DOI: 10.4274/jems.2025.34603

Prediction of Ship Trajectory and Critical Collision Zone in Sunda Strait Based on Automatic Identification System and Weather Data

Iis Dewi Ratih, Mochammad Reza Habibi, Kanugrahing Christy Sekar Arum

Institut Teknologi Sepuluh Nopember, Department of Business Statistics, Surabaya, Indonesia

Abstract

Indonesia as the largest maritime country, which dense shipping activities that increase the risk of ship accidents, especially in strategic areas such as the Sunda Strait. Extreme weather, such as storms and strong winds, increases this risk and requires special attention to improve shipping safety. This study aims to identify high-risk areas for ship encounters in the Sunda Strait, known as the Critical Collision Zone (CCZ). The CCZ is determined through ship trajectory prediction analysis using the Bi-GRU method and clustering with the DBSCAN algorithm. Trajectory data is obtained from Automatic Identification System (AIS) information and weather data. AIS data includes the position, speed, and direction of the ship in real time. Its integration with weather data allows for the formation of a more accurate trajectory. After the CCZ is identified, the probability of an encounter is calculated using the Monte Carlo Simulation method. The results show that the weather data-based prediction model performs better in identifying the CCZ, as indicated by lower MAE and MSE values and higher silhouette coefficients. These metrics improve the accuracy of identifying risky areas and estimating the probability of ship encounters in the Sunda Strait.

Keywords: Automatic Identification System (AIS), Critical Collision Zone (CCZ), Ship Encounter, Ship Trajectory, Weather

1. Introduction

Indonesia, as the world's largest archipelagic country with the second-longest coastline, plays a crucial role in international maritime trade. The Sunda Strait, one of the busiest shipping lanes in Indonesia, connects the Indian Ocean and the Java Sea, serving both domestic and international maritime transportation. The high volume of vessel traffic in this region increases the risk of ship encounters, particularly due to extreme weather conditions such as storms and strong winds. According to the Central Statistics Agency (BPS), the number of ship arrivals in Indonesian waters significantly increased in 2022, directly impacting the probability of maritime accidents. Reports from the National Transportation Safety Committee (KNKT) recorded 115 ship accidents over the last five years, highlighting the need for enhanced maritime safety measures.

To address navigation risks, the Traffic Separation Scheme (TSS) has been implemented. However, conventional ship trajectory models still struggle to incorporate both the sequential nature of Automatic Identification System (AIS) data and real-time environmental factors, reducing prediction accuracy. Various prior studies have attempted to improve ship trajectory prediction and encounter risk estimation.

To address navigation risks, the TSS has been implemented. However, conventional ship trajectory models still struggle to incorporate both the sequential nature of AIS data and real-time environmental factors, reducing prediction accuracy. Various prior studies have attempted to improve ship trajectory prediction and encounter risk estimation. For instance, Spyrou-Sioula et al. [1] demonstrated that incorporating AIS-based weather routing enables vessels to avoid hazardous routes and select safer alternatives, thereby preventing cargo losses due to extreme weather conditions. Han et al. [2] applied the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm to analyze ship movement patterns, revealing that this technique effectively identifies dense traffic areas prone to accidents. Vukša et al. [3] utilized Monte Carlo Simulation (MCS) to estimate ship encounter probability, proving its

Address for Correspondence: Iis Dewi Ratih, Institut Teknologi Sepuluh Nopember, Department of Business Statistics, Surabaya, Indonesia
E-mail: iis.dewi@statistika.its.ac.id
ORCID iD: orcid.org/0009-0003-4422-1083

Received: 17.01.2025
Last Revision Received: 14.04.2025
Accepted: 24.04.2025

To cite this article: I. D. Ratih, M. R. Habibi, and K. C. S. Arum, "Prediction of ship trajectory and critical collision zone in Sunda Strait based on automatic identification system and weather data," *Journal of ETA Maritime Science*, vol. 13(2), pp. 101-114, 2025.

ability to process complex AIS datasets and assess maritime accident risk using Bi-LSTM models.

Despite these developments, existing approaches have limitations. Traditional statistical or clustering methods often fail to capture the sequential, non-linear dynamics of ship movement. Moreover, deep learning models such as Bi-LSTM or Transformer-based predictors have not fully incorporated weather data for dynamic encounter risk modeling. Unlike those, our approach integrates weather information with a Bidirectional Gated Recurrent Unit (Bi-GRU) network for sequential learning, applies DBSCAN for spatial risk identification, and leverages MCS to simulate probabilistic encounters. This hybrid framework is designed to improve predictive accuracy while supporting proactive risk management.

This study aims to: (1) develop a Bi-GRU-based trajectory model incorporating AIS and weather data, (2) identify Critical Collision Zones (CCZs) using DBSCAN, and (3) estimate encounter probability through MCS. The proposed model not only advances predictive capability but also contributes to maritime safety planning by providing a data-driven assessment of navigational risks in complex sea routes like the Sunda Strait.

