Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

Classification Of Sea Wave Heights On The North Coast Of Central Java Using Random Forest

Aji Supriyanto*[1], Dwi Agus Diartonor[2], Budi Hartono[3], Arief Jananto[4], Afandi[5]

[1,5]Information Technology, Universitas Stikubank, Indonesia
[2,4]Information System, Universitas Stikubank, Indonesia
[3]Informatic Engeneering, Universitas Stikubank, Indonesia

Email: [1]ajisup@edu.unisbank.ac.id

Received : Jul 15, 2025; Revised : Jul 22, 2025; Accepted : Jul 31, 2025; Published : Aug 19, 2025

Abstract

Global climate change has triggered an increase in the occurrence of significant wave heights (SWH) and sea level rise (SLR) in coastal areas, including the northern coast of Central Java, Indonesia (Pantura). These phenomena directly impact maritime activities, coastal erosion, and tidal flooding. This study aims to classify and predict significant wave height (SWH) and sea level rise (SLR) trends using a machine learning approach based on the Random Forest (RF) algorithm. Daily meteorological and oceanographic observation data from 2019 to 2024, provided by BMKG, serve as the main dataset. The dataset includes wind speed, ocean current velocity, air pressure, and wave direction. SWH is categorized into three classes: Calm, Low, and Moderate. The classification model achieved excellent performance with an accuracy of 98.54%, a macro F1-score of 0.942, and maintained strong accuracy even for the minority class (Moderate) despite data imbalance. The RF Regressor for SWH prediction yielded an R^2 of 0.864, MAE of 0.067, and RMSE of 0.109 m. Visualizations such as scatter plots, boxplots, and heatmaps supported the conclusion that ocean current speed and wave period are key factors influencing SWH. The study concludes that Random Forest is effective for classifying and predicting sea conditions in tropical regions like Pantura, and it is feasible for implementation in data-driven early warning systems to mitigate coastal risks. This contributes to marine safety and coastal risk mitigation planning.

Keywords : *Central Java, Classification, Northern Costal, Random Forest, Sea Wave Heights (SWH).*

1. INTRODUCTION

In recent decades, global climate change has triggered extreme phenomena in coastal environments, including increased significant wave height (SWH) and sea level rise (SLR), particularly in Indonesian waters. These two phenomena have broad impacts on the sustainability of coastal areas, such as the disruption of shipping activities, damage to port infrastructure, and an increased risk of tidal flooding and coastal erosion. According to the Central Java Regional Disaster Management Agency (BPBD), tidal flooding has affected several regions along the northern coast of Central Java (Pantura), including the cities and regencies of Pekalongan, Semarang, Tegal, Kendal, Brebes, Demak, Batang, and Pemalang, with average flood heights ranging between 30–70 cm. Therefore, a vulnerability assessment in this region is urgently needed.

The main factor behind flooding in the northern coast of Java is rising sea levels, which are driven by factors such as lunar and solar gravitational forces and high ocean waves, particularly during severe weather events[1]. Low-lying island nations like Indonesia are especially vulnerable to Height Extremes (HEX) of sea level. When intensified by marine heatwaves, these HEX events can have even more severe ecological and social impacts[2]. In Indonesia, variations in SWH are influenced not only by local winds but also by regional factors such as air pressure, sea surface temperature, and ocean current

circulation [3]. Wave patterns along the northern coast of Java are heavily influenced by the monsoonal wind seasons, which also affect tidal patterns and can trigger tidal flooding[4].

Long-term sea level rise leads to an increase in ocean volume, thereby intensifying both the frequency and magnitude of flooding that can inundate low-lying land[5]. The correlation coefficient between wind speed and SWH during tropical cyclone periods in the Java Sea shows that wave heights rose from 0.5 meters to an average of 2 meters, with some local waves reaching 3.2 meters [6]. A more recent study found that SWH in the Java Sea ranged from 0.2 to 2.69 meters in 2017 [7], and in Semarang's coastal waters, from 0.026 to 1.255 meters, with wave periods ranging between 1.677 to 5.781 seconds [8].

Coastal land can be rapidly lost due to relative sea level rise (RSLR) caused by local land subsidence [9]. This phenomenon of sea level rise and land subsidence has been observed on the northern coast of Central Java, especially in Semarang, Pekalongan, and Demak [10][11][12]. as well as in Tegal [13] and parts of Pemalang [10].In Pekalongan, the rate of sea level rise is relatively high, reaching 10.6 mm/year, and is also driven by subsidence [13]. In Semarang City, tidal flooding (locally known as *pasut*) is caused by a combination of sea level rise and land subsidence. Meanwhile, Demak Regency, with its flat coastal topography and low elevation, is highly vulnerable to the impacts of sea level rise [14].

Prediction of significant wave height (SWH) is crucial for maritime safety, offshore operations, and coastal disaster mitigation [15]. Variations in SWH are influenced by wind speed, air pressure, and atmospheric circulation [15][16][17]. Extremely low sea-level pressure followed by increased wind speed has a strong impact on SWH [18]. Both local and global atmospheric circulation significantly control the variability of surface wind speeds, which in turn shape wind and wave patterns [19]. Thus, accurate SWH prediction is vital for ensuring safe navigation, effective ocean operations, and coastal disaster preparedness [15][20][21][22]. Monitoring and classification of SWH conditions play a key role in risk mitigation for ports and early warning services [18]. Classification or prediction of SWH has become central to marine operations, coastal engineering, and readiness for extreme wave events [23].

The use of the Random Forest (RF) model for SWH prediction on the Australian coast highlights the importance of SWH in coastal zone management and disaster mitigation [15]. The application of Random Forest Regressor for sea level rise (SLR) prediction has yielded R^2 values above 0.85 and MAPE below 5% [24]. Prediction of SWH using RF also demonstrates high accuracy, with reported $R^2 = 0.958$ and MSE = 0.0074 [15]. RF models have achieved wave prediction accuracy levels of 89.6–91.8% [22]. Although primarily used for regression, the model's high performance in short-term prediction reflects its potential to indirectly handle wave condition classification [18]. Moreover, the integration of RF and XGBoost within ensemble classification algorithms for SWH in coastal zones has proven effective, even in the context of Indonesian wave dynamics [25][26].

Wave classification models based on categorical levels (low, moderate, high) are particularly valuable for early warning systems [20]. However, such models require high-quality input data and classification methods capable of identifying minority classes such as moderate or high waves, which occur infrequently but pose high risks. These findings support the use of machine learning (ML) as a data-driven solution capable of capturing non-linear relationships between variables without relying on physical assumptions. ML has proven effective in various oceanographic applications, including wave height prediction [27], wave height classification, and sea level trend forecasting [25]. These findings suggest that RF is well-suited for modeling dynamic ocean conditions in tropical regions such as Indonesia.

Nonetheless, significant challenges remain, particularly regarding class imbalance in SWH classification. In the case of Pantura, calm wave conditions dominate the daily observations, while moderate and high waves appear only occasionally. This leads to a bias in the model toward the majority

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

class. Therefore, techniques such as class balancing, careful feature selection, and the use of appropriate evaluation metrics (e.g., F1-score, confusion matrix) are crucial in building a fair and reliable classification model.

The contribution of the field of informatics lies in the application of the Random Forest algorithm for the classification of Significant Wave Height (SWH), which integrates machine learning techniques in modeling multivariate and non-linear data for an oceanographic data-based monitoring and early warning system in coastal areas.

The study is expected to serve as a scientific foundation for decision support systems developed by institutions such as the Meteorology, Climatology, and Geophysics Agency (BMKG), the Department of Marine Affairs, and port authorities. It also aims to contribute to the applied information technology literature in the field of tropical oceanography.

2. METHOD

Based on the research background, this study adopts a machine learning (ML) approach using the Random Forest (RF) algorithm to obtain optimal performance in the classification and prediction of significant wave height (SWH) along the northern coast of Central Java. The research workflow is illustrated in Figure 2.

