Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 3, No. 2, September 2020, pp. 86 – 94
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Data Train Reduction on Data Image With K Support Vector Nearest Neighbor (Case Study: Maize Leaf Image)

[1]Marlinda Vasty Overbeek, [2]Yampi R. Kaesmetan
[1]Study Program Informatics, Faculty of Techniques and Informatics, Universitas Multimedia Nusantara Tangerang
[2]Study Program Informatics Technique, STIKOM Uyelindo Kupang
Email: [1]marlinda.vasty@umn.ac.id, [2]kaesmetanyampi@gmail.com

Article Info
Article history:
Received Aug 07th, 2020
Revised Aug 21th, 2020
Accepted Sep 02th, 2020

Keyword:
Data Train Reduction
Data Image
K Support Vector Nearest Neighbor
Maize Leaf Image

ABSTRACT
In this study, we applied the K Support Vector Nearest Neighbor algorithm to reduce data train on data image. The data image that we used is the maize leaves image infected with fungi and healthy maize leave. The aim of data train reduction in this study is to get faster and more accurate prediction results. This because by using the K Support Vector Nearest Neighbor algorithm, a support vector which is formed from the algorithm really characterize the objective function of the problem. The accuracy obtained from this study is 0.20 or 20% mean error for the value of nearest neighbor K = 3 and using K Nearest Neighbor as a model construction algorithm. The error value is smaller than when we compared to the construction the model without performing data train reduction. The error value if not doing any reduction is 0.209 or 20.9%. Whereas in terms of time efficiency, working with the K Support Vector Nearest algorithm is 24 seconds faster than without performing data train reduction.

Corresponding Author:
Marlinda Vasty Overbeek
Study Program Informatics, Faculty of Techniques and Informatics, Universitas Multimedia Nusantara Tangerang
Jl. Scientia Boulevard, Gading Serpong, Tangerang Banten
Email: marlinda.vasty@umn.ac.id
DOI: http://dx.doi.org/10.24014/ijaidm.v3i2.10451

1. INTRODUCTION

In classification cases, data train is always needed to form the teacher or model so that when there is new data with no class or target, the data can be mapped to a proper class. To get a good representation of learning, which in this case is data train, researchers often have difficulty getting the right composition, and in classification-based learning requires 'lots' of data train - which is often unpredictable how much the number of data needed to form a good model [1]. To provide solutions to these problems, we need a technique that can provide a good representation of the data used to obtain data train that truly represents the problem being worked on.

In this study, data image is used, we used the image of maize leaves affected by fungi. The data image is used because in the detection of disease in maize leaves there are similarities in the characteristics or symptoms that are seen so that if distinguished by looking directly, it is quite difficult to do [2; 3]. The characteristics of maize leaves which are affected by fungi are characterized by brown spots and look similar to one disease with another [2]. Diseases of maize leaves discussed in this study were northern leaf blight, southern leaf blight, and southern rust.

In a previous study [3] already used Multiclass Support Vector Machine with the Radial Base kernel function and using the Sobel operator as a feature extraction. The research resulted in an average accuracy of 92,225% for classification. However, there are deficiencies in the study, namely the image data used are only 140 images. There are selected and the data proportion is balance for one class to another class. Beside that, the Multiclass Support Vector Machine, even is a robust algorithm, but the structure of the algorithm is very complex and takes time to do when train a model [4].

For that, in the classification of maize leaves image to find out whether the maize is affected by a disease caused by fungi, we need a predict or recognition system. This recognition system was built

mathematically to reduce the recognition error made by humans (*false positive error*) or we called subjective classification by eyes [5]. Because of this, the development of the model as a supervision in the recognition system requires data that actually represents the data. The selection of the data can be done directly by using data segmentation and extraction on the leaves image to get the best leaf characteristics before classification or recognition of the leaves affected by disease or not [6], or data reduction is first done to be used as data train before segmentation and extraction the features.

Reduction of data train is done because the data entered initially is data that has a proportion of irrelevant and redundant features so that it can reduce the accuracy of the learning model [7; 8]. Data train reduction, actually using active learning to choose informative information from the dataset which is existing[9].

