# Crab Molting Identification using Machine Learning Classifiers

Runal Rezkiawan Baharuddin
*Department of Informatics*
*Universitas Hasanuddin*
Makassar, Indonesia
baharuddinrr17d@student.unhas.ac.id

Muhammad Niswar
*Department of Informatics*
*Universitas Hasanuddin*
Makassar, Indonesia
niswar@unhas.ac.id

Amil Ahmad Ilham
*Department of Informatics*
*Universitas Hasanuddin*
Makassar, Indonesia
amil@unhas.ac.id

Shigeru Kashihara
*Department of Network Design*
*Osaka Institute of Technology*
Osaka, Japan
shigeru.kashihara@oit.ac.jp

***Abstract*—Soft-shell crab is an export product in which foreign demand is much higher than production. The production of soft-shell crabs done by selecting the crabs just prior to molting and placing them in a box until the molting occurs. Molting is a natural process of shedding the shell when crabs respond to the lack of growth space within its shell. Shortly after molting, the new crab shells are still very soft and will be hardened in a few hours after the crabs absorb calcium from water. Farmer must harvest the crab while the crabs' shell is soft. This study investigates the initial identification of crab molting using machine learning classifier. We collected 1060 image datasets of crab molting and we divide data into 1000 training data and 60 testing data. We use three machine learning classifiers, namely K-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and the Random Forest Classifier (RFC). This study aims to compare and determine the best classification algorithm to be used for crab's molting identification. The experimental results show that, KNN is the best classification algorithm for initial identification of crab's molting.**

***Index Terms*—Crab Molting Identification,Image Processing, Classification Algorithms, Machine Learning.**

## I. Introduction

Mangrove crab (Scylla serrata Forsskål, 1775) is one of the marine biological resources with a wide and plentiful distribution in Indonesian waters. Mangrove crab is a type of crustacean with high economic value that has been widely produced by traditional farmers to supply food demands in both domestic and international markets. According to BPS (2016), it is stated that the crab export market to several countries, namely, Taiwan, China, Malaysia, Japan, America, Italy, and Singapore, is quite large. Mangrove crab exports amounted to 109,624.4 tons in 2015 and increased by 17.74% per year from 58,091.8 tons in 2010 [1].

Soft-shell crabs are a type of seafood that is well-known for its delicacy all over the world. Although many are produced in Indonesia, the Indonesian people are unfamiliar with this product. This happens because soft-shell crab is an export product where foreign demand is much higher than production. These commodities are exported to America, China, Japan, Hong Kong, South Korea, Taiwan, Malaysia, and a number of countries in Europe. Soft-shell crab production is carried out by raising individual crabs in a crab box which is placed in the pond until molting. Molting is a natural process of casting or shedding the feathers, skin, or the like, that will be replaced by a new growth, i.e. removing the old tough skin for growth purposes. Immediately after molting, the new crab shells are still very soft and will harden again a few hours later after water absorption occurs. These crabs with soft conditions are harvested as soft crabs [3].

There are numerous treatments that can be used to speed up the harvest of soft crabs, one of which is the mutilation procedure, which involves removing the crab legs to stimulate molting. According to Raden Ario et al, in his research on Differences in Mutilation Methods on the Length of Molting Time Scylla serrata, the length of molting time and using the technique had no effect on absolute weight growth [4]. In the cultivation of soft crabs, supervision will determine the quality of successful soft crabs. This is because the crabs are harvested shortly after molting. If the crabs are harvested too late, their skin will soon harden again and this causes the quality to decrease. [5]

The research above proves that the importance of identifying crabs molting in soft-shell crabs cultivation by using Machine Learning classifiers for early detection so that it can be handled more quickly when crabs soft. Therefore, this study proposes a molting crab detection using classification algorithms of machine learning and compare them to find out the best algorithms for detection.

## II. Related Works

In a related study, reference [6] conducted research on the classification and recognition of fish objects using machine learning technology with support vector machine (SVM) algorithm methods by classifying the number of 5 ornamental fish using the SVM algorithm with 250 image test data. The results show the Accruacy value of 50%, Precision 90%, Recall 47% and f1 score 63.94%.

