e-ISSN: 2614-8404
p-ISSN:2776-323

Application of Feature Selection for Identification of Cucumber Leaf Diseases (*Cucumis sativa L.*)

Lalitya Nindita Sahenda[1*)], Ahmad Aris Ubaidillah[2], Zilvanhisna Emka Fitri[3], Abdul Madjid[4], Arizal Mujibtamala Nanda Imron[5]
[1]Program Studi Teknik Komputer, Jurusan Teknologi Informasi, Politeknik Negeri Jember
[2,3]Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Jember
[4]Program Studi Budidaya Tanaman Perkebunan, Jurusan Produksi Pertanian, Politeknik Negeri Jember
[4]Program Studi Teknik Elektro, Fakultas Teknik, Universitas Jember
Email: [1]lalitya.ns@polije.ac.id, [2]arisubay06@gmail.com, [3]zilvanhisnaef@polije.ac.id, [4]abdul_madjid@polije.ac.id, [5]arizal.tamala@unej.ac.id

Abstract – According to data from BPS Kabupaten Jember, the amount of cucumber production fluctuated from 2013 to 2017. Some literature also mentions that one of the causes of the amount of cucumber production is disease attacks on these plants. Most of the cucumber plant diseases found in the leaf area such as downy mildew and powdery mildew which are both caused by fungi (fungal diseases). So far, farmers check cucumber plant diseases manually, so there is a lack of accuracy in determining cucumber plant diseases. To help farmers, a computer vision system that is able to identify cucumber diseases automatically will have an impact on the speed and accuracy of handling cucumber plant diseases. This research used 90 training data consisting of 30 healthy leaf data, 30 powdery mildew leaf data and 30 downy mildew leaf data. while for the test data as many as 30 data consisting of 10 data in each class. To get suitable parameters, a feature selection process is carried out on color features and texture features so that suitable parameters are obtained, namely: red color features, texture features consisting of contrast, Inverse Different Moment (IDM) and correlation. The K-Nearest Neighbor classification method is able to classify diseases on cucumber leaves (Cucumis sativa L.) with a training accuracy of 90% and a test accuracy of 76.67% using a variation of the value of K = 7.

Keywords – Identification, Cucumber Leaf, Disease, Color Feature, GLCM, KNN

I. INTRODUCTION

Cucumber (Cucumis sativa L.) is one of the most widely consumed fruit vegetables and the growing conditions are very flexible, because it can grow in both highlands and lowlands [1]. In Indonesia, the amount of cucumber production fluctuated (instability) from 2013 to 2017. Total cucumber production was 9.97 tons/ha in 2013, then decreased to 9.84 tons/ha in 2014 and increased to 10.27 tons/ha in 2015. In 2016, production decreased again to 10.19 tons/ha and increased again to 10.67 tons/ha in 2017 [2]. One of the causes of the fluctuating amount of cucumber production is the attack of pests and diseases on the plant. Symptoms of disease in plants can be seen from the plant body parts, such as leaves, fruit, stems and roots. Most of the cucumber plant diseases that are found in the leaf area such as antracnose lesion [3], downy mildew and powdery mildew are usually caused by fungi (fungal disease) [4].

So far, farmers have checked for cucumber plant diseases, which are still done manually based on the experience of farmers. This determination certainly has weaknesses, one of which is the lack of accuracy in determining cucumber plant diseases. To help farmers, a cucumber leaf disease identification system was created using computer vision. Computer vision is a branch of science where a system utilizes digital image processing techniques which are then analyzed using artificial intelligence.

Some studies that become the reference of this research are cucumber disease detection using diagnosis of diseases of cucumber through extracting three characteristic values of shape, texture and color [5][3], Then the research was developed by adding image smoothing and edge detection so that the segmentation results were better[6]. The feature extraction technique used is first-order statistical features and second-order statistical features such as Gray Level Co-Occurrence Matrix (GLCM) to get an accuracy of 80.45% [7], then developed the introduction of cucumber disease using sparse representation classification with an accuracy rate of 85.7% using the classification method used in this study is K-Nearest Neighbor [8]. The KNN method is also able to classify other research objects such as the classification of platelets on peripheral blood smears with an accuracy of 83.67% [9], on the classification of tuberculosis bacteria with an accuracy of 94.92% [10], Classification of white blood cell abnormalities based on shape features (area, perimeter, metric and compactness) with an accuracy rate of 94.3% [11] and classification of bacteria that cause ARI with an accuracy of 91.67% using a variation of the value of K = 3.5 and 7 [12].

