Transcendent Journal of Mathematics and Application
ISSN 2964-1845
Vol. 3, No. 1, April-May 2024, pp. 11-19

Rice Leaf Diseases Classification Using Image Processing and Machine Learning

Abdul Mahatir Najar[1,*], Maulidyani Abu[1], Resnawati[1], Syahrullah[2]
[1]Mathematics Study Program, Tadulako University, Kota Palu 94118, Indonesia
[2]Information Systems Study Program, Tadulako University, Kota Palu 94118, Indonesia
*Corresponding email: abdulmahatir@gmail.com

ARTICLE INFO

Article History
Received : 02-03-2024
Revised : 01-04-2024
Accepted : 05-04-2024

Keyword
Classification;
Disease classification;
Disease detection;
Image processing;
Machine learning;
Rice disease

ABSTRACT

Identification of diseases from images of plants is one of the interesting research areas in the agriculture field, for which machine learning concepts from the computer science field can be applied. This article presents a prototype system for the detection and classification of rice diseases based on images of infected rice plants. This prototype system was developed after detailed experimental analysis of various techniques used in image processing operations. We consider three rice plant diseases: Bacterial Leaf Blight, Blast, and Tungro. We used the Otsu method to remove the background. To enable accurate extraction of features, we combined Gabor and Sobel techniques. In the classification process, we used five machine learning techniques: Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Quadratic Discriminant Analysis (QDA). We empirically evaluated these methods, achieving 77%, 50%, 60%, and 37% accuracy, respectively.

DOI: …

1. INTRODUCTION

The significant population growth has made food security one of the primary challenges facing almost all countries worldwide, particularly those in Asia and Africa. According to research conducted by the Food and Agriculture Organization (FAO) in 2018, it is estimated that by 2050, the demand for food will double from that of 2012. This surge in demand is exacerbated by economic pressures and climate change, leading to a potential food crisis affecting 113 million people across 53 countries (FAO, 2019). Indonesia, once known as an agrarian nation, is not exempt from this threat. The uneven access to food is evident in the high prevalence of stunting, reaching 30.8 percent nationally in 2018 (Ministry of Health, 2018). Indonesia's prevalence of stunting is higher compared to other Southeast Asian countries such as Vietnam and Thailand. Additionally, according to the Global Hunger Index (GHI) in 2016, Indonesia's hunger status falls into the alarming category, with a score ranging from 35.0 to 49.9, on par with Cambodia, Laos, and Timor-Leste.

Indonesia's dependence on food imports to meet the needs of its population is also worrisome. The issue of food security encompasses a myriad of complex problems from upstream to downstream. Rice, as a staple commodity for the Indonesian people, is not exempt from these challenges. To date, the Indonesian government still relies on imports to meet the demand for rice. Declining agricultural land, environmental degradation, extreme climate change, and various diseases affecting crops often lead to frequent crop failures. According to Swain et al. (2020), diseases affecting rice can be observed through changes in leaves, particularly changes in morphology and color. Proper management of each disease is crucial to prevent crop failure. However, not all farmers have the expertise to identify the types of diseases and appropriate treatments, often resorting to treatments that may have long-term detrimental effects on the environment.

The agricultural challenges described above necessitate a paradigm shift towards agricultural transformation oriented towards ecology, economics, optimization of yields, and sustainability. In line with this, the development of technology and information can be utilized intelligently in agricultural fields. By leveraging technologies such as big data and artificial intelligence, farmers can make effective and efficient decisions (Gupta & Sarvaiya, 2020).

Machine learning is one of the artificial intelligence technologies that can be utilized in the effort to identify diseases in agricultural crops. To improve crop yields in terms of quantity and quality, disease identification must be done as early as possible so that farmers can provide appropriate treatments. Generally, monitoring and identification can be done manually by farmers, but it requires good skills and a long time, especially if the land is extensive. With the assistance of image processing technology and machine learning, the identification of disease types can be done more efficiently (Daniya & Vigneshwari, 2019; Shah et al., 2016).

