frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 December 2021
doi: 10.3389/fpls.2021.809506

Citrus Huanglongbing Detection Based on Multi-Modal Feature Fusion Learning

Dongzi Yang[1,2], Fengcheng Wang[1,2], Yuqi Hu[1,2], Yubin Lan[1,2,3,4] and Xiaoling Deng[1,2,3,4]**

[1] College of Electronic Engineering, College of Artificial Intelligence, South China Agricultural University, Guangzhou, China, [2] National Center for International Collaboration Research on Precision Agricultural Aviation Pesticide Spraying Technology, Guangzhou, China, [3] Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou, China, [4] Guangdong Engineering Technology Research Center of Smart Agriculture, Guangzhou, China

Citrus Huanglongbing (HLB), also named citrus greening disease, occurs worldwide and is known as a citrus cancer without an effective treatment. The symptoms of HLB are similar to those of nutritional deficiency or other disease. The methods based on single-source information, such as RGB images or hyperspectral data, are not able to achieve great detection performance. In this study, a multi-modal feature fusion network, combining a RGB image network and hyperspectral band extraction network, was proposed to recognize HLB from four categories (HLB, suspected HLB, Zn-deficient, and healthy). Three contributions including a dimension-reduction scheme for hyperspectral data based on a soft attention mechanism, a feature fusion proposal based on a bilinear fusion method, and auxiliary classifiers to extract more useful information are introduced in this manuscript. The multi-modal feature fusion network can effectively classify the above four types of citrus leaves and is better than single-modal classifiers. In experiments, the highest accuracy of multi-modal network recognition was 97.89% when the amount of data was not very abundant (1,325 images of the four aforementioned types and 1,325 pieces of hyperspectral data), while the single-modal network with RGB images only achieved 87.98% recognition and the single-modal network using hyperspectral information only 89%. Results show that the proposed multi-modal network implementing the concept of multi-source information fusion provides a better way to detect citrus HLB and citrus deficiency.

Keywords: convolutional neural network, citrus greening disease, machine learning, multi-modal feature fusion, hyperspectral images

OPEN ACCESS

Edited by:
Hanno Scharr, Jülich Research Center, Helmholtz Association of German Research Centres (HZ), Germany

Reviewed by:
Dimitrios Fanourakis, Technological Educational Institute of Crete, Greece
Yongqiang Zheng, Southwest University, China

****Correspondence:***
Yubin Lan
ylan@scau.edu.cn
Xiaoling Deng
dengxl@scau.edu.cn

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *05 November 2021*
Accepted: *06 December 2021*
Published: *23 December 2021*

Citation:
Yang D, Wang F, Hu Y, Lan Y and Deng X (2021) Citrus Huanglongbing Detection Based on Multi-Modal Feature Fusion Learning. Front. Plant Sci. 12:809506. doi: 10.3389/fpls.2021.809506

INTRODUCTION

Citrus Huanglongbing (HLB), also called citrus greening, is commonly believed to be citrus cancer without effective treatment. The symptoms of HLB are mainly yellow shoots, yellow leaves, and red nose fruits, among others. The infected plants easily wither and die. HLB is found all over the World, and it also occurs in China, especially in the Guangdong Sihui and Guangdong Huizhou. HLB is infectious and can be spread through insect vectors or grafting. The three most effective

Abbreviations: HLB, Citrus Huanglongbing; PCA, principal component analysis; PCR, polymerase chain reaction; UVA, unmanned aerial drone.

methods to prevent HLB are planting non-toxic seedlings, preventing and controlling citrus psyllids, and removing diseased plants (Han et al., 2021). In traditional agriculture, the prevention and control of HLB relies on the observation of experts or experienced farmers to remove diseased plants as early as possible. For plants with mild symptoms, PCR (Polymerase Chain Reaction), and other biotechnological techniques can be used to accurately identify plants. This method has high accuracy and disease can be detected and eradicated in the early stages of plant infection. However, this approach relies on experts first identifying diseased plants, and then bringing the diseased plants back to the laboratory to have disease confirmed by genetic methods. This process is lengthy and dependent on those experts. If a machine is trained as an expert and replaces the expert for identification, the detection process will be significantly accelerated.

With the development of deep learning since 2015, many useful networks for special object extraction have emerged, such as CNNs,ResNet50 (He et al., 2016), VGG16 (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015), SeNet50 (Hu et al., 2020), ResNeXt101 (Szegedy et al., 2015), VGG (Simonyan and Zisserman, 2014), and Senet50 (Hu et al., 2020). They have been very successful in modeling complicated systems, owing to their ability of distinguishing patterns and extracting regularities from data. The above-mentioned networks have been effectively incorporated in plant phenotyping projects. For example, variety identification in seeds (Taheri-Garavand et al., 2021b; Plants 10, 1406) and in intact plants by using leaves (Nasiri et al., 2021; Plants 10, 1628), weed and crop classification and recognition is the frontier and trend of agricultural artificial intelligence (Deng et al., 2020; Jiang et al., 2020), detecting crop nutritional deficiencies (Baresel et al., 2017; Tao et al., 2020), and plant disease classification (Kaya et al., 2019; Karlekar and Seal, 2020). Mostly, studies learn single-source information, and classify or identify subsequent information. These kinds of networks mostly use visual image and have rather good accuracy in specific cases. However, agriculture is a complicated system in which the shooting conditions of visual images randomly change and the crops keep growing, which leads the networks reliant on visual imaging to lack universality. Several researchers have made some efforts to improve the accuracy by continuously supplementing datasets (Picon et al., 2019), yet data collecting is a very tough work in agriculture as it is restricted by the environment and the growth cycle of plants. Therefore, how to improve the precision rate under unabundant dataset is becoming increasingly more significant.

In recent years, with the rapid development of spectroscopy, some studies adopted multispectral and hyperspectral information to detect deeper information of objects, such as using infrared to evaluate the quality of strawberry by hyperspectral images (Su et al., 2021), using hyperspectral satellite remote sensing to estimate grassland yield (Ali et al., 2014), or using UVA-based hyperspectral imagery (Feng et al., 2020) for yield prediction. Compared with RGB images, hyperspectral images combined with neural network technology can more effectively identify plant diseases, even in the early stage of disease.

