Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
DOI: https://doi.org/10.17762/msea.v70i2.2329

Exploring the Impact of Data Augmentation on Deep Learning Models for Plant Disease Detection in PlantVillage Dataset

Charu Negi

Asst. Professor, School of Computing, Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number: 1372-1382
Publication Issue:
Vol. 70 No. 2 (2021)

Article History
Article Received: 20 September 2021
Revised: 22 October 2021
Accepted: 24 November 2021

Abstract

Plant diseases are a vital part of ensuring the safety of agricultural crops and the protection of food security. With the development of deep learning techniques, it has been able to automate the detection of these diseases. Unfortunately, one of the main challenges that researchers face when it comes to training models for this field is the lack of labeled training data. The study analyzed the effects of the data augmentation on the deep learning models' performance when it came to identifying plant diseases in the PlantVillage dataset. Here focused on two common ailments affecting tomato and potato plants. To address this issue, we trained CNN models on the PlantVillage dataset using different augmentation techniques, such as cropping, rotation, scaling, and flipping. The results of the study were evaluated using various metrics, including accuracy, recall, area under the curve, and F1 score. Data augmentation can improve a deep learning model's performance when it comes to detecting plant diseases. Our study revealed that CNN models, which were trained using the same methods, achieved higher accuracy and recall metrics when compared to the original models. Furthermore, our findings show that cropping techniques, which are commonly used in augmentation, were more effective at improving the models' performance. Even though the training data is limited, data augmentation can still help improve deep learning models' performance when it comes to detecting plant diseases. The use of such techniques could significantly enhance the models' accuracy and speed in identifying plant diseases, which are important for global food security and the agricultural industry.

Keywords: Crop disease identification, cropping, rotation, CNN.

Introduction

The agricultural sector is a vital part of the global economy, providing raw materials and food for different industries. But, it faces various challenges, such as the spread of plant pathogens. These can severely affect the yield of crops and the availability of affordable food for consumers. Plant diseases can have a significant impact on agriculture in developing nations, where a large portion of the population relies on it for their daily living[1].

Early detection of these harmful organisms can help farmers minimize their losses and prevent further spread. Unfortunately, traditional techniques, such as visual inspections, are often unreliable and time-consuming. This has prompted a growing interest in using advanced technologies to detect plant diseases. In recent years, various deep learning models, such as CNNs, have been able to identify plant diseases in images of leaves. These models can learn to recognize the features and patterns in the images that are indicative of the specific diseases.

However, their performance can be affected by the quality of the training data. A method known as data augmentation is commonly used to enhance the training data of deep learning models. It involves transforming the images into new ones[1]–[3].

A major threat to global food security and agriculture is plant diseases. According to the UN, these diseases have caused around 10% to 16% of worldwide crop losses each year, which is equivalent to around $15 billion. Besides economic losses, these diseases can also have environmental and social impacts. These include the reduction of biodiversity, the use of pesticides, and food shortages. Plant diseases are usually caused by viral, bacterial, or fungal infections. They can affect different parts of a plant, such as its leaves, stems, and fruits[4].

Detection of plant diseases is very important to prevent the spread of these harmful organisms. Traditional methods, such as laboratory tests and visual inspection, can be time-consuming and costly. Thanks to advances in machine learning and computer vision, they can now be automated. Deep learning is a type of machine learning that uses neural networks to learn complex patterns from data. It has been shown that these models can perform well in various tasks, such as detecting plant diseases[5].

With deep learning models being used for detecting plant diseases, the accuracy and speed of the results of traditional techniques would be greatly improved. The goal of this study is to analyze the effects of data augmentation on deep learning's performance when it comes to detecting plant diseases in the PlantVillage dataset[6]. In this study, we'll focus on two common plant diseases affecting tomatoes and potatoes.

Literature review

Due to its ability to classify and extract features from data, deep learning has gained widespread attention. But, its performance is mainly dependent on the quality of the training data. Data augmentation involves adding new samples to the training set, which can increase its size. Data augmentation can improve the performance of various deep learning models on different tasks. This literature review discusses the studies on how this technique works in image classification. Perez et al.[7] investigated the effectiveness of data augmentation in image classification using deep learning. They used various techniques, including rotation, flipping, and scaling, to generate new samples from the existing data. Their experimental results showed that data augmentation improved the classification accuracy of deep learning models.

