Freshwater Fish Classification Based on Image Representation Using K-Nearest Neighbor Method

Suwarsito[1], Hindayati Mustafidah[2*], Tito Pinandita[2], Purnomo[3]

[1]Aquaculture, Universitas Muhammadiyah Purwokerto, Indonesia
[2]Informatics Engineering, Universitas Muhammadiyah Purwokerto, Indonesia
[3]UPTD-BIAT Kutasari, DKPP Purbalingga, Central Java, Indonesia

[1]suwarsito@ump.ac.id, [2]titop@ump.ac.id, [3]yusufizazpurnomo@gmail.com,
[2*]corr_author: h.mustafidah@ump.ac.id

Abstract - Indonesia is a maritime and agricultural country with enormous world fishery potential. The large variety of fish is often confusing for ordinary people in recognizing types of fish, especially freshwater fish. It was stated that the types of freshwater fish often consumed by the Indonesian people are *bawal* (pomfret), *betutu, gabus* (cork), *gurame* (carp), *mas* (goldfish), *lele* (catfish), *mujaer* (tilapia), *patin* (asian catfish), *tawes*, and *nila* (tilapia nilotica). Some fish types have similar shapes, so it is tricky to tell them apart. Meanwhile, in the digitalization era today, Artificial Intelligence (AI)-based technology has become a demand in all areas of life. It is overgrowing, not apart from the fisheries sector. Therefore, in this study, the K-Nearest Neighbor (KNN) method was applied as one of the methods in AI to identify and classify freshwater fish species based on their images. The KNN method classifies new data into specific classes based on the distance between the new data and the closest k data through the learning process. This KNN model is built by preparing the dataset stages, separating the dataset into data-train and data-test with a ratio of 70%:30%, then building and testing the model. The dataset is freshwater fish images, totaling 100 images from 10 freshwater fish types. Model testing is done by measuring performance using a confusion matrix. Based on the test results, the model has an accuracy performance of 70%. Thus, KNN can be used as a model to identify freshwater fish species based on their image.

Keywords: KNN; image; confusion matrix; freshwater fish

I. INTRODUCTION

Indonesia is a maritime and agricultural country with immense world fishery potential; both capture fisheries and aquaculture [1]. However, fishery growth during the Covid-19 pandemic decreased by 0.73% compared to 2019, which grew by 5.73% [2]. This predicate cannot be separated from the diversity of fish species, both freshwater, brackish and marine fish. As an agricultural country, many people cultivate freshwater fish. There are 28 species of freshwater fish in Indonesia that are known to be identified [3], but only 10 are cultivated for consumption [4]. The fish are *bawal* (pomfret), *betutu*, *gabus* (cork), *183sian183* (carp), *mas* (goldfish), *lele* (catfish), *mujaer* (tilapia), *patin* (183sian catfish), *tawes*, and *nila* (tilapia nilotica). Fish have particular shapes, sizes, colors, and textures that differ from fish to another. These characteristics are often referred to as fish images. The problem of recognizing fish species based on their appearance is more complex than identifying human facial images because fish images are more varied. The recognition of fish species is generally done manually using eye observation. This makes it difficult to distinguish the types of fish for the general public. Recognizing the type of fish based on its image requires knowledge of the process of object classification.

The types of freshwater fish were presented by [5-7]. The following describes 10 types of freshwater fish as objects in this study and examples of fish images as presented in Fig. 1.