2. Materials and Methods

2.1. Data Preprocessing

The preprocessing criteria were carefully selected to enhance the quality and reliability of the AIS dataset, ensuring accuracy in ship trajectory modeling. Given the automatic nature of AIS transmissions, data can contain errors such as missing values, noise, or unrealistic values due to equipment malfunction or spoofing. This study focuses on key AIS dynamic parameters like Speed Over Ground (SOG) and Course Over Ground (COG), which are essential for modeling realistic ship behavior.

Several studies have noted that AIS data quality can vary significantly due to poor signal coverage, data transmission delays, and vessel configuration differences. In this study, outlier removal and quality filtering were conducted to ensure that only valid and realistic AIS records are used.

1. AIS data with SOG ≥30 knots are eliminated because values beyond this threshold often result from noise or incorrect signal calibration, as typical vessel speeds in the Sunda Strait are below this range [4].

2. Removing duplicate AIS data per ship ensures that each trajectory segment is representative of actual ship movement and avoids bias from duplicated or insufficiently sampled paths [5].

3. Excluding ships with AIS transmissions fewer than 20: Vessels with fewer than 20 AIS transmissions lack sufficient trajectory data for meaningful analysis [6].

4. Filtering ships with travel duration <4 hours: Ships with short travel durations may not provide sufficient movement variation for accurate trajectory predictions [7].

5. Filtering based on empirical speed ≥40 knots: This process cross-verifies dynamic SOG values with derived speed from positional change, removing cases that show inconsistent motion [4].

By implementing this rigorous preprocessing pipeline, the AIS dataset becomes more reliable and representative of actual vessel navigation patterns. While variables such as heading and rate of turn are retained, special emphasis is given to SOG and COG due to their central role in predicting position and estimating encounter probability. This approach minimizes the propagation of data uncertainty into the predictive model and enhances the robustness of trajectory simulations.

2.2. Ship Trajectory Prediction Using Bi-GRU

Bi-GRU is a Recurrent Neural Network model that is specifically designed to process sequential data and has the ability to consider information from the past and future [8]. A Bi-GRU was employed to predict ship trajectories based on historical AIS data. The model was selected for its capability to capture sequential dependencies in time-series data while considering both past and future contexts.

The Bi-GRU architecture consists of an input layer, a hidden layer, and an output layer. In the hidden layer, there are gates, namely, the reset gate and the update gate, which play a role in determining information from the input layer. The internal structure of the Bi-GRU model is shown in Figure 1.

2.3. Model Evaluation

Every prediction has an error, so the error rate in the prediction must be calculated to determine the accuracy of the created model. According to Azmi et al. [9], indicators such as Mean Square Error (MSE) and Mean Absolute Error (MAE) can be used to calculate prediction accuracy.

The average of the positive absolute error values derived from all observational data is known as the MAE [9]. The formula for calculating MAE is written as follows (Equation 1):

$$\mathrm{MAE} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}\left|y_{ijk} - \hat{y}_{ijk}\right| \tag{1}$$

The average of the squared differences between the expected value and the observed value is known as the MSE[9]. The prediction accuracy increases as the MSE value decreases. The formula for calculating MSE is written as follows (Equation 2):

