International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol :Vol. 02, Issue 02, May 2021
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Classification of Fish Species with Image Data Using K-Nearest Neighbor

1stKaharuddin, 2ndEka Wahyu Sholeha
[1]Teknik Perangkat Lunak, [2]Teknologi Informasi
[1]Universitas Universal, [2]Politeknik Negeri Tanah Laut
[1]Batam, Indonesia, [2]Tanah Laut, Indonesia
[1]kahar.osvaldo@gmail.com,[2]ekawahyus@politala.ac.id

Abstract— Classification is a technique that many of us encounter in everyday life, classification science is also growing and being applied to various types of data and cases in everyday life, in computer science classification has been developed to facilitate human work, one example of its application is to classify fish species in the world, the number of fish species in the world is very much so that there are still many people who are sometimes confused to distinguish them, therefore in this study a study will be conducted to classify fish species using the K-Nearest Neighbor Method. 4 types of fish, all data totaling 160 data. The purpose of this study was to test the K-Nearest Neighbor method for classifying fish species based on color, texture, and shape features. Based on the test results, the accuracy value of the truth is obtained using the value of K = 7 with a percentage of the truth of 77.50%, the second-highest accuracy value is the value of K = 10, namely 76.88%. Based on the results of this study, it can be concluded that the K-Nearest Neighbor method has a good enough ability to classify, but it can be done by adding variables or adding more amount of data, and using other types of fish.

Keywords : K-Nearest Neighbor, Classification, Fish, Feature Image.

I. INTRODUCTION

The ocean area is 361 million km2 and the land area is 149 million km2 so that the ocean area is 71% and the land area is 29% of the earth's surface area. The extent and location of the oceans consist of the Ocean (Ocean), the edge sea, the Inland sea / Mediterranean sea [1].

Fish are an important part of biodiversity and one of the most widespread organisms in the world. Recently categorized into 6 classes, 62 orders, 540 fish families, and about 27,683 fish species [2,3]. There are many types of fish in this world, of course, there are many types of fish that have the same shape, color, and even size. Morphological identification has succeeded in describing nearly one million species that exist on earth by classification and species identification [4,5]. Species classification has four parts. First, differences in individual, gender, geography, phenotypic plasticity, and genetic variability can lead to misclassification [6]. Second, there is ecological damage to the environment and human activities that cause damage to the fishery environment, making it difficult to collect fish species [7,8]. Third, some fish show different shapes, patterns, colors, sizes, even though they belong to the same species. Finally, it takes taxonomic knowledge to diagnose errors in classification [9].

From the brief explanation above, the authors are interested in researching fish classification using the KNN method. 4 types of fish will be classified, Black Sea Spart, Gilt Head Bream, Horse Mackerel, and Red Mullet. These fish live in the high seas, so there are still many who do not understand these types of fish. With the amount of data as much as 160 data sourced from Kaggle, and using 1024 x 768 pixels.

II. RESEARCH METHODS

Several studies that have been conducted using the K-Nearest Neighbor algorithm, such as that conducted by Kaharudin, et al. (2019) conducted a study on the classification of types of spices in Indonesia based on shape, color, and texture feature using the K-Nearest Neighbor algorithm. accuracy reached 84% using 7 test scenarios [10].

Research by Andayani, et al. (2018) used three types of fish in the Scombridae family which were classified using the Neural Probabilistic Network method with an accuracy rate of 89.65% using 112 training data images and 29 image data testing [11].

Research by Montalbo, et al. (2019) conducted a study that aimed to classify fish species on the island of Verde using the Deep Convolutional Neural Network (DCNN) model that achieved an accuracy of 99%. Enlarged images are flipped, rotated, cropped, enlarged, and shifted to provide some powerful features for its accuracy classification [12].

