Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

STN-MobileNetV2: A Hybrid Lightweight Deep Learning Model for Maize Disease Classification

Weirong Chen
Fang Yang
11961613@qq.com

Junde Chen
MD Suzauddola
Hanlin Chen
Shuyu Jiang

Research Article

Keywords: Maize plant diseases, Lightweight Deep learning Model, STN-MobileNetV2, Image identification

Posted Date: January 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8359170/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

STN-MobileNetV2: A Hybrid Lightweight Deep Learning Model for Maize Disease Classification

Abstract: Agriculture occupies an essential role because of the demand for food, and it is a crucial source of human income in many countries, especially in developing countries. In the world, China is a large agricultural country and many people rely on agricultural production for a living. Maize is widely cultivated as a kind of the major dominant crop, while the diseases of maize not only influence the maize plantation but also the economic development. Severe maize diseases may even result in no harvest of grains. Thereupon, looking for an accurate, fast, automatic, and low-cost approach to conduct maize disease recognition is of great realistic importance. In this study, we put forward a novel image-based network architecture for maize disease identification. The proposed network integrates a Spatial Transformer Network (STN) with MobileNetV2, forming a hybrid architecture termed STN-MobileNetV2. This model leverages MobileNetV2's pre-trained efficiency while enhancing spatial invariance through STN, enabling robust recognition of maize disease types.We also improved the Focal-Loss (FL) function to enable it to handle multi-class problems and keep more attention on minor lesion characteristics. When benchmarked against other state-of-the-art (SOTA) techniques, the presented approach exhibits superior efficacy. Specifically, it attains an average recognition accuracy of 88.00% and a specificity of 92.00% using the publicly available dataset.Even when eight disease types are considered, the proposed approach realizes a mean accuracy and specificity of 92.84% and95.85% on the locally captured maize disease images. Results derived from the experiments demonstrate that the proposed model is highly effective in the identification of maize-related diseases.

Keywords: Maize plant diseases; Lightweight Deep learning Model; STN-MobileNetV2; Image identification.

1. Introduction

Maize is one of the most dominant crops which are important for the world's grain yields. Its plantation area and total output rank first in the world apart from rice and wheat (Yadav and Dutta, 2018,Zhang et al.,2018). Apart from its role as an important source of carbohydrates for the human diet, maize is cultivated for animal feed, cooking oil, cornmeal and other products, and is commonly employed as a raw material in industrial manufacturing (Kusumo et al., 2018).Whereas, maize is prone to be infected with diverse diseases. Serious maize diseases have devastating impacts on maize plants and can lead to a grain failure completely. Early recognition and detection can suppress the pandemic of maize diseases and

* **Correspondence:** Fang Yang. Email:yangfang@xit.edu.cn

decrease redundant drug use. However, traditional maize disease identification primarily depends upon the visual observation of experienced planters or plant protection specialists in this domain. This approach is labor-intensive, high-cost, subjective, and cannot be widely promoted(Al-Hiary et al.,2011, Chen et al.,2021). Along with the advancement of informatics and computer technologies, novel approaches to maize disease identification have been developed, with growing attention directed toward the research and practical application of machine learning (ML) and image processing techniques (Lu et al., 2017; Soille et al., 2000).

In the past decades, with the development of computer science and digital cameras, image recognition technique has presented promising performance in many fields, such as healthcare image analysis (Wells III,2016, Miki et al.,2017), food detection (Gökmen and Sügüt,2007, Lopez et al.,2011), manufacturing (Lee et al.,2017,Kien et al.,2019), emotion recognition (Hung et al.,2019), among others (Duan et al.,2017,Zhang et al.,2016, Mondal and Bours,2017) and has attained impressive results. Using the support vector machine (SVM) method, Song et al. (2007) classified 3 types of maize plant diseases and they achieved the highest accuracy of 89.6%. Although satisfactory results were achieved, the image dataset used is relatively limited, comprising only 262 leaf images. After comparing with the classical SVM, Zhang et al. (2015) developed a genetic algorithm (GA)-based SVM for identifying maize plant diseases and they attained the best accuracy of 90.25%. Based on the image processing technique, Chen and Wang (2011) trained a probabilistic neural network (PNN) to identify maize leaf diseases and realized the optimum recognition accuracy of 90.4%, etc. Despite the reasonably good results reported in the literature, traditional machine learning (ML) methods suffer from several drawbacks, such as their reliance on manually designed features, susceptibility to overfitting, lack of robustness, and low accuracy. Recently, a new ML technique, namely deep learning (DL), has become the preferred method to overcome some challenges (Barbedo, 2018). Nevertheless, using the DL approach in agriculture has gained little attention so far (Kamilaris and Prenafeta-Boldú, 2018). Most existing works focus on plant disease identification from publicly available plant leaf image datasets (Sardogan et al.,2018, Barbedo et al.,2018, Chen et al.,2023, Ferentinos, 2018). Ferentinos (2018) introduced a DL model to identify 14 species and 26 plant disease types. Their method achieves 99.35% accuracy on a held-out test set. Reference (Ma et al.,2018) reported a deep convolutional neural network (DCNN) to perform symptom-wise identification of four types of cucumber crop diseases. Their approach outperformed other SOTA methods and attained the highest recognition accuracy of 93.4%. **Zhang et al. (2018) proposed an optimized deep convolutional neural network (DCNN) for maize disease recognition, achieving an average identification accuracy of 98.8% with their model. Despite the relatively promising results documented in existing studies, the diversity of materials utilized remains limited to date. Notably, most experimental materials consist of plant leaf images captured under controlled laboratory conditions, as opposed to natural field environments in real-world scenarios (Syed-Ab-Rahman et al., 2022).**

In reality, crop images are taken under a variety of scenarios and the diseases have a wide-ranging effect on both the grain and leaf of maize. Thereupon, to bridge this research gap, this paper introduces a novel DL model named STN-MobileNetV2 to identify maize plant diseases. A transfer learning approach is used and the MobileNetV2 pre-trained on ImageNet is chosen as the backbone in our network model. To enhance the global invariance, the Spatial Transformer Network (STN) is embedded before the MobileNetV2, which is followed by an auxiliary 7×7×2048 convolution layer to extract high-dimensional features. Subsequently, the fully-connected (FC) layer is substituted by a global pooling (GLP) layer, and a new completely-linked (CL) ReLU layer with 1024 neutrons is embedded into the network. **Lastly, a CL Softmax layer with the actual class quantity is utilized as the network's terminal classification layer. Furthermore, the standard Cross-Entropy (CE) loss function is replaced with an improved Focal Loss (FL) function to strengthen the network's learning performance on minor lesion features. By means of this method, the proposed network reduces both the model volume and computational complexity. The subsequent section summarizes our main contributions.**

- We have collected maize disease images from real-world field scenarios and constructed a dataset, where each image was manually labeled with its specific disease category. This dataset is projected to promote the progress of follow-up research on maize disease recognition.
- A lightweight STN-MobileNetV2 model is proposed for recognizing maize disease images with complex background conditions, and it realizes an increasing performance gain compared to other SOTA methods.
- The traditional MobileNetV2 architecture is modified in this study, with the STN module integrated into the pre-trained MobileNetV2 model. Subsequent to this integration, an extra 7×7×2048 convolutional layer is added to facilitate the extraction of high-dimensional features.
- We improved the FL function to enable it to handle with multi-class problems, and to make the model keep more attention on minor lesion characteristics, we utilized the boost FL function to replace the existing CE loss function in the networks.

The remainder of this paper is organized as follows. Section 2 presents an overview of the image dataset utilized and elaborates on the methodology adopted for maize disease recognition.Section 3 devotes to algorithm experiments. Extensive experiments are performed along with the comparison analysis of experimental results. Finally, Section 4 concludes this paper with a summary and highlights avenues for further research.

