2021 IEEE 7th Information Technology International Seminar (ITIS)
Surabaya, Indonesia, October 6 - 8, 2021

An ensemble-of-classifiers approach for corn leaf based diseases detection

Wahyudi Agustiono
Engineering Faculty
University of Trunojoyo Madura
Bangkalan, Indonesia
wahyudi.agustiono@trunojoyo.ac.id
ORCID ID: 0000-0001-6816-5117

***Abstract*— This study at hand proposes an ensemble-of-classifiers approach for corn leaf based diseases detection. Instead of using a single classical machine learning method, it developed a detection model by employing two classification methods: Voting Classifier with Majority Voting and 5 Classification and Regression Tree (CART) with max depth 2, 4, 6, 8, and 10 to obtain higher accuracy in results. In particular, it benefited from PlantVillage, a public repository that provided over 8,000 corn leaf images as the dataset for the training and testing. All these images were extracted using the Global Color Histogram (GCH), Color Coherence Vector (CCV). Test results of segmented image against Red pixel using the Color Processing Detection Algorithm (CPDA) showed the best accuracy with a Precision of 82.92%, Recall 82.55%, and an f1-Score of 82.6%.**

Keywords—corn leaf-based disease, leaf image, ensemble classification, machine learning classification.

I. INTRODUCTION

The Digital Image Processing (DIP) is an area of research in the computer science that has been developing and receiving a considerable interest. In general, as shown in Fig 1, the stages of DPI include image acquisition, pre-processing, segmentation, post-processing and further analysis tasks such as object identification, recognition, quantification, and classification. One of the fundamental stages of the DIP which is frequently studied by academia and practitioners as a subject of matter is the image pre-processing step.

This is because very often the images data captured during acquisition process may contain noise or distortion originated from the devices (e.g. camera, scanner or sensor) and caused by inevitable condition (e.g. light, temperature and length of exposure) [1-3]. Further, the visual quality of the stored images may be reduced during the transmission, coding, and processing steps. These raw datasets require pre-processing step before they can be used for further analysis and convert into more meaningful and accurate results. Accordingly, image pre-processing is an important step which likely determines the overall outcome of DPI [4, 5].

Despite its pivotal position in generating enhanced image datasets suitable for analysis, image pre-processing is also a challenging task in the DPI. One of the critical task is how to remove the unwanted things of the datasets without losing of some details and essential features of the images [6, 7]. Further challenge is how to effectively carry out this de-noising function especially when dealing with parallel process, large number datasets, poor quality or other visual constraints [1, 8-10].

II. MATERIALS AND RESEARCH APPROACH

As previously mentioned, this research is aimed to provide an intelligence system that can be used to assist corn farmers in identifying disease on their crop. For this purpose, this study followed a design science research that aimed to produce an IT artefact to solve particular problem as suggested by [11] previously. Fig. 1 illustrates the overall research approach that guided this study.

Fig. 2 describes the data set used, the extraction, the data training, and testing approaches employed in this research.

A. Dataset

In his research, PlantVillage Dataset was used as data image source for training and testing. It can be downloaded from https://github.com/spMohanty/PlantVillage-Dataset that is publicly and freely available for research purpose. The PlantVillage-Dataset repository was chosen because it provides more than 8,500 images of corn leaves. The dataset was classified into three types of corn diseases including Cercospora Leaf Spot/Gray Leaf Spot, Common Rust, Northern Leaf Blight, and one category of healthy.

