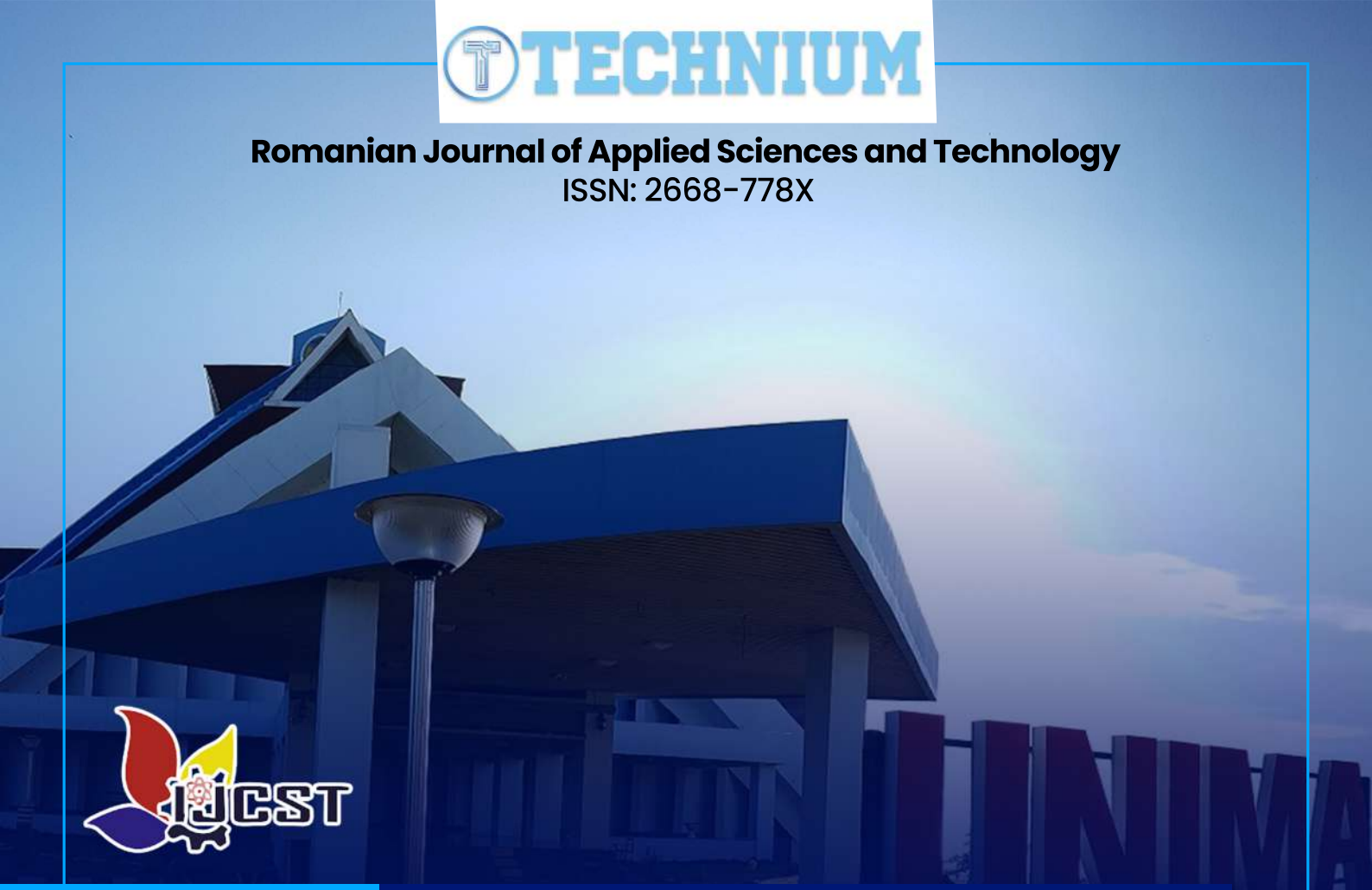
TECHNIUM

Romanian Journal of Applied Sciences and Technology
ISSN: 2668-778X



vol. 16 / 2023

The 7th International Conference on Science Technology

organized by
Faculty of Social Science and
Law Universitas Negeri Manado and
Consortium of International Conference
on Science and Technology