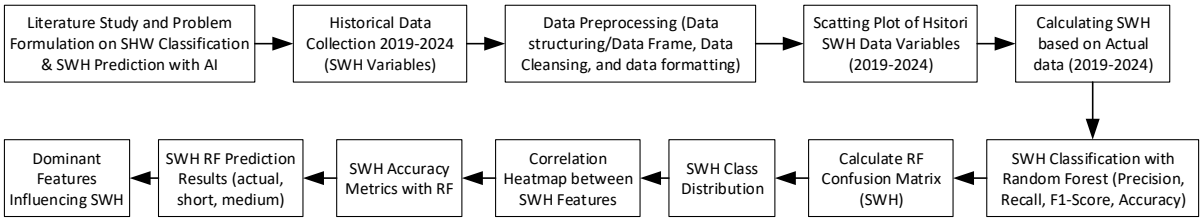

Figure 1. Research Workflow for SWH Classification and Prediction

2.1. Literature Review, Data, and Study Area

This study focuses on the development of ocean waves in the coastal region of the northern part of Central Java (Pantura), particularly covering the areas of Semarang, Demak, Pekalongan, and their surroundings. The literature review focused on SWH and SLR, especially within the northern coastal region of Java, with emphasis on classification and prediction using machine learning techniques.

The data used consist of historical daily time-series records from 2019 to 2024, encompassing meteorological and oceanographic parameters such as average and maximum wind speed, wind direction, current velocity and direction, air pressure, significant wave height (SWH), wave period, and temporal data (month and year). The data were collected from observation stations and instruments operated by BMKG Maritime Station Class II in Central Java, located at Tanjung Mas Port, Semarang. The variables involved in determining SWH include air pressure (mb), ocean current direction and speed (cm/s), significant wave height (Hs), wave period (Ts/t01), wave direction (Dir), wind speed (Kt), and local wind direction (Dir). This primary dataset is highly valuable as it reflects actual conditions at the observation site.

2.2. Data Preprocessing

In the preprocessing stage, the SWH data were structured into a dataframe format. This is crucial to define the attributes of each feature before further data processing. The data were stored in .xlsx or .csv format, and attributes serving as features or labels were added to each record. Normalization was conducted to verify whether the data were complete (non-missing) and whether the values were

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

reasonable. After normalization, it was found that out of 2,192 records, only one record contained missing data. Therefore, 2,191 valid records were used for prediction analysis.

The cleaned dataset structure, including features influencing SWH, is illustrated in Figure 1. Initial preprocessing involved checking for missing values, outliers, and formatting inconsistencies. The cleaning process included deleting or interpolating missing values, transforming time columns to datetime format, normalizing numerical values using Min-Max Scaling, and adding derived features such as seasonal labels, daily lag features, and a 30-day moving average.

```
Index: 2191 entries, 0 to 2191
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Tanggal         2191 non-null   object
 1   Kec_rata-rata   2191 non-null   float64
 2   Kec_Max         2191 non-null   int64
 3   Arah            2191 non-null   int64
 4   Kec_Arus        2191 non-null   int64
 5   Arah_arus_laut  2191 non-null   int64
 6   Tekanan         2191 non-null   float64
 7   hs              2191 non-null   float64
 8   t01             2191 non-null   float64
 9   dir             2191 non-null   float64
 10  Rata-rata       2191 non-null   float64
 11  Max             2191 non-null   int64
 12  Min             2191 non-null   int64
dtypes: float64(6), int64(6), object(1)
memory usage: 239.6+ KB
```

	Tanggal	Kec_rata-rata	Kec_Max	Arah	Kec_Arus	Arah_arus_laut	Tekanan	hs	t01	dir	Rata-rata	Max	Min
0	01/01/2019	4.0	6	340	5	90	1009.8	1.1	5.1	315.0	85.5	107	60
1	02/01/2019	3.0	9	140	4	247	1009.8	0.8	4.9	315.0	82.8	104	45
2	03/01/2019	4.0	9	10	4	215	1010.9	0.6	5.1	316.0	81.6	105	50
3	04/01/2019	5.0	13	260	7	213	1011.3	0.3	4.9	316.0	88.3	115	55
4	05/01/2019	4.0	12	310	5	210	1011.4	0.1	4.5	313.0	86.6	110	55
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2187	27/12/2024	2.0	6	110	7	82	1009.8	0.3	3.8	322.0	149.5	176	125
2188	28/12/2024	3.0	8	200	11	78	1009.2	0.4	3.7	316.0	147.5	168	122
2189	29/12/2024	3.0	7	300	10	87	1009.2	0.4	3.9	324.0	147.3	178	112
2190	30/12/2024	2.0	12	340	13	93	1008.7	0.8	4.1	323.0	154.9	182	138
2191	31/12/2024	4.0	130	300	13	99	1008.4	0.9	4.9	319.0	146.0	168	130

[2191 rows x 13 columns]

Figure 2. SWH Data Preprocessing: a. DataFrame (left), b. Feature Structure Based on SWH

2.3. Scatter Plot and SWH Classification Labeling

Scatter plots and time-series curves were generated to visualize the temporal behavior of various oceanographic and meteorological parameters from 2019 to 2025. These visualizations aimed to show the distribution of raw data (scatter points), highlight short-term trends using a 30-day moving average (MA), and depict medium- and long-term trends using linear regression. According to BMKG's classification of sea wave height: Calm: 0.1–0.5 m, Low: 0.5–1.25 m, Moderate: 1.25–2.5 m, High: 2.5–4.0 m, Very High: >4.0 m. The SWH classification in this study was calculated using daily, monthly, and annual averages derived from the actual time series data (2019–2024). Regression analysis with high R^2 values supports the suitability of linear models for projecting short- to medium-term trends in wave height.

2.4. SWH Classification Model Using Random Forest

The classification model was developed using the Random Forest Classifier algorithm, with the goal of categorizing wave conditions into three classes: Calm, Low, and Moderate. The dataset containing 2,191 records was split into 80% training data (1,752 records) and 20% test data (439 records). Model performance was evaluated using the confusion matrix, accuracy, precision, recall, F1-score, 5-fold cross-validation, and feature importance analysis for model interpretation.

a. The confusion matrix illustrates the number of model predictions compared to actual class labels. In this case, the classification includes three categories: Calm, Low, and Moderate.
b. Accuracy measures the proportion of correct predictions over the total number of predictions, as seen in (1):

$$\text{Accuracy} = \frac{Total\ amount\ of\ data}{Number\ of\ Correct\ Predictions} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

For multiclass classification as seen in (2):

$$Accuracy = \frac{\sum_{i=1}^{k} TPi}{Sample\ Total} \quad (2)$$

c. Precision evaluates the accuracy of positive predictions for each SWH class as seen in (3):

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

$$Precission = \frac{TP}{TP+FP} \quad (3)$$

d. Recall (sensitivity) assesses the model's ability to correctly detect actual positive instances as seen in (4):

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

e. F1-Score is the harmonic mean of precision and recall. It is especially important in imbalanced datasets, such as when the “Calm” class dominates and the “Moderate” class has relatively few instances as seen in (5)::

$$F1 = 2\,X\,\frac{Precision\ x\ Recall}{Precision\ x\ Recall} \quad (5)$$

f. 5-Fold Cross-Validation (5-Fold) was used to evaluate the model's generalizability and to prevent overfitting.

The distribution of SWH classes was visualized using pie charts, while boxplots illustrated the statistical distribution of wave height (hs) for each year (2019–2024). Additionally, heatmaps were generated to analyze feature correlation, which is crucial for identifying important variables for classification or prediction tasks.

3. RESULT

3.1 Scatter Plot of Significant Wave Height (SWH)

Scatter plots were used to show the distribution of original data points, depict short-term trends (via 30-day moving averages), and illustrate medium- and long-term trends (via linear regression) of significant wave height (SWH). The 30-day moving average (MA) line indicates that each point on the line represents the average value of the previous 30 days. This helps visualize short-term trends (daily and monthly), observe seasonal up-and-down patterns, and avoid misinterpretations caused by short-term spikes. The scatter plots are shown in Figures 3, 4, 5, and 6.

