**Jurnal Pengelolaan Perikanan Tropis (Journal of Tropical Fisheries Management)**

Website: http://journal.ipb.ac.id/index.php/jurnalppt
e-ISSN: 2721-1525
Vol. 10 No. 1: 24-35 (206) http://dx.doi.org/10.29244/jppt.10.1.24-35

# Performance Analysis of the K-Nearest Neighbors (KNN) Algorithm in Classification the Presence of Small Pelagic Fish during the East Season in Banten Waters

**Citra Amelia Nawati [1,*], Ayang Armelita Rosalia [2], Novi Sofia Fitriasari [3]**

Received: 20-06-2026 / Revised: 27-06-2026 / Accepted: 30-06-2026

**ABSTRACT**

The total fish catch production of Banten Province reached 71,846.71 tons in 2023, with small pelagic fish being the dominant commodity in Banten waters. However, fishermen around Banten waters still rely on conventional methods and intuition in determining fishing locations, resulting in unstable catches and inefficiencies in operational costs, time, and labor. To address these issues, this study uses a quantitative approach supported by qualitative data, where this study applies the K-Nearest Neighbors (KNN) algorithm to classify the presence of small pelagic fish during the east monsoon in Banten waters. Secondary data comes from Aqua MODIS satellite imagery for the period 2021–2025, processed through the Kaggle and Google Colab platforms. Two oceanographic parameters used are sea surface temperature (SST) and chlorophyll-a, both of which function as primary indicators of aquatic ecosystem productivity. The dataset consists of 3,382 data points divided into two classes: potential and non-potential. The model was tested with varying K values (3, 5, 7, 9) and data sharing ratios (60:40, 70:30, 80:20) using Euclidean distance and evaluated using a confusion matrix. The best configuration was obtained at K=7 with a ratio of 70:30, resulting in an accuracy of 95.66% and proportional precision, recall, and F1-score values. Permutation Feature Importance analysis showed that chlorophyll-a contributed dominantly at 25.66% and SPL at 23.58%. The results confirmed that the KNN algorithm performed well in classifying the presence of small pelagic fish in Banten waters and can be used as a reference for developing a decision support system for fishermen.

**Keywords:** Banten Waters, Chlorophyll-a, K-Nearest Neighbors, Sea Surface Temperature, Small Pelagic Fish

## INTRODUCTION

Indonesia has enormous maritime potential, but its utilization rate has only reached 58.80% (Dahuri, 2003 in Arif *et al.,* 2018). One of the strategic water areas with strong potential is the Banten Waters which is the meeting point of the water masses of the Java Strait, the Indian Ocean, and the Sunda Strait (Irnawati et al., 2020). According to the Department of Maritime Affairs and Fisheries (DKP), The total fish catch production of Banten Province reached 71,846.71 tons in 2023 with small pelagic fish (anchovies, lemuru, and mackerel) being one of the dominant groups. However, the lack of technology adoption means that conventional Banten fishermen still rely heavily on intuition to find fishing areas. This often results in uncertain catches and increased costs, time, and energy ( Ali 2020, Simbolon 2017 in Tarigan et al., 2020).

The Presence of small pelagic fish is estimated to have the greatest potential compared to other fish groups, such as large pelagic, demersal, reef fish, and commodities like shrimp and squid. Their distribution itself is highly dependent on the dynamics of marine environmental parameters, including currents, food availability, and sea surface temperature (Khatami et al., 2019). Furthermore, seasonal dynamics also significantly influence oceanographic conditions in marine waters, one of which is the east monsoon. The east monsoon (June–

[1]*Corresponding author
✉ Citra Amelia Nawati
Citraamelia2710@upi.edu

[1] Marine Information Systems Study Program, Serang Regional Campus , Indonesian University of Education, Indonesia .
[2]Marine Information Systems Study Program, Serang Regional Campus , Indonesian University of Education, Indonesia .
[3] Marine Information Systems Study Programs, Serang Regional Campus , Indonesian University of Education, Indonesia .

September) is one of the seasons that affects aquatic productivity due to upwelling in several parts of Indonesian waters. Factors such as upwelling, temperature, and chlorophyll-a distribution can influence the presence and Presence of fish in the waters so that many fish are caught while searching for food (Takwir et al., 2021). During the east monsoon, winds show an increase in their distribution and speed, which directly changes the direction of currents in the ocean's surface layer and this indicates that wind movement is a major variable that influences the physical properties of seawater masses in the region (Banjarnahor et al., 2020). The influence of winds that create upwelling phenomena such as the southeast monsoon, namely the rise of nutrients from the deep sea to the surface causes increased phytoplankton growth so that the availability of food for small pelagic fish becomes more abundant (Nufus et al., 2026). According to Nurkhairani et al. (2018) the east monsoon, it is a good season for catching pelagic fish, because the water temperature is relatively warm and the nutrient levels are high. The presence of pelagic fish in the waters is influenced by the presence of nutrients in the waters (Siringoringo et al., 2024). To identify the presence of fish, two main oceanographic parameters are sea surface temperature (SST) and chlorophyll-a. These two parameters facilitate the analysis of fishing areas, especially small pelagic fish (Sarianto, 2018 in the journal Fofied et al., 2024). One of the Artificial Intelligence technologies to solve this problem is Machine Learning. The K-Nearest Neighbors algorithm (KNN) including an algorithm that is easy to understand, efficient, and has simple data training. (Simbolon et al., 2024). The KNN algorithm will predict SST and chlorophyll-a data as x variables, then evaluate fish Presence data as y variables.

To overcome the inefficiency of capture, a Machine Learning approach is essential. Several previous studies have attempted to predict fishing grounds. Lubis *et al*. (2024) used the K-Nearest Neighbors (KNN) algorithm to predict fish abundance based on SST and rainfall with 83% accuracy. The results showed a positive correlation between SST and tuna abundance but did not use key productivity parameters such as chlorophyll-a. Another study in the coastal waters of Central Java by Sarwati *et al*. (2025) used chlorophyll-a and SST variables to estimate small pelagic fishing areas. This study also found that SST was positively correlated with small pelagic fish catches. The approach was still descriptive through regression analysis and spatial modeling. According to research by Fitrianah *et al*. (2016), the K-Nearest Neighbors (KNN) algorithm, which is based on spatial and temporal data, can identify potential fishing zones using SST and chlorophyll parameters with an accuracy of up to 87.11%. This study did not yet focus on discussing the KNN algorithm.