Along with the era of the Industrial Revolution (RI) 4.0, known as the Digital Revolution, the use of computer technology is inevitable [8], [9]. The RI 4.0 marked by the proliferation of computers and the automation of records in all fields. One of the signs is the implementation of Artificial Intelligence (AI). Machine learning (ML), as one of the methods in the field of AI (Ahmad, 2017), is a machine developed to learn by itself without direction from its users. How to know ML is to do training. One of the functions of ML is that it can be used to classify objects based on their images. Along with the current COVID-19 pandemic, it is undeniable that almost all work is done using computer technology [10]. In this era, massive changes occur in agriculture,

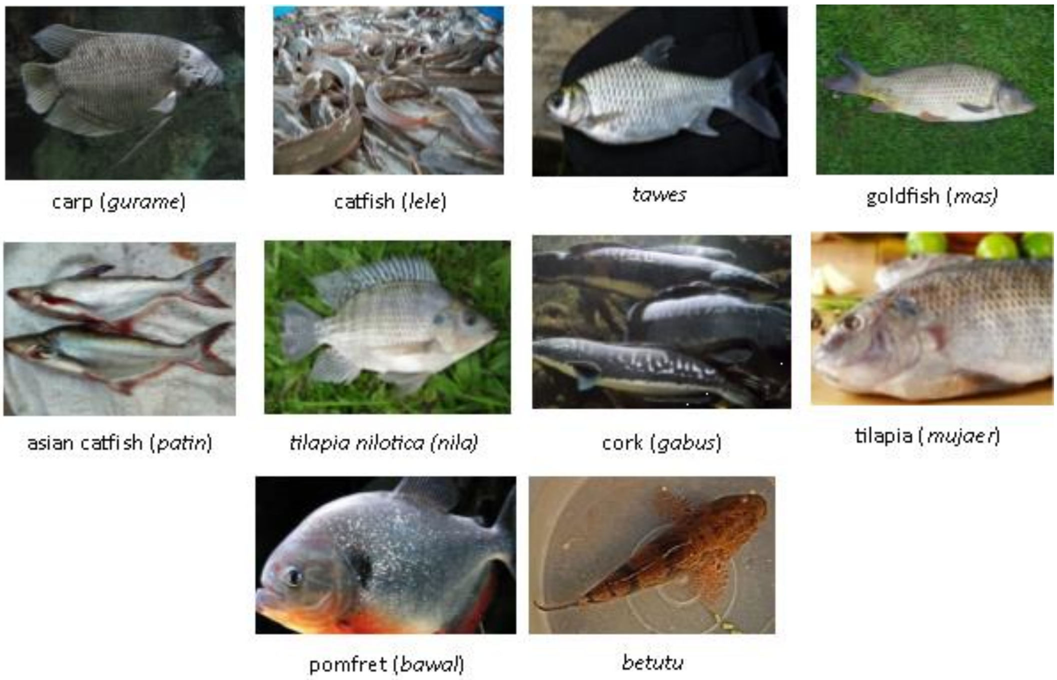


Fig. 1 Most consumed freshwater fish

manufacturing, mining, transportation, and technology and profoundly impact social, economic, and cultural conditions in the world [11].

Machine learning (ML) is interpreted as developing a machine, in this case, software, which can learn to perform specific tasks [12]. One of the tasks that ML can do is to recognize objects from digital images. One of ML's most widely used classification algorithms is K-Nearest Neighbors (KNN) [13]. In principle, KNN will classify new data into specific classes based on the distance of the new data from the closest k data through the learning process (Fig. 2).

Learning data is depicted in a multidimensional space, with each dimension representing each data feature. The new data classification is done by looking for the labels of the k nearest neighbors. Most labels that appear become new data labels. If k = 1, the new data is labeled with the nearest neighbor label. The closest distance calculation usually uses the Euclidean distance (1).

$$\mathrm{D}(\mathrm{p},\mathrm{q}) = \mathrm{d}(\mathrm{q},\mathrm{p}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1)$$

where p and q are the two position points to be distanced. In other words, the Euclidean distance between points p and q is the distance of the line segment connecting them [14].