By using the data train reduction, the accuracy and time speed in building the model is faster. Data that has passed through the preprocess then is fold into data train and data test, then the data train is reduced using an algorithm to get a faster and accurately model build. The next step is the same as the research techniques described in the study for the detection of diseases in plant leaves proposed by [6].

Data train reduction has been carried out in several studies. Such as, the researchers using Principal Component Analysis (PCA) techniques [10-15], data reduction with Artificial Neural Network algorithms [16-19], Genetic algorithms [20], Decision Tree [21], Support Vector Machine [23 ; 23], Naïve Bayes Classifier [24], Deep Learning Model [8; 25], and with Instance Based Learning [26-28].

Instance Based Learning(IBL) is an algorithm that uses data as an example for constructing a model by using the dissimilarity or distance from each data train to data that does not yet have a class or target [29]. One of the IBL technique that is often used is the Nearest Neighbor algorithm [30]. The Nearest Neighbor algorithm or sometimes referred to as the K Nearest Neighbor (KNN) has the advantage of the modeling complex objective functions with a number of local complex number estimation and stored information as data train is never lost, it is stored in the memory [29]. Because of this, there is a method developed from the KNN algorithm called K Support Vector Nearest Neighbor (KSVNN) proposed by Prasetyo [28] to reduce data train based on scored and properties of the significance degree of each data train based on nearest neighbor principles.

After the data train has been reduced, the introduction of the recognition system is done using the KNN algorithm. KNN algorithm is considered good in recognize the disease classification system in leaf images because it gives an accuracy of more than 90% [31-35].

From the previous research, in our study, before classifying the disease on the maize leaf, data train was reduced. The data used were maize leaf image data with a size of 258 x 258 pixels with the division of healthy leaves as many as 1162 images, leaves with southern rust as many as 1192 images, leaves with southern leaf blight as many as 508 images, and leaves with northern leaf blight as many as 985 images. Distribution of data train and test is 90% and 10%. It is expected that from the reduction of data train, the recognition system will be faster in classifying diseases of maize leaves caused by fungi.

2. RESEARCH METHOD

The framework in this study is the step taken to solve the problem. The details of the research procedure are shown in Figure 2 which was developed from research Kaur et al [6].
From Figure 1, the work details are as follows:

a. Data acquisition
   The data used are as much as 3847 image data and the type of the data is unbalanced data in proportion to the distribution of images of healthy maize leaves, southern leaf blight, northern leaf blight, and southern rust. Next image is the appearance of the leaf image (Figure 1) which is used for 4 different classes.

Northern leaf blight Southern leaf blight Southern rust Healthy leaf

Figure 1. Acquisition of maize leaf image

b. Pre Process

Pre-processing in this research is to color conversion from the Red Green Blue (RGB) channel to the grayscale channel and enhance the quality of the image by remove off noise from the image with the Median Filter technique with a 3x3 kernel.

Median filter is a non-linear filter that is used to produce an enhancement image [38]. The concept of the median filter is to smooth and reduce noise in the image. Noise filter contains a number of odd pixels that are shifted at a point-by-point point in the entire image area.

c. Segmentation

The segmentation used for this research is shape segmentation with Sobel Operator. It is mimicking the function of the gradient intensity of the image and giving results in the form of important things from the edge of the image and the contours of the image being clarified [36]