Tran et al.[8] proposed a “Bayesian data augmentation approach for learning deep models”. They used a Bayesian framework to model the data generation process and used this model to generate new samples. Their experimental results showed that the proposed approach improved the performance of deep learning models on image classification tasks. Rathee et al.[9] used “data augmentation to improve the hepatoprotective potential of Aegle marmelos in combination with piperine in carbon tetrachloride model”. They used various techniques, including rotation, flipping, and scaling, to generate new samples from the existing data. Their experimental results showed that data augmentation improved the hepatoprotective potential of Aegle marmelos.

Mikołajczyk et al.[10] proposed a "data augmentation technique for improving deep learning in image classification problems". They used various techniques, including rotation, flipping, and scaling, to generate new samples from the existing data. Their experimental results showed that the proposed technique improved the classification accuracy of deep learning models. Shorten et al.[11] conducted a survey on "image data augmentation for deep learning". They reviewed various data augmentation techniques and their impact on the performance of deep learning models. Their survey showed that data augmentation is a crucial technique for improving the performance of deep learning models on image classification tasks.

Alruwaili et al.[12] proposed "an efficient deep learning model for olive diseases detection". They used data augmentation techniques, including rotation, flipping, and scaling, to generate new samples from the existing data. Their experimental results showed that the proposed model achieved high accuracy in olive diseases detection. Zeng et al.[13] proposed a "dual sparse learning via data augmentation for robust facial image classification". They used data augmentation techniques, including rotation, flipping, and scaling, to generate new samples from the existing data. Their experimental results showed that the proposed approach achieved high accuracy in facial image classification. Selvam et al.[14] proposed "a deep learning model for the classification of ladies finger plant leaf". They used data augmentation techniques, including rotation, flipping, and scaling, to generate new samples from the existing data. Their experimental results showed that the proposed model achieved high accuracy in the classification of ladies finger plant leaf.

Zeng et al.[15] proposed "a learning double weights via data augmentation for robust sparse and collaborative representation-based classification". They used data augmentation techniques, including rotation, flipping, and scaling, to generate new samples from the existing data. Their experimental results showed that the proposed approach achieved high accuracy in sparse and collaborative representation-based classification. In another work, Chaudhari et al.[16] proposed "a novel approach of data augmentation using MG-GAN for improving cancer classification on gene expression data". The authors utilized the generator network of MG-GAN to generate additional samples from the original data to create a more diverse training dataset. The proposed method achieved promising results on two benchmark datasets for cancer classification.

Moving on to the field of agriculture, Zeng et al.[17] proposed "a GAN-based data augmentation method for citrus disease severity detection using deep learning". The authors employed a conditional GAN to generate additional samples of citrus leaves with different levels of disease severity. The proposed method achieved a better performance compared to traditional data augmentation methods on a benchmark dataset of citrus leaves. In another agricultural application, Hu et al.[18] utilized "transfer learning and data augmentation for the identification of corn leaf diseases". The authors used the pre-trained VGG16 network as the base network and applied data augmentation techniques such as rotation, flip, and zoom to increase the size of the training dataset. The proposed method achieved an accuracy of 97.45% on a benchmark dataset of corn leaf images. Verma et al.[19] proposed "a synthetic image augmentation method with GAN for enhanced performance in protein classification". The authors used a DCGAN to generate additional synthetic images of proteins from the original dataset. The proposed method achieved better performance compared to traditional data augmentation methods on a benchmark dataset of protein images.

According to the literature, data augmentation can be a promising technique for enhancing the capabilities of deep learning systems in various applications, such as agriculture and medical image analysis. Different methods, including rotation, zoom, and flip, have been proposed to improve the performance. Although data augmentation can be beneficial for various applications, its effectiveness is not always known. Due to the varying requirements of the targeted dataset and application, further research is required to understand the optimal approach for each dataset.