Based on the limitations of previous research, this study focuses more on analyzing the performance of the K-Nearest Neighbors algorithm for classifying the presence of small pelagic fish using oceanographic parameters such as sea surface temperature and chlorophyll-a during the peak fishing season, namely the east season. The purpose of this study is to analyze the performance of the K-Nearest Neighbors algorithm in classifying the presence of small pelagic fish during the east season in Banten Waters.

## MATERIAL AND METHOD

### Time and Place of Research

This research was carried out in Banten waters which are astronomically located between 5°7'–6°8' South Latitude and 105°1'–106°7' East Longitude. This location was chosen because it has significant fishery resource potential, with dynamic oceanographic characteristics influenced by the confluence of water masses from the Java Sea, the Indian Ocean, and the Sunda Strait.

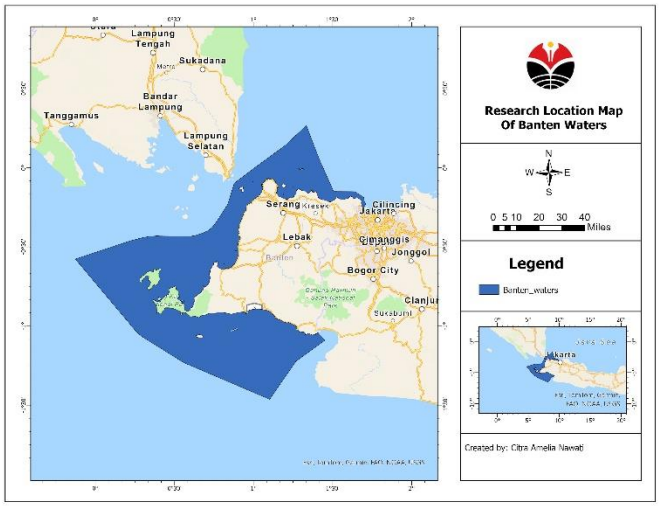


**Figure 1.** Research Location

### Tools and Materials

For data processing, the Kaggle platform is used for the pre-processing stage of data, such as cropping according to the research area and Combining the dataset between SST and chlorophyll so that

the data is ready for use. Next, Google Colab is used to implement and analyze the K-Nearest Neighbors (KNN) algorithm. The ArcGIS application is used to cut the water boundaries according to the research area and create a map layout. Microsoft Word and Excel software to support data pre-processing. The materials used include oceanographic parameters in the form of Sea Surface Temperature (SST) and Chlorophyll-a concentration sourced from the Aqua MODIS satellite, which are officially downloaded through the NASA OceanColor website.

**Methods**

This study employed a quantitative research approach supported by qualitative data. Data for this study were obtained through the official NASA Ocean Color Web (*https://ocean-color.gsfc.nasa.gov/*) downloaded on April 1, 2026 and is included in the secondary data collection method. The collected data consists of SST and Chlorophyll-a derived from satellite sensors with relevant spatial resolution, such as MODIS imagery. Spatial data on water boundaries were obtained from the Marine Regions website (*https://www.marineregions.org/*) which provides geographic information about marine areas around the world, including the boundaries of water areas, seas, straits, and established oceanographic zones.

As supporting validation, the authors also conducted interviews with fishermen at the Karangantu Fishing Port, fishermen at Lelean Beach, and the Bojonegara Fishing Port who fish around Banten waters. Information from these interviews was used to validate the model results with fishermen's empirical knowledge regarding indicators of the potential presence of small pelagic fish.

**Research Flow**

The research flow is presented in the following figure.

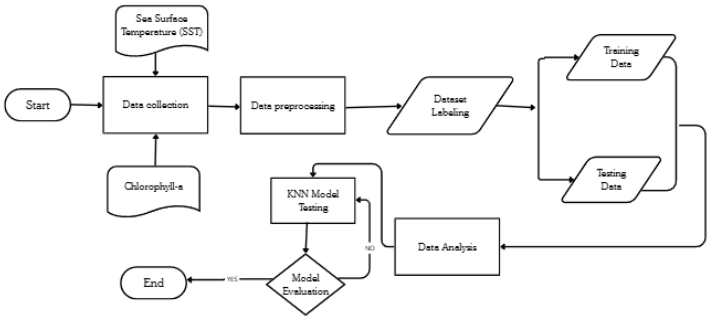


**Figure 2.** Research Flow

The research flow in Figure 2 starts from the collection of satellite image data downloaded from 2021 to 2025 with a resolution of 4 km and seasonal temporal resolution taken during the east monsoon season (June-September). The image data includes spatial data of sea surface temperature (SST) and chlorophyll-a image data to determine areas suspected of having many small pelagic fish that like the area. The pre-processing stage where the SST and chlorophyll data that have been download and then crop according to the research area to convert the format so that the data is ready to be processed. The downloaded file format in Nc format is converted to CSV using Kaggle software. After data preprocessing, the SST and chlorophyll datasets are then combined to produce a new dataset with features, year, latitude, longitude, Chlorophyll, SST, and output attributes, namely Potential. Label 1 is included in the potential class with an SST range of 28 °C and 29°C, while chlorophyll-a is 0.5 -2.5 mg /m3 while outside these criteria is classified as Label 0 (Not Potential).

**Table 1.** Determination of Fish Presence Potential Class Labels

| Potential (Class) | SST(°C) | Chlorophyll-a (mg/m³) |
|---|---|---|
| 1 (Potential) | 28 °C – 29 °C | 0.5 – 2.5 mg/m3 |
| 0 (No Potential) | <28°C or > 29°C | < 0.5 or > 2.5 mg /m3 |

Source: (Syamsuddin et al., 2018; Kurniawati et al., 2015)

The dataset was then tested with several data SST ratios, starting with 60:40, 70:30, and 80:20. Several train-test SST ratios were employed to evaluate the stability of model performance under different training data proportions. It was then the dataset is implemented into the K-Nearest Neighbors algorithm using the Euclidean Distance calculation to classify new data with training data based on its nearest neighbors, the smaller the distance, the higher the level of similarity between the data. The final stage

of the research was evaluating the model's performance using a Confusion Matrix.

Feature standardization is also performed using the StandardScaler method during the data processing stage. Standardization is performed to ensure that the original data has uniform values. The K-Nearest Neighbors (KNN) algorithm is a distance-based classification method that is sensitive to feature scaling and parameter configuration. Mismatched feature scales and suboptimal selection of the k value can lead to a decline in model performance if the optimization process is unstructured (Manurung et al., 2025). Variables that influence the model performance results are identified using the permutation importance approach.

