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 (2026) http://dx.doi.org/10.29244/jppt.10.1.24-35

Performance Analysis of the K-Nearest Neighbours (KNN) Algorithm on the Classification of Small Pelagic Fish Abundance in the East Season in Banten Waters

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

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

ABSTRACT
Small pelagic fish are the dominant commodity in the fishing sector in the waters of Banten, accounting for 51.3% of total production in 2023. However, limited access to technology means that conventional fishermen in this region still rely on experience and intuition to determine fishing locations, resulting in unstable catch yields and inefficiencies in operational costs, time, and labour. This study is designed to analyse the performance of the K-Nearest Neighbours (KNN) algorithm in classifying the abundance of small pelagic fish during the eastern season in the waters of Banten. Two oceanographic parameters were used as predictors: Sea Surface Temperature (SST) and chlorophyll-a concentration, both of which serve as indicators of aquatic ecosystem productivity. Secondary data were sourced from Aqua MODIS satellite imagery spanning 2021–2025, which were subsequently processed and integrated using the Kaggle and Google Colab platforms. The final dataset comprises 3,382 data points divided into two classes: potential and non-potential. Model testing was conducted by varying the K value (3, 5, 7, and 9) and the data split ratio (60:40, 70:30, and 80:20), using Euclidean Distance and the Confusion Matrix as performance evaluation metrics. The model with K=7 and a 70:30 split ratio yielded the best performance, with an accuracy of 95.66% and proportional values of precision, recall, and F1-score across both classes. Permutation Feature Importance analysis indicates that chlorophyll-a contributes dominantly at 25.66%, while the recorded influence of SPL is 23.58%. These results confirm the superiority of the KNN algorithm in classifying small pelagic fish in the waters of Banten.

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

INTRODUCTION

Indonesia has enormous maritime potential, but its utilisation rate has reached only 58.80% (Dahuri, 2003, in Arif *et al*., 2018). One of the strategic water areas with strong potential is the Banten Waters, where the water masses of the Java Strait, the Indian Ocean, and the Sunda Strait meet (Irnawati *et al*., 2020). According to the Department of Maritime Affairs and Fisheries (DKP), small pelagic fish groups dominate the Banten Waters, contributing around 51.3% (28,140.5 tons) of total capture fisheries production in 2023. 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 abundance 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 significantly influence oceanographic conditions in marine waters, including the east monsoon. The east monsoon

[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.

(June–September) is one of the seasons that affect aquatic productivity through upwelling in several parts of Indonesian waters. Factors such as upwelling, temperature, and chlorophyll-a distribution can influence the presence and abundance of fish in the water, leading many to be caught while searching for food (Takwir *et al*., 2021). During the east monsoon, winds show increased distribution and speed, which directly alter the direction of currents in the ocean's surface layer, indicating that wind movement is a major variable influencing the physical properties of seawater masses in the region (Banjarnahor *et al*., 2020). Changes in wind direction every three years in Indonesia result in significant adjustments in the direction and speed of ocean surface currents. These fluctuations affect oceanographic conditions, especially during the east monsoon, when currents flow from the east or southeast into western Indonesia. (Syahailatua & Wouthuyzen, 2023). According to Nurkhairani *et al*. (2018), the east monsoon is a good season for catching pelagic fish because water temperatures are relatively warm and nutrient levels are high. The abundance of pelagic fish in the waters is influenced by nutrient levels (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 for small pelagic fish (Sarianto, 2018; Fofied *et al*., 2024). One Artificial Intelligence technology to solve this problem is Machine Learning. The K-Nearest Neighbours Algorithm (KNN) is easy to understand, efficient, and has simple data preparation. (Simbolon *et al*., 2024). The KNN algorithm will predict SPL and chlorophyll-a data as x variables, then evaluate fish abundance 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 Neighbours (KNN) algorithm to predict fish abundance based on SST and rainfall, achieving an accuracy of 83%, but did not include key productivity parameters such as chlorophyll-a. In other research in the coastal waters of Central Java, Sarwati *et al*. (2025) used chlorophyll-a and SST variables to estimate small-pelagic-fish fishing grounds. However, the approach remained descriptive, relying on regression analysis and spatial modelling. According to Fitrianah *et al*. (2016), the K-Nearest Neighbours (KNN) algorithm, based on spatial and temporal data, can identify potential fishing zones with an accuracy of up to 87.11%. However, researchers used all seasons to estimate fish abundance, which is susceptible to bias because environmental conditions differ across seasons.