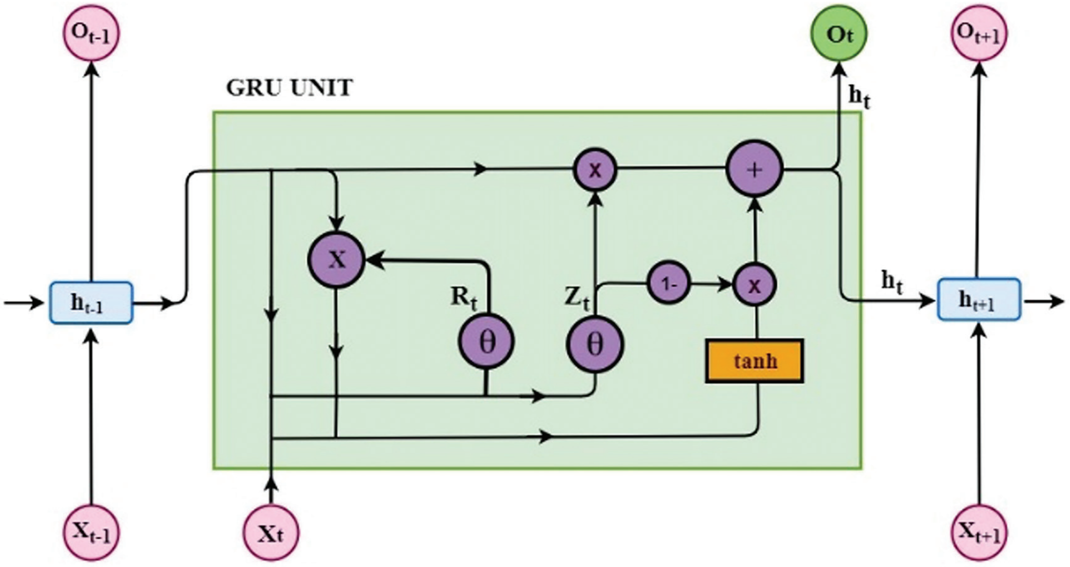


Figure 1. Internal Computing Structure of Bi-GRU

Bi-GRU: Bidirectional Gated Recurrent Unit

$$MSE = \frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}\left(y_{ijk} - \hat{y}_{ijk}\right)^2 \tag{2}$$

where m is the number of ships, n is the number of AIS data per ship, y_{ijk} is the actual value and $\hat{y}_{ijk}$ is the value of the predicted results.

2.4. CCZ Identification Using DBSCAN

To identify CCZs, the DBSCAN algorithm was applied to cluster high-density ship movements. DBSCAN is a density-based clustering method used to detect anomalous situations in spatial data by ignoring data noise. DBSCAN is designed for clustering large amounts of data. The DBSCAN algorithm takes two input parameters, namely Eps (ε) and MinPts. Eps is the maximum distance between two points to be considered neighbors, while MinPts is the minimum number of points required to form a dense region. DBSCAN identifies clusters as a collection of core points surrounded by a minimum number of other points (MinPts) within a certain radius [10].

2.5. Encounter Probability Estimation Using MCS

MCS is a statistical sampling method for estimating solutions to quantitative problems by building models based on real systems [11]. Each variable is represented by a Probability Density Function, enabling the selection of random values across thousands of simulations, depending on system complexity. In this study, the MCS process begins with the formation of a pseudo-population that represents ship movements within the CCZ. This pseudo-population accounts for key factors such as the number of ships, movement patterns, and residual distributions derived from the trajectory prediction model. Ship samples are then randomly selected, and at each coordinate point, a random residual value is added according to the predefined distribution. By repeating this process, multiple trajectory simulations are generated, capturing the inherent uncertainties in ship movement predictions. To ensure realistic encounter risk assessment, a 1 nautical mile (nm) threshold was adopted based on international maritime safety standards and validated using empirical encounter data.

2.6. Anderson Darling

Anderson Darling (AD) is a goodness-of-fit test used to test whether data come from a particular distribution [12]. Anderson Darling is a modification of the Kolmogorov Smirnov test giving more weight to the tails of the distribution, thereby making it more sensitive to changes. This method is often used to assess normal, lognormal, Weibull, and other distributions [13]. The goodness of fit value is obtained based on the following calculation.

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1]\left[\ln F\left(X_i\right) + \left(1 - \ln F\left(X_{n-i+1}\right)\right)\right] \tag{3}$$

Where X_i is a cumulative distribution function based on a certain distribution, n is the number of samples to be tested. A low Anderson Darling test statistic value indicates the suitability of the distribution tofor the observed data. The selection of data distribution can be done with two approaches: one is based on the goodness of fit value, and the other is visually using a probability plot.

Table 1. *Research variables*

Variable		Description
X_1	Latitude	Latitudes start at 0° from the Equator and end at 90° at the North Pole and 90° at the South Pole.
X_2	Longitude	Lines of longitude that stretch from 0° on the prime meridian to 180° east and west
X_3	SOG	The actual speed of a ship relative to the earth's surface
X_4	COG	The actual direction taken by a ship relative to the earth's surface
X_5	Heading	The direction in which the ship's bow is pointing, expressed in degrees from north
X_6	NAV	The ship's navigation status, such as underway, anchored, or in an emergency condition
X_7	ROT	Rate of change of direction of the ship
X_8	Datetime	Time and date information recorded when AIS data is transmitted
X_9	Wave direction	The direction of origin of ocean waves approaching a point on the sea surface
X_{10}	Wave period	The time interval measured between two successive wave crests passing a fixed point.
X_{11}	Wave height	The vertical measurement of the distance between the crest of the highest wave and the trough of the lowest wave.
X_{12}	Wind speed	The speed of air movement measured at a certain height above sea level
Y_1	Latitude prediction	Latitude prediction results
Y_2	Longitude prediction	Longitude prediction results

Probability plotused in conjunction with AD to visually illustrate the fit of a distribution. This is done by checking if the data points follow a straight line produced by the expected distribution, thus indicating a good fit between the data and that distribution. Anderson-Darling gives better results for large datasets than other tests, and is very sensitive to small changes in the data [14].