The internal information extracted from hyperspectral images can be used to compensate for the shortcomings of RGB images with only surface information. Hence, multi-source feature fusion can improve the predictive ability of the model. The purpose of the fusion model is to combine the strengths of different sub-models to compensate for any shortcomings (Zadeh et al., 2017). Deep multi-modal learning can reduce the design requirements for feature engineering and deep-learning architectures, and can achieve the required accuracy more simply and quickly (Atrey et al., 2010; Ramachandram and Taylor, 2017; Baltrusaitis et al., 2019). Yan et al. (2021) proposed a fusion scheme combining a multi-dimensional convolutional neural network with a visualization method for detection of aphis gossypii glover infection in cotton leaves using hyperspectral imaging, which achieved good development prospects in plant disease identification.

Numerous researchers have conducted laboratory investigations into the identification of HLB using different methods under different observation heights, such as using visual images in the laboratory with traditional machine-learning methods (Deng et al., 2016) and using UAV hyperspectral and multispectral images using deep-learning networks (Deng et al., 2019; Lan et al., 2020).

To increase the reliability and precision of HLB detection, in this study, a method is proposed that fuses two sources of information, namely, spectral and RGB images, by building a multi-modal deep-learning network to identify HLB leaves from four categories.

MATERIALS AND METHODS

Data Acquisition and Processing

The data used in this study were collected in the citrus test fruit field of South China Agricultural University, Tianhe, Guangdong Province (longitude 113.35875, latitude 23.15747). In early March, citrus trees are in the spring growth period and are grown in subtropical climate regions, shown in **Figure 1**. The variety of citrus is Shatangju (*Citrus reticulata Banco*). The selected tree samples were specially cultivated and PCR-tested, and Zinc deficiency was visually assessed by a field expert, and was confirmed by conducting mineral analysis. The data samples of this study include the leaves of HLB plants, of Zn-deficient plants, of healthy plants, and those with suspected HLB (in which case the surface of the leaf is uniformly yellow, which is different from the obvious symptoms of HLB). The four categories leaves are shown in **Figure 2**.

The collection environment is shown in **Figure 3**. The RGB images were taken with Sony cameras and under natural light, ensuring that the required foliage was clear, independent of the shooting location, and free of background interference. The distance between camera and leaves was controlled with 20–50 cm. The hyperspectral data of leaves were collected by a hyperspectral imager (Hypersis-VNIR-PFH, Zhuoli Hanguang, Beijing, China). The spectral range was 300 nm to 1070 nm and the exposure time for each collection was 30 ms. The running speed of the mobile platform was 5.0375 mm/s, the scanning

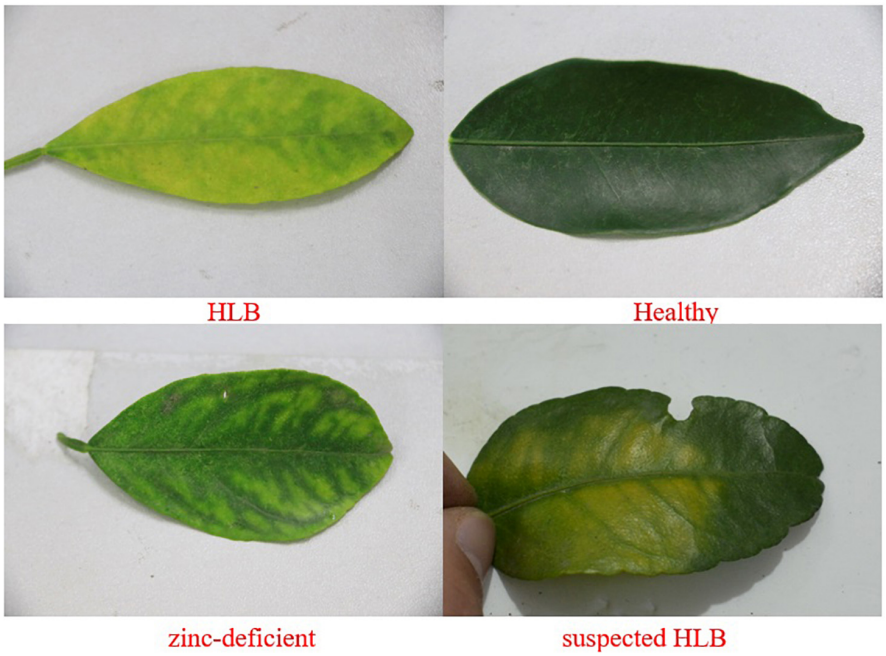

FIGURE 1 | Dataset collection location.

FIGURE 2 | Four categories leaves.

distance 120 mm, and the hyperspectral image size 100 × 200 pixels. Spectral data analysis and processing were implemented in ENVI 5.3 software (Harris Geospatial Solutions, Inc., Broomfield, CO, United States).

Figure 4 shows the method of feature area selection during the data processing step. In the process of hyperspectral image analysis, the upper, middle, and lower regions of interest of the leaf blade were chosen as the feature region, the average reflectance in the region of interest calculated, and the average reflectance used to represent the area. Finally, the hyperspectral image was converted into a hyperspectral band, and the average reflectance used to reflect the area. The frequency band of each area ranged from 300 nm to 1070 nm, removing the incomplete information about the start and the tail, leaving 768 bands in the middle. Owing to the similarity of adjacent bands of hyperspectral images, to reduce similar repetitive features, every three adjacent bands in the 300–1070-nm range were extracted and combined into a new band. After final extraction, 256 composite bands remained.

FIGURE 3 | Data collection equipment. **(A)** RGB image capture equipment. **(B)** Hyperspectral image capture equipment.

Table 1 shows the one-to-one correspondence dataset between images and spectral data. Each RGB image corresponds to a spectral sample and each piece of spectral data contains the spectral information of the upper, middle, and lower regions of the leaf.

Multi-Modal Network Architecture

The multi-modal network proposed in this study consists of two backbone networks. The architecture was divided into four parts. The first is an image feature extraction network that extracts surface features of RGB images. The second is a hyperspectral band feature extraction network that extracts the HLB feature bands. The third is a feature fusion part that fuses the two features extracted from two different networks and performs classification with an auxiliary classifier. The fourth part is classification using auxiliary classifiers. The multi-modal network structure is shown in **Figure 5**.