Research by Alsmadi, et al. (2020) conducted survey research on fish classification techniques. This survey also reviewed the use of databases such as Fish4-Knowledge (F4K), knowledge databases, and Global Information System (GIS) on Fishes and other FC databases. The study of preprocessing method of sender extraction technique and classifier was collected from recent work to increase understanding of the characteristics of pre-processing methods, feature extraction techniques, and classifiers to guide the direction of research [13].

Research by Jin, et al (2021) conducted a study with a classification approach that combines Elastic Net-Stacked Autoencode (EN-SAE) with Kernel Density Estimation (KDE) with the name ESK-model, which is proposed based on DNA coding. Whereas ESK models can accurately correlate fish from different families based on DNA [14].

Research from Adebayo, et al (2016) classified fish based on physical form processed from images, feature extraction, and classification methods. Fish feature vectors are obtained from Single Value Decomposition (SVD) extracted from fish images. Performed the test using an Artificial Neural

Network (ANN) with 36 fish images and got an accuracy of 94% [15].

2.3 Research Methodology

The research method consists of 6 stages, namely: First,

Literature study to find literature and references as a reference for conducting research both from books, journals, proceedings, and others. Furthermore, data collection was carried out by searching for the dataset to be used in this study, the data used were sourced from Kaggle.Com [16]. After the dataset in the form of an image has been collected, before processing, the data must then be cleaned first at the image pre-processing stage, at this stage the background is removed to black then adjusts the overall dimensions of the image to 1024 × 768 pixels, this is intended so that At the time of extracting the value in the image, only the value of the fish object is extracted, while the purpose of adjusting the dimensions of the image is to ensure the extraction value is taken from the same number of pixels. After pre-processing the image, then enter the feature extraction stage using an application made using MATLAB. At this stage, the color feature values are taken which consist of RGB, Texture consisting of Contrast, Correlation, Energy, and Homogeneity. Feature Form consisting of Eccentricity and Metric. After the features are successfully extracted into the CSV file, then they enter the Analysis stage, at this stage testing using the WEKA application, the testing phase using the 10 Fold Cross Validation Evaluation method, while testing is carried out by using several K values or the number of closest neighbors, including K = 1, K = 3, K = 5, K = 7, K = 9, and K = 10. After conducting the test, the final level of accuracy of the test results can be obtained, then a conclusion can be drawn.

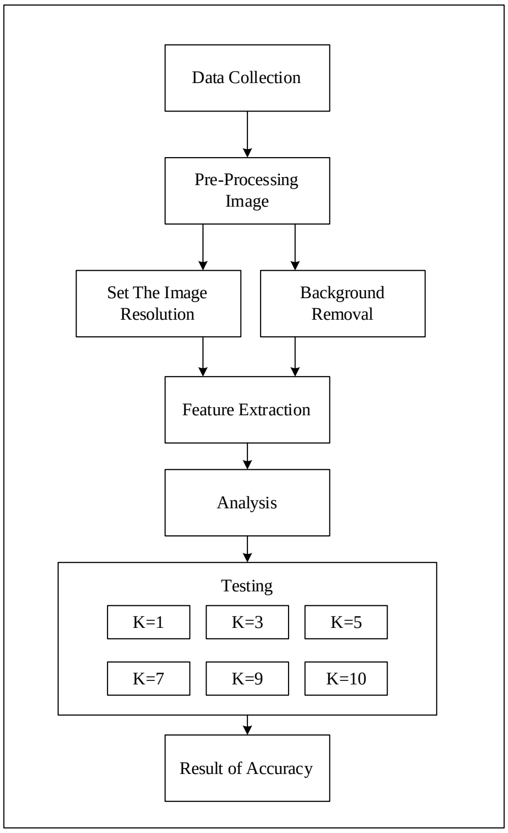


Figure 1. Research Flow

III. RESULT AND ANALYSIS

3.1 Dataset

The dataset used is an image of 160 data, each class or type of fish has 40 data. The image used has dimensions of 1024 × 768 pixels.

Here are some examples of the data used. Image data is taken using a digital camera with a view of 30-50 cm from the object.