**Data analysis**

a. K-Nearest Neighbors Model Construction

This modeling stage of data that has gone through the data preprocessing stage will be implemented in the KNN algorithm. The K-Nearest Neighbors (KNN) method is a classification technique that groups new objects into K groups based on the distance between the training data (Juniati et al., 2018). K-Nearest Neighbors as algorithm KNN was chosen in this study because it is a classification algorithm that uses the K value to determine the class of new data to be classified (Simbolon et al., 2024). This algorithm works by searching for data that has similarities or differences with the data to be classified. KNN is also an algorithm that is easy to understand, efficient, and its data training is simple, so this can be a distinct advantage. The steps of the KNN algorithm include (Arsyad, 2020 in the journal (Hasdyna & Dinata, 2020):

1) Split training data and test data from prepared data
2) Determining the number of K as many data as possible with the closest distance
3) Calculation of the distance between training data and test data
4) Sorting distances from smallest to largest
5) Determining the Y variable (model classification)
6) Model evaluation

Distance calculation techniques in KNN which is used to determine the proximity of training data namely Euclidean Distance. To calculate the square of the distance using the Euclidean method, use the following mathematical formula:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Information:
*d* = Euclidean distance
*Xi* = Testing Data
*Yi* = Training Data
*I* = Data Dimension
*n* = Total Data Dimensions

The K value in KNN is usually chosen as an odd number. This is because even values often lead to classification failure (Kusuma et al., 2023). The following mathematical formula is used to determine the ideal k value:

$$k \approx \sqrt{\frac{n}{2}}$$

b. Model Evaluation

Accuracy testing in the model evaluation stage is conducted to test and evaluate the model that has been created. Model accuracy analysis uses a Confusion Matrix. This method can be used to assess accuracy. At this stage, the information analyzed using the Confusion Matrix is checked to see whether it shows correct information according to the previous hypothesis. The Confusion Matrix produces a table used in Machine Learning to improve the performance of the classification model (Simbolon et al., 2024). This table shows the number of data that can be predicted correctly or incorrectly and consists of the following four main variables:

- True Positive (TP): Positively labeled data that is accurately predicted as positive by the algorithm.
- False Positive (FP): Negatively labeled data that is incorrectly identified as positive.
- False Negative (FN): Data that is actually positive but is instead predicted as belonging to the negative class.
- True Negative (TN): Data that is actually negative and is successfully predicted as negative.

From these main variables, various metrics can be evaluated, such as accuracy, precision, recall, and F-Score. The formula for evaluating various metrics according to Powers (2020) in the journal (Kurnianto et al., 2024) namely as follows

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision+recall}}$$

## RESULT AND DISCUSSION

### Result

### Data Processing

Data preprocessing produced a structured dataset with attributes of year, latitude, longitude, chlorophyll-a, SST, and the output variable, Potential.

| | tahun | lat | lon | chlor_a | sst | potensi |
|---|---|---|---|---|---|---|
| 0 | 2021 | -5.5208 | 106.3125 | 0.489401 | 29.394999 | 0 |
| 1 | 2021 | -5.5625 | 106.2708 | 0.471510 | 29.535000 | 0 |
| 2 | 2021 | -5.5625 | 106.3125 | 0.462717 | 29.525000 | 0 |
| 3 | 2021 | -5.6042 | 106.2292 | 0.473709 | 29.429998 | 0 |
| 4 | 2021 | -5.6042 | 106.2708 | 0.467037 | 29.435000 | 0 |
| ... | ... | ... | ... | ... | ... | ... |
| 3377 | 2025 | -7.3958 | 105.9792 | 0.245531 | 28.470000 | 0 |
| 3378 | 2025 | -7.3958 | 106.0208 | 0.259316 | 28.760000 | 0 |
| 3379 | 2025 | -7.3958 | 106.0625 | 0.285509 | 28.785000 | 0 |
| 3380 | 2025 | -7.4375 | 106.0208 | 0.232842 | 28.894999 | 0 |
| 3381 | 2025 | -7.4375 | 106.0625 | 0.242352 | 28.920000 | 0 |

3382 rows × 6 columns

**Figure 3.** Data Preprocessing Result

The number of datasets in Figure 3 consisted of 3,382 data points, with a fairly balanced class distribution of 1,700 for class 1 data and 1,682 for class 0 data. StandardScaler is also used to ensure that the distance calculation in the KNN algorithm is not biased due to the difference in units between coordinates (degrees) and chlorophyll (mg/m3).

### Model Testing Results

### K-Nearest Neighbors Method with K Value = 3

1. Data distribution ratio 60% training data and 40% testing data

   The first trial utilized 3,382 data points partitioned into a 60:40 ratio, 2,029 training samples and 1,353 testing samples with k set to 3. The system demonstrated an accuracy of 95.49%, translating to 1,292 accurate predictions and 61 errors. Evaluating the individual categories, class 0 achieved precision, recall, and F1-scores of 95%, 96%, and 96%, respectively. Meanwhile, class 1 yielded 96% precision, 95% recall, and a 95% F1-score.
2. Data distribution ratio 70% training data and 30% testing data

   The next test uses a 70:30 data sharing scheme at a value of k = 3 with a total of 3382 data. The training data is 2367 and the testing data is 1015, resulting in an accuracy of 94.97% with the number of correct predictions of 964 and 51 incorrect predictions. With the precision, recall, and f1-score values for each class, namely class 0 has a precision score of 95%, recall 95%, and f1 score 95% while class 1 has a precision score of 95%, recall 95%, and f1 score 95%.
3. Data distribution ratio 80% training data and 20% testing data

   The experiment processed 3,382 data points using an 80:20 SST ratio and a k-value of 3. By allocating 2,705 records to training and 677 to testing, the algorithm produced an accuracy of 94.38%, successfully predicting 639 instances while misclassifying 38. Evaluation shows class 0 scoring 93%, 95%, and 94% across precision, recall, and F1-score, respectively. Conversely, class 1 reached 95%, 94%, and 95% for the same metrics.

**Table 2.** Accuracy of KNN Model Testing K Value = 3

| Ratio | Accuracy (%) | Class | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|---|
| 60:40 | 95.49 | Not Potential | 95 | 96 | 96 |
| | | Potential | 96 | 95 | 95 |
| 70:30 | 94.97 | Not Potential | 95 | 95 | 95 |
| | | Potential | 95 | 95 | 95 |
| 80:20 | 94.38 | Not Potential | 93 | 95 | 94 |
| | | Potential | 95 | 94 | 95 |

### K-Nearest Neighbors Method with K Value = 5

1. Data distribution ratio 60% training data and 40% testing data

For the first run with k=5, the 3,382 data points were split using a 60:40 ratio. This means 2,029 data were used for training and 1,353 for testing. The model reached a 95.26% accuracy, getting 1,289 predictions right and only missing 64. Looking at the details per class, class 0 scored 95% for precision, 96% for recall, and 95% for the F1-score. Meanwhile, class 1 got 96% for precision, 95% for recall, and a 95% F1-score.