The Innovation Breakthrough in Digital and Disruptive Era

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 112-117, October, 2023
ISSN: 2668-778X
www.techniumscience.com

Oversampled-Based Approach to Overcome Imbalance Data in the Classification of Apple Leaf Disease with SMOTE

Eva Y Puspaningrum[11], *Yisti Vita Via*[2], *Chilyatun Nisa*[3], *Hendra Maulana*[4], *and Wahyu S.J.Saputra*[5]
[1,2,3] Informatics Department, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia
[4] Digital Busines Department, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia
[5] Data Science Department, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia

Abstract. Research on the detection of apple leaf disease has been developed. Various methods have been carried out to detect apple leaf disease, one of which is by processing digital images. In this study, the author proposes the Convolutional Neural Network (CNN) algorithm as a feature extractor and classifier of apple leaf images. CNN was chosen because it can apply learning and classification effective and automated image features than traditional feature extraction methods. The dataset used is Plant Pathology 2020 - FGV C7. In this dataset, it was found that the image size varies greatly from the entire dataset or often referred to as data imbalance. In this study, the oversampling technique was successfully applied to handle the uneven distribution of data (imbalanced) and achieved a good evaluation result. The oversampling approach method used is Synthetic Minority Oversampling Technique (SMOTE). The number of imbalanced images is carried out by SMOTE pre-processing to produce balanced data. The CNN algorithm is trained on training data and performance testing on test data with a ratio of 70:30 of the total data. The learning model on the network structure can achieve an accuracy of 92% with data that has been oversampled.

[1] Corresponding author: evapuspaningrum.if@upnjatim.ac.id

TECHNIUM

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 16, 112-117, October, 2023
ISSN: 2668-778X
www.techniumscience.com

1 Introduction

Leaf disease is one of the main obstacles in plant maintenance [1]. one of the tropical plants, especially in Indonesia, one of which is the apple. Leaf diseases can affect fruit production. Some of the most common and adversely affecting apple leaf diseases include scabies, frog's eyespot, Grey Spot, Brown Spot and rust [2]. Therefore, the detection of apple leaf disease is increasingly attracting attention and is very important for prevention.

Experts have made visual observations using photographs to diagnose plant diseases. although, there will be errors due to subjective perception [3]. In recent years, by utilizing digital camera can diagnosis of plant diseases can be automatically using machine learning because it can give satisfactory results.

Convolutional Neural Network (CNN) is a type of deep artificial neural network method commonly used in processing image data. CNN has an advantage in image processing because it uses convolution and pooling principles for automatic feature extraction from images. thus, enabling CNN to learn patterns and recognize objects better and faster. In addition, this method can be easy to train large data by optimizing the network structure [4].

Therefore, the application of CNN in the field of identification of plant diseases is increasing [5]. For example, in the study by Sladojevic et al. [6] in research that has carried out the CNN method for leaf diseases and achieved an accuracy rate of 96.3%. Then the automatic recognition method for tobacco disease images using CNN was proposed by Sun et al [5], and the tea leaf disease recognition system using CNN produced an accuracy of 93.75%, while the accuracy of other methods such as SVM and BP neural network was 89.36%. and 87.69%. So, CNN is considered as a successful learning method compared to others.

Many research on detection of apple leaf disease have been developed, such as that conducted by Nachtigall, Lucas G., et al in 2016 [7] using the CNN algorithm with AlexNet architecture to classify herbicide damage to apple leaves. The dataset was obtained by photographing each leaf consisting of 2539 images from 6 classes and yielded an accuracy of 97%. The application of the CNN algorithm carried out by Fang Tao., et al in 2019 [8] which combines batch normalization and central loss functions based on the VGG16 architecture for the classification of apple leaf diseases. The data used are 5,373 sick leaf images and 1,683 healthy leaf images. The proposed model has an accuracy of 95.0%. The application of the CNN algorithm was also carried out by Baranwal Saraansh., et al in 2019 [9] using the GoogleNet architecture to detect apple leaf disease. Using the PlantVillage dataset which consists of four classes, three of which are diseased leaves with a total of 1,526 and healthy leaves of 1,000 samples. The model gets an accuracy of 98.54%. A similar study was also conducted by Rehman et al [10] used the RCNN MASK method configured to detect infected areas and retrained ResNet-50 for classification. by using the experimental Plantvillage dataset achieved the best accuracy of 96.6%. The PlantVillage dataset was released in 2018 and many previous studies have used it.