The development of research in the AI field with the fish domain has been carried out by several researchers, including 2015 [15] has implemented a fuzzy database of the Tahani model to determine the type of fish feed based on the price and nutritional content of feed raw materials. Ref. [16] and [17] using fuzzy logic and expert system to determine the price of fish feed based on feed formulas made from local ingredients around the environment. The recommendation system for determining fish species based on environmental parameters of cultivated land using fuzzy Mamdani was developed by [18] in 2016 and continued by [19] in 2017 in the form of the development of a decision support system to determine freshwater fish species for rearing business. Furthermore [20] conducted research using an expert system to determine the type of cultured fish based on water quality, and [21] determined the kind of freshwater fish for enlargement using Multicriteria Decision Making (MCDM). Meanwhile, in 2020 [22] identified fish species using CNN (Convolution Neural Networks). Related to fish image research that has been done [23] and [24], namely classifying fish with formalin, using different methods. In addition, [25] classifies tuna images based on shape descriptors and saline points.

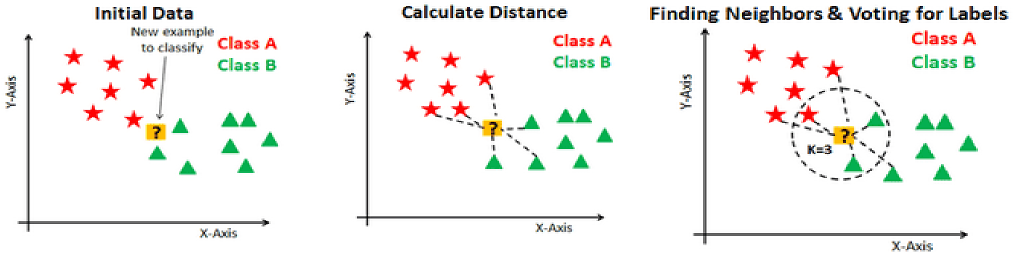


Fig. 2 Illustration of the KNN method

Based on the description that has been submitted, in this study, a machine learning-based model will be built to classify freshwater fish species based on their image. The type of freshwater fish that is classified is the consumption fish which consists of 10 types as have been mentioned. The urgency of this research is to produce a machine learning-based model that can be developed into an appropriate technology in the form of software that can help the community/users to identify the exact type of freshwater fish. This research produces a classification model for freshwater fish species using KNN method.

II. METHOD

The research stages of developing a model based on the KNN method are presented in Fig. 3. The dataset used in this study is in the form of image characteristics of freshwater fish and is taken from various sources, namely books, journal and proceedings articles, and the internet. At this stage, the dataset is separated into 2 parts, training data and testing data. Training data is used to train the model, while testing data is used to determine the performance of the trained model. The ratio between training and testing data is 70%:30%.

The testing phase is carried out after conducting training on the data train. Model testing is done using a testing data. Furthermore, the performance of the model is measured and analyzed using a confusion matrix with an explanation, as shown in Fig. 4.

Some of the symbols in Fig. 4 are described as follows:

- True Positive (TP): the number of data that has a Positive value and is predicted to be correct as Positive.
- False Positive (FP): the amount of data that is Negative but is predicted to be Positive.
- False Negative (FN): the amount of data that is Positive but is predicted to be Negative.
- True Negative (TN): the number of data that is Negative and is predicted to be true as Negative.