2. Data distribution ratio 70% training data and 30% testing data

In the subsequent trial for k=5, the 3,382 total data were split 70:30, giving us 2,367 training data and 1,015 testing data. Out of these, 967 predictions were spot on and 48 were off, resulting in an overall accuracy of 95.27%. Breaking down the performance metrics, class 0 recorded a solid 95% for precision, recall, and F1-score. On the flip side, class 1 showed 96% precision, along with 95% for both recall and F1-score.

3. Data distribution ratio 80% training data and 20% testing data

Testing for the value of k = 5 uses a data sharing scheme of 80:20 with a total of 3382 data. The training data is 2705 and the testing data is 677, resulting in an accuracy of 94.97% with the number of correct predic tions of 643 and 34 incorrect predictions. With the precision, recall, and f1-score val ues of each class, namely class 0 has a preci sion score of 94%, recall 95%, and f1 score 95% while class 1 has a precision score of 96%, recall 95%, and f1 score 95%

**Table 3.** Accuracy of KNN Model Testing K Value = 5

| Ratio | Accuracy (%) | Class | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|---|
| 60:40 | 95.26 | Not Potential | 95 | 96 | 95 |
| | | Potential | 96 | 95 | 95 |
| 70:30 | 95.27 | Not Potential | 95 | 95 | 95 |
| | | Potential | 96 | 95 | 95 |
| 80:20 | 94.97 | Not Potential | 94 | 95 | 95 |
| | | Potential | 96 | 95 | 95 |

**K-Nearest Neighbors Method with K Value = 7**

1. Data distribution ratio 60% training data and 40% testing data

The k=7 scenario, the system processed 3,382 data points partitioned 60:40 into 2,029 training and 1,353 testing data. Out of the test set, 1,286 predictions were accurate and 67 were incorrect, capping the accuracy at 95.04%. Breaking down the metrics, class 0 generated 94% precision, 96% recall, and a 95% F1-score, whereas class 1 secured 96% precision, 94% recall, and a 95% F1-score.

2. Data distribution ratio 70% training data and 30% testing data

In the next evaluation with k = 7, the model utilized a 70:30 distribution out of the 3,382 data points. Out of the 2,367 training and 1,015 testing samples, the algorithm nailed 971 predictions correctly and stumbled on just 41, bumping the accuracy to 95.66%. A closer look at the breakdown shows class 0 obtaining a 95% precision and 96% for recall and F1-score. Class 1 matched this consistency, securing 96% in precision, recall, and the F1-score.

3. Data distribution ratio 80% training data and 20% testing data

Testing for the value of k = 7 uses a data sharing scheme of 80:20 with a total of 3382 data. The training data is 2705 and the testing data is 677, resulting in an accuracy of 95.12% with the number of correct predictions of 644 and 33 incorrect predictions. With the precision, recall, and f1-score values of each class, namely class 0 has a precision score of 95%, recall 95%, and f1 score 95% while class 1 has a precision score of 96%, recall 95%, and f1 score 95%.

**Table 4.** Accuracy of KNN Model Testing K Value = 7

| Ratio | Accuracy (%) | Class | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|---|
| 60:40 | 95.04 | Not Potential | 94 | 96 | 95 |
| | | Potential | 96 | 94 | 95 |
| 70:30 | 95.66 | Not Potential | 95 | 96 | 96 |
| | | Potential | 96 | 96 | 96 |
| 80:20 | 95.12 | Not Potential | 95 | 95 | 95 |
| | | Potential | 96 | 95 | 95 |

**K-Nearest Neighbors Method with K Value = 9**

1. Data distribution ratio 60% training data and 40% testing data

   Shifting the focus to k = 9 with a 60:40 distribution, the algorithm processed 2,029 training samples and evaluated 1,353 test samples. Out of these 3,382 total instances, the system made 1,283 accurate predictions and 70 incorrect ones, landing at an accuracy rate of 94.82%. On the performance side, class 0 hit 93%, 96%, and 95% for precision, recall, and F1-score respectively. Meanwhile, class 1 logged a 96% precision, 93% recall, and 95% F1-score.
2. Data distribution ratio 70% training data and 30% testing data

   Continuing to the k=9 test with a 70:30 ratio, the 3,382 data points were divided into 2,367 for training and 1,015 for testing. The model posted a 95.36% accuracy, getting 968 predictions right and slipping up on just 47. Looking closely at the classes, class 0 scored 95% in precision, 96% in recall, and 95% in F1-score. Meanwhile, class 1 wrapped up with 96% precision, 95% recall, and a 95% F1-score.
3. Data distribution ratio 80% training data and 20% testing data

   Testing for the value of k = 9 uses a data sharing scheme of 80:20 with a total of 3382 data. The training data is 2705 and the testing data is 677, resulting in an accuracy of 94.68% with the number of correct predictions of 641 and 36 incorrect predictions. With the precision, recall, and f1-score values for each class, namely class 0 has a precision score of 94%, recall 95%, and f1 score 95% while class 1 has a precision score of 96%, recall 94%, and f1 score 95%.

**Table 5.** Accuracy of KNN Model Testing K Value = 9

| Ratio | Accuracy (%) | Class | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|---|
| 60:40 | 94.82 | Not Potential | 93 | 96 | 95 |
| | | Potential | 96 | 93 | 95 |
| 70:30 | 95.36 | Not Potential | 95 | 96 | 95 |
| | | Potential | 96 | 95 | 95 |
| 80:20 | 94.68 | Not Potential | 94 | 95 | 95 |
| | | Potential | 96 | 94 | 95 |

**Evaluation of K Value at Highest Accuracy**

The graph shows an increasing trend as the K value increases

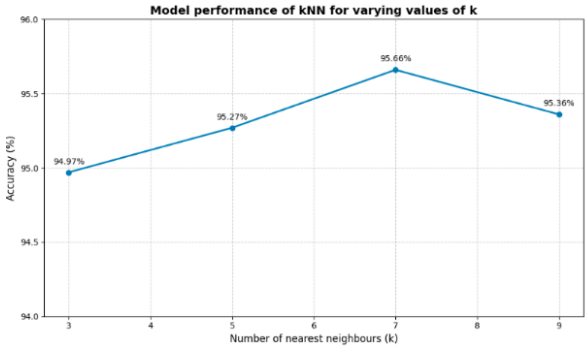


**Figure 4.** Graph of K Value at 70:30 Ratio

As shown in Figure 4 above, at K=3, the accuracy is approximately 94.97% and continues to increase until it reaches its highest value at K=7, with an accuracy of approximately 95.66%. At K=9, the accuracy drops to approximately 95.36%.