One of the datasets containing images of apple leaf disease is in the Plant Pathology2020 -FGVC7 dataset which was released in 2020. In this dataset, it was found that the image sizes vary greatly from the entire dataset. n machine learning, the number of data samples used in each dataset class used has a equal number [11] or is called balanced data [12]. But this assumption raises many obstacles in its application. For example, in agriculture, due to the impact of morbidity and subjective factors in data processing for this field. for example. There is a difference in the number of categories in the data set. If the number of one category is much more than the other categories, it can be considered as an unbalanced data set. In this case, the classifier is more likely to classify the majority class than the minority class. Therefore, in this case a dataset with an unbalanced data set becomes a challenge in machine learning [13] [14]

Data imbalance occurs when there is more data in one of the existing classes. Classification with unbalanced data can cause problems because machine learning algorithms will struggle because of the differences. Classification method will ignore small data classes [15]. Most of the classification methods will process data that has the same amount or distribution of data so that it will maximize accuracy [16]. There are many methods or solutions to solve the data imbalance problem. One of them is proposed at the data level and algorithm level. At the data level, various re-sampling techniques are applied to balance the class distribution, including over-sampling and under-sampling. This study proposes a model using the Convolutional Neural Network (CNN) architecture and using the Plant Pathology2020 -FGVC7 dataset which was released in 2020. There has been no research using this dataset except for research related to the dataset itself.

2 Methods

In this study, it starts from the data preprocessing stage. This stage is to solve the problem of unbalanced data. After the data is balanced, it will be continued with the design of the CNN architecture, then the model formed from the algorithm will be trained with the previously processed training data. After passing through the training stage the performance of the model will be tested with test data and performance results can be known by carrying out the model evaluation.

1.1 Dataset

The dataset used is taken from a study entitled "The Plant Pathology 2020 challenge dataset to classify foliar disease of apples" [17], where the dataset is public. The dataset consists of 1821 color images of apple leaves. This dataset was found on the Kaggle site under the name Plant Pathology 2020 - FGVC7 uploaded by Fine-Grained Visual Categorization. In

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 112-117, October, 2023
ISSN: 2668-778X
www.techniumscience.com

Plant Pathology 2020 - FGVC7, the Dataset is divided into 4 image classes shown Figure 1. The available classes are images of healthy leaves, leaves with various diseases, rusty leaves, and scab leaves. Healthy leaf class has data of 516 images, leaves of multiple diseases 91 images, rusty leaves 622 images, and leaf scabs have 592 images data.

In this dataset, it is found that the image sizes are very diverse from the entire dataset, so a data preprocessing stage is needed. The distribution of the amount of data for each image class is not the same (imbalanced data), so it is necessary to apply the data oversampling method to equalize the number of minority data as much as the majority data. This is done to avoid overfitting the model.

