Two-step convolutional neural network classification of plant disease

Rosni Lumbantoruan, Nico Rajagukguk, Anju Ucok Lubis, Marwani Claudia, Humasak Simanjuntak
Department of Information System, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, Medan, Indonesia

Article Info

Article history:

Received Aug 16, 2023
Revised Jul 1, 2024
Accepted Jul 26, 2024

Keywords:

Augmentation
Convolutional neural network
Disease classification
Synthetic minority oversampling technique
Two-step convolutional neural network

ABSTRACT

Indonesia is primarily an agricultural country, with farming being the primary source of income for most of its people. Unfortunately, crop production is vulnerable to plant diseases, which are usually caused by plant pests, resulting in a reduction in both the quantity and quality of the expected harvest. In addition to the large number of classes to predict, detecting and accurately classifying each disease on different plants can be difficult. We believe that limiting the number of classes to identify may improve classification accuracy. Thus, in this research, we propose a new approach, two-step convolutional neural network (CNN), which reduces the number of classes with a two-step classification approach. To begin, we identify the number of classes that can be reduced by categorizing them into different characteristics, namely, plant type classification and plant condition classification. Second, we deal with unbalanced datasets, which can result in poor performance, if overlooked. Finally, we compare the proposed two-step CNN to baseline CNN in terms of efficiency and effectiveness. Extensive experiments show that the two-step CNN outperforms the baselines, CNN and jellyfish-residual network (JF-ResNet), increasing accuracy by 4% and 2% to 99%, respectively. In addition, we also provide a simulation evaluation to ensure that this approach is applicable.

Corresponding Author:

Rosni Lumbantoruan
Department of Information System, Faculty of Informatics and Electrical Engineering
Institut Teknologi Del
St. Sisingamangaraja, Sitoluama, Laguboti, Toba, North Sumatra, Indonesia
Email: rosni@del.ac.id

1. INTRODUCTION

Indonesia is an agrarian country, with agriculture providing the majority of the population's income. Agriculture is one of the most common methods of crop production. Plant diseases are a major risk factor in crop production which may hinder growth, reducing the quantity and quality of expected yields and even resulting in production losses. As a result, plant diseases have become a major obstacle in agriculture today. One of the causes of plant diseases comes from plant pests or pathogens. In the past few decades, farmers have been able to identify plant pests and diseases with the naked eye [1]. However, it is difficult for farmers to accurately detect and classify each disease for every plant, as plant diseases have a wide range, and most farmers lack proper knowledge about plant diseases [2].

Every leaf, whether healthy or diseased, contains characteristics that can be used to detect plant diseases [2], such as the appearance of leaves, particularly the surface. Disease classification in plants aims to group leaf images based on the specific plant diseases. Deep learning, specifically convolutional neural network (CNN) algorithms, were used in this study for their ability to automatically extract significant properties from plant leaf images and their excellent generalization capabilities. There have been successful approachesin plant diseases classification using deep learning approaches such as [3], [4]. Geetharamani and Pandian [3] proposed a 9-layer deep CNN and applied image augmentation techniques to improve the classification results. The performance of the deep CNN was compared with transfer learning algorithms and conventional algorithms. Specifically, deep CNN was compared with the transfer learning algorithms such as AlexNet [5], the very deep CNNs for large-scale image recognition (VGG16) [6], Inception-v3, residual network (ResNet), and jellyfish residual network (JF-ResNet) [7]. Meanwhile, the compared conventional algorithms were k-nearest neighbours (KNN), decision tree (DT), and support vector machine (SVM). Both comparisons demonstrated that deep CNN algorithm achieved higher accuracy than the rest. Research by Panigrahi *et al.* [4], CNN was modified by adding linear activation units, using Adam optimizer, adjusting parameters, employing pooling operations, and reducing the number of classifications. This modified CNN algorithm was compared with SVM [8], probabilistic neural network (PNN) [9], deep CNN [3], and CNN [10], in which the modified CNN algorithm achieved the highest accuracy.