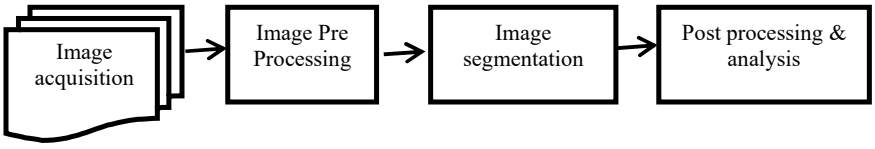


Fig. 1. DPI stages

2021 IEEE 7th Information Technology International Seminar (ITIS) | 978-1-6654-0807-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ITIS53497.2021.9791825

978-1-6654-0807-3/21/$31.00 ©2021 IEEE

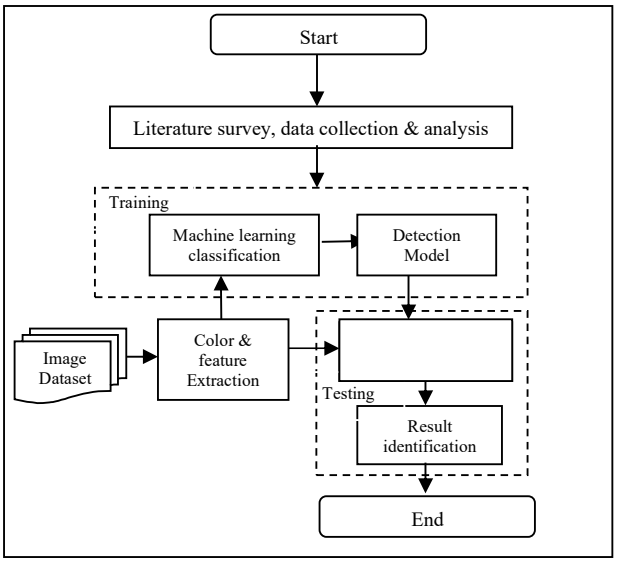


Fig. 2. Flowchart of Research Approach

B. Extraction method

As shown in the Fig. 2 above, the dataset was extracted in two ways, first color extraction and second feature extraction. Extraction is the critical point as it directly influences the accuracy of the next phase. This study employed Global Color Histogram (GCH), a popular color extraction proposed by [12] that is frequently used to image processing and computer vision. Using this GCH, all images retrieved from PlantVillage Dataset were encoded based on three global color dimensions including Red, Green, and Blue. Each color dimension was then quantified based on the nearest pixel's value. For this purpose, this study employed Color Coherence Vector (CCV) method [13] as a color descriptor to extract color-related features from the dataset that was then used in training and testing steps.

Fig. 3 shows how the color extraction process using CCV performed into three stages. First, for each of the images retrieved from the repository was blurred to eliminate random noise and then represent into pixels value of RGB color space. Second, the image dataset pixels were as if there were only three distinct color values in the image including 0, 1 and 2. Third, the color vector was grouped based on the numbers of coherent and incoherent pixels.

For the feature extraction, this study used Local binary Pattern (LBP) to help determine the texture patterns of the dataset [14] as shown in the Fig. 4. For this purpose, first, each image from the dataset was encoded into of a 3x3 matrix. The matrix represented a block size of pixels and labelled with decimal numbers, called LBP codes. The number at the center of the matrix was used as a threshold for generating the neighboring pixels value by employing the following mathematical expression (1). Second, the values generated from the equation above were arranged clockwise to form a histogram features of 2P (2^3). Finally, the encoded binary value was converted back to decimal value and used as the value of feature vectors. Once the extraction process was finished, the processed images were then used for training and testing.

Fig. 3. Color extaction process using CCV

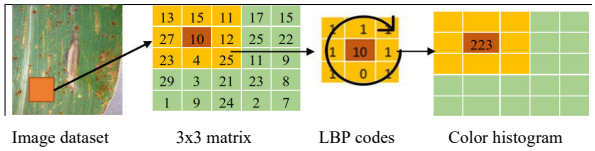


Fig. 4. Feature extraction using LBP

$$LBP = \sum_{i=0}^{P-1} s(n_i - G_c)2^i \quad (1)$$

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

P: the number of neighborhood pixels

n_i: represents the ith neighboring pixel

c: represents the center pixel

C. Training and Testing

This study used one of simple yet powerful machine learning algorithm popularly known as Regression Tree (CART) in the training stage as decision model. CART was chosen because of its ability to overcome missing data commonly occurred when working with large dataset. In particular, CART algorithm adopted in this training phase to classify the extracted dataset by building a binary tree structured rules as the basis for decision model. The decision tree model consists of a root node *(t)* derived from the dataset that had been through color and feature extraction previously and later split into two branches and leaves repeatedly. For the best attribute and threshold value, this study used Gini's impurity index *i(t)* as shown in the equation (2) where *p(wj | t)* indicated the proportion of training instance x_i allocated to class *wj* at node *t*. Each non-terminal branch is then divided into two further nodes, t_L and t_R, such that p_L, p_R are the proportions of entities passed to the new nodes t_L, t_R respectively.