Reference [7] used the Oriented Fast And Rotated Brief (ORB) and k-nearest neighbor (k-NN) for distinguishing the type of fishes. Their research results showed that the proposed



978-1-6654-0544-7/22/$31.00 ©2022 IEEE

method can classify 39 types of fish with accuracy of 97.5% using the k-NN.

Research on fish classification using machine learning technology has also been carried out by authors in [8]. They classified various fishes using machine learning algorithms, namely artificial neural networks (ANN) and k-NN and support vector machine (SVM). Their research results showed that the SVM perform well with 94% accuracy. In addition to fish classification, authors in [9] implemented a shrimp disease classification using SVM with accuracy of 81.27%.

Reference [10] proposed method to identify crabs parasites using several machine learning algorithms, namely logistic regression (LR), k-NN, Gaussian Naive Bayes (GNB), SVM, and linear discriminant analysis (LDA). The research results showed the LDA and GNB algorithms obtained the highest accuracy. In our research, we proposed a method to identify crabs molting using machine learning classifiers to help farmers to produce soft-shell crabs.

## III. RESEARCH METHOD

Research method consists of data collection, pre-processing, object detection, classification and evaluation. Figure 1 shows the research workflow. First, we collecting images data using Raspberry pi camera and split the collected data into training data and testing data. Secondly, pre-processing is carried out including image resizing, Image transformation and feature HOG extraction. After that we select three classification algorithms(k-NN, SVM, and RFC) and train the models. Finally, we evaluated and optimized the models to get the best accuracy, compare the models and then selecting the best model.

Figure 2 shows the devices used to collect data including the raspberry pi and the camera. We place the camera at a height of 2 to 4 cm from the box contain a crab. The raspberry pi take a pictures every 10 seconds and stores the image data on the hard drive.

### A. Data Collection and Pre-Processing

The dataset is separated into two categories, namely molting and non-molting classes. Figure 3.a and 3.b shows Non-molting (with mutilated leg) and molting crab, respectively.

The dataset was divided into two classes, namely molting and non-molting. From the total of 1060 crab images, each class has 530 images and they were divided into 2 datasets, i.e., 500 images of training dataset and 30 images of test datasets. After dividing the image dataset, pre-processing the dataset is conducted including image resizing, image transformation, and HOG feature extraction.

1) *Image Resizing*

Image resizing aims to establish a base size for all images because captured images vary in size. Image resizing done by using the the cv2 package available in the OpenCV-python library. The set size is 72 x 72 pixels. Figures 4 and 5 are data before resizing and after resizing, respectively.