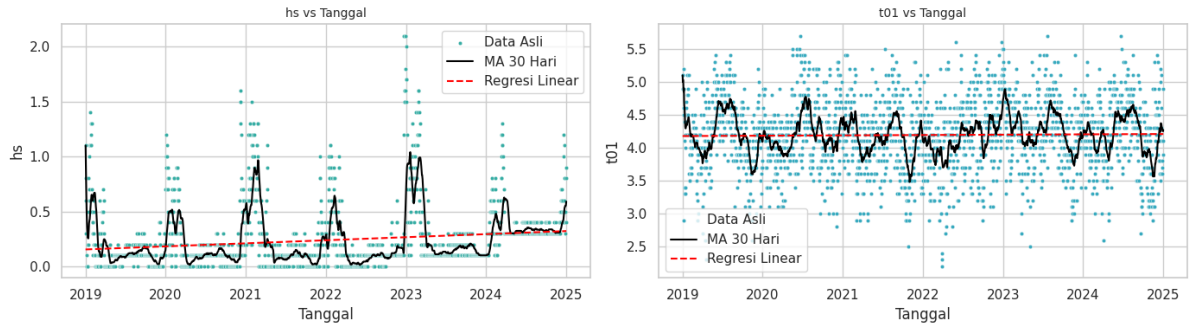

Figure 3. Scatter Plot of Significant Wave Height (Hs) and Wave Period (Ts/t01)

a. *Figure 3 (left):* Hs vs. Date — Shows seasonal wave spikes typically occurring at mid- or year-end (likely due to the west monsoon). The long-term regression trend shows a slight upward trend in Hs, which may indicate increased frequency of high waves. *Figure 3 (right):* Wave Period (t01) vs. Date — Despite seasonal fluctuations, the wave period remains relatively stable annually. (Blue dots = daily t01 spread; average period between 3.5–4.8 seconds; flat regression line indicates no significant trend.)

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

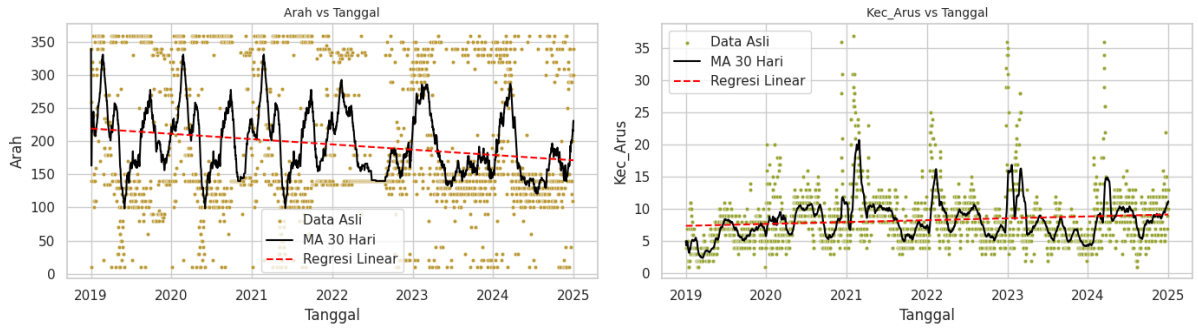

Figure 4. Scatter Plot of Wind Direction (Dir) and Current Velocity (cm/s)

b. *Figure 4 (left):* Wave Direction (Dir) vs. Date — Shows a long-term shift in dominant wind direction (from west to south/southeast), with clear seasonal cycles each year. *Figure 4 (right):* Current Velocity (cm/s) vs. Date — Indicates that ocean currents are strengthening, which may affect sediment transport and wave behavior. (Green dots = daily values; visible surges in 2021 and 2023; trend line shows long-term increase in current speed).

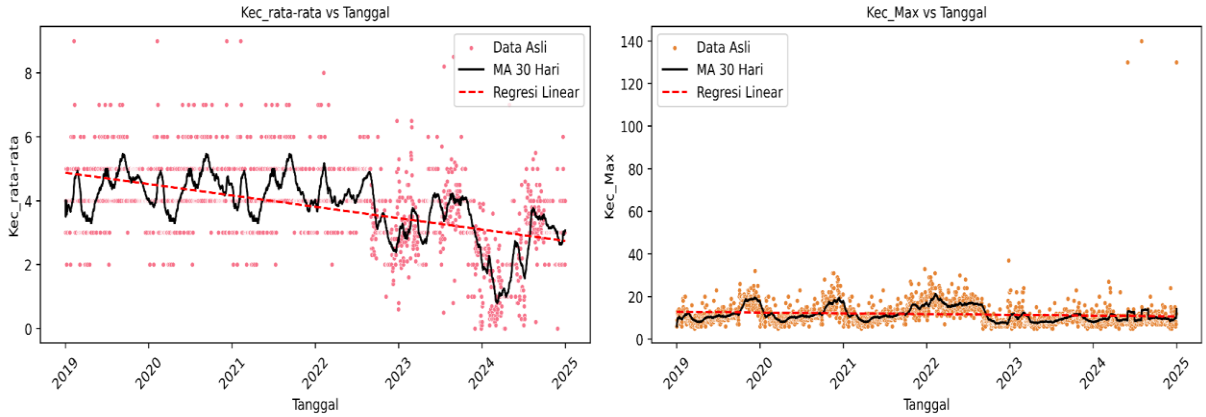

Figure 5. Scatter Plot of Wind Speed Patterns (2019–2025)

c. *Figure 5 (left):* Average Wind Speed (Kt) vs. Date — Shows a sharp drop in average wind speed, especially after 2022, which may reflect climate change effects or sensor instability. *Figure 5 (right):* Maximum Wind Speed (Kec_max) vs. Date — Regression line is relatively flat, indicating no long-term increase in extreme winds. However, some significant outliers (>100 m/s) suggest potential sensor errors.

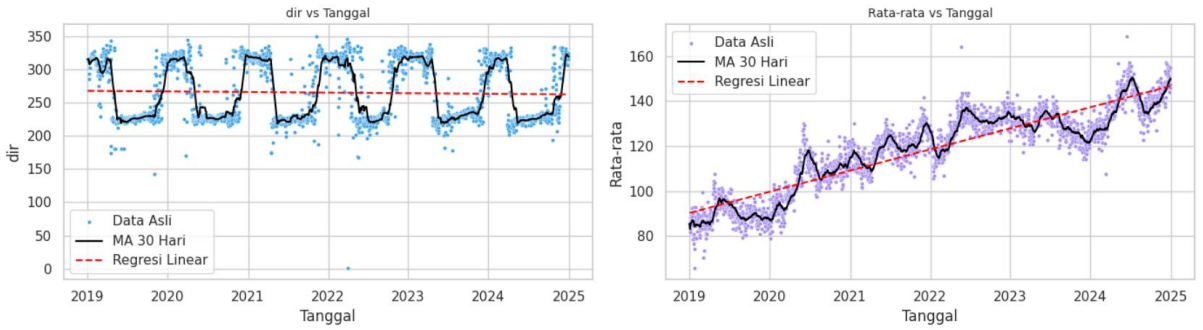

Figure 6. Scatter Plot of Wave Direction Related to Wind Direction

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

d. *Figure 6 (left):* Wave Direction (Dir) vs. Date — Reveals consistent seasonal patterns, with wave direction fluctuations repeating annually, likely linked to monsoonal wind cycles. *Figure 6 (right):* "Average" vs. Date — This metric shows a clear increasing trend from 2019 to 2024. Both trend lines and moving averages suggest continuous seasonal and long-term increases. The “average” variable likely represents daily means of parameters such as Hs, current, or wind speed.

3.2. Calculating the Actual SWH Trend

Calculating actual values of significant wave height (SWH) aims to understand the average real-world wave behavior during the 2019–2024 period. From the resulting trends, the coefficient of determination (R^2) is calculated. The higher the R^2 value, the more accurate and reliable the regression model is for predictive and coastal risk mitigation purposes. R^2-score serves as an indicator of how well the model explains variations in actual data.