RGB Image Extraction Network

In the first part of RGB image feature extraction, ResNet50, VGG16, and ResNeXt101 were selected as the candidates for the backbone network. After experimental comparison, ResNet50 was adopted because it works well and in wide use. In terms of the network structure, ResNet50 has fewer parameters, but the effect achieved is similar to that of ResNeXt101. The image in this experiment is high definition, and the amount of calculation required for the extraction of the hyperspectral band is also large. To reduce the amount of calculation and not lose too much accuracy, ResNet50 was chosen. The results of the experiment are detailed further below at **Table 2**. To enrich the diversity of samples, a data enhancement module was added to the network. During the training process, there was a 10% probability that the RGB image would be randomly rotated forward or counterclockwise by 45°. The feature dimensionality extracted from the backbone network was 2048. To reduce the dimensionality obtained by feature fusion and reduce the amount of network calculation, the fully connected layer was used for feature dimensionality reduction, and the final image feature dimensionality obtained was 256.

Feature Extraction Network for Hyperspectral Band

The second part of the multi-modal network is to extract feature band information of hyperspectral data. There are many common spectral feature band extraction methods, such as support vector machines and PCA (Principal Component Analysis), among others (Velasco-Forero and Angulo, 2013; Deng et al., 2014; Medjahed et al., 2015; Pérez et al., 2016). In this study, a simple neural network for feature extraction among the 300–1070-nm hyperspectral data is proposed, and an attention module was added in this hyperspectral feature band exaction network to increase the ability of extracting bands. After combining the three adjacent bands into one channel, the number of bands decreased from 758 to 256, which reduced the overall amount of calculation and number of parameters of the hyperspectral feature extraction network. Hyperspectral band information is one-dimensional (1D) information. Commonly used image neural networks are not suitable for 1D information extraction, and we only needed to extract the bands with large differences. Therefore, the designed neural network must be capable of 1D information extraction. Moreover, it must be able to find the bands with large differences and retain the characteristic of this large difference. As shown in **Figure 4**, the upper, middle, and lower parts of each hyperspectral image were selected, and the hyperspectral band of each hyperspectral image calculated by averaging each part of the sample. Thus, there were three pieces of hyperspectral 1D data for each channel of the hyperspectral image. Therefore, the input of the hyperspectral

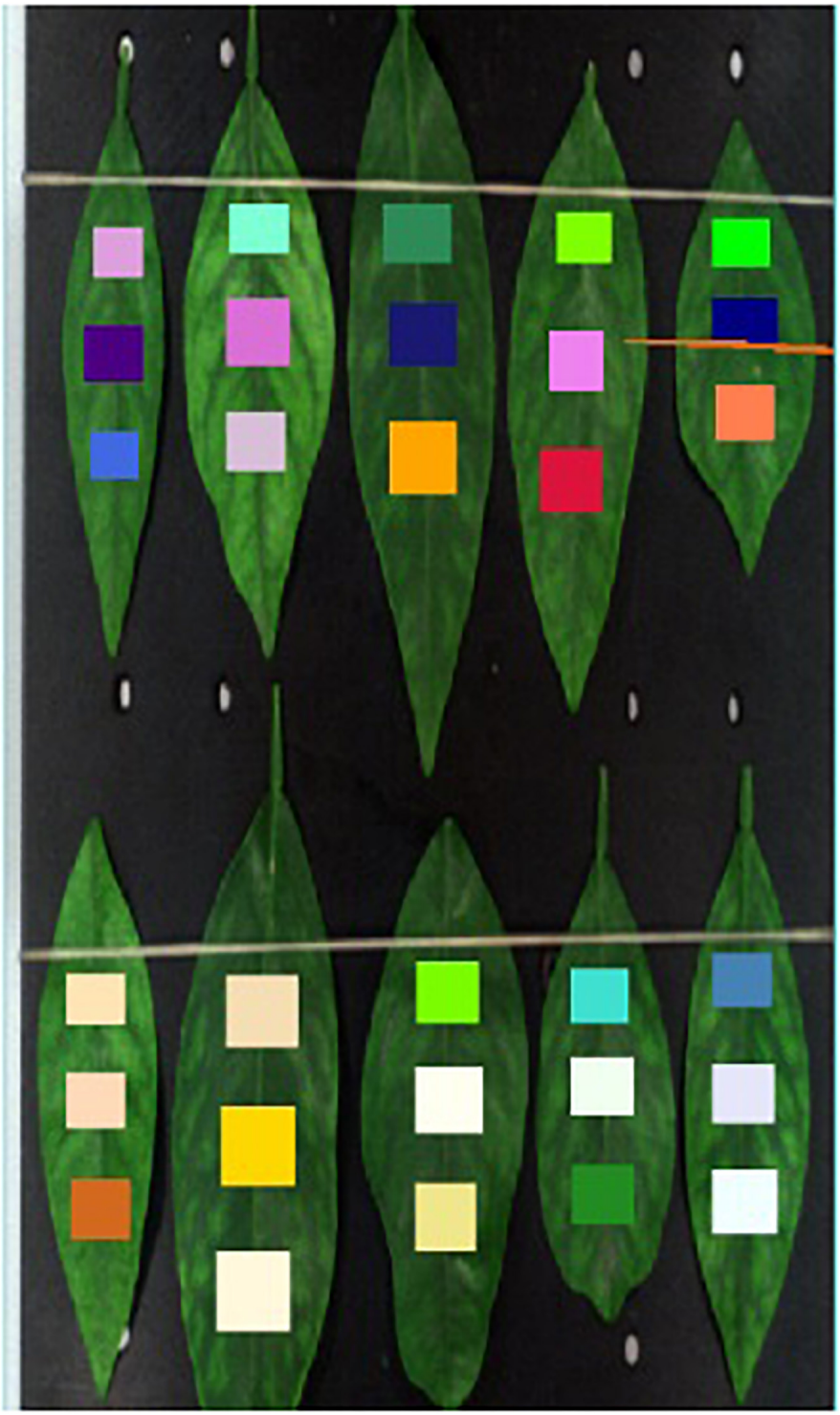

FIGURE 4 | Feature area selection during processing in Envi software.

band feature extraction network was 256 × 3. Even so, a significant amount of redundant information remains. To reduce the influence of this redundant information on the final classification results, a soft attention mechanism was adopted in the module to further extract the hyperspectral information of input data. Finally, the output size of the network was 1 × 256. The structure of the attention algorithm is shown in **Figure 6**.

