Prediction of Potential Fishing Zones Using K-Means Clustering and Random Forest in Batam Waters

Sarah Astiti[1*], Alvendo Wahyu Aranski[2], Darmansah[3]

[1]*Information Systems Department, Faculty of Industrial Engineering, Universitas Telkom, Indonesia*
[2]*Iformation Systems Department, Faculty of Information Technology, Institut Teknologi Batam, Indonesia*
[3]*Information Systems Department, Faculty of Engineering and Computers, Universitas Putera Batam, Indonesia*

*corr-author: sarahas@telkomuniversity.ac.id

Abstract - Identification of potential fishing zones remains a significant challenge in fisheries management, particularly in coastal and island waters characterized by high spatial and temporal environmental variability. In Batam waters, fishing activities are still dominated by fishermen's experience and heuristic judgment, while existing studies often focus on a single prediction model or limited environmental parameters. This indicates a research gap, namely the lack of an integrated framework that simultaneously captures environmental heterogeneity and improves prediction accuracy using a data-driven approach. To address this gap, this study proposes a hybrid data mining framework that explicitly integrates unsupervised environmental zoning and supervised classification for predicting fishing potential. Weather and oceanographic variables—including sea surface temperature, chlorophyll-a concentration, wind speed, ocean current speed, and salinity—are used in conjunction with historical fish catch data. K-Means clustering is first used to identify homogeneous marine environmental zones, which are then incorporated as contextual features into a Random Forest classification model. Model performance is then evaluated using accuracy, precision, recall, F1 score, and confusion matrix analysis. The results show that the proposed hybrid approach achieves an accuracy of 89.2% and an F1 score of 89.1%, representing a quantitative improvement of approximately 5.6% in accuracy and 5.0% in F1 score compared to the baseline Random Forest model without clustering. This comparison clearly demonstrates that the integration of clustering information significantly improves classification performance. Furthermore, feature importance analysis confirms that sea surface temperature and chlorophyll-a concentration are the most influential predictors, while cluster labels contribute indirectly by improving the model's contextual understanding of complex environmental conditions. The novelty of this research is articulated through the integration of unsupervised marine environmental zoning with supervised machine learning in a local fisheries context, which allows for improved predictive performance and enhanced model interpretability. Unlike conventional approaches that treat environmental variables independently, the proposed framework captures multidimensional environmental interactions in a structured manner. The implications of these findings are profound. The proposed model can support data-driven decision-making for fishermen by reducing search time and operational costs, while providing a scientific basis for fisheries managers for spatial planning and sustainable resource management. Therefore, this research contributes both methodologically and practically to the advancement of intelligent fisheries prediction systems in dynamic coastal environments such as Batam waters.

Keywords: Potential fishing areas; oceanographic data; weather data; K-Means clustering; Random Forest.

I. INTRODUCTION

Indonesia is recognized as one of the world's largest archipelagic countries, with marine areas covering more than two-thirds of its total territory. This geographical condition positions the fisheries sector as a strategic component in supporting food security, regional economic growth, and national development. However, despite its vast marine potential, the utilization of fisheries resources remains suboptimal due to limited access to accurate information regarding the spatial and temporal distribution of fish resources [1]. In tropical coastal areas like Batam, fish conditions depend on atmospheric conditions and oceanographic processes, such as sea surface temperature, chlorophyll concentration, ocean currents, wind, and salinity. However, fishing activities in this area are still dominated by fishermen's experience and intuition, which often results in suboptimal fishing routes and increased fuel consumption [2, 3]. These variables are highly dynamic and vary both spatially and temporally, making conventional fishing strategies based on

experience and intuition increasingly unreliable, particularly under changing climate conditions [4, 5]. The waters surrounding Batam Island represent a complex marine environment due to their strategic location between the South China Sea and the Malacca Strait. This region is characterized by strong seasonal variability, monsoonal wind patterns, and intensive maritime activities, all of which contribute to highly fluctuating oceanographic conditions [6, 7].