The purpose of this research is to analyse the performance of the K-Nearest Neighbours algorithm in classifying the abundance of small pelagic fish during the east season in Banten Waters. It is expected to optimise the effectiveness of fishing operations and provide information for the management of fishery resources 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 potential for fishery resources, with dynamic oceanographic conditions influenced by the confluence of water masses from the Java Sea, the Indian Ocean, and the Sunda Strait.

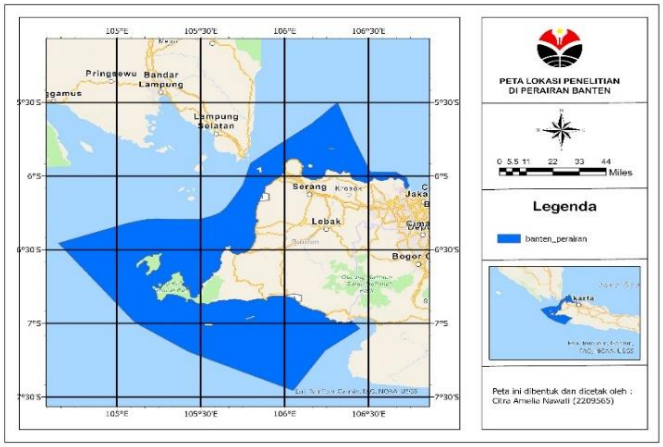

Figure 1. Research Location

Tools and Materials

For data processing, the Kaggle platform is used for the pre-processing stage, such as cropping based on the research area. Moreover, combining the SST and chlorophyll datasets makes the data ready for use. Next, Google Colab is used to implement and analyse the K-Nearest Neighbours (KNN) algorithm. The ArcGIS application is used to cut water boundaries within the research area and to create a map layout; Microsoft Word and Excel are used to support data pre-processing. The materials used include oceanographic parameters, such as Sea

Surface Temperature (SST) and Chlorophyll-a concentration, sourced from the Aqua MODIS satellite and downloaded from the NASA OceanColor website.

Methode

The researcher used a quantitative research method. supported by qualitative data. Data
For this study, data were obtained from the official NASA Ocean Colour Web (*https://oceancolor.gsfc.nasa.gov/*), downloaded on April 1, 2026, and are 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 on marine areas worldwide, including the boundaries of seas, straits, and established oceanographic zones.

The downloaded satellite image data covers 2021-2025, with a 4 km resolution and a seasonal temporal resolution during the east monsoon (June-September). The image data includes spatial data on sea surface temperature (SST) and chlorophyll-a to identify areas suspected of harbouring many small pelagic fish that like the area. The pre-processing stage involves downloading and cropping the SST and chlorophyll data for the research area, converting the formats, and preparing the data for processing. The downloaded NC-formatted file is converted to CSV using Kaggle software. After data pre-processing, 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 2- 8 °C and 29.5 °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 Abundance Potential Class Labels

Potential (Class)	SPL (°C)	Chlorophyll-a (mg/m³)
1 (Potential)	28 °C – 29.5 °C	0.5 – 2.5 mg/m3
0 (No Potential)	<28°C or > 29.5°C	< 0.5 or > 2.5 mg /m3

Source: (Hendiarti *et al*., 2004; Trenggono *et al*., 2018)

The dataset was tested with several data split ratios, starting with 60:40, 70:30, and 80:20. It was then used with the K-Nearest Neighbours algorithm, using Euclidean Distance calculations to classify coordinate points based on their nearest neighbours. The final stage of the research was evaluating the model's performance using a Confusion Matrix.

Feature standardisation is also performed using the StandardScaler method during the data processing stage. Standardisation is performed to ensure that the original data has uniform values. The K-Nearest Neighbours (KNN) algorithm is a distance-based classification method that is sensitive to feature scaling and parameter configuration. Mismatched feature scales and an unstructured optimisation process can lead to a decline in model performance if the k value is poorly selected (Manurung *et al*., 2025).