Based on the dataset (Figure 2), 2,191 records of actual SWH data involving 13 variables were used. The average increase in SWH is as follows; Daily: 0.033 m/day (indicating ±3 cm variability per day), Monthly**:** 0.058 m/month (±5.8 cm), Annually**:** 0.036 m/year (i.e., 3.6 cm/year increase). The calculated R^2-score is 0.793**,** meaning 79.3% of the variability in SWH can be explained by the linear trend model. The monthly and yearly SWH trends are shown in Figures 7 and 8.

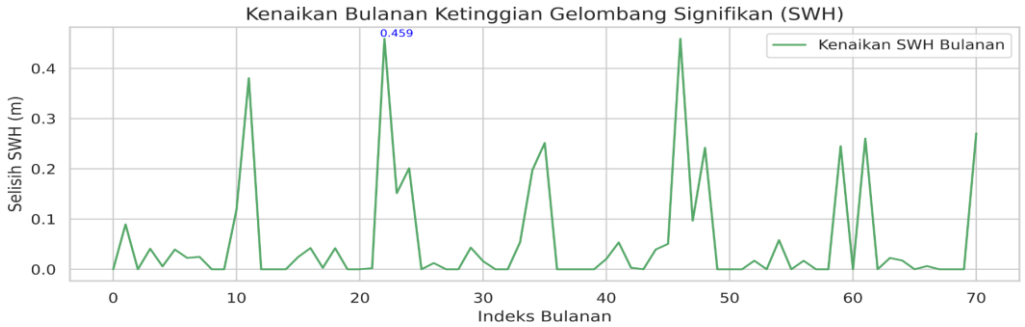

Figure 7**.** Monthly SWH Trend (2019–2024)

Demonstrates a rising trend in wave height over the years. The increase is statistically significant ($R^2 = 0.793$). The blue line shows annual mean SWH (in meters), while the red dashed line is the linear regression line, indicating an increase from 0.16 m in 2019 to 0.37 m in 2024, or a linear trend of +0.036 m/year.

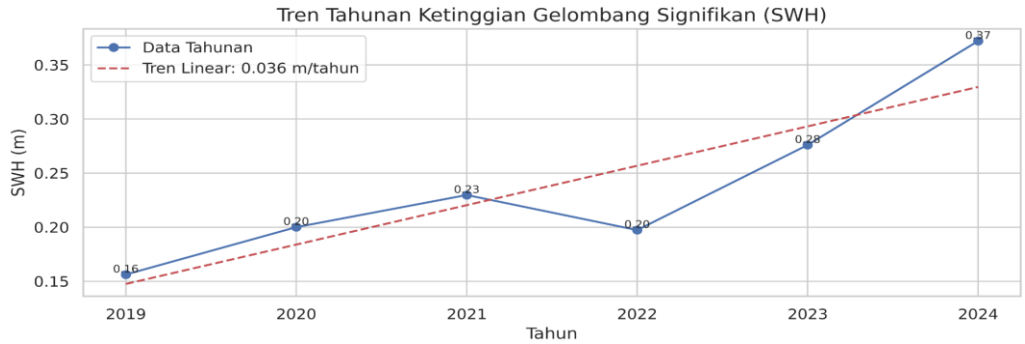

Figure 8**.** Annual SWH Trend (2019–2024)

Illustrates monthly wave height differences (in meters), with noticeable peaks in certain months up to +0.459 m (21st month). These extreme months likely correspond to west monsoon seasons, La

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

Niña/El Niño events, or tropical storms.Monthly variability is important for predictive models and early warning systems.Meanwhile, daily SWH trends (smoothed with a 7-day moving average) show sharp increases, with the highest point (+0.286 m/day) occurring in early 2023. These spikes may indicate extreme events like storms, sudden high winds, or strong currents—requiring special attention in maritime operations, aquaculture, and daily coastal activities.

3.3. Results of SWH Classification Using Random Forest

The classification results of Significant Wave Height (SWH) using the Random Forest (RF) algorithm are based on the dataset illustrated in Figure 2. This dataset contains 2,191 rows of daily data representing oceanographic and meteorological conditions in the northern coastal area of Central Java during the 2019–2025 period. The dataset was used to build and evaluate the SWH classification model using ML with RF. The total number of data rows before and after cleaning (NA handling) remained 2,191. The data was split into 80% training (1,752 rows) and 20% testing (439 rows). This dataset served two primary purposes: to classify SWH based on meteorological and oceanographic parameters targeting SWH and to perform predictive analysis and trend visualization for hs, t01, current, pressure, and wind direction. The next step is to define the dataset structure, as shown in Table 1.

Table 1. Structure of the SWH Dataset

Colom Name	Description
Tanggal	Date of observation (daily datetime format)
Kec_rata-rata	Daily average wind speed (m/s)
Kec_Max	Daily maximum wind speed (m/s)
Arah	Wind direction (0–360 degrees)
Kec_Arus	Surface ocean current speed (cm/s)
Tekanan	Surface air pressure (hPa)
hs	Significant wave height (m) – regression target
t01	Average wave period (seconds)
dir	Wave direction (0–360 degrees)
Rata-rata	Combined average value (e.g., speed or energy)
Max	Maximum combined daily parameter value
Min	Minimum combined daily parameter value
Tahun	Year of observation
Bulan	Month of observation
SWH	The sea wave class categories: ('Calm', 'Low', 'Moderate', 'High', 'Very High', 'Extreme')

The results of the classification modeling with Random Forest (RF) are presented in Table 2.

Table 2. SWH Classification Prediction in the Northern Coast of Central Java

Class	Precision	Recall	F1-Score	Support
Low	0.9495	0.8952	0.9216	210
Moderate	0.9688	0.8611	0.9118	36
Calm	0.9893	0.9974	0.9933	1945
Accuracy	-	-	0.9854	2191
Macro Average	0.9692	0.9179	0.9422	2191
Weighted Avg	0.9851	0.9854	0.9851	2191

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

Table 2 presents the performance of the classification model for significant wave height (SWH) into three categories: Low, Moderate, and Calm, based on observational data from the northern coastal region of Central Java. The evaluation uses common classification metrics, namely precision, recall, F1-score, and support (number of actual data points per class). The explanation for each class category is as follows:

a. "Calm" Class: This class has the highest support value with 1,945 data points (88.8% of the total), a precision of 0.9893 indicating highly accurate predictions for "Calm" conditions. The recall is 0.9974, meaning nearly all "Calm" data were correctly identified. The F1-score of 0.9933 shows an excellent balance between precision and recall. This suggests that the model performs exceptionally well in identifying calm sea conditions, which indeed represent the majority class.
b. "Low" Class: With 210 data points, the precision of 0.9495 indicates that about 95% of "Low" predictions are accurate. The recall of 0.8952 means nearly 90% of actual "Low" data were detected, and the F1-score of 0.9216 reflects a well-balanced performance. This implies that the classification of low wave conditions is highly effective. Although recall is slightly lower than precision, the model still correctly identifies the majority of data.
c. "Moderate" Class: This minority class has only 36 data points. With a precision of 0.9688, the predictions for "Moderate" are highly accurate when made. The recall is 0.8611, meaning about 86% of actual "Moderate" data were identified. The F1-score of 0.9118 is quite good for a minority class. This indicates that even with limited data, the model can provide high-precision predictions for the Moderate class. Although recall is slightly lower due to the small sample size, an F1-score above 0.91 demonstrates the model's robustness to class imbalance.
d. The classes "High," "Very High," and "Extreme." All results have a value of 0 (no occurrences).

Based on the performance of the "Calm", "Low", and "Moderate" classes as described in points a, b, and c, the overall accuracy is determined to be 0.9854 (98.54%), which indicates very high performance. The macro average (i.e., the unweighted mean across classes) values are: Precision = 0.9692, Recall = 0.9179, and F1-score = 0.9422. This indicates consistent performance across classes, though the model is slightly more accurate in precision than recall. The weighted average metrics, all exceeding 0.985, suggest that the model is stable and highly effective, particularly due to the dominance of the "Calm" class in the dataset.