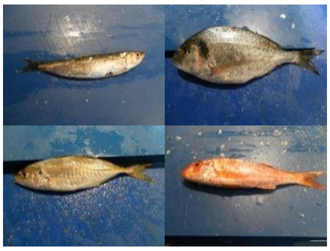


Figure 2. Initial Dataset

3.2 Pre-processing

The pre-processing stage starts from removing the background in the image, with the aim that at the time of extracting the image value will not be influenced by the background of the object so that the analysis is expected to be more accurate.

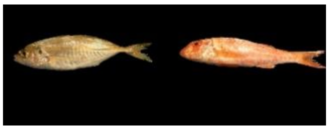
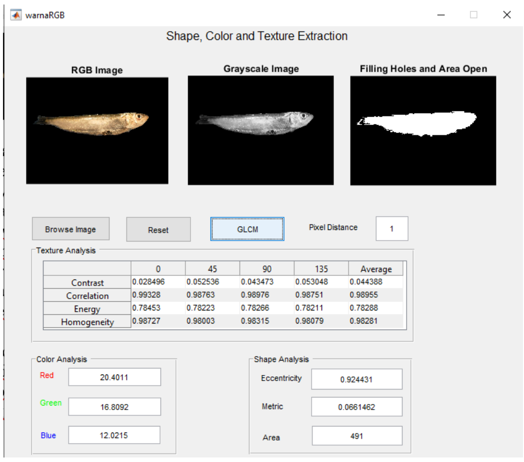


Figure 3. Dataset after going through the preprocessing process

3.3 Feature Extraction

The pre-processing stage starts from removing the background in the image, with the aim that at the time of extracting the image value will not be influenced by the background of the object so that the analysis is expected to be more accurate.



warnaRGB

Shape, Color and Texture Extraction

Pixel Distance: 1

Texture Analysis

	0	45	90	135	Average
Contrast	0.028496	0.052536	0.043473	0.053048	0.044388
Correlation	0.99328	0.98763	0.98976	0.98751	0.98955
Energy	0.78453	0.78223	0.78266	0.78211	0.78288
Homogeneity	0.98727	0.98003	0.98315	0.98079	0.98281

Color Analysis

Red: 20.4011
Green: 16.8092
Blue: 12.0215

Shape Analysis

Eccentricity: 0.924431
Metric: 0.0661462
Area: 491

Figure 4. Data Extraction Process Using Applications

In the feature extraction process, the features taken are color, texture, and shape features. Color features are taken using the RGB formula, texture features are taken using the GLCM (Gray Level Co-Occurrence Matrix) method, the first color image must be converted into a greyscale image first and then the GLCM value can be taken, the value taken for texture features is Contrast, Correlation, Energy and Homogeneity. The shape features are taken using eccentricity and metric formulas. To retrieve the features of the grayscale image, it must be converted into a Filling Holes and Area Open image first.

3.4 Analysis Using K-Nearest Neighbor

To classify fish species using K-Nearest Neighbor, modeling is necessary first, to provide an overview of how the classification process is carried out.

The following is an example of the application of the K-Nearest Neighbor to calculate the types of fish. In this example, it will only use 4 data from 4 types of fish, then only 3 variables will be used, namely red, contrast and eccentricity.

Tabel 1. Sample K-NN Training Data

No	Class	Red	Contrast	Eccentricity
1	Black Sea Spart	21.8867	0.064726	0.986765
2	Gilt Head Bream	37.6448	0.14819	0.974289
3	Horse Mackerel	29.8805	0.094623	0.973728
4	Red Mullet	21.4115	0.055717	0.976647

Tabel 2. Sample K-NN Testing Data

Class	Red	Contrast	Eccentricity
?	20.4011	0.044388	0.924431

By using existing sample data, predictions can be done with the following steps:

1. Determine the number of closest neighbors (K value). In this example, the classification results will be taken based on the value of K = 1, meaning that it is based on 1 number of closest neighbors
2. Calculating the distance between training data and testing data, distance calculation can be used with several methods, in this study we will use the euclidean distance, calculate the proximity of the testing data to all existing testing data. The formula used can be seen in equation 3:

$$D\,(x,y) = \sqrt{\Sigma_{k-1}^{n}(x_k - y_k)^2} \qquad (1)$$

First, calculate the euclidean distance testing data with the first training data:

$$D = \sqrt{(21.8867 - 20.4011)^2 + (0.064726 - 0.044388)^2 + (0.986765 - 0.924431)^2}$$
=1,46737

Furthermore, the calculation of the second data is carried out:

$$D = \sqrt{(37.6448 - 20.4011)^2 + (0.14819 - 0.044388)^2 + (0.974289 - 0.924431)^2} = 17{,}24832$$

Then the calculation of the third data is carried out:

$$D = \sqrt{(29.8805 - 20.4011)^2 + (0.094623 - 0.044388)^2 + (0.973728 - 0.924431)^2}$$
=9.47966

And finally, do the calculations on the fourth training data.

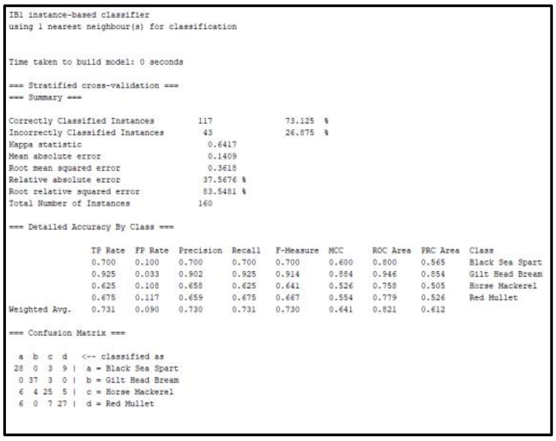
$$D = \sqrt{(21.4115 - 20.4011)^2 + (0.055717 - 0.044388)^2 + (0.976647 - 0.924431)^2}$$
= 1,07762

Based on the results of the calculation of the Euclidean Distance testing data on the four training data above, it is found that the smallest distance value is the fourth data, namely 1.07762, so if using the value of K = 1 it can be concluded that the testing data is classified as a type of Red Mullet fish.

3.5 Implementasi

Testing is done using the WEKA application, for the classification process, WEKA provides fairly complete information, testing is carried out using K = 1, K =, K = 3, K = 5, K = 7, K = 9 and K = 10.

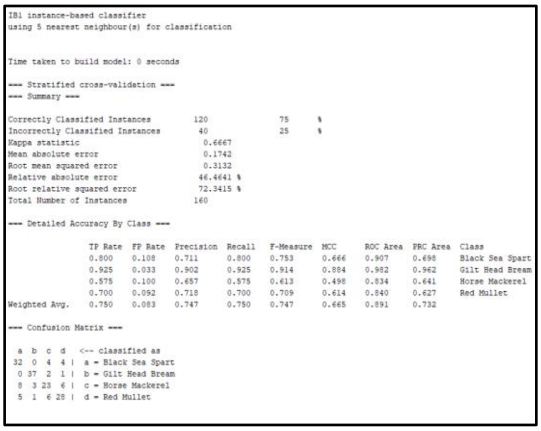
Following are some of the test results using WEKA can be seen in the image below.



```
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         117               73.125  %
Incorrectly Classified Instances        43               26.875  %
Kappa statistic                          0.6417
Mean absolute error                      0.1409
Root mean squared error                  0.3618
Relative absolute error                 37.5676 %
Root relative squared error             83.5481 %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.700    0.100    0.700      0.700    0.700      0.600    0.800     0.565     Black Sea Spart
                 0.925    0.033    0.902      0.925    0.914      0.884    0.946     0.854     Gilt Head Bream
                 0.625    0.108    0.658      0.625    0.641      0.526    0.758     0.505     Horse Mackerel
                 0.675    0.117    0.659      0.675    0.667      0.554    0.779     0.526     Red Mullet
Weighted Avg.    0.731    0.090    0.730      0.731    0.730      0.641    0.821     0.612

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 28  0  3  9 |  a = Black Sea Spart
  0 37  3  0 |  b = Gilt Head Bream
  6  4 25  5 |  c = Horse Mackerel
  6  0  7 27 |  d = Red Mullet
```