2. Materials and methods

2.1. Collected Images

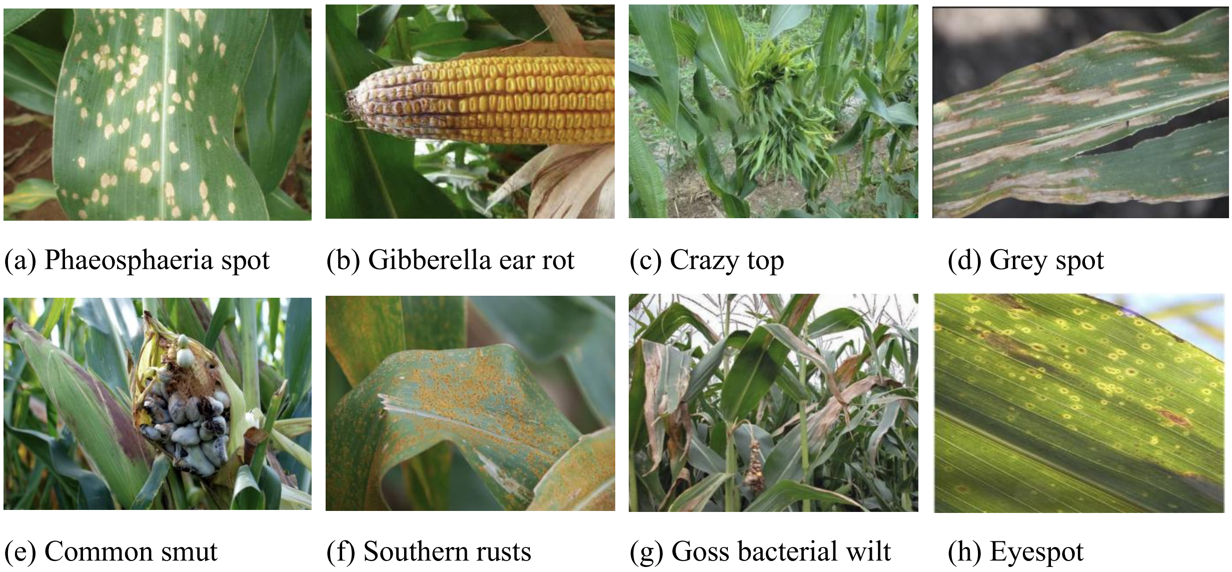
A total of roughly 466 maize plant disease images were furnished by the Plant Quarantine Department of Xiamen Greening Management Center in Xiamen City, China.These maize disease images are collected under complex backdrop conditions and nonuniform lighting strengths. Domain experts have labeled the disease category for every single image, with all the image files saved in the JPG format. The images that belong to the same category are stored in the same folder. There are eight types of maize plant diseases, including the maize eyespot, phaeosphaeria leaf spot, common smut, and others that existed in these collected disease images. For the convenience of following calculation, these maize plant images are evenly transformed into the RGB model with the Photoshop software and the sizes of each image are resized to the dimension of 224×224 pixels. The symptom features corresponding to different maize disease images are summarized in Table 1, with representative examples presented in Figure 1.

2.2. *Overview*

The complete workflow of the proposed maize disease recognition model is elaborated on in the following sections. Initially, the maize plant disease samples that are labeled by the specialists in the plant protection domains are collected for the training samples. The image preprocessing techniques are implemented in the collected images. In particular, new synthetic sample images are generated for model training by using the data augmentation scheme, which includes frequently-used methods like random rotation, flip, angle transformation, and color jittering. Subsequent to preparation, the images are inputted into the proposed network to conduct model training, and the optimal model derived therefrom is applied to maize disease recognition tasks. Finally, the recognition results of maize plant diseases are obtained, and the corresponding recognized maize disease images are employed to refresh the sample library.Figure 2 portrays an overall process of the proposed method, and the specific description is presented in subsequent sections.

Table 1. Symptom characteristics of maize lesions

Maize disease types	Maize disease symptom characteristics		
	Shape	Color	Location
Phaeosphaeria leaf	Circular or Ellipse	Offwhite or yellowish	Leaves
Gibberella ear rot	Mycelial layer on seeds	Red	Corn on cob
Goss bacterial wilt	Block or long strip	Gray or black	Leaves
Crazy top	Tassel deformity	Green or gray	Corncobs
Southern rusts	Punctate or water stain	Bronzing or rust	Leaves
Common smut	Irregular nodules	White gray	Corn on cob
Grey spot	Irregular strip	Straw-colored or brown	Leaves
Eyespot	Ellipse of sesame seed	Edge: pale yellow; Centre: brown	Leaves



(a) Phaeosphaeria spot (b) Gibberella ear rot (c) Crazy top (d) Grey spot

(e) Common smut (f) Southern rusts (g) Goss bacterial wilt (h) Eyespot

Figure 1. Samples of maize plant diseases

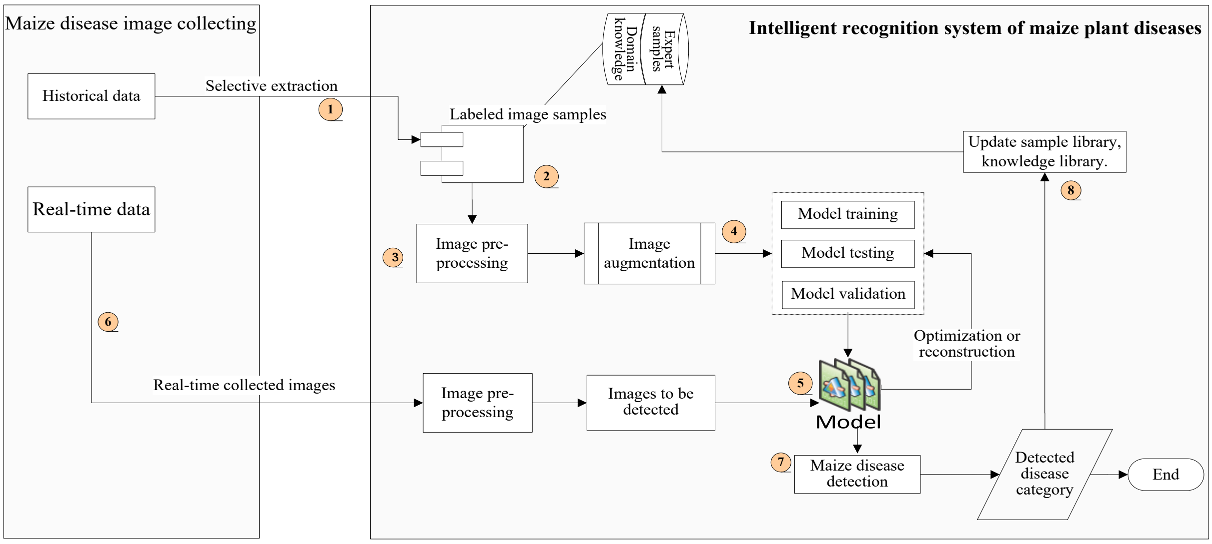


Figure 2. Overall flowchart of maize plant disease identification

2.3. *Deep Neural Networks Architecture*

2.3.1. Spatial Transformer Network

Spatial Transformer Network, or STN in short (Jaderberg et al.,2015), is designed to gain the optimum input and clearly allows spatial manipulation of data within the networks, which learns the transformation parameters automatically. To get the optimal input needed for subsequent networks, the STN uses a small network to transform a set of feature maps into other spatial domains. As shown in Figure 3(a), the module of STN primarily consists of three components: localization network, parameterized sampling grid, and image sampling. Among the three components, the localization network is essentially a regression network, and the major task of which is to input the images U∈R(C, H, W). Through several hidden layers, the localization network (floc) obtains a transformation matrix A_θ from the initial input I. In this study, we utilize a two-dimensional (2-D) affine transformation, which implements the translation, rotation, and scaling transformation determined by the following six parameters, as expressed by (Jaderberg et al.,2015)