Based on these results, the SWH classification model for the northern coastal region of Central Java demonstrates very high and balanced performance, even for the minority "Moderate" class. Evaluation metrics such as high F1-score and precision show that the model generalizes well and is suitable for integration into machine learning-based sea wave monitoring and prediction systems using Random Forest.

Following the results in Table 2, a confusion matrix was generated to provide a detailed evaluation of the model's performance, not only overall (such as accuracy) but also for each predicted sea wave class. The confusion matrix for SWH is illustrated in Figure 9. The matrix in Figure 9 illustrates the performance of the Random Forest (RF) model in classifying Significant Wave Height (SWH) into three categories: Calm, Low, Moderate, High, Very High, and Extreme.

The analysis of the confusion matrix is as follows:

a. **Calm Class:** Correctly classified: 1,940; misclassified: 5 (all misclassified as "Low"). The recall for the "Calm" class is 1940/(1940+5) ≈ 0.997, indicating that the model is highly accurate in recognizing calm sea conditions.
b. **Low Class:** Correctly classified (True Positives): 188; misclassified as "Calm": 21; misclassified as "Moderate": 1. The recall for the "Low" class is 188 / (188+1+21) ≈ 0.895, which is consistent with previous results. The primary misclassification is from

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

"Low" being predicted as "Calm," likely because the SWH values are close to the lower threshold (0.3 m).

c. **Moderate Class:** Correctly classified: 31; misclassified as "Low": 5. The recall for the "Moderate" class is 31 / (31+5) ≈ 0.861. While the model performs well, it still has some difficulty distinguishing between "Moderate" and "Low" due to the limited number of Moderate data points.
d. The classes "High," "Very High," and "Extreme" have a value of 0 (zero) because there were no wave (SWH) occurrences in those classes.

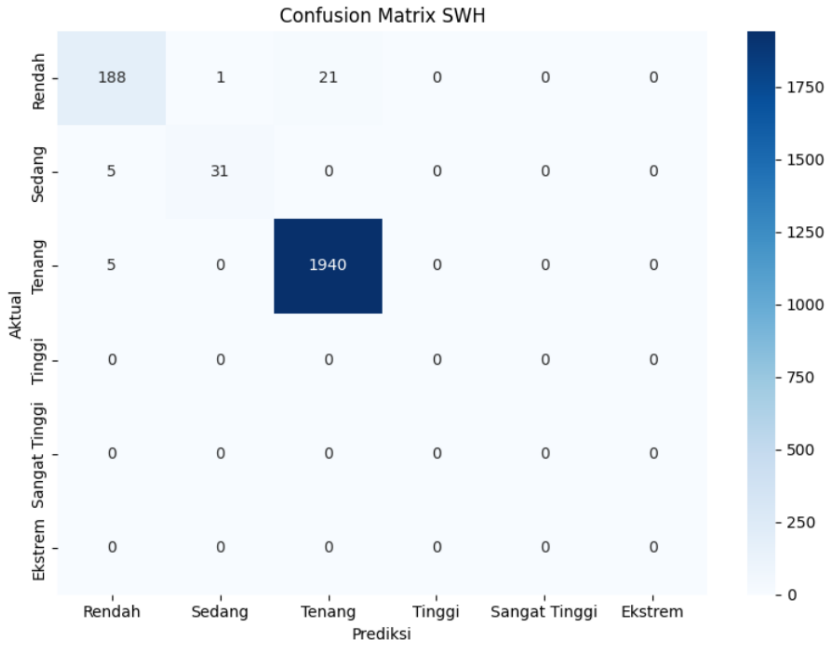

Figure 9. Confusion Matrix Results for SWH Classification on the Northern Coast of Central Java

Based on this analysis, the total correct predictions (diagonal values) are: 188 + 31 + 1940 = 2,159 out of a total of 2,191 records. This results in an overall accuracy of 2159 / 2191 ≈ 0.9854 (98.54%), which is consistent with the previous result. This indicates that the model performs exceptionally well in identifying "Calm" conditions and is reasonably strong for the "Low" class. While the "Moderate" class is more prone to confusion with "Low," it still shows satisfactory performance given the small sample size. These findings reinforce that Random Forest is capable of handling multiclass classification for SWH, even with imbalanced data.

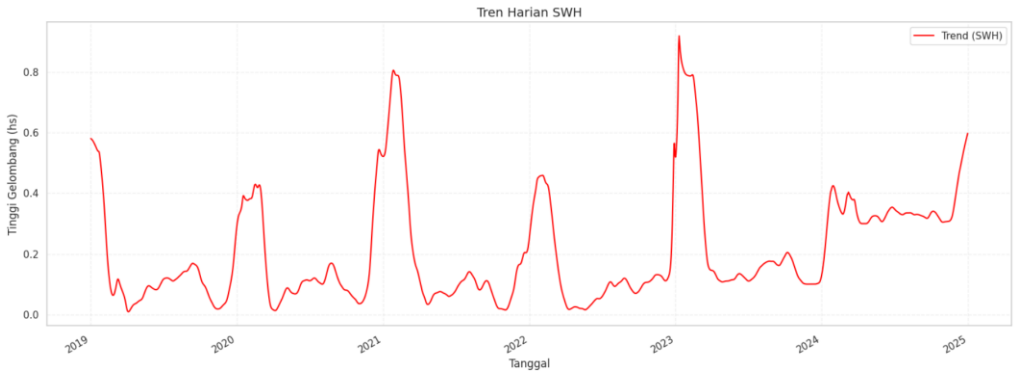

Figure 10. Daily Significant Wave Height (SWH) Trend from 2019 to 2024

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

Figure 10 presents a consistent seasonal pattern in wave height, with peaks typically occurring at the beginning of each year (January–March). The highest spikes were observed around 2021 and 2023 (>0.8 m), suggesting a strong seasonal cycle in the northern coastal region of Central Java. These recurring extreme events support the relevance of seasonal SWH classification, daily SWH prediction using machine learning, and early warning system development for high wave events in the area.

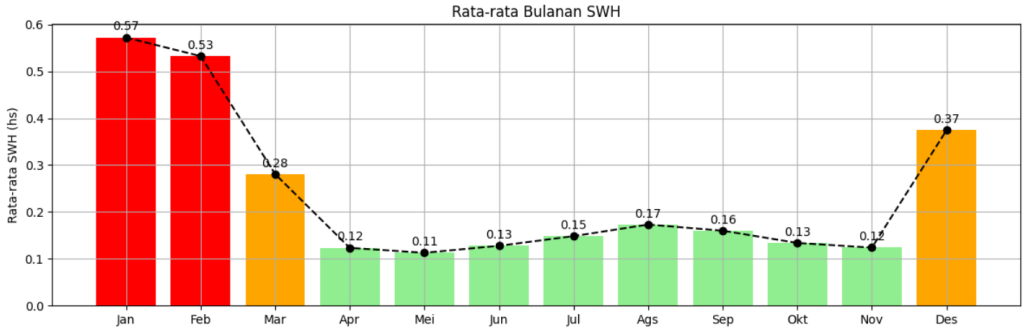

Figure 11. Monthly Average Significant Wave Height (SWH) from 2019 to 2024

The chart in Figure 11 shows a clear seasonal pattern in monthly average SWH, with peaks in January (±0.55 m) and February (±0.52 m), followed by another rise in December. This pattern reflects the influence of the westerly monsoon and is critical for seasonal classification, early warning systems, and planning of marine activities in the region. March (0.28 m) marks the end of the westerly monsoon season, while the lowest SWH values occur in May, June, October, and November, averaging only around 0.12–0.17 m. These months indicate relatively calm seas, typically during the easterly season or when no active storms are present.