Figure 5. The test results use the value of K = 1

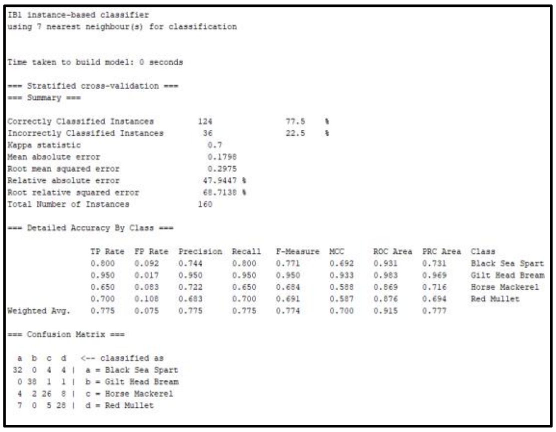
Tests using the value of K = 1 have an accuracy of the truth of 73.13%



```
IB1 instance-based classifier
using 5 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         120               75      %
Incorrectly Classified Instances        40               25      %
Kappa statistic                          0.6667
Mean absolute error                      0.1742
Root mean squared error                  0.3132
Relative absolute error                 46.4641 %
Root relative squared error             72.3415 %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.800    0.108    0.711      0.800    0.753      0.666    0.907     0.698     Black Sea Spart
                 0.925    0.033    0.902      0.925    0.914      0.884    0.982     0.962     Gilt Head Bream
                 0.575    0.100    0.657      0.575    0.613      0.498    0.834     0.641     Horse Mackerel
                 0.700    0.092    0.718      0.700    0.709      0.614    0.840     0.627     Red Mullet
Weighted Avg.    0.750    0.083    0.747      0.750    0.747      0.665    0.891     0.732

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 32  0  4  4 |  a = Black Sea Spart
  0 37  2  1 |  b = Gilt Head Bream
  8  3 23  6 |  c = Horse Mackerel
  5  1  6 28 |  d = Red Mullet
```

Figure 6. The test results use the value of K = 5

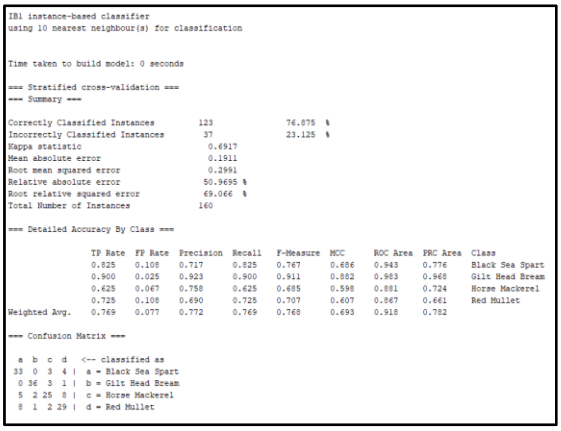
Tests using the value of K = 5 have an accuracy of 75%



```
IB1 instance-based classifier
using 7 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         124               77.5    %
Incorrectly Classified Instances        36               22.5    %
Kappa statistic                          0.7
Mean absolute error                      0.1798
Root mean squared error                  0.2975
Relative absolute error                 47.9447 %
Root relative squared error             68.7138 %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.800    0.092    0.744      0.800    0.771      0.692    0.931     0.731     Black Sea Spart
                 0.950    0.017    0.950      0.950    0.950      0.933    0.983     0.969     Gilt Head Bream
                 0.650    0.083    0.722      0.650    0.684      0.588    0.869     0.716     Horse Mackerel
                 0.700    0.108    0.683      0.700    0.691      0.587    0.876     0.694     Red Mullet
Weighted Avg.    0.775    0.075    0.775      0.775    0.774      0.700    0.915     0.777

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 32  0  4  4 |  a = Black Sea Spart
  0 38  1  1 |  b = Gilt Head Bream
  4  2 26  8 |  c = Horse Mackerel
  7  0  5 28 |  d = Red Mullet
```

