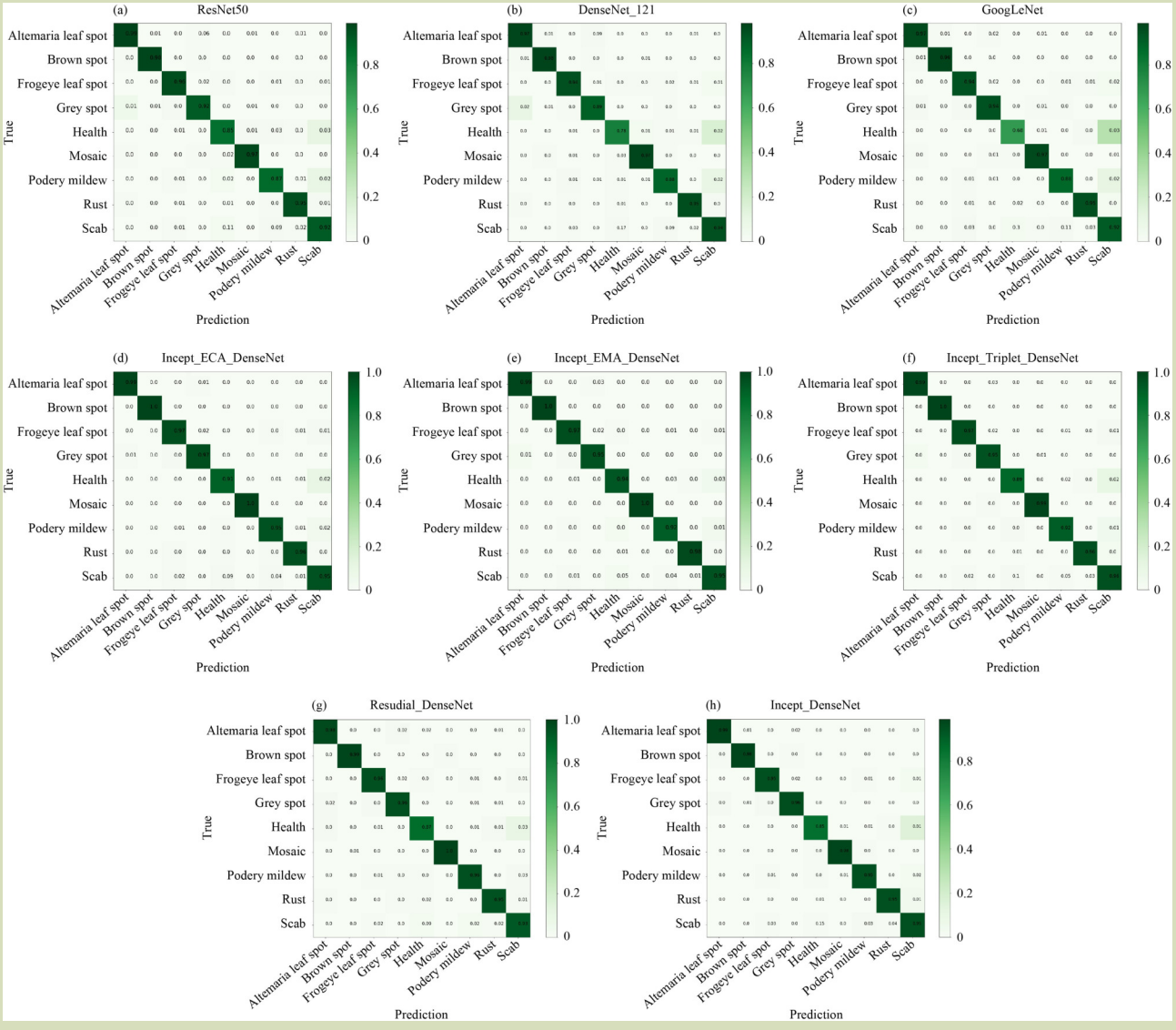


Fig. 6 Comparison of the confusion matrix of all evaluated models. (a) ResNet50; (b) DenseNet_121; (c) GoogLeNet; (d) Incept_ECA_DenseNet; (e) Incept_EMA_DenseNet; (f) Incept_Triplet_DenseNet; (g) Resudial_DenseNet; (h) Incept_DenseNet.