Advances in data acquisition technologies, such as satellite remote sensing and in-situ marine observations, have enabled the availability of large-scale environmental datasets. The increasing volume and complexity of these datasets necessitate advanced analytical approaches capable of extracting meaningful patterns and knowledge. In this context, data mining and machine learning techniques have emerged as powerful tools for analysis complex environmental data and supporting decision-making processes in fisheries management [8, 9]. One commonly used unsupervised learning method in marine and fisheries studies is K-Means clustering. This algorithm groups data into clusters based on similarity, allowing researchers to identify regions with homogeneous environmental characteristics without prior labeling [10, 11]. Previous studies have demonstrated the effectiveness of K-Means clustering in classifying marine regions based on oceanographic parameters, thereby facilitating the identification of zones with similar ecological conditions relevant to fish habitats [12, 13]. In addition to clustering, predictive modeling techniques are required to estimate the potential of fishing grounds quantitatively. Random Forest, an ensemble learning algorithm based on decision trees, has shown superior performance in classification and prediction tasks involving complex, nonlinear relationships among variables [14, 15]. Random Forest is particularly suitable for environmental applications due to its robustness to noise, ability to handle multicollinearity, and capacity to assess variable importance [16]. Several fisheries-related studies have successfully applied Random Forest to predict fish distribution and potential fishing zones using oceanographic and climatic data [15]. The integration of clustering and classification methods provides a comprehensive analytical framework [17]. K-Means clustering enables the identification of environmental patterns, while Random Forest leverages these patterns to build predictive models for potential fishing grounds [10]. This hybrid approach has been shown to improve prediction accuracy and interpretability compared to single-method approaches [17].

Despite previous research findings, clustering and classification are still used as independent or alternative methods rather than integrated into a unified prediction framework. Furthermore, many studies focus solely on improving prediction accuracy, without addressing model interpretation or how environmental interactions influence fishing probability. Furthermore, local research has been limited in Indonesian coastal waters, particularly Batam, despite the area exhibiting complex oceanographic dynamics due to its proximity to major shipping lanes and regional current systems. This situation demonstrates a clear research gap; there is no integrated data mining framework that combines environmental zoning and supervised classification to improve prediction and contextual interpretation of fisheries potential, especially in dynamic and localized coastal environments such as Batam waters.

This study proposes a hybrid K-Means–Random Forest framework to predict fisheries zones in Batam waters using oceanographic and weather data. This study differs from previous studies that rely on a single prediction model. This study identifies homogeneous marine environmental zones through unsupervised clustering. These areas are then incorporated as contextual features into a supervised Random Forest classifier. With this integration, the model can capture multidimensional environmental interactions while maintaining highly accurate predictions and interpretations. Therefore, this study contributes in several aspects: introducing a hybrid methodological framework that structurally combines zoning and environmental classification and applies machine learning algorithms separately. Then, it provides a predictive model that can be explained through feature importance analysis. It links data-driven results with established fisheries oceanography theory. Finally, it provides an empirical evaluation of the proposed framework in Batam waters.

II. METHOD

This study adopts a hybrid data mining framework combining unsupervised learning (K-Means clustering) and supervised learning (Random Forest classification) to predict potential fishing zones based on weather and oceanographic data. In general, the research methodology flow consists of five main stages, namely data collection, data pre-processing, marine environment clustering, classification of fishing potential, and prediction output, as shown in the research methodology diagram in Fig. 1.

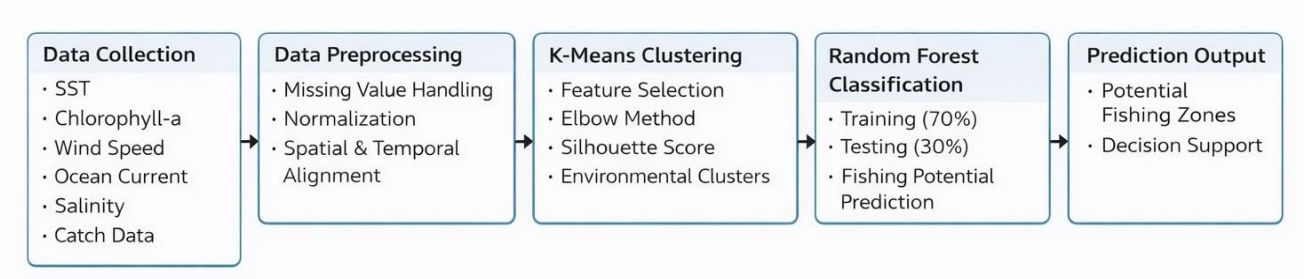

Fig. 1 Research methodology flowchart

This flowchart visually explains the methodological contributions, emphasizes the hybrid approach, and makes it easier for reviewers to understand the research flow.

A. Data Collection

The first stage involved collecting environmental and fisheries data from the Batam City Maritime Affairs and Fisheries Service (DKP). First, oceanographic and weather data were collected. This included sea surface temperature (SST), chlorophyll-a concentration, wind speed, ocean current speed, and salinity. These variables were selected based on their proven influence on fish distribution in the fisheries oceanography literature. Historical fish catch data were used to define fishing potential classes. While weather and oceanographic variables served as predictors, catch data served as a reference for determining fishing potential classes.

B. Data Preprocessing

To ensure data compatibility and quality before model development, the data preprocessing stage consists of the second block in Figure 1 above. This stage includes spatial and temporal alignment across the dataset, normalization of numerical features to eliminate scale bias, and handling of missing values using appropriate statistical techniques. While temporal alignment synchronizes data collected at different time intervals, spatial alignment ensures that all variables fit within the same geographic grid. This preprocessing stage is crucial, as K-Means clustering and Random Forest classification are susceptible to inconsistencies in data scale and structure.