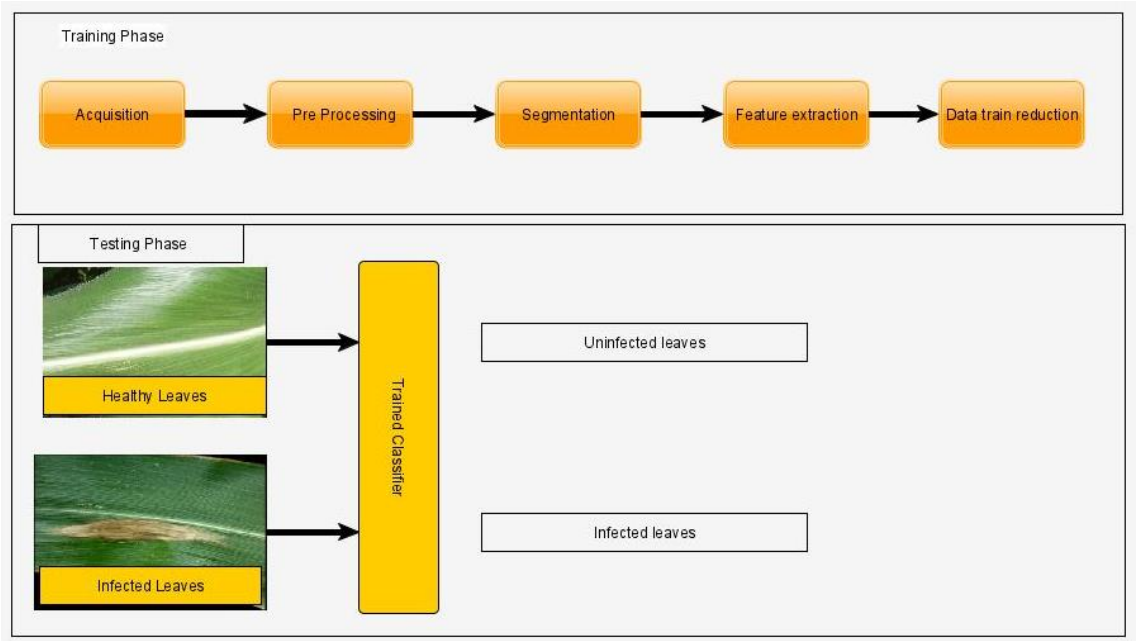


Figure 2. Research Procedure

d. Feature extraction

For feature extraction, we use the second order statistics name Gray Level Co-occurrence Matrices (GLCM). GLCM is a texture analysis introduced by Haralick in 1973 [37]. In GLCM there are 2 main variables namely the orientation of the angle and distance of the pixel. In this study we use the angle orientation are 0^o, 45^o, 90^o and 135^o and pixel distance is 1 pixel.

e. Data folding

For data fold, 10% of the total data were data test. The remaining 90% will be used as a data train.

f. Data train reduction

The algorithm used in this study to reduce data train is using the K Support Vector Nearest Neighbor (KSVNN) algorithm which combines the K Nearest Neighbor (KNN) algorithm and the Support Vector Machine (SVM) algorithm with the principle of K nearest neighbors in each data train [28] . KSVNN tries to reduce the data train based on the score property or the degree of significance in each alternative data based on the closest neighbor K principle. The two properties are the Left Value (LV) and the Right Value (RV). The left value is for data train that has the same class, while the right value is for a different class. The number of LV and RV from all training data is the same $N \times K$ as stated by equation (1).

$$\sum_{i=1}^{N} LV_i + \sum_{i=1}^{N} RV_i = N \times K \quad (1)$$

The significance degree is the value that states the relevance of the data train to the objective function described in a hyperplane area. Hyperplane area values range from 0 to 1. When the calculation get the value, then the higher value being the higher relevance as a support. Support vector values are

training data used in predictions. As for the threshold (T) used is more than 0 ($T > 0$), which means the slightest relevance value will be used as a support vector.
A value of 0 on the significance level of the training data means that the data must be discarded because it is not used as a support vector. The value of the degree of significance or significant degree (SD) is obtained by dividing LV against RV with the condition in equation (2).

$$SD_i = \begin{cases} 0 \ , SV_i = RV_i = 0 \\ \frac{SV_i}{RV_i}, \ SV_i < RV_i \\ \frac{SV_i}{RV_i}, \ SV_i > \ RV_i \\ 1, \ SV_i = \ RV_i \end{cases} \quad (2)$$