2.7. Data Sources

This study utilizes secondary data, obtained from third-party providers rather than collected directly by the researchers. The AIS data were sourced from the Indonesian Navigation District (Disnav), and the weather data were obtained from the European Centre for Medium-Range Weather Forecasts. Both sources are widely recognized for maritime studies and are highly relevant to the objectives of this research.

The dataset covers vessel traffic in the Sunda Strait from May 1 to May 30, 2021, comprising data from 1,740 ships and over 3 million AIS messages. The weather data includes wave and wind parameters at regular time intervals across the same spatial area. AIS data are categorized into Static (e.g., MMSI, ship type), Dynamic (e.g., position, SOG, COG), and Voyage-Related information (e.g., ETA). However, in the Indonesian context, publicly accessible metadata is limited. For example, over 30% of the vessels in our AIS data lacked ship type information. Although supplementary databases such as Equasis or MarineTraffic could be used to enrich metadata using the MMSI field, access to such platforms is typically restricted or subscription-based.

Despite these limitations, the AIS dataset remains suitable for trajectory analysis because the study focuses primarily on dynamic movement parameters (SOG, COG, heading, etc.) rather than vessel classification. Additionally, a preprocessing strategy was employed to improve data reliability, as described in Section 2.1. These efforts ensure that the AIS data used in this study are accurate and representative of real-world vessel behavior under normal and adverse weather conditions.

AIS transmits three main types of data: (1) Static Data, which includes ship identification, dimensions, and type, (2) Dynamic Data, which contains real-time navigational parameters such as position, speed, and COG, and (3) Voyage-Related Data, which provides information on the ship's current voyage, including destination and estimated time of arrival. Additionally, AIS supports Safety-Related Messages, which are manually transmitted alerts for emergency and navigational safety purposes.

In this study, 8 data variables and 4 weather data variables were used to model ship trajectories and assess encounter risk. The operational definitions of these research variables are detailed as follows (Table 1).

This study focuses on AIS-based ship movement data and four weather-related variables. Other environmental factors, such as visibility, time of day, and tidal currents, were not included due to data availability constraints. While these factors may influence ship trajectories, the selected variables represent widely accessible and commonly used indicators in maritime safety research. This limitation is acknowledged to clarify the study's scope and potential areas for future research.

3. Results and Discussions

3.1. Preprocessing and Compilation of AIS and Weather Data

Data preprocessing aims to clean and align data, ensuring it is ready for further processing in the prediction model.

3.1.1. AIS Data Preprocessing

Preprocessing AIS data ensures that the data used in the analysis is clean, relevant, and in accordance with the research objectives. The preprocessing stages undertaken begin with the initial input of AIS data and proceed to further filtering, which results in a reduction in the number of ships and AIS data used in the analysis. The results of preprocessing on AIS data are shown in Table 2.

The initial AIS data for one month consisted of 3,216,862 data points sent by 1,740 ships. After going through a filtering process based on speed variables and other characteristics, the amount of data used was reduced to 566,199 data points from 1,317 ships. This preprocessing eliminated about 82.4% of the total initial data, leaving about 17.6% of the data relevant for analysis in this study.

3.1.2. Weather Data Preprocessing

Preprocessing weather data ensures that the analyzed data is clean and in accordance with research needs. Weather data for one month consists of 84,240 data points and has a missing value percentage of 82.05% in: wave direction, wave period, and wave variables. This condition requires proper handling of missing values so that the analysis can be carried out accurately. The imputation method is used to overcome missing values using a weighted average that considers the distance between location points, calculated using the Haversine distance. Weather data after preprocessing is shown in Table 3.

Weather data, after preprocessing with the imputation method using weighted average, shows that missing values in the wave direction, wave period, and wave height variables have been filled with the appropriate estimated values. Each row in the weather data has a complete value; therefore, the percentage of missing values reaches 0%. These results ensure that the data are ready for further analysis without any empty values that can affect the accuracy of the study.

***Table 2.** AIS data preprocessing stages*

No	Stages	Number of ships	AIS data amount
1	Input AIS data	1,740	3,216,862
2	Eliminate data with SOG ≥30 knots	1,737	3,212,055
3	Eliminate duplicate data for each ship	1,737	573,262
4	Eliminate ships with AIS data <20	1,340	568,510
5	Eliminate ships with travel <4 hours	1,317	567,656
6	Eliminate data with empirical speed ≥40 knots	1,317	566,199

3.1.3. AIS and Weather Data Compilation

Compilation of AIS data with weather data aims to combine the two types of data so that the analysis can provide more comprehensive results. This compilation process is carried out using the K-Nearest Neighbors (KNN) method, where each AIS data point is paired with the nearest weather data based on a combination of date, time, and geographic coordinates (latitude and longitude). The AIS dataset, including weather data, is shown in Table 4.