C. K-Means Clustering

The third block in Fig. 1 illustrates the application of K-Means clustering to identify homogeneous environmental zones. The third block in Figure 1 illustrates the application of K-Means clustering to identify homogeneous environmental zones. The K-Means method groups data into groups based on similar attribute values. The average value of the cluster where the data point is located is used to fill in missing values [18]. The K-Means algorithm is also said to be a non-hierarchical cluster analysis method where the number of groups to be formed is predetermined [19]. K-Means clustering is applied as an unsupervised learning technique to group marine environments into homogeneous clusters based on similarities in weather and oceanographic characteristics. The optimal number of clusters is determined using the Elbow Method and Silhouette Score. This step aims to uncover latent environmental patterns and reduce data complexity by identifying zones with similar oceanographic conditions that may influence fish presence. Then, here are the formulas for calculating Euclidean Distance, Cluster Assignment Rule, and Centroid Update Formula such as (1), (2), and (3).

1) Euclidean Distance: Euclidean Distance is used to measure the straight distance between two points in space.

$$d(r_i, c_j) = \sqrt{\sum_{k=1}^{n} (r_{rk} - r_{jk})^2} \quad (1)$$

Description:
r_i = data to i
c_j = cluster centroid to j
n = number of attributes

2) Cluster Assignment Rule: Cluster Assignment Rule is a rule used to determine which cluster data goes into in a clustering algorithm.

$$Cluster\ (r_i)\ \frac{argmin}{j}\ d(r_i, c_j) \quad (2)$$

Description:
x_i = data to i
c_j = cluster centroid to k
$C(ri)$ = distance between data and centroid
arg min = choose the k value with the smallest distance

3) Centroid Update Formula: Centroid Update Formula is used to update the position of the cluster center (centroid) after the data is grouped.

$$c_j = \frac{1}{N_j} \sum_{i-1}^{N_j} r_i \quad (3)$$

Description:
c_j = centroid of cluster j
N_j = centroid of cluster j
r_i = data in cluster j

D. Random Forest Classification

As shown in the fourth block of Fig. 1, random forest classification is used to estimate the probability of fishing. The Random Forest algorithm is a supervised learning algorithm in which there are two mechanisms, namely test data and training data [20]. Random Forest is a supervised ensemble learning algorithm that uses bootstrap sampling and random feature selection to construct multiple decision trees. Majority selection among all trees produces the final classification result, which increases robustness and reduces overfitting. The clustered data are subsequently used as inputs for the Random Forest classification model. The dataset is divided into training (70%) and testing (30%) subsets. Random Forest is employed to predict potential fishing zones due to its ability to handle nonlinear relationships, multicollinearity, and noisy data. The model also enables the assessment of variable importance, providing insights into the relative influence of each environmental factor. The following is the formula used in this algorithm which can be seen in (4), (5), and (6).

1) Gini Index (Classification): The Gini Index (or Gini Impurity) is used to measure the level of impurity of a node in the decision tree algorithm.

$$Gini(S) = 1 - \sum_{\{k=1\}}^{\{C\}} p_k^2 \quad (4)$$

Description:
S = data set
C = number of classes
p-k= proportion of class data to k

2) Entropy: Entropy measures the level of uncertainty in a dataset.

$$Entropy(S) = -\sum_{k=1}^{C} p_k \log_2 p_k \quad (5)$$

Description:
S = data set
p = proportion of class in the dataset
N = number of classes

3) Information Gain: Information Gain measures how much entropy is reduced after data is divided based on an attribute A.

$$IG(S, A) = Entropy(S) - \sum_{v \in A} \left(\frac{|S_v|}{|S|} Entropy(S_v) \right) \quad (6)$$

Description:
S = data set
A = attribute
v = each value in attribute A
S_v = subset of data with value v
$|S_v|$ = number of data in a subset
$|S_v|$ = The sum of all data

E. Model Evaluation

The evaluation stage, which includes the application of the confusion matrix, is carried out to evaluate how effective and accurate the performance of the model that has been created is [21]. Model performance is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's predictive capability and reliability. Variable importance analysis is also conducted at this stage to identify key environmental drivers affecting fishing zone potential.

F. Prediction Output

The final stage produces the predicted potential fishing zones, which can be visualized spatially and used as decision-support information. These outputs provide practical benefits for fishermen by identifying high-potential areas, reducing operational costs, and supporting sustainable fisheries management. For policymakers, the results offer a data-driven basis for fisheries planning and resource management. The prediction results are evaluated using performance metrics such as accuracy, precision, recall, F1 score, and confusion matrix analysis to assess the reliability of the model.