$$i(t) = -\sum_{j=1}^{k} p(w_j \mid t) log p(w_j \mid t) \quad (2)$$

To obtain the best division that maximized the difference given in equation (3). According to this mathematical expression, the decision tree grows by means of the successive subdivisions until a stage is reached in which there is no significant decrease in the measure of impurity when a further additional division *s* is implemented. When this stage is reached, the node *t* is not subdivided further, and automatically becomes a terminal node. The class w*j* associated with the terminal node *t* is that which maximises the conditional probability *p(wj| t)*.

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (3)$$

The next phase was running testing upon the decision model develop during the training. This phase was the final stage that aimed to recognize and detect the type of diseases

based on dataset acquired from the earlier stages. For this purpose, this study applied ensemble classifier methods by integrating decision model, constructed previously using the CART algorithm, and a voting classifier method known as majority voting. To identify the most suitable corn leaf based disease, the performances of classification results obtained from the decision model using the CART were evaluated by majority voting with five different classifier channels as shown in Fig. 5. To check the accuracy, the results were then tested against Red pixel using the Color Processing Detection Algorithm (CPDA).

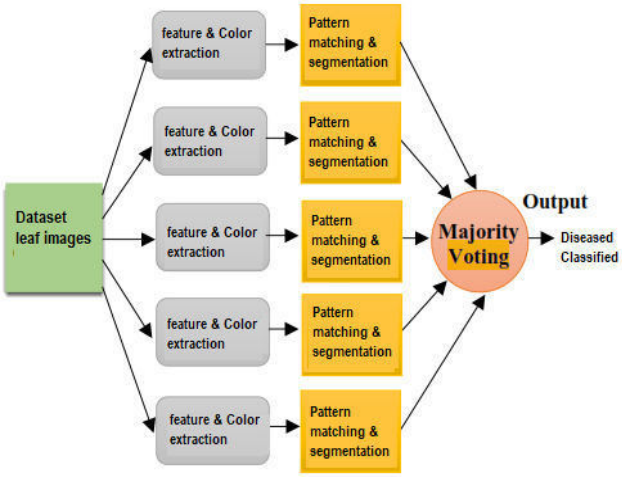


Fig. 5. The architecture of voting calssifier (majority voting)

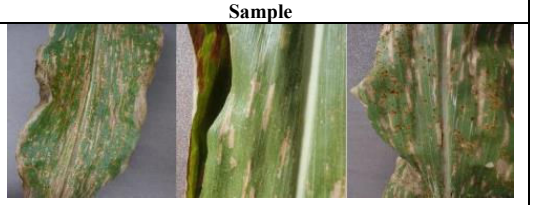
III. Results

This section reports the results of this study by applying the proposed ensemble approach. The experiment was carried out in Python 3 platform on a computer with CPU of Intel Pentium Gold G5400 3,7 GHz; 8192 RAM; NVIDIA GTX 1060 3GB Graphics Processing Unit (GPU). The next subsections discuss the results about the dataset, extraction and classification of the algorithms employed in this study.

A. Dataset and data acqusition

As mentioned above, this study benefited from an online and open access repository on crop health and disease, known as PlantVillage. This repository provides over 8500 images of healthy and infected corn leaves that have been curated and can be dowloaded through this link: https://github.com/spMohanty/PlantVillage-Dataset. All the images are provided in 256 x 256 pixels and classified into four categories: Gray Leaf Spot, Common Rust, Northern Leaf Blight, and healthy. Table I shows an example of leaf image diseases representing of disase types or healthy and the amount utilised as the dataset in this study.