As for the KSVNN training to get support vectors, the support vectors are then stored in memory for use when predictions are as follows [28]:
1) Initialization D as a set of data train, where K is the nearest neighbour and T is the threshold for significance degree (SD), LV and RV is all set to 0 in each of instance in the data train
2) For each data train $d_i \epsilon D$, go to steps 3 through 5
3) From d_i to the other data train, calculate the distance between that
4) Select d_t as K closest data train (not included d_i)
5) For each data train in d_t, if the class label is equal d_i then add 1 to LV_i, if it's not the same then add 1 to RV_i
6) For each data train d_i, calculate SD_i using equation (2)
7) Select data train with SD $\geq T$, save it in memory (variable) as a template for prediction.

g. Model testing (trained classifier)
In this study the KNN algorithm is used as an algorithm used to form models. To predict a new sample, KNN use a nearest neighbour technique to calculate the similarity or distance between [29]. The distance used is Euclidean Distance measurement. Euclidean Distance is the most common distance defined as follows:

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (3)$$

Where :
x_1 = sample data
x_2 = test data
i = variable data
d = distance
p = data dimension

The number of nearest neighbors or K used in this study are 1 - 10 as a nearest neighbor. After the model is formed, the model is tested by data test. Finally, in this study will be calculated how many errors from the system (mean error).

3. RESULT AND ANALYSIS

3.1. Dataset Description

The data used are 3847 images of maize leaves divided into unbalanced classes with the division of four classes as follows healthy leaves are 1162 images, leaves with southern rust are 1192 images, leaves with southern leaf blight are 508 images, and leaves with northern leaf blight are 985 images.

3.3 Pre Process

The preprocessing begin with converting images to images size below 100 Kb. This is done to reduce the size of the image so that the transmitted image is smaller and convert colors from the RGB channel to the grayscale channel more easy. Furthermore, noise reduction is performed with the median filter, with the kernel used is 3 x 3. Python 3 is using to define the preprocessing function. In Figure 3 shows the image that has been enhanced with median filter with kernel 3x3.

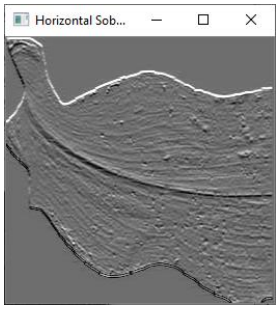
Image compression results Image of 3x3 median filter results

Figure 3. Display results before and after using median filters on maize leaves

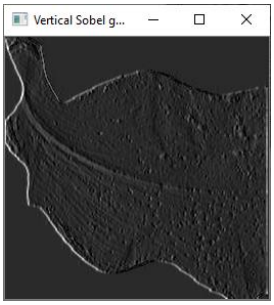
3.4 Segmentation

At this stage, input from the median filter is used to get the shape characteristics with the Sobel Operator. With the Sobel Operator, each horizontal and vertical edge of the image is emphasized. In Figure 4, using a function built in Python 3, the results of maize leaves with sharpen edge are shown.

a) Normalization of gradient image results



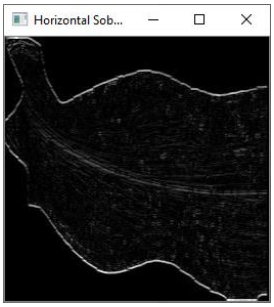
Double normalized gradient normalization



Sobel gradient vertical normalization

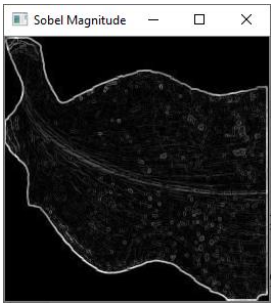
b) Edges of horizontal and vertical

Horizontal edge magnitude (from y gradient)



Vertical edge magnitude (from x gradient)

c) Finalization of image results from segmentation (Sobel Magnitude)



The combined image from vertical and horizontal

Figure 4. Results of segmentation with Sobel Operators

From Figure 4 it was found that the pattern of maize leaves affected by the patterned disease and more visible in the image of part (c) in the Sobel Magnitude.

3.5 Feature Extraction

GLCM is used in research with. As for the Haralick features used are contrast, correlation, energy, and homogeneity with angles of 0º, 45º, 90º and 135º with pixel distance are 1. Next in Figure 5 the results of the image extraction from GLCM are displayed.