We believe that more precise categorization classes will result in improved classification accuracy. Rather than recognizing the type of plant and its diseases at the same time, we intend to first identify the data's general classes before moving on to the more specific classes. In particular, we propose a two-stage classification procedure based on [4] with the first phase classifying the plant type and the second classifying the plant condition. This method was later dubbed two-step CNN. Datasets containing images of plant diseases are frequently imbalanced. Models trained on imbalanced datasets produce inaccurate predictions, favoring the majority class and ignoring the minority class, which can lead to performance issues. The majority class refers to the class with the most images in the dataset, whereas the minority class has fewer images. As a result, it is critical to address imbalanced datasets prior to classification. To address this imbalance, techniques such as synthetic minority over-sampling technique (SMOTE) [11] and augmentation [3] can be employed. SMOTE is a technique that increases the number of images in the minority class by creating synthetic data points and has been applied in many application such as for intrusion detection [12] as well as expansion and classification [13]. Meawhile augmentation involves creating variations in the dataset by modifying or manipulating images and has been employed for many tasks such as minority class augmentation using capsule adversarial networks [14] and breast cancer identification [15]. Once the data is balanced, the next step is to classify the input data into different categories to predict the plant type and condition using the two-step CNN model [16].

In summary, the major contribution of this research are as follows: i) we propose a two-step CNN algorithm that allows for two-step classifications by identifying classification labels, general to more specific labels; ii) we conduct experiments on the proposed solution to validate its effectiveness and efficiency performance, and iii) we build a simulation of the proposed method to show its applicability in real-world scenarios. The remainder of the paper is organized as follows: section 2 shows the proposed approach, the two-step CNN, as well as the scenarios of the experiments done to evaluate the proposed method's performance. Section 3 offers comprehensive experimental data and discussion, as well as a simulation of the suggested method in action. Finally, section 4 concludes the paper.

2. METHOD

On top of modified CNN algorithm [4], we propose two stages of classification approach that we named as two-step CNN. The overview of the two-step CNN as the multi-classification approach [17], is depicted in Figure 1. Specifically, given plant village dataset [18], we first handled the imbalanced dataset problem by incorporating SMOTE and augmentation techniques which involve [19]: random rotation, gaussian blur, brightness adjustment, zoom, channel shift, shear intensity, and gaussian noise to augment images as in [20], [21]. The balanced dataset, enabling the better training of the model using the two-step CNN approach. Subsequently, given the leaf images, the first phase classification process utilizes the model to predict plant species and followed by the second phase to predicts the plant conditions. Later on, we perform the evaluation of the predicted classes using precision, recall, and F1-score. The two primary components of this approach, data balancing and two-step CNN classification, are explained in more detail in the following subsections.

2.1. Data balancing

The plant village dataset consists of 28 class images of various types of leaves and 1 class background without leaves. The dataset exhibits imbalance dataset, where the number of samples in each class varies significantly. This imbalance can affect the model's performance in predicting minority classes, leading to a bias towards accurately predicting the majority classes while less accurately predicting or even ignoring the minority classes. To address this issue, we employed augmentation and SMOTE techniques to improve the model's ability to learn from the minority classes by creating additional synthetic samples.

The SMOTE technique addresses imbalance in dataset by increasing the number of samples in the minority class. It achieves this by creating synthetic samples based on the existing data points in the minority class. These synthetic samples are generated by interpolating between neighboring instances of the same class.

This way, the dataset is augmented with additional synthetic samples for the minority class, effectively balancing the class distribution. On the other hand, augmentation tackles the imbalance in dataset by creating new variations in the dataset through modifications or manipulations of the existing images. Various augmentation techniques can be used, such as rotation, flipping, zooming, cropping, and color adjustments. By introducing these variations, the dataset becomes more diverse, allowing the model to learn from a wider range of examples and improving its generalization performance, especially for classes with limited samples.

Based on [22], [23], SMOTE is performed after data splitting due to the addition of synthetic samples to the class with the fewest samples. When applied prior to the data splitting, there will be a possibility that the synthetic samples would appear simultaneously in both the training data and test data which may lead to data dependence between the training data and test data, affecting the objectivity of model evaluation and distorting the model's performance. Meanwhile, augmentation should be conducted before data splitting [24] since applying it beforehand may lead to the appearance of new variations simultaneously in both the training data and test data. The augmentation process generates the maximum number of images available across class. Thus, as the maximum number of images is 7,000 then each class will be balanced to have this number. Meanwhile, the SMOTE adds the samples to the minority class to meet the maximum number of class for each plant type. Thus, each plant type such as "apple" will have the same number of images in each class i.e., scab, black rot, and cedar apple. The example of the dataset after augmentation and smote is depicted in Table 1.