$$f_{loc} = A_\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \tag{1}$$

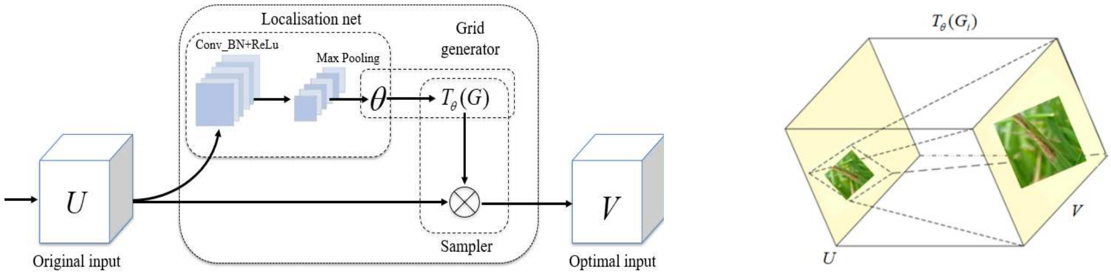
Among them, translation transformation is provided by θ_{13} and θ_{23}, rotation transformation is provided by θ_{11}, θ_{22}, θ_{12}, and θ_{21}, and scaling transformation is provided by θ_{11} and θ_{22}. The second component is the parameterized sampling grid. Using a normal space grid along with transformation matrix A_θ, this component creates a sampling grid T_θ to generate a series of points. These points represent the position of the input I that needs to be sampled to perform the output of transformation, as written in Eq. (2).

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ z_i^t \end{pmatrix} \tag{2}$$

where(x_i^s, y_i^s)represent the source location,$(x_i^t, y_i^t, z_i^t)$are the target location, and A_θ is the transformation matrix. In this way, the final feature map $V \in R(H, W, C)$ can be obtained, as portrayed in Figure 3(b). The third component is image sampling, which takes the input feature U and the sampling grid to generate the optimum inputs. With the location information of T_θ (G_i) from the affine transformation in the previous part, the image sampling must be conducted to obtain features from the input feature map U as well as generating new output V. This can be defined by

$$\begin{gathered} V_i^c = \sum_m^W \sum_n^H U_{mn}^c k(x_i^s - n; \phi_x) k(y_i^s - m; \phi_y) \\ \forall i \in [1, 2, \cdots, W'H'], \forall c \in [1, 2, \cdots, C] \end{gathered} \tag{3}$$

where U_{mn}^c denotes the value at location (n, m) in channel c of the input, $\emptyset_x$and$\emptyset_y$ indicate the interpolation parameters of sampling kernel $k(\,)$ (e.g. bilinear), and V_i^c means the output value of pixel i after T_θ (G_i) transformation. Thus, the above three components are integrated to generate an absolute space converter, which can be added to the convolutional neural networks (CNNs) as an independent module.



(a) Spatial Transformer (b) Affine transformation

Figure 3. The STN module

2.3.2. DWSC

Depth-wise separable convolution, abbreviated as DWSC, is a convolution operation that factorizes a normal convolution into two components, including the point-wise and depth-wise convolutions (Sifre and Mallat,2014). As shown in Figure 4(a), the depth-wise convolution (DWC) implements the convolutional operation on each channel using one convolution kernel for the input map. Then the results of the DWC are input into the point-wise $C = 2\pi r$convolution (PWC) to conduct the 1×1 convolution operation, thereby gaining the final results. Figure4(b) portrays the specific processes of PWC. Mathematically, the formulas of DWC and PWC can be written by Eqs. (4,5), respectively (Howard et al.,2017).

$$G_{i,j,d}^{l+1} = \sum_{i=0}^{H}\sum_{j=0}^{W}\sum_{d=0}^{D} K_{i,j,d} \times F_{i,j,d}^{l} \tag{4}$$

$$\hat{G}_{i,j,d}^{l+1} = \sum_{d=0}^{D} K_d \times \sum_{i=0}^{H}\sum_{j=0}^{W} K_{i,j} \times F_{i,j}^{l} \tag{5}$$

Among them, K is the filter, F indicates an input tensor; W, H, and D represent the width, height, and depth of the input feature map in the 1^{th} layer of the networks; G is the results of the convolution operations. As a typical architecture of mobile-first networks, MobileNet (Howard et al.,2017) has explicitly incorporated the DWSC to satisfy the device resource constraints. Suppose an intermediate feature map of $D_f \times D_f$ with the channel number M, the convolution kernel size of $D_k \times D_k$, and the output channel number of N, the calculational costs of the regular convolution (C_1) and DWSC (C_2) can be separately expressed by:

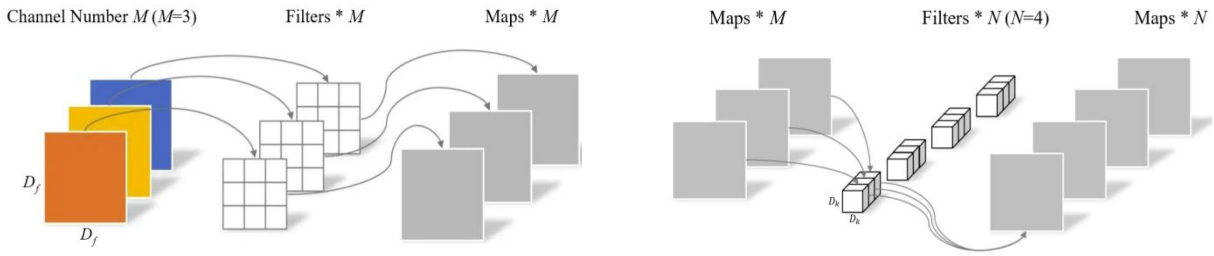
$$C_1 = D_f^2 M N D_k^2 \tag{6}$$

$$C_2 = D_f^2 M (D_k^2 + N) \tag{7}$$

Thereupon, the computing cost ratio of DWSC to that of the regular convolution can be written by:

$$R_{C_2/C_1} = \frac{1}{N} + \frac{1}{D_k \times D_k} \tag{8}$$

Practically, the 3×3 filter is usually employed in the DWSC networks and the output channel number N is big. As a consequence, the ratio of the computational cost of DWSC to that of conventional convolution is approximate 1:9.



(a) DWC (b) PWC

Figure 4. The operation process of DWSC

2.3.3. STN-MobileNetV2 network model

As described previously, STN can efficiently enhance the translation invariance of CNNs. With the capability of affine transformation, STN can locate target regions of images by dynamically implementing spatial geometric operations, thereby transforming the input images to the optimum state and improving the classification accuracy. On the other hand, MobileNets are a class of mobile-first CNNs based on DWSC and have presented excellent ability in handling both large-scale and small-scale issues. Therefore, motivated by the impressive performances, the STN and MobileNetV2 are chosen in our method. By combining the advantages of both, the STN and pre-trained MobileNetV2 are fused to form a novel networks, which we term STN-MobileNetV2, is employed to predict maize disease categories. To improve the global invariance of the networks based on the structure of the heuristic affine transformation matrix, we modify the traditional MobileNetV2 by embedding an STN module behind the input layer of the network, which is followed by a newly-added 7×7×2048 convolutional layer for extracting deep-level features. Subsequently, the fully-connected (FC) layer is substituted by a GLP layer, and a new FC ReLU layer with 1024 neutrons is incorporated into the networks. Lastly, a FC Softmax layer with the real number of classes is utilized as the tail classification component of the presented network. To summarize, the STN-MobileNetV2 primarily comprises two parts: the middle section consists of the pre-trained module acting as the backbone feature extractor, and the two ends are auxiliary layers that focus on improving the capabilities of image interpretation and high-level feature extraction.More concretely, the following steps are conducted to obtain the model.