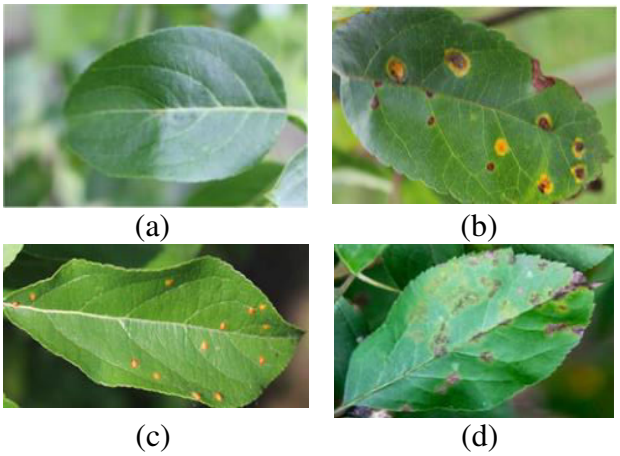


Fig. 1. Sample images (a) Healthy (b) Multiple (c) Scab (d) Rust

1.2 SMOTE

The oversampling technique used in this study is the Synthetic Minority Oversampling Technique (SMOTE). SMOTE is an oversampling technique that is different from the existing classical techniques. In the classical oversampling technique, the minority data is doubled from the minority data population. Although this classic technique will increase the amount of data, it does not provide new information or variations on machine learning models. Therefore, the SMOTE technique was used [18]. This technique explains that SMOTE works by utilizing the k-nearest neighbor algorithm to generate synthetic data. The first SMOTE begins by selecting random data from the minority class, then using the k-nearest neighbor method, synthetic data will be generated, then the data will be randomized with the k-nearest neighbors.

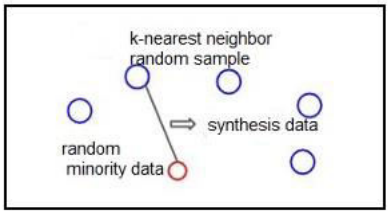


Fig. 2. Illustration of SMOTE

The process illustrated in Figure 2 will be repeated many times until the minority class has the same number as the majority class. The SMOTE sample is a combination of two similar samples from the minority class defined in Equation 1 [19].

$$s = x + u\,(x^r - x) \quad (1)$$

where $0 \leq u \leq 1$; x R is randomly selected among the 5 nearest neighbor minority classes of x.

The final stage in data preprocessing is to divide the dataset into 2, namely training data and test data. The training data consists of 1275 images data (70% dataset), while the test data consists of 546 image data (30% dataset) which is used to test the performance of the CNN model being trained.

1.3 CNN

Figure 3 is an illustration of the CNN algorithm architecture in this study. In the picture, CNN functions as a feature extractor from image data to carry out the classification process. The CNN architecture in this study consists of 3 convolutions layers and 3 pooling layers.

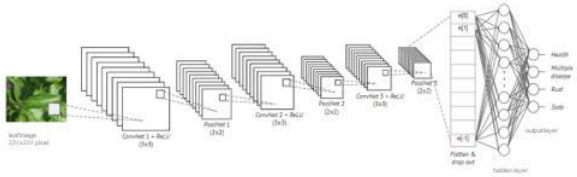


Fig. 3. Illustration of the proposed CNN Architecture

An illustration of the method architecture is shown in Fig. 3. Layer convolution is the first stage in CNN architecture. The stage uses a kernel of a certain size, the number of kernels used depends on the number of features produced. The next layer is the ReLU (Rectified Linear Unit) activation function, after the activation process is complete, the process continues at the pooling layer. This process will be repeated until a sufficient feature map is obtained to proceed to the fully connected layer, after which the layer will produce an output class. for detail the stages of the illustration are shown in Figure 4

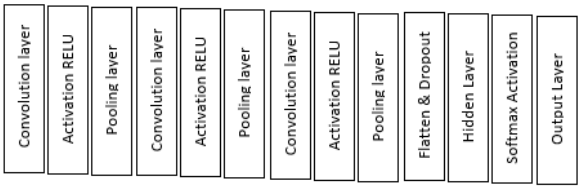


Fig. 4. Detail Illustration stage of the proposed Architecture

In this CNN architecture, starting from the input image, it enters the convolution layer, then enters the ReLU activation function, after which it is forwarded to the Max Pooling layer and the dropout regularization process. Then the data is made into a 1-dimensional matrix to be processed in the hidden layer, then the value of each image is entered into the Softmax activation function, after the value is activated, it is

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 16, 112-117, October, 2023
ISSN: 2668-778X
www.techniumscience.com

forwarded to the output layer and produces data detection results from the input image data.