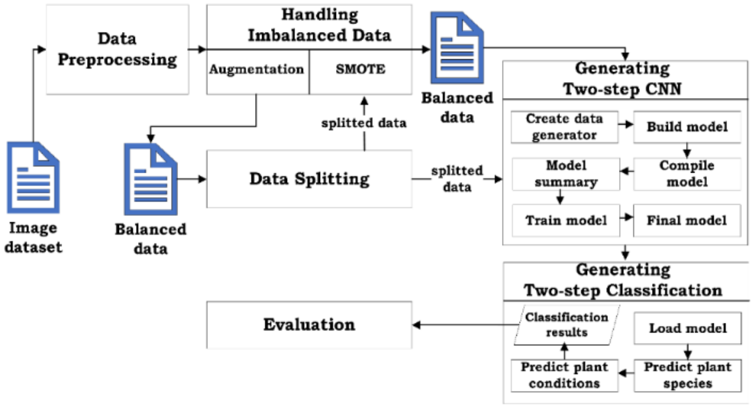


Figure 1. The proposed two-step CNN

Table 1. Data with imbalanced data handling on plant village dataset

No	Class	Number of images	Augmentation result	SMOTE result
1	Apple scab	630	7,000	1,645
2	Apple black rot	621	7,000	1,645
3	Apple Cedar apple rust	275	7,000	1,645
4	Apple healthy	1,645	7,000	1,645

2.2. Classification with two-step convolutional neural network

Two-step CNN has two separate models to classify the main class given the characteristics of the data, in this case i) plant condition and ii) plant type of Plant Village dataset [18]. Specifically, two-step CNN has two stages of classification: i) plant type classification which aims to determine the general category of the plant and; ii) plant condition classification which focuses on specific labels within each plant type to classify the plant condition. Thus, we first accurately classify the plant types, and then we classify the plant condition. There are nine plant types represented in the Plant Village datasets: i) an apple, ii) a cherry, iii) corn, iv) a grape, v) a peach, vi) a pepper bell, vii) a potato, viii) a strawberry, and ix) a tomato. Using two-step CNN, we will have one model to learn to classify the plants based on these nine types, as well as nine models, one for each plant type, to identify the diseases of each plant type. For example, model apple will identify the apple's condition as apple scab, black rot, cedar rush, or healthy apple.

In summary, two-step CNN is trained to identify the type of plant among several categories such as apple, cherry, and corn. Meanwhile, the plant condition classification identifies different conditions of apple like scab, rust, and healthy. By employing two-step CNN, the study tackles both the classification of plant types and their conditions. The architecture of the proposed model, two-step CNN is elaborated.

The CNN algorithm was used in an attempt to identify plant diseases from leaf images. The raw picture is sent through the network layer by the CNN, which then outputs the identified class. In the next subsection we elaborate the architecture of CNN for plant types classification and for plant types conditions.

– Two-step CNN architecture for plant types classification

The proposed network has 5 convolutional layers, each followed by a layer called max-pooling. Each fully connected and convolutional layer's output is subjected to the rectified linear unit (ReLU) activation function. The following is the architecture of two-step CNN model for the classification of the plant types respectively in the format of layer and its type; output shape; and number of parameters.

a) conv2d (Conv2D); (None, 62, 62, 16); 448
b) max pooling2d (MaxPooling2D); (None, 31, 31, 16); 0
c) conv2d 1 (Conv2D); (None, 29, 29, 32); 4,640
d) max pooling2d 1 (MaxPooling2D); (None, 14, 14, 32); 0
e) conv2d 2 (Conv2D); (None, 12, 12, 64); 18,496
f) max pooling2d 2 (MaxPooling2D); (None, 6, 6, 64); 0
g) conv2d 3 (Conv2D); (None, 4, 4, 128); 73,856
h) max pooling2d 3 (MaxPooling2D); (None, 2, 2, 128); 0
i) conv2d 4 (Conv2D); (None, 1, 1, 256); 295,168
j) max pooling2d 4 (MaxPooling2D); (None, 0, 0, 256); 0
k) flatten (Flatten); (None, 0); 0
l) dense (Dense); (None, 512); 512,512
m) dense 1 (Dense); (None, 512); 262,656
n) dense 2 (Dense); (None, 10); 513