TABLE I. Sample of Image Dataset

Type	Sample
Gray Leaf Spot: 1457 images	
Common Rust: 1614 images	
Northern Leaf Bligh: 958 images	
Healthy: 4450 images	

In order to develop an accurate ensemble-of-classifiers approach for corn leaf based diseases detection, each of categories were augmented. This augmentation process was also aimed to avoid common problem especially when dealing with image sources that were inconsistency in shape (some were square instead of rectangular) and often results over fitting. This study applied common data augmentation technique where the source images were rotated clockwise by a given number of degrees 90, 180, and 270. The rotated images in particular also expanded the number of data sources to improve the detection performance. Fig. 6 shows image augmentation results sample: (a) original image, (b) 90° clockwise rotation, (c) 180° clockwise rotation, (d) 270° clockwise rotation, and (e) horizontal.

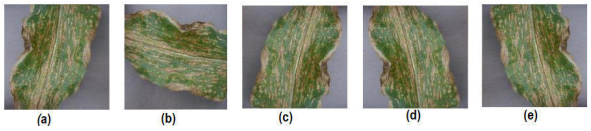


Fig. 6. Result of dataset augmentation process

B. Feature extraction

As can be seen in the image data samples (Table I), an infected crop usually has leaf with spots, rotten area or color tuning in yellowish compare to the healthy crop whose leaf is generally green and spotless. To identify those infected objects, the color feature of entire augmented image dataset need to be extracted into RGB (Red-Green-Blue) with values from 1 to 255. Finally to represent the into color histogram, the values were normalise into binary number (0 and 1). For the feature color extraction, this study used Global Color Histogram (GCH) and the result can be in the Fig. 7.

Based on Fig. 7 above, the color feature extraction using GCH can be explained as following algorithm:

1. Image acquisition
2. RGB channel extraction from the image source
3. Discretize images' colors value and store into matrix
4. Create a binary value for each color.

5. Count number of pixels for each color and store it in histogram's bins.

C. Color extraction

While GCH was able to extract the colour feature of infected leaf by counting the number of pixel of RGB channel, however, it does not capture the content and the distribution adequately. Therefore, this study also employed Local Color Histograms (LCG) to provide more information and enable the compare the distribution of color in different regions. In this study, Color Coherence Vector (CCV) technique was adopted for color extraction as shown below.

1. Image acquisition and pixels input
2. Normalisation pixels into three color spaces each of which has different between the lowest and highest values. This research quantified image color values into 0, 1 and 2 with ranges 0-9; 10-19 and 20-29 respectively
3. Labelling the pixels through the following rules:
   - Change pixel values into binary (0 and 1)
   - Scan each binary value
   - If no label found
     - → Using 4 neighbouring labelling (see Fig. 8)
       - IF *u* AND **l** = 0, THEN **p** = new label
       - IF u =1 AND has label, THEN p = u
       - IF l = 1AND has label, THEN p=l
       - IF u = 1 AND has different label THEN p = u OR l
     - → Using 8 neighbouring labelling (see Fig. 9)
       - IF lr, lu, u AND ru = 0 THEN p = new label
       - IF lr =1 AND has label, THEN p = lr
       - IF lr= 1AND has label, THEN p=lu
       - IF u = 1 AND has label THEN p = u
       - IF ru=1 AND has label THEN p= ru
       - IF lr, lu, u OR ru = 1 AND has different label AND more than 1 lable THEN p = lr, lu, u OR ru
   - Repeat until all pixel has label

D. Training and Testing

Following the feature and color extraction of all image sources from the dataset, the next step was running training and testing. This step was aimed to develop a detection model that was able to learn the texture patterns of image sources and recognise the type of diseases closest in accuracy with the dataset and then tested the accuracy. This study employed two classification methods to develop the detection model: Voting Classifier with Majority Voting and 5 Classification and Regression Tree (CART) with max depth 2, 4, 6, 8, and 10. Whereas for the testing, this study used k-fold Cross Validation as suggested by [15]. Finally to determine the performance of the model, this study used F1-Score.

