khazanah
informatika

Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika is a national scientific journal that publishes scientific research papers/articles or reviews in the field of Computer Systems and Informatics. The journal is accredited “Sinta 2” according to the decree of DRPM Ministry of Research and Higher Education number 21/E/KPT/2018 dated July 9th, 2018 which is valid since vol 2 issue 1 until vol 6 issue 2. The scope of this journal includes software engineering, information systems development, computer systems and computer networking. The journal is published by Muhammadiyah University Press (MUP), Universitas Muhammadiyah Surakarta.

ISSN: 2621-038X e-ISSN: 2477-698X

khazanah informatika

Vol. 6, No. 1 | June 2020

Table of Contents

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Writer Identification of Lampung Handwritten Documents Based on Selected Characters

Akmal Junaidi*, Syifa Trianingsih, Muhammad Iqbal
Computer Science Department, FMIPA
Universitas Lampung
Bandar Lampung
*akmal.junaidi@fmipa.unila.ac.id

Abstract-Writer identification is a sub-field in handwriting recognition which its objective is to determine the identity of the writer based on handwriting input. The goal is usually for forensic purposes such as finding the perpetrators of crimes that leave traces of evidence in the form of written messages. In addition, writer identification can also be used to determine the identity of a historical actor if he or she leaves a valuable written artefact. The object of this research is the traditional character of the Lampung region which is so-called Had Lampung by the local community. The traditional character of Lampung consists of 20 main characters and 12 diacritics. Based on selected characters, the writer will be recognized using the Principal Component Analysis (PCA) feature. PCA is one linear feature extraction method of an object in pattern recognition. The PCA algorithm consists of several stages, namely the calculation of the average dataset, the subtraction of the vector dataset with averages, the calculation of covariance, the calculation of eigenvectors and eigenvalues, eigenvector reduction, and the projection of the dataset against reduced eigenvector space. PCA in this paper is used as a feature in image recognition. The dataset utilized in this study is the Lampung Dataset which is a handwritten character recognition (HWCR) dataset. Lampung Dataset consists of 82 Lampung handwritten documents. All Lampung character images in the dataset were extracted from these documents using the connected component extraction algorithm and eventually generated 32,140 images. Furthermore, these images are converted into grayscale images. In this research, as many as 12,500 grayscale images of Lampung handwriting characters were chosen to represent 82 different writers. This data is employed as training and testing data on the proposed method. The highest accuracy of the identification of the writer using this PCA feature is 82.92%, while the lowest accuracy is 28.29%.

Keywords: Lampung Script; Writer Identification; Principal Component Analysis; Lampung Dataset

1. Introduction

Recently, writer identification has become a popular research topic in the area of pattern recognition. An interesting factor in the research topic of writer identification is the handwritten style of each individual who at a glance is similar, but it has its own uniqueness. Handwritten character patterns are an important element for forensic experts to identify the writers. In addition to forensic purposes, writer identification can also be used for the benefit of scientific development. The core process in identifying writers is the process of extracting features from handwritten character image that will be recognized. The object in writer identification can be in the form of a modern or contemporary script as well as a traditional script. This research uses a traditional script originating from the Lampung region, one of a few regions in Indonesia that has a traditional script. The Lampung script, or locally called Had Lampung, has 20 main characters and 12 diacritics. The Lampung handwritten data compiled in Lampung Dataset consists of 82 raw images of Lampung handwritten documents, 82 text files containing annotations for each document and 32,140 grayscale images of single Lampung character [1]. The grayscale image in the dataset is the result of two stages of preprocessing of these documents. Those stages are connected component extraction followed by converting images into grayscale format. Some character samples of Lampung handwriting in Lampung Dataset can be seen in Figure 1.

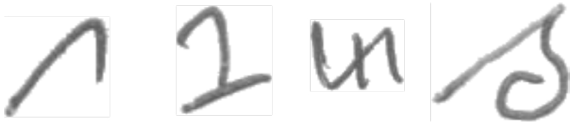


Figure 1. Handwritten Character Samples of 4 Different Writers in the Dataset

Character recognition is a research topic that has been developing for more than two decades. In the upstream side, many researchers provide a dataset to facilitate HWCR research. Some examples are the providing of the Lampung handwritten character dataset [1], historical handwritten digit documents of church records by priests in Sweden [2], and Arabic handwriting from historical manuscripts [3]. The methods and approaches in HWCR have also been applied for various scripts, for instance the Kurdish Text Classification of Sorani dialect [4], Slavic Historical Documents containing Glagolitic and Cyrillic character [5], printed Arabic [6], handwritten Bangla characters from India [7], handwritten Kanji characters [8], offline handwritten Chinese characters [9] and so on. The use of PCA specifically for handwritten character recognition is quite difficult to be found. So far, The research related to PCA has been used for recognizing Urdu characters [10]. The accuracy obtained in the study reached 96.2%. In other research, the use of PCA was not directly applied to handwritten character recognition but was employed as a method of feature space reduction before classification stage [11]. The study only addressed the recognition of digit from the MNIST dataset [12] and CVL Single Digit [13].

PCA analysis is one of the classic methods that has been widely applied to various researches. The purpose of using PCA analysis is to reduce information features that are redundant (and large) in order to obtain feature components with lower dimensions while still maintaining the values of discriminative features [14]. This analysis has been widely used to reduce feature dimensions in various pattern recognition tasks, especially for object recognition. In the field of pattern recognition for medical data, PCA is used to reduce the dimension of large size features in dynamic contrast enhanced MR imaging (DCE-MRI) data of hypoxia tumors [15]. In the context of the study, the use of PCA is intended to find the number of components that can distinguish the overall variability of data. The results of the study concluded that the 99% level of overall data variability can be described by only the first three principal components obtained by PCA. Another similar study in the medical field also uses PCA for selection of spectral entropy (SE) features from a 64-channel electroencephalogram (EEG) recording for the detection of alcoholics [16]. The role of the PCA in the study was to evaluate the size of the feature-set at the top-rank position before classification. Variations of the various sizes of these top-rank features indicate a ranking of those features after dimensional reduction. The effect of ranking and PCA during classification by k-NN has shown an improvement in accuracy compared to without ranking. Another unique PCA analysis has been applied for pattern recognition in the gym / fitness center [17]. In that study, PCA was used in gesture recognition (Action Recognition) for 770 fitness exercise movements. PCA analysis is used to reduce the feature dimension of the dataset into 3-7 features. Reduction does not decide a single feature because the number of features in that range is under scrutiny to determine the correlation between number of principal components (PCs) and features which are the most relevant for correction recognition subsets. With the PCA analysis, the best accuracy achieved was 97 ± 14%. The area of pattern recognition that also utilizes PCA a lot is in the face recognition [18], [19]. The study in [18] uses PCA to recognize faces even at different facial poses and orientations. Whereas research in [19] uses a comparison of PCA and Linear Discriminant Analysis (LDA) for each face recognition. The results concluded that the face recognition performance with PCA was superior compared to LDA.

Writer identification as well as pattern recognition, can use various features for the identification process. The use of global features and local features as one characteristic of handwriting is an appropriate feature combination for writers identification [20]. The dataset was distinguished of 650 and 225 writers, respectively. The identification performance with these features achieved 86% accuracy for the dataset of 650 writers and 79% for the dataset of 225 writers. The method proposed in the study is claimed to be applicable toward the writer identification of non-Latin handwriting such as Asian or Arabic. Other study regarding the writer identification was applied for handwritten Chinese character [21]. The research uses 16,000 Chinese words from 40 different writers. PCA in the study was applied to these words to find unique personal handwriting style characteristics and the best discriminative representation to other personal. The identification process did not use the entire writing texts but only a pieces of words. Beside achieving a high identification accuracy of 97.5%, the use of this approach has impact on time reduction during identification process.

Most of the previous scientific studies have been carried out using the entire document or complete words as identification input, while in this study the identification input used is piece of images of the single character image or double or five Lampung character images in grayscale format. The use of the grayscale format in this study is based on two reasons. First, the grayscale image is fairly well to store the characteristics of the Lampung characters. Grayscale image representation consists of intensity values with a range of 0-255 so it is quite significant to represent variations and details of the image. Second, the grayscale image consists of one channel only which will lighten the computational workload compared to the RGB image which consists of three channels. The use of features extracted from RGB images will increase computational workload by threefold.

One important step in pattern recognition is feature extraction which is defined as a process to extract the special value of the object so that the object can be recognized using this feature. This writer identification process also applies a feature extraction process to identify the writer of handwriting. A variety of feature extraction methods can be used to identify writers, such as the histogram method, line-based representation method, to linear transformation [22]. The feature extraction method implemented in this study uses the PCA approach, which is a linear transformation method that works by reducing the feature dimensionality of the object as a feature extraction stage. PCA is generally applied to data that has very high dimensionlity.

Based on the explained introduction, the purpose of this research is to implement PCA as a feature extraction method in identifying the writers of the Lampung handwritten documents. In addition, this study also aims to determine the accuracy of the PCA feature in identifying the writers of handwritten documents based on selected characters. This research is useful to understand the stages of PCA implementation in identifying writers of Lampung handwritten documents. So far, there have been no results of studies on writer identification of the Lampung handwritten character. The result reported in this paper is a novelty that has never been published formerly. Thus, this paper is expected to be useful for reference to other similar studies. In a wider scope, this research is expected to be able to contribute to the development of science in the pattern recognition domain especially in handwritten character objects.

2. METHOD

a. Writer identification

Writer identification is the recognition of the writer based on handwritten-character by matching of an unknown handwriting sample to the sample of data for which the writers have been known previously. The identification process is done by counting and comparing feature of handwritten samples with a feature database that has been stored. The results of identification of the most similar writers will have the highest level of similarity [23]. Two approaches for writer identification are to analyze based on characters and the approach by using textures from documents [24]. In this study, the proper approach is to use Principal Component Analysis (PCA) analysis to character images.

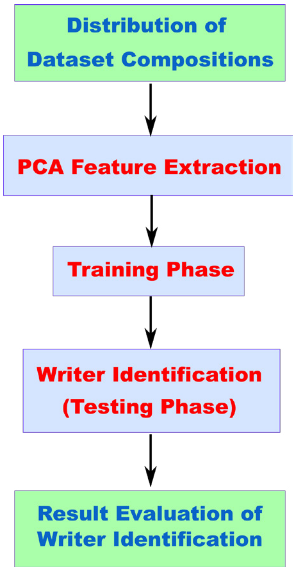

Distribution of Dataset Compositions
PCA Feature Extraction
Training Phase
Writer Identification (Testing Phase)
Result Evaluation of Writer Identification

Figure 2. Research Stage of Writer Identification

PCA is a linear transformation method which is also known as the Karhunen-Loeve Transform (KLT) method. The feature of the PCA approach is the result of feature extraction which has been reduced in a simpler form. Dimensionality reduction is done by compressing the specific information that characterizes the object. This special feature set is represented as an eigenvector and the writer identification is evaluated from projected eigenvector of training and testing images.

The steps of the research on the writer identification on the Lampung handwritten documents is illustrated in Figure 2.

The procedure as shown in Figure 2 is in principle a pattern recognition framework, but it is adapted for the purpose of writer identification. Detailed descriptions of each stage are given in the following sub-sections.

b. Extracting PCA Features

The PCA feature extraction process is carried out in several steps. The calculation of average, subtraction, eigenvectors, eigenvalues, elimination of eigenvalues, and its projections in this study refers to the steps of the standard PCA algorithm [18]. The algorithm is explained briefly in the following.

1) Step 1: Prepare the image objects.
The object image is the image with representations as I1, I2, I3, I4, ..., IM. These image objects must have the same dimension.

2) Step 2: Prepare the dataset.
Each image object Ii is transformed into a vector and used as a training set S, where S = {Γ1, Γ2, Γ3, Γ4, ..., ΓM}.

3) Step 3: Calculate the average of dataset.
The average vector of dataset (Ψ) can be calculated using formulas:

$$\Psi = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n \qquad (1)$$

where:
Ψ: average vector of dataset
M: number of data
n: index of data, n lies from 1 to M
Γ_n: n^th training image vector

4) Step 4: Subtract of dataset vector and its average.
The dataset vector (Γi) is subtracted from average vector of dataset (Ψ) and stored in the Φi variable. The formula is given below:

$$\Phi_i = \Gamma_i - \Psi \qquad (2)$$

where:
Φ_i: subtraction vector
Γ_i: i^th image vector
Ψ: average vector of dataset

5) Step 5: Calculate the covariance matrix.
The formula to compute the covariance matrix C is denoted in the following:

$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi_n \Phi_n^T \tag{3}$$

where:
C: covariance matrix
M: number of data
Φ_n: nth subtraction vector
Φ_n^T: transpose of nth subtraction vector

Equation (3) can be simplified into $C = A \times A^T$ with $A = \{\Phi_1, \Phi_2, \Phi_3, \ldots \Phi_M\}$.

6) Step 6: Calculate eigenvectors and eigenvalues of covariance matrix.
Calculating the eigenvectors and eigenvalues can consider the formula $L = A^T \times A$ for reasons of efficiency and reducing the dimensions of the matrix during the computation process.

7) Step 7: Eliminate eigenvectors.
Eigenvalues obtained from previous step are then eliminated partially and hold the most relevant values that can significantly represent the objects.

8) Step 8: Calculate the dataset projection into the eigenvector space.
The next step is to calculate the dataset projections to eigenvector space using the formula below:

$$\omega_i = L^T \times (\Gamma_i - \Psi) \tag{4}$$

where:
ω_i: image projection
Γ_i: ith training image vector
L^T: transposed eigenvector
Ψ: average vector of dataset

These steps have been implemented in a computer program to carry out the two processes in this study. The first process is training data as an effort to learn the characteristics of handwriting by the system. The next process is identification as a system decision in recognizing the writer of the document from the handwriting contained in the document.

c. Training phase

After all steps of feature extraction have been carried out, the next stage is the training stage. This stage is the training of available features obtained from former stage to be used at the writer identification stage, i.e. by projecting each eigenvector on the Lampung feature vector. The results of this projection will be used as a decision reference at the writer identification stage.

Data samples in the training process are arranged into 3 different configuration schemes. These configuration differences are arranged such that the use of sample data represents units of data consisting of 1, 2 and 5 single character images. The selection of this configuration is decided based on trial and error. A detailed explanation of this configuration can be found in subsection III.

d. Writer Identification Matching Scheme (Phase Testing)

A number of images of which the writer(s) to be identified must go through a matching scheme as described in Figure 3 [25]. Stages of writer identification aim to predict writer from input sample of handwritten images. The decision of the identification result is determined by comparing the value of the projected training image to the input image which is a sample of the testing image based on a minimum value of Euclidean distance.

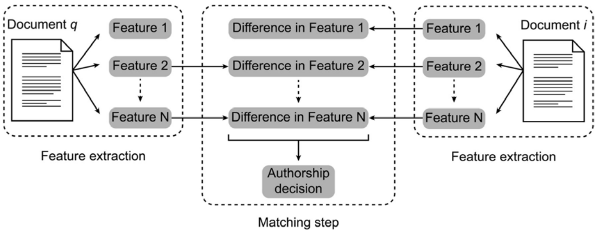


Figure 3. Matching Scheme on the Writer Identification System

If all decisions on the results of the matching steps have been obtained, the next step is to measure their accuracy using equation (5).

$$\text{Accuracy} = \left(\frac{\text{Amount of Correct Identification}}{\text{The amount of data}}\right) \times 100\% \tag{5}$$

The level of accuracy of identification is calculated after the overall results of the matching decisions are obtained completely. The counting of the matching is done with the aim to know the number of documents that are recognized correctly as an indicator of the accuracy on the writer identification process. The accuracy can be used as an evaluation whether the proposed method and the features indicate a good performance or not.

3. Results

Lampung dataset in this study is image collection of Lampung handwritten character in grayscale format with size 32x32 pixels. This dataset is distributed into two parts, one group as a training set for development of an identification model and another as a testing set for writer identification matching scheme. The character image sample is randomly selected as many as 12,424 images out of the total 32,140 character images in the dataset. The selected characters are further divided into two parts, 11,768 character images as the training set and the remaining 656 character images as the testing set. The characters from each part are then randomly selected to be divided into three sample groups for writer identification. Details of these sample distributions are listed in Table 1.

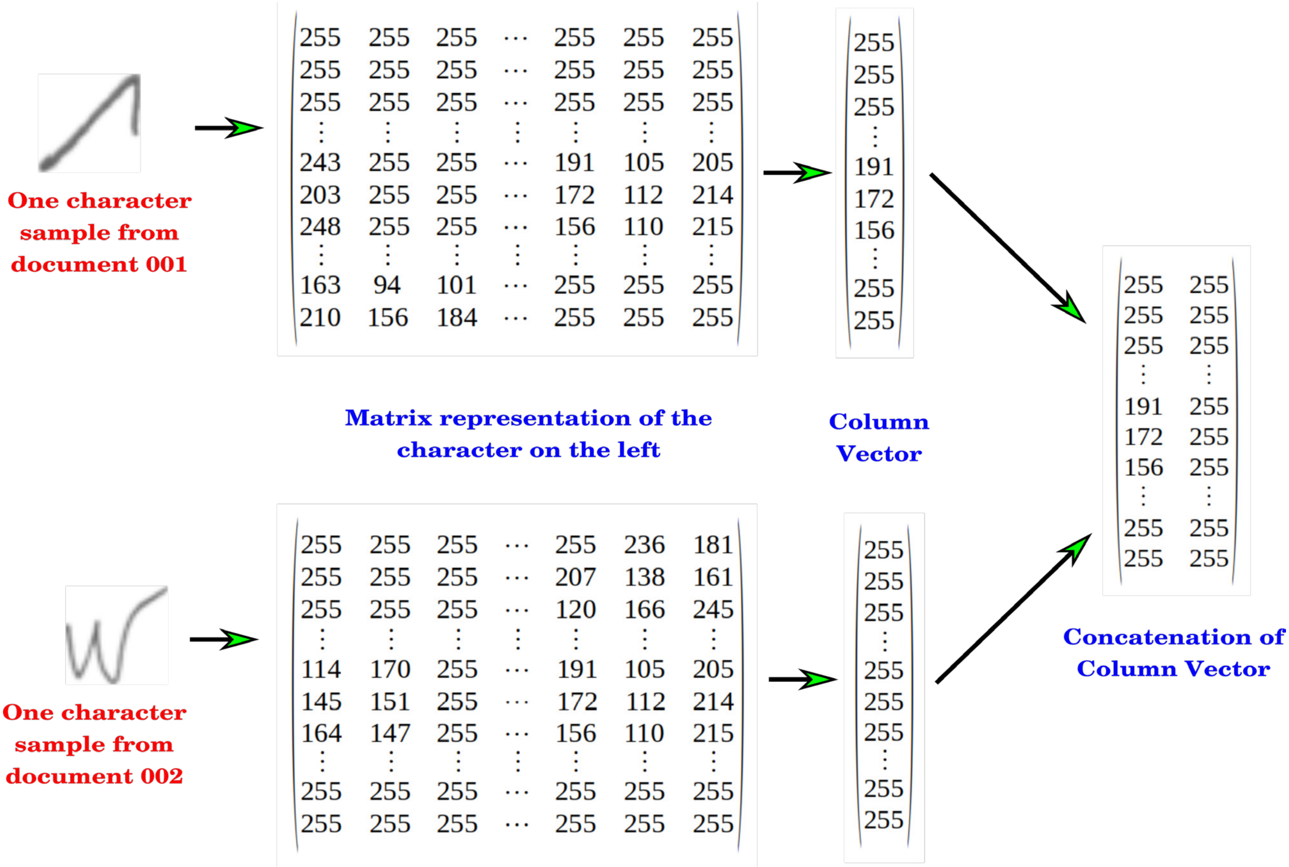

One character sample from document 001
Matrix representation of the character on the left
Column Vector
Concatenation of Column Vector
One character sample from document 002

Figure 4. Tranforming Process of Lampung Handwriting Sample into Column Vector.

Table 1. Distribution of Training and Testing Set of Writer Identification

No	Sample Name	Number of Training Set	Number of Testing Set
1	Sample I	1.548	82
2	Sample II	4.920	164
3	Sample III	5.300	410
	Total sample group	11.768	656
	Total amount of sample	12.424	

Sample I is a sam ple group consisting of one character image as one unit of data. The second sample group is a sample with 2 character images as one unit of data. The last sample group uses 5 character images as one unit of data. The elements of one unit of data in sample II and III are also randomly assigned.

Before the feature extraction stage using PCA is performed, the entire image sample is converted into a column vector. This is conducted for the sake of efficiency and convenience during the computational process as well as dimensionality reduction. An example of transforming process of two image samples of Lampung handwritten character into a column vector is illustrated in Figure 4.

After the entire selected character in each sample is converted into a column vector, the next step is feature extraction using PCA. Each sample is processed by following the algorithm and steps described in section II.B. These steps are calculating the average, then subtraction of the training set and its average, followed by calculating the covariance of subtraction and finally calculating the character projection. The final step is conducted as a reference at the writer identification stage. All these steps are the training phase on the writer identification system.

The writer identification is decided based on the Euclidean distance between the sample image of the testing set with all the images of the writer in the training set. The smallest distance among pairs indicates a high degree of similarity and implies the writer identity. The formula to compute this Euclidean distance is given in equation (6).

$$\mathrm{Euc} = \sqrt{\sum_{n=1}^{M} \left(\Gamma_{in} - \Gamma_{jn}\right)^2} \qquad (6)$$

where:
Euc: Euclidean distance
M: number of data

Γ_{in}: training image vector
Γ_{jn}: testing image vectorajunaidi

By using the formula in equation (6), Euclidean distance is computed for all three sample groups. Then the computation outcome is evaluated to find the minimum Euclidean distance as the closest pair among testing and training data. The final results of the writer identification by this procedure are summarized in Table 2.

Table 2. The Accuracy of Writer Identification on the Lampung Handwritten Document.

Sample Name	Number of Testing Set	Number of Correct Identification	Accuracy (%)
Sample I	82	68	82,92
Sample II	164	82	50,00
Sample III	410	118	28,29

Evaluation of Euclidean distance is a testing phase or matching scheme of writer identification through a series of system processes. The evaluation shows that the highest accuracy obtained from sample I. Two other samples show a significantly decreasing accuracy rate.

4. Discussion

The observations in Table 2 show that the highest accuracy of the writer identification is found in Sample I with total 68 correct identifications, resulting the accuracy of 82.92%. The lowest accuracy occurs in the result of Sample III as many as 118 correct identifications or 29.29%. The research prediciton for the best accuracy of proposed method was at least 75%. This means that the target research has met the expectations and the performance of the PCA method has provided adequate results for the writer identification of the Lampung handwritten character. However, the direct implication of this result is that there is still a space for improvement of the accuracy and the writer identification of Lampung handwriting is relatively a new "brand" in the subdomain of writer identification. Some interesting research opportunities are explained in the advice section.

Based on the accuracy noticed in Table 2, observations and analyzing were carried out on the samples that were incorrectly identified. Several possible causes of writer identification failure were successfully observed. Three main reasons of the failure in this identification process are explain in the following:

a. The combinations and permutations of characters for each unit of data in sample III is quite large

The first factor has a significant effect on the accuracy of sample III because of the large difference in the combination of characters in the training set and one unit of data that should be formed. With the total number of Lampung characters as many as 18 characters, the unit data that must be arranged with member of 5 characters for the training data should be in total 8,568 units of data. This amount is obtained from the calculation on the combination concept of ${}_{18}C_5$. This phenomenon results in a state space explosion of character variations for units of data. While the number of random selected training set is only 5,300 single characters (see Table 1) which must be arranged in a 5-character formation without replacement. With this arrangement, only 1,060 units of data have been formed. The difference as of 7,508 is the minimum number of combination arrangements that are lack in this training set. Consequently, the representation of the model for the writer identification developed during the training process does not reflect all possible combinations. If there is a newly identification data (the one from testing set) belonging to the 7508 group, it is most likely that the accuracy of the writer identification will be biased. The possibility of this bias is quite large because of the number of this group category is also large. Moreover, the arrangement of characters in one unit of data has a lot of permutations. Although there are 5 characters in one unit of data, the total composition of the arrangement of 5 images is 5!. This case doubles the bias which is already large by the former circumstance. Both have a direct impact on the decreasing in accuracy of the writer identification.

b. The similar shapes of some Lampung characters

Antoher factor leading to an improper prediction of a writer is similarities among of Lampung characters. The basic shape of many characters resembles each other. It is like a quadratic curve and mainly flows to upright. As the results, the Euclidean distance will be small so that two character images with essentially similar in its basic shape will be identified as the same character. Thus, the identification process generates an error for this case. Samples of similar shape of the Lampung character are shown in Figure 5.

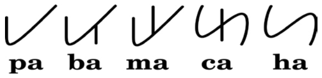


Figure 5. Similar Basic Shape of Some Lampung Characters

The three leftmost characters of Lampung Script in Figure 5 have a basic shape that is similar to the shape of a parabolic curve that opens upward. The basic shape of other similar characters is shown in the last two characters in Figure 5. Both characters have a basic shape like the letter S rotated to 90° clockwise. The main difference between the characters is only the presence or absence of a short line in the middle of the basic shape of the character. Apart from these two basic forms, the Lampung writing system still has some basic forms of the two or more characters.

c. Similarity in writing style among writers

In addition to both cases, writing style among writers who are similar each other is also the most likelihood the evidence for wrong prediction of the writers. Two examples of the Lampung handwritings in the sample that are similar in appearance but written by two different writers are visualized in Figure 6.

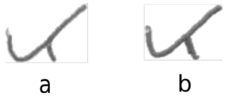


Figure 6. Two Character "Ba" Written by Different Writers

In this example, the character "ba" in part (a) was written by the first writer whereas part (b) was written by the 19th writer. Both images do not exhibit much significant variation so that they look alike. Therefore, the difference of the Euclidean distance between both samples is reasonably small. As a result, both images are considered as the characters derived from the same person during the writer identification process. This type of mistake occurs quite a lot in the testing sample during this final stage. Consequently the accuracy of the writer identification encounters a considerable degradation even in Sampel I as well. This kind of mistake is not triggered by the system operation but instead it is a purely independent factor.

5. Conclusion

The study on the writer identification of the Lampung handwritten documents based on selected characters notice some conclusion:

a. The PCA feature extraction method has been successfully applied to the identification of writers on the Lampung handwritten documents.
b. The highest performance was obtained from the evaluation of Sample I containing of 82 testing images with the accuracy of 82.92%. The lowest performance was confirmed from 410 testing images of Sample III with an accuracy of 28.29%. This highest performance is moderately convincing for the Lampung characters as the new initiating characters on the writer identification sub-field.
c. The implication of the results shown in Table 2 indicates that taking one character image from a Lampung handwriting document as one unit of data is fairly enough to identify the writer.

Based on the analysis during the study, the research team also managed some interesting challenges from this study. Some prospects can be considered as a new subject in the next study or enhanced the existing approach for future development. The topic can be one of the following:

a. Implement the Principal Component Analysis (PCA) feature extraction method as an image-based writer identification extracted from a complete character or line-based handwriting from a document.
b. Use the PCA approach to perform features selection that are the most relevant to Lampung handwritten characters and compare the results of some feature configurations of those selected features.
c. Promote the writer identification process on the Lampung handwriting document using other classification methods such as k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Naïve Bayes, Decision Tree or Hidden Markov Model (HMM).

References

[1] A. Junaidi, S. Vajda and G. A. Fink, "Lampung - A New Handwritten Character Benchmark: Database, Labeling and Recognition," dalam *Join Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, Beijing, 2011.

[2] H. Kusetogullari, A. Yavariabdi, A. Cheddad, H. Grahn and J. Hall, "ARDIS: A Swedish Historical Handwritten Digit Dataset," *Neural Computing and Applications,* 2019.

[3] K. Adam, A. Baig, S. Al-Maadeed, A. Bouridane and S. El-Menshawy, "KERTAS: Dataset for Automatic Dating of Ancient Arabic Manuscripts," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 21, no. 4, p. 283–290, December 2018.

[4] A. M. Saeed, T. A. Rashid, A. M. Mustafa, R. A. A.-R. Agha, A. S. Shamsaldin and N. K. Al-Salihi, "An Evaluation of Reber Stemmer with Longest Match Stemmer Technique in Kurdish Sorani Text Classification," *Iran Journal of Computer Science,* vol. 1, no. 2, p. 99–107, June 2018.

[5] D. Brodić, Z. N. Milivojević and Č. A. Maluckov, "Script Characterization in the Old Slavic Documents," dalam *International Conference on Image and Signal Processing,* Cherbourg, France, 2014.

[6] M. Alghamdi and W. Teahan, "Experimental Evaluation of Arabic OCR Systems," *PSU Research Review,* vol. 1, no. 3, pp. 229-241, 2017.

[7] M. Z. Alom, P. Sidike, M. Hasan, T. M. Taha and V. K. Asari, "Handwritten Bangla Character Recognition Using the State-of-the-Art Deep Convolutional Neural Networks," *Computational Intelligence and Neuroscience,* 2018.

[8] M. Grębowiec and J. Protasiewicz, "A Neural Framework for Online Recognition of Handwritten Kanji Characters," dalam *Federated Conference on Computer Science and Information Systems (FEDCSIS)*, Poznań, Poland, 2018.

[9] Z. Zhong, L. Jin and Z. Xie, "High Performance Offline Handwritten Chinese Character Recognition Using GoogLeNet and Directional Feature Maps," dalam *International Conference on Document Analysis and Recognition (ICDAR)*, Nancy, France, 2015.

[10] K. Khan, R. Ullah, N. A. Khan and K. Navid, "Urdu Character Recognition using Principal Component Analysis," *International Journal of Computer Applications,* vol. 60, no. 11, pp. 1-4, 2012.

[11] A. Das, T. Kundu and C. Saravanan, "Dimensionality Reduction for Handwritten Digit Recognition," *EAI Endorsed Transactions on Cloud Systems,* vol. 4, no. 13, 2018.

[12] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based Learning Applied to Document Recognition," dalam *Proceedings of the IEEE*, 1998.

[13] M. Diem, S. Fiel, A. Garz, M. Keglevic, F. Kleber and R. Sablatnig, "ICDAR2013 Competition on Handwritten

Digit Recognition (HDRC 2013)," dalam *International Conference on Document Analysis and Recognition*, Washington DC, 2013.

[14] X. Cui, P. Zhou and W. Yang, "Local Dominant Orientation Feature Histograms (LDOFH) for Face Recognition," *Applied Informatics,* vol. 5, no. 1, December 2017.

[15] M. Venianaki, O. Salvetti, E. de Bree, T. Maris, A. Karantanas, . E. Kontopodis, K. Nikiforaki and K. Marias, "Pattern Recognition and Pharmacokinetic Methods on DCE-MRI Data for Tumor Hypoxia Mapping in Sarcoma," *Multimedia Tools and Applications,* vol. 77, no. 8, p. 9417–9439, April 2018.

[16] T. K. Padma Shri and N. Sriraam, "Pattern Recognition of Spectral Entropy Features for Detection of Alcoholic and Control Visual ERP's in Multichannel EEGs," *Brain Informatics,* vol. 4, p. 147–158, January 2017.

[17] T. Hachaj and M. R. Ogiela, "Human actions recognition on multimedia hardware using angle-based and coordinate-based features and multivariate continuous hidden Markov model classifier," *Multimedia Tools and Applications,* vol. 75, p. 16265–16285 , 2016.

[18] L. C. Paul and A. Al-Sumam, "Face Recognition Using Principal Component Analysis Method," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* vol. 1, no. 9, pp. 135-139, 2012.

[19] A. Kaur, S. Singh and Taqdir, "Face Recognition Using PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) Techniques," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 4, no. 3, pp. 308-310, 2015.

[20] I. Siddiqi and N. Vincent, "A Set of Chain Code Based Features for Writer Recognition," dalam *10th International Conference on Document Analysis and Recognition*, Barcelona, 2009.

[21] L. Zuo, Y. Wang and T. Tan, "Personal Handwriting Identification Based on PCA," dalam *Second International Conference on Image and Graphics*, Hefei, 2002.

[22] M. Cheriet, N. Kharma, C.-L. Liu and C. Y. Suen, Character Recognition Systems A Guide for Students and Practitioners, New Jersey: John Willey & Sons, Inc., 2007.

[23] G. Louloudis, N. Stamatopoulos and B. Gatos., "ICDAR 2011 Writer Identification Contest," dalam *11th International Conference on Document Analysis and Recognition*, Beijing, 2011.

[24] S. Fiel and R. Sablatnig, "Writer Retrieval and Writer Identificatin using Local Features," dalam *10th IAPR International Workshop ond Document Analysis Systems*, Queensland, 2012.

[25] S. Al-Maadeed, A. Hassaine, A. Bouridane and M. A. Tahir, "Novel Geometric Features for Off-line Writer Identification," *Pattern Analysis and Applications,* vol. 19, no. 3, pp. 699-708, 2016.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Performance of Methods in Identifying Similar Languages Based on String to Word Vector

Herry Sujaini
Department of Informatics
Universitas Tanjungpura
Pontianak
hs@untan.ac.id

Abstract-Indonesia has a large number of local languages that have cognate words, some of which have similarities among each other. Automatic identification within a family of languages faces problems, so it is necessary to learn the best performer of language identification methods in doing the task. This study made an effort to identification Indonesian local languages, which used String to Word Vector approach. A string vector refers to a collection of ordered words. In a string vector, a word is represented as an element or value, while the word becomes an attribute or feature in each numeric vector. Among Naïve Bayes, SMO, J48, and ZeroR classifiers, SMO is found to be the most accurate classifier with a level of accuracy at 95.7% for 10-fold cross-validation and 94.4% for 60%: 40%. The best tokenizer in this classification is Character N-Gram. All classifiers, except ZeroR shows increased accuracy when using Character N-Gram Tokenizer compared to Word Tokenizer. The best features of this system are the TriGram and FourGram Character. The TriGram is preferred because it requires smaller training data. The highest accuracy value in the combination experiment is 0.965 obtained at a combination of IDF = FALSE and WC = TRUE, regardless the conditions of the TF.