Figure 7. The test results use the value of K = 7

Testing using the value of K = 7 has an accuracy of the truth of 77.5%



```
IB1 instance-based classifier
using 10 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         123               76.875  %
Incorrectly Classified Instances        37               23.125  %
Kappa statistic                          0.6917
Mean absolute error                      0.1911
Root mean squared error                  0.2991
Relative absolute error                 50.9695 %
Root relative squared error             69.066  %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.825    0.108    0.717      0.825    0.767      0.686    0.943     0.776     Black Sea Spart
                 0.900    0.025    0.923      0.900    0.911      0.882    0.983     0.968     Gilt Head Bream
                 0.625    0.067    0.758      0.625    0.685      0.598    0.881     0.724     Horse Mackerel
                 0.725    0.108    0.690      0.725    0.707      0.607    0.867     0.661     Red Mullet
Weighted Avg.    0.769    0.077    0.772      0.769    0.768      0.693    0.918     0.782

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 33  0  3  4 |  a = Black Sea Spart
  0 36  3  1 |  b = Gilt Head Bream
  5  2 25  8 |  c = Horse Mackerel
  8  1  2 29 |  d = Red Mullet
```

Figure 8. The test results use the value of K = 10

Tests using the value of K = 10 have an accuracy of the truth of 76.88%

The following can be seen a diagram of the percentage of the truth of all test results

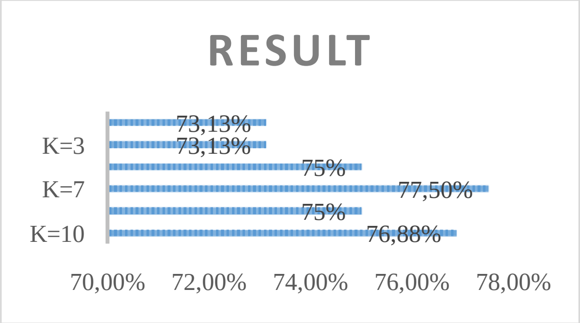


Figure 9. Test results diagram

Based on the diagram above, it can be seen that the value of K with the highest accuracy of truth is K = 7 with a percentage of 77.50%, then K = 10 with a percentage of the truth of 76.88%.

VI. CONCLUSIONS

Based on the above test, it can be concluded that the K-Nearest Neighbor method has a fairly good ability to classify fish types based on color, texture, and shape, with an accuracy value of 77.50%, further research is expected to be able to use other features or use classification methods. others and use more training data so that the accuracy value is better.

REFERENCES

[1] Arifin, Muhammad Zainul, dkk (2019) *Modul Massa Daratan dan Lautan* http://www.pusdik.kkp.go.id/

[2] Fautin D, Dalton P, Incze LS, Leong J-AC, Pautzke C, Rosenberg A, et al. An Overview of Marine Biodiversity in United States Waters. PloS one.2010;5(8): e11914. doi: 10.1371/journal.pone.0011914

[3] Xu L, Wang X, Van Damme K, Huang D, Li Y, Wang L, et al. Assessment of fish diversity in the South China Sea using DNA taxonomy. Fisheries Research. 2021; 233:105771. doi: 10.1016/j.fishres.2020.105771.

[4] Thu PT, Huang WC, Chou TK, Van NQ, Liao TY. DNA barcoding of coastal ray-finned fishes in Vietnam. PloS one. 2019;14(9):e0222631. doi: 10.1371/journal.pone.0222631.