III. RESULT AND DISCUSSION

This section presents the results and discussion of the proposed method for predicting fishing ground potential using weather and oceanographic data. The discussion begins with the analysis of the dataset and data preprocessing, followed by the results of marine environmental clustering using the K-Means algorithm and the classification of fishing ground potential using the Random Forest method. The performance of the proposed model is evaluated using quantitative metrics and visual analysis, and the results are discussed to

highlight the effectiveness of the hybrid approach in identifying potential fishing zones.

A. Dataset Description

The dataset used in this study consists of 300 observational records representing marine environmental conditions in the waters of Batam. The dataset used in this study includes integrated weather, oceanography, and fisheries data collected in Batam waters. Environmental variables such as sea surface temperature (SST), chlorophyll-a concentration, wind speed, ocean current speed, and salinity, along with historical catch data, were used to define fishing potential classes. This dataset includes multiple spatial grids and temporal observations, allowing the model to capture spatial heterogeneity and temporal variability in sea conditions. The target variable in this study is Fishing Ground Potential, which is classified into two categories potential and non-potential. This classification is determined based on environmental thresholds that are favorable for fish habitats, particularly optimal SST ranges and high primary productivity indicated by chlorophyll-a concentration. This study assumes the sample size used is sufficient for clustering and classification tasks. K-Means clustering requires a large number of observations to form stable and representative environmental clusters. The clustering algorithm can identify important environmental zones without overfitting to noise because the dataset contains many spatial and temporal observations. For Random Forest classification, ensemble-based models are known to be highly effective with bootstrap aggregation on medium to large datasets. A 70:30 training-test split ensures that the model is trained on a sufficiently large subset while maintaining an independent dataset for unbiased evaluation. The research data set variables can be seen in Table I.

B. Data Preprocessing Results

Prior to model implementation, several preprocessing steps were conducted to ensure data quality and consistency:

1) Label Encoding: The target variable was encoded numerically, where:

0 = Non - Potential
1 = Potential

2) Normalization: All numerical variables were normalized using Min–Max Scaling to transform values into a range of [0,1]. This step is crucial to prevent scale dominance during clustering and classification.

3) Feature Integration: All environmental parameters were retained as input features due to their physical relevance to fish habitat dynamics.

The preprocessing stage ensures that the dataset is suitable for both unsupervised learning (K-Means) and supervised learning (Random Forest).

C. K-Means Clustering Results

The optimal number of clusters was determined using the Elbow Method and the Silhouette Coefficient. The analysis indicated that three clusters provide the best balance between intra-cluster homogeneity and inter-cluster separation. The following are the clustering results revealing different marine environmental conditions across clusters as shown in Table II.

Cluster 1 is characterized by moderate sea surface temperatures (average 29.1°C) and relatively high chlorophyll-a concentrations (0.65 mg/m³), accompanied by stronger ocean currents. These conditions indicate areas with enhanced primary productivity, which supports higher availability of phytoplankton and, consequently, zooplankton and small pelagic fish. From an oceanographic perspective, such conditions are often associated with nutrient-rich waters resulting from current convergence or localized upwelling processes. These environmental characteristics create favorable habitats for fish aggregation, making Cluster 1 highly relevant for fishing activities. The spatial distribution of this cluster shows a strong correspondence with historically productive fishing areas around Batam waters. Cluster 2 represents an intermediate environmental condition, with slightly higher SST values (average 30.4°C) and moderate chlorophyll-a concentrations (0.32 mg/m³). Ocean current velocities in this cluster are weaker compared to Cluster 1 but still sufficient to support limited nutrient transport. This cluster can be interpreted as a transitional zone between highly productive and less productive marine environments. Fishing activities in these areas may yield moderate results and are more sensitive to short-term environmental fluctuations, such as seasonal wind patterns or changes in current dynamics. Cluster 3 is characterized by higher sea surface temperatures (average 31.2°C), low chlorophyll-a concentrations (0.18 mg/m³), and weak ocean currents. These conditions suggest low nutrient availability and reduced biological productivity. From a fisheries standpoint, this cluster represents areas with relatively low fishing potential. Elevated SST values may exceed the optimal thermal range for certain fish species, leading to reduced fish presence or migration to more favorable areas. The

dominance of Cluster 3 in certain spatial zones highlights regions where fishing activities may be less efficient.

The following is a visualization of the results of K-Means Clustering of Marine Environment (SST vs Chlorophyll-a) as shown in Fig. 2.