At the forefront of agricultural innovation, advanced technologies like big data analytics, artificial intelligence (AI), and machine learning (ML) are transforming farming practices. Big data analytics collects and analyzes vast agricultural data, offering insights for decision-making. AI, including neural networks and deep learning, develops predictive models for crop yield estimation, pest detection, and irrigation management (Kamilaris et al., 2017). ML algorithms such as random forests and support vector machines aid in crop disease identification and prediction based on image analysis (Mohanty et al., 2016). Remote sensing technologies like drones and satellite imaging provide high-resolution data for monitoring crop health and optimizing practices. Precision agriculture, integrating GPS and sensor technologies, enables targeted resource management for enhanced productivity. Additionally, advancements in crop genetics, biotechnology, and blockchain technology for supply chain transparency further propel agricultural innovation (Zheng et al., 2020). These technologies promise to revolutionize agriculture, addressing global food security challenges by improving productivity, sustainability, and resilience.

2. METHODS

The research was conducted in several stages as illustrated in Figure 1. It began with data collection, followed by preprocessing, segmentation, feature extraction, and classification processes.

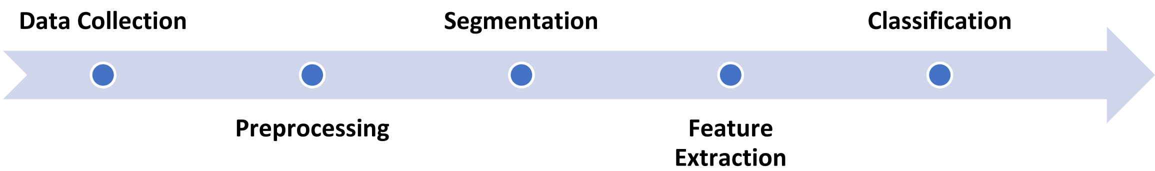

Figure 1. Research stages

2.1. Data Collection

The data used in this research is open data obtained from www.kaggle.com. The data consists of images of rice leaves classified into 3 groups, namely leaves identified with *blight, blast,* and *tungro* diseases.

(a) (b) (c)

Figure 2. Rice leaves with (a) *blast*, (b) *blight*, and (c) *tungro* disease

2.2. Preprocessing

Before further processing, the image data is first resized to 128 x 128 pixels. Next, each pixel is normalized so that its value is between 0 and 1. This is done to ensure uniform image sizes and facilitate processing in the subsequent stages.

2.3. Segmentation

Segmentation is performed with the aim of focusing on the object under study. In this research, the Otsu method is used. This method comprehensively searches for a threshold value that minimizes the variance within the defined classes, which is defined as the weighted sum of the variances of both classes (Li et al., 2020).

$$\sigma_w^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t) \tag{1}$$

where ω_0 and ω_1 is a weight that shows the probability of two classes separated by a threshold t, and ${\sigma_0}^2$ and ${\sigma_1}^2$ is the variance of the two classes.

2.4. Feature Extraction

Feature extraction is one of the crucial processes in image classification. This process involves extracting distinctive characteristics from the image in terms of color, shape, and texture domains. The extracted features represent the local and global statistics of the image (Azim et al., 2021). In this research, feature extraction is performed using several methods. The first method is the Gabor Filter, which is known as a successful feature detector due to its ability to eliminate variability caused by contrast lighting and slight image shifts and deformations.

$$g(x, y, \theta, \phi) = \exp(-\frac{x^2 + y^2}{\sigma^2})\exp(2\pi\theta i(x\cos\phi + y\sin\phi))) \tag{2}$$

Next, for edge detection, the Sobel method is used with the following formulas:

$$G = \sqrt{S_x^2 + S_y^2} \tag{3}$$

2.5. Classification

In the classification process, we applied several machine learning algorithms to the extracted data.

1. Random Forest

Random forest is a method that combines multiple random decision trees and aggregates their predictions by averaging them. This method performs very well, even in situations where the number of variables is much larger than the number of observations. Additionally, it is quite versatile for application to large-scale problems and can be easily adapted to various learning tasks. Random Forest is a predictor that consists of a collection of M random regression trees (Biau & Scornet, 2016).

$$m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n) = \sum_{i \in \mathcal{D}_n^{\star}(\Theta_j)} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)} Y_i}{N_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)} \tag{4}$$

where, $\mathcal{D}_n^*(\Theta_j)$ is the data set selected before the tree is formed. $A_n(\mathbb{x}: \Theta_j, \mathcal{D}_n)$ is the containing cell **x**.

