

---

## Application of K-Nearest Neighbor Algorithm for Estimating Fishery Product Quality at BPPMHKP Pontianak

Michelson Febrianto<sup>1</sup>, Achmad Zakki Falani<sup>2</sup>

<sup>1,2</sup>Narotama University, Faculty of Computer Science, Department of Informatics Engineering, Jalan Arief Rahman Hakim No. 51, Surabaya, Indonesia  
[mich.febrian@gmail.com](mailto:mich.febrian@gmail.com), [achmad.zakki@narotama.ac.id](mailto:achmad.zakki@narotama.ac.id)

---

### Keywords

Accuracy; Fishery Products; K-Nearest Neighbor; Prediction; Total Plate Count.

**\*Corresponding Author:**  
[mich.febrian@gmail.com](mailto:mich.febrian@gmail.com)

### Abstract

Ensuring the microbiological quality of fishery products is essential for food safety and market competitiveness. One of the main indicators used in testing is the Total Plate Count (TPC), or *Angka Lempeng Total* (ALT) in Indonesia, which measures the concentration of aerobic microorganisms. This study proposes a predictive system for ALT values using the K-Nearest Neighbor (KNN) algorithm to support rapid and accurate quality assessment. The dataset, consisting of 622 samples from microbiological tests conducted at the BPPMHKP Laboratory in Pontianak (2020–2024), was preprocessed by converting ALT values into numeric format, cleaning and validating entries, applying Min-Max normalization, and splitting the data into training (80%) and testing (20%). The KNN algorithm with  $K = 3$  was implemented, where predictions were obtained by calculating Euclidean distances, selecting the three nearest neighbors, and averaging their ALT values. The system achieved a prediction accuracy of 98.66% with an average error of 1.34%. Moreover, 97.6% of samples were correctly classified below the microbiological safety threshold of  $5.0 \times 10^5$  cfu/g. These results confirm that the model can provide reliable predictions of product safety while reducing the reliance on fully manual laboratory testing. The proposed system has the potential to be developed into an application-based decision-support tool for laboratories and fisheries stakeholders to improve efficiency in microbiological quality monitoring.

---

## 1. Introduction

Fishery products are among the most essential commodities in providing high-quality food that is safe for consumption. To ensure their quality and safety, microbiological testing plays a crucial role, with one of the key parameters being the Total Plate Count (TPC), locally known in Indonesia as *Angka Lempeng Total* (ALT). ALT measures the number of aerobic microbial colonies that grow on a specific medium. A high ALT value indicates potential microbial contamination. Products with elevated ALT levels tend to spoil more quickly, which affects their quality and market competitiveness.

At the BPPMHKP Laboratory in Pontianak, ALT testing is still performed manually. This process requires high precision and becomes less efficient as the amount of data increases. Therefore, there is a need for a decision support system that can automatically and accurately predict the quality of fishery products based on ALT values. The K-Nearest Neighbor (KNN) algorithm is one of the suitable methods for this classification task. KNN calculates the distance between the test sample and all training data, then generates a prediction based on the average or majority of the nearest neighbors. This algorithm is well known for its simplicity and effectiveness, especially in numerical and image-based classification tasks [1], [2]

Several previous studies have successfully applied the KNN algorithm in the fisheries domain. Khairunnisa et al. and Arief and Rahmadewi applied KNN to classify fish freshness by analyzing eye color features from digital images showing that visual-based classification can yield accurate results[3], [4]. Further improved this approach by utilizing HSV color features which provided better representation of natural color variations[5]. Beyond visual classification Ramadhan and Prasetyo implemented KNN for aquaculture water quality monitoring using parameters such as turbidity, pH, temperature and ammonia demonstrating its flexibility in environmental data processing[6]. Similarly Jadid et al. designed a milkfish quality detection system using physical and chemical attributes confirming that KNN can effectively process multivariate numerical data[7].