Multi-Modal Feature Fusion

Typical fusion methods mainly comprise early and late fusion. As the name suggests, early fusion is used to fuse features at feature levels, using operations such as concatenation and addition of different features (Chaib et al., 2017), and then inputting the fused features into a model for training. Late fusion refers to fusion on the score level. Methods such as a feature pyramid network (Pan et al., 2019) train multiple models, and each model will have a prediction score. The results of all models are fused to obtain the final prediction results. In this study, the 1D hyperspectral band information and 3D RGB picture information were fused before detection. ResNet50 and a hyperspectral band feature extraction network (spectrum) were used in the present work as the fusion network to carry out three different feature fusions, all of which are examples of early fusion. These three methods are feature addition, feature multiplication, and feature bilinear fusion. From **Figure 7** shows that the accuracy of addition is 94.58%, that of multiplication is 93.85%, and that of bilinear fusion is 95.1%. It can also be seen from **Figure 7** that the fitting speed of bilinear fusion was also faster than that of the other two methods.

The bilinear fusion method (Yu et al., 2018) was adopted to fuse the features between different networks. The original bilinear fusion is shown in Eqs. (1) and (2). The two input modes are **X** and **Y**, and the bilinear fusion can thus be expressed as:

$$Z_i = X^T W_i Y, \tag{1}$$

where, **W** is the projection matrix and **Z** the output of the bilinear model. **W** is decomposed into two low-rank **U** and **V** matrices, with ° indicating a matrix dot product:

$$Z_i = X^T U_i V_i^T Y = U_i^T X^{\circ} V_I^T Y. \tag{2}$$

The specific fusion formula is shown in Equation (3), where Z ($Features_{Mix}$) represents the fusion features, I ($Features_{Image}$) the features extracted by the image network, and B ($Features_{Spectrum}$) the features extracted by the spectral network. A is an N × N matrix and Bias an N × 1 matrix; in the experiments detailed herein, $N = 256$.

$$Z\ (Features_{mix}) = I\ (Features_{Image})\ A\ B\ (Features_{band}) + Bias \tag{3}$$

TABLE 1 | Four different types of data and amounts of each.

Species	Number of images	Number of spectral images
Healthy	300	300
HLB	375	375
Zn-deficient	350	350
HLB suspected	300	300

HLB, Citrus Huanglongbing.

Auxiliary Classifier

After feature fusion, the samples were modeled using auxiliary classifier based on the fused feature values. The final classification effect of the network is affected by the two backbone feature extraction networks. To improve the feature extraction effects of the RGB image feature extraction network and hyperspectral band feature extraction network, the auxiliary classifiers were modified as shown in **Figure 8**, where the loss of the overall network consists of the loss of the fusion feature classifier and one of each backbone network classifier. The specific loss calculation formula is as in Equation (2), where *Total Loss* represents the overall loss value of the network, $Loss_{mix}$ the loss value of the fusion feature classifier, $Aux\ Loss_1$ the loss value of the image auxiliary classifier, and $Aux\ Loss_2$ the spectral auxiliary classifier The loss values of μ_1 and μ_2 are the auxiliary classifier loss

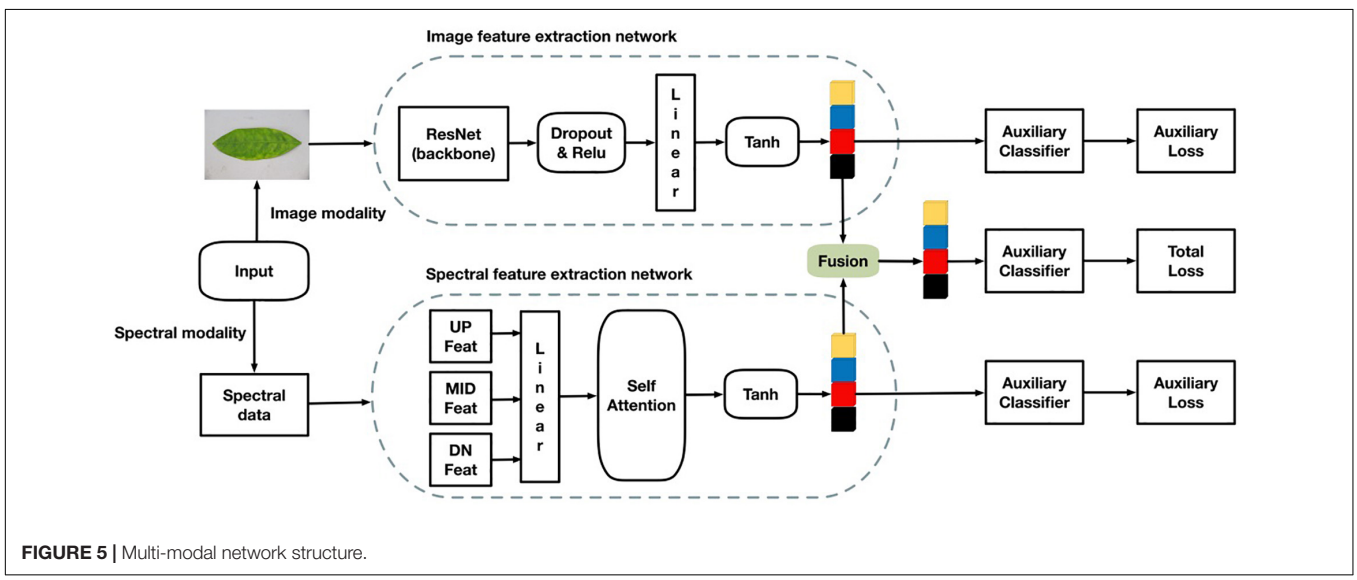

FIGURE 5 | Multi-modal network structure.

TABLE 2 | Single-network classification and multi-modal network classification accuracy.