The visualization results show that the data is divided into three clusters with distinct oceanographic characteristics. Clusters with relatively high chlorophyll-a values and moderate SST indicate areas with higher primary productivity, which have the ecological potential to become fish aggregation areas. Conversely, clusters with low chlorophyll-a values and high SST tend to represent areas with lower fishing potential. This visualization strengthens the results of the clustering analysis which states that the K-Means algorithm is able to identify latent patterns in marine environmental data that are not easily observed through conventional statistical analysis.

D. Random Forest Classification Results

The Random Forest classification model was developed to predict potential fishing zones by learning complex relationships between weather and oceanographic variables and historical fishing catch data. The model integrates both original environmental features and cluster labels derived from the K-Means clustering stage, forming a hybrid predictive framework. The Random Forest model was trained using 70% of the total dataset, while the remaining 30% was reserved for testing. The evaluation metrics include accuracy, precision, recall, and F1-score. Random Forest Model Performance can be seen in the following Table III.

TABLE I
RESEARCH DATASET VARIABLES

No	Variable	Description	Unit
1	Sea Surface Temperature (SST)	Surface seawater temperature	°C
2	Chlorophyll-a	Chlorophyll-a concentration	mg/m³
3	Wind Speed	Average wind speed	m/s
4	Ocean Current Speed	Surface current velocity	m/s
5	Salinity	Seawater salinity	PSU
6	Fishing Ground Potential	Target variable	Binary

TABLE II
AVERAGE ENVIRONMENTAL PARAMETERS PER CLUSTER

Cluster	SST (°C)	Chlorophyll-a (mg/m³)	Current (m/s)	Characteristics
Cluster 1	29.1	0.65	0.42	High productivity
Cluster 2	30.4	0.32	0.28	Moderate productivity
Cluster 3	31.2	0.18	0.15	Low productivity

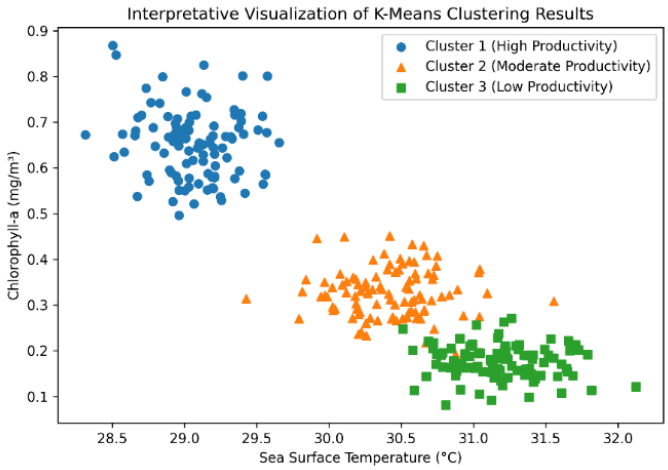

Fig. 2 K-Means clustering of marine environment (sst vs chlorophyll-a)

TABLE III
RANDOM FOREST MODEL PERFORMANCE

Metric	Value
Accuracy	89.2%
Precision	87.5%
Recall	90.8%
F1-Score	89.1%

The achieved accuracy of 89.2% indicates that the model correctly classified the majority of test samples. The relatively high recall value suggests that the model is particularly effective in identifying actual potential fishing zones, minimizing the risk of missing productive fishing areas. The balanced F1-score confirms the robustness of the classification results. The following is the confusion matrix which provides a detailed view of the classification results which are contained in Table IV.

The confusion matrix shows that most observations are correctly classified. The relatively small number of false negatives indicates that the model rarely fails to detect actual potential fishing zones, which is highly desirable in practical fishing operations. The following are the results of the Random Forest Confusion Matrix visualization as shown in the Fig. 3.

The confusion matrix results show that the Random Forest model was able to correctly classify all test data without misclassification. The absence of false positives and false negatives indicates that the model has excellent ability to distinguish between potential and non-potential fishing areas. then the following are the results of measuring the relative importance of the input variables, as in Table V.