Table 4 shows the results of the compiled dataset from AIS data and weather data based on time and location using the KNN method. The weather data, which was originally limited to 84,240 data, became the same as the number ofmatched the number of AIS data, which was 566,199 data. This shows that each ship point has the closest weather information at the appropriate time and location, so it can be used for further analysis.

3.2. Ship Characteristics

The analysis of the characteristics of ships passing through the Sunda Strait from May 1, 2021 to May 30, 2021 aims to

***Table 3.** Weather data after preprocessing*

No	Wave direction	Wave period	Wave height	Wind speed
1	179,4654	11,8968	1,3693	1,9334
2	179,4986	11,8659	1.3565	1,3667
3	179,2352	11,7715	1,3391	1,5820
4	178,5879	11,7487	1.3244	1.6272
5	177,5606	11,6754	1,3159	2,9260
⋮	⋮	⋮	⋮	⋮
2,881	90.9995	4.2245	0.2581	0.7522
2,882	91,6771	4,2210	0.2573	0.7185
2,883	92,3630	4,2182	0.2646	0.1997
2,884	92.6256	4,2138	0.2744	0.1625
2,885	92.6849	4,2055	0.2720	2,9414
⋮	⋮	⋮	⋮	⋮
84,236	164,0356	11,3876	1,2186	0.7987
84,237	164,3008	11,3604	1,2064	0.8433
84,238	164,3805	11,3357	1,1949	1,0780
84,239	164,2945	11,3253	1,1862	1,3776
84,240	164,2379	11,2994	1,1816	1,5518

provide an overview of the variety of ships passing through the Sunda Strait. Ship characteristics include the type of ship shown in Figure 2 and the country of departure shown in Figure 3.

Figure 2 shows that the most dominant ship type is “Unknown” with a total of 465 ships, indicating many ships do not send ship type information. The second most common ship type is the Bulk Carrier with 174 ships, followed by the Container Ship with 101 ships. This indicates that the Sunda Strait is an important route for the transportation of large amounts of cargo, highlighting import and export activity. Chemical and oil/products Tanker ships are also prevalent in the Sunda Strait, reflecting the demand for the transportation of chemicals and oil, thereby creating business opportunities for fuel providers and maritime security.

The distribution of flag states among vessels transiting through the Sunda Strait highlights its role as both a domestic and international maritime corridor. The high number of Indonesia-flagged vessels aligns with the country’s national maritime policies aimed at enhancing inter-island connectivity and domestic trade efficienc. Indonesia’s maritime connectivity policies focus on reducing logistics costs and improving trade flow between islands, which explains the significant presence of domestically

Table 4. *Dataset AIS and weather data compilation*

No	Latitude	Longitude	SOG	…	Wave period	Wave height	Wind speed
1	-5,9302	106,1183	2.0	...	9,4526	0.7834	4,4977
2	-5,9299	106,1181	1.0	...	9,4526	0.7834	4,4977
3	-5,9291	106,1169	0.5	...	9,4526	0.7834	4,4977
4	-5,9294	106,1171	1.0	...	9,4526	0.7834	4,4977
5	-5,9283	106,1184	1.0	...	9,4526	0.7834	4,4977
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
566,195	-5,9742	106,1197	6.0	...	11,7166	1,0164	1.4459
566,196	-5,9787	106,1184	7.0	...	11,7166	1,0164	1.4459
566,197	-5,9818	106,1161	5.5	...	11,7277	1,0350	1,4905
566,198	-5,9785	106,1157	7.0	...	11,7277	1,0350	1,4905
566,199	-5,9602	106,1146	15.5	...	11,7277	1,0350	1,4905