- The first step is to initialize the network weights. The pre-trained MobileNetV2 that is inferred on large-scale ImageNet dataset is used for the initialization and it presents excellent performance for image identification and classification.

- The second step performs the adaption of the model. In order to resolve new problems on the objective dataset, adaptive optimization is carried out by freezing the weights of the lower convolutional layers and removing the original fully connected (FC) layer.

Implement the extension on the model. After performing step 2, the STN module is added behind the input layer for the pre-trained MobileNetV2, which is extended by a 7×7×2048 convolutional layer as well as a GLP layer followed by a new CL ReLU layer with 1024 neutrons. Then, inference is performed on the objective image dataset using all layers of the newly generated network, and the last FC Softmax function is utilized to classify maize plant diseases. Figure 5 depicts the proposed STN-MobileNetV2 structure, and the primary parameters of the presented model are outlined in Table 2.

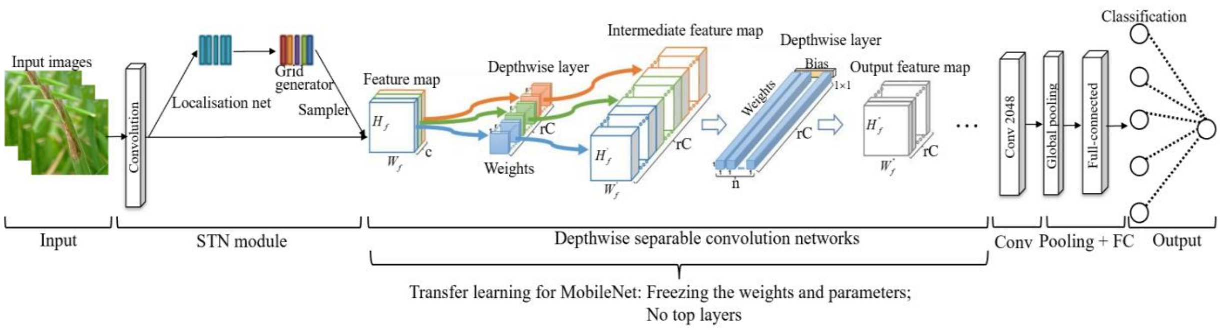


Figure 5. The proposed STN-MobileNetV2 architecture

Table 2. The primary parameters of STN-MobileNetV2

Modules	Input shapes	Expansion factors	No. of output channel	Repeats	Strides
Input	224×224×3	-	3	1	-
STN	224×224×3	-	3	1	-
Conv2d	224×224×3	-	32	1	2
Boottle-neck	112×112×32	1	16	1	1
Boottle-neck	112×112×16	6	24	2	2
Boottle-neck	56×56×24	6	32	3	2
Boottle-neck	28×28×32	6	64	4	2
Boottle-neck	14×14×64	6	96	3	1
Boottle-neck	14×14×96	6	160	3	2
Boottle-neck	7×7×160	6	320	1	1
Conv2d 1×1	7×7×320	-	1280	1	1
Avgpool 7×7	7×7×1280	-	-	1	1
Conv2d 7×7	7×7×1280	-	1280	1	1

Conv2d 7×7	7×7×2048	-	2048	1	1
Globalpool	1×1×2048	-	-	1	1
FC	1×1×1024	-	-	1	1
Softmax	1×1×k	-	k	1	1

We apply the Adam optimizer which is a frequently-used stochastic optimization solver in DL models for updating network weights.

$$w_c = w_{c-1} - r * \hat{m}_c / (\sqrt{\hat{v}_c} + \varepsilon) \tag{9}$$

In Eq.(9), w represents the weight parameters, c indexes of classes, r implies the learning rate, and $\hat{m}_c$ as well as $\hat{v}_c$ denote the bias-corrected first and second moment, respectively. In general, CE function is usually utilized as the loss function in deep networks, and the CE function can be formulated by:

$$L = -\sum_{c=1}^{N} y_c \log(p_c) \tag{10}$$

where y_c represents an indicator variable (1 or 0), and p_c denotes the probability of sample predicted as class c. On the ground of this, Lin et al. (2017) introduced a FL function since the CE function regarded the weights of the negative and positive samples as the same. Mathematically, the FL function can be formulated as:

$$L = -w_c (1 - p_c)^{\gamma} \log(p_c) \tag{11}$$

Nevertheless, the classical FL function is designed for handling binary classification problems and the maize disease recognition is a multi-class task. Therefore, to address this challenge, we expand it to be suitable for multi-class tasks. The enhanced FL function can be expressed by:

$$L_{mult} = -\sum_{c=1}^{N} w_c (1 - p_c)^{\gamma} y_c \log(p_c) \tag{12}$$

$$y_c = \begin{cases} 1 & c = acutal\_class \\ 0 & c \neq acutal\_class \end{cases} \tag{13}$$

3. Experimental Analysis

In the experiments, apart from the image preprocessing implemented with Photoshop software, the core algorithms are performed using Python 3.6, where the widely used libraries such as Keras-GPU (2022), Scikit-Learn, Tensorflow, and OpenCV-python3 are employed and sped by graphic processing unit (GPU). The hardware configurations for operating the Python DL architecture contains Intel 2.10GHz E5-2620 CPU, RTX 2080 GPU(2022), and 64GB RAM, which is applied for algorithm operation.

3.1. Experiments on Public Datasets

As an openly accessible repository, the PlantVillage database is widely utilized to test algorithms in the field of crop disease identification (Hughes and Salathé, 2015).To probe the efficacy of the proposed model, we implement the experiments on the Maize crop dataset of the Plantvillage library. 3,852 color maize leaf images are provided in this dataset, which is classified into 4 types, such as gray_spot (513 samples), common_rust (1,192 samples), northern_blight (985 samples), and healthy category (1,162 samples).It is noteworthy that the images were captured under conditions of simple backgrounds and regular light intensities, with some samples being identical leaf images taken from different angles. All images have been resized uniformly to 256×256 pixels to fit the model. Figure 6 summarizes the representative sample images. Because of the uneven distribution of samples in this dataset, the number of gray_spot disease class is lower than that of other classes in this dataset. Therefore, to guarantee the diversity of samples and suppress the risk of overfitting, the data enhancement techniques, such as random flipping, rotation, scale transformation, and color jittering are implemented on the raw data, and no less than 500 images per category are ensured for the model training. Figure 7 displays the representative synthetic samples.