Sample	Model	Accuracy (%)
RGB image	ResNet50	85
	VGG16	84.51
	ResNeXt101	87.98
hyperspectral data	Hyperspectral feature extraction network	89
RGB image +	Multi-modal network M1	96
hyperspectral data	Multi-modal network M2	95.1
	Multi-modal network M3	97.89

**M1, ResNet50+hyperspectral feature extraction network; M2, VGG16+hyperspectral feature extraction network; M3, ResNeXt101+hyperspectral feature extraction network.*

weight coefficients ($0 \leq \mu_1 < 1, 0 \leq \mu_2 < 1$). By testing different groups of weight coefficient values, it was found that the best classification effect is obtained when the coefficient $\mu_1 = 0.25$ and the coefficient $\mu_2 = 0.20$.

$$Total\ Loss = Loss_{mix} + \mu_1 \times Aux\ Loss_1 + \mu_2 \times Aux\ Loss_2 \quad (4)$$

RESULTS

The experimental hardware environment of this study is listed in **Table 3**. The software environment was set as the following: python, Ubuntu 16.04, CUDA, CUDNN, and OpenCV. In this study, **F1** score and accuracy were used to evaluate the trained model. The formulas are given in Equations (3)–(6), where ***P*** is the precision rate, ***R*** the recall rate, ***TP*** the number of true positive samples, ***FP*** the number of false positive samples, ***FN*** the number of false negative samples, ***true_num*** the number of samples that are classified correctly, and ***total_num*** the total number of tests and total number of samples.

$$P = P = \frac{TP}{TP + FP} \quad (5)$$

$$R = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = \frac{2P * R}{P + R} \quad (7)$$

$$Accuracy = \frac{true\_num}{total\_num} \quad (8)$$

Experimental Results

The experimental comparison results between single-network and multi-modal network classification are shown in **Table 2**. The recognition accuracies of the single network using RGB images were 85, 84.51, and 87.98% based on ResNet50, VGG16, and ResNeXt101, respectively. The recognition accuracy of the hyperspectral data dimensionality reduction network based on the soft attention mechanism was 89%, while that of the multi-modal networks designated M1 (ResNet50+hyperspectral feature extraction network), that designated M2 (VGG16+hyperspectral feature extraction network), and that designated M3 (ResNext101 +hyperspectral feature extraction network) reached 96, 95.1, and 98% respectively, all significantly higher than that of a single network. Compared with the F1 score of 85% using only the image network and that of 89% using only the spectral network, increases of 13 and 9%, respectively, were found. It can be clearly seen that feature fusion based on the bilinear fusion method and the multi-modal network of the auxiliary classifier can extract more useful information, and can better classify items with similar features.

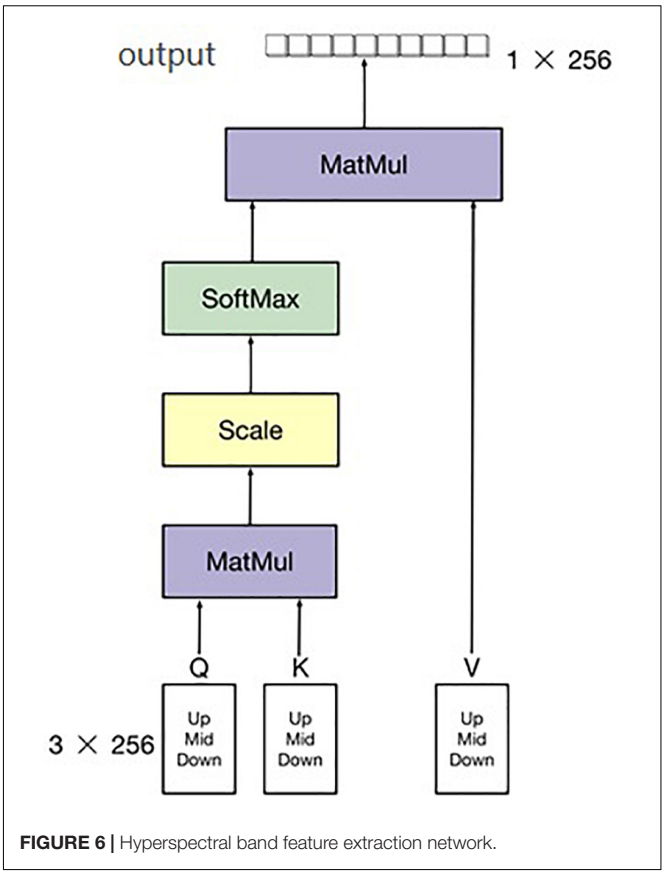

FIGURE 6 | Hyperspectral band feature extraction network.

To verify the performance of the multi-modal networks, ResNet50 with medium recognition accuracy (**Table 2**) was selected as the basic network to better reflect the improvement of recognition accuracy of multi-modal networks. **Table 4** shows the detection performance of each category based on multi-modal network M1, where the F1 scores of HLB, healthy, Zn-deficient, and suspected HLB-diseased leaves reached 95, 98, 96, and 94%, respectively, showing that average recognition accuracy reached over 95%.

Visualization Analysis of Models

Figure 9 shows the change of loss and accuracy with epoch during the training process of each network. It can be seen

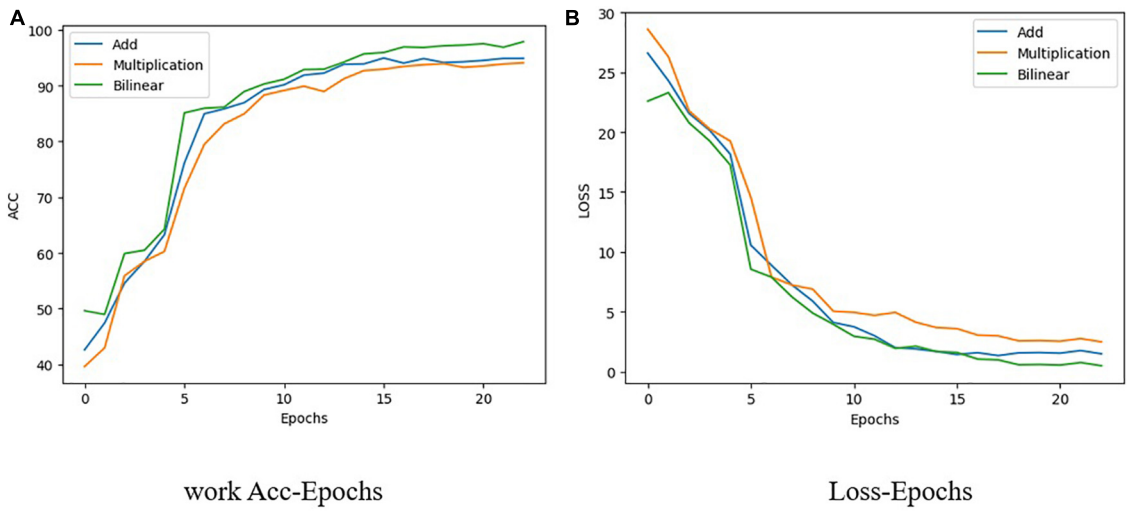

FIGURE 7 | Change with epoch of loss and accuracy of three feature fusion methods used in present. **(A)** Work Acc-Epochs. **(B)** Loss-Epochs.