This study uses the ReLU activation function and the Sigmoid activation function on the CNN architecture. In the ReLU activation function, if there is a value less than 0 then ReLU will change it to a value of 0 while if there is a value more than 0 then ReLU will not change that value.

At the training stage, the model is carried out with 1457 images from the training data that have gone through the preprocessing stage. The image sequence, the kernel value in the convolution layer and the weight and bias values on CNN every time you start training will always be initialized randomly, therefore these values will be stored after the training process is complete. Then for the results of this training in the form of a CNN algorithm model with kernel values, weights, and biases that have been trained. In the model testing phase, 364 color images were used. After the testing process, the predicted value will be compared with the original value based on the test data, then it can be determined how many correct predictions are. The performance of a model can be seen from the evaluation results based on the confusion matrix. The confusion matrix is used to determine the performance results of the model that has been made, namely with accuracy, precision, recall, and f1-score.

2 Result And Discussion

The process of transforming the image from BGR image to RGB image and changing the image size to 224×224 pixels, followed by defining the data into an array and performing normalization. Then do oversampling to handle data that is not evenly distributed (imbalanced) by using the SMOTE method, so that the dataset which initially amounted to 1024 turned into 2488 data. Finally, at this preprocessing stage, the dataset is divided into two data sets, namely training data and test data with a ratio of 70:30 so that the training data is 1742 data, and the training data is 746 data.

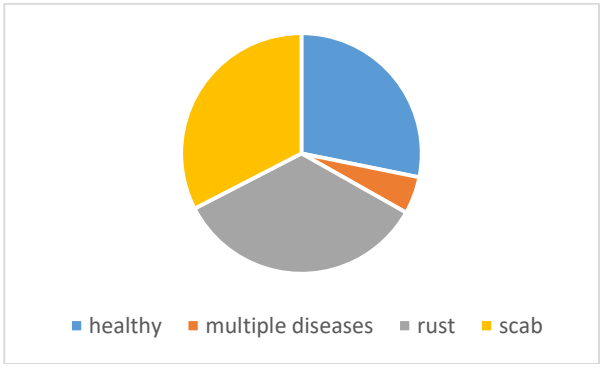


Fig. 5. Distribution of Data Before SMOTE

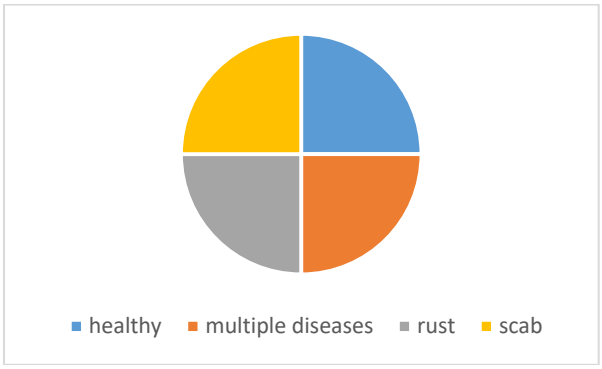


Fig. 6. Distribution Data with SMOTE

In Fig.5 show the data distribution is imbalanced in each category. Each category includes healthy 512 data (blue), multiple diseases 91 data (orange), rust 622 data (gray), and scab 592 data (yellow), with the healthy category as the majority data and multiple data and multiple diseases as minority data. The total data is 1821 image data. Because the distribution of this data will be averaged using the SMOTE method, later all minority data will be duplicated by number of majority data (in this case the rust category), so that later all categories will have the same amount of data as the majority data, which is 622 data. Fig. 6 is the distribution of the data after going through the oversampling stage with the SMOTE method, the problem of the uneven distribution of data can be resolved. The distribution of data is evenly distributed with each category totaling 622 data according to the majority data and the total data being a total of 2488 ideals data.