However, most of these studies primarily focused on visual, physical, or environmental parameters. The application of KNN to microbiological data such as Total Plate Count (ALT) remains rarely explored, even though it has significant potential for laboratory automation and decision support. While demonstrate the effectiveness of KNN in handling various non-microbiological attributes, none specifically address microbiological indicators[5], [6], [7]. This gap highlights the novelty of the present study, which develops a predictive system for assessing fishery product quality based on ALT values using the KNN algorithm, evaluated with real laboratory test data from the BPPMHKP Laboratory in Pontianak.

## 2. Research Method

This study developed a predictive system for assessing the quality of fishery products based on Total Plate Count (TPC/ALT) values using the K-Nearest Neighbor (KNN) algorithm. The methodology consists of three main stages: data preprocessing, KNN algorithm implementation and prediction result evaluation. The overall system workflow is illustrated in Figure 1.

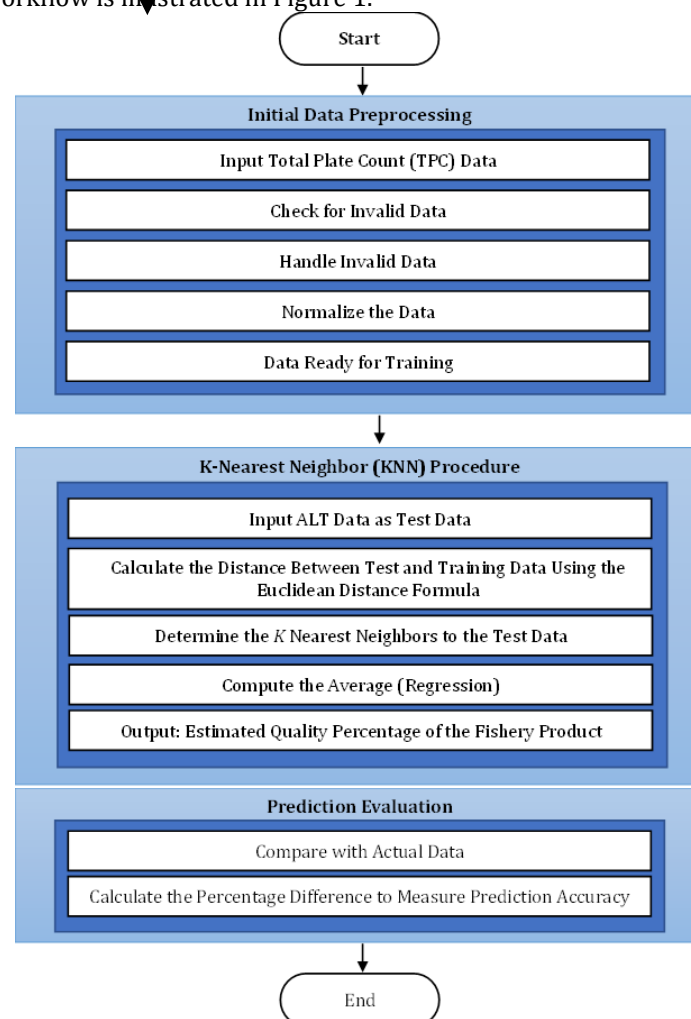


Figure 1. System flowchart for predicting fishery product quality using the K-Nearest Neighbor (KNN) algorithm

In addition to the system flowchart, the detailed steps involved in each stage of the methodology are summarized in Table 1. This table highlights the processes carried out during data preprocessing, algorithm implementation, and evaluation to ensure systematic and reliable prediction results.