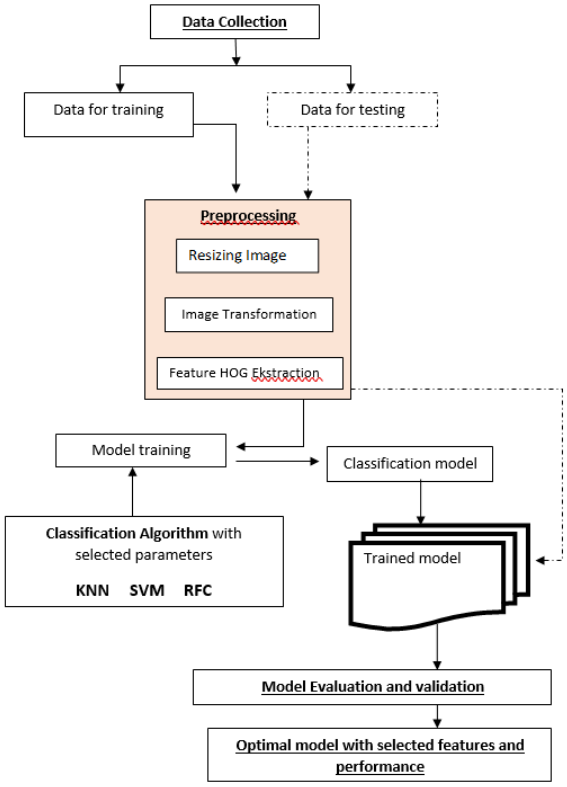


Fig. 1. Research methodology flow

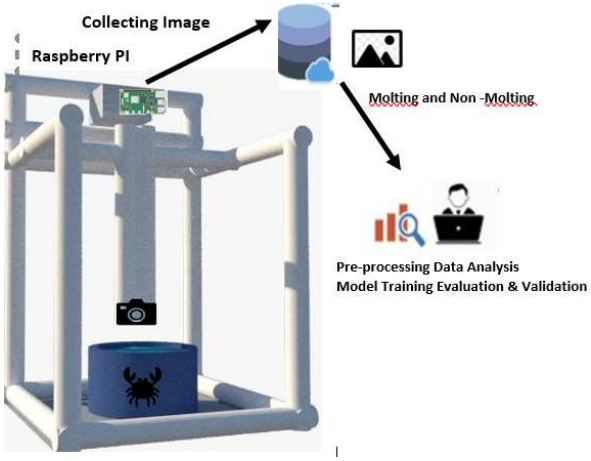


Fig. 2. Devices for Collecting Data

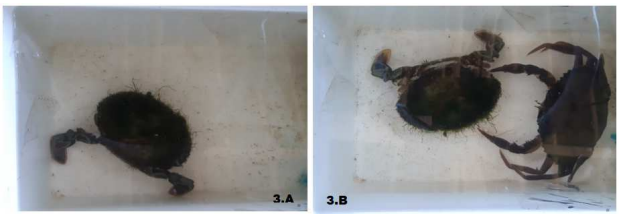


Fig. 3. Non-Molting and Molting

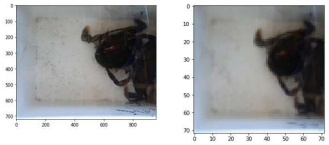

Fig. 4. Molting image data before resizing and after resizing

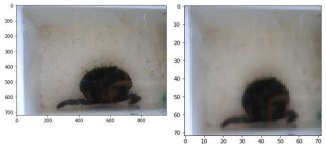

Fig. 5. Non-Molting image data before resizing and after resizing

2) *Image Transformation*

The next stage is image transformation. The color of the resized image will be changed to grayscale. This is done to keep the pixel value that will be taken from being too wide [11]. The image that is transformed into a grayscale color only has a pixel range from 0 to 1, 0 as white, 1 as black, and the value between them is gray. After becoming a grayscale image, the pixel value of the color will be taken. Figures 6 and 7 are images that has not been transformed and that has been transformed to greyscale.

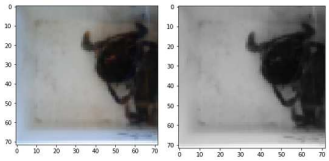

Fig. 6. molting image data that has not been transformed and that has been transformed to greyscale

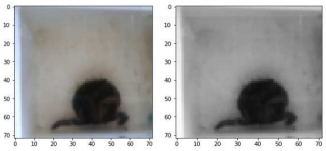

Fig. 7. non-molting image data that has not been transformed and that has been transformed to greyscale

3) *Feature HOG Extraction*

At this stage, the HOG feature extraction of the training image and the test image were compeleted. In this process, the previous size was 72x72 pixels. After resizing the data as many as 1060 images, then the molting and non-molting crab images on the training data and test data were converted to grayscale images. HOG is a form of local object and the value is used from the gradient intensity to extract features used in computer vision and image processing. HOG has the advantage of capturing edges or gradient structures that are characteristic to the actual shape [12]. The gradient distribution indicates the features of each image. The image is divided into small areas called cells.
Figure 8 shows result of feature HOG extraction.