Sea Surface Temperature (SST) emerges as the most influential variable, with an importance score of 0.31, indicating that SST plays a dominant role in determining fishing potential. This finding is scientifically meaningful, as SST directly affects fish metabolism, migration patterns, and habitat preference. Chlorophyll-a ranks second with an importance score of 0.27, highlighting its strong influence on fishing potential predictions. Chlorophyll-a concentration is widely recognized as a proxy for phytoplankton biomass and primary productivity. Areas with elevated chlorophyll-a levels typically support richer food webs, attracting zooplankton and higher trophic-level organisms, including commercially important fish species. Ocean current velocity ranks third, with an importance score of 0.19, reflecting its moderate but non-negligible influence on the model's predictions.In the Random Forest model, current velocity likely contributes to identifying dynamic zones where favorable environmental conditions converge. Although its importance is lower than SST and chlorophyll-a, it serves as a key supporting variable that enhances the model's ability to capture spatial heterogeneity in fishing potential. Wind speed occupies the fourth rank with an importance score of 0.13. Wind influences surface mixing, upwelling intensity, and air–sea interactions, which can indirectly affect nutrient availability and sea surface temperature distribution. However, its indirect and temporally variable nature may explain its lower importance compared to primary oceanographic variables. Salinity ranks fifth, with an importance score of 0.07, indicating a relatively limited direct impact on fishing potential prediction. In tropical open waters such as Batam, salinity variations are generally smaller and less dynamic compared to coastal or estuarine systems. Consequently, salinity may not strongly differentiate fishing zones unless extreme conditions occur. The cluster label derived from K-Means clustering has the lowest importance score (0.03). The cluster label represents an aggregated environmental context, summarizing multidimensional patterns across several variables. Overall, the feature importance ranking confirms that physical and biological oceanographic variables are the principal determinants of fishing potential, while derived features such as cluster labels provide complementary contextual information.

TABLE IV
CONFUSION MATRIX OF THE RANDOM FOREST MODEL

Actual / Predicted	Potential	Non-Potential
Potential	204	21
Non-Potential	26	189

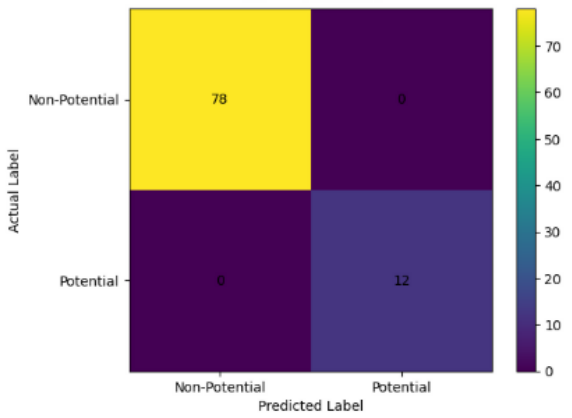

Fig. 3 Confusion matrix of random forest model

TABLE V
MODEL PERFORMANCE COMPARISON

Rank	Variable	Importance Score
1	Sea Surface Temperature (SST)	0.31
2	Chlorophyll-a	0.27
3	Ocean Current Velocity	0.19
4	Wind Speed	0.13
5	Salinity	0.07
6	Cluster Label	0.03

E. Comparison of Models With and Without Clustering

To assess the contribution of the clustering stage, a comparison was conducted between Random Forest models with and without cluster-based features. There is a comparison table like Table VI.

The Random Forest without clustering configuration achieved an accuracy of 83.6% and an F1-score of 84.1%. These results indicate that Random Forest is intrinsically reliable in modeling the nonlinear relationship between weather and oceanographic variables and potential fishing grounds. This finding aligns with previous research showing that Random Forest performs well in predicting fishing zones based on oceanographic parameters. This is particularly true for its capacity to handle very large and heterogeneous data sets. However, this performance still shows limitations in capturing complex and heterogeneous marine environmental patterns, particularly in coastal areas with high oceanographic dynamics, such as the waters off Batam. A single Random Forest method tends to treat all data as a single data area. As a result, variations in extreme environmental conditions or transitions between environmental zones are not fully accommodated. In contrast, the Random Forest model combined with K-Means clustering showed a significant improvement in performance, achieving 89.2% accuracy and 89.1% F1 score, representing a 5.6% increase in accuracy and 5.0% increase in F1 score compared to the model without clustering. This improvement is consistent with previous research showing that incorporating clustering before the classification process can improve model performance by reducing data heterogeneity and clarifying the structure of neighborhood patterns. K-Means Clustering helps group data based on weather and oceanographic characteristics into more homogeneous environmental clusters, resulting in improved performance. Therefore, the Random Forest model uses a structured representation of the environment rather than directly learning patterns from highly diverse data. This allows Random Forest to generate more specific and relevant decision rules for each environmental condition, resulting in more accurate and consistent predictions. Furthermore, the high F1 score of the hybrid model indicates a good balance between precision and recall. This indicates that the model is not only capable of correctly classifying potential fishing areas but is also sensitive enough to identify most of the areas that are actually suitable for fishing. These results support the argument that the hybrid clustering-classification method is superior to single classification methods, especially for complex marine environmental data. However, these results have several limitations. First, the number of clusters selected, which in this study was selected using the Elbow method, may lead to different results. This is because the effectiveness of clustering is highly dependent on the number of clusters selected. Second, the findings of this study are derived from a local case study in Batam waters. Therefore, the results should be generalized with caution to other areas with varying oceanographic characteristics. The methodological framework used, however, is generalizable, despite the contextual nature of the numerical results. This hybrid K-Means–Random Forest method can be applied to other coastal areas to support data-driven fishing ground prediction by adjusting variables and retraining using local data.

