APPLY K-MEANS IN OUTLIER REMOVAL TO IMPROVE RICE LEAF DISEASE CLASSIFICATION PERFORMANCE USING MOBILENETV3 MODEL

Vu Huy Luong*, Nguyen Thi Mai Phuong
TNU - University of Information and Communication Technology

ARTICLE INFO

Received: 06/10/2023
Revised: 30/10/2023
Published: 30/10/2023

KEYWORDS

Deep learning
K-means clustering
MobileNetV3
Rice leaf disease
Classification

ABSTRACT

The population is continuously increasing, urbanization is accelerating, and agricultural land is shrinking. Diseases on rice leaves cause significant yield losses, necessitating early diagnosis to mitigate their impact on productivity and ensure food security. The application of technology in detecting and diagnosing rice leaf diseases is essential. This study proposes a method to enhance the accuracy of the MobileNetV3 deep learning model. The technique involves removing duplicate and outlier images to improve the efficiency of the MobileNetV3 model using K-Means. The dataset used in the experiment is acquired from a secondary source. The data consists of 5932 images of four common diseases on rice leaves. Three sets of data are created (Set-1, Set-2 and Set-3), corresponding to the outlier thresholds set at 0.00, 0.05, and 0.06, respectively. The results show a considerable increase in accuracy when applying this method. The Top-1 accuracy rose from 80.11% to 86.85% and 87.69%, respectively.

ÁP DỤNG THUẬT TOÁN K-MEANS TRONG LOẠI BỎ DỮ LIỆU NGOẠI LỆ ĐỂ NÂNG CAO HIỆU SUẤT PHÂN LOẠI CÁC BỆNH VỀ LÁ LÚA SỬ DỤNG MÔ HÌNH MOBILENETV3

Vũ Huy Lượng*, Nguyễn Thị Mai Phương
Trường Đại học Công nghệ thông tin và Truyền thông – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 06/10/2023
Ngày hoàn thiện: 30/10/2023
Ngày đăng: 30/10/2023

TỪ KHÓA

Học sâu
Phân cụm K-means
MobileNetV3
Bệnh trên cây lúa
Phân loại

TÓM TẮT

Dân số tăng, quá trình đô thị hóa diễn ra nhanh chóng và diện tích đất nông nghiệp bị thu hẹp. Bệnh trên lá lúa được xem là nguyên nhân chính gây mất mùa, vì vậy việc chẩn đoán bệnh lá lúa sớm là quan trọng để giảm ảnh hưởng của chúng đối với năng suất, bảo đảm an ninh lương thực. Sử dụng công nghệ trong việc phát hiện và chẩn đoán các bệnh trên lá lúa là cần thiết. Nghiên cứu này đề xuất một phương pháp nhằm nâng cao độ chính xác của mô hình MobileNetV3. Phương pháp này bao gồm việc loại bỏ các hình ảnh trùng lặp và ngoại lệ sử dụng thuật toán K-Means. Tập dữ liệu sử dụng trong thí nghiệm được lấy từ một nguồn thứ cấp. Dữ liệu bao gồm 5932 hình ảnh của bốn loại bệnh phổ biến trên lá lúa. Ba tập dữ liệu khác nhau được tạo ra (Set-1, Set-2 và Set-3), tương ứng với ngưỡng bất thường 0,00, 0,05 và 0,06. Kết quả cho thấy, độ chính xác tăng đáng kể khi áp dụng phương pháp này. Độ chính xác Top-1 đã tăng từ 80,11% lên 86,85% và 87,69% tương ứng với Set-1, Set-2 và Set-3.

DOI: https://doi.org/10.34238/tnu-jst.8919

* Corresponding author. *Email: vhluong@ictu.edu.vn*

1. Introduction

Rice is the main staple food for the people of Vietnam. The population scale and urbanization rate in Vietnam have significantly increased in recent years. Data shows that Vietnam is undergoing a rapid urbanization process, with a substantial growth of the urban population compared to previous years. Urbanization has led to a narrowing of agricultural land and a shortage of labour in the agricultural sector, particularly in labour-intensive rice production [1].