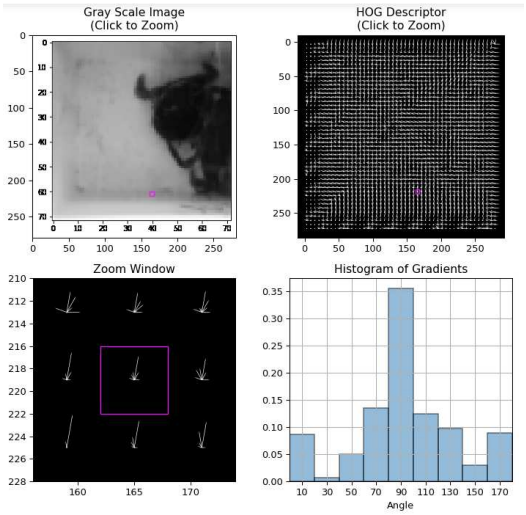


Fig. 8. Result of Feature HOG Extraction

## *B. Molting Identification with Classification Algorithms*

The object to be detected consists of 2 categories, namely molting and non-molting crabs. The object classification process uses machine learning algorithms including the k-NN, random forest classifier (RFC), and SVM. In this stage, these machine learning algorithms finds the pattern in the training datasets and build a model that captures these patterns.

1) Classification

Classification airms to categorize data into a given number of classes. This research aims to categorize data into two classes, namely molting and non-molting. We use three classification algorithms and compare which algorithm has the highest accuracy in classifying.

a) k-Nearest Neighbors (k-NN)

KNN is a classification algorithm that works by grouping data based on adjacent to the other data [13]. Calculation of the distance between the data and the group using the Euclidian and Minkowski distance formulas as shown in equation 2 [14].

- Distance Euclidian function:

$$d\,(a,b) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

- Distance Minkowski function:

$$d\,(a,b) = \sqrt[p]{(x_1 - x_2)^p + (y_1 - y_2)^p} \quad (2)$$

b) Support Vector Machine (SVM)

SVM can be used to create a model for both classification and regression cases [15]. SVM works by finding the outermost point of each data

and then drawing the most optimal dividing line [16]. From the dividing line that has been obtained, there will be two or more data groups. SVM can perform both linear and non-linear classification. SVM has many defined kernel functions according to data classification. Some of the main kernel functions is [17] :

- Linear kernel function:

$$K\left(x_i, x_j\right) = x_i^T x_j \tag{3}$$

- Polynomial Kernel function:

$$K\left(x_i, x_j\right) = \left(y x_i^T, x_j + r\right)^d, \gamma > 0\Big) \tag{4}$$

- Radial Basis function(RBF) function:

$$\left(x_i, x_j\right) = exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0\Big) \tag{5}$$

Equation (3-5) explain the working of kernels where $x_i$ dan $x_j$ are the inputs and $\gamma$ is regularization factor. The efficiency of the model can be improved by the appropriate selection of kernel and certain parameters such as $\gamma$, $C$ dan $\epsilon$.

c) Random Forest Classifier (RFC)

Random Forest is an integration of bagging and several Decision Trees [18]. Several tree combinations will produce several votes which are then taken by the most votes or voting. This voting result is the final result of the classification. Breiman has used Gini index as the goodness measure to split the attributes on. The Gini index is a statistical measure for quantifying the heterogeneity of a dataset [19]. However, it has been first introduced by the Italian statistician Corrado Gini in 1912. The index is a function that could be used to measure the impurity of the data, i.e., how uncertain we are if an event will occur. In classification, this event would be the determination of the class label. The impurity is measured by the Gini index, which has the following form [20]:

$$G = 1 - \sum_{i=1}^{k} \left(p(c_i|t)\right)^2 \tag{6}$$

where $t$ is a condition, $k$ the number of classes in the data set, and $C_i$ is the $i^th$ class label in the data set.

2) Parameters Tuning

At this stage, each algorithm is configured using several parameters to determine the effect of parameters on the resulting performance. In the KNN algorithm, the parameters tested are the distance type and the number of neighbors. Table 1 shows the parameters of the KNN used in this study. The number of neighbors used are 1, 7, and 9, while the distance types used are Euclidean [21] and Minkowski [22].