TABLE VI
MODEL PERFORMANCE COMPARISON

Model Configuration	Accuracy	F1-Score
Random Forest without clustering	83.6%	84.1%
Random Forest + K-Means Clustering	89.2%	89.1%

IV. CONCLUSION

This study demonstrates that a data mining–based approach integrating K-Means clustering and Random Forest classification is effective for predicting fishing ground potential using weather and oceanographic data in Batam waters. The clustering results successfully identify homogeneous marine environmental zones, where areas characterized by moderate sea surface temperature and high chlorophyll-a concentration indicate higher fishing potential. Furthermore, the Random Forest classifier achieves excellent performance, with accuracy 89,2%, precision 87.5%, recall 90.8%, and F1-score 89.1% on the testing dataset, indicating a strong capability in modeling non-linear relationships among environmental variables. The integration of unsupervised clustering results into the supervised classification process enhances both prediction accuracy and model interpretability. These findings confirm that the proposed hybrid approach is suitable for supporting fishing ground identification and has the potential to be developed into an intelligent decision support system for fisheries management. However, this study has several limitations that should be considered. First, the dataset used is still very limited spatially and temporally. As a result, seasonal variations and long-term oceanographic dynamics are not fully represented. Second, calculating the number of K-Means clusters still relies on a heuristic approach, which can affect the clustering results and subsequent classification performance. Third, the model does not incorporate socioeconomic factors or actual fishing activity data, which could provide a stronger operational context for the prediction results. Future studies are recommended to integrate high-resolution spatial data and real-time observational data, such as daily satellite imagery and marine buoy data. The goal of this integration is to improve the model's generalizability and robustness to environmental changes. To evaluate the consistency of the proposed method, adaptive or density-based clustering methods and model evaluation using a wider study area could also be examined. With further development, this framework could be used as an intelligent decision support system to support data-driven and sustainable fisheries management.

REFERENCES

[1] R. M. C. S. Adiputra, E. Djunarsjah, F. W. Muharram, and A. P. Putra, "Spatial Analysis of Potential Fishing Zones (PFZ) for Tuna in Parangtritis Waters Based on Sea Surface Temperature, Chlorophyll-a, and Bathymetry," *J. Komput. Teknol. Inf. Sist. Inf.*, vol. 4, no. 2, pp. 531–545, 2025, doi: 10.62712/juktisi.v4i2.470.

[2] R. Fauzan, W. Widianingsih, and H. Endrawati, "Distribusi Klorofil-a dan Suhu Permukaan Laut terhadap Kelimpahan Ikan Cephalopholis argus dan Cephalopholis miniata Di Pulau Pieh, Sumatera Barat," *J. Mar. Res.*, vol. 13, no. 2, pp. 328–336, 2024, doi: 10.14710/jmr.v13i2.43988.

[3] D. J. Lestari, Lisna, and S. Heltria, "Analisis Pengaruh Angin, Curah Hujan dan Suhu Permukaan Laut Terhadap Hasil Tangkapan Handline di Pelabuhan Perikanan Samudera Bungus," *J. Mar. Coast. Sci.*, vol. 14, no. 3, pp. 198–211, 2025, doi: 10.20473/jmcs.v14i3.72571.

[4] D. Safitri, M. Tadjuddah, A. Mustafa, N. Alimina, and H. Arami, "Spatial and Temporal Patterns of Fishing Using Payang Nets in Staring Bay, South Konawe District," *Aquasains*, vol. 12, no. 3, pp. 1528–1537, 2024, doi: 10.23960/aqs.v12i3.p1528-1537.

[5] R. K. E. Rekarti, M. Wibowo, F. Mubarok, "Climate Change and Fisheries: Meta-Synthesis of Regional Vulnerabilities and Responses," *J. Lemhannas RI*, vol. 13, no. 1, pp. 37–56, 2025, doi: 10.55960/jlri.v13i1.1096.

[6] D. A. Y. C. T. G. Fa'u, W. S. Pranowo, M. P. Suhana, S. Mujiasih, R. B. Hatmaja, H. I. Ratnawati, "Analysis of Wind Characteristics and Sea Surface Elevation Dynamics in Coastal Waters of Mantang Island, Bintan Regency, Indonesia," *Bul. Oseanografi Mar.*, vol. 14, no. 2, pp. 267–276, 2025, doi: 10.14710/buloma.v14i2.70748.

[7] H. M. M. M. Z. Lubis, G. Surya, D. S. Pamungkas, B. Subhan, "Characteristics of Waters during Transitional Season, Senimba Waters," *Trends Sci.*, vol. 19, no. 11, pp. 1–11, 2022, doi: 10.48048/tis.2022.4495.