The excellent and balanced performance of the model in classifying both classes is shown through the visualization of the Confusion Matrix at the k value with the highest accuracy, namely K=7.

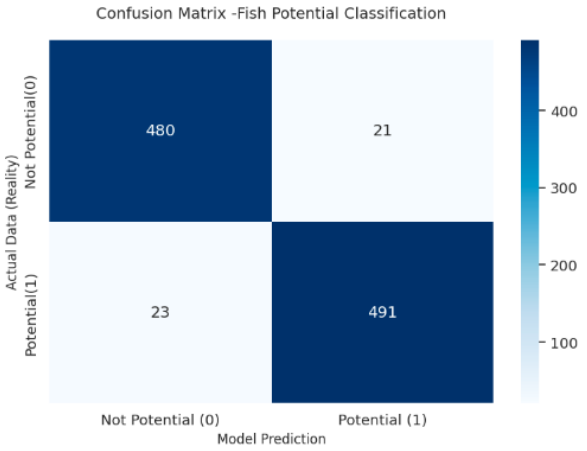


**Figure 5.** Confusion Matrix Value K=7

In figure 5 show the model performance on the test data of 1,015, the model predicted 480 True Negative (TN) points as non-potential, 491 True Positive (TP) points as potential, 21 False Positive (FP) points, and 23 False Negative (FN) points.

The parameters that have the most influence on fish Presence in Banten waters, which are calculated using Permutation Importance.

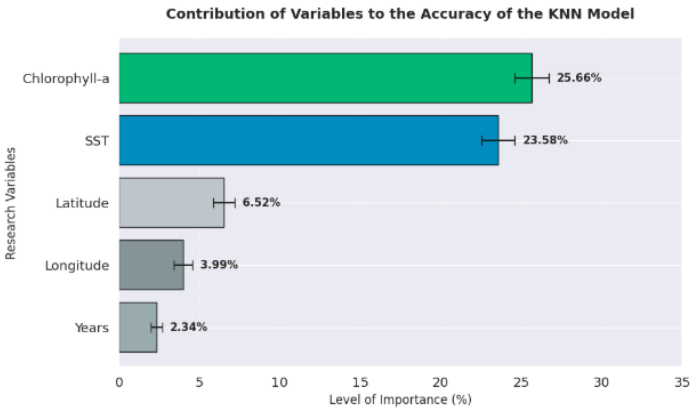


**Figure 6.** Permutation Importance Graph

The influence of each variable is shown in Figure 6. If this variable is removed, the model will be affected by chlorophyll-a reaching 25.66%, SST reaching 23.58%, 6.52% for latitude, 3.99% longitude, and 2.34% for year.

## K-Nearest Neighbors Modeling Results at Best Accuracy

Evaluation of the classification model performance using the KNN algorithm on 1,015 test points successfully predicted potential and non-potential areas.

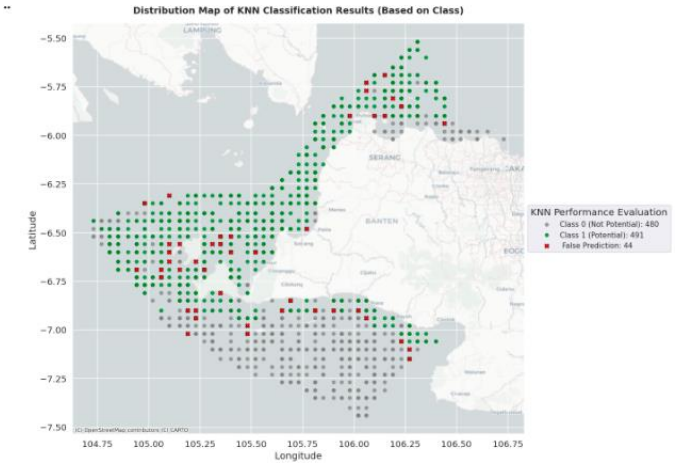


**Figure 7.** KNN Modeling Results

Overall, the model demonstrated high consistency, with a total of 971 correct predictions and only 44 incorrect predictions, supported by a stable f1-score of 0.96 for both categories. The visualization of the KNN model can be seen in Figure 7 above, the green dots depict potential classes that were successfully predicted correctly by the model, while the gray dots depict non-potential classes that were successfully predicted as non-potential by the model, and the red dots are classes that were predicted incorrectly by the model.

## Discussion

### The Effect of Data Sharing Ratio on Accuracy

The results of tests conducted at various K values (3, 5, 7, and 9) with three data sharing schemes (60:40, 70:30, and 80:20) show that the data sharing ratio has an impact on the performance of the K-Nearest Neighbors algorithm. Overall, the data sharing scheme with 70% training data and 30% testing data has the most stable accuracy, the highest accuracy level is 95.66% at a value of K = 7. In contrast, the 60:40 scheme with a value of k = 3 produces results with an accuracy of 95.49%. However, the accuracy trend in this ratio decreases as the K value increases. This occurs even though there is more test data (1,353 data), the limited training data (2,029 data) makes it difficult for the model to generalize when the K value increases. The quality of a model's predictions is highly dependent on the amount of training data available. When a shortage of training data prevents an algorithm from learning diverse patterns, the resulting poor performance is known as underfitting. Because the model's architecture is excessively simple, it struggles to grasp the complex relationships present in the data (Sivakumar et al., 2024).

The 80:20 test results show that this ratio has the lowest accuracy compared to other schemes but is still considered good. This indicates that the test

data with an 80:20 ratio shows greater variability or contains more complex outliers, this can make the model difficult to generalize, especially if the test and training data patterns are different (An et al., 2021). Research shows that the 70:30 ratio more often produces better accuracy than other ratio schemes, this is due to the less risk of overfitting due to a more balanced representation of training and testing (Sivakumar et al., 2024).

**Selection and Evaluation of K Value at the Best Ratio**

KNN classification heavily depends on how accurately the value of K is chosen. This is due to the fact that a big K number may lead to classification mistakes, making the prediction results erroneous, whereas a small K value is sensitive to outliers (Ujianto et al., 2025). The best accuracy occurs in the 70:30 ratio scheme with a value of k = 7, this indicates that adding more k values helps the model in generalizing more stable data patterns.