Rice plant diseases have a significant impact on yields, and if not detected early, they can lead to crop failure. Currently, rice diseases are often detected based on the experience and knowledge of experts and the frequency of field visits by farmers or agricultural engineers [2]. The shortage of labour, coupled with the increasing demand for productivity and reducing crop losses, necessitates the application of scientific and technical advancements in agriculture [3].

With the support of machine learning, rice plant diseases can be detected early and more easily using images captured in the fields from recording devices or smartphones. These diagnoses have the potential to provide farmers with an opportunity to treat rice plants early, thereby minimizing the damage caused by diseases and reducing instances of reduced yields or crop failure [4].

Recent studies have applied various methods and techniques to improve the accuracy of classifying different types of diseases on rice leaves, thereby enhancing the effectiveness of disease diagnosis for rice plants. Notably, various approaches and techniques for disease classification on rice leaves are using image analysis and deep learning methods.

Mulyani et al. proposed a fast and efficient classification model for diseases on rice leaves based on leaf color and texture. Two algorithms, GLCM (Gray-Level Co-occurrence Matrix) and K-NN (K-Nearest Neighbours), were tested. The images were pre-processed with brightness enhancement before being used. The experimental accuracy achieved 89% [5].

In Cherukuri et al.'s study, rice leaf images were captured using a digital camera. The images were then pre-processed by dividing them into groups with similar features to facilitate further analysis. First, the images were converted from RGB to HSV. Then, the K-means algorithm was applied to enhance the processing of image features. The pre-processed dataset was fed into a CNN (Convolutional Neural Network) for learning. The experimental results showed an accuracy of 95% for diagnosing healthy rice leaves and 90% for Sheath Blight disease [3].

The image data of various diseases on rice leaves were sourced by the authors from the Kangle platform. The data was augmented by creating a new dataset from the original one, comprising images with different angles and sizes for each disease type. This augmented dataset was used to train a basic deep learning model. The results showed that the proposed model could diagnose with an accuracy of 80.94% [6].

In the study by Ghosal and Sarkar, and Luong et al., data was collected and resized to 224*224 pixels. Subsequently, several image augmentation techniques, such as zooming, rotation, vertical and horizontal shifts, were applied [7] – [8]. The authors used the ImageDataGenerator tool in Keras to create four new datasets from the original data and then applied transfer learning on these four datasets. The results showed that the accuracy with transfer learning was 92.46%, whereas without transfer learning, the accuracy was 74% [7]. Applying the same data pre-processing method as in the previous study, the authors created a new dataset from the original data and then used transfer learning combined with CNN to build a disease diagnosis model for rice plants. The results showed that the accuracy of the method was 96.09% in training and 94.44% in testing [9].

In the work by Zakzouk et al., the authors resized the images from 64*64 to 227*227 to match the architecture of AlexNet. The dataset consists of 16,000 images divided into three parts: 70% for training, 15% for testing, and 15% for evaluation. The results showed that the proposed model using AlexNet achieved an accuracy of 99.71% [10].

It can be observed that recent studies on classifying diseases on rice leaves for disease diagnosis have been applying deep learning techniques. The data pre-processing mainly focuses on methods such as zooming, rotation, and using tools to augment the existing dataset, but there has been little attention given to controlling repeated images or handling outliers within the dataset.

In this paper, a pre-processing data method is proposed, which involves detecting outliers and identifying duplicate images. By eliminating these images from the training set, the training process becomes more efficient. The detection of duplicate and outlier images is accomplished using the well-known K-means algorithm. After the data has been pre-processed, the MobileNetV3 deep learning model is employed to be trained on the datasets, allowing for the analysis and evaluation of the effectiveness of the proposed method.

2. Methodology

2.1. Dataset

The data in this study was acquired from secondary source. The dataset is a shared data on the Kangle. It consists of raw images of rice plants. The images of rice leaves include the background of the rice field, which helps to increase the reliability of the diagnosis in real-world scenarios. The input data was captured quickly and without any processing by any recording device available to the farmers. The data consists of 5932 images of 4 common diseases on rice leaves. The four disease classes include: Bacterial Blight (1584 samples), Blast (1440 samples), Brown Spot (1600 samples), and Tungro (1308 samples). An example of each class of the dataset is shown in Figure 1.