*Table 1. Summary of preprocessing, training, and evaluation steps in the predictive system*

Stage	Process	Description
<b>Data Preprocessing</b>	Data collection	Collecting Total Plate Count (TPC/ALT) values of fishery products from laboratory test results.
	Data validation	Checking data consistency, removing duplicates, and handling missing or invalid values.
	Normalization	Scaling data to ensure uniform ranges for optimal distance calculation in KNN.
<b>KNN Algorithm</b>	Distance calculation	Computing the Euclidean distance between test data and all training data.
	Neighbor selection	Selecting the $k$ nearest neighbors based on the smallest distance values.
	Prediction	Performing regression to estimate the predicted TPC value of the test sample.
<b>Evaluation</b>	Accuracy measurement	Comparing predicted values with actual values to calculate prediction accuracy.
	Threshold analysis	Applying the TPC threshold of $5.0 \times 10^5$ colonies/gram to determine product quality status.

To strengthen the reliability of the method, each stage is supported by established practices in machine learning. First, data preprocessing was carried out by converting ALT values, initially presented in scientific notation (e.g., " $6.5 \times 10^4$ "), into numeric format. Data validation was then performed to remove duplicates, detect input errors, and handle missing entries, ensuring dataset integrity[8]. Subsequently, Min-Max normalization was applied to scale the values within the range [0,1]. This normalization is widely used to avoid bias in distance-based algorithms such as KNN[2].

Second, the KNN algorithm was implemented by computing the one-dimensional Euclidean distance between each test data point and all training data. For multi-dimensional data, the general Euclidean distance formula is also applicable[9]. The nearest  $k$  neighbors were then selected based on the smallest distances [10]. In this study,  $K = 3$  was chosen to balance prediction stability and sensitivity to local data variations.

Finally, the evaluation stage involved assessing prediction accuracy by comparing the predicted ALT values against the actual laboratory test results. Threshold analysis was also applied using the microbiological quality standard of  $5.0 \times 10^5$  cfu/g established by BPPMHKP Pontianak to determine product quality status (safe vs. unsafe). In addition, cross-validation was performed to minimize bias and improve the robustness of the evaluation [11].

Through this methodology, the system ensures accurate prediction of microbial contamination levels in fishery products while also providing an initial evaluation of their quality based on laboratory standards.

## 2.1 Data Preprocessing

The data used in this study were obtained from microbiological tests on fishery product samples conducted at the BPPMHKP Laboratory in Pontianak during the 2020 - 2024 period. The initial data consisted of ALT values presented in scientific textual format (e.g., " $6.5 \times 10^4$ "), which were converted into numeric format to enable computational processing.

To standardize the scale of the data, the Min-Max normalization method was applied to rescale all ALT values into the range [0, 1]. Normalization is essential for distance-based algorithms such as KNN to avoid bias caused by large variations in feature values[2]. The normalization is expressed as follows:

$$X_{Norm} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \quad (1)$$

Where  $X$  is the original ALT value,  $X_{Min}$  is the minimum value in the dataset,  $X_{Max}$  is the maximum value, and  $X_{Norm}$  is the normalized ALT value.

After normalization, the dataset was randomly divided into two groups: 80% as training data and 20% as testing data. The training data were used to build the prediction model, while the testing data were used to evaluate the performance of the system. This ratio was chosen to ensure sufficient data for model training while maintaining an adequate portion for evaluation [11].

## 2.2 Implementation of the K-Nearest Neighbor Algorithm

The prediction of ALT values was carried out using the K-Nearest Neighbor (KNN) algorithm with the parameter  $K = 3$ . In this approach, the system calculates the distance between each test data point and all training data using the Euclidean distance method [1], [12]. Since the data consists of a single numerical value obtained from normalized ALT, the one-dimensional Euclidean distance is calculated as:

$$d(x, y) = |x - y| \quad (2)$$

Where  $x$  is the normalized ALT value from the test data,  $y$  is the normalized ALT value from the training data and  $d(x, y)$  is the Euclidean distance [6].

Although only a single ALT feature ( $n = 1$ ) is used in this study, the general Euclidean distance formula for multi-dimensional data is expressed as:

$$d(x, y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2} \quad (3)$$

where  $x = (x_1, x_2, \dots, x_n)$  represents the feature vector of the test sample  $y = (y_1, y_2, \dots, y_n)$  represents the feature vector of a training sample, and  $n$  is the number of features. For one-dimensional data ( $n = 1$ ) this formula reduces to  $|x - y|$  [9].