Data analysis

a. K-Nearest Neighbours Model Construction

This modelling stage, which has undergone data pre-processing, will be implemented using the KNN algorithm. The K-Nearest Neighbours (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 Neighbours as an 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 is similar to or dissimilar to the data to be classified. KNN is also an easy-to-understand, efficient algorithm, and its data training is simple, which can be a distinct advantage. The steps of the KNN algorithm include (Arsyad, 2020) in the journal (Hasdyna & Dinata, 2020):

1) Split the training data and test data from the prepared data

2) Determining the number of K as many data points 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 are used to determine the proximity of training data, namely the Euclidean Distance. To calculate the square of the distance using the Euclidean method, use the following mathematical formula:

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

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:

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

b. Model Evaluation

Accuracy testing during the model evaluation stage assesses 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 analysed using the Confusion Matrix is checked to determine whether it aligns with 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 points that can be predicted correctly or incorrectly and consists of the following four main variables:

- True Positive (TP): Positively labelled data that is accurately predicted as positive by the algorithm.
- False Positive (FP): Negatively labelled 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, including accuracy, precision, recall, and F-score. The formula for evaluating various metrics, according to Powers (2020) in the journal (Kurnianto *et al*., 2024), is as follows.

$$\text{accuracy} = \frac{TP+TN}{/(TP+TN+FP+FN)} \quad (3)$$

$$\text{precision} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{recall} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}} \quad (6)$$

RESULTS AND DISCUSSION

Results

Data Processing

Data pre-processing 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 2. Data Pre-processing Result

The number of datasets in Figure 2 was 3,382, with a fairly balanced class distribution: 1,700 for class 1 and 1,682 for class 0. StandardScaler is also used to ensure that the distance calculation in the KNN algorithm is not biased by differences in units

between coordinates (degrees) and chlorophyll (mg/m3).

Model Testing Results

K-Nearest Neighbours Method with K Value = 3

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

   The first trial utilised 3,382 data points, partitioned into a 60:40 ratio, with 2,029 training samples and 1,353 testing samples, and 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 with k = 3, totalling 3382 data points. The training data is 2367, and the test data is 1015, resulting in an accuracy of 94.97% with 964 correct predictions and 51 incorrect predictions. With the precision, recall, and F1-score values for each class: class 0 has precision 95%, recall 95%, and F1 score 95%; class 1 has precision 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 split ratio and a k-value of 3. By allocating 2,705 records to training and 677 to testing, the algorithm achieved 94.38% accuracy, correctly predicting 639 instances and 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 Neighbours 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 points were used for training and 1,353 for testing. The model reached 95.26% accuracy, getting 1,289 predictions right and missing only 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 achieved 96% precision, 95% 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 total dataset of 3,382 was split 70:30, yielding 2,367 for training and 1,015 for testing. Of these, 967 predictions were spot-on, and 48 were off, yielding 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 achieved 96% precision, 95% recall, and an F1-score of 95%.
3. Data distribution ratio: 80% training data and 20% testing data

   Testing for k = 5 uses an 80:20 data-sharing scheme with a total of 3382 data points. The training data is 2705, and the test data is 677, resulting in an accuracy of 94.97% with 643 correct predictions and 34 incorrect predictions. With the precision, recall, and F1-score values of 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 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 Neighbours Method with K Value = 7

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

   In the k=7 scenario, the system processed 3,382 data points partitioned 60:40 into 2,029 training and 1,353 testing data. Of the test set, 1,286 predictions were accurate, and 67 were incorrect, yielding an accuracy of 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 used a 70:30 split across the 3,382 data points. Out of the 2,367 training and 1,015 testing samples, the algorithm nailed 971 predictions and stumbled on just 41, boosting accuracy to 95.66%. A closer look at the breakdown shows that class 0 achieves 95% precision, 96% recall, and 96% F1-score. Class 1 matched this consistency, achieving 96% precision, recall, and F1-score.
3. Data distribution ratio: 80% training data and 20% testing data

   Testing for the value of known = 7 uses an 80:20 data-sharing scheme with a total of 3382 data points. The training data is 2705, and the test data is 677, resulting in an accuracy of 95.12% with 644 correct predictions and 33 incorrect predictions. With the precision, recall, and F1-score values for each class: class 0 has precision 95%, recall 95%, and F1 score 95%; class 1 has precision 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 Neighbours Method with K Value = 9

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

   With k = 9 and a 60:40 distribution, the algorithm processed 2,029 training samples and evaluated 1,353 test samples. Out of the 3,382 total instances, the system made 1,283 accurate predictions and 70 incorrect ones, for 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 with the k=9 test and a 70:30 split, the 3,382 data points were split 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 achieved 96% precision, 95% recall, and 95% F1-score.
3. Data distribution ratio: 80% training data and 20% testing data

   Testing for k = 9 uses an 80:20 data split with a total of 3382 data points. The training data is 2705, and the test data is 677, resulting in an accuracy of 94.68% with 641 correct predictions and 36 incorrect predictions. With the precision, recall, and F1-score values for each class: class 0 has precision 94%, recall 95%, and F1 score 95%; class 1 has precision 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