(a) Bacterial Blight *(b) Blast* *(c) Brown Spot* *(d) Tungro*

Figure 1. *Four leaf disease types in the dataset*

Bacterial Blight (Figure 1a) is a common and devastating disease that affects rice plants. It is caused by the bacterium Xanthomonas oryzae pv. oryzae. The disease can lead to significant yield losses.

Blast (Figure 1b) is a destructive fungal disease that affects rice plants. It is caused by the fungus Magnaporthe oryzae and is one of the most significant diseases impacting rice crops worldwide.

Brown Spot (Figure 1c) is a fungal disease that commonly affects rice plants. The disease is caused by the fungus Bipolaris oryzae.

Tungro (Figure 1d) is a viral disease that affects rice plants and is caused by two viruses: Rice tungro bacilliform virus and Rice tungro spherical virus. The disease is transmitted by insect vector.

2.2. Image Processing

After the data is acquired, it undergoes a pre-processing stage that includes duplicate and outlier detection. The algorithm used for duplicate and outlier detection is the K-means algorithm [11]. The tool used to perform duplication and outlier detection is FastDup [12].

The parameter Threshold for Similarity clusters determines whether two images belong to the same similarity cluster and is set to 0.96. The Similarity threshold is the minimal value of similarity that is meaningful for analysis. Any two images that do not reach this threshold will not be considered together in any calculations and will not be categorised as outlier data. This parameter is set to 0.9. The K Nearest neighbours is set to 2. Thus, only 2 neighbours participate in the classification process for each image in the image set. The outlier percentile parameter is

set to 0.05. The outlier threshold sets the lower percentage of similarity values that are regarded as outliers during analysis. At the default value of 0.05 (5%), images whose nearest neighbour falls within the bottom 5% of similarity values (indicating that the closest image to them is farther away than 95% of other image pairs) are considered outliers.

Figure 2. *Duplicate images in the dataset which have distance equal 1.0*

For images with a distance of 1.0 to another image, they are identified as duplicates. Duplicate images are removed, and only one image is retained for each sample. Figure 2 demonstrates duplicated images which are detected by K-means.

After removing duplicate images, the histogram depicts the number of image samples based on their distances to the cluster centre, which is represented on the Figure 3.

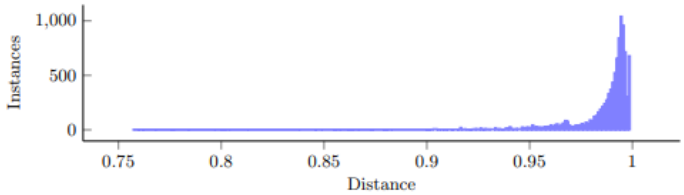

Figure 3. *The histogram chart illustrates the number of images based on the distances between images and their neighbours after removing duplicate images.*

Images with a distance less than a threshold to the dataset are considered outliers and are also removed from the dataset. The thresholds used to remove outliers are set at 0.05 and 0.06, respectively. The four dataset Figure 4 shows outlier images (threshold = 0.05) which are detected by K-means.

Figure 4. *Outlier images in the dataset where threshold is set at 0.05*

2.3. Deep Learning Model Selection

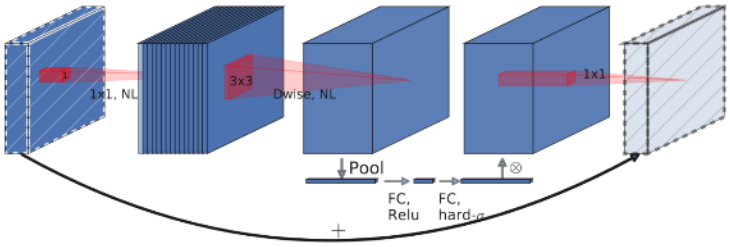

Figure 5. *Basic Block diagram of MobileNetV3 Architecture [13]*

MobileNetV3 is utilized in this study for rice leaf disease classification. MobileNetV3 is a deep learning model that employs a streamlined architecture with depth-wise separable convolutions. It was chosen for this research due to several reasons, among which MobileNetV3-

Small achieves 6.6% higher accuracy compared to a MobileNetV2 model with similar latency. Additionally, MobileNetV3-Large detection is over 25% faster at nearly the same accuracy as MobileNetV2 on COCO detection. Furthermore, MobileNetV3 can easily be embedded into handheld devices, making it suitable for diagnosing diseases on rice leaves [14].