Keywords: identification of languages, regional languages, string to word vector

1. Introduction

Language identification functions to identify or recognize the language (or dialect) of a text. Language identification, whose task is to predict the natural language of a written text, is not one of the most challenging problems in computational linguistics but is very necessary for supporting the implementation of other computational linguistics such as machine translators. The accuracy of a Language Identification system is strongly influenced by the similarity of the languages that will be the target of predictions. This research will discuss the identification of very similar languages, namely Indonesian and Malay. Educational figures from Yogyakarta, Ki Hadjar Dewantara, revealed that the basic Indonesian language is the Malay language which is adjusted to its growth in Indonesian society [1]. This is what makes it sometimes difficult to distinguish between Indonesian and Malay. In this study, regional Malay languages, Malay Pontianak and Malay Sambas are used.

Malay Pontianak is one of the languages used by people in West Kalimantan Province. There is no accurate data that can show the number of speakers of languages spoken by Malay people in the city of Pontianak. Malay Pontianak language, in many of its vocabularies, is almost the same as Indonesian. This fact is because the Indonesian language originates and is rooted in Malay [2]. Malay Sambas or Sambas Dialect Malay (BMDS) is one of the regional languages in Indonesia. This language is spread throughout the Sambas Regency, West Kalimantan Province. Sambas Regency, with an area of 6,394.70 km2 or around 4.36% of the area of West Kalimantan Province, has a population of around 505,444 inhabitants [3].

Goutte [4] described the results of evaluations of language identification systems that are trained to recognize various languages. They investigate the progress made from one study to the next. They estimated the upper limit on the performance that can be achieved using voting and oracle plurality, and identify some very challenging sentences. The research uses many diverse languages, including Bosnian, Croatian, Serbian, Indonesian, Malaysian, Czech, and Slovak. The results of this study indicate that the learning curve can help to identify how the task is being studied and which language groups need to be further considered.

There is much research on language identification. One of these studies was presented by Zaidan and Callison-Burch [5]. The authors took resources taken from social media to create large data sets of informal Arabic that are rich in dialect content (more than 100,000 sentences) on three Arabic dialects: Levantine, Gulf, and Egypt. They marked the big data manually to dialect. The authors then use the collected labels to train and evaluate automatic classifiers for dialect identification and observe interesting linguistic aspects of the tasks and behaviour of annotators. By using an approach based on the assessment of language models, they developed a classification that significantly outperformed the baseline using large amounts of MSA data, even close to the level of accuracy shown by human annotators. Lu and Muhammed [6] in another study developed a system called LAHGA, which was positioned to classify HIV, the LEV dialect, the dialect, and the MAG dialect. The author identifies features manually by using these features from interesting devices using Tweets as a dataset for the training and testing process. They use three different classifications, namely the Naïve Bayes classification, the Logistic Regression classification, and the Support Vector Machine classification. During the manual testing process, they eliminate all noise and choose 90 tweets, 30 from each dialect, whereas, in 10-fold cross-validation, there are no human interventions. LAHGA's performance showed 90% in manual tests and 75% in cross-validation.

Other researchers conducted experiments using sentence level approaches to classify whether the sentence was MSA or Egyptian dialect on the task of classifying Arabic dialects [7]. They based their research on a supervised approach using the Naïve Bayes classification. The authors present a supervised approach to the identification of Arabic dialects at the sentence level. This approach uses the features of the underlying system for identification of the token level of Arabic Egyptian Dialect in addition to the core and other meta-features. The method used by them to decide on the choice of sentence given is MSA or EDA. They vary the size of LM on the performance of their approach and study the impact of two types of preprocessing techniques. The approach they used yielded much better accuracy than the previous approach. Safitri [8] conducted a study on the identification of spoken languages with phonotactics in Minangkabau, Sundanese, and Javanese languages, concluding that the PRLM Method showed the highest accuracy using telephone identifiers trained for English and Russian with an average of 77.42% and 75.94%.

Some researchers who have written the results of their research on String To Word Vector include Jhao et al. [9] who proposed a word insertion model at the sub-word level and a word vector generalization method that allows the addition of pre-training word insertions with fixed size vocabularies to estimate "word embeddings" of words that are outside the vocabulary. Other studies found that the F-measure of rhetorical categorization performance in scientific articles can be improved by using word labeling and semantic word representation by Word2Vec [10].

This study has a specific specification that is the application of the String To Word Vector method to identify local languages that have similarities with Indonesian. String To Word Vector methods encode documents into string vectors, not numeric vectors. The traditional approach to text categorization usually requires document encoding into numerical vectors. The approach used is machine learning-based for text categorization, where string vectors are accepted as input vectors, not numeric vectors. As a result, it can improve the performance of text categorization [11].

This paper discusses the performance of the Language Identification method, specifically for languages that have similarities based on the String to Word Vector.

2. Method

The data in this study used sentences in the three languages tested, namely Indonesian, Malay Pontianak, and Malay Sambas. Each language consists of 1,000 sentences so that a total of 3,000 sentences is used. Sentences in Indonesian are taken from internet sources and translated into Malay Pontianak and Malay Sambas.

The research instrument or tool used for data begging is the Waikato Environment for Knowledge Analysis (WEKA). WEKA provides an implementation of learning algorithms that can be applied easily to data sets. WEKA also includes various tools for changing datasets, such as algorithms for discretization and sampling. We can process data, process it into learning schemes, and analyze the classifiers they produce and their performance. All algorithms take their input in the form of a single relational table that can be read from a file or generated by a database request. One way to use WEKA is to apply the learning method to the dataset and analyze the results to learn more about the data [12].

This study uses a set of classifications provided by WEKA [13] by measuring the performance of several classifications with the research steps carried out can be seen in Figure 1.

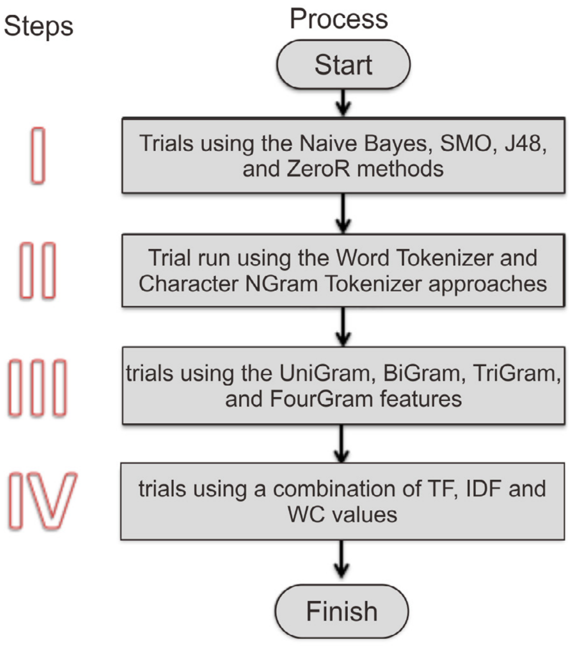

Steps
Process
Start
I
Trials using the Naive Bayes, SMO, J48, and ZeroR methods
II
Trial run using the Word Tokenizer and Character NGram Tokenizer approaches
III
trials using the UniGram, BiGram, TriGram, and FourGram features
IV
trials using a combination of TF, IDF and WC values
Finish

Figure 1. Research steps

In stage I, the experiment used 4 types of classification methods, namely Naïve Bayes, SMO, J48, and ZeroR. Each uses 10-fold cross-validation and 60%: 40% training data: test. In stage II, the experiment used 4 types of classification methods, namely Naïve Bayes, SMO, J48, and ZeroR, using 60%: 40% of training data: test. Each uses Word Tokenizer and Character NGram Tokenizer. In stage III, the experiment used 3000 sentences using 4 Character NGram features, namely UniGram, BiGram, TriGram, and FourGram, using the best classification algorithm from the results of step one and two experiments. Each using 10-fold cross-validation and 60%: 40% training data: test. In stage IV, the experiment uses a combination of different TF, IDF, and WC values using the best classification algorithm from experimental results 1 and 2, using the best Character NGram feature based on the experimental results in step three.

SVM (Support Vector Machines) works to find the hypothesis that reducing the boundary between correct errors in h will make it in the test data that is not visible, and errors in the training data. Sequential Minimal Optimization (SMO) is an implementation of the SVM classification of WEKA tools. SMO was developed for numerical prediction and data classification by building N-dimensions by optimally separating data into two categories [14]-[15]. SVM achieves the best performance in text classification tasks because SVM's ability to eliminate the need for feature selection means that SVM eliminates the high dimensional feature space that results from frequent occurrences of words in the text. Besides, SVM automatically finds proper parameter settings.

Naïve Bayes is one of the statistical classifiers, which can predict the probability of class membership of tuple data under the calculation of the probability of going into a particular class. The classifier discovered by Thomas Bayes in the 18th century is based on the Bayes theorem. In a comparative classification research report, a simple bayesian or commonly known as the Naïve Bayes classifier, shows high accuracy and speed when used in large databases [16].

J48 is one of the classifiers in data mining and part of a simple C4.5 decision tree. C4.5 builds a decision tree based on a set of labeled data inputs. A decision tree is a prediction model that uses tree structure or hierarchical structure. The decision tree has a concept in turning data into trees and decision rules [17].

ZeroR is the simplest classification method that depends on the target and ignores all predictors. ZeroR only predicts the majority category (class). Although there is no predictability in ZeroR, it is useful to determine baseline performance as a benchmark for other classification methods. Algorithm Build frequency tables for targets and choose their values most often. Contributors of Predictors Nothing can be said about the contributions of predictors to the model because ZeroR does not use one of them. The ZeroR evaluation model only predicts the majority class correctly. As mentioned earlier, ZeroR is only useful for determining baseline performance for other classification methods [18].

Term Frequency (TF) represents the frequency of specific keywords. Based on the data in the table, several words are usually found more often in one dialect than another dialect. So the weight of TF is used to show the level of importance of words in the text of the sentence. Inverse Document Frequency (IDF) scales how often a word appears in different sentence text (more than one dialect), which means words that appear in many dialects that cannot be used as features [19].

3. Results and Discussion

The data used are sentences in three languages, namely Indonesian, Malay Pontianak, and Malay Sambas. Each language consists of 1,000 sentences, so the total sentences used are 3,000 sentences, as in Table 1. The length of sentences used in this study ranged from 1-30 words, with an average of 18 words. The number of attributes (tokens) used is 4,349 tokens. Figure 2 shows the distribution of the number of words in the sentences used.

Table 1. Number of sentences used

Language	Sentence
Indonesian	1.000
Malay Pontianak	1.000
Malay Sambas	1.000

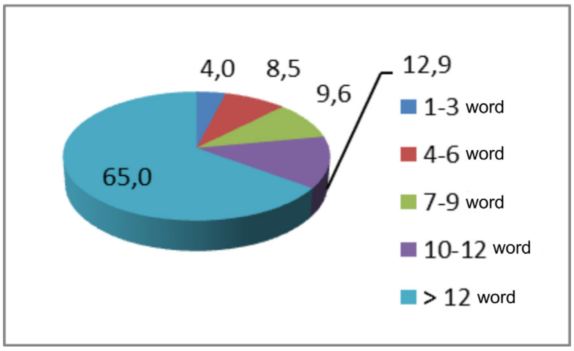
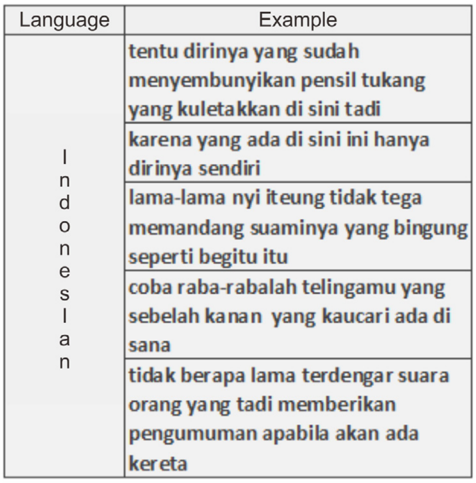


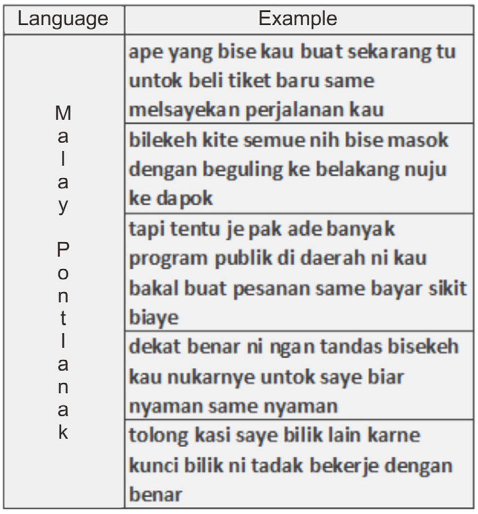
Figure 2. Distribution of the number of words in a sentence

Similarities between languages are characterized by words in sentences in a language. Similarities between Indonesian, Malay Pontianak, and Malay Sambas can be seen in the example of a few sentences in Figure 3-5. From these examples, it can be seen that several regional words are the same as the Indonesian language, for example, you, me, right, of course, an instrument, etc..



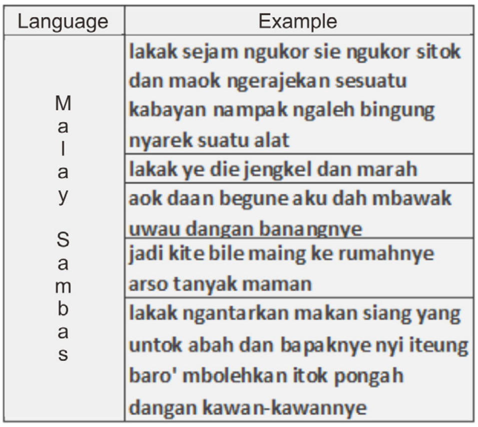
Language	Example
Indonesian	tentu dirinya yang sudah menyembunyikan pensil tukang yang kuletakkan di sini tadi
	karena yang ada di sini ini hanya dirinya sendiri
	lama-lama nyi iteung tidak tega memandang suaminya yang bingung seperti begitu itu
	coba raba-rabalah telingamu yang sebelah kanan yang kaucari ada di sana
	tidak berapa lama terdengar suara orang yang tadi memberikan pengumuman apabila akan ada kereta

Figure 3. Examples of Indonesian sentences



Language	Example
Malay Pontianak	ape yang bise kau buat sekarang tu untok beli tiket baru same melsayekan perjalanan kau
	bilekeh kite semue nih bise masok dengan beguling ke belakang nuju ke dapok
	tapi tentu je pak ade banyak program publik di daerah ni kau bakal buat pesanan same bayar sikit biaye
	dekat benar ni ngan tandas bisekeh kau nukarnye untok saye biar nyaman same nyaman
	tolong kasi saye bilik lain karne kunci bilik ni tadak bekerje dengan benar

Figure 4. Example of Malay Pontianak sentence



Language	Example
Malay Sambas	lakak sejam ngukor sie ngukor sitok dan maok ngerajekan sesuatu kabayan nampak ngaleh bingung nyarek suatu alat
	lakak ye die jengkel dan marah
	aok daan begune aku dah mbawak uwau dangan banangnye
	jadi kite bile maing ke rumahnye arso tanyak maman
	lakak ngantarkan makan siang yang untok abah dan bapaknye nyi iteung baro' mbolehkan itok pongah dangan kawan-kawannye

Figure 5. Example of Malay Sambas sentence

Table 2 reports the results for various classifications that were tried using the StringToWordVector filter with WordTokenizer, which is one of the WEKA features for extracting words as a feature of sentence strings. From table 2, it appears that SMO is the best classifier with an accuracy rate of 95.7% for 10-fold cross-validation and 94.4% for 60%: 40% of test and training data. While the lowest accuracy is obtained on the use of ZeroR for both experimental methods. From the ZeroR baseline, using SMO can increase accuracy by (0.957-0.333) / 0.333 = 187%.

Table 2. Classifier accuracy with different training methods

Classifier	10-fold cross-validation	60% : 40%
NaïveBayes	0,923	0,921
SMO	0,957	0,944
J48	0,836	0,812
ZeroR	0,333	0,325

The results of the classifier using Character NGram Tokenizer with Min = 3 and Max = 3 can be seen in Table 3. The Word Tokenizer method is a method for separating a series of words into tokens in the form of words or punctuation. The results of using this method show that the SMO classifier has a higher yield than the other classifier. Ngram Word Tokenizer has a function similar to Word Tokenizer. The difference lies in the function to enter the word order with the maximum and the minimum number of words, while Character NGram Tokenizer counts the combination of first, second, and so on, in sentence strings. The results of using this method show that the SMO classifier has a higher yield than the other classifier too.

Table 3. Classifier accuracy with word tokenizer and NGram tokenizer characters

Classifier	Word Tokenizer	Character NGram Tokenizer (3-gram)
NaïveBayes	0,921	0,931
SMO	0,944	0,965
J48	0,812	0,915
ZeroR	0,325	0,325

60% Experiment: 40% of this test and training data shows that all classifiers, except ZeroR have increased accuracy when using Character NGram Tokenizer compared to Word Tokenizer.

The first experiment to choose the best classification to identify Indonesian and Malay languages shows that the best classification of machine learning is the SMO algorithm. This study uses a WEKA StringToWordVector filter with Word Tokenizer that enters text into words between delimiters. But it was recommended to try n-Gram characters as units, not words as units. We used Character N GramTokenizer to divide strings into n-grams with maximum and minimum values. We set the Max value to 1, as well as the Min value on the model based on uni-gram; on the bigram model, we set Max to be 2, as well as the value of Min; on the tri-gram model, we set Max to be 3, as well as the Min value; in the 4-gram model, we set the Max value to 4, as well as the Min value. The results show that 4-gram models may not be appropriate because the training data is not large enough. Experiment using gram values that vary when evaluating by percentage of 60:40 from the training set and 10-fold cross-validation are shown in table 4.

Table 4. Feature accuracy with different training methods

Features	60% : 40%	10-fold cross-validation
Charater UniGram	0,809	0,828
Charater BiGram	0,935	0,943
Charater TriGram	0,965	0,966
Charater FourGram	0,965	0,967

The last experiment used a combination of different TF, IDF, and WC values using the best classification algorithm from experimental results 1 and 2, using the best Character NGram feature based on the results of previous experiments. The results show that the greatest accuracy value, 0.965, is obtained in combination TF = TRUE, IDF = FALSE dan WC = TRUE and TF = FALSE, IDF = FALSE and WC = TRUE. So it can be dreamed that a combination of values TF, IDF, and WC the best is IDF = FALSE dan WC = TRUE.

Table 5. Combination accuracy TF/IDF/WC

TF	IDF	WC	Precision
TRUE	TRUE	TRUE	0,960
TRUE	TRUE	FALSE	0,965
TRUE	FALSE	TRUE	0,969
TRUE	FALSE	FALSE	0,965
FALSE	TRUE	TRUE	0,960
FALSE	TRUE	FALSE	0,965
FALSE	FALSE	TRUE	0,969
FALSE	FALSE	FALSE	0,965

From the WEKA Confusion Matrix data testing 300 sentences each of 100 sentences that have been labeled as discussed, it is found that out of 100 Indonesian languages, two sentences are recognized as Malay Pontianak and two other sentences identified as Malay Sambas. Out of 100 Pontianak languages, one sentence is recognized as Indonesian and three sentences as Malay Sambas. While from 100 Malay Sambas languages, two sentences are recognized as Indonesian, and four sentences are recognized as Malay Pontianak.

4. Conclusion

This study classifies regional languages that are similar to Indonesian, namely Malay Pontianak and Malay Sambas, for the purpose of language identification. From Naïve Bayes, SMO, J48, and ZeroR classifiers, it was found that SMO was the most accurate classifier with an accuracy rate of 95.7% for 10-fold cross-validation and 94.4% for 60%: 40%. The best tokenizer in this classification is Character Ngram. All classifiers, except ZeroR have increased accuracy when using Character NGram Tokenizer compared to Word Tokenizer. The best features of this system are the TriGram and FourGram Character. TriGram is preferred because it requires smaller training data. The last experiment showed that the highest accuracy value, 0.965, was obtained in the combination of IDF = FALSE and WC = TRUE, regardless of the condition of TF.

References

[1] S. Sudaryanto, "Tiga Fase Perkembangan Bahasa Indonesia (1928—2009): Kajian Linguistik Historis", Aksis Jurnal Pendidikan Bahasa dan Sastra Indonesia , Vol. 2. No 1, 2018.

[2] E. Novianti, "Menilik Nasib Bahasa Melayu Pontianak". International Seminar Language Maintenance and Shiff. Pp. 70- 74. 2011.

[3] M.Z. Wiguna, "Tindak Tutur Bahasa Melayu Dialek Sambas di Kabupaten Sambas", Jurnal Pendidikan Bahasa, Vol. 5, No. 2, Desember 2016

[4] C. Goutte, S. Léger, S. Malmasi, and M. Zampieri, "Discriminating Similar Languages: Evaluations and Explorations", LREC, 2016.

[5] F. Omar, Zaidan and C.C. Burch. "Arabic dialect identification". Computational Linguistics, 40(1):171–202. 2014.

[6] M. Lu and M. Mohamed. "Lahga: Arabic dialect classifier". Report, December 13, 2011.

[7] H. Elfardy and M. Diab. "Sentence level dialect identification in arabic". In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, page 456-461. 2013.

[8] N.E. Safitri, A. Zahra, and M. Adriani, "Spoken Language Identification with Phonotactics Methods on Minangkabau, Sundanese, and Javanese Languages," Procedia Computer Science, vol. 81, pp. 182–187, 2016.

[9] J. Zhao, S. Mudgal, and Y. Liang, "Generalizing Word Embeddings using Bag of Subwords," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[10] G. H. Rachman, M. L. Khodra, and D. H. Widyantoro, "Word Embedding for Rhetorical Sentence Categorization on Scientific Articles," *Journal of ICT Research and Applications*, vol. 12, no. 2, p. 168, 2018.

[11] T.O. Ayodele. "Types of machine learning algorithms". 2010.

[12] M. Hall, E.Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. "The weka data mining software: An update". SIGKDD

Explorations, 11(1):10–18. 2009.

[13] I. H. Witten, and E. Frank. “Data mining: Practical machine learning tools and techniques with Java implementations”. San Francisco, CA: Morgan Kaufmann. 2016.

[14] T. Jo, “Representation of Texts into String Vectors for Text Categorization”. Journal of Computing Science and Engineering, 4(2), 110-127. 2010

[15] T. Joachims. “Text categorization with support vector machines: Learning with many relevant features”. 1998.

[16] F. Handayani and F. S. Pribadi, “Implementasi Algoritma Naïve Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110,” J. Tek. Elektro, 2015.

[17] S. Diwandari and N. A. Setiawan, “Perbandingan Algoritme J48 dan Nbtree untuk Klasifikasi Diagnosa Penyakit Pada Soybean,” Semin. Nas. Teknol. Inf. dan Komun., 2015.

[18] C. Nasa and S. Suman, “Evaluation of Different Classification Techniques for WEB Data,” Int. J. Comput. Appl., 2012.

[19] B.G. Gebre, M.Zampieri, P. Wittenburg, and T. Heskes. “Improving native language identification with tf-idf weighting”. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 216–223. Association for Computational Linguistics. 2013.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Mapping Land Suitability for Sugar Cane Production Using K-means Algorithm with Leaflets Library to Support Food Sovereignty in Central Java

Pramudhita Tunjung Seta, Kristoko Dwi Hartomo*
Informatics Engineering Study Program
Universitas Kristen Satya Wacana
Salatiga
*kristoko@uksw.edu

Abstract-Indonesia is the largest sugar importer country in the world, this is contrary to the government's desire to realize sugar self-sufficiency. To overcome the dependence on sugar imports in order to support national food sovereignty, geographic information system technology (GIS) can be used to present information as material for consideration by the government in determining policies on the management of sugar cane land resources. The K-means algorithm is used to group regions according to production level, while the Matching method is for evaluating the suitability of sugarcane land. Presentation of data in the form of map visualization on the web using a new model in processing land data, where this model processes production grouping data, and land suitability class data in the form of GeoJSON then mapped with the help of Leaflets. This new model enables dynamic land data processing and visualization in the form of interactive maps. The results of the EUCS test for GIS mapping of Land Suitability and Cane Production are 3.23 (Satisfied) of the total score of 4, so this system can be accepted by the user.

Keywords: mapping; land suitability; sugarcane production; K-means; food sovereignty

1. Introduction

Sugar cane (Saccharum officinarum L.) is a type of plantation commodity as a raw material for making sugar. Based on the Decree of the Coordinating Minister for the Economy No. Kep-28 / M.EKON / 05/2010 concerning the Staple Food Stabilization Coordination Team, which includes staples as rice, sugar, cooking oil, flour, soybeans, beef, chicken meat, and chicken eggs [1]. As a source of calories other than rice, corn and tubers, sugar is one of the basic needs that is consumed in large quantities both from home to industrial scale [2]. The high level of consumption sugar which is not balanced with the amount of sugar production has resulted in Indonesia having to import to meet the demand for sugar consumption.

Based on data released by Statista, for the period of 2017/2018 Indonesia is the largest sugar importer country in the world with a number of sugar imports reaching 4.45 million tons. This figure beat China and the United States (US), which were 4.2 million tons and 3.11 million tons, respectively [3]. If this is not immediately addressed, Indonesia will suffer heavy losses due to having to depend on imports.

Table 1. Indonesian Sugar Cane Area and Its Production Quantity 2017 [4]

Province	Area (Ha)	Production (Tons)
North Sumatra	7,806	29,664
South Sumatra	21,967	99,860
Lampung	121,346	768,939
West Java	20,774	86,2016
Central Java	52,106	202,956
Yogyakarta	3,155	12,226
East Java	203,471	1,186,515
Gorontalo	7,764	44,298
South Sulawesi	11,690	34,786

Based on Table 1, Central Java is the third largest sugar production center after East Java and Lampung, so that Central Java is making a significant contribution to the sugar industry in Indonesia. However, sugar production in Central Java is still less than the government's target for sugar self-sufficiency. So that in 2018, the Governor of Central Java established the Central Java Provincial Regulation, number 1 Regarding Increasing Sugarcane Productivity, in which there was a program of empowerment of farmers by the local government, one of which was to provide or expand the area of sugarcane land [5].

According to Surono (2006), in a journal article entitled Sugar Self-Sufficiency Policy in Indonesia, one of the things behind the importance of sugar self-sufficiency in Indonesia is to maintain food sovereignty, because sugar is one of the main foodstuffs with a high level of need, so it needs to always available in sufficient quantities and reasonable price levels. To achieve the sugar self-sufficiency target, a method that is able to map the yield is needed to find out areas with less than maximum sugar cane production so that effective handling can be done [6].

One technology that can be used to map spatial data is Geographic Information Systems (GIS). The use of GIS in planning crop production management is needed in agriculture [7]. Government agencies such as the Ministry of Agriculture currently have a GIS for Sugar Cane Monitoring. However, the GIS only focuses on the composition of the area based on the growth phase of sugarcane, so the analysis of sugarcane production in each region is limited. With the K-means algorithm, a grouping of regions based on the level of sugarcane production can be done. In an effort to maximize the production of sugarcane, agricultural planning is also needed in accordance with the capabilities of the land. For land suitability assessment for sugarcane, it can be identified through land suitability evaluation by classifying potential land into S1 (very suitable), S2 (suitable) and S3 (marginal appropriate) classes [8].

Based on these problems, we conducted a study using Library Leaflets and OpenStreetMap to cluster regions based on the level of sugarcane production using the K-means algorithm and to map land suitability for sugarcane in Central Java Province, so that web-based GIS is produced which is expected to be able to present information that can used as a material consideration for the government to determine policies on the management of sugar cane land resources in order to realize food sovereignty in the case of sugar commodities.

Previous research that has a connection with this research, titled Clustering using K-means and Fuzzy C-Means on Food Productivity clustered the productivity of one food commodity namely rice using the K-means algorithm and Fuzzy C-means with Excel Software for processing data. Obtained three clusters with cluster 1 and cluster 2 having low productivity, so it can be concluded that the majority of rice productivity per province in Indonesia is classified as low [9].

Previous research, Implementation of K-means Algorithm for Mapping of Harvest Productivity in Karawang District applies a cluster technique using K-means algorithm to map rice harvest productivity data by dividing data into 3 groups: less than the target, according to the target, and exceeding the target using attributes rice planting and production area. The results of the mapping are visualized into a map on the web [10].

The study entitled Spatial Model Design of Landslide Vulnerability Early Detection with Exponential Smoothing Method Using Google API produces spatial models of early detection of landslide disasters based on rainfall data and soil condition data using the Single Exponential Smoothing method which is implemented using the Google API. This model is able to predict areas prone to landslides [11].

The study titled Cluster Analysis Using Fuzzy C-means and K-means Algorithms for Clustering and Mapping of Agricultural Land in Southeast Minahasa conducted cluster analysis to determine the area of agricultural land for paddy, paddy, corn and cassava commodities based on the attributes of harvested area, area planting, and production. The results of this study are the grouping of agricultural areas based on commodities and their attributes. Further studies need to be done by calculating the slope of the land, soil structure, compatibility of the commodity with the land [12].

The study entitled Evaluation of Land Suitability for Rice Commodities by Utilizing the Application of Geographic Information Systems (GIS) in Central Lombok Regency aims to determine the land suitability classes for lowland rice and upland rice in Central Lombok Regency based on topographic aspects, soil type and climate. The method used is the Matching method by adjusting the existing land suitability class criteria. The results of this study are visualization using ArcView GIS based on the suitability class for paddy and upland rice plants and their extent [13].

The renewal and superiority of this research compared to previous research is optimizing the process of presenting and processing land data by utilizing a new model in processing land data. With this new model, the data from the processing of the K-means algorithm to determine groups of regions based on the level of production and the Matching method to determine areas based on land suitability, can be converted into the form of GeoJSON. By utilizing Library Leaflet technology and OpenStreetMap as its basemap, the GeoJSON data is visualized in the form of an interactive web-based map, which makes data processing can be done dynamically and efficiently. Information generated from the combination of the visualization of production maps and land suitability is expected to be an input in the process of managing sugar cane land resources.

K-means algorithm is an algorithm that groups data into several groups based on similar characteristics, so that one group with another group has different characteristics. The function of an object in K-means can be determined by equation (1).

$$\boldsymbol{d_{ij}} = \sqrt{\textstyle\sum_{k=1}^{p}\{\boldsymbol{x_{ik}} - \boldsymbol{x_{jk}}\}^{2}} \qquad (1)$$

note :
d_{ij} = Distance between object i and j
P = Data dimension
x_{ik} = The coordinates of object i on dimension k
x_{jk} = The coordinates of object j on the dimension k

Table 2. Land Characteristics and Suitablility Level [17]

Numb	Land Characteristics	Land Suitability Class				
		S1	S2	S3	N1	N2
1	Annual average temperature (^{o}C)	24-30	>30-32 22 < 24	>32-34 21-<22	Td	>34 >21
2	Rainfall / year (mm)	1200-2500	1300-<1500	>2500-3000 1000-<1300	-	>3000 <1000
3	Slope (%)	<8	8-15	>15-30	>30	-

Food sovereignty, cannot be separated from the potential of land resources for plant growth. Surveys and inventory of land resources need to be emphasized to support agriculture. However, this land resource survey is still limited. Land evaluation needs to be done on Land and Land Resource Data Data so that it can produce land suitability information [14].

One of the land suitability classification systems according to FAO (1976) in Sarwono and Widiatmaka (2011) [15] consists of land suitability classes. Consisting of :

- Class S1: Very suitable.
- Class S2: Quite appropriate or moderate compatibility.
- Class S3: Marginal fit or low suitability.
- Class N1: Not suitable at this time.
- Class N2: Does not suit permanently.

To determine land suitability must be based on plant growth requirements. Land suitability evaluation is carried out using the Matching method based on the requirements for growing sugarcane in Table 2.

The elements that are important for the growth of sugarcane are rainfall, sunlight, wind, temperature, and slope. Therefore in this study the parameters used are temperature, rainfall, and slope.

2. Method

The research process of mapping sugarcane production using the K-means algorithm and land suitability with the Matching method is carried out through several stages that are interrelated with each other. Following the research methodology chart shown in Figure 1.

- Stage 1: The first stage is to identify problems regarding the high import of sugar cane which is contrary to the government's desire for self-sufficiency in sugar. Then a literature study is conducted to find references to solve problems regarding the case.
- Stage 2: The data used in this study are secondary data. To determine the suitability of land for sugarcane, data on rainfall, annual average temperature, and slope are obtained from the Central Statistics Agency, and the Exploration Soil Resource Map scale of 1: 1,000,000 (Center for Soil and Agro-climate Research, 2000).

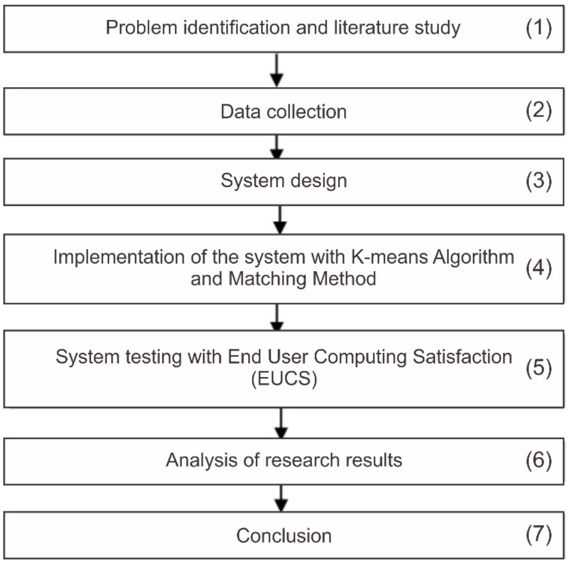

Problem identification and literature study (1)
Data collection (2)
System design (3)
Implementation of the system with K-means Algorithm and Matching Method (4)
System testing with End User Computing Satisfaction (EUCS) (5)
Analysis of research results (6)
Conclusion (7)

Figure 1. Research Stages

While the data for regional grouping based on the level of sugarcane production uses variables in the form of planting area (ha), harvested area (ha), and production (tons) obtained from the 2015-2017 Indonesian Plantation Statistics.