TABLE I. PARAMETERS K-NN

| **Parameters** | ***Description*** |
|---|---|
| *N*umber Neighbors | 1,7,9 |
| *T*ype Distance | Euclidean,Minkowski |

TABLE II. PARAMETERS SVM

| **Parameters** | ***Description*** |
|---|---|
| *K*ernel | Linear,Poly, RBF |

In the SVM algorithm, the parameter being tested is the type of kernel. Table 2 shows the parameters of the SVM.

The Random Forest Classifier (RFC) algorithm being tested was estimators, namely 20, 30, 40. Table 3 shows the parameters of the random forest classifier (RFC).

TABLE III. PARAMETERS RFC

| **Parameters** | ***Description*** |
|---|---|
| *E*stimators | 20, 30, 40 |

### *C. Performance Evaluations & Results*

We conduct the performance evaluation of the KNN, SVM, and RFC algorithms for crab molting identification. The performances matrics are accuracy, precision, recall, and F1 Score. Precision is the ratio of true positive predictions compared to the overall positive predictive result. Recall is the ratio of true positives compared to all positive data. f1-score is a comparison of precision and weighted mean gain [23]. The Confusion Matrix allows to compare the model's classification results to the real classification results. Accuracy is the ratio of true predictions (positive and negative) to the overall data. Equations (7),(8),(9),(10) respectively show the formula for calculating accuracy. TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} * 100\% \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1Score = \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

Table 4 shows the performance results of the k-NN algorithm. NN is Number Neighbors (Number of Neighbors), the matrix is the distance counter parameter. The value taken is AVG or the average performance of each class.

Based on the performance calculations, the k-NN algorithm with Minkowski and Euclidean distances with 1 neighbor has

TABLE IV. PERFORMANCE K-NN

| **Distance .** | NN . | **Confusion Matrix** | | | | Total Data **Testing** | **Accuracy .** |
|---|---|---|---|---|---|---|---|
| | | ***TP*** | ***TN*** | ***FP*** | ***FN*** | | |
| Euclidean | 1 | 30 | 29 | 0 | 1 | 60 | **98.3%** |
| Euclidean | 7 | 29 | 29 | 1 | 1 | 60 | 96.7% |
| Euclidean | 9 | 29 | 28 | 1 | 2 | 60 | 95.0% |
| Minkowski | 1 | 30 | 29 | 0 | 1 | 60 | **98.3%** |
| Minkowski | 7 | 29 | 29 | 1 | 1 | 60 | 96.7% |
| Minkowski | 9 | 29 | 28 | 1 | 2 | 60 | 95.0% |

the best performance. The difference in performance is not significant, but k-NN with distance Minkowski and Euclidean with the number of neighbors 1 will be compared their performance with SVM and RFC.

Table 5 shows the performance of the SVM algorithm. SVM with Linear kernel has the best performance when compared to other kernels. All performance metrics of the Linear kernel outperform other kernels. SVM with Linear kernel will be compared its performance with k-NN and RFC.

TABLE V. PERFORMANCE SVM

| **Kernel .** | **Confusion Matrix** | | | | Total Data Testing | **Accuracy .** |
|---|---|---|---|---|---|---|
| | ***TP*** | ***TN*** | ***FP*** | ***FN*** | | |
| Linear | 29 | 29 | 1 | 1 | 60 | **96.7%** |
| Poly | 25 | 25 | 5 | 5 | 60 | 83.3% |
| RBF | 25 | 25 | 5 | 5 | 60 | 83.3% |

Table 6 shows the performance results of the RFC algorithm. It can be seen that the RFC with 30 estimators has a good performance, all the resulting performance metrics have values above 95%.

TABLE VI. PERFORMANCE RFC

| **Estimators .** | **Confusion Matrix** | | | | Total Data Testing | **Accuracy .** |
|---|---|---|---|---|---|---|
| | ***TP*** | ***TN*** | ***FP*** | ***FN*** | | |
| 20 | 29 | 28 | 1 | 2 | 60 | 95.0% |
| 30 | 29 | 29 | 1 | 1 | 60 | **96.7%** |
| 40 | 29 | 28 | 1 | 2 | 60 | 95.0% |

The performance comparison between k-NN, SVM, and RFC algorithms. For the k-NN algorithm, this study chose k-NN with Minkowski and Euclidean distances with the number of neighbors 1 because it has the best performance compared to k-NN with neighbors 7 and 9 parameters. As for the SVM algorithm, this study chose SVM with LINEAR kernel because it has the best performance among SVM with parameters. On the other hand, RFC with estimators 30 has the best performance among RFCs with other parameters.