After the oversampling process is carried out, the data will be processed on CNN. The convolution layer is the main process that underlies a CNN architecture network. In this study, the function for the convolution layer in the library uses TensorFlow. This Conv2D layer creates a convolution kernel with the input layer to generate an output tensor. After that the Max pooling layer will sample the input representation by taking the maximum value over the specified window for each dimension along the feature axis. The window shifts with a step in each of its dimensions. In the test determine the number of filters in the convolution layer by setting the value on the variable with n as the order of the convolution layer and set the number of nodes in the hidden layer by setting the value on the variable. In addition, there is also a Dropout layer which in this study uses a value of 0.5 and an output layer whose number is adjusted to the number of classes, in this study there are 4 classes. Don't forget to also use ReLU activation at the convolution, pooling, and hidden layers as well as Softmax activation at the output layer.

Tests were carried out with different scenarios according to the choice of the number of filters in the convolution layer and the number of nodes in the hidden layer in the CNN architecture. The variations in the number of nodes in the hidden layer are 256, 512, and 1024 nodes, the variations in the number of filters in the three convolution layers of each layer 8, 16, 32 and 16, 32, 64. It should be noted that the more nodes in the hidden layer and the number of filters in the

CNN convolution layer, the more memory (RAM) the computer must allocate.

The best result of training uses a model with three convolution layers (each layer consists of 8, 16, and 32 filters) and a hidden layer with a total of 1024 nodes. After being trained with 15 epochs, the loss value is 0.0468, accuracy is 0.9847, loss validation is 0.0094, and accuracy validation is 0.9994 with a training time of 845,011 seconds.

After the training process is carried out, the model is ready to be tested. To test the CNN model that has been trained previously by using the model predict function on the test variable. The output of the testing process is the prediction result stored in the prediction result variable.

Furthermore, in the evaluation of the model in this study, the confusion matrix function is used to evaluate the performance of the model in testing. The results of the evaluation can be seen in Fig. 7 and Fig.8.

	precision	recall	f1-score	support
0	0.90	0.86	0.88	132
1	0.93	0.99	0.96	501
2	0.84	0.72	0.78	36
3	0.85	0.60	0.70	77
accuracy			0.92	746
macro avg	0.88	0.79	0.83	746
weighted avg	0.91	0.92	0.91	746

Fig. 7. The Result With SMOTE

	precision	recall	f1-score	support
0	0.52	0.49	0.50	167
1	0.07	0.18	0.10	11
2	0.42	0.52	0.47	155
3	0.58	0.46	0.52	213
accuracy			0.48	546
macro avg	0.40	0.41	0.40	546
weighted avg	0.51	0.48	0.49	546

Fig. 8. The Result Without SMOTE

3 Conclusion

In this study, the oversampling technique was successfully applied to handle the uneven distribution of data (imbalanced) and achieved a good evaluation result. The oversampling approach method used is SMOTE. The next step is to enter the CNN network design stage which manages many hyper-parameters such as the convolution layer, pooling layer, activation function selection, dropout, optimizer, the number of nodes in the hidden layer that can affect the results. The performance of the learning model from testing using the CNN algorithm obtained from the value of precision, recall, F1-score, and the best accuracy is a structure that has 8, 16, 32 filters in the convolution layer and 1024 nodes in the hidden layer on CNN. The learning model on the network structure can achieve an accuracy of 92% with data that has been oversampled before. While for data that has not been preprocessed, SMOTE only reaches 48% accuracy.

References

[1] Q. Yan, B. Yang, W. Wang, B. Wang, P. Chen, and J. Zhang, "Apple leaf diseases recognition based on an improved convolutional neural network," Sensors (Switzerland), vol. 20, no. 12, pp. 1–14, 2020, https://doi.org/10.3390/s20123535

[2] X. Chao, G. Sun, H. Zhao, M. Li, and D. He, "Identification of apple tree leaf diseases based on deep learning models," Symmetry (Basel)., vol. 12, no. 7, pp. 1–17, 2020, doi:10.3390/sym12071065

[3] [3] P. Jiang, Y. Chen, B. Liu, D. He, and C. Liang, "Real-Time Detection of Apple Leaf Diseases Using Deep Learning Approach Based on Improved Convolutional Neural Networks," IEEE Access, vol. 7, pp. 59069–59080, 2019, doi: 10.1109/ACCESS.2019.2914929