- Stage 3: The software design process is carried out using modeling from the needs analysis. At this stage three Unified Modeling Language (UML) Diagrams are generated, namely Usecase Diagrams, Activity Diagrams, and Class Diagrams.
- Stage 4: System Design for grouping production areas with K-means algorithm and land suitability with Matching method is built using the PHP programming language, MYSQL as Database, and mapping using Library Leaflet and OpenStreetMap as its basemap.
- Stage 5: System feasibility testing uses End user Computing Satisfaction (EUCS) with dimensions of content, accuracy, format, ease of use, timeliness. With a Likert scale as a rating scale consisting of Very Satisfied (4), Satisfied (3), Dissatisfied (2) and Very Dissatisfied (1).
- Stage 6: Conduct analysis of the results of the implementation of the K-means algorithm and the Matching method and the results of the EUCS test.
- Stage 7: Drawing conclusions from the results of research

UML is used in making Use Case Diagrams, Activity Diagrams, and Class Diagrams for system design. The Usecase Diagram of the application to be built can be seen in Figure 2.

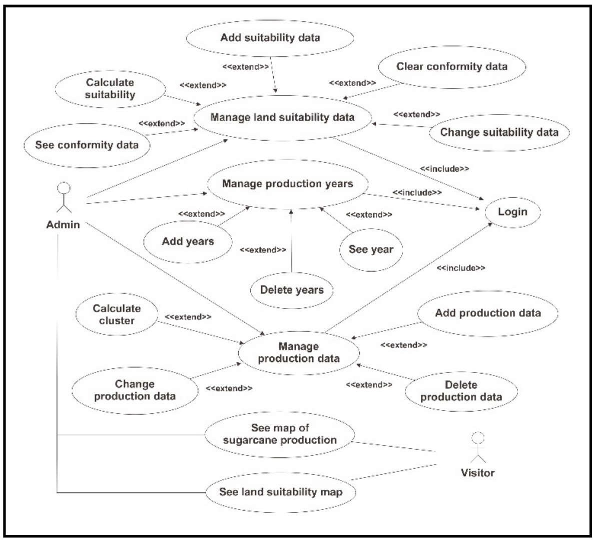


Figure 2. Usecase Diagram

Figure 2 explained that the system has two users namely admin and user. Admin has access rights to manage sugarcane production data, manage land suitability data, manage users, perform cluster calculations for sugarcane production data, perform land suitability calculations, and view map visualizations. While users can only see maps that have been visualized on web pages. Class Diagrams are used to illustrate the structure of a system that is defined through classes according to the system requirements described in Figure 3.

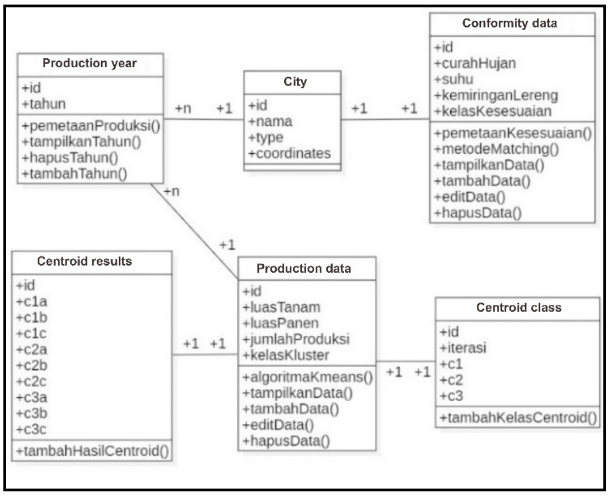


Figure 3. Class Diagram

Based on Figure 3, the City Class contains attributes about cities or districts in Central Java Province along with polygon coordinates to map each area on the map. Production Data Class contains sugarcane production data. The Cluster class attribute contains the results of class grouping from the K-means algorithm. The Centroid Result Class contains the minimum data distance to the centroid while the Centroid Class Class contains the data grouping on the centroid. Conformity Data Class serves to accommodate the suitability data in each city that only has one Land Suitability data.

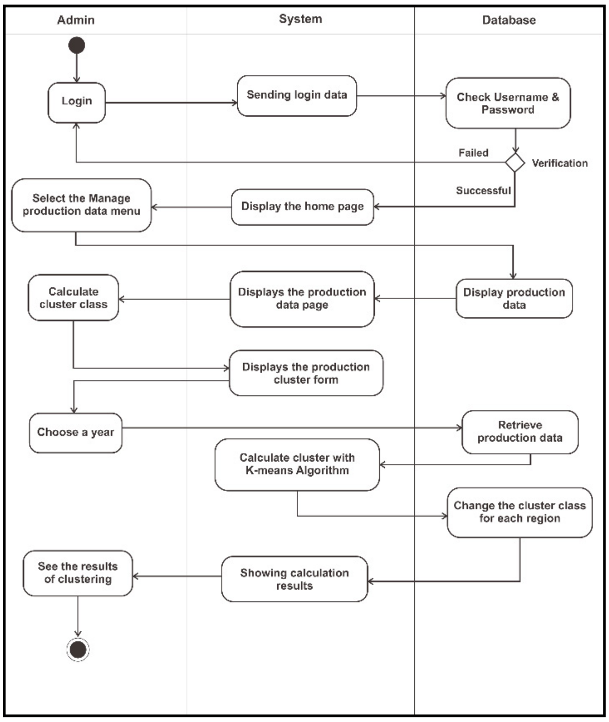

Admin
System
Database
Login
Sending login data
Check Username & Password
Failed
Verification
Successful
Select the Manage production data menu
Display the home page
Calculate cluster class
Displays the production data page
Display production data
Displays the production cluster form
Choose a year
Retrieve production data
Calculate cluster with K-means Algorithm
Change the cluster class for each region
See the results of clustering
Showing calculation results

Figure 4. Administrator Activity Diagram to Perform Clustering

Figure 4 is an Activity Diagram when the admin does clustering on sugarcane production data. Previously, the admin must log in first. After successfully logging in, the admin can manage the data by selecting the Manage data menu. To cluster the sugarcane production data, the Admin enters the sugarcane production data management menu and chooses the calculate class cluster menu. Admin then chooses the year of the data to be processed, after that the system will perform K-means calculations on sugarcane production data in accordance with the year chosen by the Admin. The grouping data is then stored in a database.

3. Result and Discussion

The result of the system built is Geographic Information System (GIS) which is able to manage production data using the K-means algorithm and determine the land suitability class using the Matching method then visualize it on the web. The process of presenting data in the form of map visualization on the web uses a new model as illustrated in Figure 5.

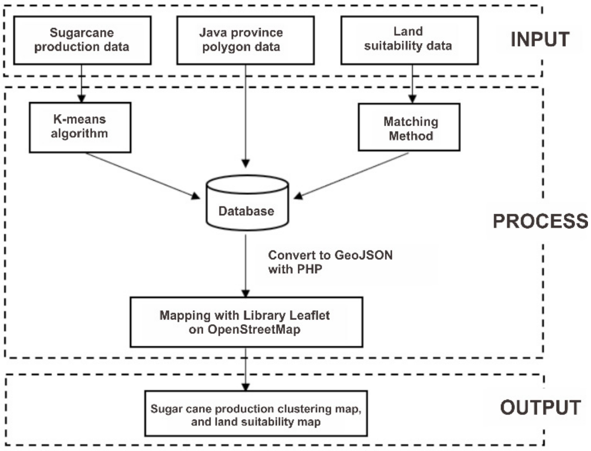

Sugarcane production data
Java province polygon data
Land suitability data
INPUT
K-means algorithm
Matching Method
Database
PROCESS
Convert to GeoJSON with PHP
Mapping with Library Leaflet on OpenStreetMap
Sugar cane production clustering map, and land suitability map
OUTPUT

Figure 5. Land Data Management Model

Figure 5 is a land data management model where the input is sugarcane production data and land suitability data. Data management is done using the PHP programming language in the system. Cluster calculations use the K-means algorithm while evaluating land characteristics using the Matching method. The result of data processing is numerical data which is then stored in a database and processed by the system in the form of GeoJSON. Data that has been changed to GeoJSON is then displayed interactively using the Library Leaflet in the form of a map in the system so that the information can be analyzed by experts in land management.

1) K-means Algorithm and Matching Method

The K-means algorithm is modeled based on the K-means algorithm flowchart illustrated in Figure 6.

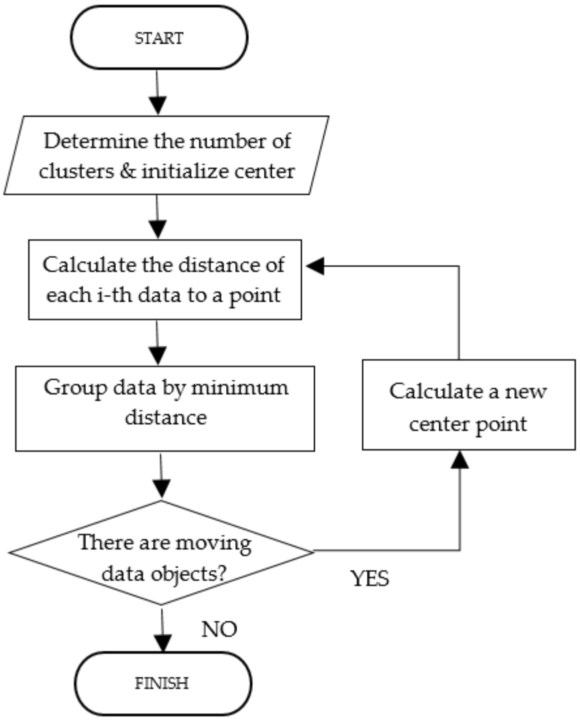

START
Determine the number of clusters & initialize center
Calculate the distance of each i-th data to a point
Group data by minimum distance
Calculate a new center point
There are moving data objects?
YES
NO
FINISH

Figure 6 K-means Algorithm Flowchart

Based on Figure 6, the K-means algorithm program flow chart is arranged. For initial point initialization the initial centroid is determined using the Simple Random Sampling method, which is taking random sample data as seen in pseudocode 1.

```
Input :pusat[0] ← min(data)
       pusat[1] ← random(data)
       pusat[2] ← max(data)
   c1a ← data[pusat[0]][luas_tanam]
   c1b ← data[pusat[0]][luas_panen]
   c1c ← data[pusat[0]][jumlah_produksi
   c2a ← data[pusat[1]][luas_tanam]
   c2b ← data[pusat[1]][luas_panen]
   c2c ← data[pusat[1]][jumlah_produksi]
   c3a ← data[pusat[2]][luas_tanam]
   c3b ← data[pusat[2]][luas_panen]
   c3c ← data[pusat[2]][jumlah_produksi]
```

Pseudocode 1 K-means Algorithm Initialization of Center Point

Pseudocode 1 is the initial initialization process for the center point. The centroid center 1 is taken from the smallest production data center, the random centroid 2 data center, and the biggest centroid data center 3 production. Next is the calculation of the distance of each i-th data to the central point.

```
n ← jumlah data
FOR i ← 0 TO i < n DO
    hc1 ← sqrt(pow(data[luas_tanam]-
    c1a),2)+pow((data[luas_panen]-c1b),2)+
    pow((data[jumlah_produksi]-c1c),2)
    hc2 ← sqrt(pow(data[luas_tanam]-
    c2a),2)+pow((data[luas_panen]-c2b),2)+
    pow((data[jumlah_produksi]-c2c),2)
    hc3 ← sqrt(pow(data[luas_tanam]-
    c3a),2)+pow((data[luas_panen]-c3b),2)+
    pow((data[jumlah_produksi]-c3c),2)
    IF hc1 <= hc2
      IF hc1 <= hc3
        Data[i][kelas_klaster] ← 1
        Arr_c1[i] ← 1
      Else Arr_c1[i] ← 0 END IF
      ELSE Arr_c1[i] ← 0
      END IF
      IF hc2 <= hc1
        IF hc2 <= hc3
        Data[i][kelas_klaster] ← 1
          Arr_c2[i] ← 1
      Else Arr_c2[i] ← 0 END IF
        ELSE Arr_c2[i] ← 0
        END IF
      IF hc3 <= hc1
        IF hc3 <= hc2
          Data[i][kelas_klaster] ← 1
          Arr_c3[i] ← 1
        Else Arr_c3[i] ← 0
      END IF
```

```
25.       ELSE Arr_c3[i] ← 0
26.       END IF
27.    arr_c1_temp[i] ← data[luas_tanam]
28.    arr_c2_temp[i] ← data[luas_panen]
29.    arr_c3_temp[i] ← data[jumlah_produksi]
30. END FOR
```

Pseudocode 2 K-means Algorithm Calculating Euclidean Distance

Euclidiean distance is calculated in lines 2 to 4 to measure the distance from the center of the cluster so that the distance obtained is hc1, hc2, and hc3. Then do the comparison of the distance of each class and then grouped into classes based on the minimum distance in lines 5 to 26. this group shows that the data has a distance from the nearest cluster center. After the members of each cluster are known, the next process is to determine the center of the new cluster.

```
1.  FOR i ← 0 TO count(arr_c1) < n DO
2.    Arr[i] ← arr_c1_temp[i]*arr_c1[i]
3.    IF arr_c1[i] == 1 THEN jum++ END IF
4.  END FOR
5.  C1a_b ← array_sum(arr)/jum
6.  FOR i ← 0 TO count(arr_c2) < n DO
7.    Arr[i] ← arr_c2_temp[i]*arr_c1[i]
8.    IF arr_c1[i] == 1 THEN jum++ END IF
9.  END FOR
10. C1b_b ← array_sum(arr)/jum
11. FOR i ← 0 TO count(arr_c3) < n DO
12.   Arr[i] ← arr_c3_temp[i]*arr_c1[i]
13.   IF arr_c1[i] == 1 THEN jum++ END IF
14. END FOR
15. C1c_b ← array_sum(arr)/jum
```

Pseudocode 3 K-means Algorithm Calculating Central Centroid 1

Pseudocode 3 is a new centroid calculation process for centroid center 1. To calculate centroid centers 2 and 3, the same calculation is done with Pseudocode 3. The next process is checking the data that moves class.

```
1. IF c1_sebelum == c1_sesudah OR c2_sebelum ==
c2_sesudah OR c3_sebelum == c3_sesudah
2.    selesai ← TRUE
3.  ELSE selesai ← FALSE
4.  Iterasi++
```

Pseudocode 4 K-means Algorithm for Data Transfer Check

Pseudocode 4 is a process of checking data transfer. Comparison with the previous group was conducted. If there is a data transfer then the calculation is performed again on Pseudocode 2, while if there is no data transfer, the iteration process stops.

Matching methods to determine land suitability classes are arranged as a program flow model based on sugarcane growing requirements in Table 2, so that rules are obtained as seen in Pseudocode 5.

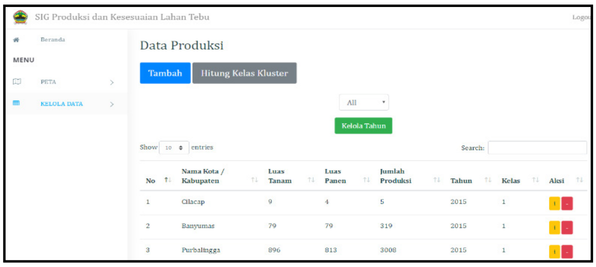
```
Input: curahHujan ← data curah hujan tiap wilayah
kemiringanLereng ← data kemiringan lereng tiap
wilayah
suhu ← data suhu tiap wilayah
n ← jumlah data
1. FOR i ← 0 TO i < n DO
2.   IF curahHujan[i] >= 1500 AND curahHujan[i]
<=2500 AND suhu[i] >= 24 AND suhu[i] <= 30 AND
     kemiringanLereng[i] < 8 THEN
        kelas[i] ← 'S1'
3.    END IF
4.   ELSE IF curahHujan[i] >3000 OR
curahHujan[i]  < 1000 OR suhu[i] > 34 OR
suhu[i] < 21 THEN
     Kelas[i] ← 'N2'
5.   END IF
6.     ELSE IF kemiringanLereng[i] > 30 THEN
        Kelas[i] ← 'N1'
7.      END IF
8.   ELSE IF curahHujan[i] > 2500 AND
curahHujan [i]<= 3000 OR curahHujan[i]
>= 1000 AND curahHujan[i] < 1300 OR
suhu[i] > 32 AND suhu[i] <= 34 OR suhu[i]
== 21 OR kemiringanLereng[i] > 15 AND
kemiringanLereng[i] <= 30 THEN
        kelas[i] ← 'S3'
9.    END IF
10.  ELSE IF curahHujan[i] >= 1300 AND
curahHujan[i] <= 1500 OR suhu[i] > 30 AND
suhu[i] <= 32 OR suhu[i] >= 22 AND suhu[i]
< 24 OR kemiringanLereng[i] >= 8 AND
kemiringanLereng[i] <= 15 THEN
        kelas[i] ← 'S2'
11.   END IF
12.  ELSE
        kelas[i] ← "Tidak Terklasifikasi"
13.   END IF
14. END FOR
```

Pseudocode 5 Matching Methods

Pseudocode 5 is the result of implementation of the Matching method. Classification of land suitability classes is carried out through the process matching stage so that rules are obtained for each land suitability class. Line 1 is matched for all land data so that each land is classified in the appropriate land suitability class.

2) Mapping of Sugar Cane Production and Conformity

The process of managing data in order to be visualized in the form of interactive maps requires the help of Library Leaflets with OpenStreetMap as the base map. The following is a mapping process for sugarcane production data in 2015, which begins with entering the production data management menu and inputting production data in the form of harvested area, planted area, and the amount of production obtained from the 2015-2017 Indonesia Plantation Statistics. manage production data menu. In this menu, the admin can delete, edit and add production data. The admin can then calculate the cluster class from data in the selected year by selecting the calculate class cluster menu.



No	Nama Kota / Kabupaten	Luas Tanam	Luas Panen	Jumlah Produksi	Tahun	Kelas	Aksi
1	Cilacap	9	4	5	2015	1	
2	Banyumas	79	79	319	2015	1	
3	Purbalingga	896	813	3008	2015	1	

Figure 7. Application Interface of Sugar Cane Production Management

Table 3. Cluster Start Center

Cluster	City	A	B	C
c1	Cilacap	9	4	5
c2	Grobogan	2212	3621	14586
c3	Rembang	11697	10515	50345

Information: A = Planting Area; B = Harvested area; C = Production amount

Table 4. Result of Production Data Cluster

Numb	City	Centroid 1	Centroid 2	Centroid 3
1	Cilacap	2692,49	13562,72	43997,71
2	Banyumas	2362,21	13232,50	43667,51
3	Purbalingga	568,27	10340,91	40770,97
4	Banjarnegara	1679,37	12549,64	42984,20
5	Kebumen	2174,07	13044,94	43479,29
6	Purworejo	391,36	10725,77	41150,15
7	Kab. Magelang	303,54	11059,93	41489,09
8	Boyolali	1084,19	11952,43	42385,58
9	Klaten	2058,72	9083,53	39495,35
10	Wonogiri	1866,47	9020,69	39449,72
11	Karanganyar	6637,18	4257,79	3475,05
12	Sragen	32577,33	21717,39	8761,66
13	Grobogan	4161,14	6809,36	37218,28
14	Blora	12503,17	2011,79	28912,97
15	Rembang	22448,94	11586,40	18862,69
16	Pati	50045,39	39174,59	8761,66
17	Kudus	8195,83	2769,69	33143,10
18	Jepara	6887,26	4101,86	34477,74
19	Demak	2670,13	13540,40	43975,20
…				
28	Semarang	1950,56	12820,83	43255,87

A description of the steps for manually calculating the K-means algorithm for sugarcane production data in Central Java in 2015 will be explained at this stage. The initial center of the cluster is composed of three clusters.

Table 3 is the initial cluster center table where c1 is taken from the smallest production data, c2 random data, and c3 is the largest production data. The next step is to calculate the euclidean distance, i.e. the distance of each i-th data to the center point. The following is an example of calculating the distance of Banyumas city production data to the cluster center point.

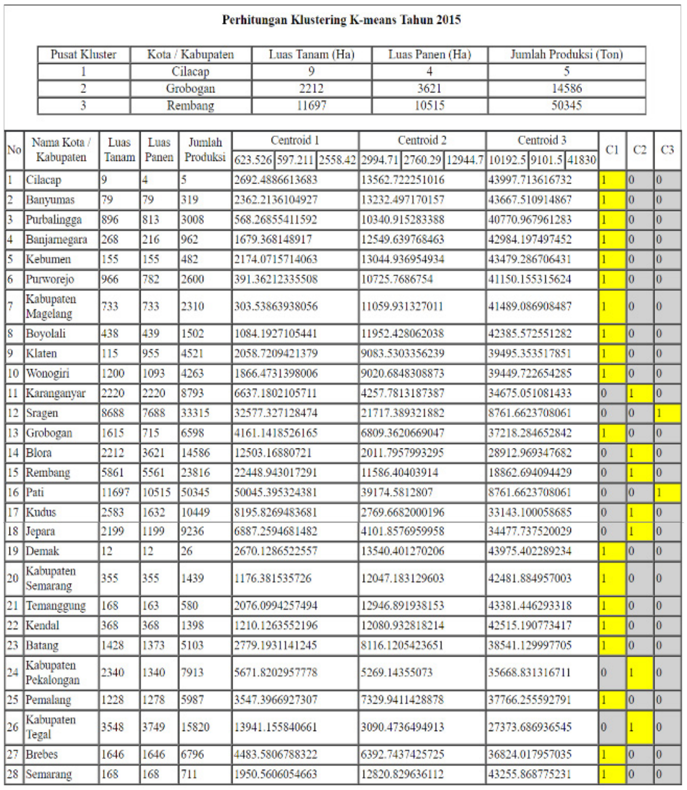
$$
\begin{aligned}
C1 &= \sqrt{(A - c1a)^2 + (B - c1b)^2 + (C - c1c)} \\
&= \sqrt{(79 - 9)^2 + (79 - 4)^2 + (319 - 5)^2} \\
&= \sqrt{4900 + 5625 + 98596} \\
&= 303.334 \\
C2 &= \sqrt{(A - c2a)^2 + (B - c2b)^2 + (C - c2c)} \\
&= \sqrt{(79 - 2212)^2 + (79 - 3621)^2 + (319 - 14586)^2} \\
&= \sqrt{4549689 + 12545764 + 203547289} \\
&= 14854.048 \\
C3 &= \sqrt{(A - c3a)^2 + (B - c3b)^2 + (C - c3c)} \\
&= \sqrt{(79 - 11697)^2 + (79 - 10515)^2 + (319 - 50345)^2} \\
&= \sqrt{134977924 + 108910096 + 2502600676} \\
&= 52406.952
\end{aligned}
$$

After knowing the euclidean distance values, a comparison of the distance of each class is performed to determine the minimum distance to each cluster center.

Lowest distance = min(C1:C2:C3)
= min(303.334:1485.048:52406.952)
= 303.334

After knowing the minimum distance, group the data according to its cluster, that is based on data that has a minimum distance. After that, do a new centroid calculation by finding the average by adding up all the members of each cluster and dividing the number of members. Then do the same calculation to find the distance of data to the new cluster center point. The process will continue to repeat until there is no data transfer between classes.

Table 4 is the final cluster result of sugarcane production data. The results of the K-means calculation in the application are the same as the K-means calculations done manually. The following are the results of the cluster in the application.

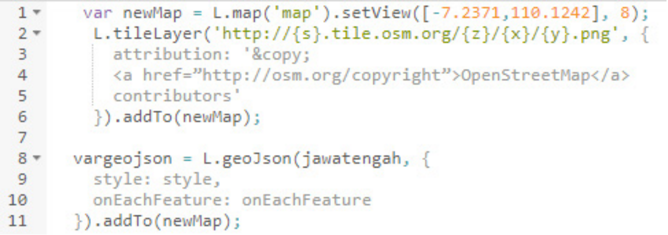


Perhitungan Klustering K-means Tahun 2015

Pusat Kluster	Kota / Kabupaten	Luas Tanam (Ha)	Luas Panen (Ha)	Jumlah Produksi (Ton)
1	Cilacap	9	4	5
2	Grobogan	2212	3621	14586
3	Rembang	11697	10515	50345

No	Nama Kota / Kabupaten	Luas Tanam	Luas Panen	Jumlah Produksi	Centroid 1			Centroid 2			Centroid 3			C1	C2	C3
					623.526	597.211	2558.42	2994.71	2760.29	12944.7	10192.5	9101.5	41830			
1	Cilacap	9	4	5	2692.4886613683			13562.722251016			43997.713616732			1	0	0
2	Banyumas	79	79	319	2362.2136104927			13232.497170157			43667.510914867			1	0	0
3	Purbalingga	896	813	3008	568.26855411592			10340.915283388			40770.967961283			1	0	0
4	Banjarnegara	268	216	962	1679.368148917			12549.639768463			42984.197497452			1	0	0
5	Kebumen	155	155	482	2174.0715714063			13044.936954934			43479.286706431			1	0	0
6	Purworejo	966	782	2600	391.36212335508			10725.7686754			41150.155315624			1	0	0
7	Kabupaten Magelang	733	733	2310	303.53863938056			11059.931327011			41489.086908487			1	0	0
8	Boyolali	438	439	1502	1084.1927105441			11952.428062038			42385.572551282			1	0	0
9	Klaten	115	955	4521	2058.7209421379			9083.5303356239			39495.353517851			1	0	0
10	Wonogiri	1200	1093	4263	1866.4731398006			9020.6848308873			39449.722654285			1	0	0
11	Karanganyar	2220	2220	8793	6637.1802105711			4257.7813187387			34675.051081433			0	1	0
12	Sragen	8688	7688	33315	32577.327128474			21717.389321882			8761.6623708061			0	0	1
13	Grobogan	1615	715	6598	4161.1418526165			6809.3620669047			37218.284652842			1	0	0
14	Blora	2212	3621	14586	12503.16880721			2011.7957993295			28912.969347682			0	1	0
15	Rembang	5861	5561	23816	22448.943017291			11586.40403914			18862.694094429			0	1	0
16	Pati	11697	10515	50345	50045.395324381			39174.5812807			8761.6623708061			0	0	1
17	Kudus	2583	1632	10449	8195.8269483681			2769.6682000196			33143.100058685			0	1	0
18	Jepara	2199	1199	9236	6887.2594681482			4101.8576959958			34477.737520029			0	1	0
19	Demak	12	12	26	2670.1286522557			13540.401270206			43975.402289234			1	0	0
20	Kabupaten Semarang	355	355	1439	1176.381535726			12047.183129603			42481.884957003			1	0	0
21	Temanggung	168	163	580	2076.0994257494			12946.891938153			43381.446293318			1	0	0
22	Kendal	368	368	1398	1210.1263552196			12080.932818214			42515.190773417			1	0	0
23	Batang	1428	1373	5103	2779.1931141245			8116.1205423651			38541.129997705			1	0	0
24	Kabupaten Pekalongan	2340	1340	7913	5671.8202957778			5269.14355073			35668.831316711			0	1	0
25	Pemalang	1228	1278	5987	3547.3966927307			7329.9411428878			37766.255592791			1	0	0
26	Kabupaten Tegal	3548	3749	15820	13941.155840661			3090.4736494913			27373.686936545			0	1	0
27	Brebes	1646	1646	6796	4483.5806788322			6392.7437425725			36824.017957035			1	0	0
28	Semarang	168	168	711	1950.5606054663			12820.829636112			43255.868775231			1	0	0

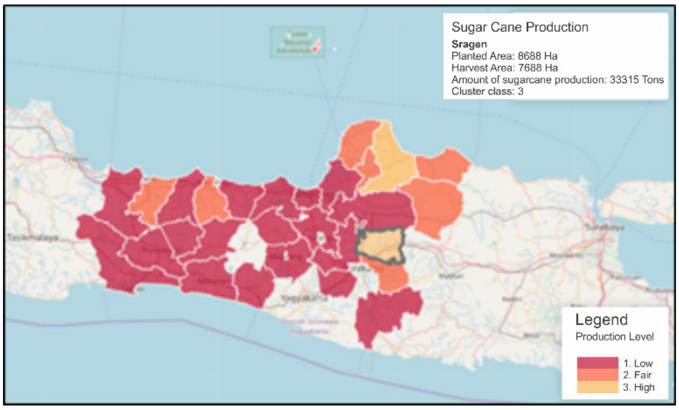
Figure 8. Results of K-means calculations in applications



```
  var newMap = L.map('map').setView([-7.2371,110.1242], 8);
   L.tileLayer('http://{s}.tile.osm.org/{z}/{x}/{y}.png', {
     attribution: '&copy;
     <a href=”http://osm.org/copyright”>OpenStreetMap</a>
     contributors'
   }).addTo(newMap);

 vargeojson = L.geoJson(jawatengah, {
   style: style,
   onEachFeature: onEachFeature
 }).addTo(newMap);
```

Figure 9. Pieces of the Library Leaflet Script



Sugar Cane Production
Sragen
Planted Area: 8688 Ha
Harvest Area: 7688 Ha
Amount of sugarcane production: 33315 Tons
Cluster class: 3

Figure 10. Cane Production Map Interface

Figure 8 is the result of the K-means calculation process in the application using three clusters. Obtained the results of the first cluster with a mean value of 623,526 planting area, 597,211 harvested area, and total production of 2558.42. The second cluster has a mean cluster value of planting area of 2994.71, harvested area of 2760.29, and total production of 12944.7. While the third cluster with the mean cluster size is 10192.5, 9101.5 harvested area, and total production is 41830.

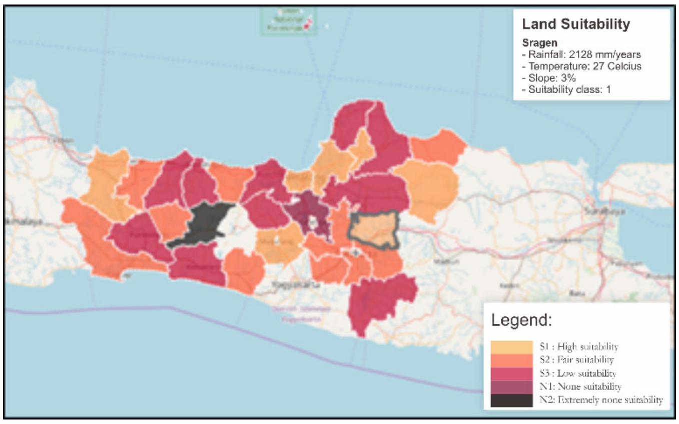


Figure 11. Interface of Land Suitability Map

The clustered data is then processed by the system by changing the data in the database into GeoJSON. The data that has been changed to GeoJSON is then visualized into the OpenStreetMap map by using the Library Leaflet with a script like in Figure 9.

Figure 9 is a piece of code Library Leaflet. In line 1 a OpenStreetMap map is called which is focused on the Province of Central Java. In line 8 a mapping is done based on GeoJSON that has been converted from the database. So we get results like Figure 10.

Figure 10 is a display of the mapping of sugarcane production levels in Central Java Province in 2015, which has a color attribute based on the value of the cluster members in each region. The number of clusters used in the designed application consists of three clusters namely low, sufficient, and high. The color attribute on the map is one of the features that makes it easy for users to distinguish one cluster from another.

The regions of Cilacap, Banyumas, Purbalingga, Banjarnegara, Kebumen, Purworejo, Magelang Regency, Boyolali, Klaten, Wonogiri, Grobogan, Demak, Semarang Regency, Temanggung, Kendal, Batang, Pemalang, Brebes and Semarang fall into the category of low production levels marked by dark red color. The Karanganyar, Blora, Rembang, Kudus, Jepara, Pekalongan and Tegal Regencies are included in the production level, which is marked in orange. While the Region of Sragen, and Pati are areas that are classified as high production levels marked by yellow.

This also applies to the mapping of land suitability of sugarcane, where land suitability data that has been processed by the Matching method is then visualized in the form of a map as illustrated in Figure 11.

Figure 11 is a map of land suitability for sugarcane land in Central Java province which is divided into five classes, namely S1 (High suitability), S2 (Fair suitability), S3 (Low suitability), N1 (, None suitability), and N2 (Extremely none suitability). Where the Districts of Magelang, Sragen, Blora, Pati, Kudus, Demak, Brebes belong to the S1 class, Cilacap, Purbalingga, Purworejo, Boyolali, Klaten, Sukoharjo, Karanganyar, Rembang, Batang, and Tegal Regencies enter the S2 class, Banyumas Region, Kebumen, Wonogiri, Grobogan, Jepara, Temanggung, Kendal, Pekalongan Regency and Pemalang enter S3 class, Semarang Regency enter N1 class, and Banjarnegara enter N2 class.