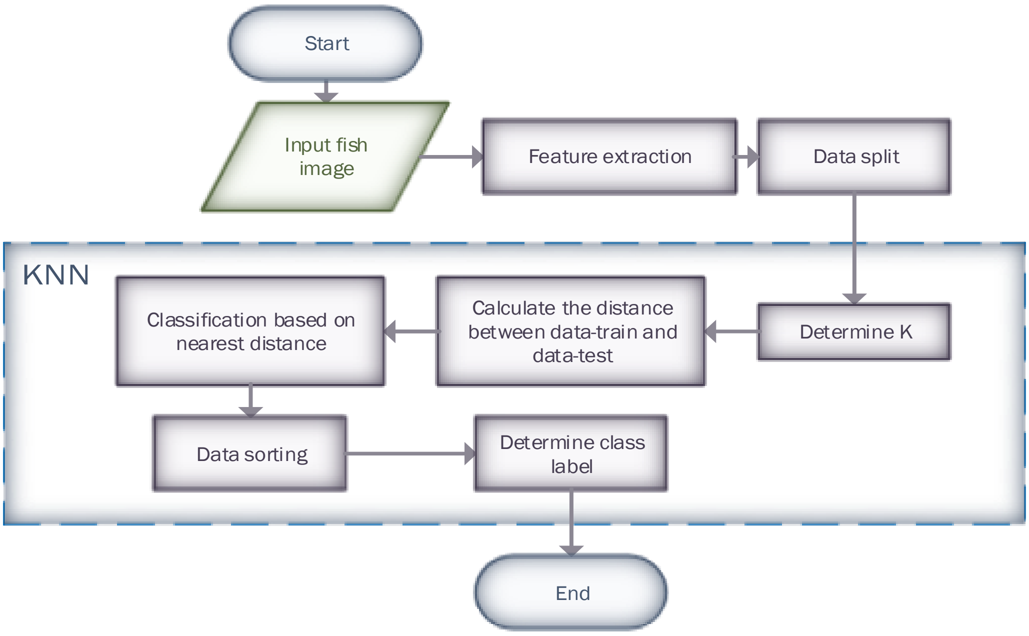
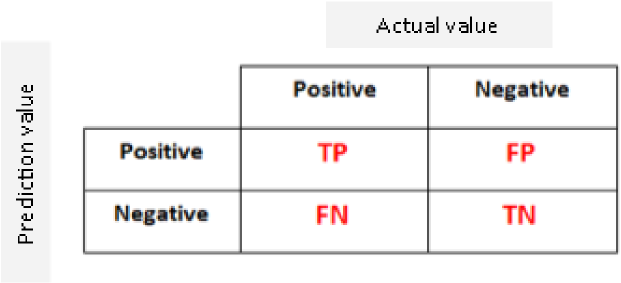


Fig. 3 KNN method development flow



		Actual value	
		Positive	Negative
Prediction value	Positive	TP	FP
	Negative	FN	TN

Fig. 4 Illustration of confusion matrix analysis

Performance analysis using a confusion matrix has 4 indicators: accuracy, precision, recall, and F1_Score. The accuracy value is obtained from the number of positive data predicted to be positive and negative data that is expected to be negative divided by the total data in the dataset as in (2). At the same time, precision is the probability of a positive predicted case that belongs to the positive category case using (3). Meanwhile, recall is the probability of a case with a positive category correctly predicted to be positive using (4). Finally, the value of F1_Score, also known as F_Measure, is obtained from the precision and recall results between the predicted category and the actual category as in (5).

$$Accuracy = \frac{TP+TN}{N} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F\_Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \tag{5}$$

with N = total data

III. RESULTS AND DISCUSSION

The classification method used is an algorithm in supervised learning because, in its development, a target in the form of freshwater fish is needed. Supervised learning is a machine learning algorithm that, in its learning process, requires a series of correct input-output examples as a supervisor or coach. This series of input-output samples are then used to train the algorithm to produce an output in the form of an appropriate label when given new data input. In this study, the input is the image of freshwater fish, while the label is the type of fish.

A. Dataset

Research data on freshwater fish images were obtained from various sources, namely books, internet websites, and journal articles—image data obtained as many as 10 kinds for each type of fish, totaling 10. The ten types of fish are pomfret, betutu, cork, carp, mas, catfish, mujaer, catfish, tawes, and tilapia. Thus, the fish image dataset used is 100, divided into a data train of 70 image data (70%) and a data-test of 30 image data (30%). The research dataset is shown in Table I.