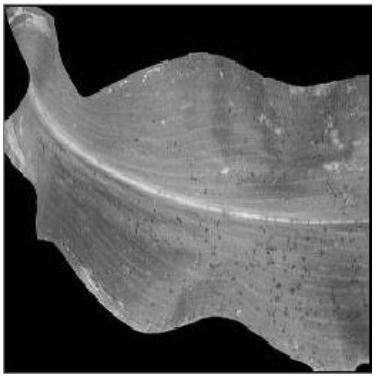


Figure 5. Image of GLCM results

3.6 Data Train Reduction

The implementation of the KSVNN algorithm is used on the system to reduce data train using Python 3. The following is a code snippet image used for developing data train reduction shown in Figure 6.

```
class KSVNN:
    def __init__(self, K, T):
        self.K = K
        self.T = T
```

Figure 6. Making the initial class

For the first time, a class called KSVNN was created, the number of K (nearest neighbors) and the T threshold was determined first, then the value of the Significant Degree (SD) was calculated using equation (2). The code snippet is shown in Figure 7.

```
LV = round(rest_proba * self.K)
SD = 0
if LV == RV:
    SD = 1
elif LV > RV:
    SD = RV / LV
else:
    SD = LV / RV

if SD <= self.T:
    self.support_vectors.append(X_trans[idx])
    self.support_vectors_class.append(y[idx])
```

Figure 7. Implementation of equation (2) on the system

From the code snippet in Figure 7, it is seen that if the SD is smaller or equal to the T threshold value, then the data is taken and stored in memory to be used as predictive data (data train). This data is referred to as

a support vector. The selected data (support vector) is then combined to be used as data for the development of prediction models. Figure 8 shows the function for merging selected support vector values.

```
def get_support_vectors(self):
    return np.array(list((zip(self.support_vectors,self.support_vectors_class))))
```

Figure 8. Implementation of the support vector merging function

The results provided from the overall merging of support vectors, which are used as data train for model development in this study are obtained as many as 3419 support vector support vectors. Initial data used (90% of the total dataset) were 3462 leaf images. With the KSVNN algorithm, reduced as many as 43 leaf images. The nearest neighbor value is K = 3 and the threshold is T = 0.5.

3.7 Model Testing

KNN algorithm is used to build the model. By folding the data train and data test by 90% and 10%, we get the mean error value of the system for the training data that has been reduced by the KSVNN algorithm shown in Table 1 for the 4 class labels used for the nearest neighbor are as many as K = 1 - 10.

Table 1. Mean errors resulting from the introduction of the system

Number of K	Mean Error
1	0.22
2	0.21
3	0.20
4	0.21
5	0.22
6	0.215
7	0.235
8	0.24
9	0.23
10	0.23

From Table 1 it can be seen that at K = 3 the recognition system gives the smallest error value which is 0.20 or 20%. So for the accuracy of the system is 80%. In terms of efficiency, the time given in doing work is 134 seconds. If the system does not implement data train reduction, then the best nearest neighbor value is also K = 3 with an error value of 0.209 or 20.9%. Whereas for the duration of the process, a longer processing time is obtained for 24 seconds from a system that uses data train reduction, which is for 158 seconds.

4. CONCLUSION

Based on the research, we get the results that the reduction of data train using the KSVNN algorithm is able to increase the efficiency and effectiveness of the recognition and classification system of diseases in maize leaves caused by fungi. This can be seen from the length of time, if you use data train reduction with the KSVNN algorithm, the model training time is 24 seconds faster than if you do not do data train reduction. As for the accuracy side, prediction errors for the recognition of healthy and infected maize leaves can be seen that a system with reduced training data provides better accuracy. This is because data train is reduced, and the selected support vector is really characterized the objective function so that the prediction process becomes faster and more accurate.