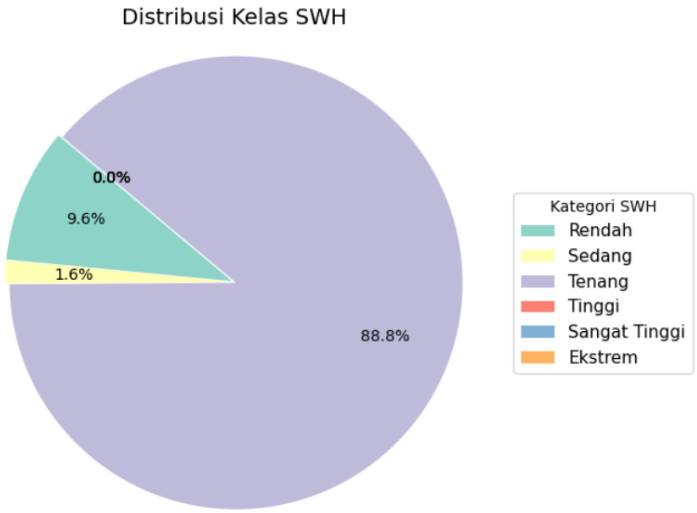

Figure 12. Pie Chart of SWH Classification on the Northern Coast of Central Java

The pie chart in Figure 12 illustrates the data distribution based on SWH classification: Calm (88.8%), Low (9.6%), and Moderate (1.6%). This class imbalance must be addressed in classification modeling to ensure that higher-risk wave conditions are accurately detected and not overlooked. The majority of sea conditions in the region fall under "Calm" (SWH ≤ 0.3 m). The "Low" class represents a small portion of data when waves start to increase but are not yet extreme (0.3 m < SWH ≤ 1.0 m). The "Moderate" class is rare but carries high risk, often associated with coastal disturbances and early warnings. To address this imbalance, machine learning models should implement techniques such as

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

resampling, class weighting, or use specialized evaluation metrics (F1-score, class-wise recall) to avoid misleading high accuracy due only to class dominance.

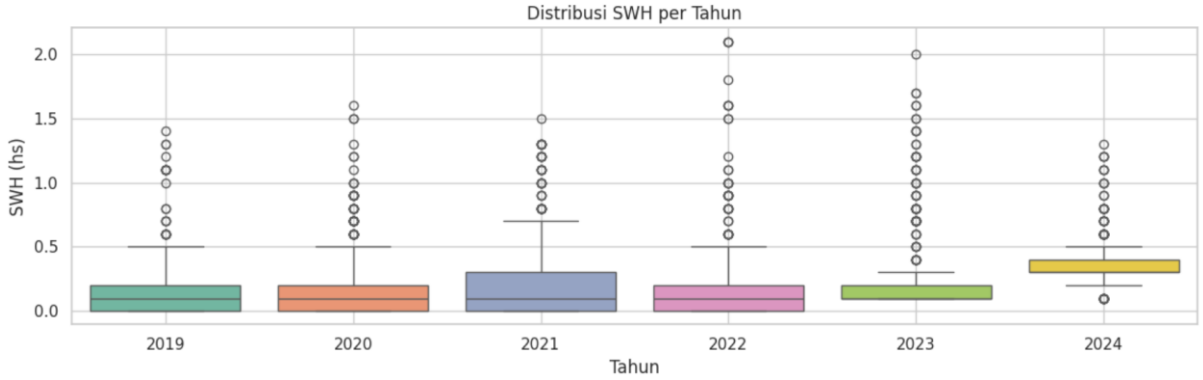

Figure 13. Boxplot of SWH Distribution (2019–2024)

Figure 13 shows annual boxplots of SWH, indicating that most significant wave heights (hs) fall below 0.5 meters, with high wave outliers (>1.0 m) appearing every year. The year 2023 exhibited the widest spread and most extreme outliers, suggesting more intense extreme wave events compared to other years. In 2024, the highest median supports the trend of increasing wave height. Extreme outliers (>1.5 m) occurred nearly every year, reflecting high waves caused by extreme weather. Increasing variability (IQR range) from 2019–2023 suggests a more dynamic and fluctuating ocean environment.

The Pearson correlation heatmap in Figure 14 indicates that the most positively correlated variable with SWH is the significant wave height (hs) itself, with a perfect correlation of 1.00, followed by surface current velocity (Kec_Arus) at 0.64, and wave period (t01) at 0.25. These findings suggest that increasing current speeds and longer wave periods contribute to higher SWH values, making them essential oceanographic features for both classification and prediction models.

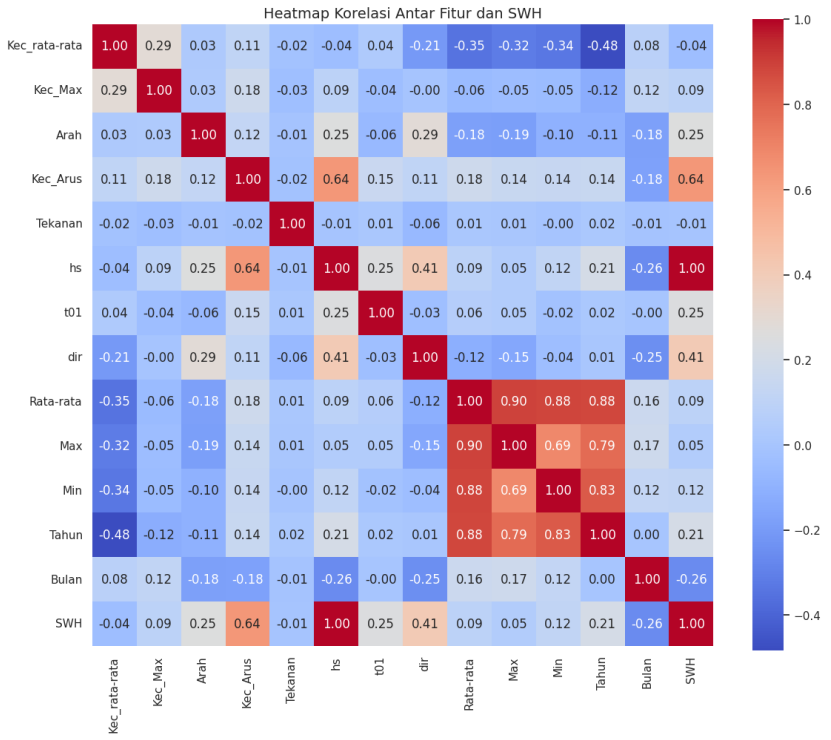

Figure 14. Heatmap of Variables Influencing SWH

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

On the other hand, some variables show negative correlations with SWH, such as minimum wave parameters (-0.26) and month (-0.26), indicating lower SWH values during certain months outside the westerly season. Meanwhile, variables like wind direction, air pressure, and year exhibit very weak correlations, suggesting insignificant linear relationships with SWH. Therefore, a machine learning approach capable of capturing non-linear feature interactions is required to accurately model sea wave conditions, especially in complex coastal environments like the northern coast of Java.

3.4. Evaluating SWH Accuracy Metrics

The Random Forest Regressor model used to predict Significant Wave Height (SWH) demonstrated excellent performance based on four key evaluation metrics:

1. **Mean Absolute Error (MAE) = 0.067**. MAE represents the average absolute difference between predicted and actual values. A value of 0.067 means that, on average, the prediction error is only around 6.7 cm, which is considered very small in oceanographic contexts.
2. **Mean Squared Error (MSE) = 0.012**. MSE measures the average squared difference between predicted and actual values. This low value indicates that extreme errors (outliers) are rare, as MSE is sensitive to large deviations.
3. **Root Mean Squared Error (RMSE) = 0.109**. RMSE is the square root of MSE and provides a standard deviation of prediction error in meters. An RMSE of 0.109 m suggests that the model's predictions deviate by about 11 cm on average from the actual values, which remains well within acceptable tolerance for maritime applications.
4. **R-squared (R^2) = 0.864**. R^2 indicates the proportion of variability in the actual data that can be explained by the model. With $R^2 = 0.864$, the model can explain 86.4% of the variation in SWH, demonstrating a high predictive capability.