3) System Feasibility Analysis

The system feasibility analysis was carried out using the End User Computing Stations (EUCS) model with five respondents from the Central Java province's agriculture service staff. The variables used are Content, Accuracy, Format, Ease of Use, and Timeliness [16]. With assessment scores based on Likert Scale namely Very Satisfied (4), Satisfied (3), Dissatisfied (2) and Very Dissatisfied (1). The questions asked are as follows

CONTENT:

- GIS provide information correctly and correctly
- GIS provide information as needed
- GIS provides information that is easy to understand
- GIS provides useful information for users

ACCURACY :

- GIS provide accurate information in accordance with the wishes of the user
- GIS provide information in accordance with user access rights
- GIS feedback results are in accordance with the functions on the website
- GIS presents the results of data processing correctly and in accordance with user requirements

FORMAT :

- GIS display is attractive and interactive
- GIS display does not confuse users
- GIS display feedback results are in accordance with the functions available on the website
- Content layout is eligible

EASE OF USE :

- GIS provides instructions for using the system
- Users can process data on GIS without difficulty
- The navigation in GIS does not confuse users
- SIG can be operated easily by users

TIMELINESS :

- GIS sites can be accessed quickly
- GIS website performs data calculation quickly (calculate cluster and calculate land suitability)
- The GIS website displays maps quickly
- GIS sites can process data quickly (add, change, delete)

Table 5. Results of the EUCS Questionnaire

Respondents	Mean				
	Content	Accuracy	Format	Ease Of Use	Timeliness
Dwi	3	3	3.75	3.25	2.75
Agus	3.77	3.5	4	3.5	3
Kurniawan	3	3	3.5	3	2.75
Fajar	3.25	3	3.75	3.25	2.5
Agung	3	3.25	4	3	3
Average total	3.2	3.15	3.8	3.2	2.8
Results			3.23		

The level of user satisfaction with the Geographic Information System Mapping the Production and Conformity of the Cane is determined through the conversion of the scale of user satisfaction levels based on Table 5. Based on Table 5, the results of the EUCS test earned a score of 3.23 out of a maximum total score of 4, so that it is classified as satisfied. The score shows that the system is feasible and can be accepted by the user.

4. Conclusion

Based on the test results, the system has been able to map the area based on sugarcane production data in Central Java using the K-means algorithm into three groups namely low, sufficient, and high. The regions of Cilacap, Banyumas, Purbalingga, Banjarnegara, Kebumen, Purworejo, Magelang Regency, Boyolali, Klaten, Wonogiri, Grobogan, Demak, Semarang Regency, Temanggung, Kendal, Batang, Pemalang, Brebes and Semarang fall into the category of low production levels. The Karanganyar, Blora, Rembang, Kudus, Jepara, Pekalongan and Tegal Regencies are included in the production level. While the Region of Sragen, and Pati are areas that are classified as high production levels. As well as mapping land suitability data using the Matching method with suitability classes S1, S2, S3, N1, and N2. The EUCS test with a score of 3.23 (Satisfied) of the maximum total score of 4, shows that the system is feasible and can be accepted by the user.

The Department of Agriculture can make the Geographic Information System Mapping the Production and Suitability of Cane Land as a consideration for determining policies in sugarcane management both for sugarcane expansion and making efforts to overcome the inhibiting factors of sugarcane growth so that sugarcane production in areas that are not optimal can increase in the future. So with these efforts, it is expected that sugarcane production from year to year will increase so that it can realize sugar self-sufficiency in order to support food sovereignty in Central Java Province.

Further research, it is necessary to use a combination of other algorithms to further analyze sugarcane production such as using the Backpropagation Neural Network algorithm to predict sugarcane production in subsequent years in each region in Central Java Province.

Determination of land suitability classes for sugarcane can be refined further by adding other assessment parameters such as water availability, root media, nutrient retention, toxicity, etc.

References

[1] Menko Perekonomian, "Keputusan Menteri Koordinator Bidang Perekonomian No. Kep28/M. EKON/05/2010 tentang Tim Koordinasi Stabilisasi Pangan Pokok," 2010. [Online]. Available: http://www.setneg.go.id/. [Accessed: 10-Feb-2019].

[2] R. I. Kurniasari, D. H. Darwanto, and S. Widodo, "Permintaan Gula Kristal Mentah Indonesia," *Ilmu Pertan.*, vol. 18, no. 1, pp. 24–30, 2015.

[3] Statista, "Principal sugar importing countries in 2017/2018 (in million metric tons)," 2018. [Online]. Available: https://www.statista.com/statistics/273438/principal-sugar-importing-countries/. [Accessed: 06-Feb-2019].

[4] D. P. Kementerian Pertanian RI, *Statistik Perkebunan Indonesia 2015-2017 Tebu*, 2015. 2016.

[5] Perda, *Peraturan Daerah Provinsi Jawa Tengah Nomor 1 Tahun 2018 Tentang Peningkatan Produktivitas Tanaman Tebu*. Indonesia, 2018.

[6] S. Surono, "Kebijakan Swasembada Gula di Indonesia," *J. Ekon. Dan Pembang. Indones.*, vol. 7, no. 1, pp. 63–79, 2006.

[7] Herniwati, "Peranan Geographic Information System (GIS) Dalam Perencanaan Pengembangan Pertanian," *Balai Pengkajian Teknologi Pertanian Sulawesi Selatan*, *Buletin*, No. 6, 2012.

[8] F. Rizal, G. Herdiyansyah, "Analisis Potensi Lahan Pertanian Pangan Untuk Mendukung Ketahanan Pangan Kota Bandung," *J. Teknotan*, vol. 10, no. 1, pp. 61–67, 2016.

[9] Adriyendi, "Clustering using K-Means and Fuzzy C-Means on Food Productivity," *Int. J. u- e- Serv. Sci. Technol.*, vol. 9, no. 12, pp. 291–308, 2017.

[10] M. R. Ridlo, S. Defiyanti, and A. Primajaya, "Implementasi Algoritme K-Means Untuk Pemetaan Produktivitas Panen Padi Di Kabupaten Karawang," in *Conference on Information Technology and Electrical Engineering (CITEE)*, 2017, vol. 9, pp. 426–433.

[11] K. D. Hartomo, S. Yulianto, and J. Maruf, "Spatial Model Design of Landslide Vulnerability Early Detection with Exponential Smoothing Method Using Google API," *Int. Conf. Soft Comput. Intell. Syst. Inf. Technol. (ICSIIT),* vol. 1, pp. 102–106, 2017.

[12] J. Tamaela, E. Sediyono, and A. Setiawan, "Cluster Analysis Menggunakan Algoritma Fuzzy C-means dan K-means Untuk Klasterisasi dan Pemetaan Lahan Pertanian di Minahasa Tenggara," *J. Buana Inform.*, vol. 8, no. 3, pp. 151–160, 2017.

[13] A. F. Hidayat, Z. W. Baskara, W. Werdiningsih, and Y. Sulastri, "Evaluasi Kesesuaian Lahan untuk Komoditas Padi dengan Memanfaatkan Aplikasi Sistem Informasi Geografis (SIG) di Kabupaten Lombok Tengah," *J. Ilm. Rekayasa Pertan. dan Biosist.*, vol. 6, no. 1, pp. 69–75, 2018.

[14] Widiatmaka, "Integrasi Informasi Geografis dan Informasi Sumberdaya Lahan Pertanian Mendukung Kedaulatan Pangan Nasional," in *Seminar Nasional Peran Geografi Dalam Mendukung Kedaulatan Pangan*, 2015, pp. 109–117.

[15] Widiatmaka and S. Hardjowigeno, "Evaluasi Lahan dan Perencanaan Tataguna Lahan," September, Gadjah Mada University Press, 2011.

[16] W. J. Doll and G. Torkzadeh, "The Measurement of End-User Computing Satisfaction," *MIS Q.*, vol. 12, no. 2, pp. 259–274, 1988.

[17] D. Wahyunto, Himatullah, E. Suryani, *Petunjuk Teknis Pedoman Penilaian Kesesuaian Lahan untuk Komoditas Pertanian Strategis Tingkat Semi Detail Skala 1 : 50.000*, April, 201. Bogor: Balai Besar Litbang Sumberdaya Lahan Pertanian, Badan Penelitian dan Pengembangan Pertanian, 2016.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Detection of Cyber Malware Attack Based on Network Traffic Features Using Neural Network

Ventje Jeremias Lewi Engel[1*], Evan Joshua[2], Mychael Maoeretz Engel[2]

[1]Computer Engineering Department
Institut Teknologi Harapan Bangsa
Bandung

[2]Informatics Department
Institut Teknologi Harapan Bangsa
Bandung

[*]ventje@ithb.ac.id

Abstract-Various techniques have been developed to detect cyber malware attacks, such as behavior based method which utilizes the analysis of permissions and system calls made by a process. However, this technique cannot handle the types of malware that continue to evolve. Therefore, an analysis of other suspicious activities – namely network traffic or network traffic – need to be conducted. Network traffic acts as a medium for sending information used by malware developers to communicate with malware infecting a victim's device. Malware analyzed in this study is divided into 3 classes, namely adware, general malware, and benign. The malware classification implements 79 features extracted from network traffic flow and an analysis of these features using a Neural Network that matches the characteristics of a time-series feature. The total flow of network traffic used is 442,240 data. The results showed that 15 main features selected based on literature studies resulted in F-measure 0.6404 with hidden neurons 12, learning rate 0.1, and epoch 300. As a comparison, the researchers chose 12 features based on the nature of the malware possessed, with the F-measure score of 0.666 with hidden neurons 12, learning rate 0.05, and epoch 300. This study found the importance of data normalization technique to ensure that no feature was far more dominant than other features. It was concluded that the analysis of network traffic features using Neural Network can be used to detect cyber malware attacks and more features does not imply better detection performance, but real-time malware detection is required for network traffic on IoT devices and smartphones.

Keywords: cyberattacks; malware detection; neural network; network traffic feature

1. Introduction

As the adoption of society towards technology increases, the number of IoT (Internet of Things) devices and smartphones usage has been increasing and widespread. Security threats on IoT devices and smartphones also increase. Various cyberattacks can be committed on IoT devices and smartphones, ranging from taking access rights, destructing the data, thieving important information, and recording personal activities of users when using IoT devices and smartphones [1]. Most of these cyberattacks enter the system through malicious software or malware that are successfully planted on IoT devices and smartphones.

Malware is an application that has a negative purpose, such as corrupting data, stealing important information, disrupting device performance, and taking over the system. This threat continues to increase every year. In 2017, it is found around 3.5 million new malwares only on Android smartphone devices [2]. One of the suspicious activities of malware is the use of network traffic – can be applied as a medium for sending confidential information in the form of PINs, bank account information, personal messages, and passwords to malware makers [3]. Malware can also utilize network traffic as a backdoor for other malwares to enter.

The network traffic on IoT devices and smartphones has the same basis as network traffic in general, which contains packets that have a header and data section [4]. Data is obtained and processed at the application layer, while headers are added at each layer. The size of each data and header varies with the specified limits. The packet contains the data that the sender wants to send from source to destination. The header contains the destination IP address, sender's IP address, source port, destination port, and several other related information. Most network traffic features are time-series.

In general, malware detection system classifies applications into adware, general malware, and benign [5]. Adware is a type of malware that displays advertisements on running software. Adware aims to increase revenue for software developers so that the advertised company pays for the adware. Each type of general malware is confirmed to have a negative purpose, such as damaging or stealing data. Benign is a normal type of application that does not have dangerous purposes; it runs according to what the application developer has written in the documentation section.

There are several efforts in detecting mobile malware that have been carried out using various approaches. Behavior-based approach that uses permissions and system calls as features, produces accuracy that is still relatively low with an average of 60%. Specifically, Simple Logistic 65.29%, Naive Bayes 65.29%, SMO 70.31% and Random Tree 54.79% [6]. Other studies using network traffic features using the Neural Network (NN) method to detect malware on smartphones have successfully detected malware botnets with a precision level of around 88.3% [7]. This result is much higher compared to the Naive Bayes and Logistic Regression methods, each of which has a value of 7% and 32% [7]. In addition, the NN method successfully outperformed the Support Vector Machine (SVM) method in classifying network traffic [8]. NN method is often used as a classification method because of its robust characteristics. It can even be used for quality classification [9]. Detecting malware through network traffic analysis – which is mostly in time-series data – suits with the NN machine learning method.

The weakness of the previous research is that the NN method is carried out on all network traffic features, despite there are several network features that has a more important role than other network traffic features. For example, the network destination port is more important than the length of the header contents. Second, the use all network traffic features results in the increase of the internal errors that carried in the data. Third, features with large values automatically weigh higher, for example the port values commonly used are much smaller, when being compared to the value of data flow across the network [5].

The difference between this study and previous research is the network traffic dataset, the combination of features, and the iteration of the NN configuration applied. The dataset applied in this research obtained from the Canadian Institute for Cybersecurity, University of New Brunswick [10] combined with sample data collected at the Harapan Bangsa Institute of Technology Computer Laboratory (ITHB). A total of 1900 android applications with a percentage of 20% malware and 80% benign. Malware is divided into two types including adware and general malware. The combination of features is carried out based on literature studies to obtain the intersection of network traffic features that are frequently used in malware detection system. The iteration of the NN configuration is conducted by programming that concern to learning rate, epoch, and parameter evaluation. The purpose of this study is to obtain the configuration of the NN model to detect cyber-type malware attacks and to investigate the combination of network traffic features that can result in high precision, recall, and F-measure in the detection of mobile malware using NN.

2. Methods

a. Research Flowchart

The research steps are arranged in the form of a flowchart, which begins with preprocessing. The preprocessing conducted is the normalization of features that will be used by dividing the features' values by the maximum value of each feature. Hence, this process will minimize features, so that a feature does not dominate other features.

Next, the learning stage applies the Neural Network method with backpropagation algorithm and the testing phase uses feed-forward method. In the initial phase, the weights will be randomly assigned in accordance with the previous provisions and they are stored in the file weight. Learning outcomes will give new weight values. The test will use the weight in the previous learning file. The test output is divided into 3, namely benign, adware, or general malware.

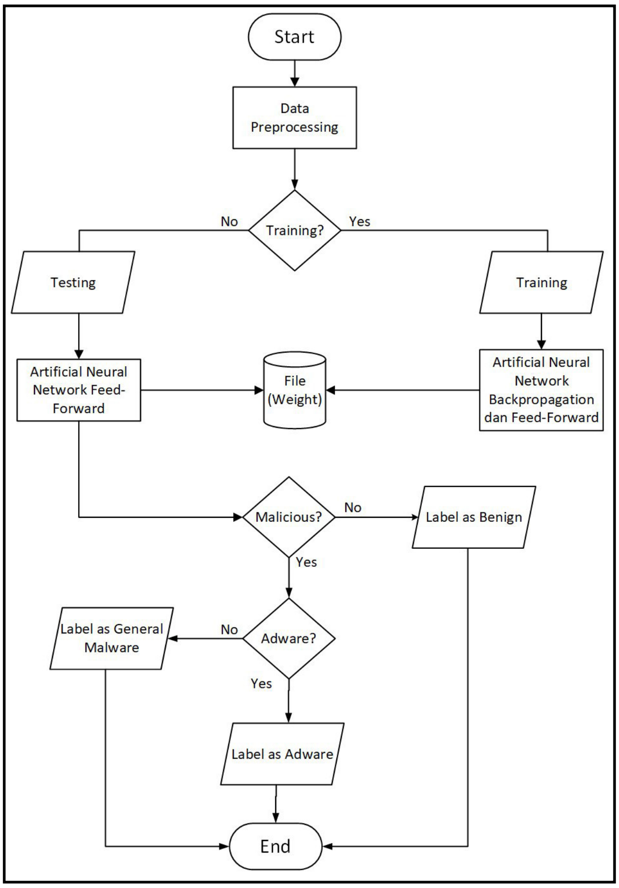


Figure 1. Research Flowchart

b. Neural Network Architecture

Neural Network (NN) or often also called Artificial Neural Network is one of machine learning techniques.

Neural networks are included in supervised learning, with the resulting model in the form of weight [11]. The weights are used at the test stage and the output is mapped to the activation function to determine which label the output refers to.

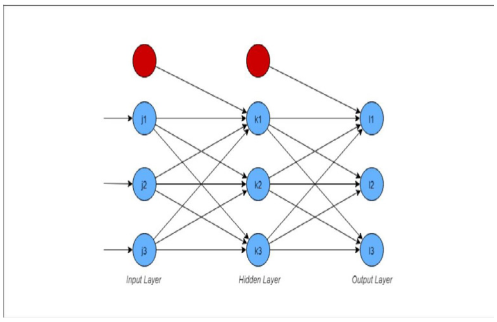


Figure 2. Neural Network Component

As shown in Figure 2, There are 3 main layers in the Neural Network, namely the input layer, hidden layer and output layer. It is also drawn several circles of various colors according to their role. The blue circles are called nodes or neurons, while the red circles are bias – benefit to increase the flexibility of the model.

The input layer acts as the layer that receives initial input. The input obtained is processed to produce output on the hidden layer. The hidden layer is situated between the input and output layers and is useful for supporting neural networks learning complex features. The hidden layer itself can contain several layers. Each layer in the hidden layer may have different number of neurons. The hidden layer will produce output which then subjects to an activation function, to be mapped to the class in the output layer.

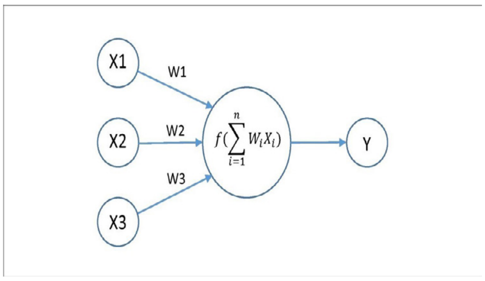


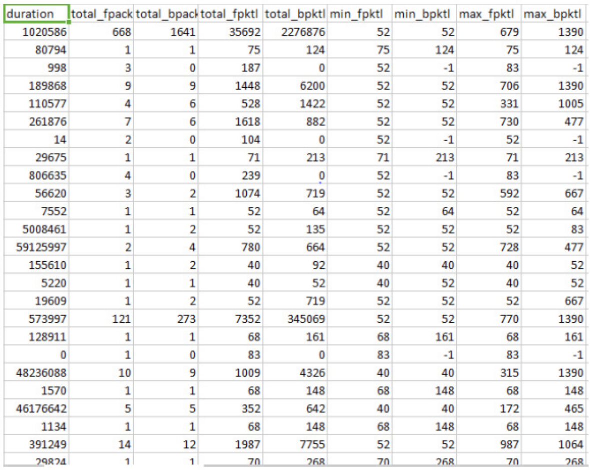
Figure 3. Calculation Process of Output at Layer

Figure 3 shows that each neuron has a weight according to the number of connections with other neurons. Output calculation is influenced by the weight and input values, which the results will then be processed with an activation function. According to Stevanovic [7], this mechanism makes Neural Network able to read and analyze simultaneously many features of network traffic for detection of malware with a high degree of precision.

In this study, three layers will be used, including the input layer, hidden layer and output layer. The input layer has a number of neurons according to the number of features used. In the hidden layer, only one layer will be used with the number of neurons tested, such as 4, 5, 6, and 12. The output layer will produce output in the form of 3 classes, namely benign, adware, and general malware. The test will apply several combinations of Neural Network parameters including learning rate, hidden neurons and the number of epochs. The learning rates tested are 0.1, 0.05, and 0.01 with the number of epoch 100, 200, and 300.

c. Dataset

Dataset used is a pcap (packet capture) file that contains network traffic packets with a total of 79 features. The pcap file was earned from a total of 1900 android applications with a percentage of 20% malware and 80% benign. The malware dataset is divided into three groups, including 250 adware applications, 150 general malware applications, and 1,500 benign applications. In the training data, there are 2,312 network traffic flows from general malware, 149,871 for adware, and 201,609 for benign, while in data testing there are 1,626 general flow malware, 24,271 flow adware, and 62,551 flow benign. The total flow of network traffic used is 442,240 data. By using the CICFlowMeter application, the pcap file is converted to CSV file, so that one flow means one line of data.



duration	total_fpack	total_bpack	total_fpktl	total_bpktl	min_fpktl	min_bpktl	max_fpktl	max_bpktl
1020586	668	1641	35692	2276876	52	52	679	1390
80794	1	1	75	124	75	124	75	124
998	3	0	187	0	52	-1	83	-1
189868	9	9	1448	6200	52	52	706	1390
110577	4	6	528	1422	52	52	331	1005
261876	7	6	1618	882	52	52	730	477
14	2	0	104	0	52	-1	52	-1
29675	1	1	71	213	71	213	71	213
806635	4	0	239	0	52	-1	83	-1
56620	3	2	1074	719	52	52	592	667
7552	1	1	52	64	52	64	52	64
5008461	1	2	52	135	52	52	52	83
59125997	2	4	780	664	52	52	728	477
155610	1	2	40	92	40	40	40	52
5220	1	1	40	52	40	52	40	52
19609	1	2	52	719	52	52	52	667
573997	121	273	7352	345069	52	52	770	1390
128911	1	1	68	161	68	161	68	161
0	1	0	83	0	83	-1	83	-1
48236088	10	9	1009	4326	40	40	315	1390
1570	1	1	68	148	68	148	68	148
46176642	5	5	352	642	40	40	172	465
1134	1	1	68	148	68	148	68	148
391249	14	12	1987	7755	52	52	987	1064
29824	1	1	70	268	70	268	70	268

Figure 4. Screenshot of CSV Datasheet File Contents

d. Feature Combination Analysis

The combination of features that will be used in the Neural Network is chosen based on the analysis, obtained from the literature study. The results of the literature study can be observed in Table 1.

Table 1. Key Features Network Traffic Literature Study Results

Numb	Feature Name	Reference
1.	Source port	[12][13]
2.	Destination port	[12][13]
3.	L3 / L4 Protocol identifier	[12][13]
4.	Total number of packets	[7][12][13][14][15]
5.	Total number of bytes	[7][12][14]
6.	Mean of number of bytes per packet	[7][12][15]
7.	Standard deviation of number of bytes per packet	[7][12]
8.	Number of packets per second	[7][12]
9.	Number of bytes per second	[7][12][13]
10.	Flow duration	[7][12][13][15]
11.	Mean of inter-arrival time (IAT)	[7][12]
12.	Standard deviation of IAT	[7][12]
13.	Ratio of number of packets OUT/IN	[12][13][14]
14.	Ratio of number of bytes OUT/IN	[7][12]
15.	Ratio of IAT OUT/IN	[7][12]

Table 2. Researcher's Selected Features

Numb	Feature Name	Category
1.	Forward packets	Packet based
2.	Total forward packets	Packet based
3.	Forward packet length max	Byte based
4.	Active mean	Time based
5.	Backward packets / second	Packet based
6.	Forward IAT standard deviation	Time based
7.	Max packet length	Packet based
8.	Total backward packets	Packet based
9.	Total length of backward packets	Byte based
10.	Backward IAT standard deviation	Time based
11.	FIN flag count	Flow based
12.	Packet length variance	Byte based

As a comparison, researchers chose 12 features according to researchers' understanding regarding malware. Adware variant has characteristics that interrupts the application to display the advertisements which is actually malicious code. This causes a lot of flow in the forward and backward packages. The twelve features selected by the researchers did not overlap with the features of the literature study results, and are informed in Table 2.

e. Objective and Evaluation

From the analysis of dataset, it was found that class imbalance occurred in malware label data, which was only 20% compared to benign (80%) [10] resulting in an evaluation computed with ordinary accuracy metric to be insufficient. Therefore, in this case, F-measure was used as a metric instead of accuracy. The F-measure is used to help in drawing conclusions about which Neural Network parameters are best implemented. The advantage of the F-measure is able to consider precision and recall into a single unit that is interconnected with one another. Table 2 shows the confusion matrix used to obtain the values of True Positive, False Positive, True Negative and False Negative.

Table 3. Confusion Matrix

Prediction Value	True Value	
	TRUE	***FALSE***
TRUE	True Positive (TP)	False Positive (FP)
FALSE	False Negative (FN)	True Negative (TN)

Table 4 Neural Network Results Using the Literature Study Features vs. Researcher Features vs. Combined Features

Numb.	Combination of Features	Number of Features	Hidden Neuron	Learning Rate	Epoch	Training time	Testing time	Precision	Recall	F-measure
1	**Literature Study Features**	15	12	0.1	300	28m 50s	6s	47.58%	97.88%	0.6404
2	**Researcher Features**	12	12	0.05	300	27m 43s	4s	55.89%	82.39%	0.6660
3	**Combined Features**	27	12	0.1	300	30m 3s	4s	47,40%	98.26%	0.6395

Table 5. Comparison of Feature Combinations in Hidden Neurons of 12

Hidden Neuron = 12	Learning Rate	Epoch	Precision	Recall	F-measure
Literature Study Features	0.1	100	47.34%	98.07%	0.6386
		200	47.44%	98.14%	0.6396
		300	**47.58%**	**97.88%**	**0.6404**
	0.05	100	63.83%	49.85%	0.5598
		200	64.46%	49.78%	0.5618
		300	65.30%	49.75%	0.5648
	0.01	100	70.69%	42.65%	0.532
		200	71.30%	42.57%	0.5331
		300	71.68%	42.38%	0.5326
Researcher Features	0.1	100	0.00%	0.00%	0
		200	0.00%	0.00%	0
		300	0.00%	0.00%	0
	0.05	100	53.93%	85.13%	0.6603
		200	54.94%	83.45%	0.6626
		300	**55.89%**	**82.39%**	**0.6660**
	0.01	100	0.00%	0.00%	0
		200	0.00%	0.00%	0
		300	0.00%	0.00%	0

Formulas (1), (2), and (3) are employed for determining the value of precision, recall, and F-measure, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

3. Results and Discussion

The implementation and testing environment is conducted in cloud computing since the CSV data that must be processed is quite large, both for training and testing. Weight configurations on the Neural Network are randomly generated. Then, the first training process is carried out – the weights are updated. The training process is conducted continuously until the specified epoch is finished. After that, testing is carried out with feed forward. Table 4 shows a comparison of Neural Network results with features obtained from literature studies and features earned from researchers' knowledge.

Complete test results for each combination of features are given in the supplement of this article. The highest F-measure was achieved for hidden neurons number 12. These results are consistent with Stevanovich's research [7, 12] which states that the more hidden neurons used, the Neural Network performance tends to be better until it finds a saturation point. This is different for learning rate. Comparison of learning rate and epoch for each combination of features in hidden neurons totaling 12 can be seen in Table 5.

A higher learning rate does not guarantee that the F-measure results will also be better. In the combination of researchers' features, the best results are achieved when the learning rate is 0.05 only. The combination of literature study features does achieve the best results with maximum configuration of Neural Network parameters (learning rate 0.1 and epoch 300). Technically, learning rate is the magnitude of change given to the weight which is changed according to the error value. Whereas, the epoch indicates the number of iterations performed by the computer. Learning rate that is overly high or low might result in new weights at further deviation than the expected weights. From Table 5, it is shown that in the combination of researchers' features, there are several learning processes that produce a value of 0 for precision, recall, and F-measure. This is assumed that the model produced with these parameters experienced underfitting when the learning rate is 0.01 and 0.1.

The F-measure score of the combination of 12 researchers' features is greater than the combination of 15 features of literature studies (0.6660 > 0.6404). This shows that using more features does not necessarily improve the accuracy of malware detection on the Neural Network. It is obvious that the two sets of feature combinations do not intersect, but have slightly different F-measure values. It means that there are still combinations of features that are likely to produce F-measure values better than both. For this reason, researchers merged the two combinations of features and conducted testing and training once again. The results of the merged combination of 27 features earned the highest score of F-measure on the number of hidden neurons 12, learning rate 0.1, and epoch 300; the resulted F-measure is 0.6395 (see Table 4). This score is lower than the results of a combination of literature study features. These results once again show that more features do not necessarily improve detection accuracy. This is because the more features used might result in more internal errors which were involved in the learning process. Each feature has internal errors, such as errors due to measurement or errors due to rounding values [13]. Another factor is that each feature has its own contribution in malware detection and there is a possibility that features that are combined together have the effect of eliminating each other, so that the detection accuracy might decrease [15].

4. Conclusion

Detection of cyber malware attacks based on network traffic features using Neural Network results in different F-measure values for different combinations of features. A combination of features based on literature studies (15 features) produces an F-measure of 0.6404, a combination of researchers' analysis features (12 features) produces an F-measure of 0.6660, and a combination of the two combined features (27 features) produces an F-measure of 0.6395. The conclusion is that the number of features does not mean that the accuracy of malware detection will increase. Instead, an improper combination of features can reduce detection accuracy.

This research uses Dataset with 442,240 data which is a combination of existing Dataset and the results of laboratory experiments, for the learning process. It is recommended that the existing Neural Network model can be applied to detect malware in real time on IoT devices and smartphones. Additionally, further research is also needed on the analysis of the combination of network traffic features to produce even better accuracy.

Acknowledgement

This research was funded through a grant from the Ministry of Research, Technology and Higher Education *(Kementerian Riset, Teknologi, dan Pendidikan Tinggi.)*

References

[1] Kaspersky, "Mobile Malware Threatens Smartphones & Tablets," *Kaspersky Lab ZA*, 2015. [Online]. Available: https://www.kaspersky.co.za/resource-center/threats/mobile-malware. [Accessed: 18-Jul-2018].

[2] C. Lueg, "8,400 new Android malware samples every day," *G Data Security Blog*, 2017. [Online]. Available: https://www.gdatasoftware.com/blog/2017/04/29712-8-400-new-android-malware-samples-every-day. [Accessed: 18-Jul-2018].

[3] Y. Zhou and X. Jiang, "Dissecting Android malware: Characterization and Evolution," in *Proceedings - IEEE Symposium on Security and Privacy*, 2012, no. 4, pp. 95–109.

[4] B. A. Forouzan, *TCP/IP Protocol Suite*, 4th ed. New York: McGraw-Hill Companies, Inc., 2010.

[5] A. H. Lashkari, A. F. A. Kadir, H. Gonzalez, K. F. Mbah, and A. A. Ghorbani, "Towards a Network-Based Framework for Android Malware Detection and Characterization," in *Proceeding of the 15th international conference on privacy, security and trust*, 2017.

[6] P. Kaushik and A. Jain, "Malware Detection Techniques in Android," *Int. J. Comput. Appl.*, vol. 122, no. 17, pp. 22–26, 2015.

[7] M. Stevanovic and J. M. Pedersen, "An analysis of network traffic classification for botnet detection," in *2015 International Conference on*

Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), 2015, pp. 1–8.

[8] J. Zhang, Y. Xiang, and Y. Wang, “Network Traffic Classification Using Correlation Information,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 1, pp. 104–117, 2013.

[9] F. Wibowo and A. Harjoko, “Klasifikasi Mutu Pepaya Berdasarkan Ciri Tekstur GLCM Menggunakan Jaringan Saraf Tiruan,” *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 3, no. 2, pp. 100–104, 2018.

[10] Canadian Institute for Cybersecurity, “Android Adware and General Malware Datasheet ,” *University of New Brunswick*, 2017. [Online]. Available: https://www.unb.ca/cic/Datasheet s/ android-adware.html. [Accessed: 18-Nov-2018].

[11] T. Rashid, *Make Your Own Neural Network: A Gentle Journey Through the Mathematics of Neural Networks*. CreateSpace Independent Publishing Platform, 2016.

[12] M. Stevanovic and J. M. Pedersen, “An efficient flow-based botnet detection using supervised machine learning,” in *2014 International Conference on Computing, Networking and Communications (ICNC)*, 2014, pp. 797–801.

[13] H. Lim, Y. Yamaguchi, H. Shimada, and H. Takakura, “Malware Classification Method Based on Sequence of Traffic Flow,” in *2015 International Conference on Information Systems Security and Privacy (ICISSP)*, 2015, pp. 394–401.

[14] D. Jiang and K. Omote, “An approach to detect remote access trojan in the early stage of communication,” in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, 2015, pp. 706–713.

[15] Z. B. Celik, R. J. Walls, P. Mcdaniel, and A. Swami, “Malware Traffic Detection using Tamper Resistant Features,” in *MILCOM 2015-2015 IEEE Military Communications Conference*, 2015, pp. 330–335.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Market Basket Analysis to Identify Stock Handling Patterns and Item Arrangement Patterns Using Apriori Algorithms

Sunardi*, Abdul Fadlil, Tresna Yudha Prawira
Masters Informatics Technology
Universitas Ahmad Dahlan
Yogyakarta
*sunardi@mti.uad.ac.id

Abstract-The process of managing the pattern of handling stock of goods and the pattern of arranging goods on store shelves requires an identification process by utilizing data from sales transaction results. Market basket analysis of sales transaction data using Apriori Algorithm stages produces an information in the form of association rules with a minimum support value of 50% and a minimum confidence of 60%. It can be a reference in the arrangement of items on store shelves by referring to a combination of items that are often bought by consumers simultaneously. In addition, the stock inventory pattern can take advantage of the results of determining the high frequency value in the combination pattern 1 - itemset C1 with a minimum support value of 50% which is compared with the initial inventory.

Keywords: Apriori Algorithm, arrangement of goods, stock

1. Introduction

In the management of a minimarket, it is necessary to have a system management in the process of handling the patterns of stock handling and patterns of arrangement of goods, with the aim of structuring the goods to make it easier for customers to shop and patterns of handling stock of goods to provide the availability of goods needed by customers. Of the two, if not handled by an analysis of sales transaction data, it can be a problem for the development of minimarkets. Many business sellers of goods that assume that the sales transaction data report is only to be a report on several things, such as how many items are sold, how many items are still available, and how much profit is obtained. Sales transaction data can be used to help decisions in predicting the layout of goods so that consumers easily find the items sought and determine the prediction of the amount of stock in the future. These problems can be solved by market basket analysis using the stages of Apriori Algorithm, namely by identifying the value of support and confidence of goods sold at the minimarket. It can be a preference in the pattern of arrangement of goods based on customer habits in buying goods simultaneously and can also be a prediction of the stock of goods in the future [1].

Market basket analysis is the process of analyzing customer buying habits on each sales transaction data, by identifying associations between different items of the consumer shopping basket [2].