To classify data, the KNN method uses the closest distance to the object so that the method is often known as lazy learning. The basic principle of KNN is to find the value of K where the value of K is the closest amount of data that will determine the classification results and to calculate the closest distance using Euclidean distance calculations. Based on this description, the K-Nearest Neighbor (KNN) method is able to classify cucumber leaf diseases with a good level of accuracy.

II. RESEARCH METHODOLOGY

This research was conducted in a cucumber field in Lumajang Regency, East Java. Data collection is done with the help of direct sunlight, so the resulting image is very bright and does not cause shadows (noise). This will affect the results in the image processing process. The system flow consists of starting from the image of cucumber leaves, then the image is processed using digital image

e-ISSN: 2614-8404
p-ISSN:2776-323

processing techniques, the result of feature extraction becomes the input of the KNN classification method which will be classified into 3 classes as shown in Figure 1.

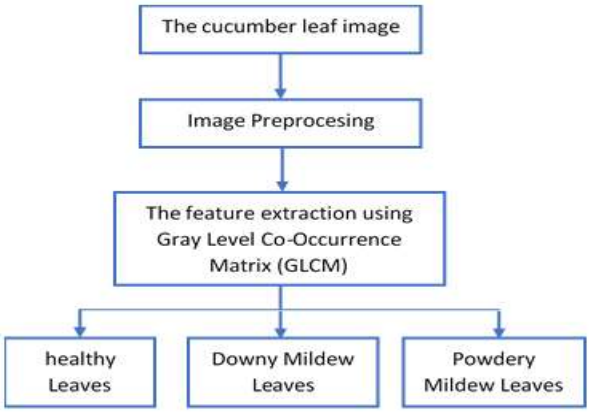


Figure 1. The Cucumber Leaf Identification System Diagram

A. *The Cucumber Leaf Images Data*

Image data was taken using a Canon EOS 1100D camera with a camera resolution of 12 MP, using a tripod and studio light box. The data is divided into 3 classes, namely healthy leaf images, downy mildew leaf images and powdery mildew leaf images as shown in Figure 2. This research used 90 training data consisting of 30 healthy leaf data, 30 powdery mildew leaf data and 30 downy mildew leaf data. while for the test data as many as 30 data consisting of 10 data in each class.

(a) **(b)** **(c)**

Figure 2. Image of cucumber leaves (a) healthy, (b) downy mildew and (c) powdery mildew

B. *Image Processing*

The image processing process is the initial stage of the image processing process which aims to improve image quality and the data normalization process. This process begins with a cropping process whose aim is to obtain a smaller image size to reduce the computational load [13]. The initial image size of 4272 x 2848 pixels is cropped to 1001 x 1001 pixels as shown in Figure 3.

(a) **(b)**

Figure 3. Image (a) original and (b) after cropping process

C. *Feature Extraction*

Feature extraction aims to extract the unique characteristics of the cucumber leaf image. Before performing feature extraction, the original image is an RGB color space image, then the process of splitting the color components into red, green and blue as shown in Figure 4.

In this study, two features were used, namely color and texture. The color features used are red, green and blue color features, while the texture features used are texture features from the gray level co-occurrence matrix (GLCM) method.

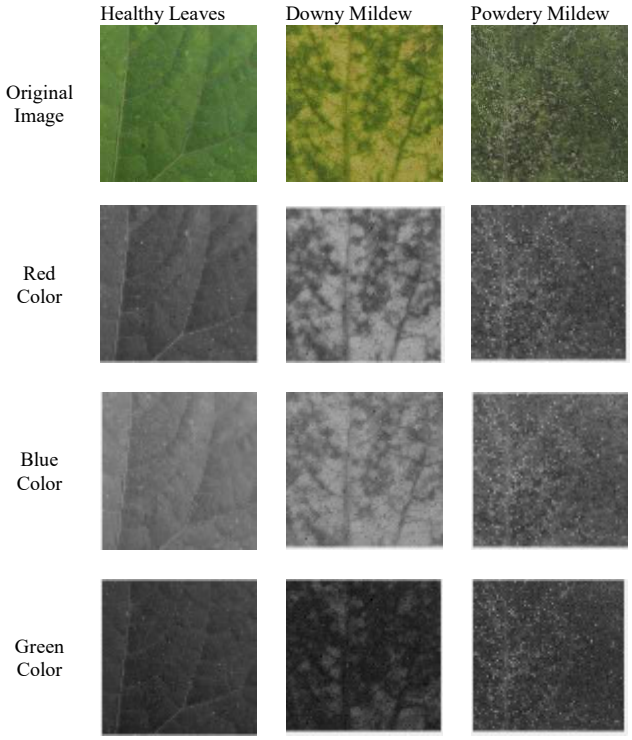


Figure 4. Image of the RGB components based on each classes

GLCM is a gray degree matrix that represents the frequency of occurrence of two pixels with a certain intensity in a distance d and a certain angle direction θ. Therefore, the matrix provides information that differs from the difference in distance between pixels [14]. The angles used are 0°, 30°, 45°, 90° and 135°. The formula equation used is as follows [15]:

$$\text{ASM} = \sum_{i=1}^{L}\sum_{j=1}^{L}(\text{GLCM}(i,j))^2 \quad (1)$$

$$\text{contrast} = \sum_{i}^{L}\sum_{j}^{L}|i-j|^2\text{GLCM}(i,j) \quad (2)$$

$$\text{IDM} = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{(\text{GLCM}(i,j))^2}{1+(i-j)^2} \quad (3)$$

$$\text{Entropy} = -\sum_{i=1}^{L}\sum_{j=1}^{L}(\text{GLCM}(i,j))\log(\text{GLCM}(i,j)) \quad (4)$$

$$\text{Correlation} = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{(i-\mu i')(i-\mu j')(\text{GLCM}(i,j))}{\sigma i \sigma j} \quad (5)$$

D. *K-Nearest Neighbor Classification*

One of the easy classification methods is the K-Nearest Neighbor classification method. This method has the basic

e-ISSN: 2614-8404
p-ISSN:2776-323

principle of looking for a constant K value where the K value is the number of closest distances that affect the classification results. The calculation of distance using Euclidean distance calculation with the equation [12] :

$$d_{(xi,xj)} = \sqrt{\sum_{r=1}^{n} (Xir - Xij)^2} \quad (6)$$

III. RESULTS AND DISCUSSION

In this research, the results of the cropping process are taken for the color features of each component of the RGB color space as shown in Figure 4. The image shows that in the blue and green images there is no significant difference in the classes: healthy leaves, downy mildew and powdery mildew. On the other hand, the red image shows a significant difference, especially in the powdery mildew leaf class, so that the red image is the best in representing the textures of the three classes. In the image of the red component, the gray level values on the downy mildew and powdery mildew leaves are clearly visible so that the red image becomes the input for texture feature extraction based on the Gray Level Co-Occurrence Matrix (GLCM) value.

A. Color Feature Extraction

In this research, color features were taken so that each class was as shown in Table 1. The table shows that there is a closeness of values for the Blue features of the Healthy leaf class and the downy mildew leaf class. Whereas in the Green feature, there is also a closeness of values in the Healthy leaf class and the powdery mildew leaf class, so that the color feature used is the red image feature.

Table 1. The Average Value of The RGB Color Feature in Each Class

Class	*Red*	*Green*	*Blue*
Healthy	89.29	110.83	39.80
Downy mildew	118.84	125.34	39.64
Powdery mildew	98.27	111.81	57.44

B. Texture Feature Extraction

This research also takes texture features using GLCM texture features. To take the GLCM feature, a red image is used as the input image, as previously explained that the red image best represents the texture in the three classes. The average value of texture features is shown in Table 3, where the features used are Angular Second Moment (ASM), Contrast, Inverse Different Moment (IDM), Entropy and Correlation. Table 3 shows that there is a closeness of values for the ASM features of the three classes, while for the entropy features there is also a closeness of values for the downy mildew and powdery mildew class.

Table 2. The Average Value of GLCM Features in Each Class

Class	ASM	Contrast	IDM	Entropy	Correlation
Healthy	0.0068	4.8496	0.6109	5.6273	0.0076
Downy mildew	0.0026	5.5866	0.5265	6.4159	0.0024
Powdery mildew	0.0030	21.5099	0.3883	6.6609	0.0044

C. Feature Selection

Based on the discussion on color feature extraction and texture feature extraction, it was found that some features have close values and this will affect the classification process. There will be errors in the classification process due to the proximity of these values, so the researcher conducts a feature selection process and the features used are red color features, contrast texture features, IDM, and Correlation. An example of selecting a red feature is shown in Figure 5.