Based on these evaluation metrics, the Random Forest Regressor model can be considered accurate, stable, and reliable in predicting significant wave height (SWH) in the northern coastal region of Central Java. The low prediction error and high R^2 value confirm the model's suitability for sea condition monitoring, early warning systems, and decision-making in maritime and coastal sectors. The previous sections have discussed both the actual SWH trends and the SWH classification using the Random Forest model. Their connection can be summarized as follows:

a. **SWH prediction serves as the basis for classification:** The RF regression model predicts continuous SWH values (in meters), which are then used to classify sea conditions into three categories: “Calm” (SWH ≤ 0.3 m), “Low” (0.3 m < SWH ≤ 1.0 m), and “Moderate” (SWH > 1.0 m). Therefore, the accuracy and precision of classification results directly depend on the accuracy of the regression model.
b. **The results show consistency between regression and classification:** The actual vs. predicted SWH plot reveals that most points lie close to the diagonal line, indicating that predictions closely match actual values. The high correlation ($R^2 = 0.864$) and low error (MAE = 0.067) confirm that the SWH predictions are sufficiently accurate. Consequently, as seen in previous classification results, the high classification accuracy (98.54%) and excellent classification performance for the majority class (“Calm”), as well as good precision and recall for minority classes (“Low”, “Moderate”), are well supported.

These points illustrate that the high accuracy of the regression model enhances the success of the classification model—particularly in maintaining precision at class boundaries (e.g., distinguishing between SWH 0.28 m vs. 0.32 m to separate “Calm” and “Low”). The actual vs. predicted SWH plot (Figure 14) shows that the regression model has high accuracy, which directly contributes to the

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

successful classification of SWH into three categories. The model's stable predictive accuracy ensures clear identification of class thresholds, resulting in precise classification—even for minority classes that are harder to distinguish.

The plot in Figure 15 shows that predictions at higher SWH values (>1.0 m) are slightly more scattered, which could potentially affect classification accuracy for the "Moderate" class. However, as previously reported, the "Moderate" class was still classified with 96.9% precision and 86.1% recall, meaning the model remains capable of effectively handling minority classes.

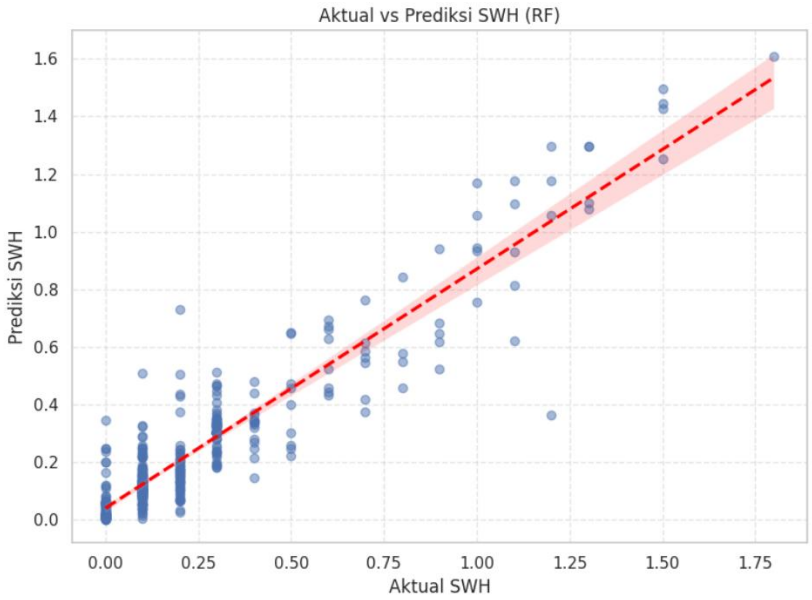

Figure 15. Scatter Plot of the Relationship between Actual and Predicted Significant Wave Height (SWH)

3.5. The Relationship Between Classification and Regression of SWH

The results of the classification and regression models for predicting Significant Wave Height (SWH) are deeply interconnected, with the regression model laying the foundation for the classification process. The Random Forest Regressor, used for regression, provided highly accurate predictions of continuous SWH values, with key performance metrics such as $R^2 = 0.864$, MAE = 0.067, and RMSE = 0.109, indicating that the model can explain 86.4% of the variability in the actual SWH data and maintain a low prediction error. These regression results directly influenced the classification process, where the predicted SWH values were categorized into three distinct classes: "Calm," "Low," and "Moderate." The accuracy and precision of this classification depend heavily on the accuracy of the regression model, as even small discrepancies in predicted values at class boundaries (e.g., SWH values between 0.3 m and 1.0 m) can affect the classification outcome.

The high accuracy of the regression model enhanced the success of the classification, particularly in terms of correctly identifying the class thresholds for each category. For example, the classification results demonstrated a very high accuracy (98.54%) for the majority class ("Calm"), with good precision and recall for the minority classes ("Low" and "Moderate"). Although there was some scattering in the predicted values for higher SWH values (>1.0 m), which slightly affected the classification of the "Moderate" class, the model still achieved 96.9% precision and 86.1% recall for this category. This illustrates that the regression model's stability and high accuracy ensured clear identification of the class boundaries, which supported precise classification, even for minority classes that are typically more difficult to distinguish. Thus, the seamless integration of both the regression and classification models contributes to a robust, reliable, and highly accurate SWH prediction system for maritime applications.

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

4. DISCUSSIONS

The results of this study demonstrate that the Random Forest algorithm is effective in classifying sea wave conditions and predicting significant wave height (SWH) with high accuracy, even under class imbalance conditions. The model successfully captures the non-linear relationships between oceanographic and meteorological variables and SWH. The strong regression performance ($R^2 = 0.864$) indicates its high potential for integration in monitoring and early warning systems. Despite the severe imbalance in the dataset, the model effectively identifies minority classes such as "Moderate", which pose higher risks to maritime activities. Based on these findings, the following recommendations are proposed:

a. The integration of this model into operational systems of institutions such as BMKG (Indonesian Meteorological, Climatological, and Geophysical Agency) for daily sea condition monitoring.
b. The application of data balancing techniques such as SMOTE to improve the model's performance on minority classes.
c. Expansion of the model to other coastal regions of Indonesia for broader validation. d. The inclusion of additional features such as sea surface temperature and the ENSO index to enhance prediction capabilities.
d. Random Forest (RF) is the best choice for SWH in regions like Pantura, which exhibit variations and non-linear relationships between oceanographic and meteorological factors. RF is also more stable and easier to optimize on complex multivariate data. Although other models such as XGBoost and SVM show good performance in SWH prediction, RF has the advantage in terms of implementation speed and clear feature interpretation. This contributes to marine safety and coastal risk mitigation planning.

5. CONCLUSION

This study successfully developed a classification and prediction model for significant wave height (SWH) and sea level rise (SLR) in the northern coastal region of Java (Pantura) using the Random Forest algorithm. Based on daily data from 2019 to 2024, the classification model achieved an accuracy of 98.54% in categorizing wave conditions into Calm, Low, and Moderate classes. Meanwhile, the regression model demonstrated strong predictive performance with $R^2 = 0.864$, MAE = 0.067, and RMSE = 0.109. These results indicate that the RF model is effective in handling multivariate and non-linear relationships. Despite the dataset being heavily dominated by the "Calm" class, the model maintained reliable performance in identifying the minority "Moderate" class. Additionally, long-term trends indicate an increase in SWH of approximately 3.6 cm per year, suggesting a growing risk of coastal hazards in the future. This study confirms that machine learning approaches, particularly Random Forest, offer practical and accurate solutions for maritime monitoring systems, early warning systems, and climate adaptation planning in tropical coastal regions such as Pantura. The model can also be adapted to similar oceanographic phenomena in other regions. This contributes to marine safety and coastal risk mitigation planning.

REFERENCES

[1] T. Djamaluddin *et al.*, "Simple model of sea level peak potentially trigger coastal flood on north coast of Java," *AIP Conference Proceedings*, vol. 3074, no. 1, 2024, doi: 10.1063/5.0211328.