This research focuses on the event identifying the process of managing stock patterns of handling goods and structuring patterns of goods using Apriori Algorithm. This algorithm is used to find rules or measure the relationship between two or more items. Associative rules are expressed in terms of if they are antecedents, so they are consistent with the amount of support and confidence associated with the rules [3].

In a study entitled “Implementation of Apriori Algorithms with market basket analysis for product layout settings” explained that the association rules formed from the results of the discussion are used to regulate product placement in stores. Products that have high associations with other products will be placed close together, so as to facilitate consumers in buying products and store management in managing stock [4]. Maharani, et al. Also conduct research under the heading “Implementation of data mining for minimarket layout by applying association rule”. The research aims to apply the association rule in the preparation of product layouts. From the rules generated, it can help companies in the preparation of product layouts [5]. Meanwhile, according to research conducted by Adyawangkara, et al with the title "Analysis of the rules of association between items in purchase transactions using data mining with Apriori Algorithm (case study: Gunungan minimarket, Central Java)", in that study aims to find the rules of association in the purchase of items in minimarkets to solve the problem of procurement of goods stock, determining promotional strategies, and arranging goods in minimarkets [6]. Research on the rules of association between items in identifying products that are sold and the relationship between products based on the conditions of the sales transaction data. [7] [8] [9] [10] [11] [12] [13] [14] [15].

2. Research Methods

This study conducted an experimental process with the aim of understanding the steps in the market basket analysis process using Apriori Algorithms in identifying patterns of handling stock items and patterns of arrangement of goods on store shelves. The stages in this study can be shown by Figure 1:

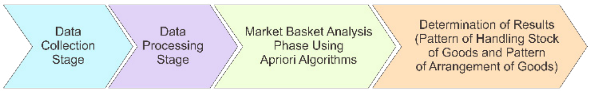

Data Collection Stage
Data Processing Stage
Market Basket Analysis Phase Using Apriori Algorithms
Determination of Results (Pattern of Handling Stock of Goods and Pattern of Arrangement of Goods)

Figure 1. Research Process Stage

a. Data Collection Stage
The data used in this study were taken from the sales transaction database that is available at the Surya Mart minimarket. The mini market is a campus cooperative located at STMIK Muhammadiyah Paguyangan Brebes which is located on Jl. Pangeran Diponegoro, Kec. Paguyangan, Kab. Brebes. Postal Code 52276.

b. Data Processing Stage
At this stage the database scan process is carried out from the sales transaction table in order to calculate the number of purchases that occur on each item of goods. After that, the process is implemented into tabular data in the form of binary data with the statement 0 if the item was not purchased and 1 if the item was purchased on each item sales transaction identity.

c. Market Basket Analysis Phase Using Apriori Algorithms.
From the tabular data generated at the data processing stage, a market basket analysis process is performed using Apriori Algorithm to find high frequency values of sales transaction data by identifying support values and determining a minimum support value of 50%. The next step is to establish a 2-item A association containing B by finding a confidence value, and determining a minimum confidence value of 60%. Figure 2 shows the completion stage of the market basket analysis process using apriori algorithm:

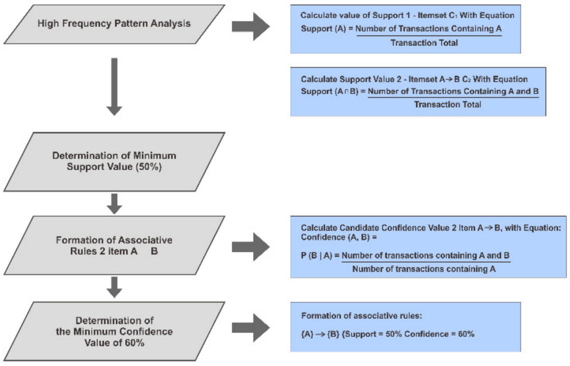

High Frequency Pattern Analysis
Calculate value of Support 1 - Itemset C₁ With Equation
Support (A) = Number of Transactions Containing A / Transaction Total
Calculate Support Value 2 - Itemset A→B C₂ With Equation
Support (A∩B) = Number of Transactions Containing A and B / Transaction Total
Determination of Minimum Support Value (50%)
Formation of Associative Rules 2 item A B
Calculate Candidate Confidence Value 2 Item A→B, with Equation:
Confidence (A, B) =
P (B | A) = Number of transactions containing A and B / Number of transactions containing A
Determination of the Minimum Confidence Value of 60%
Formation of associative rules:
{A} → {B} {Support = 50% Confidence = 60%

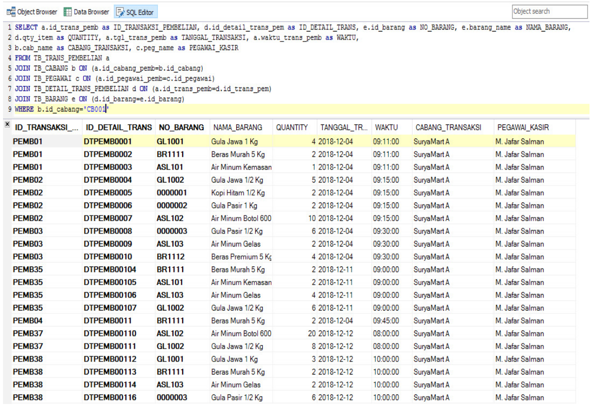
Figure 2. Market Basket Analysis Phase Using Apriori Algorithms.

d. Results Determination Stage
The final stage of this research is the process of determining the pattern of handling stock of goods and the arrangement of goods on store shelves from the results of the market basket analysis process using Apriori Algorithm in sales transaction data.

3. Results snd Discussion

a. Data collection stage

At this stage the data collection process is carried out from the database, in the form of sales data at the solar mart minimarket during the 1 (one) month transaction period as shown in Figure 3:



```
SELECT a.id_trans_pemb as ID_TRANSAKSI_PEMBELIAN, d.id_detail_trans_pem as ID_DETAIL_TRANS, e.id_barang as NO_BARANG, e.barang_name as NAMA_BARANG,
d.qty_item as QUANTITY, a.tgl_trans_pemb as TANGGAL_TRANSAKSI, a.waktu_trans_pemb as WAKTU,
b.cab_name as CABANG_TRANSAKSI, c.peg_name as PEGAWAI_KASIR
FROM TB_TRANS_PEMBELIAN a
JOIN TB_CABANG b ON (a.id_cabang_pemb=b.id_cabang)
JOIN TB_PEGAWAI c ON (a.id_pegawai_pemb=c.id_pegawai)
JOIN TB_DETAIL_TRANS_PEMBELIAN d ON (a.id_trans_pemb=d.id_trans_pem)
JOIN TB_BARANG e ON (d.id_barang=e.id_barang)
WHERE b.id_cabang='CB001'
```

ID_TRANSAKSI_...	ID_DETAIL_TRANS	NO_BARANG	NAMA_BARANG	QUANTITY	TANGGAL_TR...	WAKTU	CABANG_TRANSAKSI	PEGAWAI_KASIR
PEMB01	DTPEMB0001	GL1001	Gula Jawa 1 Kg	4	2018-12-04	09:11:00	SuryaMart A	M. Jafar Salman
PEMB01	DTPEMB0002	BR1111	Beras Murah 5 Kg	2	2018-12-04	09:11:00	SuryaMart A	M. Jafar Salman
PEMB01	DTPEMB0003	ASL101	Air Minum Kemasan	1	2018-12-04	09:11:00	SuryaMart A	M. Jafar Salman
PEMB02	DTPEMB0004	GL1002	Gula Jawa 1/2 Kg	5	2018-12-04	09:15:00	SuryaMart A	M. Jafar Salman
PEMB02	DTPEMB0005	0000001	Kopi Hitam 1/2 Kg	2	2018-12-04	09:15:00	SuryaMart A	M. Jafar Salman
PEMB02	DTPEMB0006	0000002	Gula Pasir 1 Kg	2	2018-12-04	09:15:00	SuryaMart A	M. Jafar Salman
PEMB02	DTPEMB0007	ASL102	Air Minum Botol 600	10	2018-12-04	09:15:00	SuryaMart A	M. Jafar Salman
PEMB03	DTPEMB0008	0000003	Gula Pasir 1/2 Kg	6	2018-12-04	09:30:00	SuryaMart A	M. Jafar Salman
PEMB03	DTPEMB0009	ASL103	Air Minum Gelas	2	2018-12-04	09:30:00	SuryaMart A	M. Jafar Salman
PEMB03	DTPEMB0010	BR1112	Beras Premium 5 Kç	4	2018-12-04	09:30:00	SuryaMart A	M. Jafar Salman
PEMB35	DTPEMB00104	BR1111	Beras Murah 5 Kg	2	2018-12-11	09:00:00	SuryaMart A	M. Jafar Salman
PEMB35	DTPEMB00105	ASL101	Air Minum Kemasan	2	2018-12-11	09:00:00	SuryaMart A	M. Jafar Salman
PEMB35	DTPEMB00106	ASL103	Air Minum Gelas	4	2018-12-11	09:00:00	SuryaMart A	M. Jafar Salman
PEMB35	DTPEMB00107	GL1002	Gula Jawa 1/2 Kg	6	2018-12-11	09:00:00	SuryaMart A	M. Jafar Salman
PEMB04	DTPEMB0011	BR1111	Beras Murah 5 Kg	2	2018-12-04	09:45:00	SuryaMart A	M. Jafar Salman
PEMB37	DTPEMB00110	ASL102	Air Minum Botol 600	20	2018-12-12	08:00:00	SuryaMart A	M. Jafar Salman
PEMB37	DTPEMB00111	GL1002	Gula Jawa 1/2 Kg	8	2018-12-12	08:00:00	SuryaMart A	M. Jafar Salman
PEMB38	DTPEMB00112	GL1001	Gula Jawa 1 Kg	3	2018-12-12	10:00:00	SuryaMart A	M. Jafar Salman
PEMB38	DTPEMB00113	BR1111	Beras Murah 5 Kg	2	2018-12-12	10:00:00	SuryaMart A	M. Jafar Salman
PEMB38	DTPEMB00114	ASL103	Air Minum Gelas	2	2018-12-12	10:00:00	SuryaMart A	M. Jafar Salman
PEMB38	DTPEMB00116	0000003	Gula Pasir 1/2 Kg	6	2018-12-12	10:00:00	SuryaMart A	M. Jafar Salman

Figure 3. Data from the simulation of sales transactions.

After that, the data processing stage is carried out by scanning the database and changing it into tabular data.

b. Data Processing Stage

Data processing stage is done by scanning the database of sales transaction data, namely by identifying the number of items purchased at each transaction id, information Y and N on each item explains if "Y" means the item was purchased and if "N" means the item is not bought. Data is presented in tabular form as in table 1:

Table 1. Number of items in each transaction id.

Transaction ID	Black Coffee 1/2 Kg	Granulated Sugar 1 Kg	Granulated Sugar 1/2 Kg	Gallon Bottled Water	600 ml Bottled Water	Glass drinking water	Cheap rice 5 kg	Premium Rice 5 Kg	Javanese Sugar 1 Kg	Java Sugar 1/2 Kg
PEMB01	Y	Y	Y	Y	N	N	Y	N	Y	N
PEMB02	N	Y	N	N	Y	N	N	Y	Y	Y
PEMB03	N	Y	N	N	N	Y	N	Y	Y	N
PEMB04	Y	Y	Y	N	N	N	Y	Y	N	N
PEMB10	Y	N	N	Y	N	N	N	N	Y	Y
PEMB11	Y	N	Y	N	N	N	Y	Y	Y	Y
PEMB15	Y	N	Y	Y	N	N	Y	Y	Y	Y
PEMB16	N	Y	Y	Y	Y	Y	Y	N	N	N
PEMB19	N	Y	N	Y	Y	N	N	N	N	N
PEMB20	Y	Y	Y	N	N	N	Y	Y	Y	N
PEMB23	N	Y	Y	Y	N	N	Y	N	Y	N
PEMB24	Y	N	N	Y	Y	N	Y	Y	N	Y
PEMB27	N	Y	Y	N	N	N	Y	N	Y	Y
PEMB28	Y	N	N	N	N	N	Y	Y	Y	N
PEMB31	N	N	N	Y	N	N	Y	Y	Y	N
PEMB32	Y	Y	Y	Y	N	N	Y	Y	N	Y
PEMB35	N	N	Y	Y	N	Y	Y	N	Y	Y
PEMB37	Y	Y	Y	Y	Y	N	Y	N	N	Y
PEMB38	N	N	Y	N	N	Y	Y	N	Y	N
PEMB41	Y	Y	Y	Y	N	Y	N	N	Y	N
PEMB43	Y	Y	Y	Y	Y	N	N	N	N	N
PEMB44	Y	Y	N	N	N	N	Y	Y	N	Y
PEMB47	Y	Y	N	N	N	Y	N	Y	N	N
PEMB49	N	N	N	N	N	N	Y	Y	N	N
PEMB50	N	N	Y	Y	N	N	Y	N	Y	N
PEMB53	Y	Y	N	N	Y	N	Y	N	Y	N
PEMB55	N	N	N	Y	N	N	N	Y	N	Y
PEMB56	Y	Y	Y	N	N	Y	N	N	N	N
PEMB59	N	N	N	N	N	N	Y	Y	N	N
PEMB61	Y	Y	N	N	N	Y	N	Y	N	N
PEMB62	N	N	Y	N	Y	N	Y	N	N	N
PEMB65	N	N	N	Y	N	N	N	Y	Y	N
PEMB67	N	Y	N	N	N	N	Y	N	N	N
PEMB68	N	N	Y	Y	N	N	N	N	Y	N
PEMB71	N	Y	N	N	N	Y	Y	N	N	N
PEMB73	N	N	N	Y	N	N	Y	Y	Y	N
PEMB74	Y	Y	N	N	N	N	N	N	Y	N
PEMB77	N	N	Y	Y	N	N	N	Y	Y	N
PEMB79	Y	N	Y	Y	N	N	Y	N	N	Y
PEMB80	N	Y	N	N	N	Y	N	N	N	N
PEMB81	Y	Y	Y	Y	N	N	N	Y	Y	N
PEMB82	N	N	N	N	N	Y	Y	Y	N	N
PEMB88	Y	N	Y	Y	N	N	Y	N	Y	N
PEMB89	N	N	N	Y	N	Y	N	Y	N	N
PEMB91	Y	Y	Y	Y	N	N	N	Y	N	N
PEMB92	Y	Y	Y	Y	Y	N	N	Y	Y	N
PEMB95	N	N	Y	Y	N	N	N	N	Y	N
PEMB97	Y	Y	Y	Y	N	N	N	Y	Y	N
PEMB98	Y	Y	Y	Y	N	N	N	Y	Y	N
PEMB99	Y	N	Y	N	N	N	N	Y	N	N

From the data in Table 1 which contains the items column with information Y and N, it is changed in binary form to Y = 1 and N = 0 with the aim that it is easy to count each item on each transaction, as shown in Table 2:

Table 2. The number of transaction items is changed in binary form.

Transaction ID	Black Coffee 1/2 Kg	Granulated Sugar 1 Kg	Granulated Sugar 1/2 Kg	Gallon Bottled Water	600 ml Bottled Water	Glass drinking water	Cheap rice 5 kg	Premium Rice 5 Kg	Javanese Sugar 1 Kg	Java Sugar 1/2 Kg
PEMB01	1	1	1	1	0	0	1	0	1	0
PEMB02	0	1	0	0	1	0	0	1	1	1
PEMB03	0	1	0	0	0	1	0	1	1	0
PEMB04	1	1	1	0	0	0	1	1	0	0
PEMB10	1	0	0	1	0	0	0	0	1	1
PEMB11	1	0	1	0	0	0	1	1	1	1
PEMB15	1	0	1	1	0	0	1	1	1	1
PEMB16	0	1	1	1	1	1	1	0	0	0
PEMB19	0	1	0	1	1	0	0	0	0	0
PEMB20	1	1	1	0	0	0	1	1	1	0
PEMB23	0	1	1	1	0	0	1	0	1	0
PEMB24	1	0	0	1	1	0	1	1	0	1
PEMB27	0	1	1	0	0	0	1	0	1	1
PEMB28	1	0	0	0	0	0	1	1	1	0
PEMB31	0	0	0	1	0	0	1	1	1	0
PEMB32	1	1	1	1	0	0	1	1	0	1
PEMB35	0	0	1	1	0	1	1	0	1	1
PEMB37	1	1	1	1	1	0	1	0	0	1
PEMB38	0	0	1	0	0	1	1	0	1	0
PEMB41	1	1	1	1	0	1	0	0	1	0
PEMB43	1	1	1	1	1	0	0	0	0	0
PEMB44	1	1	0	0	0	0	1	1	0	1
PEMB47	1	1	0	0	0	1	0	1	0	0
PEMB49	0	0	0	0	0	0	1	1	0	0
PEMB50	0	0	1	1	0	0	1	0	1	0
PEMB53	1	1	0	0	1	0	1	0	1	0
PEMB55	0	0	0	1	0	0	0	1	0	1
PEMB56	1	1	1	0	0	1	0	0	0	0
PEMB59	0	0	0	0	0	0	1	1	0	0
PEMB61	1	1	0	0	0	1	0	1	0	0
PEMB62	0	0	1	0	1	0	1	0	0	0
PEMB65	0	0	0	1	0	0	0	1	1	0
PEMB67	0	1	0	0	0	0	1	0	0	0
PEMB68	0	0	1	1	0	0	0	0	1	0
PEMB71	0	1	0	0	0	1	1	0	0	0
PEMB73	0	0	0	1	0	0	1	1	1	0
PEMB74	1	1	0	0	0	0	0	0	1	0
PEMB77	0	0	1	1	0	0	0	1	1	0
PEMB79	1	0	1	1	0	0	1	0	0	1
PEMB80	0	1	0	0	0	1	0	0	0	0
PEMB81	1	1	1	1	0	0	0	1	1	0
PEMB82	0	0	0	0	0	1	1	1	0	0
PEMB88	1	0	1	1	0	0	1	0	1	0
PEMB89	0	0	0	1	0	1	0	1	0	0
PEMB91	1	1	1	1	0	0	0	1	0	0
PEMB92	1	1	1	1	1	0	0	1	1	0
PEMB95	0	0	1	1	0	0	0	0	1	0
PEMB97	1	1	1	1	0	0	0	1	1	0
PEMB98	1	1	1	1	0	0	0	1	1	0
PEMB99	1	0	1	0	0	0	0	1	0	0
Total Items Sold	**26**	**27**	**28**	**28**	**9**	**12**	**27**	**27**	**27**	**12**

1) Market Basket Analysis Phase Using Apriori Algorithms.

a) Analysis of high frequency patterns

High frequency pattern analysis is done by finding a combination of items that meet the minimum requirements of the support value in the transaction data. The data in Table 3 shows the results of the number of transactions in each itemset:

Table 3. List of itemset.

Numb	Code BR	Name BR	Total Transactions	Initial Inventory	Stock Available
1	0000001	Black Coffee 1/2 Kg	26	45	19
2	0000002	Granulated Sugar 1 Kg	27	30	3
3	0000003	Granulated Sugar 1/2 Kg	28	40	12
4	ASL101	Gallon Bottled Water	28	80	52
5	ASL102	600 ml Bottled Water	9	240	231
6	ASL103	Glass drinking water	12	40	28
7	BR1111	Cheap rice 5 kg	27	40	13
8	BR1112	Premium Rice 5 Kg	27	40	13
9	GL1001	Javanese Sugar 1 Kg	27	40	13
10	GL1002	Java Sugar 1/2 Kg	12	60	48

The next step is to calculate the value of support in combination with 1 - itemset or C_1 using equation (1):

$$Support\ (A) = \frac{Number\ of\ Transactions\ Containing\ A}{Transaction\ Total} \quad (1)$$

The results of determining the support value of each item can be seen in table 4:

Table 4. Results for 1- itemset support values

NUMB	Code BR	Name BR	Total Transactions	Support (%)
1	0000001	Black Coffee 1/2 Kg	26	**52**
2	0000002	Granulated Sugar 1 Kg	27	**54**
3	0000003	Granulated Sugar 1/2 Kg	28	**56**
4	ASL101	Gallon Bottled Water	28	**56**
5	ASL102	600 ml Bottled Water	9	**18**
6	ASL103	Glass drinking water	12	**24**
7	BR1111	Cheap rice 5 kg	27	**54**
8	BR1112	Premium Rice 5 Kg	27	**54**
9	GL1001	Javanese Sugar 1 Kg	27	**54**
10	GL1002	Java Sugar 1/2 Kg	12	**24**

The next step is to identify a combination pattern with a minimum support of 50%, then the process of forming 2 - itemset or C_2 with a minimum support of 50% as shown in table 5.

Table 5. Combination patterns of C_1 with a minimum support of 50%.

Numb	Code BR	Name BR	Total Transactions	Support (%)
1	0000001	Black Coffee 1/2 Kg	26	**52**
2	0000002	Granulated Sugar 1 Kg	27	**54**
3	0000003	Granulated Sugar 1/2 Kg	28	**56**
4	ASL101	Gallon Bottled Water	28	**56**
5	BR1111	Cheap rice 5 kg	27	**54**
6	BR1112	Premium Rice 5 Kg	27	**54**
7	GL1001	Javanese Sugar 1 Kg	27	**54**

Calculation of support values using a 2-itemset combination using equation (2):

Support(A,B)=P(A∩B)

$$\text{Support}(A, B) = \frac{\Sigma\ transaction\ contains\ A\ and\ B}{\Sigma\ Total\ Transactions} \quad (2)$$

Table 6. Minimum support values of 50% of the 2-itemset combination

Numb	Item Name	Amount	Support (%)
1	Black Coffee 1/2 Kg → Granulated Sugar 1 Kg	19	38
2	Black Coffee 1/2 Kg → Granulated Sugar 1/2 Kg	18	36
3	Black Coffee 1/2 Kg → Gallon Bottled Drinking Water	5	10
4	Black Coffee 1/2 Kg → Cheap Rice 5 Kg	12	24
5	Black Coffee 1/2 Kg → Premium Rice 5 Kg	16	32
6	Black Coffee 1/2 Kg → Java Sugar 1 Kg	14	28
7	Granulated Sugar 1 Kg → Granulated Sugar 1/2 Kg	16	32
8	Granulated Sugar 1 Kg → Gallon Bottled Drinking Water	13	26
9	Granulated Sugar 1 Kg → Cheap Rice 5 Kg	12	24
10	Granulated Sugar 1 kg → Premium Rice 5 kg	13	26
11	Granulated Sugar 1 kg → Javanese Sugar 1 kg	13	26
12	Gallon Bottled Drinking Water → Cheap Rice 5 Kg	13	26
13	Gallon Bottled Drinking Water → Premium Rice 5 Kg	14	28
14	Gallon Bottled Drinking Water → Java Sugar 1 Kg	18	36
15	Cheap Rice 5 Kg → Premium Rice 5 Kg	13	26
16	Cheap Rice 5 Kg → Java Sugar 1 Kg	15	30
17	Premium Rice 5 Kg → Java Sugar 1 Kg	12	24

In the process of forming a 2-itemset combination pattern no one meets the minimum support value of 50%, then a 1-itemset combination can be met for the formation of associative rules.

2) **Establishment of associative rules**

After a minimum support value of 50% of the C_1 combination is determined, an associative rule is sought that meets the minimum confidence requirement of candidate 2 item A containing B in each itemset. The confidence value of the associative rule 2 - item A → B is obtained using equation (3):

Confidence (A,B)=

$$P(B|A) = \frac{Number\ of\ transactions\ containing\ A\ and\ B}{Number\ of\ transactions\ containing\ A} \qquad (3)$$

Table 7. Confidence values from forming 2-item associative rules A → B.

Numb	Itemset	Amount	Confident (%)
1	(0000001, Black Coffee 1/2 Kg), (0000002, 1 Kg Granulated Sugar)	18	69,23
2	(0000001, Black Coffee 1/2 Kg), (0000003, Granulated Sugar 1/2 Kg)	18	69,23
3	(0000001, Black Coffee 1/2 Kg), (ASL101, Gallon Bottled Drinking Water)	15	57,69
4	(0000001, Black Coffee 1/2 Kg), (BR1111, Cheap Rice 5 Kg)	13	50,00
5	(0000001, Black Coffee 1/2 Kg), (BR1112, Premium Rice 5 Kg)	16	61,53
6	(0000001, Black Coffee 1/2 Kg), (GL1001, Javanese Sugar 1 Kg)	14	53,84
7	(0000002, 1 Kg Granulated Sugar), (0000003, 1/2 Kg Granulated Sugar)	16	59,25
8	(0000002, 1 Kg Granulated Sugar), (ASL101, Gallon Bottled Drinking Water)	13	48,14
9	(0000002, 1 Kg Granulated Sugar), (BR1111, Cheap Rice 5 Kg)	12	44,44
10	(0000002, 1 Kg Granulated Sugar), (BR1112, Premium Rice 5 Kg)	13	48,18
11	(0000002, 1 Kg Granulated Sugar), (GL1001, 1 Kg Java Sugar)	13	48,14
12	(0000003, Granulated Sugar 1/2 Kg), (ASL101, Gallon Bottled Drinking Water)	20	71,42

Numb	Itemset	Amount	Confident (%)
13	(0000003, Granulated Sugar 1/2 Kg), (BR1111, Cheap Rice 5 Kg)	16	57,14
14	(0000003, Granulated Sugar 1/2 Kg), (BR1112, Premium Rice 5 Kg)	12	42,85
15	(0000003, Granulated Sugar 1/2 Kg), (GL1001, Java Sugar 1 Kg)	18	64,28
16	(ASL101, Gallon Bottled Water), (ASL102, 600ml Bottled Water)	6	21,42
17	(ASL101, Gallon Bottled Drinking Water), (BR1111, Cheap Rice 5 Kg)	13	46,42
18	(ASL101, Gallon Bottled Drinking Water), (BR1112, Premium Rice 5 Kg)	14	50,00
19	(ASL101, Gallon Bottled Drinking Water), (GL1001, Java Sugar 1 Kg)	18	64,28
20	(BR1111, Cheap Rice 5 Kg), (BR1112, Premium Rice 5 Kg)	13	48,14
21	(BR1111, Cheap Rice 5 Kg), (GL1001, Java Sugar 1 Kg)	14	51,85
22	(BR1112, Premium Rice 5 Kg), (GL1001, Java Sugar 1 Kg)	14	51,85

Next is determining the minimum confidence value of 60%, the confidence value of 2 - items A → B that will qualify for the determination of association rules has a confidence value above 60%.

Table 8. Value of Association rules with a minimum value of 60% confidence.

Numb	Itemset	Amount	Confident (%)
1	(0000001, Black Coffee 1/2 Kg) → (0000002, 1 Kg Granulated Sugar)	18	69,23
2	(0000001, Black Coffee 1/2 Kg) → (0000003, Granulated Sugar 1/2 Kg)	18	69,23
3	(0000001, Black Coffee 1/2 Kg) → (BR1112, Premium Rice 5 Kg)	16	61,53
4	(0000003, Granulated Sugar 1/2 Kg) → (ASL101, Gallon Bottled Drinking Water)	20	71,42
5	(0000003, Granulated Sugar 1/2 Kg) → (GL1001, Java Sugar 1 Kg)	18	64,28
6	(ASL101, Gallon Bottled Drinking Water) → (GL1001, Java Sugar 1 Kg)	18	64,28

Table 9. Results of determining the pattern of handling stock of goods.

NUMB	KODE BR	NAME BR	TOTAL TRANSACTIONS	SUPPORT (%)	INITIAL INVENT-ORY	ADDITI-ON STOCK
1	0000001	Black Coffee 1/2 Kg	26	52	45	23
2	0000002	Granulated Sugar 1 Kg	27	54	30	16
3	0000003	Granulated Sugar 1/2 Kg	28	56	40	22
4	ASL101	Gallon Bottled Water	28	56	52	29
5	BR1111	Cheap rice 5 kg	27	54	40	22
6	BR1112	Premium Rice 5 Kg	27	54	40	22
7	GL1001	Javanese Sugar 1 Kg	27	54	40	22

Table 10. Association Rules

Numb	Rules	Confidence %
1	If you buy ½ Kg Black Coffee then buy 1 Kg Granulated Sugar	69,23
2	If you buy ½ Kg Black Coffee then buy ½ Kg Sugar	69,23
3	If you buy ½ Kg Black Coffee then buy 5 Kg Premium Rice	61,53
4	If you buy ½ Kg Granulated Sugar then buy Gallon Bottled Drinking Water	71,42
5	If you buy ½ Kg Granulated Sugar then buy 1 kg of Java Sugar	64,28
6	If you buy Gallon Bottled Drinking Water then buy Java Sugar 1 Kg	64,28

c. Results Determination Stage

From the results of the analysis of determining the value of support and confidence, it can be concluded the results of the pattern of handling stock of goods and arrangement of goods on the store shelf, as follows:

1. Pattern of handling stock of goods

   From the results of determining the minimum support value of 50% in the 1-itemset combination pattern, it can be identified the addition of stock items with a percentage:

$$Stock\ Addition = \frac{Initial\ inventory}{Value\ of\ Support\ (\%)} x\ 100\% \quad (4)$$

   From Table 9 results can be drawn from the stage of identifying the pattern of handling stock of goods utilizing a minimum value of support of 50% of the combination of 1-itemset, by calculating the initial inventory is compared with the value of support so that it can be identified in the item stock increase column in Table 9.

2. Pattern of Arrangement of Goods

   From the results of the analysis in identifying the handling of the layout of the goods using Apriori Algorithm, it can be determined from the minimum value of 50% support and the minimum value of Confidence 60% which produces the following association rules:

   From Table 10. it can be identified the tendency of customer purchases, it is known that the tendency of customers if buying Black Coffee items ½ Kg is most likely with a Confidence value of 69.23% customers will also buy 1 Kg Granulated Sugar. from this tendency can be a strategy to arrange the layout of the item with the pattern of Black Coffee ½ Kg brought closer to 1 Kg Granulated Sugar, Black Coffee ½ Kg closer to Granulated Sugar ½ Kg and so on, so that customers can easily find the items they need.

4. Conclusion

Based on the analysis phase of the results and discussion above, it can be concluded that identification of patterns of handling stock of goods and arrangement of goods can utilize data on the results of sales transactions. And then, analysis of high frequency patterns with a minimum support value of 50% from a combination of 1 - itemset C1 can determine the pattern of handling stock of goods, namely by balancing the initial inventory with a support value so that the prediction results of adding stock will be obtained. Beside that, the results of the formation of association rules that are determined from a minimum value of 50% support and a minimum value of 60% confidence can produce a tendency of customers to buy items, so that these tendencies can be a reference in the process of item layout by arranging items close together.

References

[1] A. Junaidi, "Implementasi Algoritma Apriori dan FP-Growth untuk Menentukan Persediaan Barang," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 8, no. 1, pp. 61-67, Maret 2019.

[2] H. Jiawei dan K. Micheline, *Data Mining : Concept and Techniques Second Edition*, Morgan Kaufman Publishers, 2006.

[3] D. T. Larose, Discovering knowledge in data : an introduction data mining, Jhon Wiley & Sons Inc, 2005.

[4] Suprayogi, dan A. Karima, "Implemetasi Algoritma Apriori dengan Market Basket Analysis untuk Pengaturan Tata Letak Produk," *Jurnal SISFOTENIKA*, vol. 9, no. 2, pp. 169-179, Juli 2019.

[5] Maharani, N. A. Hasibuan, dan N. Silalahi, "Implementasi Data Mining untuk Pengaturan Layout Minimarket dengan Menerapkan Association Rule," *Jurnal Riset Komputer*, vol. 4, no. 4, pp. 6-11, Agustus 2017.

[6] A. K. Prasidya, dan C. Fibriani "Analisis Kaidah Asosiasi Antar Item Dalam Transaksi Pembelian Menggunakan Data Mining Dengan Algoritma Apriori (Studi Kasus Gun Bandungan Jawa Tengah," *Jurnal Ilmiah Teknologi Informasi*, vol. 14, no. 2, pp. 173-184, Juli 2017.

[7] M. Sholik, dan A. Salam, "Implementasi Algoritma Apriori untuk Mencari Asosiasi BArang yang Dijual di E-commerce OrderMAs," *Jurnal Techno. Com*, vol. 17, no. 2, pp. 158-170, Mei 2018.

[8] G. Abdurrahman, "Analisis Aturan Asosiasi Data Transaksi Supermarket Menggunakan Algoritma Apriori," *Jurnal Sistem dan Teknologi Informasi Indonesia,* vol. 2, no.2, pp. 100-111, Agustus 2017.

[9] R. Yanto, dan R. Khorirah, "Implementasi Data Mining dengan Metode Algoritma Apriori dalam Menentukan Pola Pembelian Obat," *Creative Information Technology Journal*, vol. 2, no. 2, pp. 102-113, April 2015.

[10] S. Wahyuni, Suherman, dan L. P. Harahap, "Implementasi Data Mining dalam Memprediksi Stok Barang Menggunakan Algoritma Apriori," *Jurnal Teknik dan Informatika*, vol. 5, no.2, pp. 67-71, Juli 2018.

[11] E. D. Sikumbang, "Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori," *Jurnal Teknik Komputer*, vol. 4, no. 1, pp. 156-161, Februari 2018.

[12] Y. Suardi, B. F. Ahmad, M. Ita, dan W. Bambang,

"Penerapan Data Mining Pengaturan Pola Tata Letak Barang pada Berkah Swalayan untuk Strategi Penjualan Menggunakan Algoritma Apriori," *Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD,* vol. 2, no. 1, pp. 69-75, Januari 2019.

[13] A. Cep, H. Nila, dan W. Wiwiek, "Implementasi Data Mining Penjualan Kosmetik Pada Toko Zahrani Menggunakan Algoritma Apriori," *Sentra Penelitian Engineering dan Edukasi,* vol. 11, no. 2, pp. 1-7, Mei 2019.