2. Support Vector Machine

Support Vector Machine (SVM) is a machine learning method that has a relatively simple concept yet is flexible enough to address various classification problems (Pisner & Schnyer, 2020). The SVM model built for classification is called Support Vector Classification (SVC). The concept of SVC can be explained simply as an attempt to find the best hyperplane that serves as a separator between two classes in the input space. Figure 3(b) shows several patterns that are members of two classes: +1 and -1. Patterns belonging to the -1 class are symbolized by red color (squares), while patterns in the +1 class are symbolized by yellow color (circles). The classification problem can be translated into an effort to find a line (hyperplane) that separates between the two groups. Various alternative separation lines (discrimination boundaries) are shown in Figure 3(a).

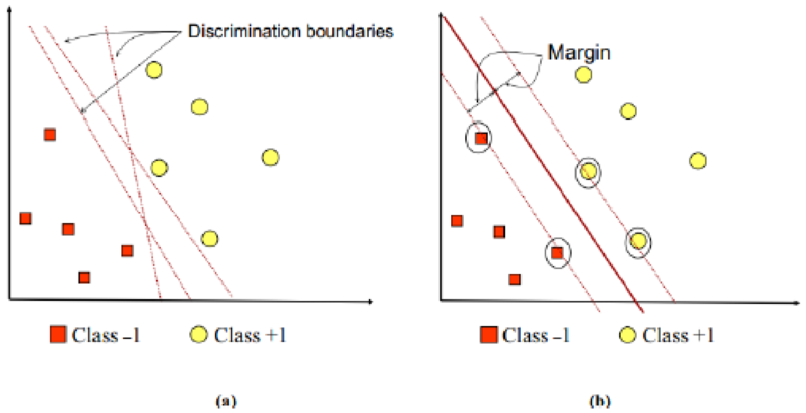

Figure 3. Hyperplane on SVC

3. Naïve Bayes

Naïve Bayes is a simple machine learning algorithm that utilizes Bayes' rule along with the strong assumption that the attributes are conditionally independent based on their class. Although this independence assumption is often violated in practice, naïve Bayes often provides competitive classification accuracy. Coupled with computational efficiency and many other desirable features, this has led to the widespread application of naïve Bayes in various fields. For example, given a problem with a domain consisting of n attributes, the mathematical mapping validity between the vector of attributes $(A_1, A_2, \ldots, A_n)$ and the set of classifications

$(C_1, C_2, \dots, C_3)$. will be calculated. This validity can be computed using Bayesian inference (Yang, 2018).

$$
\begin{aligned}
&P(C_i|(A_1 = v_1 \cap A_2 = v_2 \cap \dots \cap A_n = v_n)) \\
&= \frac{P(C_i \cap (A_1 = v_1 \cap A_2 = v_2 \cap \dots \cap A_n = v_n))}{P(A_1 = v_1 \cap A_2 = v_2 \cap \cap A_n = v_n)} \\
&= \frac{P((A_1 = v_1 \cap A_2 = v_2 \cap \dots \cap A_n = v_n) \cap C_i)}{P(A_1 = v_1 \cap A_2 = v_2 \cap \cap A_n = v_n)} \\
&= \frac{P((A_1 = v_1 \cap A_2 = v_2 \cap \dots \cap A_n = v_n)|C_i) \times P(C_i)}{P(A_1 = v_1 \cap A_2 = v_2 \cap \cap A_n = v_n)} \\
&where \quad 1 \le i \le m
\end{aligned}
\tag{5}
$$

After performing a series of probabilistic calculations, the best value between the given attribute vector and its classification probability can be quantitatively indicated.

4. Quadratic Discriminant Analysis

Discriminant Analysis is a classifier with quadratic decision boundaries, which are derived by fitting class conditional densities to the data and using Bayes' rule. The model fits Gaussian densities to each class (Tharwat, 2016).