The results of the KNN algorithm testing show that the value of k = 7 has better accuracy compared to other k values. This indicates that, when compared to other k values depending on the dataset, the value of k = 7 provides a better balance of bias and variance. The model is more sensitive to changes in pixel values and noise in MODIS image data when k is smaller, namely k = 3 and k = 5. This is due to the fact that MODIS images with a spatial resolution of 4 km represent a fairly large area of about 16 km² in one pixel so that the values of sea surface temperature (SST) and chlorophyll-a are only average values and are unable to describe environmental variations at a smaller scale. This resolution is relatively coarse so it is difficult to identify changes in water conditions on a small scale, such as significant temperature differences between areas or upwelling zones that often occur in water areas (Drozd et al., 2025). Conversely, a large value of k, such as k = 9 makes the model consider more of its nearest neighbors including neighbors or points with different actual environmental conditions. Consequently, oversmoothing blurs the boundaries between potential and non-potential areas, reducing the classification rate. This aligns with the bias-variance principle in KNN a small k value tends to make the model highly sensitive to noise but has low bias, conversely a high k value improves generalization ability but can potentially lead to high bias (Mladenova & Valova, 2023).

A k=7 value results in more neighbors and successfully mitigates the influence of inaccurate data, meaning the model does not lose its sensitivity to patterns around the location. Conversely, at k=9, too many neighbors are included, including data from various sea conditions, reducing accuracy from 95.66% to 95.36% (Figure 4). Therefore, a k=7 value provides the best balance in reducing the impact of inaccurate data at 4 km resolution while maintaining the model's ability to identify patterns in Banten waters.

As a form of validation, the evaluation was carried out using a Confusion Matrix, which is a table that compares predicted labels with actual labels to calculate the level of classification accuracy. The Confusion Matrix has Precision, Recall, and F1-score values. The precision metric defines how many positive classifications are actually correct, divided by the total number of positive predictions. Pratiwi et al. (2020) state that recall demonstrates the system's ability to capture the percentage of genuinely positive data. To provide a balanced assessment of the model, the F1-Score combines both of these measurements.

Based on the confusion matrix results, the model only experienced a classification error of 4.3% (44 of 1,015 test data). This small and balanced error proportion between FN and FP indicates that there is no systematic bias in the model, but is caused by the uncertainty of KNN voting on border data due to the application of hard thresholds and the effect of pixel mixtures in coastal areas due to the resolution of MODIS 4 km imagery which shows the average value and is reinforced by cloud cover interference on satellite image sensors which can cause spatial data to be lost at certain points so that the interpolation results at the data processing stage may not represent the actual conditions at that point.

The model evaluation succeeded in achieving the best accuracy, namely 95.66 %. Al Iman et al. (2023) used KNN to classify various types of fish based on 2-Dimensional Linear Discriminant Analysis (2D-LDA), which showed that by using the ideal nearest neighbor parameter at k=9, the KNN algorithm can detect fish species with a very high accuracy of 93.12%. Similarly, Wiastopo & Imelda, (2024) using KNN to detect the freshness of mackerel based on color and texture features with an accuracy of 96%.

The model's success in accurately predicting potential classes demonstrates a strong linear correlation between SST and chlorophyll-a variables and

fish distribution. These results align with research showing that Bhavan et al. (2025) abiotic parameters such as Sea Surface Temperature (SST) and biotic variables such as chlorophyll-a, as indicators of water fertility, can be effectively used to estimate fish Presence.

**The Most Influential Variables on Fish Presence**

The results of the tests conducted using the Permutation Feature Importance method identified the chlorophyll-a variable as the most important feature with an importance level of 25.66%. This importance level indicates a significant decrease in model accuracy if the chlorophyll-a variable is removed from the classification process. This is in line with the statement of Sayad (2023) that two crucial indicators of aquatic productivity and fish Presence are sea surface temperature and chlorophyll-a.

The relationship between chlorophyll-a and SST influences the distribution of marine biota. Biologically, environmental changes affect the lives of small pelagic fish, which are highly sensitive to environmental changes. Therefore, the existence of small pelagic fish depends on changes in water temperature and food availability. The upwelling phenomenon, which produces high nutrient levels in the waters, acts as a primary trigger for phytoplankton growth because small pelagic fish are predominantly phytoplankton-feeders. According to Kurniawati *et al*. (2017), the chlorophyll-a range of 0.5–2.5 mg/m³ is the threshold for fertile waters without experiencing excessive algal blooms. If chlorophyll-a concentrations are high, the distribution of small pelagic fish is also high, but if chlorophyll-a levels are very high, there is concern that algal blooms will occur, which will negatively impact aquatic ecosystems (Prayitno et al., 2021). Small pelagic fish also tend to prefer relatively warm waters, with temperatures ranging from 28°C to 29°C for small pelagic fish survival (Syamsuddin *et al*., 2018). Prayitno *et al*. (2021) also stated that catches will increase when sea surface temperatures decrease, and when temperatures rise above the ideal limit, catches will relatively decrease. Small pelagic fish can adapt to temperatures between 28°C and 30°C in waters because temperature affects fish metabolism, and pelagic fish tend to move according to warmer temperatures (Rasyid, 2010 in the journal Sarwati *et al*., 2025). However, due to the dynamic nature of ocean waters, where water conditions can change and the constant movement of fish will adapt to water conditions, fishing activities can still occur outside of ideal fish conditions. This shows that all waters have the potential to catch fish, especially small pelagic fish.

An analysis conducted using the Generalized Additive Model (GAM) showed that the combination of chlorophyll-a and SST was the most suitable variable to explain the distribution pattern of mackerel in Banten Bay (Putri et al., 2025). This condition confirms the reason why these two variables were ranked highest in feature importance in the KNN model built**.** The study found Dewi et al. (2023) that chlorophyll- a had a stronger influence on fish catches than SST, further highlighting the importance of this variable in predicting fish Presence. The high contribution chlorophyll-a in this study proves that the KNN algorithm relies heavily on this primary productivity parameter to identify fishing ground patterns in Banten waters.

The results of interviews with local fishermen also stated that to determine the presence or absence of fish can be seen from the dark green color of the water which indicates the amount of chlorophyll and warm water temperature, this is done in a traditional way without tools which further highlights that the parameters of sea surface temperature and chlorophyll-a are important parameters in determining the presence of small pelagic fish. However, habitat selection also depends on the fish species, small pelagic fish have a lower tendency to prefer excessive chlorophyll content compared to large pelagic fish (Sarifah et al., 2024).