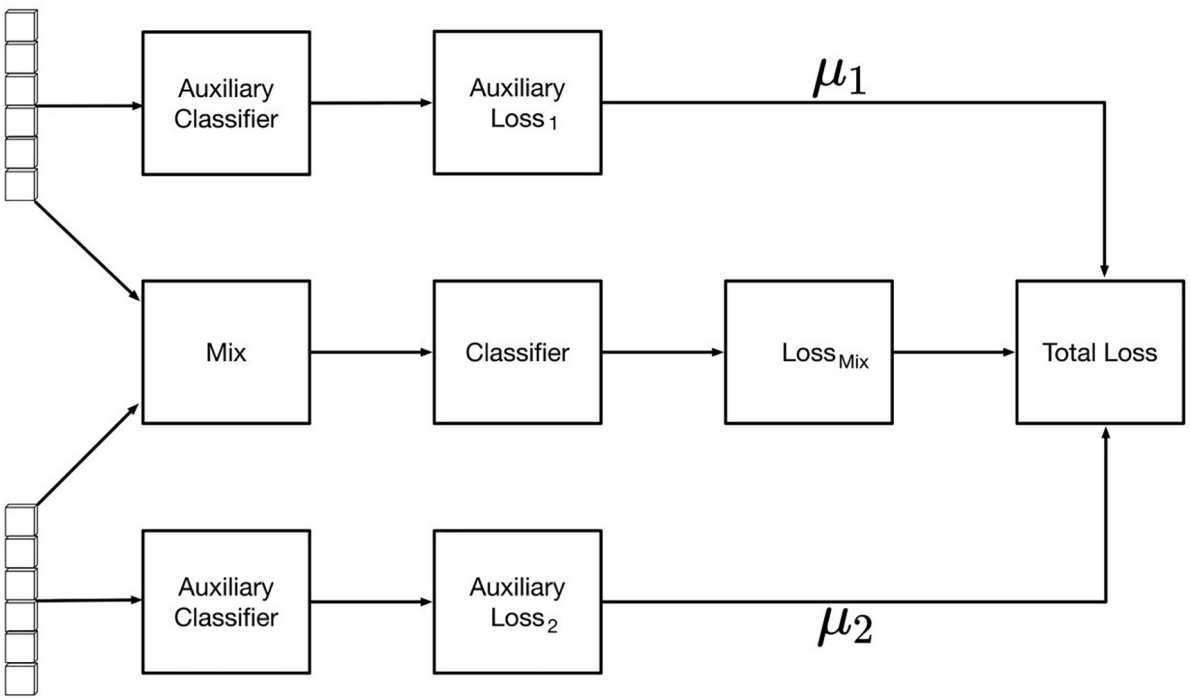

FIGURE 8 | Loss calculation method based on auxiliary and mixture loss.

from **Figure 9** that with increasing epoch loss, the fitting effect of the multi-modal model is obviously better than that of the RGB image network, and both tend to stabilize after 20 epochs. Compared with single-modal networks, including the spectrum network and RGB image networks using VGG16, ResNet50, and ResNeXt101, the three multi-modal networks achieved significantly better performance with faster convergence (as shown in **Figure 9A**) and higher accuracy (as shown in **Figure 9B**).

Figure 10 shows the confusion matrix of the three models. **Figures 10B,C** is the confusion matrix of the RGB image network and the hyperspectral network. It can be seen that the classification effects of the RGB image network and the hyperspectral image network have complementary aspects, especially for zinc deficiency. Classification of symptoms and HLB symptoms. **Figure 10A** is the effect of the final multi-modal network. It can be seen that the final confusion matrix has achieved a good effect.

DISCUSSION

Most existing networks can significantly improve the recognition accuracy by increasing the depth of the network, the dimensionality of the network, and the size of the data set. Such as ResNet, from ResNet50 to ResNet101, its recognition accuracy is improved, but the recognition speed and calculation amount are increased. When more than 101 layers are added, the recognition accuracy is not improved. This shows that although only increasing network depth can increase the accuracy, the cost is too high. The GoogleNet is to increase the width of each layer without increasing the depth of the network, but this improvement is also limited. Besides, dataset is difficult work in agriculture as it is restricted by the environment and the growth cycle of plants. Multi-modal networks can expand the data dimension through network fusion and fusion of features extracted from different data. Under the condition of insufficient data for a deep-learning network, it is relatively simple to combine other sources of information to improve the accuracy of the network from a horizontal perspective rather than a vertical perspective.

In the present study, the four testing categories discussed have similar symptoms, and are difficult to discriminate only by visual imaging. Hyperspectral data can reflect the internal information of plants to a certain extent, such as chlorophyll or element content, and can make up for the lack of RGB imagery and solve the discrimination problem resulting from the similar appearance of leaves.

Regarding the multi-feature fusion part, fusion weight coefficients were introduced to the weigh the output result,

TABLE 3 | Experimental environment.

Hardware	Brand	Number
CPU	I7–10700	1
Storage	Kingston, 16 GB	2
Graphics card	GeForce GTX3070	1
Hard disk	West Statistics, 1 TB	1
Main board	Dell Precision 3640 tower	1

TABLE 4 | Four classification results of multi-modal network M1.