(a) Normal sample (b) Gray Spot (c) Common Rust (d) Northern Blight

Figure 6. Representative samples in maize diseases dataset

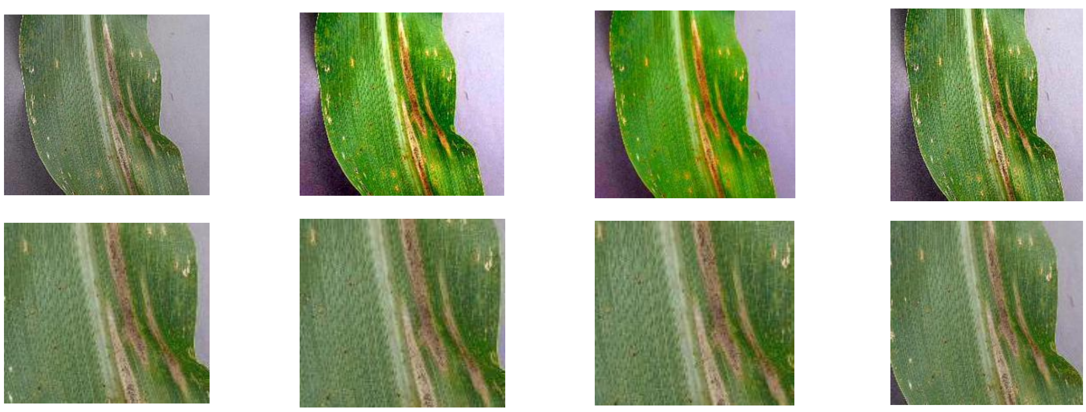


Figure 7. The augmented maize disease sample images

Table 3. The hyperparameters used for this study

Parameters	Values
Fitness function	Enhanced FL function (see Eq. (12))
Mini-batch size	32
Epoch	30
Learning rate	0.001
Activation function	ReLU; Softmax
Optimizer	Adam

By adopting the methodology introduced in Section 2.3, we conducted training and testing of the model on the maize disease image dataset.Except that the 100 pcs images per class are retained as the test dataset to verify the efficacy of the model, the random allocation ratio of samples between the validation set and the training set is set to 3:7. Based on the experimental comparison, we have selected the Adam (Kingma and Ba,2014) optimizer to train the model, with a learning rate of 1.0×10^{-3} and a min-batch size of 32. Concretely, the hyper-parameters of the model applied in the experiments are summarized in Table 3. Moreover, to compare the proposed method, the four SOTA methods, including the DenseNet (Huang et al.,2017), VGGNet (Simonyan and Zisserman,2014), Inception V3 (Szegedy et al.,2016), and ResNet (He et al.,2016), are chosen to compare models. The transfer learning (TL) method is used for the training of these models, where the weight parameters pre-trained from ImageNet are embedded into the networks, and the tail layers are substituted by new fully-connected (FC) classification layers with the actual number of categories. A series of extensive experiments were implemented on the dataset, and the training performance of various methods is compiled in Table 4.

Table 4. Computational results on the public maize dataset

Models	10 epochs			30 epochs			
	Training *Acc* %	Test *Acc* %	Training loss	Training *Acc* %	Test *Acc* %	Training loss	Test loss
Inception V3	86.06	67.67	0.4567	93.35	66.00	0.2538	1.0591
VGGNet-19	79.06	**78.17**	0.8700	85.20	**83.83**	0.5466	0.5428
ResNet-50	71.84	66.50	0.9285	74.77	70.50	0.7270	0.7627
DenseNet-201	**92.28**	70.17	0.2652	97.07	69.00	0.1200	0.8826

MobileNet-V2	90.99	66.67	0.3610	96.21	67.00	0.2443	0.8790
STN-MobileNetV2	87.56	69.83	0.4437	**97.43**	72.17	0.2172	0. 7429

Table 5.A summary of Some related works

ID	Models	Recognition Accuracy	No.of training samples	Superior to other maize disease detection models
1	YOLOv8-GO (Jiang et al., 2024)	88.4%	5793	YOLOv8
2	ICPNet(Yang et al., 2024)	87.3 %	4107	MobileNetV2, DenseNet-121, ResNeXt-50, ResNet-50,CMT, ICAI-V4,et al.
3	MC-ShuffleNetV2(Zhu & Gao, 2024)	99.86 %	2725	LeNet, AlexNet, MobileNetV2, and EfficientNetV2.
4	LSANNet(Zhang et al., 2024)	94.35%	1246	LSANNet, MobileNetV3S, EfficientNetB0, VGG16, DenseNet201, Xception, and ResNet50
5	VGG16((Askale et al., 2025)	95%	4188	VGG16,AlexNet,ResNet50
6	EfficientNet(Xiaowei & Hua, 2025)	98.32%	4187	ResNet34, DenseNet121, MobileNet V2, SqueezeNet, EfficientNet B0 ， EfficientNetB4

7	ICS-ResNet(Ji et al., 2024)	98.87%	8280	CSPNet, InceptionNet_v3, EfficientNet, ShuffleNet, MobileNet, ResNet50, ResNet101, and ResNet152
8	LFMNet(Hu et al., 2024)	94.12%	3141	ResNet50， Inception v3， DenseNet 121， MobileNet (V3-large)， ShuffleNet V2
9	STN-MobileNetV2	97.43%	3852	Inception V3， VGGNet-19， ResNet-50， DenseNet-201, MobileNet-V2

It can be observed from Table 4 that the proposed method has outperformed the other influential CNNs on the publically-available maize dataset except for the VGGNet-19. After 10 and 30 epochs of training, the accuracy of the STN-MobileNetV2 has realized 87.56% and 97.43%, respectively, and the test accuracy achieves 72.17% after 30 epochs of training. By contrast, the training accuracy of VGGNet separately realizes 79.06% and 85.20% after 10 and 30 epochs of training, and the test accuracy reaches 83.83% after training for 30 epochs. Whereas the model size of VGGNet is more than that of our proposed model. In addition, we select both the proposed model and VGGNet to perform inference on unseen images for classification tasks, with Figure 8 depicting the confusion matrices corresponding to their predicted outcomes. It can be also inferred from Table 5 that due to various factors have an impact on the final results of various models leading to the results do not absolutely comparable, including the influence of data sources, data quality and diversity, environmental conditions and model parameter settings.This applies especially to the number of model parameters. To probe into the efficacy of the proposed model, we employed a set of widely used metrics—Accuracy (Acc), Recall (Rec), and Specificity (Spe)—for our experimental analysis, and their calculation formulas are listed below.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{14}$$

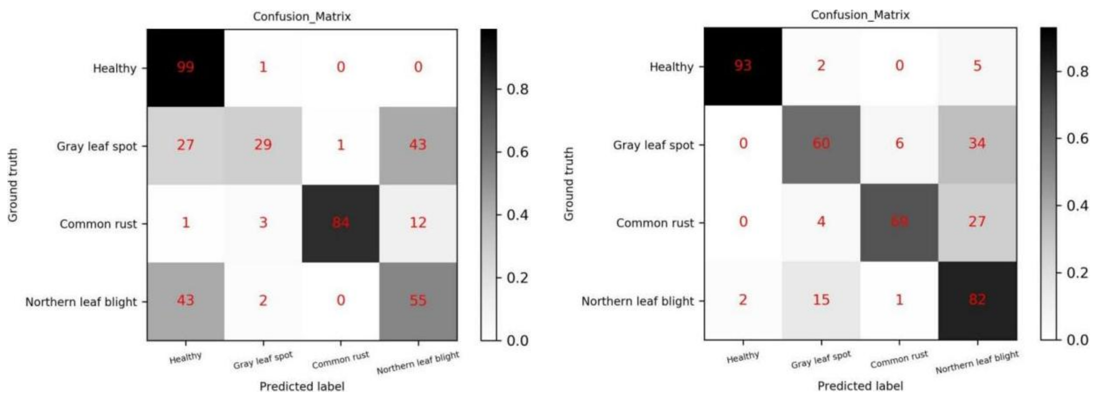
$$Recall = \frac{TP}{FN + TP} \tag{15}$$

$$Specificity = \frac{TN}{FP + TN} \tag{16}$$

Among them, *TP*, *TN*, *FP*, and *FN* represent the true positive, true negative, false positive, and false negative, respectively. Table 6 summarizes the evaluation indicator value of the recognition results for the proposed model.