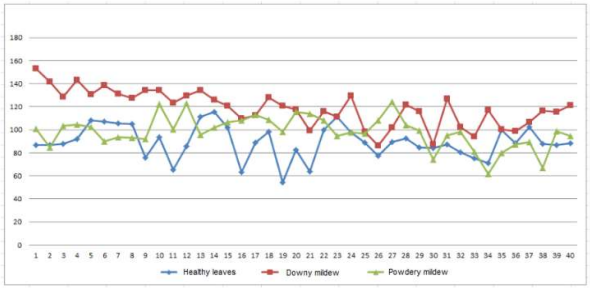
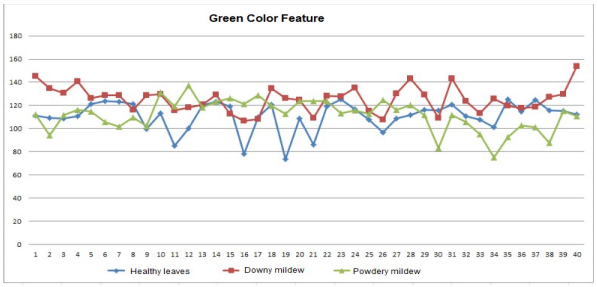
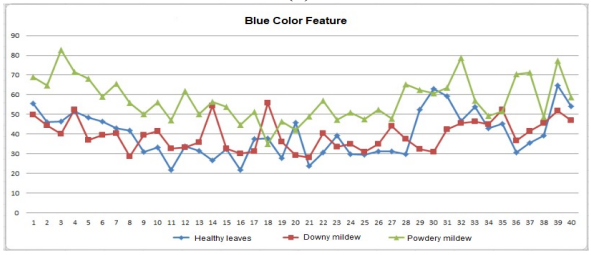


Figure 5. Graph of data distribution in each class on the red feature

Figure 5 shows that the blue color describes the distribution of healthy leaf data, the red color describes the distribution of downy mildew data and the green color describes the distribution of powdery mildew data. The graph will be different when compared to the green and blue color features as shown in Figure 6.



(a)

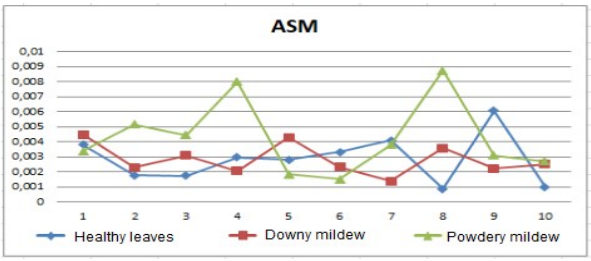


(b)

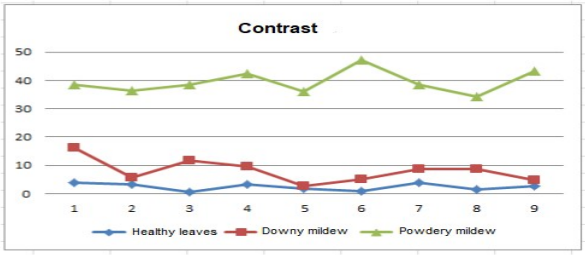
Figure 6. Graph of data distribution in each class on (a) green and (b) blue color features

On the graph (Figure 6), there is still a lot of data that occurs in the values between the three classes so that the graphs of the three classes intersect. While the GLCM features a graph depicting the value of each feature in the three classes (Figure 7).

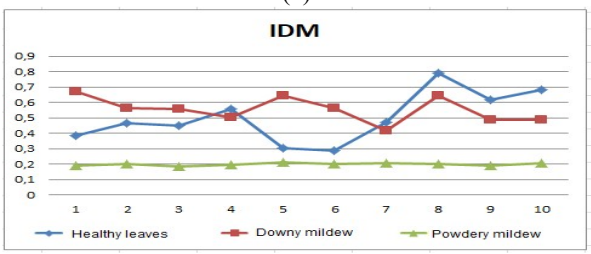
e-ISSN: 2614-8404
p-ISSN:2776-323



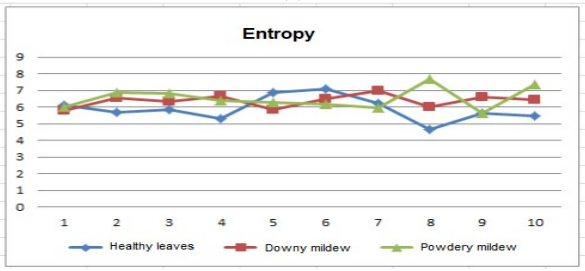
(a)