REFERENCES

[1]. Bengio Y, Courville A, Vincent P. 2013. Representation Learning : A review and new perspectives. IEEE Transaction on Pattern Analysis and Machine Intelligence. 35(8). Pp. 1798-1828

[2]. Wakman W, Burhanuddin. 2016. Pengelolaan penyakit prapanen jagung[online] balitsereal.litbang.pertanian.go.id. Tersedia pada : http://balitsereal.litbang.pertanian.go.id/wpcontent/uploads/2016/11/satutujuh.pdf. [diakses, Juli 11, 2020]

[3]. Overbeek MV, Kaesmetan YR, Tobing FAT. 2019. Identification of maize leaves diseases cause by fungus with digital image processing (case study : Bismarak village, Kupang District – East Nusa Tenggara). 2019 5th International Conference on New Media Studies (CONMEDIA), Bali – Indonesia.pp.125-128. DOI : 10.1109/CONMEDIA46929.2019.8981843.

[4]. Azlah M Z, Lee S C, Rahmad F R, Abdullah F I, Alwi S R W A. 2019. Review on techiques for plant leaf classification and recognition. Computers vol 8, issue 4.pp.77

[5]. Sabu A, Sreekumar K. 2017. Literature review of image features and classifiers used in leaf based plant recognition through image analysis approach. Proc.of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore-India 10 – 11 March 2017.pp.145-149
[6]. Kaur S, Pandey S, Goel S. 2019. Plant disesase identification and classification through leaf images : a survey. Arch Computat Methods Eng 26. Pp.207-530. DOI : https://doi.org/10.1007/s11831-018-9255-6
[7]. Lei Y, Liu H. 2003. Feature selection for high dimensional data : a fast correlation based filter solution. ICML vol 3.pp.856-863
[8]. Zheng J, Yang W, Li X. 2017. Training data reduction in deep neural networks with partial mutual information based feature selection and correlation matching based active learning. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA. Pp.2362-2366. DOI : 10.1109/ICASSP.2017.7952579
[9]. Doyle S, Monaco J, Feldman M Tomaszewski, Madubhusi A. 2011. An active learning based classification strategy for the minority class problem : application to histopathology annotation. BMC bioinformatics, vol.12, no.1. pp.424
[10]. Zhang Y, Zhao Z. 2017. Fetal state assessment based on cardiotocogralhy parameters using PCA and ADA boost. 10th international congress on image and signal processing. BioMedical Engineering and Informatics (CISP-BMEI), pp.1-6.IEEE
[11]. Zhu C, Uwa C, Idemudia, Feng W. 2019. Improved logistic regression model for diabetes prediction by integrating PCA and KMeans techniques. Informatics in Medicine Unlocked, page 100-179.
[12]. Kaya I E, Pehliyanl A C, Sekizkarde E G, Ibrikci T. 2017. PCA based clustering for brain tumor segmentation of T1W MRI images. Computer Methods and Programs in BioMedicine. 14 : 19-28.
[13]. Hu L,Cui J. 2019. Digital image recognition based on fractional – order PCA SVM coupling algorithm. Measurement. 145:150-159.
[14]. Bhattacharya S, Kaluri R, Singh S, Alazab M, Tariq U. 2020. A novel PCA-Firefly based XG Boost classification model for intrusion detection in networks using GPU. Electronics, 9(2):219
[15]. Gadekallu T R, Khare N, Bhattacharya S, Singh S, Maddikunta P K R, Ra I H, Alazab M. 2020. Early detection of diabetic retinopathy using PCA-Firefly based deep learning model. Electronics. 9(2) : 272
[16]. Li Z, Ma X, Xin H. 2017. Feature engineering of machine-learning chemisorption models for catalyst design. Catalyst Today. 280 : 232 – 238.
[17]. Cheng C A, Chiu H W. 2017. An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national wide database. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Pp 2566-2569.
[18]. Zaman S, Toufiq R. 2017. Codon based back propagation neural network approach to classify hypertension gene sequences. 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE). Pp.443-446.
[19]. Tang H, Wang T, Li M, Yang X. 2018. The design and implementation of cardiotocography signals classification algorithm based on neural network. Computational and Mathematical Methods in Medicine.
[20]. Tao Z, Huiling L, Wenwen W, Xia Y. 2019. GA – SVM based feature selection and parameter optimization in hospitalization expense modelling. Applied Soft Computing. 75:323-332
[21]. Karolis M A, Moutiris J A, Hadjipanayi D, Pattichis C S. 2010. Assessment of the risk factors of coronary hearth event based on data mining with decision tree. IEEE Transaction on Information Technology in Biomedicine. 14(3): 559-566
[22]. Abdar M, Makarenkov V. 2019. CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement. 146:557-570.
[23]. Sartakhti J S, Zangooei M H, Mozafari K. 2012. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM – SA). Computer methods and programs in BioMedicine. 108(2):570-579
[24]. Orphanou K, Dagliati A, Sacchi L, Stassopoulou A, Keravnou E, Bellazi R. 2018. Incorporating repeating temporal association rules in naïve bayes classifiers for coronary heart disease diagnosis. Journal of BioMedical Informatics. 81:74-82.
[25]. Qummar S, Khan F G, Shah S, Khan A, Shamshirband S, Rehman Z U, Khan I F, Jadoon W. 2019. A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access. 7:1500530-150539
[26]. Srisawat A, Phientrakul T, Kijisrikul B. 2006. SV – KNNC : an algorithm for improving the efficiency of K Nearest Neighbor. Qian Yang, Geoffrey I. Webb, the 09th Pacific RIM International Conference on Artificial Intelligence (PRICAI-2006). Guilin, China, 7-11 August 2006. Springer – Verlag Berlin Heidelberg