Type	Precision (%)	Recall (%)	F1 score (%)
HLB	96	94	95
Health	98	99	98
Zn-deficient	97	94	96
HLB suspected	92	96	94

**M1, ResNet50+ hyperspectral feature extraction network. HLB, Citrus Huanglongbing.*

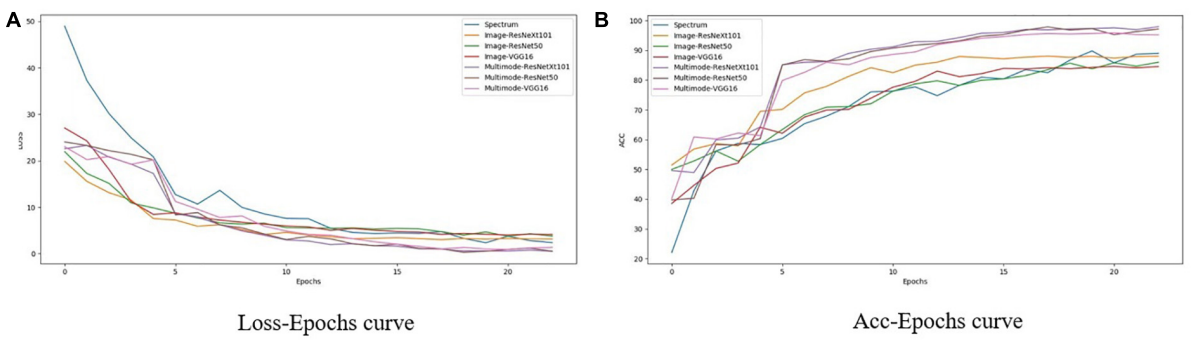

FIGURE 9 | Change with epoch of loss and accuracy of different networks in training process. **(A)** Loss-Epochs curve. **(B)** Acc-Epochs curve.

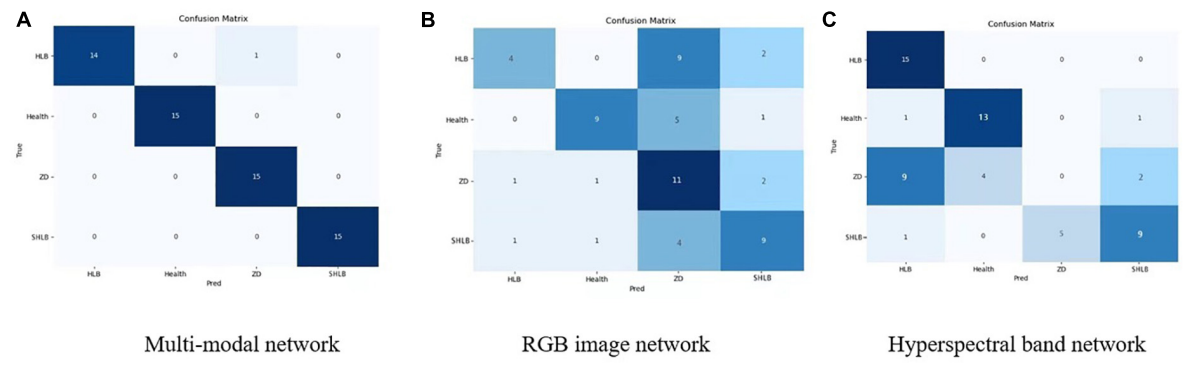

FIGURE 10 | Confusion matrix of the three models. **(A)** Multi-modal network. **(B)** RGB image network. **(C)** Hyperspectral band network.

thus improving the fitting effect of the proposed multi-modal network. The image recognition accuracy of the multi-modal model can be even improved by adding more dimensional information or improving the performance of the constituent network. The proposed method can also be applied to other agricultural applications, such as pest and disease detection with similar symptoms or appearances.

On a commercial scale, evidently, a capital investment is initially required for adopting the employed approach (Taheri-Garavand et al., 2021a Industrial Crop Prod 171, 113985). Nevertheless, the wide-ranging large-scale commercial applications can provide high returns through considerable improvements in process enhancement and cost reduction. Spectroscopy is a high-cost and high-tech imaging device, and its application areas are still being developed. However, through the research in this article, it can further expand its application fields and improve its technology. Through the neural network fusion method and the combination of RGB images, the recognition and classification of agricultural pests or agricultural diseases are enhanced.

CONCLUSION

A multi-modal network for citrus HLB detection and a bilinear fusion method based on RGB images and hyperspectral information are proposed in this study. Four HLB types with similar symptoms of leaves (HLB, suspected HLB, Zn-deficient, and healthy) were tested experimentally to verify the effectiveness of the multi-modal network. Results show that the F1-score of HLB detection based on multi-modal network reached 95%, that of healthy leaves reached 98%, that of Zn-deficient leaves reached 96 %, and that of suspected HLB diseased leaves reached 94%. The image recognition accuracy of the multi-modal model can effectively improve the recognition accuracy of the model when the size of the dataset is limited.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

DY conceptualized the experiment, selected the algorithms, collected and analyzed the data, and wrote the manuscript. FW and YH trained the algorithms, collected and analyzed data, and wrote the manuscript. XD and YL supervised the project. All authors discussed and revised the manuscript.

FUNDING

This work was supported by Key-Area Research and Development Program of Guangdong Province (Grant No. 2019B020214003), Key-Area Research and Development Program of Guangzhou (Grant No. 202103000090), Key-Areas of Artificial Intelligence in General Colleges and Universities of Guangdong Province (Grant No. 2019KZDZX1012), Laboratory of Lingnan Modern Agriculture Project (Grant No. NT2021009), National Natural Science Foundation of China (Grant No. 61675003), National Natural Science Foundation of China (Grant No. 61906074), and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515011276).

ACKNOWLEDGMENTS

DY thank FW for supporting and helping, XD and YL for supervision, and YH for revising the article.

REFERENCES

Ali, I. C. F., Green, S., and Dwyer, N. (2014). "Application of statistical and machine learning models for grassland yield estimation based on a hypertemporal satellite remote sensing time series," in *Proceedings of The 2014 IEEE Geoscience and Remote Sensing Symposium*, Quebec City, QC, 5060–5063.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimed. Syst.* 16, 345–379. doi: 10.1007/s00530-010-0182-0

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 423–443. doi: 10.1109/tpami.2018.2798607

Baresel, J. P., Rischbeck, P., Hu, Y., Kipp, S., Hu, Y., Barmeier, G., et al. (2017). Use of a digital camera as alternative method for non-destructive detection of the leaf chlorophyll content and the nitrogen nutrition status in wheat. *Comput. Electron. Agric.* 140, 25–33. doi: 10.1016/j.compag.2017.05.032

Chaib, S., Liu, H., Gu, Y., and Yao, H. (2017). Deep feature fusion for VHR remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* 55, 4775–4784. doi: 10.1109/tgrs.2017.2700322

Deng, R., Jiang, Y., Tao, M., Huang, X., Bangura, K., Liu, C., et al. (2020). Deep learning-based automatic detection of productive tillers in rice. *Comput. Electron. Agric.* 177:105703. doi: 10.1016/j.compag.2020.105703