[14] B. A. Pilipus, R. J. Ekik, dan L. Yonata, "Prediksi Pola Pembelian *Customer* dengan *Market Basket Analysis* pada PT. Capella Medan," *Jurnal Sistem Informasi dan Komputer,* vol. 2, no. 2, pp. 59-66, Maret 2019.

[15] Y. S. Haysrif, Rismayani, dan L. S Novita, "Data Mining Menggunakan Algoritma Apriori untuk Analisis Penjualan," Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi, vol. 6, no. 1, pp. 217-226, Agustus 2019.

khazanah informatika
Jurnal Ilmu Komputer dan Informatika

User Experience Design in Mobility Assistant Application for the Physically Disabled Using the Wheel Method

Azman Fatahillah*, **Auzi Asfarian**
Department of Computer Science
Institut Pertanian Bogor
Bogor, Indonesia
*azman.fatahillah1@gmail.com

Abstract-This research was conducted to design a mobility assistant application for persons with physical disability using The Wheel method. This application can assist mobile activities for people with disability. The application called Kuygo was made in four stages: analysis, design, prototype, and evaluation. Analysis of user needs is carried out at Loka Bina Karya (LBK) Bogor through observation and interviews. The interaction design requirements are generated in the form of requirements statements and inventory tasks with three main tasks, namely mobility preparation, mobility comfort, and providing a review of a location. Afterwards, discussions were held in the form of design thinking and ideation sessions, with people who care about people with disabilities and the Senyum Difabel community, which produced sketches, storyboards, and wireframes. Furthermore, the design implementation is carried out by making a prototype of medium-fidelity and the results are tested using cognitive walkthrough techniques. Of the four questions that must be answered with cognitive walkthrough techniques, the average success rate of all the tasks tested is 94.62%. Based on these results, no major usability errors were found in the prototype medium-fidelity and the Kuygo application can be further developed.

Keywords: mobility assistant, disability, mobility, physical disability

1. Introduction

Everyone hopes for a perfect physique with the completeness of the limbs, but not everyone gets a gift from God to get a complete body and mental fitness. There are also those who have physical disabilities or illnesses so there are limitations in their activities. They are known as people with disabilities. Physically disabled person is a person who has a physical/body deficiency or abnormality that causes interference with their personal activities and development [1]. Disabilities has two categories, namely ambulant-disabled and wheelchair-bound disabled. Ambulant-disabled persons are persons with disability who can still move around using tools without using a wheelchair, while wheelchair-bound disabled are those who have limitations in mobilization and are certain to use a wheelchair to carry out daily activities [2].

Data from the Ministry of Health of the Republic of Indonesia states that there are 6,515,500 people with disabilities in Indonesia in 2012 and as many as 10.26% of people experience physical disability [3]. Over time with the increasing number of accidents in Indonesia, it is predicted that people with disabilities in Indonesia will also increase [4]. Indonesia, with a population of around 265 million, should have good enough public services, especially for accessibility. In Law Number 4 of 1997 article 1 (paragraph 1) and Government Regulation Number 43 of 1998 in particular article 1 (paragraph 1) it has been explained that citizens who are both disabled and non-disabled have the right to get equality of position, rights and obligations, and play a role in the society according to his ability. This accessibility is very important to help people with disabilities carry out activities independently. However, often people with disabilities still find it difficult to get services and information about accessibility for people with disabilities. In addition to information, they also sometimes need a companion or assistant in their activities. Trained animals are sometimes used to assist them in order to imrpove independence and carry out social activities [5].

Mobility is the ability to manipulate objects and the ability of an individual to move in their environment. Mobility assistants help people who have difficulty in mobility because of physical impairments or disabilities. This mobility assistant is applied by using technology-based also increasingly developed such as electric powered wheelchairs, prosthetics, functional electrical stimulation, and exoskeleton robots [6]. Mobility assistant can not only be applied in the form of hardware technology, but can also be applied to software, one of which is application technology. This last form can be used more practically by various walks of life because it can run smart gadgets in general which are owned by the majority of the Indonesian population.

This research will design a mobile application that makes it easy for persons with disabilities to obtain information and get appropriate accessibility when doing mobility. This research resulted in a prototype application of mobility assistant for the disabled with visual impairment through the process of designing the user experience of The Wheel [7].

2. Method

This research was conducted using The Wheel [7] cycle method which consists of four stages, namely analysis, design, prototype, and evaluation. The analysis is done by doing contextual inquiry to the user, describing it in the form of a work activity affinity diagram, and modeling it in the form of a hierarchical task inventory (HTI). Based on the HTI produced, the application design is made through a process of design thinking and ideation in groups to produce many possible solutions. The best solution is made in the form of sketches, storyboards, and wireframes. Wireframe is made in the form of a medium-fidelity prototype that will be evaluated by users using cognitive walkthroughs.

a. Analysis

In the first stage of the analysis is to do contextual analysis with direct observation and interviews to three respondents with disabilities. Observation was carried out at Loka Bina Karya (LBK) Bogor by observing the mobility activities carried out by persons with disability and conducting interviews by asking 12 questions related to experiences when mobility to the response to the system to be made. The results of these observations and interviews are made into a label that contains the user's work activity notes.

Activities that have been written in the work activity notes are then combined with work activity affinity diagrams (WAAD) which aim to visualize the contextual data obtained by grouping work activities. WAAD consists of four levels: the first level is the overall scope of the objectives of the research activity, the second level contains questions, the third level (group label) which is the first grouping for work activity, and the fourth level which is the work activity (work activity) of the user.

b. Design

At this stage, design thinking and ideation is carried out with a team from the community concerned with disabilities and the Senyum Difabel community to explore as many solutions as possible from the issues raised. The problem is divided into three main tasks: "getting information to prepare for mobility", "getting information and security while traveling", and "giving a review of a location". This stage produces user persona, sketches, storyboards, and wireframes.

c. Prototype

The results of the design which includes three main tasks are implemented in a medium-fidelity prototype. The medium-fidelity prototype was created to make it easier for users to interact with the system. The medium-fidelity prototype is carried out after the initial design is used for more detailed design purposes and uses validation. This medium-fidelity presents detailed information such as navigation, functionality, content and layout, as well as form estimates [8]

d. Evaluation

At this stage, testing and evaluation of prototypes that have been made in the previous stage are carried out. The prototype was tested using cognitive walkthrough (CW) techniques to get direct feedback from users. CW is one of the testing methods used to test system usability and find the cause of problems that occur in the system usability in a design process [9]. CW assesses the ease of new users in completing tasks with a system. In evaluations using this CW technique relies on a series of detailed questions that must be answered by the user when evaluating specifically through exploration.

Participants in this test will evaluate the interface provided based on the specified task. Input in one walkthrough session includes the design of the system interface, scenarios for each task, and the sequence of actions that must be successfully performed to fulfill each task tested. At each step of the action in a task, the evaluator must observe what the participant is doing and answer the four questions that are the success rate of this evaluation. The questions are as follows: (1) 'Will the user try to achieve the right results?' (2) 'Will the user see that the correct action is available for them?' (3) 'Will the user associate the correct action with their expected results? ', and (4)' If the correct action is taken, will the user see that progress is being made towards the expected results? '[9]. This evaluation was tested on two participants with physical disabilities who used wheelchairs.

3. Results and Discussion

a. Analyst

During the analysis phase interviews were conducted with respondents asking 12 questions to three respondents with disability related to their daily mobility. Several points are obtained based on the results of interviews and observations, including regarding location, transportation, and social information. Respondents often experience difficulties when using transportation are wheelchair user respondents because there are still many transportation facilities that is not so disable-friendly. In addition, information on disability facilities related to a location is needed for persons with disabilities, especially wheelchair users to facilitate their needs. The lack of empathy of non-disabled people is still felt by people with disabilities so they have to struggle alone in carrying out their activities.

Data from interviews and observations in the previous stage were analyzed by making notes of work activities carried out by respondents. The relationship between notes of user work activities on the Kuygo application is interpreted into a work activity affinity diagram (WAAD) that has a research objective, namely the user mobility activity reaches the destination location written at the first level. Each label is coded to make it easier to trace at a later stage. The code is Z for the purpose of research activities (first level), A ... Z for the second level, (AA, AB, ...) for the third level, and (AA1, AB1, ...) for the fourth level. The WAAD diagram can be seen in full in Figure 1. Furthermore, making a requirement document to determine the system interaction design needs based on user work activities. In addition, the feasibility of implementing these needs is also estimated. Full results can be seen in Table 1.

After making the requirements statement, a hierarchical task inventory (HTI) is made which is focused on the user's role, namely the physical disability and can be seen in Figure 2. This HTI is made to show the structure of the user's tasks and actions and guide the overall interaction design. In this HTI, it looks supertask in the form of 'doing mobility activities to the destination location' which is divided into three main tasks. Each main task is divided into smaller tasks.

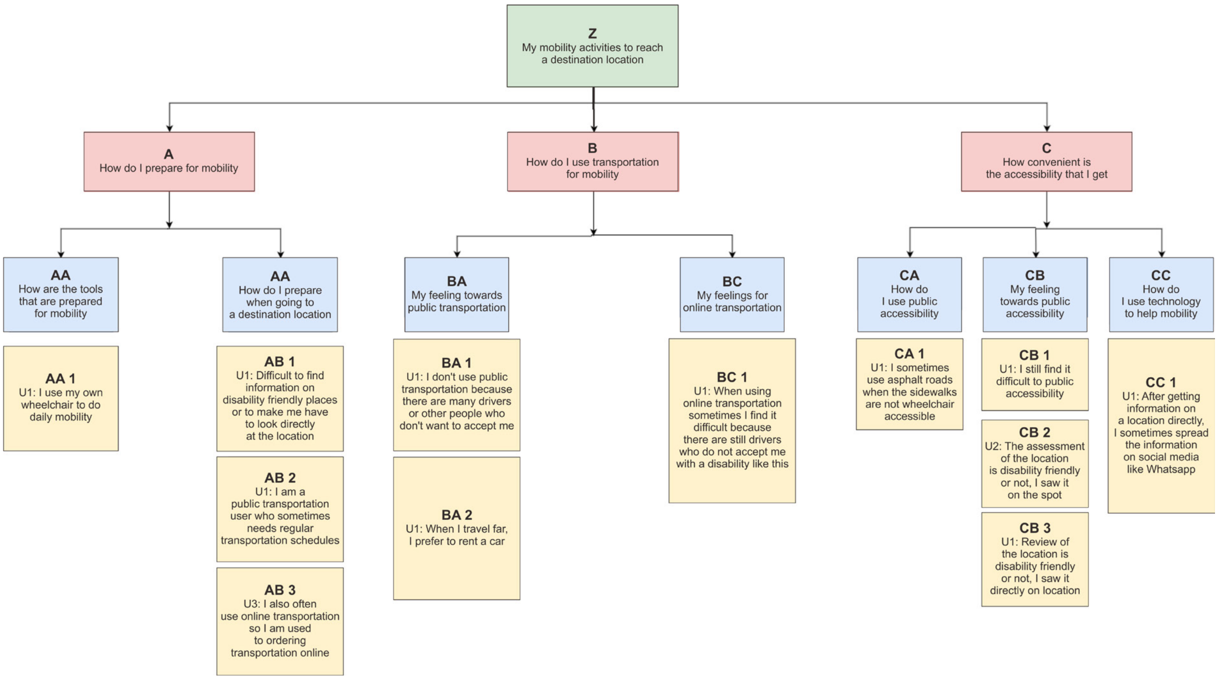

Z
My mobility activities to reach a destination location
A
How do I prepare for mobility
B
How do I use transportation for mobility
C
How convenient is the accessibility that I get
AA
How are the tools that are prepared for mobility
AA
How do I prepare when going to a destination location
BA
My feeling towards public transportation
BC
My feelings for online transportation
CA
How do I use public accessibility
CB
My feeling towards public accessibility
CC
How do I use technology to help mobility
AA 1
U1: I use my own wheelchair to do daily mobility
AB 1
U1: Difficult to find information on disability friendly places or to make me have to look directly at the location
AB 2
U1: I am a public transportation user who sometimes needs regular transportation schedules
AB 3
U3: I also often use online transportation so I am used to ordering transportation online
BA 1
U1: I don't use public transportation because there are many drivers or other people who don't want to accept me
BA 2
U1: When I travel far, I prefer to rent a car
BC 1
U1: When using online transportation sometimes I find it difficult because there are still drivers who do not accept me with a disability like this
CA 1
U1: I sometimes use asphalt roads when the sidewalks are not wheelchair accessible
CB 1
U1: I still find it difficult to public accessibility
CB 2
U2: The assessment of the location is disability friendly or not, I saw it on the spot
CB 3
U1: Review of the location is disability friendly or not, I saw it directly on location
CC 1
U1: After getting information on a location directly, I sometimes spread the information on social media like Whatsapp

Figure 1. Work activity affinity diagram (WAAD) of the Kuygo application

Table 1. Requirements document for the Kuygo application

ID	Work activity	Requirement statements	Feasibility
AA1	Find out mobility information for wheelchair users	Wheelchair users will receive wheelchair mobility information	√
AB1	Knowing disability friendly location information	Wheelchair users must be able to find disability-friendly location information	√
AB2	Knowing public transportation schedule information	Wheelchair users will receive transportation schedule information to be on time to get transportation	√
AB3	Look for transportation online	Wheelchair users must be able to find transportation information online so they can rent it	√
BA1	Knowing disability-friendly public transportation	Wheelchair users will receive any transportation information that can be used with disabilities	√
BA2	Looking for vehicle rental when traveling far away	Seat users must be able to find a vehicle to rent when going long distance	√

ID	Work activity	Requirement statements	Feasibility
BC1	Knowing disability-friendly online transportation	Wheelchair users will receive disability-friendly online transportation information for convenience	√
CA1	Knowing wheelchair lane information	Wheelchair users must be able to find a route that can be passed by a wheelchair for mobility convenience	√
CB1	Knowing information about disability friendly facilities in a place	Wheelchair users must be able to find disability-friendly location information based on location criteria	√
CB2	Rating a place	Wheelchair users must be able to provide ratings so that other users can find out the location that has been assessed	√
CB3	Write a review somewhere	A wheelchair user must be able to give a reason for a location to be seen by other users for his assessment	√
CC1	Share disability-friendly location information	Wheelchair users must be able to disseminate disability-friendly location information so other users are able to mobility comfortably	√

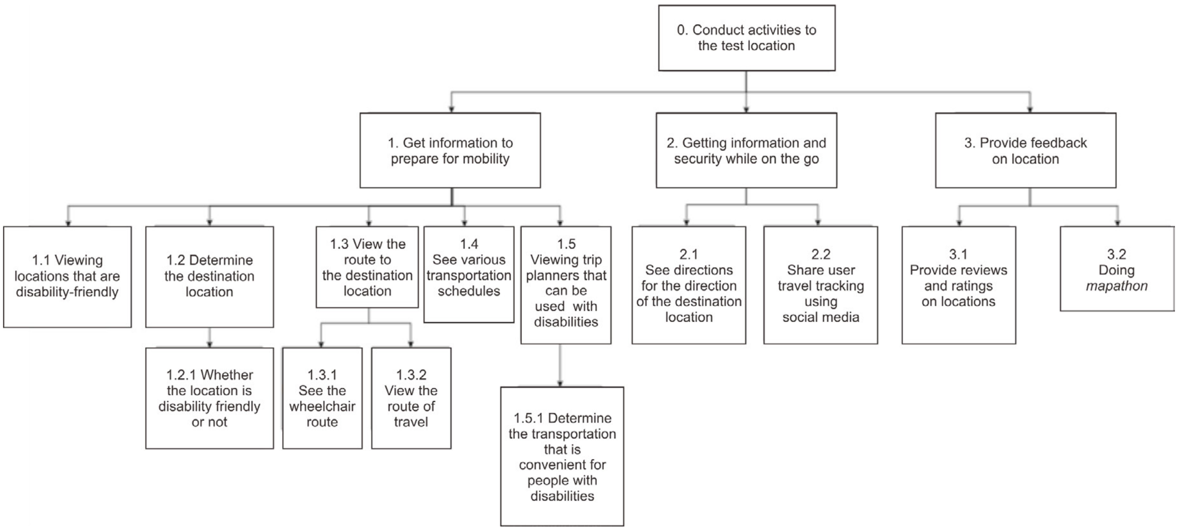

0. Conduct activities to the test location
1. Get information to prepare for mobility
2. Getting information and security while on the go
3. Provide feedback on location
1.1 Viewing locations that are disability-friendly
1.2 Determine the destination location
1.3 View the route to the destination location
1.4 See various transportation schedules
1.5 Viewing trip planners that can be used with disabilities
2.1 See directions for the direction of the destination location
2.2 Share user travel tracking using social media
3.1 Provide reviews and ratings on locations
3.2 Doing mapathon
1.2.1 Whether the location is disability friendly or not
1.3.1 See the wheelchair route
1.3.2 View the route of travel
1.5.1 Determine the transportation that is convenient for people with disabilities

Figure 2. Kuygo's Hierarchical task inventory (HTI) application.

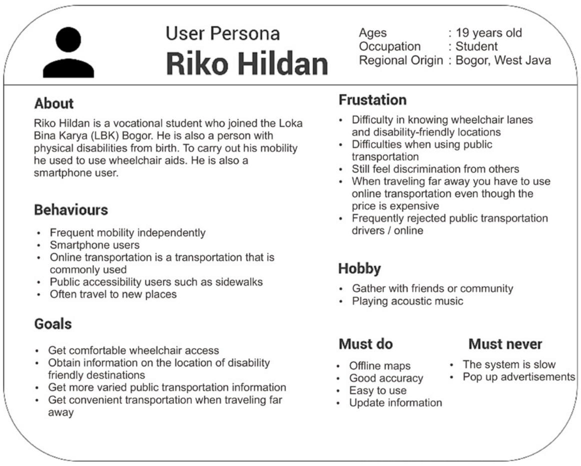
b. Design

In the first stage the design is done by making persona to describe the characteristics of the target user of the application to be developed. Personas are made based on the process of analysis of the results of the interview in the previous stage. User persona for the Kuygo application in Figure 3.

Then the idea was carried out by brainstorming with a team consisting of two people who care about people with disabilities and one of the Senyum Difabel community. In this process of ideation is based on three main tasks on HTI, namely 'getting information to prepare for mobility', 'getting information and security while traveling', and 'giving a review of a location'. The results of the design of ideas and solutions from brainstorming are described in the form of eight sketches (crazy eight) for each main task. The results of eight sketches for the three main tasks can be seen in Figure 4.

After making eight sketches for each of the main tasks, voting is then conducted to determine the best sketch for each task. For the first task, the chosen sketch is a sketch that displays location information, accessibility information to the destination location, as well as information on transportation recommendations that can be accessed by people with disabilities. For the second task, the selected sketch provides information on directions and user safety while traveling by providing emergency calls and can share the user's journey to others through social media. Finally, the selected sketch for the third task is a sketch that provides a place for users to do a review and mapping of a location and its path whether disability friendly or not. All three sketches contain a list of tasks as in Table 2.

After the three sketches have been selected, a storyboard for each sketch will be made. Storyboard is made to know the flow of the use of the system as well as contests of its use from the user's side. One of the storyboards can be seen for the first task, which is mobility preparation in Figure 5, which tells the user who is going to a location but wants to know if the location is disability friendly or not by using the Kuygo application.



User Persona

Riko Hildan

Ages : 19 years old
Occupation : Student
Regional Origin : Bogor, West Java

About

Riko Hildan is a vocational student who joined the Loka Bina Karya (LBK) Bogor. He is also a person with physical disabilities from birth. To carry out his mobility he used to use wheelchair aids. He is also a smartphone user.

Behaviours

- Frequent mobility independently
- Smartphone users
- Online transportation is a transportation that is commonly used
- Public accessibility users such as sidewalks
- Often travel to new places

Goals

- Get comfortable wheelchair access
- Obtain information on the location of disability friendly destinations
- Get more varied public transportation information
- Get convenient transportation when traveling far away

Frustation

- Difficulty in knowing wheelchair lanes and disability-friendly locations
- Difficulties when using public transportation
- Still feel discrimination from others
- When traveling far away you have to use online transportation even though the price is expensive
- Frequently rejected public transportation drivers / online

Hobby

- Gather with friends or community
- Playing acoustic music

Must do

- Offline maps
- Good accuracy
- Easy to use
- Update information

Must never

- The system is slow
- Pop up advertisements

Figure 3. User persona of the Kuygo application.

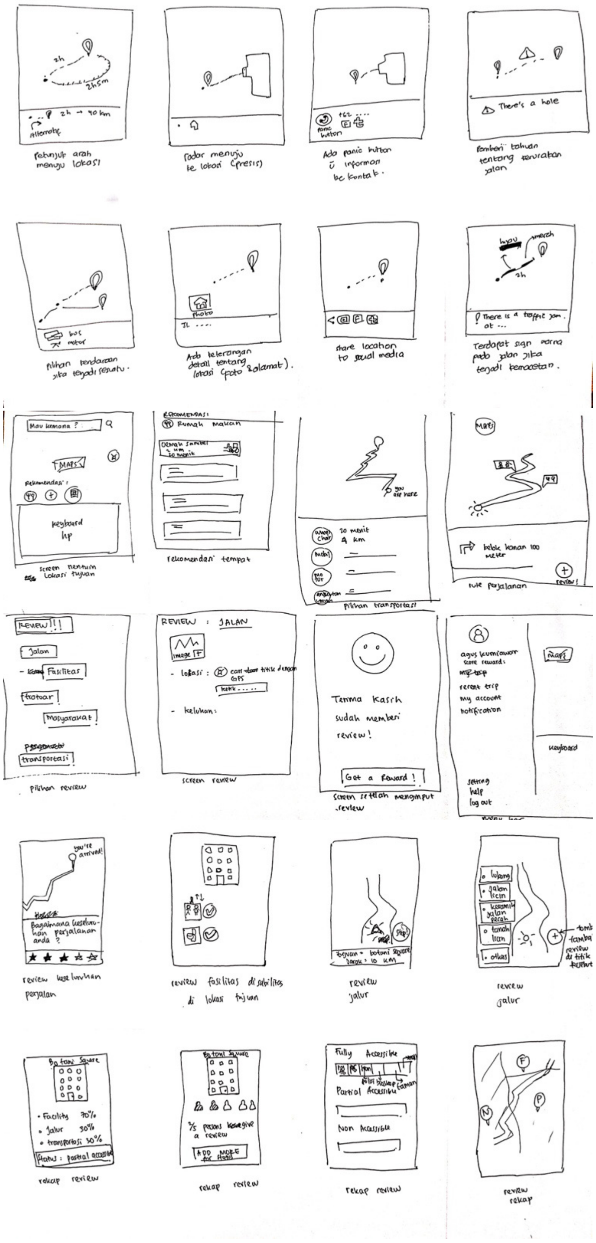


Figure 4. The results of crazy eight activities for the three main tasks in the Kuygo application.

Based on the storyboard produced, the application design is made in a higher level of fidelity, namely wireframes. This form focuses on the interface that appears in an application and shows more clearly the information displayed along with its layout on the screen. Wireframes are created using Adobe Experience Designer software and follow Google Material Design standards. Wireframes will be increased fidelity into a medium to show more visual elements and interactions that will be felt by users when using this application. An example of a wireframe created can be seen in Figure 6, which is for the task of preparing for mobility.

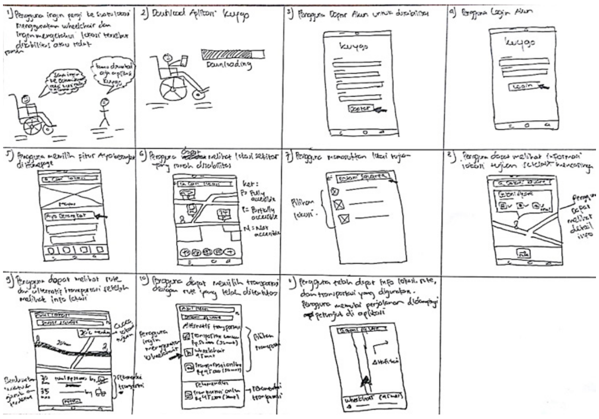


Figure 5. Storyboard for mobility preparation tasks

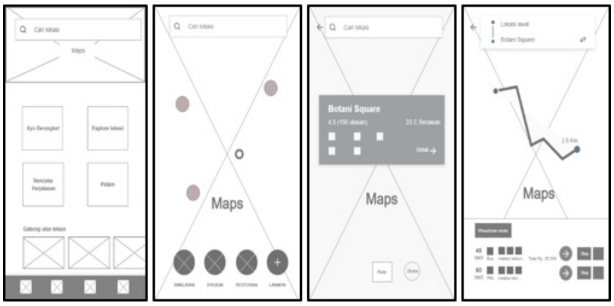


Figure 6. Kuygo Wireframes application

c. Prototype

The prototype stage is the stage of implementing the results of the previous stage. The current prototype is a feature providing information for preparing user mobility with location and transportation information that can be used. This prototype was made with a medium-fidelity level. The prototype can be accessed at bit.ly/kuygo application and the following sample prototype of the Kuygo application in Figure 7.

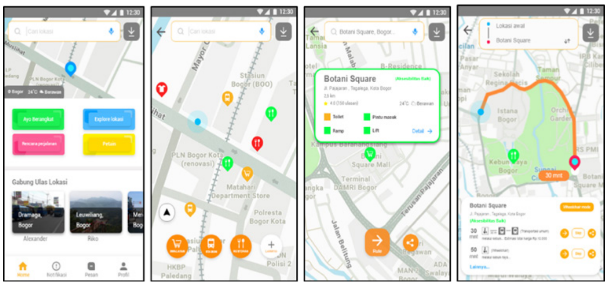


Figure 7. Kuygo application medium-fidelity prototype

d. Evaluation

The overall results of the Kuygo application prototype were evaluated using cognitive walkthrough (CW) techniques. In this study, testing with 16 tasks that must be completed by participants. This evaluation involved two participants who came from Loka Bina Karya (LBK) Bogor who are wheelchair users. This 16 task test is a test based on HTI that has been made.

There are 4 walkthrough questions to evaluate the test results based on the application interface. The first question is to analyze whether the user will try to achieve the right results. This relates to the understanding of users in completing an action. The second question is to analyze whether users will see that the correct actions are available for them. This second question relates to the availability of interface controls. The third question is to analyze whether the user will associate the right action with the results they expect. Relates to labels on available controls. The fourth question is to analyze what if the correct action is taken, will the user see that progress is being made towards the expected results. This relates to the user interface feedback that should be received.

Table 2. Testing tasks

Task
Main Task 1 Get information for mobilization preparation.
1. The user is logged in as a registered user.
2. Users explore locations to see locations that are friendly with disabilities.
3. The user searches for the destination location based on the last search.
4. The user sees destination location information.
5. Users see the recommended wheelchair route.
6. The user searches for the desired transportation alternative along with the travel route.
7. The user saves a trip planner for the user's scheduled trip.
Main Task 2 Get information and security on the way
8. The user starts the trip by looking at the directions.
9. Users share trips for the user's travel security.
10. Users make an emergency button for emergencies when traveling.
11. The user searches for the location of his destination when he is around the location with the Augmented Reality (AR) maps feature.
Main Task 3 Provide feedback on location
12. The user ends the trip by giving a review and rating of the location that has been done.
13. The user does a mapathon / location mapping.
14. Users get mapathon notifications.
15. The user starts the mapathon.
16. The user terminates mapathon.

To measure the success of each action on the task, a representation of the results is carried out, labeled 'Yes' with a weight of 1 if the action was successful and labeled 'No' with a weight of 0 if the action was failed [10]. Every action that is labeled 'No' must be accompanied by a description of the cause of failure for the action (Table 3). Then count the number of successful actions on each task (Table 4). After calculating the number of successful actions, each number of normalized actions by counting the number of successful actions divided by the results of the total actions on a task multiplied by the number of partisans. The results of normalization are summed and the average is taken from each walkthrough question (Table 5). The calculation results obtained success rate based on four questions raised. The success rates for the four questions are 97.25%, 90.87%, 96.18% and 94.18%, respectively.

Based on the average value of the results of the normalization of success there is the lowest average normalized value of each walkhtrough question. In the first question, which is "Will the user try to achieve the right results?" Has the lowest value of 0.80 on task 7 because the user does not understand the save button available so the user fails to save the itinerary.

Task 10 has the lowest value, which is 0.50 for the second question, which is 'Will the user see that the correct action is available for them?'. That is because the user is not too visible with the available text that is too contrasting in color and also the button does not have the word 'emergency' so that users find it difficult to find the 'emergency' button.

In the third question, which is 'Will the user associate the correct action with the results they expect?', There is the lowest value on task 10 which is 0.66. That is because the user feels that there are too many stages for emergency calls so that they cannot contact emergency calls as soon as possible.

The final question, which is 'If the correct action is taken, will the user see that progress is being made towards the expected results?' There is the lowest value on task 7 of 0.80. This is again because the user did not complete the storage because he did not know the save button and also after being notified of the actual action, the user needed a message after the itinerary was saved.

Based on these results, it can be concluded that no major usability errors were found in the prototype medium-fidelity and the Kuygo application can be further developed. However, for the next prototype or product, further testing is needed using the number of participants in the top two people which allows for greater reusability of major problems.

Table 3. Example of cognitive walkthrough test results

Task 15: The user starts the mapathon					
Numb	**Action**	**a[1]**	**b[2]**	**c[3]**	**d[4]**
1	The user selects the correct navigation ('Petain' on 'Home' navigation)	Y	Y	Y	Y
2	The user selects 'current pet'	Y	Y	Y	Y
3	The user presses the 'Start' button to start mapathon	Y	Y	Y	Y
4	The user added a new location mapathon	Y	Y	Y	Y
5	The user fills in the mapathon form	Y	N	Y	Y
6	The user presses the 'submit' button	Y	Y	Y	N

(1) Will users try to achieve the right results ?;

(2) Will users see that the right action is available for them ?;

(3) Will the user associate the right action with the results they expect ?; and

(4) If the correct action is taken, will the user see that progress is being made towards the expected results?

Table 4. Number of successful actions

	Walkthrough Question							
	1		**2**		**3**		**4**	
Participation	1	2	1	2	1	2	1	2
Task 1	3	3	3	3	3	3	2	3
Task 2	3	3	3	3	3	3	3	3
Task 3	3	3	3	3	3	3	3	3
Task 4	3	3	3	2	3	3	3	3
Task 5	2	2	2	1	2	2	2	2
Task 6	2	2	2	2	2	2	2	2
Task 7	4	4	4	4	4	4	4	4
Task 8	2	2	2	2	2	2	2	1
Task 9	2	2	2	2	2	2	2	2
Task 10	3	2	2	1	2	2	3	3
Task 11	3	3	3	3	3	3	2	3
Task 12	2	2	2	2	2	2	2	2
Task 13	8	7	8	8	8	7	7	7
Task 14	3	3	3	3	3	3	3	3
Task 15	6	6	6	5	6	6	6	5
Task 16	2	2	2	2	2	2	2	2

Table 5. Average results of the normalization of success and success rate

0,83	Walkthrough Question			
	1	**2**	**3**	**4**
Task 1	1,00	1,00	1,00	0,83
Task 2	1,00	1,00	1,00	1,00
Task 3	1,00	1,00	1,00	1,00
Task 4	1,00	0,83	1,00	1,00
Task 5	1,00	0,75	1,00	1,00
Task 6	1,00	1,00	1,00	1,00
Task 7	0,80	0,80	0,80	0,80
Task 8	1,00	0,75	1,00	0,83
Task 9	1,00	1,00	1,00	1,00
Task 10	0,83	0,50	0,66	1,00
Task 11	1,00	1,00	1,00	0,83
Task 12	1,00	1,00	1,00	1,00
Task 13	0,93	1,00	0,93	0,87
Task 14	1,00	1,00	1,00	1,00
Task 15	1,00	0,91	1,00	0,91
Task 16	1,00	1,00	1,00	1,00
Average	0,9725	0,9087	0,9618	0,9418
Succes rate (%)	97,25	90,87	96,18	94,18

4. Conclusion

The conclusion of this study is the needs and desires of respondents have been analyzed based on contextual inquiries that produce work activities, requirements documents, and task inventory. The design stage has resulted in user persona and ideation in the form of ideas, sketches and wireframes for all the main tasks, namely the task of 'getting information to prepare for mobility', the task of 'getting information and security while traveling', and the task of 'giving a review of a location 'based on the results of brainstorming. The results of the analysis and design are implemented as a medium-fidelity prototype. The prototype was tested using cognitive walkthrough (CW) techniques that produce success rates based on four walkthrough questions.

From the calculation results of the average normalization of the success of a task obtained success rate for the question 'Will the user try to achieve the right results?' Obtained a success rate of 97.25%, the question 'whether the user will see that the correct action is available for them?' Is 90.87 %, the question 'Will the user associate the correct action with the expected results?' by 96.18%, and the question 'If the correct action is taken, will the user see that progress is being made towards the expected results?' of 94.18%. Based on these results overall the user has understood and can run this Kuygo application correctly.

Based on these results, it can be concluded that no major usability errors were found in the prototype medium-fidelity and the Kuygo application can be further developed. However, the authors suggest that for the next prototype or product, further testing is needed by using the number of participants in the top two people which allows for greater reusability of major problems..

References

[1] Karyana A, Widiawati S. 2013. *Pendidikan Anak Berkebutuhan Khusus Tuna Daksa*. Jakarta (ID): PT Luxima Metro Media.

[2] Murdiyanti D. 2012. Aksesibilitas sarana prasarana transportasi yang ramah bagi penyandang disabilitas (TransJakarta) [skripsi]. Depok (ID): Universitas Indonesia.