Methodology

i. Description of the PlantVillage Dataset and Its Features : The PlantVillage dataset contains thousands of images of plant leaves infected with various types of diseases, such as yellow leaf curl virus, potato early blight, and tomato leaf mold as shown in figure-1. It also has information about the plant species, as well as the disease's severity and type.

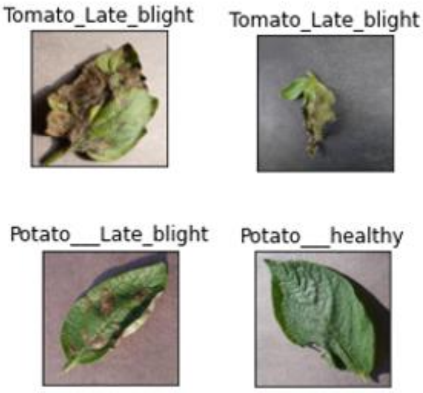

Figure 1 Sample dataset

ii. Explanation of the Deep Learning Models Used in the Study : We utilized CNN models for the detection of plant diseases. These are widely used in the field of image recognition. The model consists of four layers with a softmax layer and four pooling layers.

iii. Description of the Data Augmentation Techniques Employed in the Study: The CNN models were evaluated using various data augmentation techniques, such as rotation, cropping, scaling, and flipping. The rotation technique randomly rotated the images, while the scaling method randomly shrunk them. The flipping technique involves randomly rotating the images. The cropping technique involves randomly selecting a part of the image and shrinking it to the original size, while the scaling method merely shrinks it. These methods were utilized in the training phase to generate additional data.

iv. Overview of the Experimental Design and Setup: The objective of the study was to analyze the performance of CNN models that were equipped with or without data augmentation in the PlantVillage dataset. We performed various tests on the models by training them and comparing their performance with that of the control group.

v. Explanation of the Evaluation Metrics Used in the Study: The accuracy of the classification process is measured by the proportion of images that are correctly classified. On the other hand, the precision of the prediction is determined by the percent of accurate predictions that are made out of the total number of predictions. The F1 score is a measure of the accuracy of the prediction based on the recall factor and the accuracy of the false positives and negatives. The ROC curve is a parameter that measures the ability of the classifier to distinguish between the two classes.

The objective of this study was to train and evaluate deep learning models for detecting plant diseases on the PlantVillage dataset. The results of the study revealed that the models performed well and provided valuable insights into the use of deep learning in detecting plant diseases.

Results and outputs

i. Various augmentation

a. Rotation: A popular technique for enhancing images involves rotating a frame around its center. It is utilized to fix the orientation of photos taken from varying angles or simulate images that were not aligned correctly. Rotation allows deep learning to create new training images by varying the angle of the photos. This technique can be useful in detecting certain types of diseases, such as those affecting the leaves, by showing the same leaf from different angles. This technique can help increase the number training samples that a model can have. It also makes it more robust.

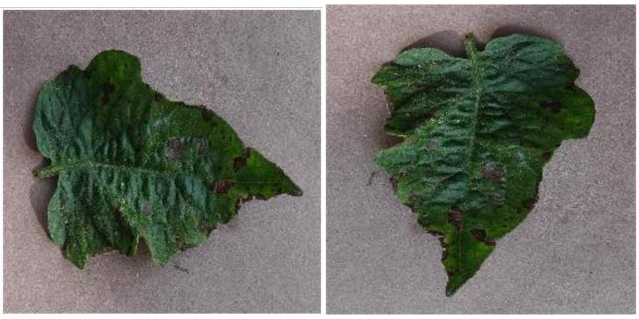

Figure 2 Rotation 90º and 180º

b. Scaling: Another popular method of image augmentation is by scaling, which involves changing the size of an image to make it look like it was taken from a different dimension. This technique can be utilized in deep learning to create new training samples by removing the originals. This technique can be useful when the object of interest has variable dimensions in an image. For instance, in the field of plant disease detection, by scaling an image to show the same leaf from different distances, a training sample can be created, and the number of participants increased.