Deng, X., Huang, Z., Zheng, Z., Lan, Y., and Dai, F. (2019). Field detection and classification of citrus Huanglongbing based on hyperspectral reflectance. *Comput. Electron. Agric.* 167:105006. doi: 10.1016/j.compag.2019.105006

Deng, X., Lan, Y., Hong, T., and Chen, J. (2016). Citrus greening detection using visible spectrum imaging and C-SVC. *Comput. Electron. Agric.* 130, 177–183. doi: 10.1016/j.compag.2016.09.005

Deng, X.-L., Li, Z., Deng, X.-L., and Hong, T.-S. (2014). Citrus disease recognition based on weighted scalable vocabulary tree. *Precis. Agric.* 15, 321–330. doi: 10.1007/s11119-013-9329-2

Feng, L., Zhang, Z., Ma, Y., Du, Q., Williams, P., Drewry, J., et al. (2020). Alfalfa yield prediction using UAV-based hyperspectral imagery and ensemble learning. *Remote Sens.* 12:2028. doi: 10.3390/rs12122028

Han, H.-Y., Cheng, S.-H., Song, Z.-Y., Ding, F., and Xu, Q. (2021). Citrus huanglongbing drug control strategy. *J. Huazhong Agr. Univ.* 40, 49–57. doi: 10.13300/j.cnki.hnlkxb.2021.01.006

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference Computing Vision Pattern Recognition (CVPR)*, Las Vegas, NV, doi: 10.1109/cvpr.2016.90

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2011–2023. doi: 10.1109/tpami.2019.2913372

Jiang, H., Zhang, C., Qiao, Y., Zhang, Z., Zhang, W., and Song, C. (2020). CNN feature based graph convolutional network for weed and crop recognition in smart farming. *Comput. Electron. Agric.* 174:105450. doi: 10.1016/j.compag.2020.105450

Karlekar, A., and Seal, A. (2020). SoyNet: soybean leaf diseases classification. *Comput. Electron. Agric.* 172:105342. doi: 10.1016/j.compag.2020.105342

Kaya, A., Keceli, A. S., Catal, C., Yalic, H. Y., Temucin, H., and Tekinerdogan, B. (2019). Analysis of transfer learning for deep neural network based plant classification models. *Comput. Electron. Agric.* 158, 20–29. doi: 10.1016/j.compag.2019.01.041

Lan, Y., Huang, Z., Deng, X., Zhu, Z., Huang, H., Zheng, Z., et al. (2020). Comparison of machine learning methods for citrus greening detection on UAV multispectral images. *Comput. Electron. Agric.* 171:105234. doi: 10.1016/j.compag.2020.105234

Medjahed, S. A., Saadi, T. A., Benyettou, A., and Ouali, M. (2015). Binary cuckoo search algorithm for band selection in hyperspectral image classification. *IAENG Int. J. Comput. Sci.* 42, 183–191.

Nasiri, A., Taheri-Garavand, A., Fanourakis, D., Zhang, Y., and Nikoloudakis, N. (2021). Automated grapevine cultivar identification via leaf imaging and deep convolutional neural networks: a proof-of-concept study employing primary iranian varieties. *Plants* 10:1628. doi: 10.3390/plants10081628

Pan, H., Chen, G., and Jiang, J. (2019). Adaptively dense feature pyramid network for object detection. *IEEE Access* 7, 81132–81144. doi: 10.1109/access.2019.2922511

Pérez, M. R. V., Mendoza, M. G. G., Elías, M. G. R., González, F. J., Contreras, H. R. N., and Servín, C. C. (2016). Raman spectroscopy an option for the early detection of citrus Huanglongbing. *Appl. Spectrosc.* 70, 829–839. doi: 10.1177/0003702816638229

Picon, A., Seitz, M., Alvarez-Gila, A., Mohnke, P., Ortiz-Barredo, A., and Echazarra, J. (2019). Crop conditional Convolutional Neural Networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions. *Comput. Electron. Agric.* 167:105093. doi: 10.1016/j.compag.2019.105093

Ramachandram, D., and Taylor, G. W. (2017). Deep multimodal learning: a survey on recent advances and trends. *IEEE Signal Process. Mag.* 34, 96–108. doi: 10.1109/msp.2017.2738401

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint] arXiv: 1409.1556, doi: 10.3390/s21082852

Su, Z., Zhang, C., Yan, T., Zhu, J., Zeng, Y., Lu, X., et al. (2021). Application of hyperspectral imaging for maturity and soluble solids content determination of strawberry with deep learning approaches. *Front. Plant Sci.* 12:736334. doi: 10.3389/fpls.2021.736334

Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference Computing Vision Pattern Recognition (CVPR)*, Boston, MA, doi: 10.1109/cvpr.2015.7298594

Taheri-Garavand, A., Nasiri, A., Fanourakis, D., Fatahi, S., and Omid, M. (2021b). Automated in situ seed variety identification via deep learning: a case study in chickpea. *Plants* 10:1406. doi: 10.3390/plants10071406

Taheri-Garavand, A., Mumivand, H., Fanourakis, D., Fatahi, S., and Taghipour, S. (2021a). An artificial neural network approach for non-invasive estimation of essential oil content and composition through considering drying processing factors: a case study in *Mentha aquatica*. *Ind. Crops Prod.* 171:113985. doi: 10.1016/j.indcrop.2021.113985

Tao, M., Ma, X., Huang, X., Liu, C., Deng, R., Liang, K., et al. (2020). Smartphone-based detection of leaf color levels in rice plants. *Comput. Electron. Agric.* 173:105431. doi: 10.1016/j.compag.2020.105431

Velasco-Forero, S., and Angulo, J. (2013). Classification of hyperspectral images by tensor modeling and additive morphological decomposition. *Pattern Recognit.* 46, 566–577. doi: 10.1016/j.patcog.2012.08.011

Yan, T., Xu, W., Lin, J., Duan, L., Gao, P., Zhang, C., et al. (2021). Combining multi-dimensional convolutional neural network (CNN) with visualization method for detection of aphis *Gossypii* glover infection in cotton leaves using hyperspectral imaging. *Front. Plant Sci.* 12:604510. doi: 10.3389/fpls.2021.604510

Yu, C., Zhao, X., Zheng, Q., Zhang, P., and You, X. (2018). "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proceedings of the European Conference Computing Vision (ECCV)*, Cham, doi: 10.1007/978-3-030-01270-0_35

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, doi: 10.18653/v1/d17-1115

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.