[3] [KEMENKES] Kementrian Kesehatan, Republik Indonesia. 2014. Situasi penyandang disabilitas. *Bul Jendela Data dan Info Kesehatan*. 2:1-57.

[4] Nuansa AW. 2014. Kesetaraan hak pilih untuk penyandang disabilitas [internet]. [diunduh 2018 Des 19]. Tersedia pada: https://www.kompasiana.com/anwibisono/54f8022ba33311ea638b487f/kesetaraan-hak-pilih-untuk-penyandang-disabilitas-.

[5] Davis BW, Nattrass K, O'Brien S, Patronek G, MacCollin M. 2004. Assistance dog placement in the pediatric population: benefits, risks, and recommendations for future application. *Anthrozoös*. 17(2):130-145.

[6] Cowan RE, Fregly BJ, Boninger ML, Chan L, Rodgers MM, Reinkensmeyer DJ. 2012. Recent trends in assistive technology for mobility. *NeuroEngineering and Rehabilitation*. 9(20):1-8.

[7] Hartson R, Pyla PS. 2012. *The UX Book Process and Guidelines for Ensuring a Quality User Experience*. Waltham (US): Morgan Kaufmann.

[8] Engelberg D, Seffah A. 2002. A Framework for rapid mid-fidelity prototyping of web sites. *IFIP World Computer Congress*. 13:203-215.

[9] Polson PG, Lewis C, Rieman J, Wharton C. 1992. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*. 36(5):741-773.

[10] Mano A, Campos JC. 2006. Cognitive walkthroughs in the evaluation of user interfaces for children. Di dalam: Chambel T, Nunes NJ, Romão T, Campos JC, editor. *Conferência Nacional Em Interacção Pessoa-Máquina* ; 2006 Okt 16-18; Braga, Portugal. Porto (PT): Grupo Portugues de Computacao Grafica. hlm 195-198.

khazanah informatika
Jurnal Ilmu Komputer dan Informatika

Usability Testing on QR Code Scanner Application for Lecture Presence

Eli Pujastuti[*]**, Arif Dwi Laksito**
Informatics Department
Amikom University
Yogyakarta
[*]eli@amikom.ac.id

Abstract-Universities are obliged to facilitate lectures. Many students require universities to provide a fair teaching and learning services with a minimum of student cheating. In order to improve the quality of teaching and learning, each university develops a lecture presence system electronically. The QR code scanner application became a solution offered for leak problems that previously existed on the magnetic card system. Before the application applied on a large scale, developers needed to conduct an assessment of the usability of the QR scanner application. The assessment aimed to make lectures go smoothly and to maintain the good reputation of the university. The method used is usability testing. The result of this study is a usability system at the level of 65%. This value consists of an effectiveness value of 70%, an efficiency value of 54.31%, and a satisfaction value of 70.85%. The improvements of user interface recommended in this study include adding of placeholders to inform the correct NIM format, changing the QR scanner icon into a titled icon and choosing a stimulating color, providing a zoom feature on the scanner camera, and applying a more familiar logout icon according to the mental model of the user.

Keyword: Usability Testing, Usability, Presence Application, QR Code Scanner.

1. Introduction

Universities and colleges are required to facilitate lectures. Higher education institutions have a different lecture system, but still refer to the 2015 government regulation on Permenristekdikti 44 concerning SN Dikti Article 3 [1] where the University must meet the national standards of higher education to achieve the specified quality. The quality in the teaching and learning process can be achieved, one of which is the facilities and infrastructure that can be accessed by students and lecturers. In terms of teaching and learning, the facility that must be provided is a student presence system in the classroom. Presence is the evidence of the teaching and learning process that needs to be well-conducted.

This study engaged a private university as a case study, which is the University of AMIKOM Yogyakarta. AMIKOM University is a university that has 16 study programs and more than 10,000 active students. Many students demanded that AMIKOM provides fair teaching services with a minimum of student cheating. Hence, the presence system for students had already applied using an electronic system. As time goes by, the magnetic cards were considered less capable to ensure that students are actually present or only entrust their magnetic cards to others. This leak problem must be resolved, so that fair lectures could be conducted to generate quality college graduates. Therefore, the QR Code scanner application was developed to replace the magnetic cards. The QR code scanner applications required testing to ensure that the application could actually be used. Then, the usability testing was carried out to determine the usability level of the application. The testing was conducted to maintain the good reputation and sustainability of the university.

All actions in the human factor were considered to have their impact towards sustainability [2], [3], [4]. The usability testing was an important step in the preparation of scanner applications, because the success or failure of the implementation would greatly affect sustainability in the teaching and learning process at the university. The problem formulation of this research is "what is the usability level of the QR code scanner application for student attendance?". The purpose of this study is to provide some recommendations for improving the QR code scanner application for student attendance.

Research on usability testing was conducted by Shafrida, R et al [5]. This study applied cognitive walkthrough methods to test applications for blind people, namely B-Smart. This research focused on the interface testing. Shafrida's study used one testing method. The number of respondents from Shafrida's study were 30 respondents – who were potential users of the B-Smart application. The next research was a study conducted by Zuntriana, A [6]. This study examined the Indonesian national library web portal. The method used is the Remote Usability Test. The findings in this study indicate that there were difficulties in finding the service schedules. Another finding is the dissatisfaction of the user interface design and layout on the home page. Another study was conducted by Farouqi et al [7] who tested usability testing in online transportation applications in Indonesia. Farouqi used the SUS questionnaire as well as measuring effectiveness and efficiency. Nielsen [8] defines usability as a quality that assesses how easy the user interface is used. International Standards (ISO) [9] defines usability as effectiveness, efficiency, and user satisfaction in achieving certain goals in certain environments. Usability testing of computer systems for complex tasks must include measures of efficiency, effectiveness, and user satisfaction [10]. Usability testing refers to evaluating a product or service by testing it with a representative user. The aim is to identify usability problems, gather qualitative and quantitative data, and determine participant satisfaction with the product [11]. Usability is a trait that depends on interactions between users, products, tasks, and the environment [12] p. 1267. Usability is very important for all types of interfaces [13]. This helped engineers to design user interfaces [14]. This study refers to the usability handbook written by Rubin J, and Chisnell, D [15]. This book guides how usability researchers begin the research. One of the methods discussed in the book is usability testing implemented in this study.

2. Method

a. Usability Testing Method

According to the book "Handbook of Usability" [15], methods for evaluating usability can be done in various ways, one of which is Usability Testing. This method was held by observing the user. Researchers provided realistic tasks using both formal and informal approaches.

The usability testing process is explained in Figure 1.

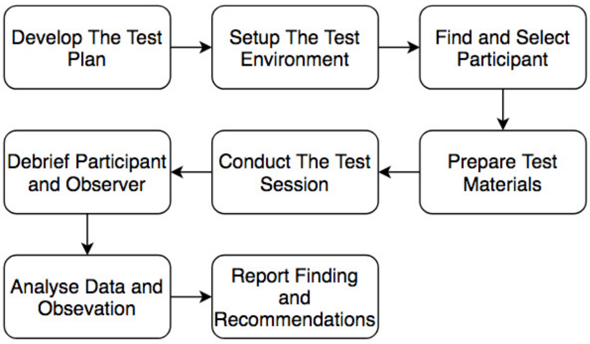

Develop The Test Plan
Setup The Test Environment
Find and Select Participant
Prepare Test Materials
Conduct The Test Session
Debrief Participant and Observer
Analyse Data and Obsevation
Report Finding and Recommendations

Figure 1. Usability Testing Process [15]

The test in Figure 1 consists of 8 steps, including:

1. Develop a test plan
   At this stage, the researcher planned an overall test of how, when, where, who, why.
2. Prepare a test environment
   At this stage, the researcher established a "usability-lab" to accommodate testing by respondents.
3. Find and select participants
   Participants in this study were determined at this stage.
4. Prepare test material
   This step contains the implementation of step 2. Step 4 aimed to reduce testing failures due to lack of equipment preparation in testing.
5. Running a test session
   This step is the beginning of testing.
6. Question and answer by participants and observers
   At this stage, participants began to be explored with questions from the observer.
7. Analysis and observation of data
   Data obtained at the question and answer stage, was followed by analysis to produce findings.
8. Report findings and recommendations
   Reports on the findings of the analysis were written and documented.

b. Sampling technique

There are 2 sample groups required in this study, including:

1) Qualitative sample group

This sample was taken from students who had never used the QR scanner application. This sample aimed to observe the mistakes made by participants from the level of success and efficiency. The sample required for this group was 5 people. The number of samples recommended by Nielsen is 3-5 people [16]. Students – who were participants – were asked to carry out the following three task scenarios:

- T01 - Task 1 [Login]
- T02 - Task 2 [Presence with QR Code Scanner]
- T03 - Task 3 [Logout]

There were three assessments of the success rate of the participants that became the measurements in this study, including:

- Success: the participants completed the task with the right steps.
- Delayed: the participants completed the task, but with steps that do not fit the scenario.
- Failed: the participant failed to complete the task.

Participants were arranged to sit in the front, middle and back area of the class. The first and second rows were the front area category, the third and fourth rows were the middle area category, and the fifth and sixth rows were the rear area category. Participants were observed by observers in terms of their level of success and efficiency.

2) Quantitative sample group

Quantitative sample groups in this study were used to measure satisfaction of users of the application. The sampling technique in this quantitative group is the Probability Sampling Technique. This technique was taken since the population is countable, so that the sample data could become a representation of population data. Participants in this study were students who were active and still doing lectures in the theoretical class. The list of students who took lectures in class can be seen in the Study Plan Card (KRS) which is taken in the even semester of the 2018/2019 academic year. Slovin formula was applied to determine the number of samples. The data obtained was used to measure application user satisfaction. However, the satisfaction itself could only be measured on users who had already used the application.

c. Data analysis method

The level of success, efficiency, and satisfaction were measured using the measurements conveyed by Nielsen [17]. The measurements to be analyzed were the success rate (1), efficiency (2), and satisfaction.

The calculation formula is shown as follows:

Success Rate = (Success + (Delayed Success x 0.5)) / (Total Tasks) x100% (1)

Efficiency was determined by the amount of time required to complete the given task [18]. The comparison of time between novice users and standard time to complete tasks became a percentage of efficiency. The efficiency was based on time, which can be calculated by the formula:

Efficiency = (Standard time / time spent completing assignments) x100% (2)

Satisfaction was calculated using the SUS (System Usability Scale) [19] formula, informed in the following steps:

1. Odd-numbered questions' score was reduced by 1.
2. Even-numbered questions' score was calculated by subtracting a value of 5 with the respondent's answer.
3. SUS score = the scores of each question were added up, then multiplied by 2.5.
4. The average of the score was calculated.

3. Results and Discussion

QR code scanner application could be applied in a theoretical class with QR code processing as follows:

1. Lecturer Login
   Each class was already provided a PC for a lecturer to login on the class presence system and projector.
2. Generating QR Code
   When the lecturer logged in, the system would generate a QR code.
3. Displaying the QR Code
   The projector would display the generated code.
4. Refreshing the QR Code Display
   The system would periodically update the displayed QR code in every 15 seconds.
5. Displaying students who had successfully attended.

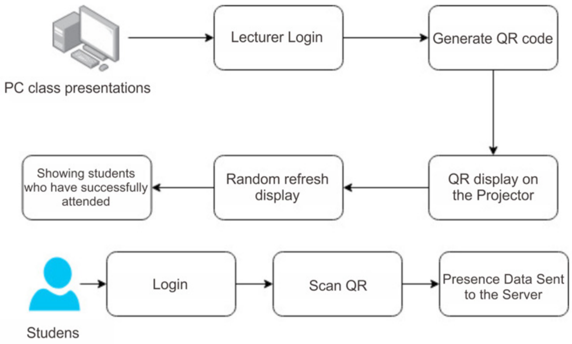

PC class presentations
Lecturer Login
Generate QR code
QR display on the Projector
Random refresh display
Showing students who have successfully attended
Studens
Login
Scan QR
Presence Data Sent to the Server

Figure 2. Process of QR Code Processing into Presence Data

Students could send the information regarding name, NIM (Student Identification Number), and scan time, by logging in to the QR code scanner application – installed in each student's smartphone. After that, the process of scanning the code on the projector screen by pointing the smartphone's camera to the screen was conducted. Finally, the data was sent to server via the internet.

Based on the process in Figure 2, the usability test can be started with the following steps:

a. Develop a test plan

The test plan was divided into several steps:

1) Testing Purpose
   The purpose of testing a QR code scanner application is to find out whether the application could be used by Amikom students for attendance (marked by success). Another purpose is to find out which part of the application became the point of the most errors, and then creating recommendations for improvement.
2) Research Questions
   What is the usability level of the QR code scanner application for lecture attendance?
3) Testing Method (Table 1)
4) Task List
- Login to the QR code scanner application
- Presence using the QR code scanner application
- Logout from QR code scanner application

5) Testing, Equipment and Logistics Environment
- Testing environment: Classroom
- Equipment: smartphone, internet connection, projector screen, presentation computer.

6) Role of Testing Moderators
- Introducing the testing objectives
- Giving assignments according to the assignment plan

7) Data collected and evaluation measurements
   The data collected is in the form of qualitative data and quantitative data. Qualitative data was obtained from interviews with participants; the goal was to explore the failure reasons of the application. Quantitative data was obtained from questionnaire results.

Table 1. Testing Method

	Description
Methodology	Usability in this study was measured from the ratio of success, efficiency, and satisfaction conducted by the participants during completing the task. The data collected was in the form of qualitative and quantitative data. Qualitative data was taken from direct observation to participants who had never used the application. The tasks given to the participants include login, try presence, and logout. Quantitative data was obtained from the questionnaire answers from 384 active students who were taking theoretical courses and had used the QR Code scanner application for the presence before.
Inter-Subject Plan	Each participant was given the task to do a presence in class using the presence application. Participants sat on a chair that was prearranged in a sequence from the front to the back side of the classroom. The order of the seat shows whether the distance determines presence failure or not. The participant was asked to carry out the task, when the participant had completed the tasks given by the observer. The form was filled in by the observer, containing how long the participant completed the task, what failures were encountered, and what made it difficult for the participants to complete the tasks.
Time	The total time required is 25 minutes. The introduction session was held in 5 minutes, the task implementation was conducted in 10 minutes, and the interview session upon the test completion was carried out in 10 minutes. The test is located in the regular classroom at Amikom University in Yogyakarta.
Introduction (5 minutes)	Introduction to observer and explanation of the observation activities. Filling in participant's personal data.
Execution of tasks (10 minutes)	The participants began to work on the tasks by using a QR code scanner for class presence.
Interview after test (10 minutes)	Asking participants regarding the problems they encountered during the presence.

Table 2. Layout Usability Lab

	QR Code Screen
Front - Left	Front - Right
Middle - Left	Middle - Right
Rear - Left	Rear - Right

8) Report content and presentation
The report contains a Google form spreadsheet and photo documentation.

9) Set up the test environment
At this stage, the researcher arranged "usability-lab" to accommodate the testing performed by the participants (Table 2).

b. Participant Selection

1) Qualitative Group

There were 5 participants who became respondents in this study. Participants 1-5 were observed in a class. Each participant represents the front area of the right and left sides, the middle of the right and left sides, and the back of the right side.

Table 3. Seating Design

Name of Participant	Seat
Participant 1 [Par 1]	Front right side
Participant 2 [Par 2]	Front left side
Participant 3 [Par 3]	Middle right side
Participant 4 [Par 4]	Middle left side
Participant 5 [Par 5]	Back right side

2) Quantitative Group

According to data taken from the Directorate of Innovation Center in the even semester of 2018/2019, there were 9,659 students who were active, were taking theoretical courses, and had used the QR code scanner application. The number of students represents the total population in this study.

The sampling technique employed is simple random sampling method. This method was conducted by determining the number of samples using the Slovin formula. The simple random sampling method was preferred since the population is large and each member of the population has the same opportunity to be sampled. The number of samples taken was based on the Slovin formula, explained as follows:

n= 9659/(9659*(0,05)^2+1)
n= 9659/(24,1475+1)
n= 384,093846

The number of samples used was 384.093846, rounded up to 384 students.

c. Test Materials Preparation

This step contains the implementation of step 2. This step aimed to reduce the failure of the test due to lack of equipment preparation in the test.

The materials prepared for the testing session includes:

- Projector Screen
- Monitoring Form
- Internet connection

d. Running a Test Session

This step is the initial step of the test session. The testing session was carried out in parallel between success, efficiency, and satisfaction. The testing session is divided into 2 groups, including the success and efficiency testing as well as the satisfaction testing.

The success and efficiency testing was held in class with 5 participants and 1 observer. The testing process of the 5 participants was not conducted simultaneously, but one at a time. First, the participants were given a briefing regarding the tasks that need to be completed by the participants. Participants did not receive an explanation or clue related to completing the given tasks.

The satisfaction testing was conducted online by filling out questionnaires that are accessible for 10 days – to meet the number of samples, 384 respondents.

e. Participants and Observer Questions and Answers

At this stage, participants began to be explored with questions from the observer. After the participant completed the task, the observer conducted an interview with the participant. The interview questions are mentioned as follows:

1) Did you successfully login to the application?
2) Were there any difficulties experienced in the login task?
3) Did you successfully use QR code?
4) Were there any difficulties in assigning attendance?
5) Did you successfully logout from the application?
6) Were there any problems experienced in the logout task?

f. Analyzing Data and Observation

1) Success

Table 4. Task Success Ratios - Login

Task [TO1]	Par 1	Par 2	Par 3	Par 4	Par 5
Login	D	S	S	S	S

Success= (4+ (1x0.5))/5 x100%
= 90%

The data obtained in Task 1 is that participants were asked to login to the QR code scanner application. Participant 1 has a status of D or Delayed. It means that participant 1 successfully completed the task but did the wrong steps. An error occurred when participant 1 tried to enter a NIM that was supposed to apply a dot. Participant 1 did not use a dot, so that participant 1 was unable to enter the application successfully. The application did not inform the user regarding the location of the error. However, participant 1 tried to add a dot to the NIM (Student Identification Number) text field and finally succeeded.

Table 5. Task 2 Presence Success with QR Code

Task [TO2]	Par 1	Par 2	Par 3	Par 4	Par 5
Scan the Code	S	D	D	D	F

Success = (1+ (3x0.5))/5 x100%
= 50%

Participants 2, 3, and 4 are categorized as Delayed due to the same mistake, i.e. participants mistakenly tapped on the "presence" icon that contains the attendance history. Participants thought that the icon was an icon for attendance using QR scanner. After realizing that the participant was wrong, the participant returned and then chose the QR code scanner icon and finally succeeded. Participant 5 failed because participant 5 could not scan the QR code from the point where the participant sat. Participant 5 sat at the back-right side of the classroom. Participant 5 tried to enlarge the camera, and expected that the application has a zoom feature. The success rate for this task was only 50% because there were 3 delayed and 1 failed.

Table 6. Task 3 Success Ratios - Exit

Task [TO2]	Par 1	Par 2	Par 3	Par 4	Par 5
Logout	D	S	D	S	D

Success = (2+ (3x0.5))/5 x100%
= 70%

Participants 2 and 4 conducted the logout process without any problems. Participant 1 was delayed because participant 1 thought that the logout button was on the profile menu, so that participant 1 chose the profile photo icon. Participant 3 lack knowledge regarding which icon to choose to exit the application. Participant 3 tried to logout and finally succeeded. Participant 5 tapped the code icon when being asked to exit. The error that occurred in participants 1, 3, and 5 is basically due to the lack of information about the logout icon. An uncommon icon is used to log out from the application. This application uses icons that are less familiar to users.

Table 7. Average Success

Task Scenarios	Success Ratio (%)
Task 1	90
Task 2	50
Task 3	70

Average success = (90+50+70)/3 = 70%

According to Table 7, the lowest success rate is task 2 with the success rate of only 50%. Assignment 2 is the main feature, where students were able to scan QR codes for their presence. There were two mistakes that were often made including a mistake of placing the presence history in the presence menu. The next fatal error is that the QR code could not be scanned from the seat at the back of the classroom.

2) Efficiency

Table 8. Task 1 Time Efficiency

	Par 1	Par 2	Par 3	Par 4	Par 5
Success	1	1	1	1	1
Time	22	15	11	21	13

Standard time = 11 second

The longest time needed to complete task 1 is 22 seconds and the fastest is participant 3, completed in 11 seconds. The standard time is the time taken by people who have already used the application.

Table 9. Task 2 Time Efficiency - Presence with QR Codes

	Par 1	Par 2	Par 3	Par 4	Par 5
Success	1	1	1	1	0
Time	34	15	29	40	18

Standard time = 15 second

The fastest time to complete a task is 15 seconds, performed by participant 2. The longest time is 34 seconds, carried out by participant 1. The default time is 15 seconds.

Table 10. Task 3 Time Efficiency - Logout

	Par 1	Par 2	Par 3	Par 4	Par 5
Success	1	1	1	1	1
Time	20	15	4	3	7

Standard time = 3 second

Logout was able to be completed in 3 seconds, if the icon in the application was already known. The most time required by participant 1 is 20 seconds because participant 1 did not recognize the logout icon.

The average overall time to complete task 1 to task 3 for novice users is 52.4 seconds. Whereas, the standard time is 29 seconds or less.

Table 11. Average Time for Completing Tasks

Task	New user	Standard Time
Task 1	16.4	11
Task 2	27.2	15
Task 3	9.8	3
Average	53.4	29

The time comparison between novice users' completion time and standard time, produces calculations as follows:

Total Time Efficiency = 29/(53.4) x100%
= 54.31%

The application is stated to be efficient when the value is close to 100%. In this application, the efficiency value is 54.31% – obtained because the participant undertook the wrong steps in selecting the menu for completing the task.

3) Satisfaction

Table 12. SUS Standard Values

Score	Grades
>81	A
68-81	B
68	C
51-67	D
<51	E

Table 13. Recapitulation of Respondents

Value	Number of Respondents
A	87
B	133
C	0
D	137
E	27

Most of the data belongs to class D, which is 137 respondents. Class D contains respondents whose score between 51-67. The number of respondents and the scores of respondents are informed in the following table. The individual values obtained in table 13 were then calculated altogether, resulting in a SUS score of 70.85. The score implies OK/Fair satisfaction level, which means good but required an improvement to become Excellent.

Table 14. Grading SUS Key

Score	Grade
92	Best imaginable
85	Excellent
72	Good
52	OK/Fair
38	Poor
25	Worst imaginable

SUS score = 70.85

Usability Value calculation results:

Usability = (70+54.31+70.85)/3
= 65%

g. Report Findings and Recommendations

1) Findings

The findings gained from this study were some errors that arose due to the design of the user interface that did not sufficiently depict the function of each icon. Some error findings are informed in Table 15.

Table 15. Table of Errors

No.	Task	Error	Cause
1	Login	Error occurred when user entered NIM (Student Identification Number)	The user lack knowledge that inputting NIM must apply dots to separate numbers
2	Presence	The user chose a larger icon and can attract more attention.	The user thought the icon is an icon to start scanning the QR code
3	Presence	The user failed to scan the QR code because the seat is located at the back side of the classroom	The camera could not reach the code, so that it could not read the code properly
4	Logout	The user was trying around by selecting the profile icon	The user thought logout is situated in the profile menu; the logout icon is not familiar.

2) Recommendation

The recommendations of the study findings were produced according to errors found in this study. The recommendations are informed in Table 16.

Table 16. Table of Recommendations

No.	Error	Recommendation
1.	Error when user entered NIM (Student Identification Number)	Placeholder for informing the correct format to the user
2.	The user chose a larger icon and able to attract more attention.	Changing the QR scanner icon with a titled icon and change the color to a stimulating color
3.	Error during scanning QR code	Providing a zoom feature on the scanner camera
4.	Error during logging out	Using a logout icon that is more familiar according to the user's mental model

4. Conclusion

Usability testing showed the usability level of the QR Code scanner results of this study is 65% – with effectiveness at the level of 70%, efficiency at the level of 54.31% and satisfaction at the level of 70.85%. There were four mistakes that were often made by novice users. The first mistake occurred when user entered NIM (Student Identity Number). The second mistake occurred when user trying to scan a QR code, but the user chose a more attracting icon. The third mistake occurred when user tried to scan a QR code that was too far from the seat. The fourth mistake was a mistake in selecting the profile icon when the user tried to exit. These errors resulted in inefficient use of time. The recommendations of developing a QR code scanner application in lectures include adding a placeholder to inform the correct format of NIM, changing the QR scanner icon to a titled icon and choosing a stimulating color, providing a zoom feature on the scanner camera, and using a logout icon that is more familiar according to user's mental model. The future studies are expected to provide comparative results after some interface improvements based on the results of this study.

Acknowledgement

Thanks to AMIKOM Yogyakarta University.

References

[1] KEMENRISTEKDIKTI, "Keputusan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia," *No 257/M/Kpt/2017*, vol., no., p. , 2017.

[2] N. A. Stanton and R. B. Stammers, "Bartlett and the future of ergonomics," *Ergonomics*, vol. 51, no. 1, pp. 1–13, 2008.

[3] R. Haslam and P. Waterson, "Ergonomics and Sustainability," *Ergonomics*, vol. 56, no. 3, pp. 343–347, 2013.

[4] V. G. Duffy, "Erratum to: Chapters 41 and 55 in," no. December, p. 8519, 2016.

[5] R. S. Kurnia, E. Utami, and H. Al Fatta, "Pengujian Usability Antarmuka Aplikasi Braille Smart pada

Siswa Tunanetra," *J. Inf. Interaktif Univ. Janabadra*, vol. 2, no. 1, pp. 21–28, 2017.

[6] A. Zuntriana, P. Universitas, I. Negeri, M. Malik, and I. Malang, "Uji Usabilitas Jarak Jauh (Remote Usability Testing) pada Portal Web Perpustakaan Nasional Republik Indonesia Remote Usability Testing in Portal Web Perpustakaan Nasional Republik Indonesia," vol. 1, no. 1, pp. 68–76, 2015.

[7] M. I. Farouqi, I. Aknuranda, and A. D. Herlambang, "Evaluasi Usability pada Aplikasi Go-Jek Dengan Menggunakan Metode Pengujian Usability," vol. 2, no. 9, pp. 3110–3117, 2018.

[8] J. Nielsen, "Usability 101: Introduction to Usability." [Online]. Available: https://www.nngroup.com/articles/usability-101-introduction-to-usability/. [Accessed: 03-Aug-2019].

[9] I. Standard, "INTERNATIONAL STANDARD," vol. 2010, 2010.

[10] E. FrØkjaer, M. Hertzum, and K. Hornbæk, "Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated?," *Conf. Hum. Factors Comput. Syst. - Proc.*, no. January, pp. 345–352, 2000.

[11] A. S. for P. Affairs, "Usability Testing," Nov. 2013.

[12] G. Salvendy, *Handbook of Human Factors and Ergonomics: Fourth Edition*. 2012.

[13] F. Paz, F. A. Paz, D. Villanueva, and J. A. Pow-Sang, "Heuristic Evaluation as a Complement to Usability Testing: A Case Study in WEB Domain," *Proc. - 12th Int. Conf. Inf. Technol. New Gener. ITNG 2015*, pp. 546–551, 2015.

[14] R. Kaur and B. Sharma, "Comparative Study for Evaluating the Usability of Web Based Applications," *Proc. - 4th Int. Conf. Comput. Sci. ICCS 2018*, pp. 94–97, 2019.

[15] D. Chisnell, *Handbook of Usability Testing How to Plan , Design , and Conduct Effective Tests.* .

[16] Jakob Nielsen, "Success Rate: The Simplest Usability Metric," 2001. [Online]. Available: https://www.nngroup.com/articles/success-rate-the-simplest-usability-metric/. [Accessed: 27-Jul-2019].

[17] J. Nielsen, "Usability Metrics." [Online]. Available: https://www.nngroup.com/articles/usability-metrics/. [Accessed: 05-Aug-2019].

[18] D. Alfageh and F. Demir, "The Usability Evaluation of a Digital Book Application for Elementary School The Usability Evaluation of a Digital Book Application for Elementary School Students," vol. 3, no. November, pp. 27–33, 2018.

[19] J. Brooke, "SUS - A quick and dirty usability scale," Earley, READING RG6 2UX United Kingdom, 1986.

khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Drought Analysis and Forecast Using Landsat-8 Sattelite Imagery, Standardized Precipitation Index and Time Series

Musa Marsel Maipauw*, Eko Sediyono, Sri Yulianto Joko P
Graduate Study in Information Systems, Faculty of Information Technology
Universitas Kristen Satya Wacana
Salatiga 50711, Jawa Tengah
*marselmaipauw@yahoo.com

Abstract-A drought is a phenomenon of shortages in water supply in an area for a long time. Drought usually occurs in areas that has little rain for a long time or in areas with low precipitation. Drought have negative impacts on many sectors such as agriculture, plantations, water resources and environment. This paper describes the results of a research that aims to analyze data to get the level of drought during four yearly periods, and predict the likelihood of drought to occur in the future. The level of drought was analyzed using the Inverse Distance Weighted (IDW) method and the Standardized Precipitation Index (SPI). Least square time series was utilized to forecast the level of drought in the near future. Data consists of drought data collected from electronic news, rainfall data from BMKG, and anual Landsat-8 satellite imagery. All data are for Western Southeast Mallucas in the range of 2015-2018. Analysis using IDW and SPI methods produce similar interpretation for year 2015, i.e. mild dryness, and fro year 2018, i.e. no drought. However, the two methods show discrepancy in analysis of data for 2016 and 2017. The use of least square time series to forecast drought in 2019 gives SPI value of 0.03 which intepretes as normal weather (no drought) that is consistent with the result of field observation.

Keywords: Drought, inverse distance weighted, standard precipitation index, least square, time series

1. Introduction

A drought is a phenomenon of shortages in water supply in an area for a long time for example in the range of months or even years. Drought usually occurs in areas that has little rain for a long time or in areas with low precipitation. Drought often occurs in many part of the world, including regions in Indonesia. Indonesia's geographical position between two continents and two oceans and its position on the equator are factors which contribute to dynamic weather that may cause flooding and drought. The archipelago has a tropical monsoon climate that is very sensitive to the El Nino Southern Oscillation (ENSO) anomaly [1], [2]. Nugroho (2018), Head of the BNPB Data and Public Relations Data Center, stated that an area that has once experienced a drought had a good chance to experience similar disaster in the upcoming years [3].

Drought may be considered as an imbalance between water need and water supply that the nature can be provide. Drought gives negative impact in various sectors such as agriculture, animal husbandry, plantations and forestry because the absence of water inhibits the growth of plants and slows down biological metabolism of humans and animals. A such, water availability affects economic productivity in the sectors.

Drought is a calamity to anticipate. It is very difficult, if possible, to avoid being exposed to drought and if that happens, people will suffer from many undesirable impacts. However, people can make preparations so that the effect of a drought is minimized. Such preparations will be effective if we can predict as when the disaster will take place. Efforts to forecast the occurrence of drought has become possible through the utilization of methods of remote sensing which work on satellite imagery data. Satellites are celestial bodies, such as the moon, that revolve in an orbit surrounding planets. Since the 60s, people have made artificial satellites, which are spacecrafts that orbit the earth, to meet many purposes such as telecommunication and the acquisition of surface images of the earth [4]. Earth surface imagery are useful to analyze the shape of the earth and the changes of its surface and objects thereon. From satellite imagery, people may reveal the occurrence of forest fire, flood, landslide, and drought.

In addition to satellite imagery, drought forecast may be conducted based on rainfall data. A popular drought calculation method is based on rainfall, which is known as the Standardized Precipitation Index (SPI). According to Adidarma et al., calculation of SPI is based on the amount of monthly rainfall and is widely used in the world [5]. Lincoln Declaration states that there has been an increase in the frequency and level of drought so that standard benchmarks are needed to monitor the disaster. These benchmarks are important because they are used together in monitoring and managing climate risks around the globe. Based on the SPI benchmarks, drought is declared to start if SPI is below zero (negative) and the drought ends when SPI shifts to positive [6], [7].

This paper describes the results of a research that observes methods to analyze drought. The methods used include the Inverse Distance Weigted which processes remote sensing data, and least square time series which processes annual Standard Precipitation Index (SPI) data. The data processed is climate data for the Western Southeast Mollucan regency, in the province of South Maluku.

The Western Southeast Mollucan regency (MTB) often suffers from drought. The region is a tropical area close to the Pacific ocean. Generally, the weather in Indonesia is very sensitive to a climate anomaly called El Nino Southern Oscillation (ENSO). ENSO will cause low rainfall in Indonesia when the surface temperatures of the equatorial Pacific ocean in the middle area to the eastern part warms up [2]. According to the Head of the Center for Data, Information and Public Relations of the National Disaster Management Agency (BNPB), ENSO may cause drought in more than 4000 villages in Indonesia affecting around 4.87 million people [8]. The perceived impacts include the lack of clean water and the decline of food production. MTB regency is one of the areas that suffers the hardest hit by the ENSO climate anomaly. Drought may persists over a long period of time resulting in crop failure for maize, rice and tubers [9].

2. Literature Review

We observe several studies that have been carried out related to drought disasters. One of them was carried out by [10] who observed remote sensing for monitoring drought in paddy fields. They observed vegetation index (VI) of paddy fields in East Java and Bali during July to December 2011. The level of greenness of vegetation during eleven 8-day periods showed a domination of low VI. Low rainfall during July 2011 caused a shortage of water supply to paddy fields in several districts, especially during August and September. A low level VI indicates that the observed area suffers a drought.

Another study has examined the relationship between vegetation index and spatial position. The results of analysis suggest that drought phenomenon has spatial connectivity among the observed regions. K-means analysis shows that under the high vegetation density, the weight of the distance between vegetation data points and the centroid is shorter. That is, data are concentrated in an area. Under conditions of low vegetation density, the weight of the distance between the data points to the centroid becomes wider. Under such condition, the data looks more distributed. The dominance of the green data group shows the dominance of the grid that marks the vegetation. The dominance of brown and red data groups shows the dominance of the grid marking the non-vegetation surface area or low vegetation growth, which indicates drought [11].