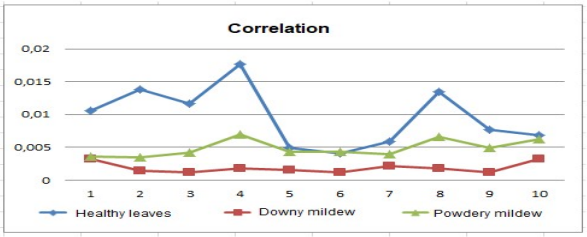
(b)



(c)



(d)



(e)

Figure 7. Graph of data distribution in each class on (a) ASM, (b) contrast, (c) IDM, (d) entropy and (e) correlation features

On Figure 7 shows that in features (a) ASM and (d) entropy features, there are many data that have values that are tangent to each other in the three classes. this affects the accuracy of the system in classifying data so that the two features become features that will reduce the system in classifying the three classes. Whereas in features (b) contrast, (c) IDM and (e) correlation, there is a difference in value even though there are features whose values intersect on healthy leaf data with downy mildew data or on healthy leaf data with powdery mildew data. Based on the results of feature selection, this study used 4 features consisting of color features and texture features, namely red color features, texture features (contrast, IDM and correlation) as shown in Table 3.

Table 3. The Examples of Feature Selection Results for Each Class

Feature	Class			
	Red	**Contrast**	**IDM**	**Correlation**
Healthy Leaves	86.71	8.7433	0.3836	0.0105
	86.94	5.2779	0.4650	0.0138
	87.96	5.4021	0.4503	0.0116
Downy mildew	153.04	0.6716	0.0030	2.3340
	141.83	0.5647	0.0014	4.5463
	128.70	0.5571	0.0027	4.3620
Powdery Mildew	93.70	0.4850	0.0069	6.9519
	92.82	0.2257	0.0021	7.5547
	92.22	0.6759	0.0021	5.4755

D. K-Nearest Neighbor Classification

The number of training data is 90 cucumber leaf image data consisting of 30 healthy leaf data, 30 powdery mildew leaf data, 30 downy mildew leaf data. While 30 test data consisting of 10 data in each class. The results of the accuracy of the KNN classification method. Based on Table 4, the highest accuracy of system training is 100% at the variation of the value of K = 1. However, the accuracy of the test is 66.77%, so that judging from the results of the highest test accuracy, it is 76.67% with a training accuracy of 90% at the variation of the value of K = 7.

Table 4. The Percentage of System Training and Testing Accuracy

K Value	Training Accuracy (%)	Testing Accuracy (%)
1	100	66.67
3	96.67	73.33
5	93.33	70
7	90	76.67
9	88.89	73.33
11	90	70
13	90	66.67
15	87.78	66.67

The calculation of system accuracy is obtained based on ROC calculations with a confusion matrix as shown in Table 5 for the training process and Table 6 for the testing process.

Table 5. The Confusion Matrix in The Training Process

Classification Results			Target
Healthy	Downy Mildew	Powdery Mildew	
24	0	6	Healthy
0	30	0	Downy Mildew
3	0	27	Powdery Mildew

$$Accuracy\ of\ training = \frac{24 + 30 + 27}{24 + 30 + 27 + 3 + 6} x100\%$$

$$= 90\%$$

e-ISSN: 2614-8404
p-ISSN:2776-323

Table 6. The Confusion Matrix in The Testing Process

Classification Results			Target
Healthy	Downy Mildew	Powdery Mildew	
8	0	2	Healthy
0	10	0	Downy Mildew
4	1	5	Powdery Mildew

$$Accuracy\ of testing = \frac{8+10+5}{8+10+5+2+4+1} x100\%$$

$$= 76.67\%$$

Based on the results of these calculations, it can be seen that there is a significant difference in accuracy in the training and testing process. The system training accuracy rate is 90% while the system testing accuracy is 76.67%. This can happen due to the lack of data used or there is a significant difference in the value of the training data and testing data.

IV. CONCLUSION

The K-Nearest Neighbor (KNN) classification method is able to classify diseases on cucumber leaves (Cucumis sativa L.) with a training accuracy of 90% and a test accuracy of 76.67% using a variation of the value of K = 7. The lack of data also affects the classification results because the system has limitations in recognizing patterns from healthy leaf classes, downy mildew and powdery mildew. In addition, this research must also compare other classification methods.

REFERENCES

[1] 1. KL, A., Napitupulu, M., & Jannah, N. (2015). RESPON TANAMAN MENTIMUN (Cucumis sativus L.) TERHADAP JENIS POC DAN KONSENTRASI YANG BERBEDA. *AGRIFOR*, *XIV*(1), 15–26.