2.4. Experiment Setup

The models were trained on a computer running the Ubuntu 20.2 operating system with the following specifications: AMD® Ryzen 7 5800x 8-core processor × 16 4.7Ghz, 64GB RAM, and a 195TB hard drive.

The four datasets include: the original data taken from the source (Set-1); data with outliers filtered at the threshold equal 0.9 (Set-2); data with outliers filtered at the threshold equal 0.8 (Set-3); and data with outliers filtered at the threshold equal 0.7 (Set-4). These datasets will be further divided into training, validating, and testing datasets. All four datasets have been processed to remove duplicate images.

Training data accounts for 70%, validating data for 10%, and testing data for 20% of the total samples for each label (Figure 6). The percentage ratios are calculated based on the samples of each label.

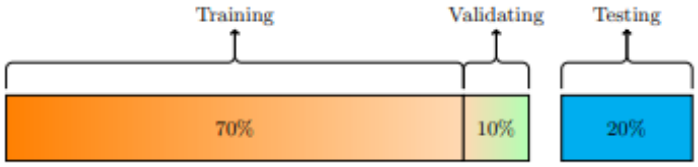

Figure 6. *Training, validating and testing data*

The size of each image is resized to 300x300 pixels before being used for training, validating, and testing. The learning rate is initialized to 0.004; alpha is initialised to 0.9; momentum is initialised to 0.9; gamma is initialised to 0.973; and the batch size is set to 5. All experiments set the maximum epoch to 450. The tool used to perform training, validating, and testing is mm pretrain [13].

3. Results

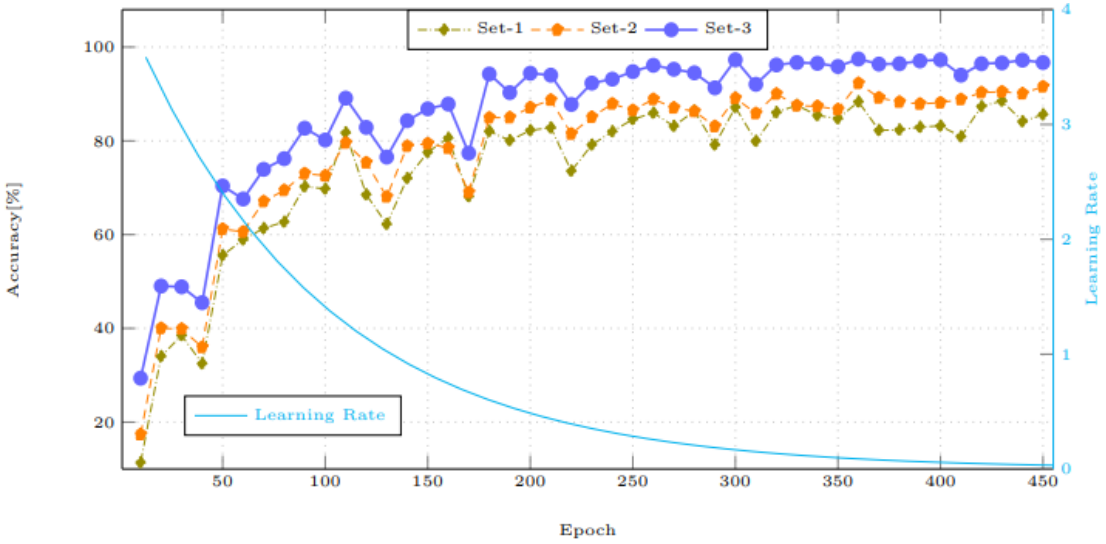

Figure 7. *Top 1 Validating Accuracy and Learning Rate Versus Epoch*

Three datasets, Set-1 (outlier threshold = 0.0), Set-2 (outlier threshold = 0.05), and Set-3 (outlier threshold = 0.06), are used to train, validate, and test the MobileNetV3 deep learning

model. The deep learning model is evaluated at each epoch using the separate validating datasets and testing datasets, distinct from the training dataset. The evaluation results at each epoch of the model are illustrated in Figure 7 for Top-1 and Figure 8 for Top-2. The test result of the model is also shown in the Table 1 for both Top-1 and Top-2.