B. KNN Model

The KNN model, as shown in Fig. 3, can be explained as follows. Input the model in the form of fish images in the dataset. Each fish image is feature extracted using the GLCM (Gray-Level Cooccurrence Matrix) method. This method is used to extract image features based on their texture. An example of an extracted image feature is presented in Fig. 5. Split data divides the dataset into 70 data-train and 30 data-test.

The process in the KNN model begins with determining the value of k = 3 to take the number of neighbors generated when calculating the distance. The calculation of the distance between the data-test and the data-train was used in the Euclidean method as in (6). An example of the distance calculation using the Euclidean method is presented in Table II.

$$Euclidean = \sqrt{\sum_{i=1}^{n}\left(x_{training_i} - x_{testing_i}\right)^2} \tag{6}$$

C. Model Testing

The classification results using KNN are then evaluated by measuring the model's performance using a confusion matrix. The measurement components include accuracy for all classes, precision, recall, and F1_score for each class. The evaluation results are presented in Fig. 6. The KNN classification using the Euclidean method predicts 24 true positive data out of 30. For false positive prediction data, there are 6 (calculated from the number of positive predictions column) and 6 false negative prediction data (calculated

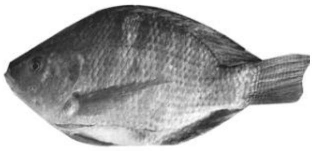


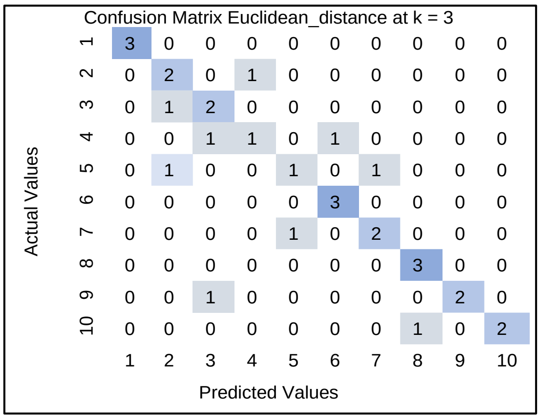
Fig.5 A fish extracted image using GLCM

TABLE I
RESEARCH DATASET

No.	Fish image	Fish type	Class
1.		*nila*	1
2.		*nila*	1
3.		*lele*	2
4.		*lele*	2
5.		*bawal*	3
…			
100.		*betutu*	10

TABLE II
EXAMPLE OF DISTANCE CALCULATION RESULTS USING THE EUCLIDEAN METHOD WITH K = 3

Data-test (i)	Euclidean distance	Neighbor's class	Target class	Class	Information
1	0,2416	9	9	9	True
	0,2687	6			
	0,3270	9			
2	0,2240	7	7	7	True
	0,2812	7			
	0,3335	7			
3	0,2805	2	2	2	True
	0,2980	2			
	0,3336	3			
…					
30	0,4686	6	5	6	False
	0,6690	9			
	0,6715	4			



Confusion Matrix Euclidean_distance at k = 3

Actual Values \ Predicted Values	1	2	3	4	5	6	7	8	9	10
1	3	0	0	0	0	0	0	0	0	0
2	0	2	0	1	0	0	0	0	0	0
3	0	1	2	0	0	0	0	0	0	0
4	0	0	1	1	0	1	0	0	0	0
5	0	1	0	0	1	0	1	0	0	0
6	0	0	0	0	0	3	0	0	0	0
7	0	0	0	0	1	0	2	0	0	0
8	0	0	0	0	0	0	0	3	0	0
9	0	0	1	0	0	0	0	0	2	0
10	0	0	0	0	0	0	0	1	0	2

Fig. 6 The results of the calculation of the confusion matrix in the Euclidean method with k = 3

from the true positive count line). Based on equations (2) to (5), then obtained values as follows:

$$accuracy = \frac{21}{30} x100\%= 70\%$$

$$precision = \frac{9}{9+3} x100\% = 75\%$$

$$recall = \frac{9}{9+6} x\ 100\% = 60\%$$

$$F1\_score \quad = \frac{2\ x\ 75\%\ x\ 60\%}{75\% + 60\%} = 67\%$$

IV. CONCLUSION

As a machine learning method, the K-Nearest Neighbor (KNN) classification can be used to identify freshwater fish species based on their image. The testing results show an accuracy of 70%, which means that the model's performance is quite good. The limitation of this research is that the input image is a single fish image, not in groups. Therefore, it is recommended to continue recognizing or identifying fish species based on images in groups.