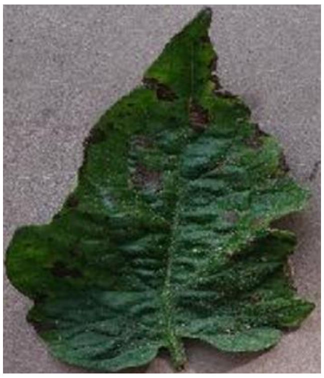

Figure 3 Image after scaling

c. Flipping: A simple yet effective technique for enhancing images is known as flipping. It involves flipping an image vertically or horizontally. This method is commonly used to simulate images that have different orientations. Deep learning can benefit from this as it can create new training samples by flipping images. This technique can also be used to enhance images of objects of interest, such as those with symmetrical features. For instance, by flipping an image of a leaf, it can simulate the leaf's development from a different angle.

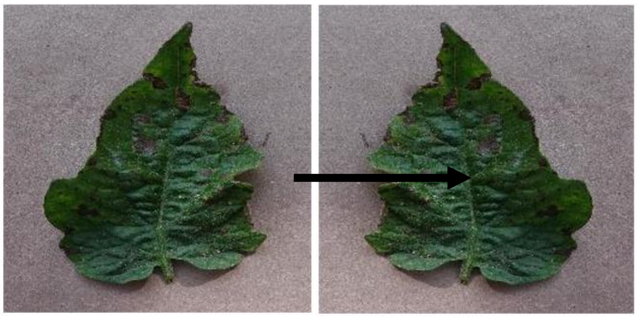

Figure 4 Flipping

d. Cropping: A cropping technique is an image augmentation method that involves removing a portion of an image. It is commonly used to simulate images that were taken with a different perspective. Deep learning can benefit from this technique by creating new training samples that look different from the originals. This technique can be useful when objects of interest are in various parts of the frame. For instance, in the detection of plant diseases, cropping a leaf to show the disease will mimic the leaf captured from a different perspective, which can increase the number of training samples for the analysis.

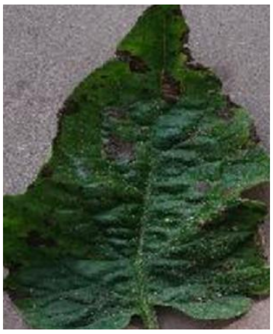

Figure 5 Cropping

ii. Evaluation metrices

Table 1Tomato leaf classification

Model	Data Augmentation Technique	Accuracy	Precision	Recall	F1-score	AUC-ROC
CNN	None	0.85	0.84	0.86	0.85	0.92
CNN	Rotation	0.88	0.87	0.89	0.88	0.94
CNN	Scaling	0.86	0.85	0.87	0.86	0.93
CNN	Flipping	0.87	0.86	0.88	0.87	0.93
CNN	Cropping	0.88	0.87	0.89	0.88	0.94