[2] BPS Kabupaten Jember. (2020). *Kecamatan Arjasa Dalam Angka Tahun 2020*.

[3] Pixia, D., & Xiangdong, W. (2013). Recognition of Greenhouse Cucumber Disease Based on Image Processing Technology. *Open Journal of Applied Sciences*, *03*(01), 27–31. https://doi.org/10.4236/ojapps.2013.31b006

[4] Pawar, P., Turkar, V., & Patil, P. (2016). Cucumber disease detection using artificial neural network. *Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016*, *2016*. https://doi.org/10.1109/INVENTIVE.2016.7830151

[5] Wei, Y., Chang, R., Wang, Y., Liu, H., Du, Y., Xu, J., & Yang, L. (2012). A study of image processing on identifying cucumber disease. *IFIP Advances in Information and Communication Technology*, *370 AICT*(PART 3), 201–209. https://doi.org/10.1007/978-3-642-27275-2_22

[6] Jiannan, J., & Haiyan, J. (2013). Recognition for cucumber disease based on leaf spot shape and neural network. *Transactions of the Chinese Society of Agricultural Engineering*, *29*(1).

[7] Kaur, S., Pandey, S., & Goel, S. (2019). Plants Disease Identification and Classification Through Leaf Images: A Survey. *Archives of Computational Methods in Engineering*, *26*(2), 507–530. https://doi.org/10.1007/s11831-018-9255-6

[8] Khan, M. A., Akram, T., Sharif, M., Javed, K., Raza, M., & Saba, T. (2020). An automated system for cucumber leaf diseased spot detection and classification using improved saliency method and deep features selection. *Multimedia Tools and Applications*, *79*(25–26), 18627–18656. https://doi.org/10.1007/s11042-020-08726-8

[9] Fitri, Z. E., Purnama, I. K. E., Pramunanto, E., & Purnomo, M. H. (2017). A comparison of platelets classification from digitalization microscopic peripheral blood smear. *2017 International Seminar on Intelligent Technology and Its Application: Strengthening the Link Between University Research and Industry to Support ASEAN Energy Sector, ISITIA 2017 - Proceeding*, *2017-Janua*, 356–361. https://doi.org/10.1109/ISITIA.2017.8124109

[10] Sahenda, L. N., Pumomo, M. H., Purnama, I. K. E., & Wisana, I. D. G. H. (2018). Comparison of Tuberculosis Bacteria Classification from Digital Image of Sputum Smears. *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2018 - Proceeding*, 20–24. https://doi.org/10.1109/CENIM.2018.8711386

[11] Fitri, Z. E., Syahputri, L. N. Y., & Imron, A. M. N. (2020). Classification of White Blood Cell

e-ISSN: 2614-8404
p-ISSN:2776-323

Abnormalities for Early Detection of Myeloproliferative Neoplasms Syndrome Based on K-Nearest Neighborr. *Scientific Journal of Informatics*, *7*(1), 136–142. https://doi.org/10.15294/sji.v7i1.24372

[12] Fitri, Z. E., Sahenda, L. N., Puspitasari, P. S. D., Destarianto, P., Rukmi, D. L., & Imron, A. M. N. (2021). The Classification of Acute Respiratory Infection (ARI) Bacteria Based on K-Nearest Neighbor. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *12*(2), 91–101.

[13] Fitri, Z. E., Baskara, A., Silvia, M., Madjid, A., & Imron, A. M. N. (2021). Application of backpropagation method for quality sorting classification system on white dragon fruit (Hylocereus undatus). *IOP Conf. Series: Earth and Environmental Science*, *672*(IT Agriculture), 1–6. https://doi.org/10.1088/1755-1315/672/1/012085

[14] Fitri, Z. E., Nuhanatika, U., Madjid, A., & Imron, A. M. N. (2020). Penentuan Tingkat Kematangan Cabe Rawit (Capsicum frutescens L.) Berdasarkan Gray Level Co-Occurrence Matrix. *Jurnal Teknologi Informasi Dan Terapan*, *7*(1), 1–5. https://doi.org/10.25047/jtit.v7i1.121

[15] Fitri, Z. E., Rizkiyah, R., Madjid, A., & Imron, A. M. N. (2020). Penerapan Neural Network untuk Klasifkasi Kerusakan Mutu Tomat. *Jurnal Rekayasa Elektrika*, *16*(1), 44–49. https://doi.org/10.17529/jre.v16i1.15535