ACKNOWLEDGEMENT

Thanks to the Research and Community Service Board (LPPM) Universitas Muhammadiyah Purwokerto for providing financial support and facilities in carrying out this research based on contract no. A.11-III/575-S.Pj./LPPM/XII/2021.

REFERENCES

[1] Wantimpres, *Potensi Perikanan Indonesia*. 2017. Accessed: Nov. 05, 2021. [Online]. Available: *https://wantimpres.go.id/id/potensi-perikanan-indonesia/*

[2] N. A. Romfiz, "Potensi Perikanan, Konsumsi Ikan, dan Kesejahteraan Nelayan," *https://news.detik.com/kolom/d-5521785/potensi-perikanan-konsumsi-ikan-dan-kesejahteraan-nelayan*, 2021. (accessed Nov. 10, 2021).

[3] H. Mustafidah, S. Suwarsito, and E. Puspitasari, "Case-Based Reasoning System to Determine the Types of Fish Farming Based on Water Quality," in *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 2020, pp. 1–5.

[4] PintarPet, "Berkenalan dengan 23 Jenis-Jenis Ikan Air Tawar Populer di Indonesia," *https://petpintar.com/ikan/jenis-jenis-ikan-air-tawar*, 2020. (accessed Nov. 09, 2021).

[5] Surya Mina Farm, "Mengenal Ikan Tawes (Barbonymus Goniono Bleeker)," *https://www.bibitikan.net/mengenal-ikan-tawes-barbonymus-goniono-bleeker/*, 2013. (accessed Nov. 05, 2021).

[6] P. M. D. K. H. P. BADAN KARANTINA IKAN, "Mengenal Ikan Betutu, Si Gabus Malas Berkhasiat Tinggi," *https://kkp.go.id/bkipm/artikel/9051-mengenal-ikan-betutu-si-gabus-malas-berkhasiat-tinggi*, 2019. (accessed Nov. 06, 2021).

[7] ResepKoki, "10 Jenis Ikan Air Tawar di Indonesia Yang Sering Dikonsumsi," *https://resepkoki.id/10-jenis-ikan-air-tawar-di-indonesia-yang-sering-dikonsumsi/*, 2021. (accessed Nov. 07, 2021).

[8] M. E. Auer, D. Guralnick, and I. Simonics, *Teaching and Learning in a Digital World: Proceedings of the 20th International Conference on Interactive Collaborative Learning – Volume 2*. Switzerland: Springer International Publishing, 2018.

[9] M. E. Auer, H. Hortsch, and P. Sethakul, "The Impact of the 4th Industrial Revolution on Engineering Education"*: Proceedings of the 22nd International Conference on Interactive Collaborative Learning (ICL2019) – Volume 2*. Switzerland: Springer International Publishing, 2020.

[10] Kompasiana, "Peran Teknologi Dalam Pandemi Covid-19," *https://www.kompasiana.com/waju27020/5ea90c97d541df0aac6be9c2/peran-teknologi-dalam-pandemi-covid-19*, 2020. (accessed May 30, 2020).

[11] A. D. Syafaati, "Revolusi Industri dari Generasi 1.0 hingga 4.0," *https://www.academia.edu/37491240/REVOLUSI_INDUSTRI_DARI_GENERASI_1.0_HINGGA_4.0*, 2019. (accessed Feb. 19, 2019).