Table 6. Performance of recognition results

Types	Recognized number	Accurate number	Accuracy (%)	Recall (%)	Specificity (%)
Healthy	100	93	97.75	93.00	**99.33**
Gray spot	100	60	84.75	60.00	93.00
Common rust	100	69	90.50	69.00	97.67
Northern blight	100	82	79.00	82.00	78.00
Average	100	76	88.00	76.00	92.00



(a) Predicted results of VGGNet-19 (b) Predicted results of STN-MobileNetV2

Figure 8. Confusion matrices of the VGGNet and the proposed method

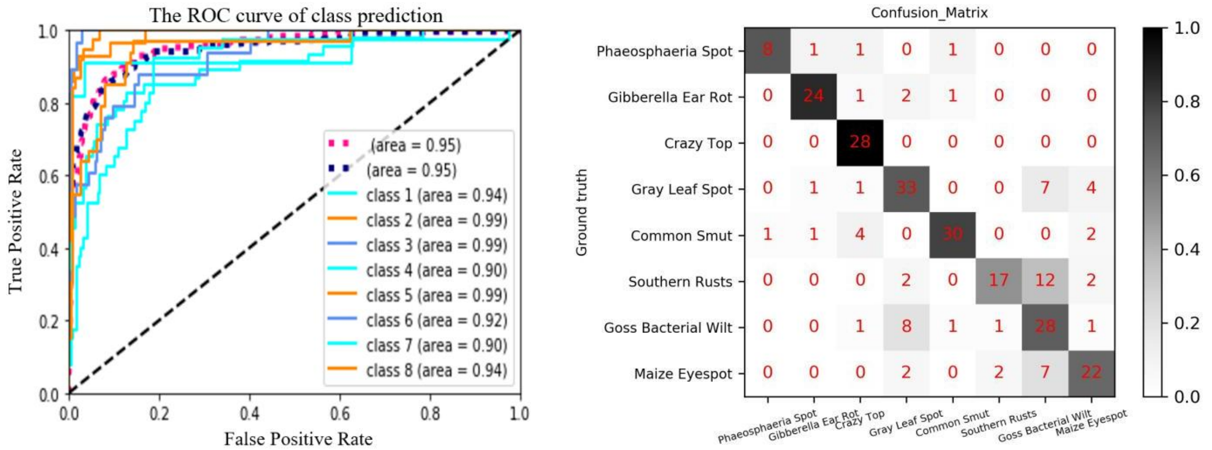
As shown in Figure 8(a), while VGGNet exhibits better performance than STN-MobileNetV2 on the training set, the latter surpasses the former in the classification of maize disease categories. However, when the model trained by VGGNet was employed for the class prediction of unseen images, there was a significant drop in accuracy. That is, the generalization ability of the VGGNet is weaker than the proposed STN-MobileNetV2 model, which may be caused by the over-fitting of VGGNet. Additionally, as shown in Table 6, the average recognition accuracy reaches no less than 88.00%, indicating that the STN-MobileNetV2 possesses excellent proficiency in identifying maize disease images from the public dataset. Notably, beyond distinguishing between normal and diseased maize plants, the proposed model can further discriminate specific types of maize crop diseases—achieving an average specificity of 92.00% in this fine-grained classification task, which fully verifies the model’s validity.

3.2. Experiments on Local Dataset

To further verify the validity, the proposed model is tested on the locally collected maize disease image dataset, which is captured under a wild field scenario with cluttered backdrops and uneven illumination intensity. In line with the experiments performed in Section 3.1, the proportion of samples in the training set to those in the test set is maintained at 7:3.except for a small portion of images used to validate the model. Also, it is essential to emphasize that the data enhancement scheme, such as color jittering, random flipping, rotation, and scale transformation, is implemented on the raw data. No less than 200 images per category are guaranteed for the augmented training samples. Concretely, the specific processes are presented as follows.

- Resize the images: For the raw images and augmented synthetic samples, the dimensions of them have all been resized to a fixed size of 224×224 pixels for fitting models.
- Image pre-processing: To retain the original information and avoid the distortion of raw images, the image pre-processing technique that blackens the shorter sides of images is implemented to keep the images in the same proportion.
- Model training: The proposed STN-MobileNetV2 model is trained using the locally collected maize image dataset and extensive experiments are implemented with the shuffled images.

After implementing the above processes, the obtained satisfactory model has been employed to recognize the maize disease types from new images. Figure 9 portrays the Receiver Operating Characteristic (ROC) and confusion matrix charts of the identification results. Table 7 presents the corresponding estimation metrics, and the test accuracy of the proposed STN-MobileNetV2 compared to those of other methods are summarized in Table 8.



(a) ROC curve (b) Confusion matrix

Figure 9. The performance of recognition results

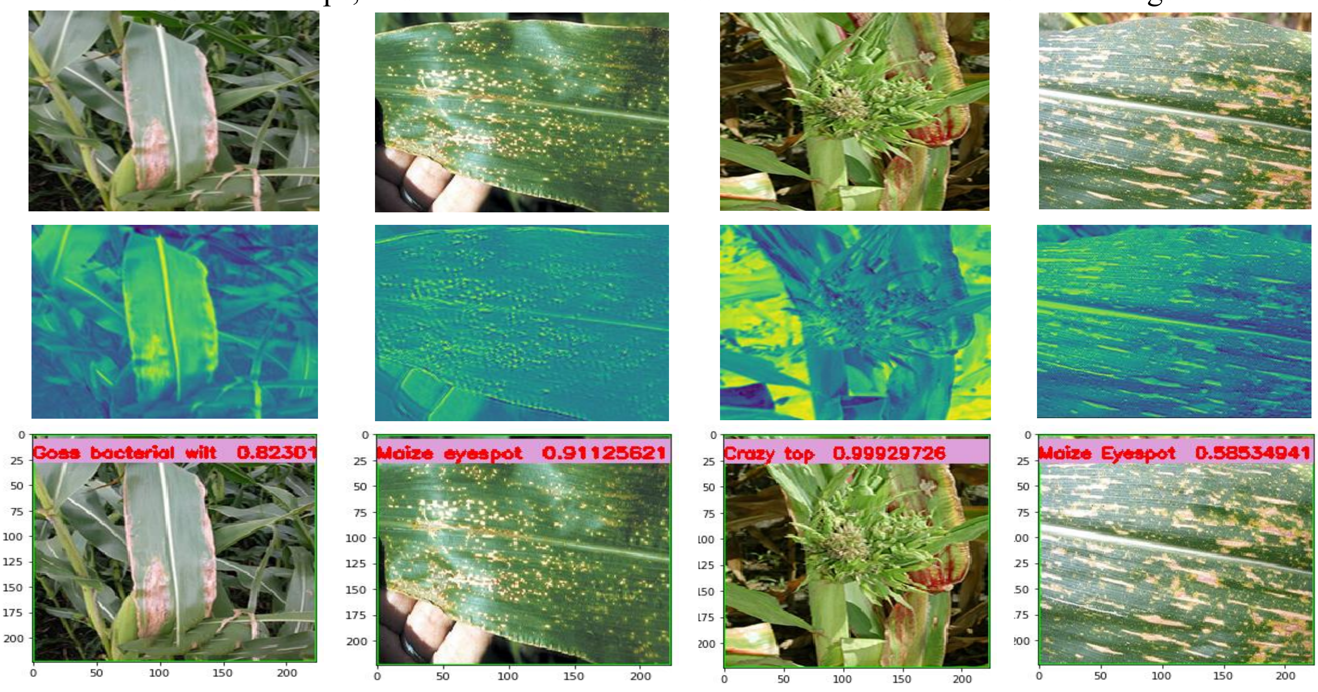
Table 7. The specific test performance of the proposed model