For each test sample the three training data points with the smallest Euclidean distances are selected as the nearest neighbors. The predicted ALT value is then obtained by averaging the ALT values of these neighbors through a regression approach [11].

The choice of  $K = 3$  was made to balance sensitivity to local data patterns and overall prediction stability. Smaller  $K$  values such as  $K = 1$ , are often too sensitive to noise while larger  $K$  values may reduce the influence of relevant neighbors [12], [13]. The performance comparison of different  $K$  values will be presented in the Results and Discussion section.

## 2.3 Prediction Evaluation

The evaluation was conducted by comparing the predicted ALT values with the actual ALT values from the test data. The error rate was measured using the percentage error formula:

$$\text{Percentage Error} = \left( \frac{|\text{Predicted ALT} - \text{Actual ALT}|}{\text{Actual ALT}} \right) \times 100\% \quad (4)$$

where *Predicted ALT* is the system-generated value, and *Actual ALT* is the laboratory measurement

The percentage error was calculated for all test samples and then averaged to estimate the overall system accuracy. In addition, the system was evaluated based on the ALT threshold of  $5.0 \times 10^5$  colony-forming units per gram (cfu/g), in accordance with the microbiological quality standards set by the BPPMHKP Laboratory in Pontianak. This approach enables the system to indicate whether the predicted product quality falls within the acceptable range based on microbiological criteria [11], [14].

## 3. Result and Discussions

This study developed a prediction system for Total Plate Count (ALT) values in fishery products using the K-Nearest Neighbor (KNN) algorithm. The dataset was obtained from microbiological test results conducted by the BPPMHKP Laboratory in Pontianak during the 2020 to 2024 period. A total of 622 samples were collected and divided into 497 training data and 125 testing data. Based on the microbiological quality threshold ( $5.0 \times 10^5$  cfu/g) established by BPPMHKP Pontianak [3], the dataset consisted of 619 safe samples ( $\text{ALT} \leq 5.0 \times 10^5$  cfu/g) and 3 unsafe samples ( $\text{ALT} > 5.0 \times 10^5$  cfu/g). This distribution highlights that the dataset is imbalanced, with the majority of products categorized as safe, reflecting the actual conditions commonly encountered in laboratory testing.

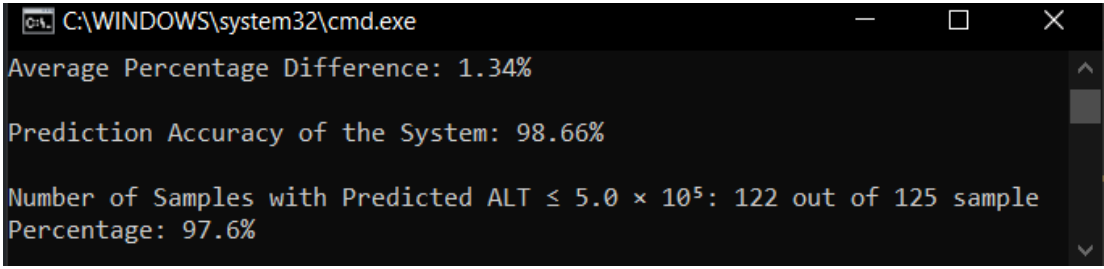
Initially, ALT values were presented in scientific textual formats (e.g., “ $4.5 \times 10^5$ ”) and then converted into numerical form. The converted data were subsequently validated, cleaned, and normalized using the Min-Max method [2],[8] to ensure a consistent scale for classification purposes. Prediction was carried out by calculating the one-dimensional Euclidean distance between each test sample and all training data. The three nearest neighbors ( $K = 3$ ) with the smallest distances were then selected, and regression was performed to determine the predicted ALT value based on the average of the ALT values of those three neighbors. This process was implemented using the Python programming language, with spreadsheet tools employed for verification purposes.