Table 7 shows that the KNN algorithm perform better than SVM and RFC in classifying crab molting and non-molting. KNN has an accuracy of 98.3%, a precision of 100.0%, a recall of 96.8%, and F1 score of 100.0%, therefore the KNN Classifier is recommended to be used in crab molting identification on soft shell crab farms.

TABLE VII. COMPARISON OF THE PERFORMANCE OF THE K-NN, SVM, AND RFC ALGORITHMS

| **Machine Learning Algorithms** | **Comparison of the performance** | | | |
|---|---|---|---|---|
| | ***Acuracy*** | ***Precision*** | ***Recall*** | ***F1 Score*** |
| k-NN Minkowski and Euclidean 1 | 98.3% | 100.0% | 96.8% | 100.0% |
| SVM LINEAR | 96.7% | 96.7% | 96.7% | 96.7% |
| RFC Estimators 30 | 96.7% | 96.7% | 96.7% | 96.7% |

## IV. CONCLUSION

This study aims to identify crab molting in the box using machine learning classifiers. We have collected crab image data, conducted data pre-processing, identified crab molting with classification algorithms, and performance evaluation. We compare three machine learning classifiers, namely KNN, SVM, and RFC, to find the best classifiers for crab molting identification. The experiments results show that the KNN algorithm perform better than SVM and RFC in classifying crab molting and non-molting. KNN has an accuracy of 98.3%, a precision of 100.0%, a recall of 96.8%, and F1 score of 100.0%, therefore the KNN Classifier is recommended to be used in crab molting identification on soft shell crab farms.

## ACKNOWLEDGEMENT

This work was supported by The Telecommunications Advancement Foundation, Japan.

## REFERENCES

[1] BPS, Mangrove Crab Production Data. Jakarta, 2016.
[2] H. Kudsiah, S. W. Rahim, M. A. Rifa'i, and Arwan, "Demonstration of Development of Soft Shell Crab Cultivation in Salemba Village, Ujung Loi District, Bulukumba Regency, South Sulawesi," J. Panrita Abdi Univ. Hasanuddin, vol. 2, no. 2, pp. 151–164, 2018, [Online]. Available: https://journal.unhas.ac.id/index.php/panritaabdi/issue/view/518.
[3] Y. Fujaya, "Growth and molting of mud crab administered by different doses of vitomolt," J. Indones. Aquac., vol. 10, no. 1, pp. 24–28, 2011.
[4] R. Ario, A. Djunaedi, I. Pratikto, P. Subardjo, and F. Farida, "Differences in Mutilation Methods on Molting Time of Scylla serrata," Bul. Oseanografi Mar., vol. 8, no. 2, p. 103, 2019, doi: 10.14710/buloma.v8i2.24886.
[5] N. A. Yushinta Fujaya, Siti Aslamyah, Letty Fudjaja, Soft Crab Cultivation and Business: Stimulation of Molting With Spinach Extract. Surabaya: Firstbox Media, 2019.
[6] F. F. Ferdiansyah, B. Rahmat, and I. Yuniar, "Classification and Recognition of Fish Objects Using the Support Vector Machine (Svm) Algorithm," J. Inform. dan Sist. Inf. ( JIFoSI ), vol. 1, no. 2, pp. 522–528, 2020.
[7] M. Ramadhani and D. H. Murti, "Fish Classification Using Oriented Fast and Rotated Brief (Orb) and K-Nearest Neighbor (Knn)," JUTI J. Ilm. Teknol. Inf., vol. 16, no. 2, p. 115, 2018, doi: 10.12962/j24068535.v16i2.a711.
[8] M. M. M. Fouad, H. M. Zawbaa, N. El-Bendary, and A. E. Hassanien, "Automatic Nile Tilapia fish classification approach using machine learning techniques," 13th Int. Conf. Hybrid Intell. Syst. HIS 2013, pp. 173–178, 2014, doi: 10.1109/HIS.2013.6920477.
[9] L.-D. Quach, L. Q. Hoang, N. D. Trung, and C. N. Nguyen, "Towards Machine Learning Approaches To Identify Shrimp Diseases Based on Description," 2020, doi: 10.15625/vap.2019.00063.