– Two-step CNN architecture for plant conditions classification

The following is the architecture of two-step CNN model for the classification of the plant conditions respectively in the format of layer and its type, output shape, and number of parameters.

a) conv2d (Conv2D); (None, 62, 62, 32); 896
b) max pooling2d (MaxPooling2D); (None, 31, 31, 32); 0
c) conv2d 1 (Conv2D); (None, 29, 29, 64); 18,496
d) max pooling2d 1 (MaxPooling2D); (None, 14, 14, 64); 0
e) conv2d 2 (Conv2D); (None, 12, 12, 64); 36,928
f) max pooling2d 2 (MaxPooling2D); (None, 6, 6, 64); 0
g) conv2d 3 (Conv2D); (None, 4, 4, 64); 36,928
h) max pooling2d 3 (MaxPooling2D); (None, 2, 2, 64); 0
i) conv2d 4 (Conv2D); (None, 1, 1, 64); 36,928
j) max pooling2d 4 (MaxPooling2D); (None, 0, 0, 64); 0
k) flatten (Flatten); (None, 0); 0
l) dense (Dense); (None, 64); 64
m) dense 1 (Dense); (None, 2); 2

The CNN model for plant condition classification consists of two steps: convolutional layers (Conv2D) and max pooling layers (MaxPooling2D) to reduce the dimensions of the feature maps. These layers extract and compress spatial information from the input images. The model also includes a flatten layer that converts the feature maps into a 1D vector, which is then processed by two dense layers, each of which has 64 units with ReLU activation functions. The final dense layer matches the number of output classes and uses the softmax activation function for classification. Notably, the model has 130,370 trainable parameters and 0 non-trainable parameters, implying that all weights will be adjusted during training.

2.3. Parameter setting and evaluation metrics

The experiment is conducted using the images with pixels of 256×256. Later, we split the dataset to training, validation, and testing respectively 70%, 20%, and 10%. For the CNN model, we set the learning rate to 0.0001, the epochs to 25, and the batch size to 64. The effectiveness is assessed in terms of accuracy, precision, recall, F1-score, and efficacy in terms of the cost of model training.

3. RESULTS AND DISCUSSION

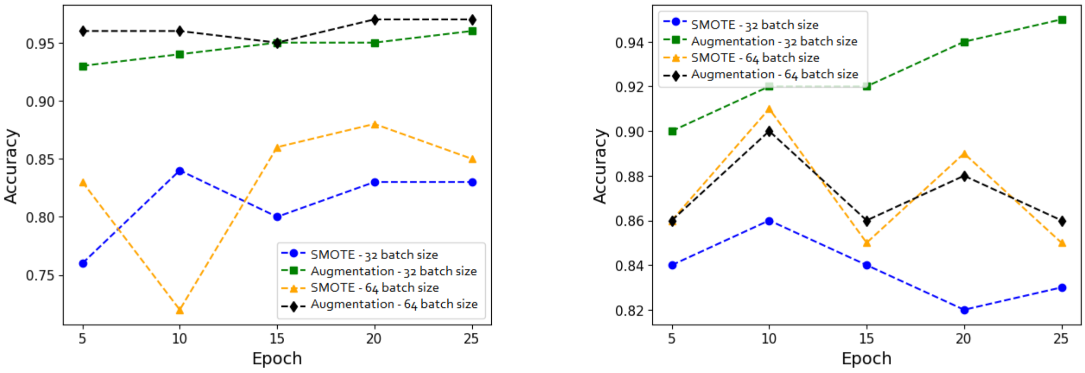
In this section, we evaluate the impact of batch size and data balancing on CNN and the suggested two-step CNN, along with their training duration. First, as indicated in Table 2 and Figure 2, we compare the performance of data balancing approaches, SMOTE, and augmentation in terms of accuracy, precision, recall, and F1-score. Next, we compare the time costs for balancing data with SMOTE and augmentation, as shown in Table 3, and the effect of batch size on the model performance, as shown in Figure 3, to determine how long it takes to train both two-step classification models. Afterwards, as shown in Table 4, we contrast our suggested two-step CNN with competitors CNN [4] and JF-ResNet [7]. The web simulation of our suggested strategy is then displayed in Figure 4, illustrating how our proposed approach can be used in a real-world scenario.