[12] T. P. Trappenberg, *Fundamentals of Machine Learning*. Oxford University Press, 2020. doi: 10.1093/oso/9780198828044.001.0001.

[13] S. Ray, "Commonly used Machine Learning Algorithms (with Python and R Codes)," *https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/*, 2017. (accessed Oct. 20, 2021).

[14] MATH VAULT, "Compendium of Mathematical Symbols," *https://mathvault.ca/hub/higher-math/math-symbols/*, 2020. (accessed Nov. 10, 2021).

[15] M. A. Sofia, H. Mustafidah, and S. Suwarsito, "Basis Data Fuzzy Model Tahani untuk Menentukan Jenis Pakan Ikan Berdasarkan Harga dan Kandungan Gizi Bahan Baku Pakan," *JUITA (Jurnal Inform.*, vol. III, no. 3, pp. 143–155, 2015, doi: http://dx.doi.org/10.30595/juita.v3i3.870.

[16] S. Suwarsito and H. Mustafidah, "Determination of Feed Fish Price Based on Feed Formulation with Local Raw Materials using Fuzzy Logic Implementation," *Int. J. Fish. Aquat. Stud.*, vol. 3, no. 2, pp. 1–5, 2015.

[17] S. Suwarsito and H. Mustafidah, "Formulasi Pakan Ikan Menggunakan Sistem Pakar Metode Perunutan Maju," in *Prosiding Seminar Nasional Hasil - Hasil Penelitian dan Pengabdian LPPM Universitas Muhammadiyah Purwokerto*, 2015.

[18] K. K. Widiartha and A. A. G. Ekayana, "Penentuan Jenis Ikan Air Tawar pada Lahan Budidaya Menggunakan Fuzzy Logic Berbasis Interface Microcontroller (Determination of Freshwater Fish Species in Cultivation Land Using Fuzzy Logic Based on Microcontroller Interface)," *S@ CIES*, vol. 7, no. 1, pp. 7–14, 2016, doi: https://doi.org/10.31598/sacies.v7i1.108.

[19] T. Ningsih, "Sistem Pendukung Keputusan Penentuan Jenis Ikan Air Tawar Untuk Usaha Pembesaran Menggunakan Metode ANP-PROMETHEE II (Studi Kasus Kabupaten Nganjuk)." Universitas Brawijaya, 2017.

[20] S. Suwarsito and H. Mustafidah, "Determination of Appropriate Fish Culture Method Based on Water Quality Using Expert System," *Adv. Sci. Lett.*, vol. 24, no. 12, pp. 9178–9181, 2018, doi: https://doi.org/10.1166/asl.2018.12120.

[21] A. A. Soebroto and S. Hartati, "Penentuan Jenis Ikan Air Tawar Untuk Usaha Pembesaran Menggunakan Multicriteria Decision Making (MCDM)," 2018.

[22] A. Azis, "IDENTIFIKASI JENIS IKAN MENGGUNAKAN MODEL HYBRID DEEP LEARNING DAN ALGORITMA KLASIFIKASI," *Sebatik*, vol. 24, no. 2, pp. 201–206, 2020.

[23] A. Pariyandani, D. A. Larasati, E. P. Wanti, and M. Muhathir, "Klasifikasi Citra Ikan Berformalin Menggunakan Metode k-NN dan GLCM," in *Semantika (Seminar Nasional Teknik Informatika)*, 2019, vol. 2, no. 1, pp. 42–47.

[24] E. P. Wanti and M. Muhathir, "Pengidentifikasian Citra Ikan Berformalin Dengan Menggunakan Metode Multilayer Perceptron," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 5, no. 1, pp. 491–502, 2021, doi: http://dx.doi.org/10.30645/j-sakti.v5i1.342.

[25] R. E. Pawening, A. Z. Arifin, and A. Yuniarti, "Ekstraksi fitur berdasarkan deskriptor bentuk dan titik salien untuk klasifikasi citra ikan tuna," 2016.