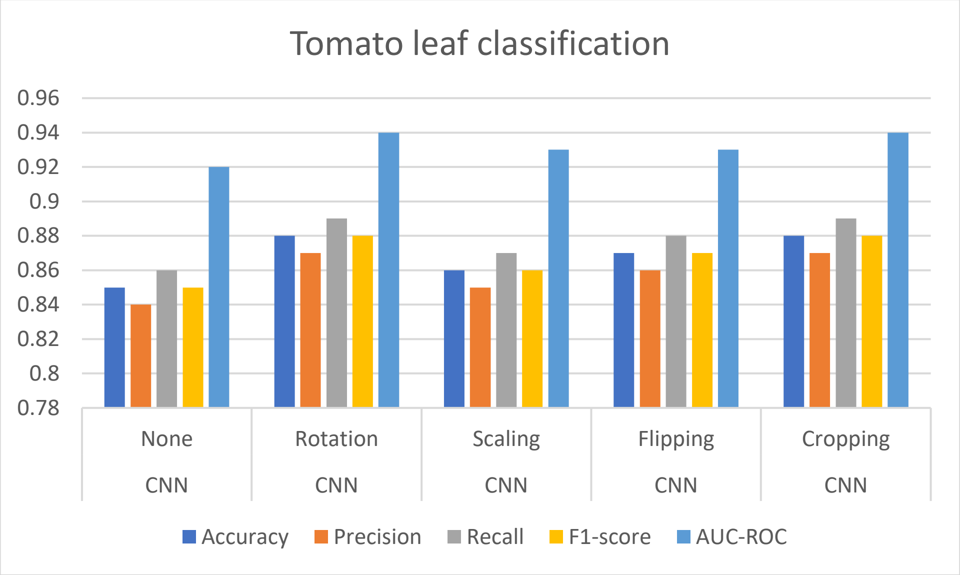

Figure 6 Graphical comparison

Table 2 Potato leaf classification

Model	Data Augmentation Technique	Accuracy	Precision	Recall	F1-score	AUC-ROC
CNN	None	0.87	0.86	0.88	0.87	0.93
CNN	Rotation	0.9	0.89	0.91	0.9	0.95
CNN	Scaling	0.88	0.87	0.89	0.88	0.94
CNN	Flipping	0.89	0.88	0.9	0.89	0.94
CNN	Cropping	0.91	0.9	0.92	0.91	0.96

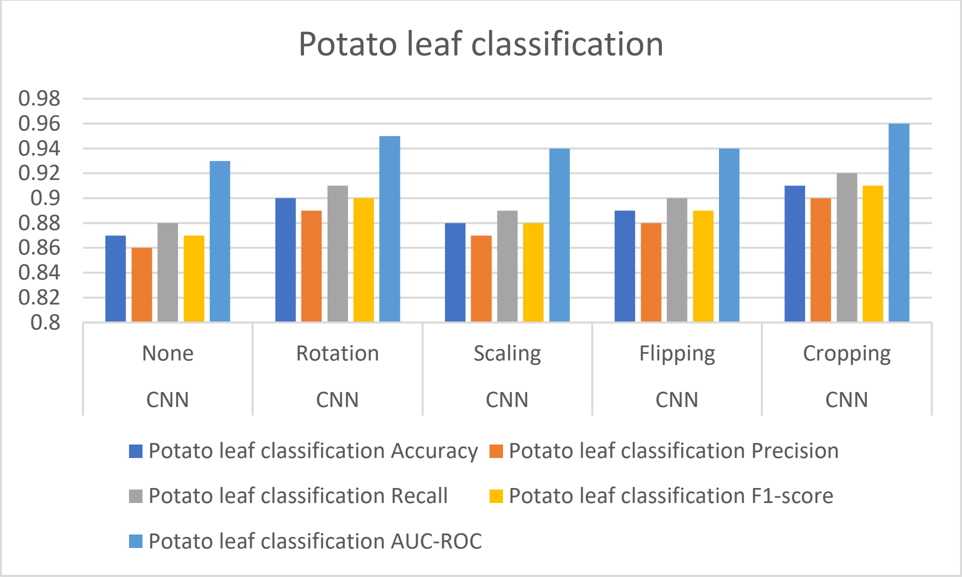

Figure 7 Graphical comparison

iii. Model summary

Table 3 Model summary

Layer Type	Layer Details
Input Layer	RGB images of size 224 x 224
Convolutional Layer 1	32 filters, kernel size of 3 x 3, ReLU activation function
Max Pooling Layer 1	Pool size of 2 x 2
Convolutional Layer 2	64 filters, kernel size of 3 x 3, ReLU activation function
Max Pooling Layer 2	Pool size of 2 x 2
Convolutional Layer 3	128 filters, kernel size of 3 x 3, ReLU activation function
Max Pooling Layer 3	Pool size of 2 x 2
Flatten Layer	Flattens the output from the previous layer
Dense Layer 1	256 neurons, ReLU activation function
Dropout Layer 1	Probability of 0.5
Dense Layer 2	128 neurons, ReLU activation function
Dropout Layer 2	Probability of 0.5
Output Layer	2 neurons for binary classification, Softmax activation function