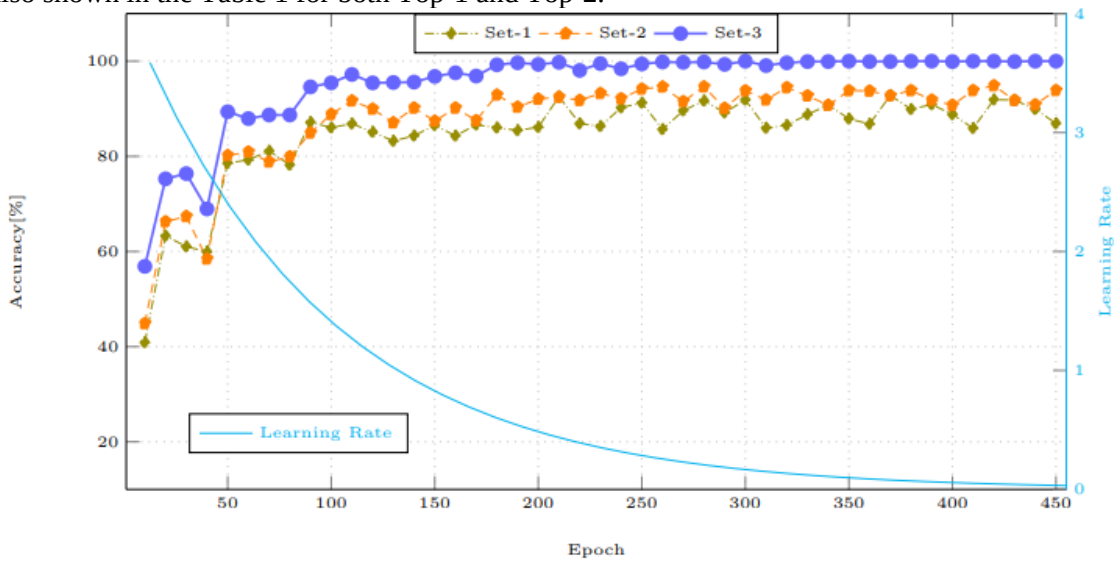

Figure 8. *Top 2 Validating Accuracy and Learning Rate Versus Epoch*

Results of the proposed model is listed in Table 1, the accuracy results for three different sets of data (Set-1, Set-2 and Set 3) are presented. The accuracy is measured using the Top-1 and Top-2 metrics, where 1 and 2 denote the rank of the correct answer among the top choices. The outlier threshold increases from Set-1 to Set-3, from 0.00 to 0.06, respectively. As the outlier threshold increases from Set-1 to Set-3, there is a noticeable improvement in both validating and testing accuracies of the models in classifying diseases on rice leaves. Higher outlier thresholds seem to lead to better model performance, suggesting that the models are more accurate when outliers are filtered. The Top-1 accuracy increases from 80.11% to 86.85% and 87.69%. There is a substantial drop in accuracy from Top-1 to Top-2, indicating that the correct answer is often among the top two predictions. Comparing validating and testing accuracy for the same outlier threshold shows how well the model generalizes to unseen data. In most cases, testing accuracy is slightly lower than validating accuracy, which is expected as the model encounters new, unseen data during testing.

Table 1. *Validating and Testing Accuracy Results*

Data	Outlier Threshold	Validating		Testing	
		Top-1[%]	Top-2[%]	Top-1[%]	Top-2[%]
Set-1	0.00	85.62	86.91	80.11	82.05
Set-2	0.05	91.58	93.87	86.85	88.27
Set-3	0.06	96.71	99.99	87.69	89.01

The Top-1 accuracy on the validating dataset shows that the model achieves its highest accuracy at epoch 300 (Figure 7), while the Top-2 chart even shows that the model's accuracy becomes stable at epoch 200 (Figure 8). Increasing the outlier threshold from 0.05 to 0.06 does not create much difference in the model's accuracy.

4. Conclusion

In this study, the performance of the MobileNetV3 deep learning model in classifying different diseases on rice leaves with various levels of removing duplicate and outlier images has been evaluated. The removal levels include: no removal, outlier threshold equal to 0.05, and outlier threshold equal to 0.06. The results demonstrate a significant improvement in the accuracy of the MobileNetV3 model when applying the proposed method. The Top-1 accuracy increases from 80.11% to 87.69%.