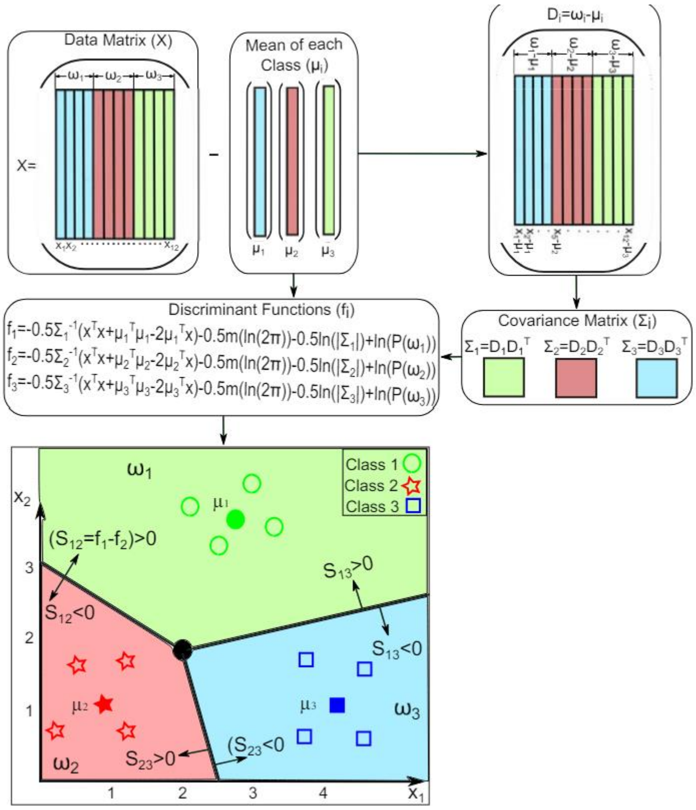

Figure 4. Quadratic Discriminant Analysis Algorithm

Figure 4 illustrates the process of developing a classification model using discriminant analysis method. Given a matrix X representing N data, the first step is to calculate the mean of each class, followed by computing the initial probabilities of each class. After that, calculate the covariance of each class and proceed to compute the discriminant function for each class.

3. RESULT AND DISCUSSION

After obtaining the features, the process continues with classifying rice leaves based on their disease characteristics. In this research, they are classified into 3 groups. The classification process is attempted using several methods.

1. Random Forest Methods

The model built using the Random Forest method achieves an accuracy of 77%, with the following confusion matrix:

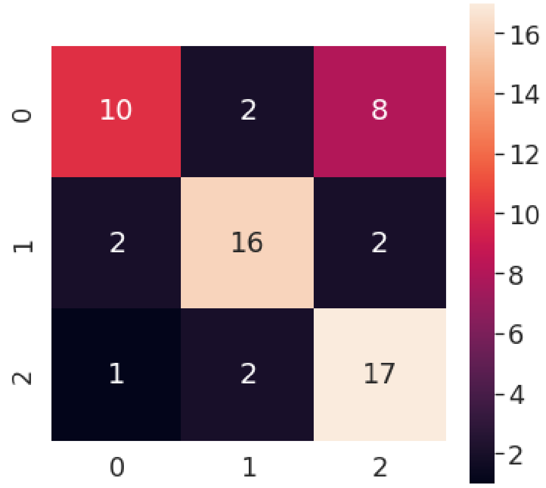

Figure 5. Random forest classification

2. Support Vector Machine Methods

The model built using the Support Vector Machine method achieves an accuracy of 50%, with the following confusion matrix:

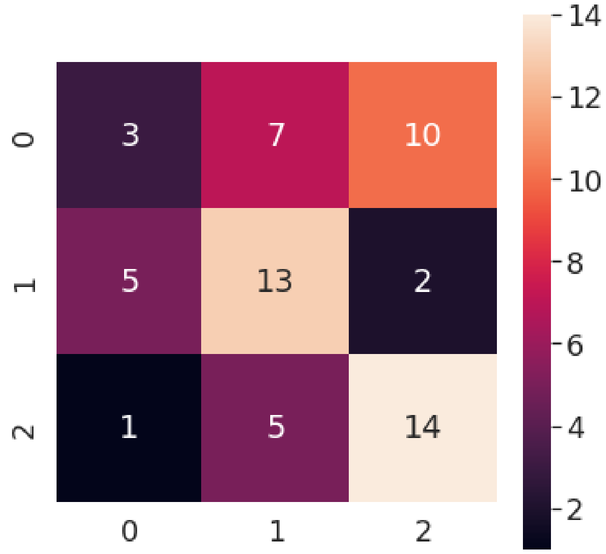

Figure 6. Support vector machine classification

3. Naïve Bayes Methods

   The model built using the Naïve Bayes method achieves an accuracy of 60%, with the following confusion matrix:

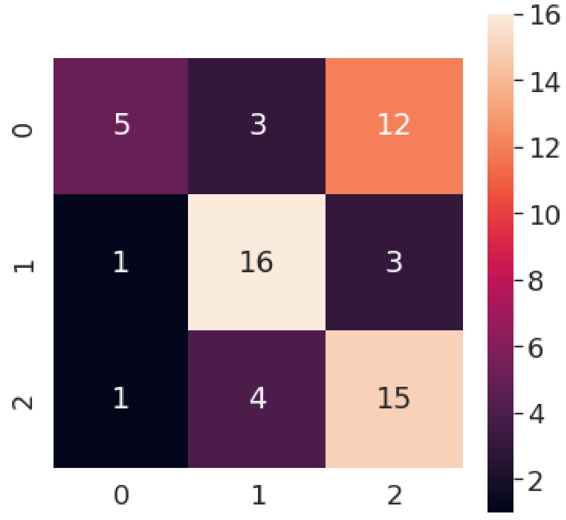

Figure 7. Naïve Bayes classification

4. Quadratic Discriminant Analysis Methods

   The model built using the Decision Tree Classifier method achieves an accuracy of 37%, with the following confusion matrix:

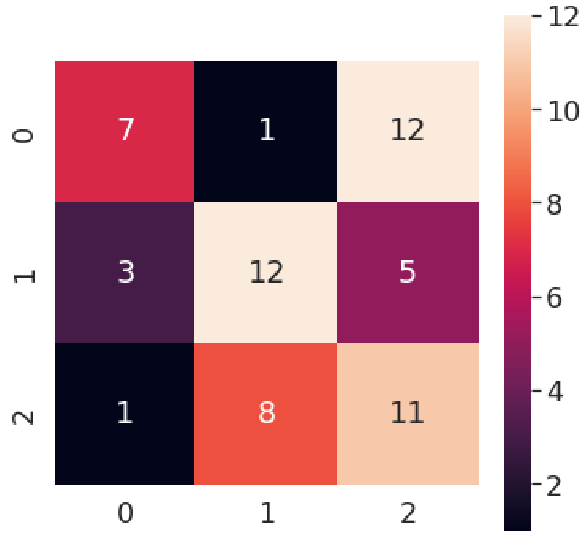

Figure 8. Quadratic discriminant analysis classification

Random Forest is a powerful ensemble learning method that combines multiple decision trees to improve classification accuracy. One of its key strengths is its ability to handle large and diverse datasets effectively. By constructing multiple decision trees and aggregating their predictions, Random Forest can mitigate the risk of overfitting, even in situations where the number of features exceeds the number of observations. However, Random Forest models can be computationally intensive to train, especially for large datasets, and their complexity may make them less interpretable compared to simpler models. Support Vector Machine

(SVM) is another popular classification algorithm known for its ability to handle high-dimensional data and nonlinear relationships. SVM works by finding the optimal hyperplane that separates classes in the feature space. It is effective in scenarios where the number of features is much larger than the number of samples, making it suitable for tasks like image recognition and text classification. However, SVM can be computationally demanding, especially for large datasets, and requires careful selection of parameters like the kernel type and regularization parameter. Naive Bayes is a simple yet powerful probabilistic classifier that is widely used for text classification and other tasks. It is based on Bayes' theorem and assumes that features are conditionally independent given the class label. Naive Bayes models are easy to implement and computationally efficient, making them suitable for large-scale applications. However, the assumption of feature independence may not hold true in practice, which can lead to suboptimal performance in some cases. Quadratic Discriminant Analysis (QDA) is a classification method that models the class conditional densities using quadratic functions. Unlike Linear Discriminant Analysis (LDA), which assumes equal covariance matrices for all classes, QDA allows for different covariance matrices for each class. This flexibility makes QDA suitable for datasets with complex relationships and varying covariance structures. However, QDA requires more parameters to be estimated compared to LDA, which can lead to overfitting, especially in high-dimensional spaces.