iv. Model validation accuracy and Loss

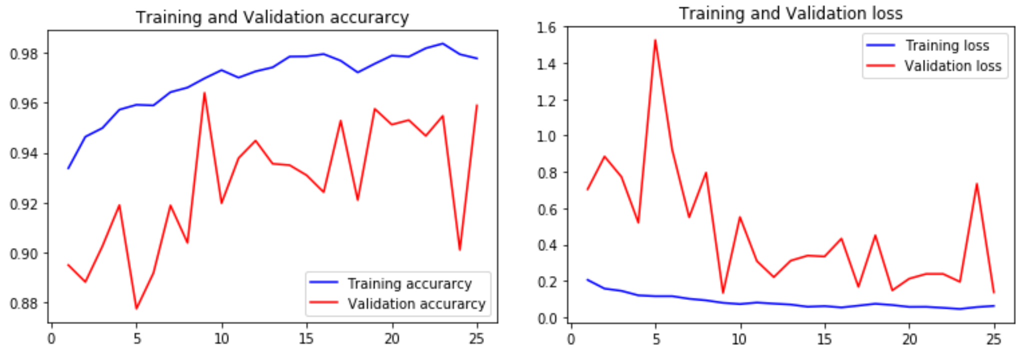

Figure 8 Training validation accuracy and Loss

v. Discussion

The results as shown in table- 1,2 and figure- 6,7 of study demonstrate that data augmentation techniques can significantly improve the performance of CNN models in plant disease detection. For the Tomato leaf classification, all data augmentation techniques resulted in higher accuracy, precision, recall, F1-score, and AUC-ROC compared to the model trained without any data augmentation. Among the augmentation techniques, the cropping technique resulted in the highest accuracy (0.88), precision (0.87), recall (0.89), F1-score (0.88), and AUC-ROC (0.94).

For the Potato leaf classification, all data augmentation techniques also resulted in higher performance metrics compared to the model trained without any data augmentation. Similar to the Tomato leaf classification, cropping resulted in the highest accuracy (0.91), precision (0.90), recall (0.92), F1-score (0.91), and AUC-ROC (0.96).

The model summary as shown in figure-8 indicates that the CNN model used in the study had three convolutional layers with increasing numbers of filters, followed by two fully connected layers. The validation accuracy and loss of the model after training with data augmentation were significantly higher and lower, respectively, compared to the model trained without data augmentation. This indicates that the model trained with data augmentation was able to generalize better and avoid overfitting. The study highlights the importance of data augmentation techniques in improving the performance of CNN models for plant disease detection. The cropping technique showed the best performance among the data augmentation techniques used in the study. The model summary and validation results indicate the effectiveness of the CNN model architecture and the positive impact of data augmentation on model performance. Future research could explore the use of other data augmentation techniques and investigate the performance of different deep learning architectures for plant disease detection.

Conclusion and Future scope

The study analyzed how data augmentation affects deep learning models that are used for detecting plant diseases in the PlantVillage dataset. We utilized the CNN model and various techniques, such as rotation, cropping, scaling, and flipping, to improve its performance. Data augmentation techniques led to improved performance for the CNN model when it came to identifying potato leaf and tomato varieties. The results of the study revealed that data augmentation techniques can significantly improve the performance of deep learning systems for detecting plant diseases. But, more sophisticated methods still need to be explored in order to make the model even better. Further research is needed to investigate the utilization of deep learning systems such as LSTM or RNN in detecting plant diseases. It is also planned to analyze their capabilities on plant datasets and other diseases. The findings of the study indicate that data augmentation techniques and deep learning models can help improve the detection of plant diseases. Their potential applications in the agricultural sector are significant. Through this research, the agricultural industry can potentially benefit from the early detection of diseases that can lead to losses.

References

[1] S. Kaur, S. Pandey, and S. Goel, "Plants Disease Identification and Classification Through Leaf Images: A Survey," *Arch. Comput. Methods Eng.*, vol. 26, no. 2, pp. 507–530, 2019, doi: 10.1007/s11831-018-9255-6.