The findings also indicate that the deep learning models with outlier removal achieves stability with high Top-1 accuracy at epoch 300 and Top-2 accuracy at epoch 200. The highest accuracy is achieved by the dataset with an outlier threshold equal to 0.06, with Top-1 accuracy of 87.69% and Top-2 accuracy of 89.01%. However, there is not much difference in accuracy when comparing the datasets with an outlier threshold equal to 0.05 and 0.06.

The removal of outliers from the dataset needs to be further investigated to assess the impact of the outlier threshold parameter on the training speed and accuracy of the model on datasets of different sizes and different magnitudes of the outlier threshold.

Acknowledgments

This research is funded by Thai Nguyen University of Information and Communication Technology under grand number T2023-07-01.

REFERENCES

[1] J. Harris, P. H. Nguyen, L. M. Tran, and P. N. Huynh, "Nutrition transition in Vietnam: changing food supply, food prices, household expenditure, diet and nutrition outcomes," *Food Security*, vol. 12, no. 5, pp. 1141–1155, Oct. 2020.

[2] M. Mavaddat, M. Naderan, and S. E. Alavi, "Classification of rice leaf diseases using cnn-based pre-trained models and transfer learning," in *2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, 2023, pp. 1–6.

[3] N. Cherukuri, G. Kumar, O. Gandhi, V. S. K. Thotakura, D. NagaMani, and C. Z. Basha, *"*Automated classification of rice leaf disease using deep learning approach," in *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2021, pp. 1206–1210.

[4] S. P. Singh, K. Pritamdas, K. J. Devi, and S. D. Devi, "Custom convolutional neural network for detection and classification of rice plant diseases," *International Conference on Machine Learning and Data Engineering*, vol. 218, pp. 2026–2040, 2023.

[5] E. D. S. Mulyani, H. J. Pramana, L. Listiani, N. N. Febriani SM, R. A. Wiyono, and F. P. Pratiwi, *"*Classification of rice leaf diseases based on texture and leaf colour," in *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, 2022, pp. 1–6.

[6] B. D. Satoto, D. R. Anamisa, M. Yusuf, M. K. Sophan, N. Alamsyah, and B. Irmawati, "Rice disease classification based on leaf damage using deep learning," in *2022 6th International Conference on Informatics and Computational Sciences (ICICoS)*, 2022, pp. 42–47.

[7] S. Ghosal and K. Sarkar, "Rice leaf diseases classification using CNN with transfer learning," in *2020 IEEE Calcutta Conference (CALCON)*, 2020, pp. 230–236.

[8] H. L. Vu, T. M. P. Nguyen, V. N. Pham, V. S. Nguyen, and V. C. Tran, "An evaluation of eye diseases classification using ResNet on fundus image dataset collected from Thai Binh hospital," *TNU Journal of Science and Technology*, vol. 228, no. 07, pp.100-107, 2023.

[9] Rukhsar and S. K. Upadhyay, "Rice leaves disease detection and classification using transfer learning technique," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2022, pp. 2151–2156.

[10] S. Zakzouk, M. Ehab, S. Atef, R. Yousri, R. M. Tawfik, and M. S. Darweesh, "Rice leaf diseases detector based on Alexnet," in *2021 9th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*, 2021, pp. 170–174.

[11] R. Bishwas, S. Yasmin, and N. A. Turzo, "Relative comparison of k-means clustering segmented rice leaves image based nitrogen, phosphorus, and potassium nutrient deficiency classification using

convolutional neural network," in *2021 International Conference on Science and Contemporary Technologies (ICSCT)*, 2021, pp. 1–6.

[12] F. Contributor, "Fastup manage, clean and curate visual data - fast and scale," 2023. [Online]. Available: https://github.com/visual-layer/fastdup. [Accessed September 5, 2023].

[13] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," *in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019, pp. 1314-1324.

[14] M. Contributors, "Openmmlab's pre-training toolbox and benchmark," 2023. [Online]. Available: https://github.com/open-mmlab/mmpretrain. [Accessed September 5, 2023].