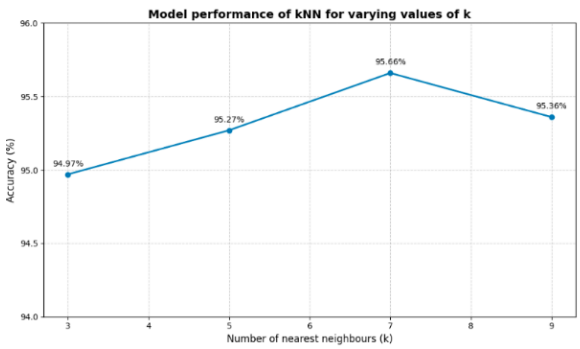

Figure 3. Graph of K Value at 70:30 ratio

As shown in Figure 3 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 model's excellent, balanced performance in classifying both classes is shown in the Confusion Matrix visualisation at the k value with the highest accuracy, namely K=7.

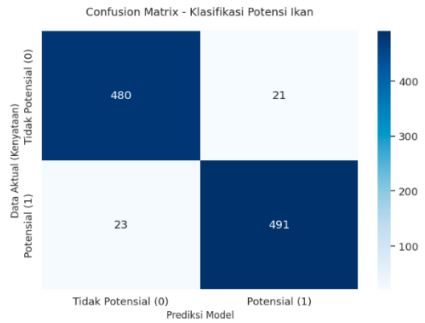

Figure 4. Confusion Matrix Value K=7

Figure 4 shows 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 greatest influence on fish abundance in Banten waters, as calculated using Permutation Importance, are chlorophyll-a at 25.66% and SPL at 23.58%.

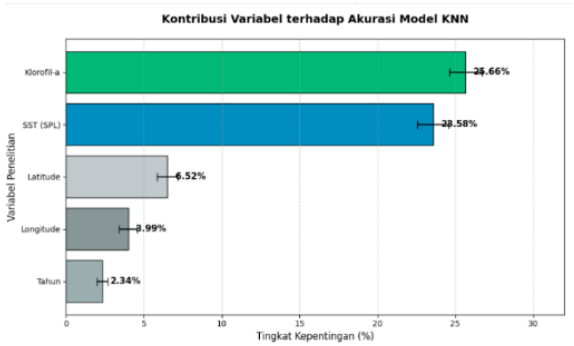

Figure 5. Permutation Importance Graph

The effect of each parameter is shown in Figure 6; if these parameters are removed, the model will be affected by 6.52% for longitude, 3.99% for latitude, and 2.34% for year.

K-Nearest Neighbours Modelling 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.

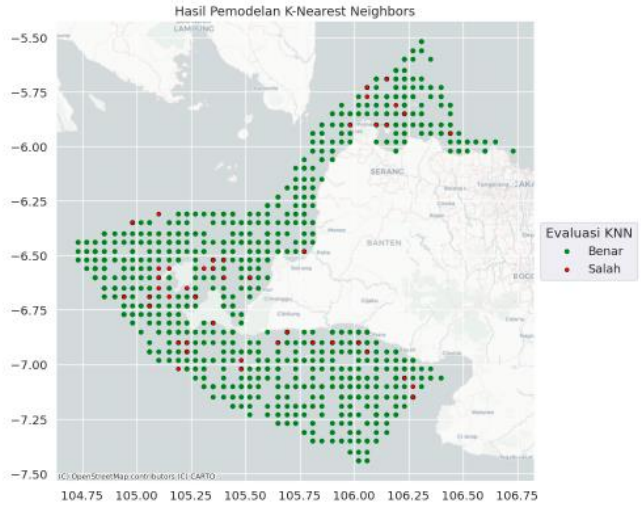

Figure 6. KNN Modelling Results

Overall, the model demonstrated high consistency, with 971 correct predictions and only 44 incorrect, supported by a stable F1-score of 0.96 for both categories. A visualisation of the KNN model is shown in Figure 6 above.

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 Neighbours algorithm. Overall, the data-sharing scheme with 70% training and 30% testing data has the most stable accuracy; the highest accuracy is 95.66% at K = 7. In contrast, the 60:40 scheme with k = 3 yields an accuracy of 95.49%. However, the trend in this ratio's accuracy decreases as the K value increases. This occurs even though there is more test data (1,353 data points). The limited training data (2,029 data points) makes it difficult for the model to generalise as K increases. The quality of a model's predictions depends heavily 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 overly simple, it struggles to capture the complex relationships in the data (Sivakumar *et al*., 2024).