Table 2. The effect of data balancing for CNN and two-step CNN

Model	Batch size	SMOTE			Augmentation		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
CNN	32	0.84	0.83	0.83	0.94	0.94	0.94
	64	0.92	0.93	0.85	0.95	0.95	0.95
Two-step CNN	32	0.92	0.92	0.87	0.99	0.99	0.99
	64	0.93	0.93	0.86	0.99	0.99	0.99

3.1. Comparison of SMOTE vs augmentation

Table 2 depicts the effect of data balancing for both CNN [9] and the proposed model, two-step CNN. We can see that there is a significant difference in performance with the increasing of precision, recall, and F1-score around 0.05 points. The results suggest that using SMOTE or augmentation to balance data improves the performance of both models in terms of precision, recall, and F1-score. Meanwhile, Figures 2(a) and 2(b) respectively show the accuracy for two-step CNN and CNN with data balancing where two-step CNN outperforms CNN in terms of accuracy with augmentation, but CNN outperforms two-step CNN with SMOTE. This makes sense given that the SMOTE technique is applied after the splitting of data into plant types, resulting in even fewer training data. Improvements to the SMOTE technique [25], [26], including user feedback to improve model performance [27], and the dynamic train of model [28] might be considered for future research.



(a) (b)

Figure 2. Comparison of SMOTE and augmentation data balancing methods for two-step CNN and CNN: (a) data balancing for two-step CNN and (b) data balancing for CNN

Figure 2 depicts the comparison of data balancing techniques, respectively SMOTE and augmentation for two-step CNN in Figure 2(a) and CNN model in Figure 2(b). Linear with the findings in [29], data balancing with augmentation outperforms SMOTE for both models. It is reasonable given that the data used for balancing is more reliable when it is augmented with samples to reach the largest possible target class. In terms of batch size, we can see that both batches perform similarly, with 64 batch size slightly outperforming the 32 batch size.

3.2. Comparison of training time of two-step CNN for augmentation vs. SMOTE

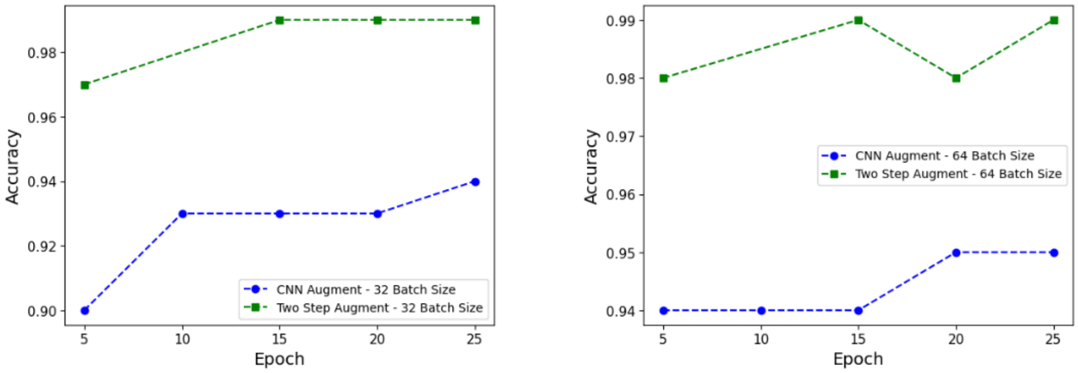
In Table 3, we compare the time cost of model training for augmentation and SMOTE. It can be seen that augmentation takes longer to train the data than SMOTE. Yet, it is still reasonable considering the improvement in accuracy and the number of parameters that must be learned during training.

Table 3. Training cost of two-step CNN for augmentation and SMOTE

Criteria	Time in train (second)	
	Augmentation	SMOTE
Plant type	1612.07	1093.02
Plant condition	429.50	79.59

3.3. The impact of batch size for the model

We also assess the influence of batch-size on accuracy between two-step CNN and baseline CNN, as shown in Figure 3. We assess the impact of batch size on accuracy for both the two-step CNN and the baseline CNN models, as illustrated in Figure 3. A comparison of the results from Figure 3(a) with a batch size of 32 and Figure 3(b) with a batch size of 64 shows that a batch size of 64 yields higher accuracy across both models. Additionally, the two-step CNN consistently outperforms the baseline CNN regardless of the batch size.