[5] Hebert PD, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proceedings Biological sciences. 2003;270(1512):313-21. doi: 10.1098/rspb.2002.2218. PubMed PMID: 12614582; PubMed Central PMCID: PMC1691236.

[6] Ramirez JL, Rosas-Puchuri U, Canedo RM, Alfaro-Shigueto J, Ayon P, Zelada-Mazmela E, et al. DNA barcoding in the Southeast Pacific marine realm: Lowcoverage and geographic representation despite high diversity. PloS one. 2020;15(12): e0244323. doi: 10.1371/journal.pone.0244323. PubMed PMID: 33370342; PubMed Central PMCID: PMC7769448.

[7] Liang H, Meng Y, Luo X, Li Z, Zou G. Species identification of DNA barcoding based on COI gene sequences in Bagridae catfishes. Journal of Fishery Sciences of China. 2018;25(4):772. doi: 10.3724/sp.j.1118.2018.18036.

[8] Xu L, Van Damme K, Li H, Ji Y, Wang X, Du F. A molecular approach to the identification of marine fish of the Dongsha Islands (South China Sea). Fisheries Research. 2019; 213:105-12. doi: 10.1016/j.fishres.2019.01.011

[9] Ren BQ, Xiang XG, Chen ZD. Species identification of Alnus (Betulaceae) using nrDNA and cpDNA genetic markers. Mol Ecol Resour. 2010;10(4):594-605. doi: 10.1111/j.1755-0998.2009.02815. x. PubMed PMID: 21565064.

[10] Kaharudin, Kusrini, Wati Vera, dkk. (2019) Classification of Spice Types UsingK-Nearest Neighbor Algorithm. International Conference on Information and Communications Technology. Doi: 10.1109/ICOIACT46704.2019.8938515

[11] Andayani U, Wijaya Alex, dkk. Fish Species Classification Using Probabilistic Neural Network, Journal of Physics: Conference Series, The 3rd International Conference on Computing and Applied Informatics 2018 IOP Conf. Series: Journal of Physics: Conf. Series 1235 (2019) 012094 IOP Publishing doi.org/10.1088/1742-6596/1235/1/012094

[12] Montablo Francis Jesmar P, Hernandez Alexander A. Classification of Fish Species with Augmented Data using Deep Convolutional Neural Network. 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET), 7 October 2019, Shah Alam, Malaysia

[13] Alsmadi Mutasem K, Almarashdeh Ibrahim. A survey on fish classification techniques. Journal of King Saud University – Computer and Information Sciences. doi.org/10.1016/j.jksuci.2020.07.005

[14] Jin Lina, Yu Jiong, dkk. A deep learning model for fish classification base on DNA barcode. Doi.org/10.1101/2021/02/15/431244

[15] Adebayo Daramola S, Olumide Omololu. Fish Classification Algorithm using Single Value Decomposition. Internation Journal of Innovative Research in Science, Engineering, and Technology. Vol.5, Issue 2, February 2016.

[16] O. Ulucan, D. Karakaya, and M. Turkan. (2020) A large-scale dataset for fish segmentation and classification. In Conf. Innovations Intell. Syst. Appli. (ASYU).

[17] Muslihah, I., Muqorobin, M., Rokhmah, S., & Rais, N. A. R. (2020). Texture Characteristic of Local Binary Pattern on Face Recognition with PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS. *International Journal of Computer and Information System (IJCIS)*, *1*(1).

[18] Muqorobin, M., Rokhmah, S., Muslihah, I., & Rais, N. A. R. (2020). Classification of Community Complaints Against Public Services on Twitter. *International Journal of Computer and Information System (IJCIS)*, *1*(1).

[19] Muqorobin, M., Kusrini, K., Rokhmah, S., & Muslihah, I. (2020). Estimation System For Late Payment Of School Tuition Fees. *International Journal of Computer and Information System*, *1*(1), 341475.