## CONCLUSION

The K-Nearest Neighbors algorithm performed very well in estimating the Presence of small pelagic fish during the east monsoon in Banten Waters, achieving an optimal accuracy of 95.66% at a data sharing ratio of 70:30 and a K value of 7. This modeling confirmed that primary productivity parameters, namely chlorophyll-a and Sea Surface Temperature, were influential variables in estimating fish Presence. Evaluation using a Confusion Matrix showed that the KNN model with K=7 at a ratio of 70:30 had balanced and precise classification capabilities. The use of the Euclidean Distance metric successfully confirmed that fishing grounds in Banten Waters tended to cluster, where water masses with similar temperatures (SST) and chlorophyll-a within a certain radius had the same potential probability level. Overall, the use of Machine Learning through the KNN method was able to provide good

classification regarding the estimation of the Presence of small pelagic fish. However, this study has limitations in its labeling scheme, which risks imbalanced data, which can directly degrade model performance. In data collection instruments, satellite imagery sensors are highly susceptible to cloud cover interference, potentially leading to missing data at certain water points.


## ACKNOWLEDGEMENT

Thanks are extended to all parties who have assisted in the implementation of this research, especially to the Indonesian Education University for its academic support, the fishermen at PPN Karangantu, the fishermen at TPI Bojonegara, lecturers, and fellow researchers.


## REFERENCES


Ali AA. 2020. Identifikasi dan pemberdayaan masyarakat miskin nelayan tradisional. *PONDASI*. *25*(1): 37–49.

Al Iman YD, Isnanto R, Nurhayati OD. 2023. Klasifikasi jenis ikan laut k-nearest neighbor berdasarkan ekstraksi ciri 2-dimensional linear discriminant analysis. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*. *10*(4): 919–926. https://doi.org/10.25126/jtiik2023106767

An C, Park YW, Ahn SS, Han K, Kim H, Lee S-K. 2021. Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results. *PLoS ONE*. 16(8): 1-13. https://doi.org/10.1371/journal.pone.0256152

Arif H, Saleh F, Jaya G. 2018. Pemanfaatan Citra Landsat 8 Oli/Tirs Untuk Penentuan Zona Potensi Penangkapan Ikan (ZPPI) Di Perairan Kabupaten Wakatobi. *Jurnal Geografi Aplikasi Dan Teknologi. 2(2):* 21–30.

Banjarnahor HP, Suprayogi A, Bashit N. 2020. Analisis pengaruh fenomena upwelling terhadap jumlah tangkapan ikan dengan pengamatan temporal citra aqua modis (Studi Kasus : Selat Bali). *Jurnal Geodesi Undip*. 9(2): 91–101.

Bhavan SG, Das B, Chakurkar EB, Thomas D, Vasudevan C, Don S. 2025. Modelling the Presence of key aquatic species in a tropical Indian estuary using biotic and abiotic predictors. *Regional Studies in Marine Science*. 85: 1–15. https://doi.org/10.1016/j.rsma.2025.104119

Dewi P, Sutarjo, Hermawan M, Yusrizal, Maulita M, Nurlaela E, Kusmedy, B, Danapraja S, Nugraha E. 2023. Study of sea surface temperature and chlorophyll-a influence on the quantity of fish caught in the waters of Sadeng, Yogyakarta, Indonesia. *BIOFLUX SRL*. 16(1): 110–127.

Drozd S, Kussul N, Shelestov A. (2025). Improving spatial resolution of Aqua MODIS and GCOM-C chlorophyll-a data for Cyprus coastal waters monitoring. *European Journal of Remote Sensing*, *58*(1): 1-21. https://doi.org/10.1080/22797254.2025.2551023

[DKP] Dinas Kelautan dan Perikanan Provinsi Banten. 2024. Laporan Statistik Perikanan Provinsi Banten Tahun 2024. Serang (ID): DKP Provinsi Banten.

Fitrianah D, Hidayanto AN, Gaol JL, Fahmi H, Arymurthy AM. 2016. A Spatio-Temporal Data-Mining Approach for Identification of Potential Fishing Zones Based on Oceanographic Characteristics in the Eastern Indian Ocean. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 9(8): 3720–3728. https://doi.org/10.1109/JSTARS.2015.2492982

Fofied FG, Hartoko A, Saputra SW. 2024. Analisis Sebaran Suhu Permukaan Laut, Klorofil-a, dan Zona Potensial Penangkapan Ikan Cakalang di Perairan Jayapura. *Buletin Oseanografi Marina Oktober*.13(3): 409–423. https://doi.org/10.14710/buloma.v13i3.63007

Hasdyna N, Dinata KR. 2020. Analisis Matthew Correlation Coefficient pada K-Nearest Neighbor dalam Klasifikasi Ikan Hias. *Informatics Journal*. 5(2):57-64.

Hendiarti N, Siegel H, Ohde T. 2004. Investigation of different coastal processes in Indonesian waters using SeaWiFS data. *Deep-Sea Research Part II: Topical Studies in Oceanography*. 51(1–3): 85–97.

Irnawati R, Surilayani D, Susanto A, Rahmawati A, Munandar A, Sari R, Nurdin HS. 2020. Analisis penentuan lokasi basis perikanan teri dan jalur pemasarannya di provinsi banten. *Jurnal Sosial Ekonomi Kelautan Dan Perikanan*. 15(2): 159-168. https://doi.org/10.15578/jsekp.v15i2.7989

Juniati D, Khotimah C, Wardani DEK, Budayasa K. 2018. Fractal dimension to classify the heart sound recordings with KNN and fuzzy c-mean clustering methods. *Journal of Physics: Conference Series*. 953(1): 1–9. https://doi.org/10.1088/17426596/953/1/012202

Khatami AM, Yonvitner, Setyobudiandi I. 2019. Karakteristik biologi dan laju eksploitasi ikan pelagis kecil di perairan utara jawa. *Jurnal Ilmu Dan Teknologi Kelautan Tropis*. 11(3): 637–651. https://doi.org/10.29244/jitkt.v11i3.19159

Kurnianto A, Sitanggang IS, Kusuma M, Hardhienata D. 2024. Klasifikasi Daerah Penangkapan Ikan Menggunakan Algoritma Random Forest dan Support Vector Machine Classification of Fishing Ground Using Random Forest and Support Vector Machine Algorithms. *Jurnal Ilmu Komputer & Agri-Informatika*. 11(2): 100–110.

Kurniawati F, Sanjoto TB, & Juhadi. (2015). Pendugaan Zona Potensi Penangkapan Ikan Pelagis Kecil Di Perairan Laut Jawa Pada Musim Barat Dan Musim Timur Dengan Menggunakan Citra Aqua Modis. *Geo Image*. 4(2): 9-19.