The prediction results obtained showed values that were very close to the actual ALT values. As an example, a comparison between the actual and predicted ALT values from a subset of the test data is presented in Table 2. The table illustrates that the predicted values are almost identical to the actual laboratory results, with differences ranging from 0 to 167 colonies/gram, corresponding to percentage differences of 0%–2.53%. This indicates that the proposed model demonstrates high prediction accuracy and consistency in estimating ALT values.

*Table 2. Comparison of Actual and Predicted ALT Values for a Subset of Test Data*

Sample Code	Actual ALT	Predicted ALT	Difference	Percentage Difference
S606	24000	24000	0	0,00
S544	14000	14000	0	0,00
S527	20000	20000	0	0,00
S156	27000	27000	0	0,00
S591	7400	7567	167	2,26
S596	6600	6767	167	2,53
S603	24000	24000	0	0,00
S561	20000	20000	0	0,00
S601	38000	38000	0	0,00
S102	24000	24000	0	0,00

From a total of 125 test samples, the system demonstrated an average percentage difference of 1.34%, equivalent to a prediction accuracy of 98.66%. This evaluation was performed automatically using Python, with a summary of the output results shown in Figure 2.



```

C:\WINDOWS\system32\cmd.exe
Average Percentage Difference: 1.34%

Prediction Accuracy of the System: 98.66%

Number of Samples with Predicted ALT  $\leq 5.0 \times 10^5$ : 122 out of 125 sample
Percentage: 97.6%

```

*Figure 2. System Evaluation Results Based on Python Output*

Based on the evaluation results, the average percentage error was 1.34%, which corresponds to a system prediction accuracy of 98.66%. In addition, 122 out of 125 samples (97.6%) were predicted to be below the microbiological quality threshold of  $5.0 \times 10^5$  cfu/g, indicating that the majority of products are considered safe for consumption. To further illustrate the system’s performance, Figure 3 presents a bar chart comparing the actual and predicted ALT values along with the percentage differences for 10 representative test samples.

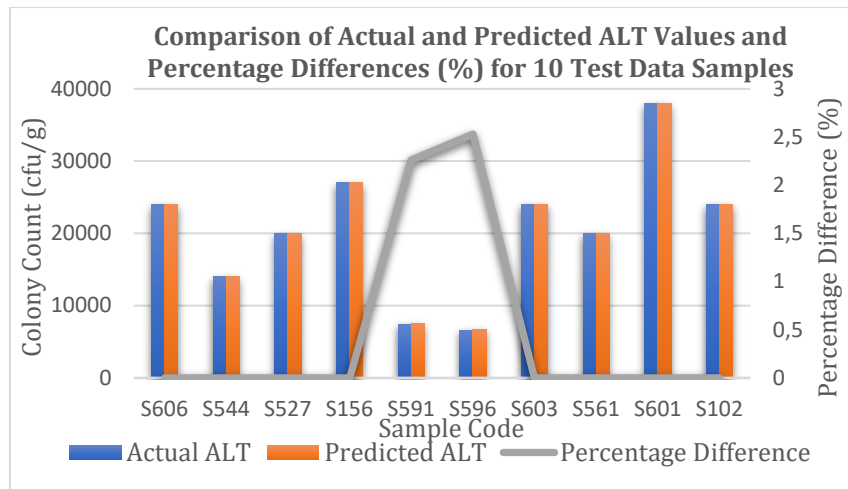


Figure 3. Bar chart comparing actual and predicted ALT values along with percentage differences for 10 test samples using the K-Nearest Neighbor (KNN) algorithm

The chart demonstrates that most predictions closely matched the actual values, with percentage differences generally remaining below 3%. The selected samples were chosen to provide a representative view of system performance while maintaining visual clarity.