[27]. Barigou F. 2016. Improving K Nearest Neighbor efficiency for tex categorization. Neural Network World. 26(1). 45
[28]. Prasetyo E. 2012. K Support Vector Nearest Neighbor untuk klasifikasi berbasis KNN. Proc. Seminar Nasional Sistem Informasi Informasi. Institut Teknologi Sepuluh November Surabaya
[29]. Han J, Kamber M, Pei J. 2012. Data mining Concepts and Techniques 3th edition. Waltham (US) : Morgan Kaufmann Publishers
[30]. Cover T, Hart P. 1967. Nearest Neighbor Pattern Classification. IEEE Transaction on Information Theory. 13.pp.21-27
[31]. Zhang S, Zhang C. 2013. Orthogonaly locally discriminant projection for classification of plant leaf disease. IEEE International Conference on Computational Intelligence and Security CIS. Leshan. Pp.241-245
[32]. Prasad S, Peddoju S K, Ghosh D. 2016. Multi – resolution mobile vision system for plant leaf disease diagnosis. Signal Image Video Process. 1092):379-388
[33]. Zhang S W, Shang Y J, Wang L.2015. Plant disease recognition based on plant leaf image. J Anim Plant Sci 25 (suppl. 1):42-45
[34]. Pujari J D, Yakkundimath R, Byadgi A S. 2015. Image processing based detection on fungal disesase in plants. Proc. Computer Sci. 26:1802-1808
[35]. Wibowo A, Hidayatno S A, Isnanto A, Rizal R. 2011. Analisis deteksi tepi untuk mengidentifikasi pola daun. Undergraduate Thesis, Teknik Elektro Universitas Diponegoro
[36]. Overbeek M V, Kaesmetan Y R. 2015. Ekstraksi tekstur benih jagung lokal Pulau Timor dengan GLCM. Proc. SEMMAU I Conference
[37]. Putra D. 2010. Pengolahan citra digital. Jogyakarta (ID) : ANDI Jogyakarta

BIBLIOGRAPHY OF AUTHORS

Marlinda Vasty Overbeek, S.Kom, M.Kom, is a lecture. Early career as lecturer at STIKOM Uyelindo Kupang in the study program Informatic Technique. Now moves to Tangerang, Banten and has a career at Universitas Multimedia Nusantara as a lecturer.

Yampi R. Kaesmetan, S.Kom, M.Kom, now work at STIKOM Uyelindo kupang as a lecturer in the study program Informatic Technique