No.	Maize disease classes	Recognized number	Accurate number	Accuracy (%)	Recall (%)	Specificity (%)
1	Phaeosphaeria leaf spot	11	8	**98.44**	72.72	**99.59**
2	Gibberella ear rot	28	24	97.27	85.71	98.69
3	Crazy top	28	28	96.89	100.00	96.51
4	Gray spot	46	33	89.49	71.74	93.36
5	Common smut	34	30	95.72	78.95	98.63
6	Southern rusts	33	17	73.61	51.52	92.31
7	Goss bacterial wilt	40	28	85.21	70.00	88.02
8	Maize eyespot	33	22	92.22	66.67	95.98
–	Average	–	–	92.84	73.93	95.85

Table 8. The test accuracy on the local maize dataset

No.	Model	Accuracy (%)	Recall (%)	Specificity (%)
1	Inception V3	88.12	67.23	89.27
2	VGGNet-19	88.63	73.52	91.17
3	ResNet-50	90.06	70.38	92.23
4	DenseNet-201	91.26	72.11	93.52
5	MobileNet-V2	89.17	69.81	91.98
6	Proposed model	**92.84**	73.93	**95.85**

From Figure 9(a) it can be easily visualized that the ROC curve tends to the left top corner of the figure, demonstrating the effectiveness of the proposed STN-MobileNetV2, which can be also revealed by the confusion matrix of Figure. 9(b). The accurate number is 8 for the recognition of “Phaeosphaeria leaf spot” disease in 11 samples. Except for 4 misidentifications, there are 24 samples correctly recognized in the class of “Gibberella Ear Rot”. More than that, all 28 samples of the “Crazy Top” class are predicted correctly. Thus, 190 maize disease images exactly match the actual categories in 253 samples, and we achieve the mean identification accuracy of 92.84% on the local maize disease images, as presented in Table 7. Moreover, from Table 8 it can be also observed that the proposed STN-MobileNetV2 shows outperformance relative to other methods. On the other hand, there are some examples of “Gray Leaf Spot” and “Phaeosphaeria Spot” is misidentified into the class of “Maize Eyespot”, which is owing to the situation that some diseases such as the “Maize Eyespot” and “Gray Leaf Spot” appear on the same leaf. Additionally, the cluttered backdrop condition and uneven lighting intensity can influence this as well. Figure 10 displays representative recognition samples: the upper section contains raw images, the middle section presents extracted feature maps, and the lower section shows the identified maize disease images.



(a) Goss bacterial wilt (b) Maize eyespot (c) Crazy Top (d) Gray leaf spot

Figure 10. The recognized maize disease samples

From Figure 10 it can be visualized that the recognized classes of these samples are generally compatible with their true classes of them, and most maize diseases have been accurately recognized by the STN-MobileNetV2 except for individual samples. The critical rationale behind the proposed approach’s remarkable performance is that the STN-MobileNetV2 incorporates both the STN and a pre-trained MobileNetV2. Here, the STN is tasked with locating target image regions through dynamic spatial geometric operations, which transforms input images to their optimal form and improves the precision of identification. Moreover, based on lightweight network architecture, MobileNetV2 reveals outstanding ability in handling both large-scale and small-scale tasks. Combining the merits of both, the proposed STN-MobileNetV2 architecture achieves a satisfactory result. Despite the promising performance, some different diseases such as the "Maize Eyespot" and "Gray Leaf Spot" appear in the same leaf, which leads to incorrect recognition. In addition, the seriously cluttered field background and inconsistent illumination strength affect the extraction of lesion image features, resulting in the individual misidentification issue. Nevertheless, most maize disease types have been effectively identified by the proposed method.As illustrated in Table 7, the average recognition accuracy is more than 92.84% and the specificity realizes 95.85% in multiple experiments, indicating the significant capability of the proposed STN-MobileNetV2 for recognizing maize plant diseases. Based on the experimental analysis, we can conclude that the STN-MobileNetV2 architecture is effective for recognizing maize diseases and can be also transplanted to other related fields like target recognition, fault detection, online failure diagnosis, and others.

4. Conclusions

This study puts forward a novel DL architecture named STN-MobileNetV2 to recognize maize plant diseases. Transfer learning is used in our solution. The STN module combined with the MobileNetV2 pre-trained from the ImageNet are selected in our model. To improve the global invariance, the existing MobileNetV2 is modified by embedding an STN module behind the input layer of the network, which is followed by an auxiliary 7×7×2048 convolutional layer to extract deep-level features. Subsequently, the completely-linked layer is replaced by a GLP layer, and a new completely-linked ReLU layer with 1024 neutrons is incorporated into the networks. Ultimately, a FC Softmax layer with the actual number of classes is employed as the tail layer for the classification. Additionally, instead of the existing CE Loss function, an enhanced FL function is used to enhance the feature-extracting capability of minor lesions. By this means, the pre-trained MobileNetV2 along with the STN module are fused to form a novel network architecture, which reduces the parameter number and calculation cost of the model. The proposed model outperforms the other compared methods and attains the average specificity of 92.00% and 95.85% on the publicly-available and local datasets, respectively. Experimental findings demonstrated the validity of the proposed model for maize disease identification. In future work, we intend to deploy the model in embedded systems and edge devices to automatically identify and monitor the broad range of maize disease information. Whilst, we hope to transplant it to more practical applications.

Availability of data and material

The data and material that support the findings of this study are available from the corresponding author upon reasonable request.

References

Askale et al., 2025 G.T.Askale, A.B.Yibel,B.M. Taye, G.D. Wubneh, Mobile based deep CNN model for maize leaf disease detection and classification.Plant Methods, 21(2025), 72.

Al-Hiary et al.,2011 H. Al-Hiary, S. Bani-Ahmad, M. Reyalat, M. Braik, Z. Alrahamneh, Fast and accurate detection and classification of plant diseases. *Int. J. Comput. Appl.*, **17** (2011), 31–38.

Barbedo, 2018 J. G. Barbedo, Factors influencing the use of deep learning for plant disease recognition. *Biosyst. Eng*., **172** (2018), 84–91.

Barbedo et al.,2018 J. G. A. Barbedo, Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Comput. Electron. Agr*., **153** (2018), 46–53.

Chen et al.,2021 J. Chen, W. Wang, D. Zhang, A. Zeb, Y.A. Nanehkaran, Attention embedded lightweight network for maize disease recognition. *Plant Pathology*, *70*(2021), 630-642.

Chen et al.,2023 J. Chen, W. Chen, Y. A. Nanehkaran, MD Suzauddola, MAM-IncNet: an end-to-end deep learning detector for Camellia pest recognition. *Multimedia Tools and Applications*. (2023), 1-6.

Chen and Wang 2011 L. Chen, L. Y. Wang, Research on application of probability neural network in maize leaf disease identification. *J. Agr. Mech*., **33** (2011), 145–148.

Duan et al.,2017 Y. Duan, F. Liu, L. Jiao, P. Zhao, L. Zhang, SAR image segmentation based on convolutional-wavelet neural network and Markov random field. *Pattern Recognition*, **64** (2017), 255–267.

Ferentinos, 2018 K. P. Ferentinos, Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agr*., **145** (2018), 311–318.

GeForce GTX 1060. Available online: https://www.nvidia.com/ en-us/geforce/ products/ 10serie/geforce-gtx-1060/specifications (accessed on 17 Jun, 2022).

Gökmen and Sügüt,2007 V. Gökmen, I. Sügüt, I. A non-contact computer vision based analysis of color in foods. *Int. J. Food Eng*., **3** (2007), 1–12.