(a) (b)

Figure 3. Comparison of SMOTE and augmentation data balancing methods for two-step CNN and CNN in terms of accuracy: (a) CNN vs. two-step CNN using augmentation with a batch size of 32 and (b) CNN vs. two-step CNN using augmentation with a batch size of 64

3.4. Performance comparison with competitors

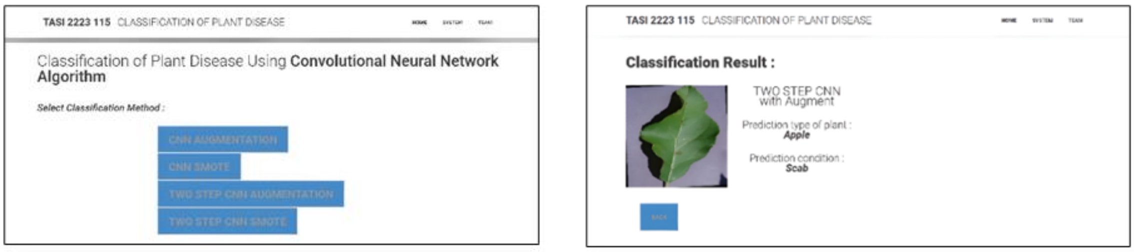
Table 4 shows the comparison of two-step CNN with competitors namely CNN [4] and JF-ResNet [7]. The more thorough classification that is carried out in two steps allows us to observe that, our proposed two-step CNN technique achieves 99% accuracy, which is superior than 95% and 97% accuracy for CNN and JF-ResNet, respectively. The improved performance is attributed to the more detailed, two-step classification process used in the proposed model.

Table 4. Performance comparison in terms of accuracy

Model	Accuracy
CNN [4]	0.95
JF-ResNet [7]	0.97
two-step CNN	0.99

3.5. Simulation of two-step convolutional neural network

In addition, we run a simulation to prove that the two-step CNN model can detect plant diseases from the leaf, as shown in Figure 4. Figure 4(a) displays the options of the models and Figure 4(b) displays the classification results given the input. The test shows that two-step CNN can identify diseases in unobserved leaves. In addition to that, the simulation returns “undetected object” for non-leaf items.


TASI 2223 115 CLASSIFICATION OF PLANT DISEASE

Classification of Plant Disease Using **Convolutional Neural Network Algorithm**

Select Classification Method :

TASI 2223 115 CLASSIFICATION OF PLANT DISEASE

Classification Result :

TWO STEP CNN
with Augment

Prediction type of plant :
Apple

Prediction condition :
Scab

(a) (b)

Figure 4. Model simulation in web-based application (a) option menu to choose the model to predict the plant diseases and (b) the prediction result for the given output

4. CONCLUSION

Based on the experiments, we can conclude that our proposed two-step CNN outperforms the baseline, CNN for almost all the evaluation settings. Thus, rather than training the data for having more accurate class prediction, our findings provide conclusive evidence that two-steps classification may improve the classification by narrowing the prediction error in each classification steps. In terms of data balancing, we found that model performance with augmentation outperforms SMOTE, although it requires more training time. In conclusion, the proposed model, two-step CNN method, outperforms the baseline, CNN, for plant disease identification. However, a more thorough assessment of the effectiveness of this two-step CNN technique may be needed by incorporating a larger dataset with various labels for classification.

ACKNOWLEDGEMENTS

The authors would like to thank Institut Teknologi Del for financially supporting this research through Institute of Research and Community Service or Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM).

REFERENCES

[1] D. S. Rao *et al.*, “Plant disease classification using deep bilinear CNN,” *Intelligent Automation and Soft Computing*, vol. 31, no. 1, pp. 161–176, 2022, doi: 10.32604/IASC.2022.017706.

[2] I. Y. Purbasari, B. Rahmat, and C. S. Putra, “Detection of rice plant diseases using convolutional neural network,” *IOP Conference Series: Materials Science and Engineering*, vol. 1125, no. 1, 2021, doi: 10.1088/1757-899x/1125/1/012021.

[3] G. Geetharamani and J. A. Pandian, “Identification of plant leaf diseases using a nine-layer deep convolutional neural network,” *Computers and Electrical Engineering*, vol. 76, pp. 323–338, 2019, doi: 10.1016/j.compeleceng.2019.04.011.