[4] [4] A. Ajit, K. Acharya, and A. Samanta, "A Review of Convolutional Neural Networks," Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020, pp. 1–5, 2020, DOI:10.1109/ic-ETITE47903.2020.049

[5] [5] X. Sun, S. Mu, Y. Xu, Z. Cao, and T. Su, "Image Recognition of Tea Leaf Diseases Based on Convolutional Neural Network," 2018 Int. Conf. Secur. Pattern Anal. Cybern. SPAC 2018, pp. 304–309, 2018, doi: 10.1109/SPAC46244.2018.8965555

[6] [6] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification," Comput. Intell. Neurosci., vol. 2016, 2016, DOI:10.1155/2016/3289801

[7] [7] L. G. Nachtigall, R. M. Araujo, and G. R. Nachtigall, "Classification of apple tree disorders using convolutional neural networks," Proc. - 2016 IEEE 28th Int. Conf. Tools with Artif. Intell. ICTAI 2016, pp. 472–476, 2017, DOI:10.1109/ICTAI.2016.0078

[8] [8] T. Fang, P. Chen, J. Zhang, and B. Wang, Identification of Apple Leaf Diseases Based on Convolutional Neural Network, vol. 11643 LNCS. Springer International Publishing, 2019, DOI:10.1007/978-3-030-26763-6_53

[9] [9] S. Baranwal, S. Khandelwal, and A. Arora, "Deep Learning Convolutional Neural Network for Apple Leaves Disease Detection," SSRN Electron. J., pp. 260–267, 2019, DOI:10.2139/ssrn.3351641

[10] [10] Z. ur Rehman et al., "Recognizing apple leaf diseases using a novel parallel real-time processing framework based on MASK RCNN and transfer learning: An application for smart agriculture," IET Image Process., vol. 15, no. 10, pp. 2157–2168, 2021. https://doi.org/10.1049/ipr2.12183

[11] [11] L. Ren and W. Zhou, "Optimization research and application of unbalanced data set multi-classification algorithm," Proc. - 2016 8th Int. Conf. Intell. Human-Machine Syst. Cybern. IHMSC 2016, vol. 2, pp. 39–42, 2016, doi: 10.1109/IHMSC.2016.272

[12] [12] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," Int. J. Emerg. Technol. Adv. Eng., vol. 2, no. 4, pp. 42–47, 2012.

[13] [13] T. Su, S. Mu, M. Dong, W. Sun, and A. Shi, "An improved tradaboost for image recognition of unbalanced plant leaf disease," ACM Int. Conf. Proceeding Ser., pp. 374–379, 2019, https://doi.org/10.1145/3373509.3373583

[14] [14] J. Luengo, A. Fernández, S. García, and F. Herrera, "Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling," Soft Comput., vol. 15, no. 10, pp. 1909–1936, 2011, DOI:10.1007/s00500-010-0625-8

[15] [15] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," 2009 IEEE Symp. Comput. Intell. Data Mining, CIDM 2009 - Proc., pp. 324–331, 2009, doi: 10.1109/CIDM.2009.4938667

[16] [16] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Diversity exploration and negative correlation learning on imbalanced data sets" ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 1–6, 2004, doi: 10.5555/1704175.1704434

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 16, 112-117, October, 2023
ISSN: 2668-778X
www.techniumscience.com

[17] [17] R. Thapa, N. Snavely, S. Belongie, A. Khan. "The Plant Pathology 2020 challenge dataset to classify foliar disease of apples" Plant Pathology and Plant-Microbe Biology Section, Cornell University, Geneva, NY, 14456, USA, DOI:10.1002/aps3.11390

[18] [18] B. Kovács, F. Tinya, C. Németh, and P. Ódor, "Unfolding the effects of different forestry treatments on microclimate in oak forests: results of a 4-yr experiment," Ecol. Appl., vol. 30, no. 2, pp. 321–357, 2020, https://doi.org/10.1002/eap.2043

[19] [19] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," BMC Bioinformatics, vol. 14, 2013, https://doi.org/10.1186/1471-2105-14-106