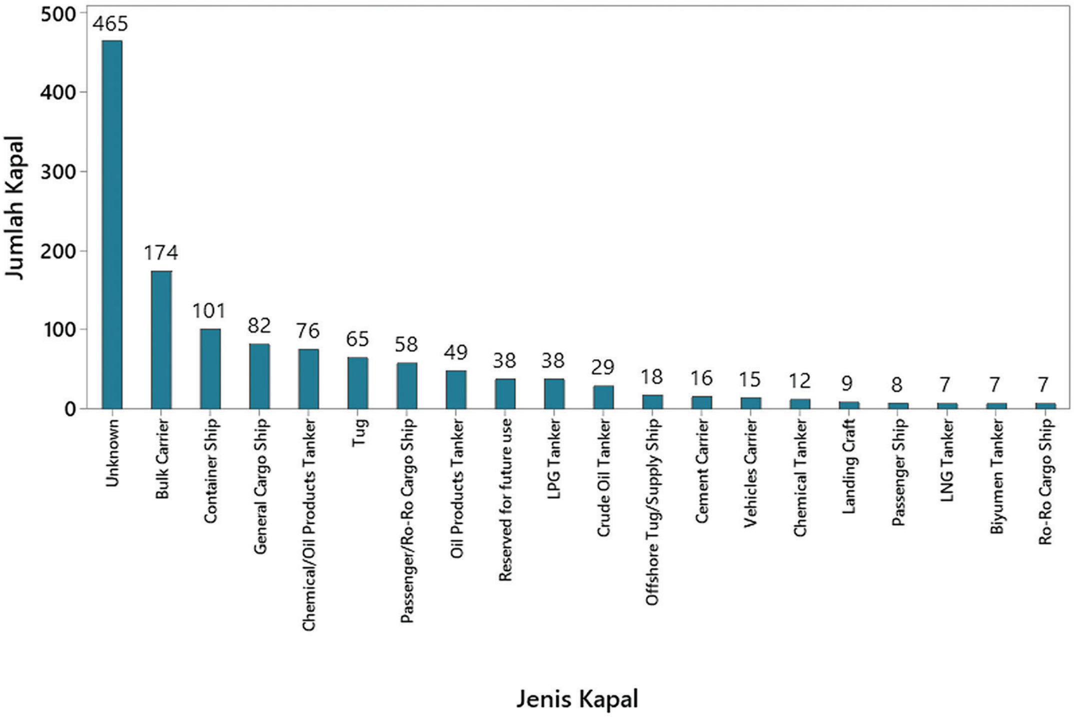


Figure 2. *Types of ships in the Sunda Strait*

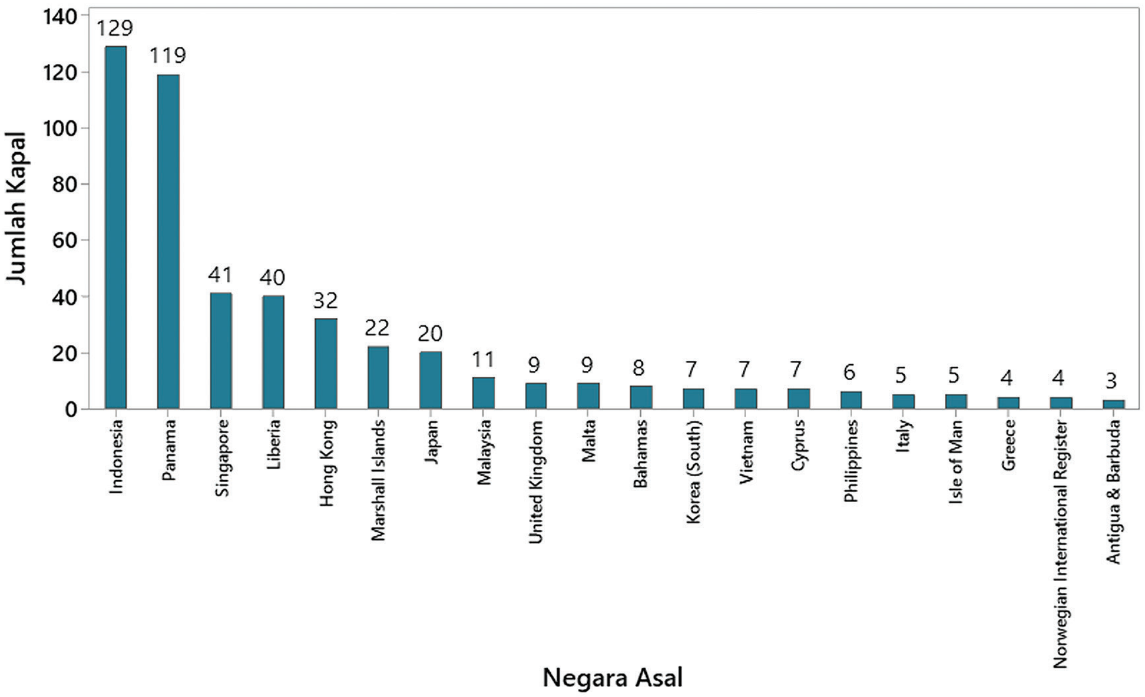


Figure 3. *Country of origin of ship departures in the Sunda Strait*

registered vessels in the strait [15]. Panama is one of the most commonly used flag states in global shipping, as shown by the number of Panama-flagged vessels transiting through the Sunda Strait. The widespread use of Panama's ship registry is due to its open registration system, which allows ships from various countries to be registered under its flag. Similar patterns are also observed in Liberia and the Marshall Islands, which have become key flag states in international maritime activities. These flag states play a crucial role in global trade, as they are widely chosen for their flexibility and operational advantages. These findings confirm that the Sunda Strait plays a crucial role in both national and international maritime activities, reinforcing its importance as a key global shipping passage.

3.2.1 Prediction of Ship Trajectory in Sunda Strait

Ship trajectory prediction is performed to understand ship movement patterns, with analysis using the Bi-GRU model to produce accurate predictions based on historical data.