[4] K. P. Panigrahi, A. K. Sahoo, and H. Das, “A CNN approach for corn leaves disease detection to support digital agricultural system,” in *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, 2020, pp. 678–683, doi: 10.1109/ICOEI48184.2020.9142871.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv-Computer Science*, pp. 1-14, Sep. 2015.

[7] G. Ezhilarasan, S. R. Ranganathan, L. S. Mani, and S. Kadry, “An intelligent deep residual learning framework for tomato plant leaf disease classification,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 3, pp. 3168–3176, 2024, doi: 10.11591/ijece.v14i3.pp3168-3176.

[8] M. A. Chandra and S. S. Bedi, “Survey on SVM and their application in image classification,” *International Journal of Information Technology*, vol. 13, no. 5, pp. 1–11, 2021, doi: 10.1007/s41870-017-0080-1.

[9] S. R. Mohanty, P. K. Ray, N. Kishor, and B. K. Panigrahi, “Classification of disturbances in hybrid DG system using modular PNN and SVM,” *International Journal of Electrical Power and Energy Systems*, vol. 44, no. 1, pp. 764–777, 2013, doi: 10.1016/j.ijepes.2012.08.020.

[10] J. Lu, L. Tan, and H. Jiang, “Review on convolutional neural network (CNN) applied to plant leaf disease classification,” *Agriculture*, vol. 11, no. 8, 2021, doi: 10.3390/agriculture11080707.

[11] D. Dablain, B. Krawczyk, and N. V. Chawla, “DeepSMOTE: fusing deep learning and SMOTE for imbalanced data,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, 2023, doi: 10.1109/TNNLS.2021.3136503.

[12] L. Tian and Y. Lu, “An intrusion detection model based on SMOTE and convolutional neural network ensemble,” *Journal of Physics: Conference Series*, vol. 1828, no. 1, 2021, doi: 10.1088/1742-6596/1828/1/012024.

[13] S. Wang, Y. Dai, J. Shen, and J. Xuan, “Research on expansion and classification of imbalanced data based on SMOTE algorithm,” *Scientific Reports*, vol. 11, no. 1, Oct. 2021, doi: 10.1038/s41598-021-03430-5.

[14] P. Shamsolmoali, M. Zareapoor, L. Shen, A. H. Sadka, and J. Yang, “Imbalanced data learning by minority class augmentation using capsule adversarial networks,” *Neurocomputing*, vol. 459, pp. 481–493, 2021, doi: 10.1016/j.neucom.2020.01.119.

[15] M. Saini and S. Susan, “Deep transfer with minority data augmentation for imbalanced breast cancer dataset,” *Applied Soft Computing Journal*, vol. 97, 2020, doi: 10.1016/j.asoc.2020.106759.

[16] Z. Iqbal, M. A. Khan, M. Sharif, J. H. Shah, M. H. U. Rehman, and K. Javed, “An automated detection and classification of citrus plant diseases using image processing techniques: a review,” *Computers and Electronics in Agriculture*, vol. 153, pp. 12–32, 2018, doi: 10.1016/j.compag.2018.07.032.

[17] F. M. Alsuhimat and F. S. Mohamad, “A hybrid method of feature extraction for signatures verification using CNN and HOG a multi-classification approach,” *IEEE Access*, vol. 11, pp. 21873–21882, 2023, doi: 10.1109/ACCESS.2023.3252022.

[18] S. P. Mohanty, “Plant village dataset,” *Kaggle,* 2018. Accessed: Nov. 23, 2018. [Online]. Available: https://github.com/spMohanty/PlantVillage-Dataset

[19] P. Kaur, B. S. Khehra, and E. B. S. Mavi, “Data augmentation for object detection: a review,” in *Midwest Symposium on Circuits and Systems*, 2021, pp. 537–543, doi: 10.1109/MWSCAS47672.2021.9531849.

[20] M. Arslan, M. Guzel, M. Demirci, and S. Ozdemir, “SMOTE and Gaussian noise based sensor data augmentation,” in *Proceedings, 4th International Conference on Computer Science and Engineering*, 2019, pp. 458–462, doi: 10.1109/UBMK.2019.8907003.