performance as they all take into account channel information, offering adaptive weights for the output of each dense block. However, a complex attention mechanism may lead to model degradation and increase the difficulty of adjusting the parameters of each attention branch, which would make it difficult for a model to achieve satisfactory predictions. Therefore, though triplet attention fuses temporary, spatial and channel attention, it resulted in poorer performance than the other two attention mechanisms. The Overall, the EMA mechanism gave a better performance due to its combination of image features from multiple scales, which enhances the capability of the model to extract information at different scales.

Generally speaking, the dense networks combined with the EMA mechanism and inception module gave extraordinary performance compared to the standard networks, achieving an accuracy of 96.76% when identifying heathy apple leaves and those with eight common leaf diseases, demonstrating its specific advantages in crop disease detection.

5 Conclusions

This paper proposes a novel model, Incept_EMA_DenseNet, which integrates a multiscale fusion module with an EMA mechanism. The model is composed of multiscale shallow

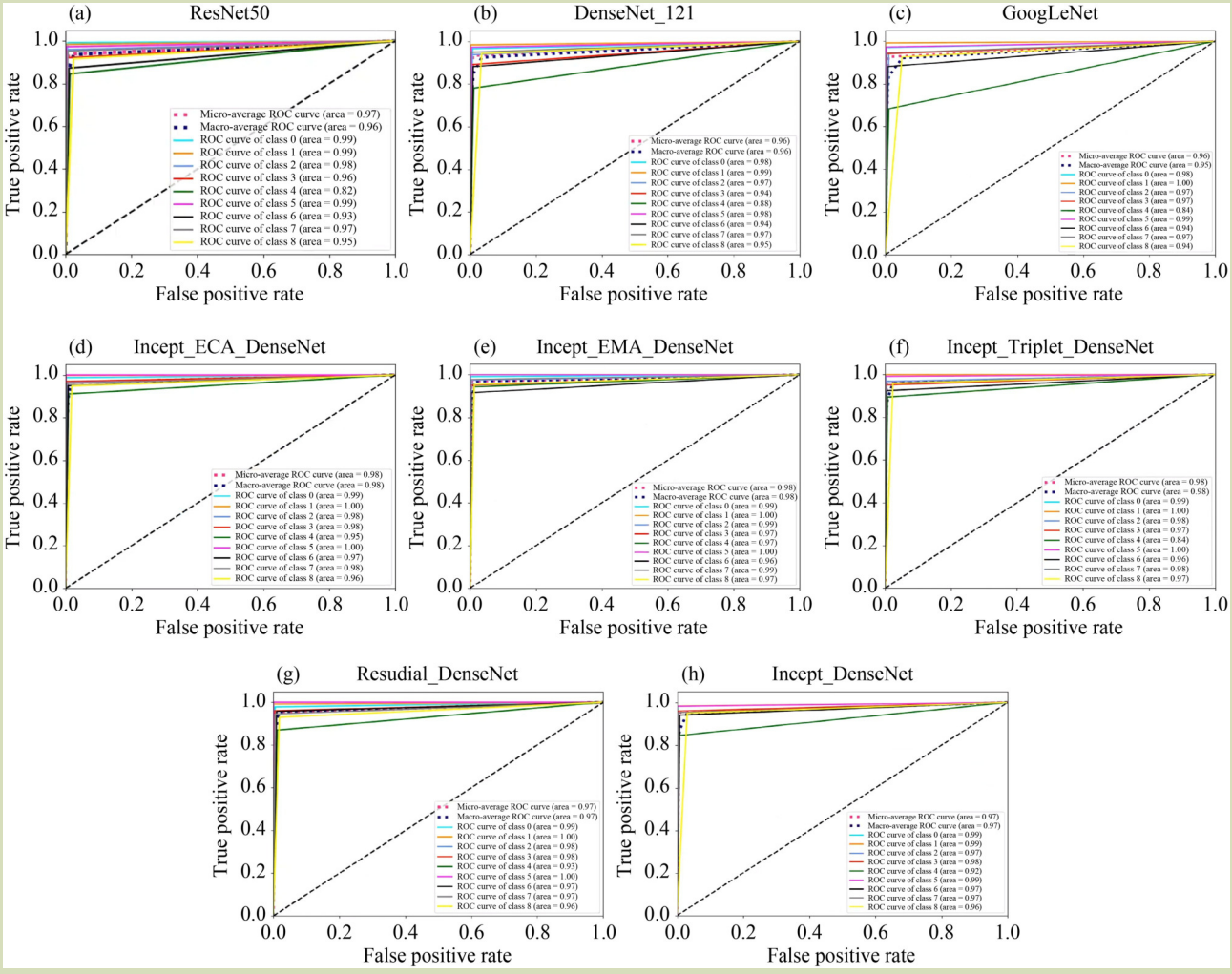


Fig. 7 Comparison of the receiver operating characteristic curve of all evaluated models. (a) ResNet50; (b) DenseNet_121; (c) GoogLeNet; (d) Incept_ECA_DenseNet; (e) Incept_EMA_DenseNet; (f) Incept_Triplet_DenseNet; (g) Resudial_DenseNet; (h) Incept_DenseNet.

feature extraction, an attention mechanism applied to dense blocks and an enhanced DenseNet architecture. The new comprehensive model was able extract detailed local and global information through different kernel sizes of convolution layers. Its EMA mechanism merges spatial information and channel information with little increase in computational load. The framework of DenseNet was also been improved compared to DenseNet_121, which saves a significant amount of computing time. In this work, the