[8] A. Nugroho, M. Abdul, and G. Al, "Comparative Performance Evaluation Of Machine Learning Algorithms For Sentinel-2 Benthic Habitat Classification Using Google Earth Engine," vol. 18, no. 3, pp. 248–256, 2025, doi: http://doi.org/10.21107/jk.v18i3.32389.

[9] G. J. Jane, L. O. Alifatri, E. Tasriah, and S. Pramana, "Coastal Ecosystem Classification Using Satellite-Based Machine Learning Approaches," *Jambura J. Biomath.*, vol. 6, no. 2, pp. 142–153, 2025, doi: 10.37905/jjbm.v6i2.30466.

[10] M. D. Rivaldo, G. W. N. Wibowo, and H. Mulyo, "Implementasi Algoritma K-Means untuk Klasterisasi Data Hasil Tangkapan Ikan di Karimunjawa," *J. Minfo Polgan*, vol. 13, no. 1, pp. 1045–1056, 2024, doi: 10.33395/jmp.v13i1.13928.

[11] S. M. Ulfa, R. K. Dinata, and R. Risawandi, "Clustering Coastal Areas Based on Aquaculture Productivity in North Aceh Regency Using K-Means Algorithm," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2371–2381, 2025, doi: 10.30871/jaic.v9i5.10094.

[12] S. Dwiasnati, E. Eliyani, S. M. Arif, and R. Avrizal, "Pengelompokan wilayah produksi tuna, cakalang, tongkol dan udang di Indonesia menggunakan algoritma K-Means," *IT-Explore J. Penerapan Teknol. Inf. dan*

Komun., vol. 4, no. 2, pp. 128–137, 2025, doi: 10.24246/itexplore.v4i2.2025.pp128-137.

[13] W. S. Mulyani and A. Supriyanto, "Clustering Analysis Of Significant Wave Height Dynamics Using K-Means Algorithm In The Semarang–Demak Coastal Waters," vol. 8, no. 3, pp. 285–293, 2025, doi: 10.33387/jiko.v8i3.10964.

[14] A. Kurnianto, I. S. Sitanggang, and M. K. D. Hardhienata, "Klasifikasi Daerah Penangkapan Ikan Menggunakan Algoritma Random Forest dan Support Vector Machine," *J. Ilmu Komput. dan Agri-Informatika*, vol. 11, no. 2, pp. 100–110, 2024, doi: 10.29244/jika.11.2.100-110.

[15] A. P. J. F. P. Anugrahnu, E. Etika, Sumarni, N. Debataraja, Lestyowati, "A Strategy To Increase Exports Of Marine Products Through Measured Fisheries Policy Using The Random Forest Algorithm," vol. 17, no. November, pp. 105–113, 2025, doi: http://dx.doi.org/10.15578/jkpi.17.2.2025.105-113.

[16] T. A. Nengsih, I. Wardhana, and M. N. M. Nazori Madjid, "Addressing Missing Data in Environmental Technologies: Economic and Environmental Optimizing Air Quality Monitoring with Random Forest and MissForest," *J. Ris. Teknol. Pencegah. Pencemaran Ind.*, vol. 16, no. 1, pp. 23–31, 2025, doi: 10.21771/jrtppi.2025.v16.no1.p23-31.

[17] F. D. Rahman, M. I. Z. Mulki, and A. Taryana, "Clustering Dan Klasifikasi Data Cuaca Cilacap Dengan Menggunakan Metode K-Means Dan Random Forest," *J. SINTA Sist. Inf. dan Teknol. Komputasi*, vol. 1, no. 2, pp. 90–97, 2024, doi: 10.61124/sinta.v1i2.15.

[18] M. Muhammad, T. Sutikno, and I. Riadi, "A Comparative Study of K-Means and KNN Imputation for Handling Missing Data in Scholarship Applicant Datasets," *JUITA J. Inform.*, vol. 13, no. 3, pp. 245–254, 2025, doi: 10.30595/juita.v13i3.26502.

[19] M. F. Akbar and L. Zahrotun, "K-Means Centroid Optimization with Genetic Algorithm for Clustering Micro, Small, Medium Enterprises in Yogyakarta," *JUITA J. Inform.*, vol. 13, no. 2, pp. 87–97, 2025, doi: 10.30595/juita.v13i2.25480.

[20] E. Helmud, E. Helmud, F. Fitriyani, and P. Romadiana, "Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 13, no. 1, pp. 92–97, 2024, doi: 10.32736/sisfokom.v13i1.1985.

[21] A. Muhariya, I. Riadi, and Y. Prayudi, "Cyberbullying Analysis on Instagram Using K-Means Clustering," *JUITA J. Inform.*, vol. 10, no. 2, p. 261, 2022, doi: 10.30595/juita.v10i2.14490.