In summary, each classification method has its own strengths and weaknesses. Random Forest is versatile and robust, but computationally intensive. SVM is effective for high-dimensional data, but requires careful parameter tuning. Naive Bayes is simple and efficient, but relies on strong independence assumptions. QDA offers flexibility and can capture complex relationships, but may suffer from overfitting. The choice of classification method depends on the specific characteristics of the dataset and the goals of the analysis.

4. CONCLUSION

Plant diseases in rice can cause significant losses in agriculture. This article presents a system for detecting and classifying three rice plant diseases: Bacterial Leaf Blight, Blast, and Tungro. We experimentally evaluate all possible techniques in the image processing section, including background removal techniques. Finally, we demonstrate that the background of an image is accurately removed by applying the Otsu thresholding method. We experimentally evaluate various segmentation and feature extraction techniques. We use four machine learning techniques: Random Forest, Support Vector Machine, Naïve Bayes, and Quadratic Discriminant Analysis. We empirically evaluate these methods, achieving accuracies of 77%, 50%, 60%, and 37%, respectively.

REFERENCES

Azim, M. A., Islam, M. K., Rahman, Md. M., & Jahan, F. (2021). An effective feature extraction method for rice leaf disease classification. *TELKOMNIKA*, *19*(2), 463. https://doi.org/10.12928/telkomnika.v19i2.16488

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Daniya, T., & Vigneshwari, D. S. (2019). A review on machine learning techniques for rice plant disease detection in agricultural research. *International Journal of Advanced Science and Technology*, *28*(13), 49-62.

FAO. (2018). *Asia and the Pacific regional overview of food security and nutrition 2018: Accelerating progress towards the SDGs*. Food & Agriculture Organization of the United Nations.

FAO. (2019). *Global Report on Food Crises*. Food & Agriculture Organization of the United Nations.

Gupta, S., & Sarvaiya, J. N. (2020). Emerging Technology Trends in Electronics, Communication and Networking: Third International Conference, *ET2ECN 2020, Surat, India, February 7–8, 2020, Revised Selected Papers (Vol. 1214)*. Springer Singapore. https://doi.org/10.1007/978-981-15-7219-7

Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, *143*, 23-37.

Kemenkes. (2018). *Hasil Utama Riskesdas 2018*. Badan Penelitian dan Pengembangan Kesehatan-Kementerian Kesehatan.

Li, N., Lv, X., Xu, S., Li, B., & Gu, Y. (2020). An improved water surface images segmentation algorithm based on the otsu method. *Journal of Circuits, Systems and Computers*, *29*(15), 2050251. https://doi.org/10.1142/S0218126620502515

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, *7*, 1419.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. *In Machine Learning (pp. 101–121)*. Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Shah, J. P., Prajapati, H. B., & Dabhi, V. K. (2016). A survey on detection and classification of rice plant diseases. *2016 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, 1–8. https://doi.org/10.1109/ICCTAC.2016.7567333

Swain, S., Nayak, S. K., & Barik, S. S. (2020). A review on plant leaf diseases detection and classification based on machine learning models. *Mukt Shabd Journal*, *9*(7), 5195-5205. https://doi.org/09.0014.MSJ.2020.V9I6.0086781.105023

Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: A tutorial. *International Journal of Applied Pattern Recognition*, *3*(2), 145. https://doi.org/10.1504/IJAPR.2016.079050

Yang, F.-J. (2018). An Implementation of Naive Bayes Classifier. *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 301–306. https://doi.org/10.1109/CSCI46756.2018.00065

Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2020). An overview of blockchain technology: Architecture, consensus, and future trends. *In IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)* (pp. 1861-1865). IEEE.