accuracy and performance of the models were evaluated in three comparison experiments, and Incept_EMA_DenseNet gave an accuracy of 96.76% in the identification of healthy apple leaves and those with eight common leaf diseases, which was 2.93%, 3.44%, and 4.16% higher than Resnet50, DenseNet_121, and GoogLeNet, respectively. Additionally, different multiscale fusion methods and attention mechanisms were also evaluated to determine the most appropriate module for apple leaf disease classification with Incept_EMA_DenseNet performing best.

However, there are also limitations in the Incept_EMA_DenseNet, such as the imbalanced data set might decrease the recognition accuracy for a certain target, and model performance on other types of data sets still needs to be validated. In addition, the attention mechanism combines holistic information depending on the degree of relevance of channels by obtaining the weights of each dense block. Concurrently, too many attention mechanisms may cause the network to give excessive attention to redundant information, hindering the model from achieving optimal performance. Therefore, future works should focus on three ways to improve the generalization of such models:

(1) More apple leaf disease and healthy pictures should be

collected to obtain a more comprehensive data set, which can improve accuracy and solve the problem of accuracy loss caused by an unbalanced data set.

(2) The EMA mechanism used in this work obtains the weights of different dense block weights, and further work should concentrate on other convolution layers of DenseNet.

(3) The Incept_EMA_DenseNet provides high performance in apple leaf disease classification. Therefore, various types of data sets could be used to train and enhance the model, enabling its application to other crops or plant species. This approach could significantly improve the accuracy of detecting leaf diseases in agricultural systems.

Acknowledgements

This work was fully supported by the National Natural Science Foundation of China (52072412).