3.2.2. Bi-GRU Hyperparameters

Systematic hyperparameter tuning allows the model to achieve a balance between accuracy and generalization ability, resulting in better predictions. The hyperparameters used in this study are shown in Table 5.

3.2.3. Ship Trajectory Prediction Results

The comparison between actual data and trajectory prediction results is shown more clearly, and this is achieved through tables and visualizations. The results of the ship trajectory prediction are shown in Table 6.

Table 6 shows that incorporating weather factors in ship trajectory prediction improves accuracy compared to predictions without weather. The comparison in Table 6 indicates that the inclusion of weather data enhances trajectory prediction accuracy. On average, the positional error for latitude is reduced by approximately 50%, from 0.0054° (without weather) to 0.0027° (with weather). Similarly, the longitude error is reduced by 43%, from 0.0107° to 0.0061°. These findings confirm that integrating weather variables into the prediction model leads to smaller differences between predicted and actual positions, improving overall trajectory estimation. The ability to achieve higher accuracy in trajectory prediction is particularly critical in high-traffic maritime zones such as the Sunda Strait, where precise ship positioning can help mitigate navigation risks and prevent potential encounters. The comparison of actual data with the trajectory prediction results is shown in Supplementary Figure 1.

Supplementary Figure 1 shows a comparison between actual data and predicted ship trajectory results in the Sunda Strait for two ships with different MMSI. The predicted points using weather variables are closer to the actual track than the predictions without weather variables, indicating an increase in prediction accuracy on complex routes. These results indicate that the predicted AIS data can be relied upon for CCZ identification through clustering.

Table 5. *Bi-GRU Hyperparameters*

No	Hyperparameter	Mark	Descriptions
1	Optimizer	Adam	Adjusts model weights based on the difference between predictions and actual values to accelerate convergence
2	Loss Function	MSE, MAE	Measures prediction errors, where MAE captures absolute errors, and MSE accounts for squared errors
3	Bi-GRU Layers	2	Determines the number of Bi-GRU layers used to capture sequential patterns in AIS data
4	Unit	24	Specifies the number of neurons per Bi-GRU layer for processing AIS movement data
5	Number of Epochs	20	Represents the total number of training iterations over the entire dataset
6	Batch Size	64	Defines the number of samples processed before model weight updates
7	Early Stopping Patience	3	Sets the number of epochs without validation improvement before training stops to prevent overfitting
8	Dense Unit	12	Determines the number of neurons in the fully connected layer before the final prediction
9	Activation Function	ReLU	Enhances computational efficiency and training stability

Table 6. *Comparison of predicted and actual ship positions with and without Weather Factors*

Current		Weather forecast without weather		Prediction with weather	
Latitude	**Longitude**	**Latitude**	**Longitude**	**Latitude**	**Longitude**
-5,9292	106,1170	-5,9353	106,1054	-5,9322	106,1121
-5.9225	106,1231	-5,9279	106,1123	-5,9253	106,1169
-5,9236	106,1222	-5,9285	106,1116	-5,9262	106,1153
-5,9065	106,1269	-5,9110	106,1177	-5,9087	106,1206
-5,9031	106,1658	-5,9091	106,1545	-5,9083	106,1597
⋮	⋮	⋮	⋮	⋮	⋮
-5,9685	106,1117	-5,9740	106,0990	-5,9715	106,1071
-5,9685	106,1118	-5,9740	106,0990	-5,9716	106,1071
-5,9672	106,1123	-5,9692	106,1049	-5,9677	106,1107
-5,9602	106,1147	-5,9568	106,1133	-5,9644	106,1357
-5,9688	106,1150	-5,9693	106,1089	-5,9694	106,1159

3.2.4. Bi-GRU Model Performance

The performance of the Bi-GRU model will be evaluated using the MSE and MAE metrics to measure how well the model predicts the ship trajectory. In addition, loss and validation loss graphs are used to monitor the training and validation process, as well as to identify whether the model is overfitting or underfitting.

Supplementary Figure 2 shows the plots that compare training loss and validation loss for trajectory prediction with (orange) and without (blue) weather factors. The inclusion of weather data results in slightly lower loss values, indicating improved model accuracy. In the loss graph the loss value decreases drastically in the first few epochs, then stabilizes at a low value. This indicates that the model has successfully learned and minimized errors in the training data. In the validation loss graph, there is a significant decrease in the first few epochs followed by stability, both for models with and without weather data. There are no signs of increasing validation loss or significant differences between loss and validation loss, which means that the model does not experience overfitting and can generalize well. This shows that the model is effective at learning data patterns and can be relied on to proceed to feature extraction for Bi-GRU modeling to improve the accuracy of ship trajectory prediction.