A series of seven scenarios for training and testing were proposed and conducted using 3,000 for each dataset category. All the image sources had been augmented and gone through features and color extraction. Using K-Fold Cross Validation techniques with 5 folds, each category contained 2,400 images for training process and 600 images for testing process. Table II shows results of evaluation from seven scenarios.

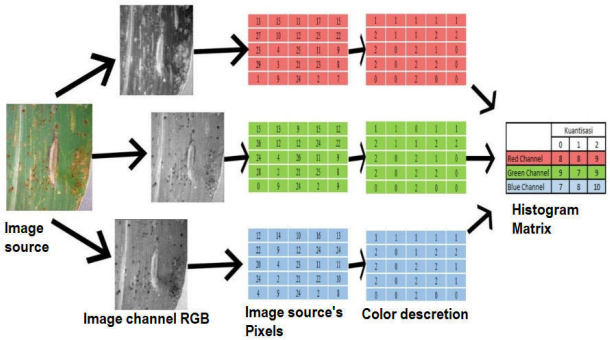


Fig. 7. Feature color extraction process.

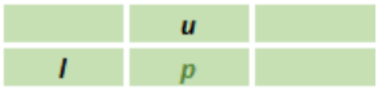


Fig. 8. Labelling 4 neighbouring

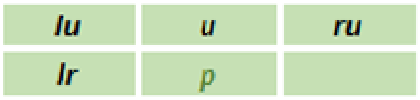


Fig. 9. Labelling 8 neighbouring

TABLE II. RESULTS OF EVALUATION

Scenario	*Precision (%)*	*Recall (%)*	*f1 - Score (%)*
1	78.33	76.32	75.09
2	73.87	72.51	71.17
3	70.16	66.35	65.69
4	79.08	76.61	75.94
5	77.73	77.32	77.24
6	82.92	82.55	82.6
7	76.03	76.28	74.52

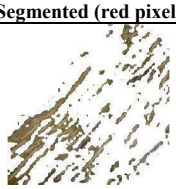
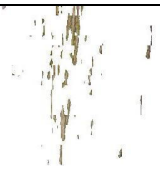
It can be seen from the results summarised in Table II, scenario 6 showed the optimum result with the accuracy 82.92%. This result has a significant positive correlation with the accuracy rate of each category as shown in Table III. Based on the results shown in Table III, scenario 6 has the highest average of accuracy among other scenarios.

TABLE III. RESULTS OF EVALUATION

Scenario	Gray Leaf Spot (%)	Common Rust (%)	Northern Leaf Blight (%)	Healthy (%)
1	45,94	98,42	88,42	67,2
2	43,44	91,68	82,4	67,14
3	51,66	94,14	71,08	45,88
4	50,9	98.36	86,48	68
5	58,14	87,7	98,1	65,2
6	67,92	94,22	99,66	68,78
7	41,7	99,08	88,7	68,6

Further analysis revealed all these results likely have strong correlation with the red pixel segmentation on the image data sources prior to training and evaluation. In the other words, the proposed classification model has resulted more accurate detection if the image sources are firstly segmented especially using red pixel segmentation. Table IV shows the results disease identification on scenario 6 using red pixel segmentation. As shown in Table IV, the results indicated 4 positive identification results out of 6 attempts of testing.

TABLE IV. RESULTS OF IDENTIFICATION USING RED PIXEL SEGEMENTED

Input	Segmented (red pixel)	Result
		IdentificationClass 0 Target = Class 0 Prediction = Class 0 **True Positive**
		IdentificationClass 0 Target = Class 0 Prediction = Class 0 **True Positive**
		IdentificationClass 0 Target = Class 0 Prediction = Class 3 **False Negative**
		IdentificationClass 0 Target = Class 0 Prediction = Class 3 **False Negative**
		IdentificationClass 3 Target = Class 3 Prediction = Class 0 **False Positve**
		IdentificationClass 3 Target = Class 3 Prediction = Class 0 **False Positvie**