Subsequent research looked at the use of Theory of Run in the Krueng watershed, Aceh. This research yields monthly rainfall calculations for various monitoring points in Aceh province [5].

3. Research Methods

a. Research stages

This research runs in several stages as shown in the flowchart in Figure 1. Satellite imagery on the flowchart is input data or observations from the remote sensing process. Remote sensing is usually defined as the science of obtaining information about an object, area or phenomenon through the analysis of data obtained through a device without direct contact with the object, area, or phenomenon [12].

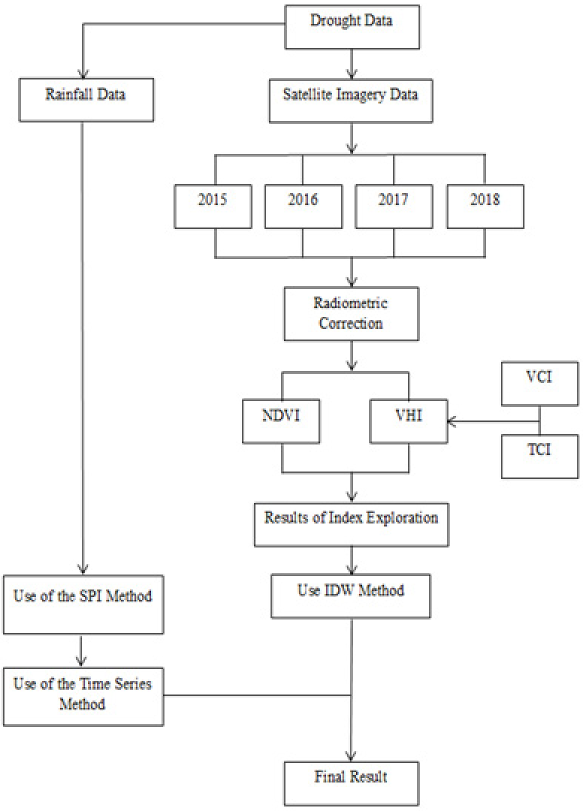


Figure 1. Research Flow; see text for description of acronyms

Normalized Difference Vegetation Index (NDVI) is a formula that is widely used to predict surface properties when the vegetation canopy is not too tight and not too rare. NDVI calculations are shown in equation (1) where

NIR is the spectral value of the near infrared channel, and RED is the spectral value of the Red channel.

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{1}$$

Vegetation Health Index (VHI) is an index that describes the health of vegetation regardless of the cause. Poor vegetation health can be caused by stress due to drought, flooding, or pest attacks [5]. The VHI calculation is shown in equation (2) where VCI is the vegetation condition index, TCI is the temperature condition index, and parameter a is the contribution coefficient of the two indices [13].

$$VHI = a \times VCI + (1 - a) \times TCI \tag{2}$$

Phase 1 is the data collection stage which includes data on drought events, both electronic news sites and BMKG, rainfall data at the location studied and satellite imagery data over an interval of four periods (2015-2018).

Phase 2 is the initial data processing stage which is the analysis phase using several formulations. At this stage, satellite image data is corrected, then processed using the index exploration processing method. The methods referred to are the Normalized Difference Vegetation Index (NDVI) and Vegetation Health Index (VHI).

Phase 3 is a drought analysis using the Inverse Distance Weighted (IDW) method which is applied to the results of index exploration processing. At this stage the SPI method is also applied to the rainfall data to get the standardized precipitation index. SPI calculation results are used to predict drought in 2019 using the least square time series method.

Phase 4 is the analysis of the results of the calculation of the drought prediction as the output of the two methods.

b. Drought Analysis Method

1) Inverse Distance Weighted

Inverse Distance Weighted (IDW) is a simple deterministic method by considering the points around it. Points closer to the estimated location will be given more weight than points further away. This weighting is the origin of the name inverse distance weighted [14]. Two points that are close together are said to be more similar than two points that are far apart.

The IDW method uses the centroid point which is the estimated sample point in each district. One point represents one weighting value for each district. In this study there were six centroid points according to the number of sub-districts in the area observed. The IDW formula is shown in equations (3) and (4) [15].

$$u(x) = \sum_{i=0}^{N} \frac{w_i(x) u_i}{\sum_{j=0}^{N} w_j(x)} \tag{3}$$

$$w_i(x) = \frac{1}{d(x, x_1)^p} \tag{4}$$

In both of these equations, $u_i = u(x_i)$, for i from 0 to N, x is an interpolated point, x_i is a known point, d is the distance of point x with respect to x_i, N is the number of points, and p is power which is a positive real number.

2) Standardized Precipitation Index

The Standardized Precipitation Index (SPI) was developed in 1993 by McKee. The aim is to determine and monitor drought [16]. SPI is calculated as follows [17]. First determine the α and β values estimated for each rain station monitoring point using equations (5) - (7).

$$\alpha = \frac{1}{4A}\left(1 + \sqrt{1 + \frac{4A}{3}}\right) \tag{5}$$

$$A = \frac{\ln(x) - \sum \ln(x)}{n} \tag{6}$$

$$\beta = \frac{x}{\alpha} \tag{7}$$

In equations (5), (6) and (7), ln is the natural logarithm, x the amount of rainfall and n the amount of rainfall observation data.

SPI value calculation is based on the gamma distribution which is defined as a function of frequency or chance of occurrence. For this reason, it is necessary to determine the value of H (x) following equation (8).

$$H(x) = q + (1 - q).\, G(x) \tag{8}$$

In equation (8), q is the number of rainfall events, 0 / n where n is the number of years of rainfall data, G is the gamma distribution, and x can be filled with *x*, α, or β.

If 0 <H (x) ≤ 0.5, the calculation of the SPI value follows equations (9) and (10).

$$SPI = -\left(t - \frac{c_0 + c_1 + c_2 t^2}{1 + d_1 + d_2 t^2 + d_3 t^3}\right) \tag{9}$$

$$t = \sqrt{In \frac{1}{(H(x))^2}} \tag{10}$$

Calculation of the SPI value for 0.5 <H (x) ≤ 1.0 follows equations (11) and (12).

$$SPI = \left(t - \frac{c_0 + c_1 + c_2 t^2}{1 + d_1 + d_2 t^2 + d_3 t^3}\right) \tag{11}$$

$$t = \sqrt{In \frac{1}{1 - (H(x))^2}} \tag{12}$$

The coefficients in equations (9) and (11) have the following values.

c_0 : 2.515517
c_1 : 0.802853

c_2 : 0.010328
d_0 : 1.432788
d_1 : 0.189269
d_2 : 0.001308

SPI values indicate the value of rainfall compared to average rainfall. A positive SPI value indicates rainfall at a time greater than average rainfall. Conversely, a negative SPI value indicates rainfall at a time smaller than average rainfall [16].

3) Time Series

Time series is a collection of data collected in a time span, or data whose independent variable is time. Time series data is usually used to forecast (value) the value of the dependent variable for future events. Forecasting is done on the basis of scientific methods (science and technology) and carried out systematically [18].

Least Square is a method for determining relationships between variables, in the form of the most appropriate mathematical equation. This method minimizes the number of errors between the data points and the equation curve points. Least Square produces the most appropriate trend line (best fit) from time series data [19], [20]. Mathematically, the trend line can be expressed with equation (13).

$$Y = a + bx \tag{13}$$

In equation (13), Y is the value of the forecast result, a basic period, b the rate of development of the predicted value, x the time calculated, and n the number of years predicted. Coefficients a and b are calculated based on past drought data by following equations (14) and (15) [20].

$$a = \frac{\sum y}{n} \tag{14}$$

$$b = \frac{\sum xy}{\sum x^2} \tag{15}$$

c. Research sites

This study takes data on drought and rainfall for Western Southeast Mollucan (MTB) districts, namely in the districts of Nirunmas, Kormomolin, Wer Tamrian, and South Tanimbar. Geographically, MTB is located at coordinates 6o34’24” - 8o24’36” South Latitude and 130o37’47” - 133o4’12” East Longitude. The total area is 52,995.20 km^2 which consists of land area of 10,102.92 km^2 (19.06%) and water area of 42,892.28 km^2 (80.94%).

4. Results And Discussion

Before calculating the exploration index, the researcher first carries out radiometric correction to correct the pixel values to match what they should be. Correction is needed to reduce image errors due to atmospheric factors which are the main source of errors. Radiometric correction is done by converting a number value into a reflectance value.

a. Exploration Index Calculation

The radiometric corrected image was processed using two exploration index methods, namely NDVI and VHI. Figure 2 shows the results of the NDVI (green) exploration index calculation and Figure 3 shows the results of the VHI (drought) calculation in the Western Southeast Mollucan (MTB) region. The four sub-districts of Nirunmas, Kormomolin, Wer Tamrian and Tanimbar Selatan are shown in a row in a row from north to south.

The final results of the calculation of vegetation index in the form of numbers are presented in table 1. The first row of the table contains the final results of calculating the index with NDVI (greenness) and the second row contains the final results of calculations with VHI (drought). The results of calculations are carried out on satellite imagery for a period of four years, namely in the range of 2015-2018.

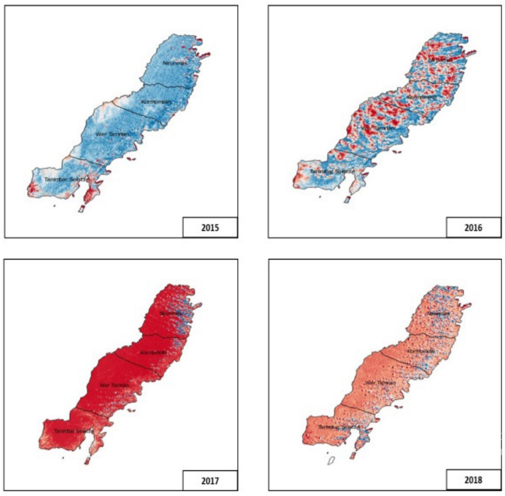


Figure 2. NDVI image during 2015-2018

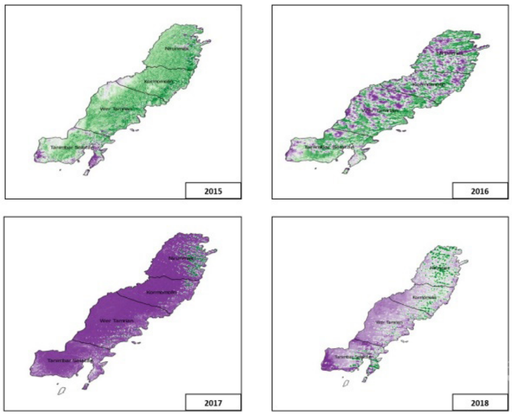


Figure 3. VHI image during 2015-2018

Table 1. Calculation result of exploration Indices

Year	2015	2016	2017	2018
NDVI	0.40	0.31	10.31	41.62
VHI	26.46	25.16	26.58	51.08

The NDVI index in 2015 and 2016 shows that vegetation is rare in MTB. Whereas the index in 2017 and 2018 shows dense vegetation. Meanwhile, the VHI index in 2015, 2016 and 2017 shows the level of severe drought occurring in the area under study. But in 2016 there was a decrease in the VHI index by 1.3, which means an increase in drought, which can be attributed to the El Nino phenomenon in 2015. In 2017, the VHI index value was higher than the previous two years, which means relatively low drought. Whereas in 2018 the VHI value is very high which indicates no drought in MTB.

The exploration index calculation for the MTB district shows that this area had a severe level of drought in 2015-2017. The NDVI index shows rare vegetation which can be interpreted as a possibility of drought. The NDVI index for 2018 shows that dense vegetation means that there is no drought. This indication is in accordance with the results of the VHI index calculation which states that in 2018 there will be no drought.

b. Drought Analysis

Drought analysis in this study was carried out using two methods namely the Distance Weighted Index (IDW) and the Standardized Precipitation Index (SPI). IDW calculation requires a centroid point value which contains a weighting value, which is a condition of applying IDW calculation in the QGIS application. Calculations using the IDW method in this study use the results of the VHI index calculation to examine the pattern of drought that occurs in each district in the area studied. The IDW calculation results are displayed in Figures 4 and 5.

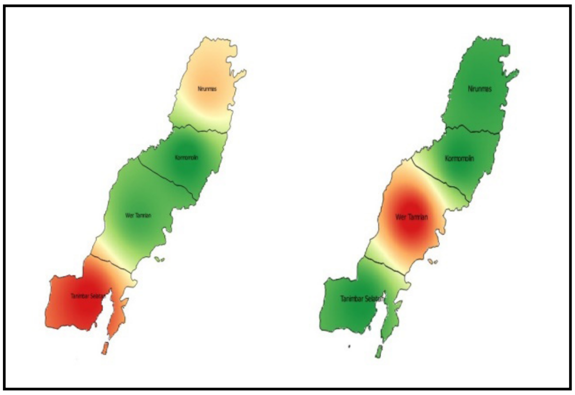


Figure 4. IDW calculation for year 2015 (left) and 2016 (right)

Figure 4 is the result of processing with the IDW method for data in 2015 and 2016. In 2015, Nirunmas sub-district (located the northernmost on the map) had moderate to mild drought levels dominated by mild drought levels. In that year the Kormomolin sub-district experienced a mild level of drought in the range of 30-40, while the Wer Tamrian sub-district experienced a moderate to mild drought in the range of 20-30. While South Tanimbar sub-district has a drought level of 10-30 which is included in the category of moderate-severe drought dominated by severe drought.

In 2016, all Nirunmas sub-districts experienced a moderate level of drought with an IDW value of 25.4. Kormomolin District has moderate to severe drought. While Wer Tamrian sub-district has a level of drought less than 10-20 which is included in the category of severe-extreme drought dominated by severe drought. Lastly, South Tanimbar district has moderate to severe drought.

Figure 5 shows the results of IDW calculations for 2017 and 2018. In 2017, Nirunmas sub-district did not experience drought even though there were some regions that experienced mild to moderate drought. Kormomolin District has mild to severe drought levels while Wer Tamrian and South Tanimbar districts have severe drought levels in the range of 10-20. While in 2018, the IDW calculation results for the four districts produce 45.5-69.3 which means that it is included in the category of no drought. The results of IDW interpretation for the four sub-districts in the MTB district over a span of four years can be seen in table 2.

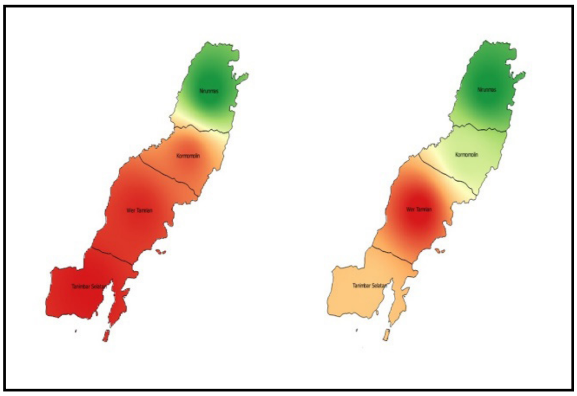


Figure 5. IDW calculation for year 2017 (left) and 2018 (right)

Table 2. Drought Analysis by IDW Method

Year	Drought Level
2015	Mild-Severe
2016	Moderate-Extreme
2017	Mild-Severe
2018	No Dryness

Table 3 shows the results of applying the SPI method to rainfall data in MTB. The MTB Regency only has one rain monitoring post so the available data is considered to represent one district data, and there is no data available per district. The data in the table states that the MTB district experienced drought only in 2015. In 2016 and 2017 there was no drought in the area, moreover in 2018 MTB experienced a wetter climate.

Table 3. Drought Analysis by SPI Method

Year	SPI Value	Conditions
2015	-1.56	Very Dry
2016	0.33	Normal
2017	0.25	Normal
2018	1.15	Quite Wet

The results of the IDW and SPI analysis show results that are not entirely uniform. Both methods produce similar analyzes for the 2015 and 2018 data. However, both methods provide different analysis results for the 2016 and 2017. data. Table 4 shows a comparison of interpretations of the level of drought produced by the two methods..

Table 4. Drought Level Comparison

Year	Drought level		Level Similarity
	IDW	SPI	
2015	Mild-Severe	Very Dry	Similar
2016	Moderate-Extreme	Normal	Different
2017	Mild-Severe	Normal	Different
2018	No Dryness	Quite Wet	Similar

For 2015, both methods provide the results of calculations that are interpreted as a occurrence of drought. Whereas in 2018, the results of the calculation of the two methods are interpreted in the same way that there is no drought. However, analysis results for 2016-2017 show different results between the IDW and SPI methods. According to the IDW method, in the range of 2016-2017 there was a drought in MTB while the results of the calculation of the SPI method stated there was no drought [11].

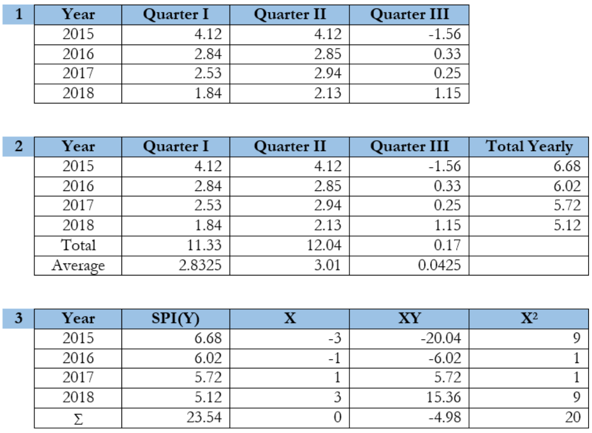
The difference in the results of the analysis of the two methods allegedly caused by differences in the sample of data used in the analysis. IDW method uses satellite imagery so that the data obtained are in the form of two-dimensional data consisting of many monitoring points. Meanwhile, the SPI method uses rainfall data at one point in the district of MTB so that it may not necessarily represent events in land areas covering more than 10 thousand square kilometers.

c. Drought Forecast

Forecasting drought is beneficial for the government and residents to anticipate the weather with activities that can be carried out in these weather conditions. Because the data obtained from drought analysis is time-based data, forecasting the level of drought in the future can be done using the time series method. One method that is quite simple to apply is least square, which models the weather with a trend line.

This study uses the least square time series approach to predict the level of drought in 2019 based on drought data in the range of 2015-2018. Annual data is refined into quarter data so that 12 drought levels are obtained over a four-year period (see Figure 6 Step 1). The least square time series calculation is done with the R-Studio and Excel applications. The step taken is to calculate the average value of each quarter and the total value per year (see figure 6 step 2). Then the values of Y and X (figure 6 step 3) are determined.

Least square time series is applied in the form of determining the trend line following equation (13) where the coefficient values a and b are calculated using equations (14) and (15). The calculation process appears in Figure 7 step 4. Using the trend line, the value of the drought level in 2019 can be predicted (see Figure 7 step 5). Furthermore, the predicted value for each quarter for 2019 is done by looking at the average index for each quarter for 2015-2018 (figure 7 step 6) where it is assumed that the value for each quarter in 2019 will be the same as the average value for each quarter in 2015- 2018. The quarterly index forecast results for 2019 are shown in figure 7 step 7.

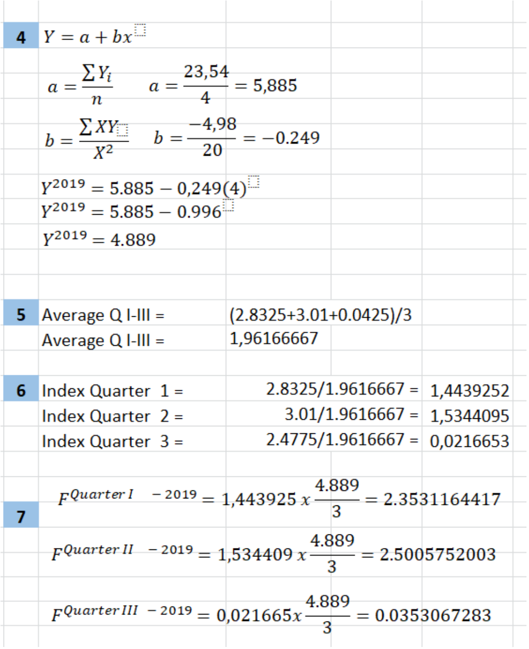


1	Year	Quarter I	Quarter II	Quarter III
	2015	4.12	4.12	-1.56
	2016	2.84	2.85	0.33
	2017	2.53	2.94	0.25
	2018	1.84	2.13	1.15

2	Year	Quarter I	Quarter II	Quarter III	Total Yearly
	2015	4.12	4.12	-1.56	6.68
	2016	2.84	2.85	0.33	6.02
	2017	2.53	2.94	0.25	5.72
	2018	1.84	2.13	1.15	5.12
	Total	11.33	12.04	0.17	
	Average	2.8325	3.01	0.0425	

3	Year	SPI(Y)	X	XY	X²
	2015	6.68	-3	-20.04	9
	2016	6.02	-1	-6.02	1
	2017	5.72	1	5.72	1
	2018	5.12	3	15.36	9
	Σ	23.54	0	-4.98	20

Figure 6. Forecasting SPI index using least square time series step 1-3



4 $Y = a + bx$

$a = \frac{\sum Y_i}{n}$ $\quad a = \frac{23{,}54}{4} = 5{,}885$

$b = \frac{\sum XY}{X^2}$ $\quad b = \frac{-4{,}98}{20} = -0.249$

$Y^{2019} = 5.885 - 0{,}249(4)$

$Y^{2019} = 5.885 - 0.996$

$Y^{2019} = 4.889$

5 Average Q I-III = (2.8325+3.01+0.0425)/3

Average Q I-III = 1,96166667

6 Index Quarter 1 = 2.8325/1.9616667 = 1,4439252

Index Quarter 2 = 3.01/1.9616667 = 1,5344095

Index Quarter 3 = 2.4775/1.9616667 = 0,0216653

7 $F^{Quarter\,I} - 2019 = 1{,}443925\,x\,\frac{4.889}{3} = 2.3531164417$

$F^{Quarter\,II} - 2019 = 1{,}534409\,x\,\frac{4.889}{3} = 2.5005752003$

$F^{Quarter\,III} - 2019 = 0{,}021665x\frac{4.889}{3} = 0.0353067283$

Figure 7. Forecasting SPI index using least square time series step 4-7

SPI index values for 2015-2018 and forecast results for 2019 are presented in table 4. The same values in graphical form are presented in figure 8. According to table 4, in 2019 from the first quarter to the third quarter there was no drought (more index values from 0). Values 0-2.5 has the meaning of normal to wet weather. The results of this forecast are in accordance with field observations made by researchers.

Table 4. SPI Index in 2015-2018 and its forecast for year 2019

Year	Quarter I	Quarter II	Quarter III
2015	4.12	4.12	-1.56
2016	2.84	2.85	0.33
2017	2.53	2.94	0.25
2018	1.84	2.13	1.15
2019	2.35	2.5	0.03

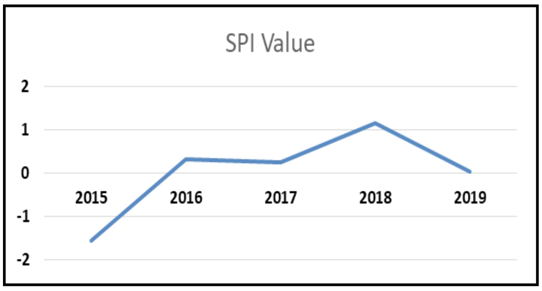


Figure 8. SPI Value during year interval 2015-2019

4. Conclusion

Description in the results and discussion section suggests that satellite imagery can be used to observe the occurrence of drought in an area. The NDVI and VHI index calculation results provide similar interpretation results related to drought in Western Southeast Mollucan (MTB). However, IDW analysis based on satellite imagery and SPI analysis using rainfall data give results that are not entirely the same. In the span of 4 years of observation, the two analytical methods give similar results for the 2015 and 2018 data, but give different results for the 2016 and 2017 data. This difference is caused by the number of data points under consideration. Satellite imagery has many data points whereas rainfall data only has one point for the whole area of MTB.

The least square time series method can be used to forecast the occurence of drought by utilizing drought index data in the past years to calculate drought index for the future. Drought forecast based on SPI data in the period of 2015-2018 gives the SPI index value at 0.03 for 2019. SPI index at such a level is interpreted as no drought, which fully agrees with the result of our field observation.

Reference

[1] R. D'Arrigo and R. Wilson, "El Nino and Indian Ocean influences on Indonesian drought: implications for forecasting rainfall and crop productivity," *Int. J. Climatol. A J. R. Meteorol. Soc.*, vol. 28, no. 5, pp. 611–616, 2008.

[2] I. G. Hendrawan, K. Asai, A. Triwahyuni, and D. V. Lestari, "The interanual rainfall variability in Indonesia corresponding to El Niño Southern oscillation and Indian Ocean Dipole," *Acta Oceanol. Sin.*, vol. 38, no. 7, pp. 57–66, 2019.

[3] L. Arumingtyas, "Jawa dan Nusa Tenggara Langganan Bencana Kekeringan, Mengapa?," *Mongabay*, 2018.

[4] A. Sulistyo, "Kombinasi Teknologi Aplikasi GPS Mobile dan Pemetaan SIG dalam Sistem Pemantauan Demam Berdarah (DBD)," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 5, no. 1, pp. 6–14, 2019.

[5] A. Syahrial, A. Azmeri, and E. Meilianda, "Analisis Kekeringan Menggunakan Metode Theory of Run di DAS Krueng Aceh," *J. Civ. Eng.*, vol. 24, no. 2, pp. 167–172, 2017.

[6] L. Wang, G. Huang, and W. Chen, "Towards a theoretical understanding of multiscalar drought indices based on the relationship between precipitation and standardized precipitation index," *Theor. Appl. Climatol.*, vol. 136, no. 3–4, pp. 1465–1473, 2019.

[7] W. Hatmoko, "Indeks Kekeringan Hidrologi untuk Alokasi Air di Indonesia." Puslitbang Sumber Daya Air, Bandung, 2012.

[8] M. Hendartyo, "BNPB: 4,87 Juta Jiwa Terdampak Kekeringan," *Tempo*, 07-Sep-2018.

[9] Saumlaki, "Maret, Kelaparan Ancam MTB Akibat Krisis Air Parah," *Dhara Pos*, 02-Mar-2016.

[10] A. Zubaidah, D. Dirgahayu, and J. M. Pasaribu, "Penginderaan jauh untuk pemantauan kekeringan lahan sawah," *J. Ilm. Widya*, vol. 1, no. 1, 2014.

[11] S. Y. J. Prasetyo, K. D. Hartomo, B. H. Simanjuntak, and D. W. Candra, "Mitigation & identification for local aridity, based of vegetation indices combined with spatial statistics & clustering k means," in *Journal of Physics: Conference Series*, 2019, vol. 1235, no. 1, p. 12028.

[12] R. P. Gupta, *Remote sensing geology*. Springer, 2017.

[13] V. A. Bento, I. F. Trigo, C. M. Gouveia, and C. C. DaCamara, "Contribution of land surface temperature (TCI) to vegetation health index: A comparative study using clear sky and all-weather climate data records," *Remote Sens.*, vol. 10, no. 9, p. 1324, 2018.

[14] S. M. Indirawati, S. Pandia, H. Mawengkang, and W. Hasan, "Inverse Distance Weighted Method and Environmental Health Risks of Plumbum Pollution in Drinking Water in Belawan Coastal Area," *Adv. Sci. Lett.*, vol. 23, no. 4, pp. 3339–3343, 2017.

[15] S. R. Fitri, E. Saadudin, B. Pranoto, and others, “Comparison of Inverse Distance Weighted (IDW), Natural Neighbour, and Spline Interpolation Methods for Downscaling Data of Solar Energy Potential Map,” *Ketenagalistrikan dan Energi Terbarukan*, vol. 13, no. 1, pp. 27–38, 2014.

[16] R. Kumar, M. Majid, S. Mir, and M. Shahzad, “Temporal analysis of drought using standard precipitation index (SPI) method,” *Indian J. Soil Conserv.*, vol. 45, no. 3, pp. 348–350, 2017.

[17] I. A. Andika, “Penerapan Metode Standardized Precipitation Index (SPI) untuk Analisa Kekeringan di DAS Ngasinan Kabupaten Trenggalek,” Universitas Brawijaya, 2016.

[18] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*. springer, 2016.

[19] P. K. Pradhan, S. Dhal, and N. K. Kamila, “Time series least square forecasting analysis and evaluation for natural gas consumption,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 5, no. 11, pp. 91–99, 2017.

[20] F. R. Hariri, “Metode Least Square Untuk Prediksi Penjualan Sari Kedelai Rosi,” *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 7, no. 2, pp. 731–736, 2016.

Peer Reviewers

The Board of Editors greatly appreciate the participation of the following reviewers that help during the review process for Khazanah Informatika since 2019.

1. Adi Supriyatna, Universitas Bina Sarana Informatika
2. Afandi Nur Aziz Thohari, Telkom Institute of Technology
3. Ahmad Luky Ramdani, Institut Teknologi Sumatera
4. Angelina Prima Kurniati, Telkom University
5. Anjar Wanto, STIKOM Tunas Bangsa, Pematangsiantar - Sumatera Utara
6. Aris Rakhmadi, Universitas Muhammadiyah Surakarta
7. Arief Wibowo, Universitas Budi Luhur
8. Arkham Zahri Rakhman, Universitas Muhammadiyah Surakarta
9. Assoc. Prof. Leon A. Abdillah, Bina Darma University
10. Ridho Ananda, Institut Teknologi Telkom Purwokerto
11. Azizah Fatmawati, Universitas Muhammadiyah Surakarta
12. Asslia Johar Latipah, Universitas Muhammadiyah Kalimantan Timur
13. Deny Jollyta, Sekolah Tinggi Ilmu Komputer (STIKOM) Pelita Indonesia
14. Devi Afriyanti Puspa Putri, Universitas Muhammadiyah Surakarta
15. Diah Priyawati, Informatika Universitas Muhammadiyah Surakarta
16. Didiek Sri Wiyono, Universitas Negeri Sebelas Maret
17. Dr. Bana Handaga, Universitas Muhammadiyah Surakarta
18. Dr. Slamet Riyadi, Universitas Muhammadiyah Yogyakarta
19. Dr. Eng. Favian Dewanta, Telkom University
20. Dwi Dwi Ely Kurniawan, Politeknik Negeri Batam
21. Dr. Kusrini Kusrini, Universitas AMIKOM Yogyakarta
22. Dwi Murdaningsih Pangestuty, Universitas Muhammadiyah Kalimantan Timur
23. Endang Wahyu Pamungkas, Universitas Muhammadiyah Surakarta
24. Eko Setiawan, Universitas Muhammadiyah Surakarta
25. Endah Sudarmilah, Universitas Muhammadiyah Surakarta
26. Frieyadie, STMIK Nusa Mandiri
27. Gunawan Ariyanto, Universitas Muhammadiyah Surakarta
28. Hari Prasetyo, Universitas Muhammadiyah Surakarta
29. Heru Supriyono, Universitas Muhammadiyah Surakarta
30. Herry Sujaini, Universitas Tanjungpura
31. Indra Waspada, Universitas Diponegoro
32. Irma Yuliana, Universitas Muhammadiyah Surakarta
33. Jan Wantoro, Universitas Muhammadiyah Surakarta
34. Lasmedi Afuan, Universitas Jenderal Soedirman
35. Lutfiyah Dwi Setia, Politeknik Negeri Madiun
36. Mardhiya Hayaty, Universitas AMIKOM Yogyakarta
37. Maryam, Universitas Muhammadiyah Surakarta
38. Mei Silviana Saputri, Universitas Indonesia
39. Muhammad Shulhan Khairy, Institut Teknologi Sepuluh Nopember
40. Naufal Azmi Verdikha, Universitas Muhammadiyah Kalimantan Timur
41. Nor Bakiah, Universiti Tun Hussein Onn Malaysia, Malaysia
42. Nu’man Normas, Universitas Muhammadiyah Surakarta
43. Puji Sari Ramadhan, STMIK Triguna Dharma Medan
44. Ramalia Narotama Putri, Sekolah Tinggi Ilmu Komputer Pelita Indonesia
45. Rajif Agung Yunmar, Institut Teknologi Sumatera
46. Ramos Somya, Universitas Kristen Satya Wacana
47. Rizki Wahyudi, Universitas AMIKOM Purwokerto
48. Rofilde Hasudungan, Universitas Muhammadiyah Kalimantan Timur
49. Sayekti Harits Suryawan, Universitas Muhammadiyah Kalimantan Timur
50. Sinar Nadhif Ilyasa, Universitas Muhammadiyah Surakarta
51. Sitaresmi Wahyu Handani, Universitas AMIKOM Purwokerto

52. Siti Helmiyah, Universitas Ahmad Dahlan
53. Siti Puspita Hida Sakti, STMIK Syaikh Zainuddin NW Anjani
54. Sri Karnila, Informatics and Business Institute Darmajaya Bandar Lampung
55. Sukirman, Universitas Muhammadiyah Surakarta
56. Tati Ernawati, Politeknik TEDC Bandung
57. Teguh Bharata Adji, Universitas Gadjah Mada
58. Titin Pramiyati, Universitas Pembangunan Nasional "Veteran" Jakarta
59. Tri Ginanjar Laksana, Institut Teknologi Telkom Purwokerto
60. Umi Fadlilah, Universitas Muhammadiyah Surakarta
61. Wiwit Supriyanti, Politeknik Indonusa Surakarta
62. Yogiek Indra Kurniawan, Universitas Jenderal Soedirman
63. Yuliant Sibaroni, Universitas Telkom
64. Yusuf Sulistyo Nugroho, Universitas Muhammadiyah Surakarta