Compliance with ethics guidelines

Dandan Dai and Hui Liu declare that they have no conflicts of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Luo X W, Liao J, Zang Y, Ou Y G, Wang P. Developing from mechanized to smart agricultural production in China. *Strategic Study of CAE*, 2022, **24**(1): 46–54 (in Chinese)
2. Restrepo-Arias J F, Branch-Bedoya W J, Awad G. Image classification on smart agriculture platforms: systematic literature review. *Artificial Intelligence in Agriculture*, 2024, **13**: 1–17
3. Zhai Z, Cao Y, Xu H, Yuan P, Wang H. Review of key techniques for crop disease and pest detection. *Transactions of the Chinese Society for Agricultural Machinery*, 2021, **52**(7): 1–18 (in Chinese)
4. Zhao J L, Du S Z, Huang L S. Monitoring wheat powdery mildew (*Blumeria graminis* f. sp. *tritici*) using multisource and multitemporal satellite images and support vector machine classifier. *Smart Agriculture*, 2022, **4**(1): 17–28
5. Nettleton F D, Katsantonis D, Kalaitzidis A, Sarafijanovic-Djukic N, Puigdollers P, Confalonieri R. Predicting rice blast disease: machine learning versus process-based models. *BMC Bioinformatics*, 2019, **20**(1): 514
6. Raj N, Perumal S, Singla S, Sharma G K, Qamar S, Chakkaravarthy A P. Computer-aided agriculture development for crop disease detection by segmentation and classification using deep learning architectures. *Computers & Electrical Engineering*, 2022, **103**: 108357
7. Yu X, Yang M, Zhang H, Li D, Tang Y, Yu X. Research and application of crop diseases detection method based on transfer learning. *Transactions of the Chinese Society for Agricultural Machinery*, 2020, **51**(10): 252–258 (in Chinese)
8. Wang Y, Li W, Xu X, Qiu C, Wu T, Wei Q, Ma F, Han Z. Progress of apple rootstock breeding and its use. *Horticultural Plant Journal*, 2019, **5**(5): 183–191
9. Iftikhar M, Kandhro I A, Kausar N, Kehar A, Uddin M, Dandoush A. Plant disease management: a fine-tuned enhanced CNN approach with mobile app integration for early detection and classification. *Artificial Intelligence Review*, 2024, **57**(7): 167
10. Thaseentaj S, Ilango S S. Deep convolutional neural networks for south Indian mango leaf disease detection and classification. *Computers, Materials & Continua*, 2023, **77**(3): 3593–3618
11. Fu M, Lu C, Mao Y, Zhang X, Wu Y, Luo H, Liu Z, Li W, Ou G. An efficient and accurate deep learning method for tree species classification that integrates depthwise separable convolution and dilated convolution using hyperspectral data. *International Journal of Digital Earth*, 2024, **17**(1): 2307999
12. Singh A K, Rao A, Chattopadhyay P, Maurya R, Singh L. Effective plant disease diagnosis using Vision Transformer trained with leafy-generative adversarial network-generated images. *Expert Systems with Applications*, 2024, **254**: 124387
13. Khan M A, Akram T, Sharif M, Awais M, Javed K, Ali H, Saba T. CCDF: automatic system for segmentation and recognition of fruit crop diseases based on correlation coefficient and deep CNN features. *Computers and Electronics in Agriculture*, 2018, **155**: 220–236
14. Nagachandrika B, Prasath R, Joe I R P. An automatic classification framework for identifying type of plant leaf diseases using multi-scale feature fusion-based adaptive deep network. *Biomedical Signal Processing and Control*, 2024, **95**: 106316
15. Chen X, Shi D, Zhang H, Pérez J A S, Yang X, Li M. Early diagnosis of greenhouse cucumber downy mildew in seedling stage using chlorophyll fluorescence imaging technology. *Biosystems Engineering*, 2024, **242**: 107–122
16. Radočaj P, Radočaj D, Martinović G. Image-based leaf disease

recognition using transfer deep learning with a novel versatile optimization module. *Big Data and Cognitive Computing*, 2024, **8**(6): 52