IV. CONCLUSION

Corn is one of the most common food crops in tropical regions such as Indonesia. There are several factors that can affect the yield of maize crops, such as pests, diseases in plants, to weeds. This factor is usually referred to as Plant Pest Organisms (OPT). The identification of pest attacks has several obstacles in its application in the agricultural world, from the lack of knowledge of the type of pest, to the extent of agricultural land that is not proportional to the number of experts who are experts in their fields. So technology is needed that can help farmers identify pests. In this study, digital image processing will be used to identify maize leaves that are attacked by pests. The proposed method is the extraction of color and texture features using the Global color histogram (GCH), Color Coherence Vector (CCV), and Local Binary Pattern (LBP) combined with the Voting Classifier containing five Classification and Regression Trees (CART) with differences in the Depth or CART level 2,4,6,8 and 10. The dataset used in this research is the PlantVillage Data set which contains images of corn leaves. Test results of segmented image against Red pixel using the Color Processing Detection Algorithm (CPDA) gets the best accuracy with a Precision of 82.92%, Recall 82.55%, and an f1-Score of 82.6%

REFERENCES

[1] L. Fan, F. Zhang, H. Fan, and C. Zhang, "Brief review of image denoising techniques," *Visual Computing for Industry, Biomedicine, and Art,* vol. 2, p. 7, 2019/07/08 2019.

[2] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation,* vol. 4, pp. 490-530, 2005.

[3] J. G. A. Barbedo, "Digital image processing techniques for detecting, quantifying and classifying plant diseases," *SpringerPlus,* vol. 2, p. 660, 2013.

[4] Y. Kim, D. Lee, and D. Kim, "Pre-Processing Images for Enhancing Reliability in Screen-to-Camera Communication," *IEEE Wireless Communications Letters,* vol. 7, pp. 934-937, 2018.

[5] S. Shang, M. Li, Y. Hou, L. Chen, Y. Yang, and J. Sun, "A novel method of ISAR image pre-processing for ship," in *2016 IEEE International Conference on Electronic Information and Communication Technology (ICEICT)*, 2016, pp. 262-265.

[6] M. Diwakar and M. Kumar, "A review on CT image noise and its denoising," *Biomedical Signal Processing and Control,* vol. 42, pp. 73-88, 2018/04/01/ 2018.

[7] O. Berezsky, O. Pitsun, B. Derish, K. Berezska, G. Melnyk, and Y. Batko, "Adaptive Immunohistochemical Image Pre-processing Method," in *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, 2020, pp. 820-823.

[8] S. Ramani and J. A. Fessler, "Parallel MR image reconstruction using augmented Lagrangian methods," *IEEE transactions on medical imaging,* vol. 30, pp. 694-706, 2011.

[9] B.-S. Kim and J.-U. Kim, "Design and Implementation of a Boundary Matching System Supporting Partial Denoising for Large Image Databases," *Journal of the Korea Society of Computer and Information,* vol. 24, pp. 35-40, 2019.

[10] H. Li, F. He, Y. Liang, and Q. Quan, "A dividing-based many-objective evolutionary algorithm for large-scale feature selection," *Soft Computing,* pp. 1-20, 2019.

[11] W. Agustiono, "Integrated Public Transportation Systems Model for Passengers' Convenience and Safety," in *2020 6th Information Technology International Seminar (ITIS)*, 2020, pp. 243-248.

[12] S. Wang, "A robust CBIR approach using local color histograms," Department of Computer Science, University of Alberta2001.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*: Routledge, 2017.

[14] M. A. Khan, M. I. U. Lali, M. Sharif, K. Javed, K. Aurangzeb, S. I. Haider, *et al.*, "An optimized method for segmentation and classification of apple diseases based on strong correlation and genetic algorithm based feature selection," *IEEE Access,* vol. 7, pp. 46261-46277, 2019.

[15] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition,* vol. 48, pp. 2839-2846, 2015.