The 80:20 test results show that this ratio has the lowest accuracy among the schemes but is still considered good. This indicates that test data with an 80:20 ratio exhibit greater variability or contain more complex outliers, which can make the model difficult to generalise, especially when the test and training data patterns differ (An *et al*., 2021). Research shows that the 70:30 ratio more often produces better accuracy than other ratios, due to a lower risk of overfitting from a more balanced training-to-testing split (Sivakumar *et al*., 2024).

Selection and Evaluation of K Value at the Best Ratio

KNN classification heavily depends on the accuracy of the chosen K value. A large K value may lead to classification errors, whereas a small K value is sensitive to outliers (Ujianto *et al*., 2025). The best accuracy occurs with the 70:30 ratio scheme and k = 7; this indicates that adding more k values helps the model generalise more stable data patterns. Small k values tend to make the model highly sensitive to noise but low-biased; conversely, high k values improve generalisation but can lead to high bias (Mladenova & Valova, 2023). The value k = 7 provides a better balance between bias and variance than other k values, depending on the dataset.

As a form of validation, the evaluation used a Confusion Matrix, a table that compares predicted and actual labels to calculate classification accuracy (Sowmiya *et al*., 2024). The Confusion Matrix has Precision, Recall, and F1-score values. The precision metric is the number of correctly classified positive instances 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. The model evaluation achieved the best accuracy, namely 95.66 %. Al Iman *et al*. (2023) used KNN to classify various fish types using 2-Dimensional Linear Discriminant Analysis (2D-LDA), showing that with the optimal k=9, the KNN algorithm achieved a very high accuracy of 93.12%. Similarly, Wiastopo & Imelda (2024) used KNN to detect mackerel freshness based on colour and texture features, achieving 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, can be used effectively to estimate fish abundance as indicators of water fertility.

The Most Influential Variables on Fish Abundance

The Permutation Feature Importance tests identified chlorophyll-a 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 aligns with Sayad's (2023) statement that two crucial indicators of aquatic productivity and fish abundance are sea surface temperature and chlorophyll-a.

The relationship between chlorophyll-a and SST influences the distribution of marine biota. An analysis using the Generalised Additive Model (GAM) showed that the combination of chlorophyll-a and SST was the most suitable variable for explaining the distribution pattern of mackerel in Banten Bay (Putri *et al*., 2025). This condition confirms why these two variables were ranked highest in feature importance in the KNN model. Dewi *et al*. (2023) found that chlorophyll-a had a stronger influence on fish catches than SST, further highlighting its importance in predicting fish abundance. The high contribution of 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. However, habitat selection also depends on the fish species; small pelagic fish tend to prefer lower chlorophyll levels than large pelagic fish (Sarifah *et al*., 2024).

CONCLUSION

The K-Nearest Neighbours algorithm performed very well in estimating the abundance 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 modelling confirmed that primary productivity parameters, namely chlorophyll-a and Sea Surface Temperature, were influential variables in estimating fish abundance. Evaluation using a Confusion Matrix showed that the KNN model with K=7 at a 70:30 split achieved balanced, precise classification performance. The use of the Euclidean Distance metric successfully confirmed that fishing grounds in Banten Waters tended to cluster, with water masses with similar temperatures (SST) and chlorophyll-a within a certain radius having the same probability level. Overall, the use of Machine Learning via the KNN method yielded good classification of small pelagic fish abundance.

ACKNOWLEDGEMENT

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

REFERENCES

Ali, A. A. (2020). Identifikasi Dan Pemberdayaan Masyarakat Miskin Nelayan Tradisional. *Pondasi*, *25*(1), 37–49.

Al Iman, Y. D., Isnanto, R., & Nurhayati, O. D. (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, Y. W., Ahn, S. S., Han, K., Kim, H., & Lee, S. K. (2021). Radiomics machine learning study with a small sample size: A single random training-test set split may lead to unreliable results. *PLoS ONE*, *16*(8 August). 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, H. P., 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, S. G., Das, B., Chakurkar, E. B., Thomas, D., Vasudevan, C., & Don, S. (2025). Modelling the abundance 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.

Fitrianah, D., Hidayanto, A. N., Gaol, J. L., Fahmi, H., & Arymurthy, A. M. (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, F. G., Hartoko, A., & Saputra, S. W. (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, K. R. (2020). Analisis Matthew Correlation Coefficient pada K-Nearest Neighbour dalam Klasifikasi Ikan Hias. In *Informatics Journal* (Vol. 5, Number 2).