[10] R. Ali, M. M. Yusro, M. S. Hitam, and M. Ikhwanuddin, “Machine Learning With Multistage Classifiers For Identification Of Of Ectoparasite Infected Mud Crab Genus Scylla,” Telkomnika (Telecommunication Comput. Electron. Control., vol. 19, no. 2, pp. 406–413, 2021, doi: 10.12928/TELKOMNIKA.v19i2.16724.
[11] F. Muwardi and A. Fadlil, “Image Processing-Based Flower Recognition System and Distance Classifier,” J. Ilm. Tek. Elektro Komput. dan Inform., vol. 3, no. 2, pp. 124–131, 2017.
[12] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, vol. 1, pp. 886–893.
[13] Mustakim and G. Oktaviani F, “Algoritma K-Nearest Neighbor Classification As an Achievement Predicate Prediction System Student,” vol. 13, no. 2, pp. 195–202, 2016.
[14] A. Pandey and A. Jain, “Comparative Analysis of KNN Algorithm using Various Normalization Techniques,” Int. J. Comput. Netw. Inf. Secur., vol. 9, no. 11, pp. 36–42, 2017.
[15] E. Ahishakiye, E. O. Omulo, D. Taremwa, and I. Niyonzima, “Crime prediction using Decision Tree (J48) classification algorithm,” Int. J. Comput. Inf. Technol., vol. 06, no. 03, pp. 188–195, 2017.
[16] T. B. Sasongko, “Comparison and Performance Analysis of SVM and PSO-SVM Algorithm Models (Case Study of Path Classification high school interest),” J. Tek. Inform. dan Sist. Inf., vol. 2, no. 2, pp. 244–253, 2016.
[17] A. Vijayakumar and A. S. Mahesh, ‘Quality assessment of ground water in pre and post-monsoon using various classification technique’, Int. J. Recent Technol. Eng., vol. 8, no. 2, pp. 5996–6003, 2019, doi: 10.35940/ijrte.B3709.078219.
[18] Y. Xu, X. Zhao, Y. Chen, and Z. Yang, “Research on a mixed gas classification algorithm based on extreme random tree,” Appl. Sci., vol. 9, no. 9, 2019.
[19] Y. Zhang and J. T. Yao, ‘Gini objective functions for three-way classifications’, Int. J. Approx. Reason., vol. 81, pp. 103–114, 2017, doi: 10.1016/j.ijar.2016.11.005.
[20] M. Bader-el-den and M. M. Gaber, ‘GARF : Towards Self-optimised Random Forests GARF : Towards Self-optimised Random Forests’, no. November 2012, 2014, doi: 10.1007/978-3-642-34481-7.
[21] DOKMANIC, I., PARHIZKAR, R., RANIERI, J. AND VETTERLI, M., 2015. Euclidean Distance Matrices: Essential Theory, Algorithms and Applications. IEEE Signal Processing Magazine, [online] 32(6), pp.12–30. Available at:¡http://arxiv.org/abs/1502.07541¿ [Accessed 29 Dec. 2020].
[22] ÇOLAKOĞLU, H.B., 2019. A generalization of the Minkowski distance and a new definition of the ellipse. [online] Available at:¡http://arxiv.org/abs/1903.09657¿ [Accessed 29 Dec. 2020].
[23] Patil, N. M., & Nemade, M. U. (2017). Music Genre Classification Using MFCC , K-NN and SVM Classifier. International Journal of Computer Applications, 4(2), 43–47