[21] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, “A comprehensive survey of image augmentation techniques for deep learning,” *Pattern Recognition*, vol. 137, 2023, doi: 10.1016/j.patcog.2023.109347.

[22] E. F. Swana, W. Doorsamy, and P. Bokoro, “Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset,” *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.

[23] T. Pan, J. Zhao, W. Wu, and J. Yang, “Learning imbalanced datasets based on SMOTE and Gaussian distribution,” *Information Sciences*, vol. 512, pp. 1214–1233, 2020, doi: 10.1016/j.ins.2019.10.048.

[24] F. Marzougui, M. Elleuch, and M. Kherallah, “A deep CNN approach for plant disease detection,” in *2020 21st International Arab Conference on Information Technology, ACIT 2020*, 2020, pp. 1–6, doi: 10.1109/ACIT50332.2020.9300072.

[25] F. Koto, “SMOTE-out, SMOTE-cosine, and selected-SMOTE: an enhancement strategy to handle imbalance in data level,” in *ICACSIS 2014: 2014 International Conference on Advanced Computer Science and Information Systems*, 2014, pp. 280–284, doi:

10.1109/ICACSIS.2014.7065849.

[26] X. Chao and L. Zhang, “Few-shot imbalanced classification based on data augmentation,” *Multimedia Systems*, vol. 29, no. 5, pp. 2843–2851, 2023, doi: 10.1007/s00530-021-00827-0.

[27] R. Lumbantoruan, X. Zhou, Y. Ren, and L. Chen, “I-CARS: an interactive context-aware recommender system,” in *IEEE International Conference on Data Mining, ICDM*, 2019, pp. 1240–1245, doi: 10.1109/ICDM.2019.00154.

[28] R. Lumbantoruan, X. Zhou, Y. Ren, and Z. Bao, “D-CARS: a declarative context-aware recommender system,” in *IEEE International Conference on Data Mining, ICDM*, 2018, pp. 1152–1157, doi: 10.1109/ICDM.2018.00151.

[29] Y. S. Won, D. Jap, and S. Bhasin, “Push for more: On comparison of data augmentation and SMOTE with optimised deep learning architecture for side-channel,” in *Information Security Applications*, 2020, pp. 227–241, doi: 10.1007/978-3-030-65299-9_18.

BIOGRAPHIES OF AUTHORS

Rosni Lumbantoruan, Ph.D. holds her Ph.D. in Computer Science from Royal Melbourne Institute of Technology (RMIT) University, Australia, in 2021, with the Dissertation “Declarative context-aware recommendation”. She earned a master’s degree in Information System Development from HAN University of Applied Sciences, the Netherlands, in 2010 and a bachelor’s degree in informatics from Bandung Institute of Technology (ITB), Indonesia, in 2007. She is a lecturer in the Informatics and Electric Engineering faculty at Institut Teknologi Del (IT Del) in Indonesia. Her research areas of interest include recommender systems, machine learning, and knowledge discovery. She can be contacted at email: rosni@del.ac.id.

Nico Rajagukguk received his bachelor’s degree in Information System from Institut Teknologi Del (IT Del), Indonesia in 2023. His research interests are data analytics, data science, and data engineering. He can be contacted at email: nikorajagukguk647@gmail.com.

Anju Ucok Lubis received his bachelor’s degree in Information System from Institut Teknologi Del (IT Del), Indonesia in 2023. His research interests are web development, mobile development, and data mining. He is currently working as mobile develover in an IT company. He can be contacted at email: anjuucoklubis@gmail.com.

Marwani Claudia received her bachelor’s degree in Information System from Institut Teknologi Del (IT Del), Indonesia in 2023. Her research interests are data analytics, data science, and machine Learning. She can be contacted at email: claudiamarwani@gmail.com.

Humasak Simanjuntak is currently pursuing his Ph.D. degree in Computer Science in The University of Sheffield, South Yorkshire, England. He received his master’s degree in information system development from HAN University of Applied Sciences, the Netherlands in 2010 and his B.Sc. degree of Informatics from Bandung Institute of Technology (ITB), Indonesia in 2007. He is currently a lecturer in the faculty of Informatics and Electric Engineering at Institut Teknologi Del (IT Del), Indonesia. His research interests include machine learning, data mining, data science, big data analytics, and artificial intelligence. He can be contacted at email: humasak@del.ac.id.