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, H. S. (2020). ANALISIS PENENTUAN LOKASI BASIS PERIKANAN TERI DAN JALUR PEMASARANNYA DI PROVINSI BANTEN. *Jurnal Sosial Ekonomi Kelautan Dan Perikanan*, *15*(2), 159. https://doi.org/10.15578/jsekp.v15i2.7989

Juniati, D., Khotimah, C., Wardani, D. E. K., & 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/10.1088/1742-6596/953/1/012202

Khatami, A. M., 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, I. S., 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. http://journal.ipb.ac.id/index.php/jika

Kusuma, F. H., Ubaidillah Ms, A., Ibadillah, A. F., Nahari, R. V., Joni, K., & Saputro, A. K. (2023). Sistem Identifikasi Kesegaran dan Jenis Ikan dengan Metode K-Nearest Neighbor Berdasarkan Citra Mata dan Bentuk Ikan. *Journal FORTECH*, *4*(1), 33–41. https://doi.org/10.56795/fortech.v4i1.383

Lubis, M. G. A., Palupi, R., Astriani, A., & Khotimah, P. A. (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. http://jfmr.ub.ac.id

Manurung, J., Saragih, H., Prabukusumo, M. A., & Firdaus, E. A. (2025). Optimising the performance of the K-Nearest Neighbours algorithm using grid search and feature scaling to improve data classification accuracy. *Jurnal Mandiri IT*, *14*(2), 260–268. www.ejournal.isha.or.id/index.php/Mandiri

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

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

Putri, D., Yulius, Y., Rosalia, A. A., Arifin, T., Putra, A., Heriati, A., Prihantono, J., Purbani, D., Salim, H. L., Hartati, S. T., 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., Adhi Suryono, C., Ilmu Kelautan, D., Perikanan dan Ilmu Kelautan, F., Diponegoro Jl Jacub Rais, U., & Tengah, J. (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 Maret*, *27*(1), 187–196. https://doi.org/10.14710/jkt.v27i1.22269

Sarwati, D. E., Suryono, C. A., & 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, Y. O. (2023). Mapping Potential Fishing Zones Using Remote Sensing Data and GIS: A Case Study of Moroccan Waters. In *Resbee Publishers doi: 1* (Vol. 6, Number 2).

Simbolon, I. N., Siburian, H., & Manik, W. A. (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, E. O. H., Simbolon, D., Wahju, R. I., & 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. (2024a). 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

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

Sowmiya, N. K. J., Manimaran, & Asaithambi. (2024). Using Decision Risk and Decision Accuracy Metrics for Decision Making for Remote Sensing and GIS Applications. *Springer Science and Business Media Deutschland GmbH*, *398*, 125–136.

Syahailatua, A., & Wouthuyzen, S. (2023). Implikasi Upwelling terhadap Produktivitas Perikanan Laut di Indonesia dan Upaya Konservasinya. In *Pengelolaan Sumber Daya Perikanan Laut Berkelanjutan* (pp. 221–266). Penerbit BRIN. https://doi.org/10.55981/brin.908.c758

Takwir, A., Rondonuwu, A. B., Wahidin, N., Rahman, A. A., Giu, L. O., & Erawan, M. T. F. (2021). Analisis Kejadian Upwelling Dan Daerah Potensial Penangkapan Ikan Tuna Di Perairan Teluk Tolo. *Jurnal Enggano*, *6*(2), 238–252.

Tarigan, D. J., Cahyadi, D. F., Agung, S. S., Yonanto, L., & Rahayu, D. B. (2020). DAERAH PENANGKAPAN IKAN KEMBUNG (Rastrelligger sp) DI SELAT SUNDA PADA MUSIM PERALIHAN POTENTIAL FISHING ZONES Rastrelligger sp IN SUNDA STRAIT IN TRANSITIONAL SEASON. In *Jurnal Teknologi Perikanan dan Kelautan* (Vol. 11, Number 1). http://www.oceancolor.gsfc.nasa.

Trenggono, M., Amron, A., Avia Pasha, W., & Lazuardy Rolian, D. (2018). Effects of El Niño on the distribution of chlorophyll-a and sea surface temperature in the northern to southern Sunda Strait. *E3S Web of Conferences*, *47*, 1–11. https://doi.org/10.1051/e3sconf/20184705004

Ujianto, N. T., Gunawan, F., H., F., A. P., S., A. D., & Ramadhan, I. G. (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.pp33-43

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