[2] G. Brunet *et al.*, "Advancing Weather and Climate Forecasting for Our Changing World," *Bulletin of the American Meteorological Society*, vol. 104, no. 4, pp. E909–E927, 2023, doi: 10.1175/BAMS-D-21-0262.1.

[3] R. D. Susanto and R. D. Ray, "Seasonal and Interannual Variability of Tidal Mixing Signatures in Indonesian Seas from High-Resolution Sea Surface Temperature," *Remote Sensing*, vol. 14,

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

no. 8, 2022, doi: 10.3390/rs14081934.

[4] A. W. Nirwansyah and B. Braun, "Mapping impact of tidal flooding on solar salt farming in northern Java using a hydrodynamic model," *ISPRS International Journal of Geo-Information*, vol. 8, no. 10, 2019, doi: 10.3390/ijgi8100451.

[5] W. Setianingsih, B. Sasmito, and N. Bashit, "Analisis Level Rise di Laut Utara Jawa Terhadap Perubahan Garis Pantai Wilayah Demak Pada tahun 2006-2016," *Jurnal Geodesi Undip*, vol. 7, no. April, pp. 53–64, 2018.

[6] L. Q. Avia, "A comparative analysis of the wind and significant wave height on the extreme weather events (TC cempaka and TC Dahlia) in the Southern Sea of Java, Indonesia," *IOP Conference Series: Earth and Environmental Science*, vol. 572, no. 1, 2020, doi: 10.1088/1755-1315/572/1/012033.

[7] H. T. Mudho, I. A. Azies, J. Setiyadi, E. A. Kisnarti, and W. S. Pranowo, "Karakteristik Tinggi Gelombang Laut di Perairan Halmahera Utara dan Morotai pada Periode Waktu ENSO Tahun 2012-2021," *Jurnal Kelautan Tropis*, vol. 28, no. 1, pp. 11–24, 2025, doi: 10.14710/jkt.v28i1.25192.

[8] F. Anggraeni, D. Adytia, and A. W. Ramadhan, "Forecasting of Wave Height Time Series Using AdaBoost and XGBoost, Case Study in Pangandaran, Indonesia," *2021 International Conference on Data Science and Its Applications, ICoDSA 2021*, pp. 97–101, 2021, doi: 10.1109/ICoDSA53588.2021.9617524.

[9] C. Tay *et al.*, "Sea-level rise from land subsidence in major coastal cities," *Nature Sustainability*, vol. 5, no. 12, pp. 1049–1057, 2022, doi: 10.1038/s41893-022-00947-z.

[10] S. Susilo *et al.*, "GNSS land subsidence observations along the northern coastline of Java, Indonesia," *Scientific Data*, vol. 10, no. 1, pp. 1–8, 2023, doi: 10.1038/s41597-023-02274-0.

[11] A. A. Sidiq and F. W. Christanto, "Algoritma Naive Bayes Untuk Penentuan PKH (Program Keluarga Harapan) Berbasis Sistem Pendukung Keputusan," *Riptek*, vol. 14, no. 1, pp. 65–71, 2020.

[12] K. Triana and A. J. Wahyudi, "Sea level rise in Indonesia: The drivers and the combined impacts from land subsidence," *ASEAN Journal on Science and Technology for Development*, vol. 37, no. 3, pp. 115–121, 2020, doi: 10.29037/AJSTD.627.

[13] C. Murtiaji, M. Irfani, I. Fauzi, A. S. D. Marta, C. I. Sukmana, and D. A. Wulandari, "Methods for addressing tidal floods in coastal cities: An overview," *IOP Conference Series: Earth and Environmental Science*, vol. 1224, no. 1, 2023, doi: 10.1088/1755-1315/1224/1/012019.

[14] I. Rudiarto and D. Pamungkas, "Spatial exposure and livelihood vulnerability to climate-related disasters in the North Coast of Tegal City, Indonesia," *International Review for Spatial Planning and Sustainable Development*, vol. 8, no. 3, pp. 34–53, 2020, doi: 10.14246/irspsd.8.3_34.

[15] A. Durap, "Data-driven models for significant wave height forecasting: Comparative analysis of machine learning techniques," *Results in Engineering*, vol. 24, no. December, p. 103573, 2024, doi: 10.1016/j.rineng.2024.103573.

[16] L. Wu, E. Sahlée, E. Nilsson, and A. Rutgersson, "A review of surface swell waves and their role in air–sea interactions," *Ocean Modelling*, vol. 190, no. March, p. 102397, 2024, doi: 10.1016/j.ocemod.2024.102397.

[17] B. Hou, H. Fu, X. Li, T. Song, and Z. Zhang, "Predicting significant wave height in the South China Sea using the SAC-ConvLSTM model," *Frontiers in Marine Science*, vol. 11, no. August, pp. 1–14, 2024, doi: 10.3389/fmars.2024.1424714.

[18] T. Song, R. Han, F. Meng, J. Wang, W. Wei, and S. Peng, "A significant wave height prediction method based on deep learning combining the correlation between wind and wind waves," *Frontiers in Marine Science*, vol. 9, no. October, pp. 1–19, 2022, doi: 10.3389/fmars.2022.983007.

[19] O. Omonigbehin, E. O. O. Eresanya, A. Tao, V. E. Setordjie, S. Daramola, and A. Adebiyi, "Long-Term Evolution of Significant Wave Height in the Eastern Tropical Atlantic between

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108

1940 and 2022 Using the ERA5 Dataset," *Journal of Marine Science and Engineering*, vol. 12, no. 5, 2024, doi: 10.3390/jmse12050714.

[20] R. M. Campos, M. O. Costa, F. Almeida, and C. G. Soares, "Operational wave forecast selection in the atlantic ocean using random forests," *Journal of Marine Science and Engineering*, vol. 9, no. 3, 2021, doi: 10.3390/jmse9030298.

[21] J. Shi *et al.*, "A Machine-Learning Approach Based on Attention Mechanism for Significant Wave Height Forecasting," *Journal of Marine Science and Engineering*, vol. 11, no. 9, 2023, doi: 10.3390/jmse11091821.

[22] P. Pokhrel, E. Ioup, J. Simeonov, M. T. Hoque, and M. Abdelguerfi, "A Transformer-Based Regression Scheme for Forecasting Significant Wave Heights in Oceans," *IEEE Journal of Oceanic Engineering*, vol. 47, no. 4, pp. 1010–1023, 2022, doi: 10.1109/JOE.2022.3173454.

[23] W. Shen, Z. Ying, Y. Zhao, and X. Wang, "Significant wave height prediction in monsoon regions based on the VMD-CNN-BiLSTM model," *Frontiers in Marine Science*, vol. 11, no. November, pp. 1–17, 2024, doi: 10.3389/fmars.2024.1503552.

[24] H. Ding, "Using Random Forest for Future Sea Level Prediction," *SHS Web of Conferences*, vol. 174, p. 03008, 2023, doi: 10.1051/shsconf/202317403008.

[25] D. Demetriou, C. Michailides, G. Papanastasiou, and T. Onoufriou, "Coastal zone significant wave height prediction by supervised machine learning classification algorithms," *Ocean Engineering*, vol. 221, no. December 2020, p. 108592, 2021, doi: 10.1016/j.oceaneng.2021.108592.

[26] Y. Zhan, H. Zhang, J. Li, and G. Li, "Prediction Method for Ocean Wave Height Based on Stacking Ensemble Learning Model," *Journal of Marine Science and Engineering*, vol. 10, no. 8, 2022, doi: 10.3390/jmse10081150.

[27] N. A. Hazrin *et al.*, "Predicting sea levels using ML algorithms in selected locations along coastal Malaysia," *Heliyon*, vol. 9, no. 9, p. e19426, 2023, doi: 10.1016/j.heliyon.2023.e19426.

Vol. 6, No. 4, August 2025, Page. 2263-2280
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5108