In addition, to validate the selection of  $K = 3$ , experiments were also conducted using different  $K$  values (1, 5, and 7). The accuracy results are summarized in Table 3

Table 3. Accuracy results with different  $K$  values

K Value	Accuracy (%)
1	99.10%
3	98.66%
5	97.81%
7	97.01%

Although  $K = 1$  produced the highest accuracy (99.10%), this configuration is highly sensitive to noise and outliers because it relies solely on one neighbor. In contrast,  $K = 3$  provided slightly lower accuracy (98.66%) but offered more stable and reliable predictions. Larger  $K$  values (5 and 7) resulted in a gradual decline in accuracy due to the inclusion of less relevant neighbors, which diluted the influence of the closest data points. Therefore,  $K = 3$  was selected as the optimal parameter, balancing both accuracy and robustness for predicting ALT values in fishery products.

Despite these promising results, this study has several limitations. First, the dataset size was relatively small, consisting of only 622 samples (497 training and 125 testing), which may not fully represent the variability of fishery products in broader contexts. Second, the data were collected exclusively from the BPPMHKP Laboratory in Pontianak, reflecting only one laboratory within a single region. This limited scope may introduce potential bias, as the findings might not be directly generalizable to other geographic areas or laboratories with different testing practices. Future research should therefore consider expanding the dataset across multiple laboratories and regions to improve the robustness and generalizability of the predictive system.

Furthermore, an analysis of misclassifications revealed that most prediction errors were minor and occurred near the decision boundary of the threshold ( $5.0 \times 10^5$  cfu/g). For example, some samples with ALT values close to the threshold showed slightly higher percentage differences between actual and predicted values. While these cases did not significantly affect the classification outcome (safe vs. unsafe), they highlight the algorithm's sensitivity when dealing with borderline values. Such conditions could lead to potential misclassification if the dataset contained a larger proportion of samples near the threshold. Addressing this issue may require incorporating additional features, larger datasets, or hybrid approaches that combine KNN with other machine learning techniques to enhance robustness in borderline cases.

To further validate the reliability of the proposed model, a 10-fold cross-validation was performed. In this procedure, the dataset was randomly partitioned into ten equal subsets, with nine subsets used for training and the remaining one for testing in each iteration. This process was repeated ten times, and the results were averaged to minimize bias from any single partition. The cross-validation results indicated that the model consistently maintained high accuracy across all folds, confirming its robustness and reducing the likelihood of overfitting to a specific training–testing split.

In addition, an error distribution analysis was conducted to examine the magnitude of prediction deviations. The majority of prediction errors were concentrated within  $\pm 5\%$  of the actual ALT values, demonstrating that the system provides not only correct classifications (safe vs. unsafe) but also accurate regression estimates of ALT levels. Only a small number of outliers were observed, primarily in samples with ALT values close to the microbiological threshold of  $5.0 \times 10^5$  cfu/g. These findings emphasize the importance of considering uncertainty near the decision boundary, as even minor deviations may influence quality assessments in practical applications.

Based on evaluations of all test samples and applying the ALT threshold of  $5.0 \times 10^5$  cfu/g established by BPPMHKP Pontianak, the system provides a reliable initial evaluation of product quality.. This makes the model suitable for supporting laboratory workflows and offers a foundation for future development of decision support systems in fishery product quality monitoring across wider datasets and contexts.

## **4. Conclusions and Future Works**

### **4.1 Conclusions**

This study successfully designed and implemented a predictive system for assessing the quality of fishery products based on Total Plate Count (TPC), locally known as Angka Lempeng Total (ALT), using the K-Nearest Neighbor (KNN) algorithm. The system was developed utilizing microbiological test data obtained from the BPPMHKP Laboratory in Pontianak for the period of 2020–2024.