17. Liu M, Liang H, Hou M. Research on cassava disease classification using the multi-scale fusion model based on EfficientNet and attention mechanism. *Frontiers in Plant Science*, 2022, **13**: 1088531
18. Wen C, He W, Wu W, Liang X, Yang J, Nong H, Lan Z. Recognition of mulberry leaf diseases based on multi-scale residual network fusion SENet. *PLoS One*, 2024, **19**(2): e0298700
19. Zeng T, Li C, Zhang B, Wang R, Fu W, Wang J, Zhang X. Rubber leaf disease recognition based on improved deep convolutional neural networks with a cross-scale attention mechanism. *Frontiers in Plant Science*, 2022, **13**: 829479
20. Zhou H, Su Y, Chen J, Li J, Ma L, Liu X, Lu S, Wu Q. Maize leaf disease recognition based on improved convolutional neural network ShuffleNetV2. *Plants*, 2024, **13**(12): 1621
21. Wang G, Xie R, Mo L, Ye F, Yi X, Wu P. Multifactorial tomato leaf disease detection based on improved YOLOv5. *Symmetry*, 2024, **16**(6): 723
22. Liu W, Yu L, Luo J. A hybrid attention-enhanced DenseNet neural network model based on improved U-Net for rice leaf disease identification. *Frontiers in Plant Science*, 2022, **13**: 922809
23. Li D, Zhang C, Li J, Li M, Huang M, Tang Y M C C M. Multi-scale feature extraction network for disease classification and recognition of chili leaves. *Frontiers in Plant Science*, 2024, **15**: 1367738
24. @H. Classification of apple leaf diseases. *Fei Jiang AI studio*, 2022. Available at Fei Jiang AI studio website on October 20, 2024 (in Chinese)
25. Yang Q, Shu D, and Li W. Efficient Identification of Apple Leaf Diseases in the Wild Using Convolutional Neural Networks. *Agronomy*, 2022, **12**(11): 2784
26. Hu J, Jiang X, Gao J, Yu X. LFMNet: a lightweight model for identifying leaf diseases of maize with high similarity. *Frontiers in Plant Science*, 2024, **15**: 1368697
27. Cai Z, Ou Y, Ling Y, Dong J, Lu J, Lee H. Feature detection and matching with linear adjustment and adaptive thresholding. *IEEE Access: Practical Innovations, Open Solutions*, 2020, **8**: 189735–189746
28. Yang L, Yu X, Zhang S, Long H, Zhang H, Xu S, Liao Y. GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases. *Computers and Electronics in Agriculture*, 2023, **204**: 107543
29. Wang Y, Zhao G, Xiong K, Shi G. MSFF-Net: multi-scale feature fusing networks with dilated mixed convolution and cascaded parallel framework for sound event detection. *Digital Signal Processing*, 2022, **122**: 103319
30. Liu B, Huang X, Sun L, Wei X, Ji Z, Zhang H. MCDCNet: multi-scale constrained deformable convolution network for apple leaf disease detection. *Computers and Electronics in Agriculture*, 2024, **222**: 109028
31. Tian Y, Li E, Liang Z, Tan M, He X. Diagnosis of typical apple diseases: a deep learning method based on multi-scale dense classification network. *Frontiers in Plant Science*, 2021, **12**: 698474
32. Wang K, Jiang P, Meng J, Jiang X. Attention-based DenseNet for pneumonia classification. *IRBM*, 2022, **43**(5): 479–485
33. Gera D, Balasubramanian S. Landmark guidance independent spatial-channel attention and complementary context information based facial expression recognition. *Pattern Recognition Letters*, 2021, **145**: 58–66
34. Hao W, Ren C, Han M, Li F, Liu Z. Cattle body detection based on YOLOv5-EMA for precision livestock farming. *Animals*, 2023, **13**(22): 3535
35. Chen B, Zhang Z, Liu N, Tan Y, Liu X, Chen T. Spatiotemporal convolutional neural network with convolutional block attention module for micro-expression recognition. *Information*, 2020, **11**(8): 380
36. Zaryabi E H, Moradi L, Kalantar B, Ueda N, Halin A A. Unboxing the black box of attention mechanisms in remote sensing big data using XAI. *Remote Sensing*, 2022, **14**(24): 6254