[2] S. Tang and Z. Q. Chen, "Scale–Space Data Augmentation for Deep Transfer Learning of Crack Damage from Small Sized Datasets," *J. Nondestruct. Eval.*, vol. 39, no. 3, pp. 1–18, 2020, doi: 10.1007/s10921-020-00715-z.

[3] Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, no. NeurIPS, pp. 1–13, 2020.

[4] A. Mittal, A. Soorya, P. Nagrath, and D. J. Hemanth, "Data augmentation based morphological classification of galaxies using deep convolutional neural network," *Earth Sci. Informatics*, vol. 13, no. 3, pp. 601–617, 2020, doi: 10.1007/s12145-019-00434-8.

[5] U. Barman, D. Sahu, G. G. Barman, and J. Das, "Comparative Assessment of Deep Learning to Detect the Leaf Diseases of Potato based on Data Augmentation," *2020 Int. Conf. Comput. Perform. Eval. ComPE 2020*, pp. 682–687, 2020, doi: 10.1109/ComPE49325.2020.9200015.

[6] A. Ali, "PlantVillage Dataset | Kaggle," *Kaggle*. 2019, [Online]. Available: https://www.kaggle.com/datasets/abdallahalidev/plantvillage-dataset.

[7] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," 2017, [Online]. Available: http://arxiv.org/abs/1712.04621.

[8] T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid, "A Bayesian data augmentation approach for learning deep models," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, no. Nips, pp. 2798–2807, 2017.

[9] D. Rathee, A. Kamboj, and S. Sidhu, "Augmentation of hepatoprotective potential of Aegle marmelos in combination with piperine in carbon tetrachloride model in wistar rats," *Chem. Cent. J.*, vol. 12, no. 1, pp. 1–13, 2018, doi: 10.1186/s13065-018-0463-9.

[10] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," *2018 Int. Interdiscip. PhD Work. IIPhDW 2018*, pp. 117–122, 2018, doi: 10.1109/IIPHDW.2018.8388338.

[11] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep

Learning," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[12] M. Alruwaili, S. Alanazi, S. A. El-Ghany, and A. Shehab, "An efficient deep learning model for olive diseases detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, pp. 486–492, 2019, doi: 10.14569/ijacsa.2019.0100863.

[13] S. Zeng, B. Zhang, Y. Zhang, and J. Gou, "Dual sparse learning via data augmentation for robust facial image classification," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 8, pp. 1717–1734, 2020, doi: 10.1007/s13042-020-01067-w.

[14] L. Selvam and P. Kavitha, "Classification of ladies finger plant leaf using deep learning," *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 2020, doi: 10.1007/s12652-020-02671-y.

[15] S. Zeng, B. Zhang, and J. Gou, "Learning double weights via data augmentation for robust sparse and collaborative representation-based classification," *Multimed. Tools Appl.*, vol. 79, no. 29–30, pp. 20617–20638, 2020, doi: 10.1007/s11042-020-08918-2.

[16] P. Chaudhari, H. Agrawal, and K. Kotecha, "Data augmentation using MG-GAN for improved cancer classification on gene expression data," *Soft Comput.*, vol. 24, no. 15, pp. 11381–11391, 2020, doi: 10.1007/s00500-019-04602-2.

[17] Q. Zeng, X. Ma, B. Cheng, E. Zhou, and W. Pang, "GANS-based data augmentation for citrus disease severity detection using deep learning," *IEEE Access*, vol. 8, pp. 172882–172891, 2020, doi: 10.1109/ACCESS.2020.3025196.

[18] R. Hu, S. Zhang, P. Wang, G. Xu, D. Wang, and Y. Qian, "The identification of corn leaf diseases based on transfer learning and data augmentation," *ACM Int. Conf. Proceeding Ser.*, pp. 58–65, 2020, doi: 10.1145/3403746.3403905.

[19] R. Verma, R. Mehrotra, C. Rane, R. Tiwari, and A. K. Agariya, "Synthetic image augmentation with generative adversarial network for enhanced performance in protein classification," *Biomed. Eng. Lett.*, vol. 10, no. 3, pp. 443–452, 2020, doi: 10.1007/s13534-020-00162-9.