The data processing involved several stages, including the conversion of ALT values from scientific text format to numeric values, normalization using the Min-Max method, splitting the data into training and testing sets, and prediction using the K-Nearest Neighbor (KNN) algorithm with  $K = 3$ . The predicted ALT value was obtained by averaging the ALT values of the three nearest neighbors based on Euclidean distance.

The system evaluation showed a prediction accuracy rate of 98.66%, with an average percentage error of 1.34% between the predicted and actual ALT values. Additionally, the system successfully predicted that 97.6% of the total test data fell below the microbiological quality threshold of  $5.0 \times 10^5$  colony-forming units per gram (cfu/g), indicating that the majority of the products were classified as having good quality.

This system has the potential to be used as a rapid analytical tool for evaluating the microbiological quality of fishery products, especially in laboratories that routinely handle large volumes of test data. Its reliability also opens opportunities for further development into an automated quality evaluation application.

### **4.2 Future Works**

To enhance the performance and broaden the applicability of the proposed prediction system, several directions for future research are recommended:

1. Expansion of the dataset is essential by increasing the number of samples, diversifying the types of fishery products, and extending the testing period. A larger and more heterogeneous dataset will allow the system to capture wider variability and generate more representative predictions.
2. Development of a user-friendly application or interface (UI) should be considered to improve accessibility. Such an implementation would enable Technical Implementation Units (UPT) under BPPMHKP in different regions to utilize the prediction system for rapid and accurate evaluation of microbiological quality in fishery products.
3. Further exploration of different  $K$  values in the KNN algorithm is required to identify the optimal configuration that maximizes prediction accuracy. This investigation will provide valuable insights into the sensitivity of the system to parameter variations and strengthen the validity of the model.
4. Integration with database and automated reporting systems is strongly recommended. Leveraging programming languages such as Python for this purpose would enhance efficiency in laboratory workflows and support the modernization of microbiological evaluation systems by incorporating historical data for continuous improvement.

## 5. References

- [1] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.
- [2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin: Springer, 2006.
- [3] Khairunnisa, Munawir, and N. Fadillah, "Pengenaln Kualitas Ikan Berdasarkan Warna Mata Menggunakan Metode K-Nearest Neighbor," 2020.
- [4] A. I. Arief and R. Rahmadewi, "Penerapan Algoritma KNN pada Kesegaran Ikan Menggunakan Citra Digital," 2021.
- [5] C. Y. Jerandu, P. Batarius, and A. A. J. Sinlae, "Identifikasi Kualitas Kesegaran Ikan Menggunakan Algoritma K-Nearest Neighbor Berdasarkan Ekstraksi Ciri Warna Hue, Saturation, dan Value (HSV)," *Biosainstek*, vol. 4, no. 3, 2022, doi: 10.47065/bits.v4i3.2613.
- [6] D. Ramadhan and B. H. Prasetio, "Sistem Klasifikasi Kualitas Air Kolam Ikan Lele dengan Metode K-Nearest Neighbor," 2023.
- [7] M. Jadid, A. S. Adani, and P. H. Susilo, "Implementasi Metode K-Nearest Neighbor Sebagai Sistem Pendeteksi Kualitas Ikan Bandeng," *G-Tech Journal*, vol. 8, no. 1, 2024, doi: 10.29407/gj.v8i1.21131.
- [8] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco: Morgan Kaufmann, 2011.
- [9] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans Inf Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
- [10] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.
- [11] D. Aha, D. Kibler, and M. Albert, "Instance-based learning algorithms," *Mach Learn*, vol. 6, no. 1, pp. 37–66, 1991, doi: 10.1023/A:1022689900470.
- [12] L. F. Cunningham, J. Gerlach, and M. D. Harper, "Perceived risk and e-banking services: An analysis from the perspective of the consumer," *Journal of Financial Services Marketing*, vol. 10, no. 2, pp. 165–178, 2005, doi: 10.1057/palgrave.fsm.4770183.
- [13] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Cambridge, MA: MIT Press, 2018.
- [14] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1145.