He et al.,2016 K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recogn*., (2016), pp. 770–778.

Howard et al.,2017 A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint* (2017) arXiv:1704.04861.

Huang et al.,2017 G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, (2017), pp. 4700–4708.

Hu et al., 2024 J. Hu, X. Jiang, J. Gao, X. Yu, LFMNet: a lightweight model for identifying leaf diseases of maize with high similarity. Frontiers in Plant Science, 15(2024), 10.3389.

Hughes and Salathé,2015 D. Hughes, M. Salathé, An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint*, (2015), arXiv:1511.08060.

Hung et al.,2019 J. C. Hung, K. C. Lin, N. X. Lai, Recognizing learning emotion based on convolutional neural networks and transfer learning. *Appl. Soft Comput.*, **84** (2019), 105724.

Jaderberg et al.,2015 M. Jaderberg, K. Simonyan, A. Zisserman, Spatial transformer networks. *Adv. Neural Inform. Proc. Syst.*, **28** (2015).

Jiang et al., 2024 T. Jiang,X. Du, YOLOv8-GO: A Lightweight Model for Prompt Detection of Foliar Maize Diseases. Applied Sciences (Switzerland),14 (2024),10004.

Ji et al., 2024 Z. Ji, S. Bao, M. Chen,L. Wei, ICS-ResNet: A Lightweight Network for Maize Leaf Disease Classification.Agronomy , 14(2024), 10.3390.

Kamilaris and Prenafeta-Boldú, 2018 A. Kamilaris, F. X. Prenafeta-Boldú, Deep learning in agriculture: A survey. *Comput. Electron. Agr.*, **147** (2018), 70–90.

Keras-GPU. Available online: https://anaconda.org/anaconda/keras-gpu, accessed on 17 Jun, 2022.

Kien et al.,2019 B. H. Kien, D. Iba, Y. Ishii, Y. Tsutsui, N. Miura, T. Iizuka, I. Moriwaki, Crack detection of plastic gears using a convolutional neural network pre-learned from images of meshing vibration data with transfer learning. *Forschung im Ingenieurwesen*, **83** (2019), 645–653.

Kingma and Ba,2014 Kingma, D.P., and J. Ba, Adam: A method for stochastic optimization. *arXiv preprint*, (2014), arXiv:1412.6980 .

Kusumo et al.,2018 B. S. Kusumo, A. Heryana, O. Mahendra, H. F. Pardede, Machine learning-based for automatic detection of corn-plant diseases using image processing. In: *2018 Int. Conf. Comput., Control, Inform. Appl.*, (2018), pp. 93–97.

Lee et al.,2017 S. J. Lee, J. P. Yun, G. Koo, S. W. Kim, End-to-end recognition of slab identification numbers using a deep convolutional neural network. *Knowledge-based Syst.*, **132** (2017), 1–10.

Lin et al., 2017 T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection. In: *Proc. IEEE Int. Conf. Comput.Vis.*, (2017), pp. 2980–2988.

Lopez et al.,2011 J. J. Lopez, M. Cobos, E. Aguilera, Computer-based detection and classification of flaws in citrus fruits. *Neural Comput. Appl.*, **20** (2011), 975–981.

Lu et al.,2017 Y. Lu, S. Yi, N. Zeng, Y. Liu, Y. Zhang, Identification of rice diseases using deep convolutional neural networks. *Neurocomputing*, **267** (2017), 378–384.

Ma et al.,2018 J. Ma, K. Du, F. Zheng, L. Zhang, Z. Gong, Z. Sun, A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network. *Comput. Electron. Agr.*, **154** (2018), 18–24.

Miki et al.,2017 Y. Miki, C. Muramatsu, T. Hayashi, X. Zhou, T. Hara, A. Katsumata, H. Fujita, Classification of teeth in cone-beam CT using deep convolutional neural network. *Comput. Biol. Med.*, **80** (2017), 24–29.

Mondal and Bours,2017 S. Mondal, P. Bours, A study on continuous authentication using a combination of keystroke and mouse biometrics. *Neurocomputing*, **230** (2017), 1–22.

Sardogan et al.,2018 M. Sardogan, A. Tuncer, Y. Ozen, Plant leaf disease detection and classification based on CNN with LVQ algorithm. In: *2018 3rd Int. Conf. Comput. Sci. Eng.*, (2018), pp. 382–385.

Sifre and Mallat,2014 L. Sifre, P. S. Mallat, *Rigid-motion scattering for image classification author*. Supervisor: Prof. Stéphane Mallat. Ph.D. Thesis. Ecole Polytechnique, **2** (2014).

Simonyan and Zisserman,2014 K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, (2014), arXiv:1409.1556.

Soille et al.,2000 P. Soille, Morphological image analysis applied to crop field mapping. *Image Vision Comput.*, **18** (2000), 1025–1032.

Song et al., 2007 K. Song, X. Y. Sun, J. W. Ji, Corn leaf disease recognition based on support vector machine method. *Trans. CSAE*, **23** (2007), 155–157.

Sun & Huo, 2025 X.Sun, H. Huo, Corn leaf disease recognition based on improved EfficientNet. IET Image Processing,19 (2025), e13288.

Syed-Ab-Rahman et al.,2022 Syed-Ab-Rahman, S.F., Hesamian, M.H. and Prasad, M., Citrus disease detection and classification using end-to-end anchor-based deep learning model. *Applied Intelligence*, 52 (2022), 927-938.

Szegedy et al.,2016 C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, (2016), pp. 2818–2826.

Wells III,2016 W. M. Wells III, Medical image analysis–past, present, and future. *Med. Image Anal.*, **33** (2016), 4–6.

Yadav and Dutta, 2018 A. Yadav, M.K. Dutta, An automated image processing method for segmentation and quantification of rust disease in maize leaves. In: *2018 4th Int. Conf. Comput. Intell. Communicat. Technol.*, (2018), pp. 1–5.

Yang et al., 2024 J. Yang,W. Zhu,G. Liu,W. Dai, Z. Xu, L. Wan, G. Zhou, ICPNet: Advanced Maize Leaf Disease Detection with Multidimensional Attention and Coordinate Depthwise Convolution. Plants,13 (2024),2277.

Zhang et al., 2015 Z. Y. Zhang, X. Y. He, X. H. Sun, L. M. Guo, J. H., Wang, F. S. Wang, Image recognition of maize leaf disease based on GA-SVM. Chemical Engineering Transactions, **46** (2015), 199–204.

Zhang et al.,2016 L. Zhang, K. Mistry, S. C. Neoh, C. P. Lim, Intelligent facial emotion recognition using moth-firefly optimization. *Knowledge-based Syst.*, **111** (2016), 248–267.

Zhang et al.,2018 X. Zhang, Y. Qiao, F. Meng, C. Fan, M. Zhang, Identification of maize leaf diseases using improved deep convolutional neural networks. *IEEE Access*, **6** (2018), 30370–30377.

Zhang et al., 2024 F. Zhang, R. Bao,B. Yan, M. Wang, Y. Zhang,S. Fu, LSANNet: A lightweight convolutional neural network for maize leaf disease identification. Biosystems Engineering,248 (2024),pp. 97–107.

Zhu & Gao, 2024 S. Zhu, H. Gao, MC-ShuffleNetV2: A lightweight model for maize disease recognition. Egyptian Informatics Journal , 27(2024), 100503.

Statements & Declarations

Funding

This work was supported by the Natural Science Foundation of Hunan Province, China and the Research Funds on Industry-University Cooperative Education Project of the Ministry of Education of China (Grant numbers 2023JJ30593 and 241203242051839). The authors have received research support from the above funding institutions.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

All authors contributed to the study conception and design.