Kusuma FH, Ubaidillah MA, Ibadillah AF, Nahari RV, Joni K, Saputro AK. 2023. Sistem Identifikasi Kesegaran dan Jenis Ikan dengan Metode K-Nearest Neighbor Berdasarkan Citra Mata dan Bentuk Ikan. *Jurnal FORTECH*. 4(1): 33–41. https://doi.org/10.56795/fortech.v4i1.383

Lubis MGA, Palupi R, Astriani A, Khotimah PA. 2024. Model prediksi kelimpahan ikan tongkol (euthynnus affinis) menggunakan metode k-nearest neighbor (k-nn) di laut jawa. *Journal of Fisheries and Marine Research*. 8(3): 1–7.

Manurung J, Saragih H, Prabukusumo MA, Firdaus EA. 2025. Optimizing the performance of the K-Nearest Neighbors algorithm using grid search and feature scaling to improve data classification accuracy. *Jurnal Mandiri IT*. 14(2): 260–268.

Mladenova T, Valova I. 2023. Classification with K-Nearest Neighbors Algorithm: Comparative Analysis between the Manual and Automatic Methods for K-Selection. *International Journal of Advanced Computer Science and Applications*. 14(4): 396–404.

Nufus NI, Yulius, Agus SB, Arifin T, Putra A, Salim HL, Rahmania R, Prihantono J, Purbani D, Purwanto AD, Ramdhan M. (2026). Spatio-Temporal Variation Of Sea Surface Temperature, Chlorophyll-A, And Wind Direction On Euthynnus Affinis Fishing Productivity In Fisheries Management Area 712. *Egyptian Journal of Aquatic Biology & Fisheries*. 30(1): 63-81.

Nurkhairani Y, Supriatna S, Susiloningtyas D. 2018. Wilayah Potensi ikan pelagis pada variasi kejadian ENSO dan normal di Selat Sunda. *Jurnal Geografi Lingkungan Tropik*. 2(1):52–63. https://doi.org/10.7454/jglitrop.v2i1.32

Pratiwi BP, Handayani AS, & Sarjana. (2020). Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix. *Jurnal Informatika Upgris*. *6*(2): 66–75.

Prayitno L M, Rahman A, & Yasmi, Z. (2021). Aplikasi Data Citra Satelit Aqua-Modis Untuk Menentukan Produktivitas Primer Perairan Dengan Metode Sebaran Klorofil-A Dan Suhu Permukaan Laut Di Perairan Kalimantan Selatan. *AQUATIC*. 4(1): 1-11.

Putri D, Yulius Y, Rosalia AA, Arifin T, Putra A, Heriati A, Prihantono J, Purbani D, Salim HL, Hartati ST, Ramdhan M, Wahyono A, Rahmania R. 2025. Influence of sea surface temperature and chlorophyll-a on mackerel productivity in banten bay, indonesia: analysis using aqua modis data (2014–2023). *Geographia Technica* 20(1): 44–63. https://doi.org/10.21163/GT_2025.201.05

Sarifah A, Pratikto I, Suryono CA. 2024. Potensi Perikanan di Perairan Selatan Yogyakarta Ditinjau dari Sebaran Klorofil-a, Suhu Permukaan Laut, dan Particulate Organic Carbon Berbasis Citra Satelit Aqua MODIS. *Jurnal Kelautan Tropis*. 27(1): 187–196. https://doi.org/10.14710/jkt.v27i1.22269

Sarwati DE, Suryono CA, Suryono S. 2025. Pendugaan Daerah Tangkapan Ikan Pelagis Kecil di Perairan Pesisir Utara Jawa Tengah Berdasarkan Paremater Lingkungan Laut. *Jurnal Kelautan Tropis*. *28*(1): 107–117. https://doi.org/10.14710/jkt.v28i1.26262

Sayad YO. 2023. Mapping Potential Fishing Zones Using Remote Sensing Data and GIS: A Case Study of Moroccan Waters. In *Resbee Publishers Multimedia Research (MR)*. 6(2): 1-13.

Simbolon IN, Siburian H, Manik WA. 2024. Prediksi kualitas air sungai di jakarta menggunakan knn yang dioptimalisasi dengan pso. *Jurnal Informatika Dan Teknik Elektro Terapan*. 12(2): 1193–1203. https://doi.org/10.23960/jitet.v12i2.4191

Siringoringo EOH, Simbolon D, Wahju RI, Purwangka F. 2024. Produktivitas dan pola musim penangkapan cakalang di wilayah pengelolaan perikanan 572 productivity and seasonal pattern of skipjack tuna in fisheries management area 572. *Jurnal penelitian perikanan indonesia*. 30(2): 99–109. https://doi.org/10.15578/jppi.30.2.2024.99-109

Sivakumar M, Parthasarathy S, Padmapriya T. 2024. Trade-off between training and testing ratio in machine learning for medical image processing. *PeerJ Computer Science*. *10*: 1–17. https://doi.org/10.7717/PEERJ-CS.2245

Syamsuddin M, Sunarto & Yuliadi L. (2018). How Do El Niño Southern Oscillation Events Impact On Small Pelagic Fish Catches In The West Java Sea. *Iop Conference Series: Earth And Environmental Science*. *176*(1): 1-9. Https://Doi.Org/10.1088/1755-1315/176/1/012014

Takwir A, Rondonuwu AB, Wahidin N, Rahman AA, Giu LO, Erawan, MTF. 2021. Analisis Kejadian Upwelling Dan Daerah Potensial Penangkapan Ikan Tuna Di Perairan Teluk Tolo. *Jurnal Enggano*. 6(2): 238–252.

Tarigan DJ, Cahyadi DF, Agung SS, Yonanto L, Rahayu DB. 2020. daerah penangkapan ikan kembung (*rastrelligger sp*) di selat sunda pada musim peralihan. *Jurnal Teknologi Perikanan dan Kelautan*. 11(1): 63-79.

Ujianto NT, Gunawan, Fadillah H, Fanti AP, Saputra AD, Ramadhan IG. 2025. Penerapan algoritma K-Nearest Neighbors (KNN) untuk klasifikasi citra medis. *IT-Explore: Jurnal Penerapan Teknologi Informasi Dan Komunikasi*, 4(1): 33–43. https://doi.org/10.24246/itexplore.v4i1.2025.

Wiastopo JP, Imelda I. 2024. Deteksi Kesegaran Ikan Kembung dengan Metode KNN Berdasarkan Fitur GLCM dan RGB-HSV. *Jurnal TICOM: Technology of Information and Communication*. 13(1): 10–16.