Table 7 presents the evaluation metrics for the Bi-GRU trajectory prediction model, comparing results with and without weather variables. The MAE represents the absolute difference between predicted and actual ship locations, measured in degrees (°) of latitude and longitude. This metric provides a direct interpretation of the typical positional error in trajectory prediction. Meanwhile, the MSE quantifies the squared positional differences before averaging, making it more sensitive to large deviations. Unlike MAE, MSE does not have a direct unit of measure and serves as a relative measure of prediction accuracy. The results show that

***Table 7.** Bi-GRU model evaluation metrics*

Model	Without weather variables		With weather variables	
	MAE	MSE	MAE	MSE
Bi-GRU	0.00976	0.00116	**0.00709**	**0.00112**

incorporating weather variables reduces prediction errors, with MAE decreasing from 0.00976° to 0.00709°, while MSE shows a slight improvement from 0.00116 to 0.00112. This confirms that the inclusion of weather data enhances the model's predictive accuracy by reducing both absolute errors and large deviations in ship trajectory estimation.

3.3. Identification of CCZ

CCZ identification in the Sunda Strait was conducted using the DBSCAN clustering method to detect areas with high ship traffic density and significant encounter risks. Encounter risk refers to the probability of ships encountering each other at close distances in congested maritime zones, thereby increasing the likelihood of navigational accidents. This study quantifies risk based on ship density in CCZs and projected movement patterns over time. By integrating AIS and weather data, the model enhances risk estimation, helping identify hazardous zones and improve maritime safety measures. The clustering process focused on intersecting domestic and international shipping routes, particularly near Merak and Bakauheni Ports, where heavy traffic makes precise risk assessment crucial.

DBSCAN clustering, which relies on distance matrices between AIS data points, demands high computing power due to the large dataset involved. Identifying CCZs in this area aids in implementing targeted risk mitigation strategies at key navigational chokepoints. While a 1 nautical mile (nm) threshold is a standard benchmark in maritime risk assessments, actual encounter risks depend on factors such as ship speed, heading, and relative positioning. The model dynamically incorporates trajectory predictions to reduce false positives, enhancing accuracy. Future work will integrate vessel maneuverability parameters to further refine encounter risk assessments and improve proactive maritime safety measures.

3.3.1. DBSCAN Parameters

The main parameters of DBSCAN, namely epsilon (ε) and minimum points (minPts), affect the clustering results and the amount of noise generated. The optimal values of epsilon and minPts are determined using a k-distance plot, with the elbow point on the curve as a guide. The analysis was performed on a random sample of 10,000 data points to fit within the researcher's computing resources. The k-distance plot of the prediction results with and without weather data, using k = 175, is shown in Figure 4.

Figure 4 shows the distance of each data point to the 175 nearest points, sorted from the smallest to the largest value. Significant changes occur in the data interval between 8,000 and 10,000, forming an elbow point around the value of 300. This elbow point indicates a larger distance between points, so the optimal epsilon for DBSCAN is approximately 300.

3.3.2. CCZ Clustering Results

The clustering results with DBSCAN produce the geographical boundaries of the CCZ in the Sunda Strait, which are used to group coordinate points into relevant clusters. The clustering results using the optimal parameters are shown in Figure 5.

Figure 5 shows three main categories, namely outliers, cluster 1, and cluster 2, in both models without and with weather data. Outliers indicate ships outside the high-density areas and not following the main traffic patterns. Cluster 1 is located in the west, reflecting the high concentration of ships around domestic lane intersections, while Cluster 2 is located in the east, indicating a major shipping lane with significant congestion. Both clusters indicate high-risk areas that require attention in mitigating encounter risks.

Table 8 shows the formation of two clusters, indicating the CCZ in the Sunda Strait, in both the models without and with weather data. Cluster 1 includes 8,549 points in the model without weather and 8,540 points in the model with weather, while cluster 2 includes 1,087 points in the model without weather and 1,100 points in the model with weather. The silhouette score of 0.5425 in the model without weather and 0.5489 in the model with weather indicates a fairly good separation. Higher values in the model with weather data indicate increased accuracy in identifying the CCZ.

3.4. Probability of Ship Encounter in the CCZ

MCS is used to calculate the probability of ship encounter in the CCZ considering uncertainties of variables, such as ship motion and weather conditions. MCS generates ship movement scenarios based on relevant probability distributions, so that interactions between ships are modeled realistically. Each ship in both CCZs is simulated with a trajectory that reflects the uncertainties in the prediction model.

3.4.1. Ship Trajectory Simulation

The ship trajectory simulation is obtained by adding random values from the residual distribution of the prediction model to each ship coordinate point, based on historical data in the CCZ. The most appropriate distribution for the residuals of the Bi-GRU prediction model is determined through a goodness of fit test using the Anderson-Darling method. The distribution that gives the best Anderson-Darling value is selected for use in the simulation, allowing for an accurate

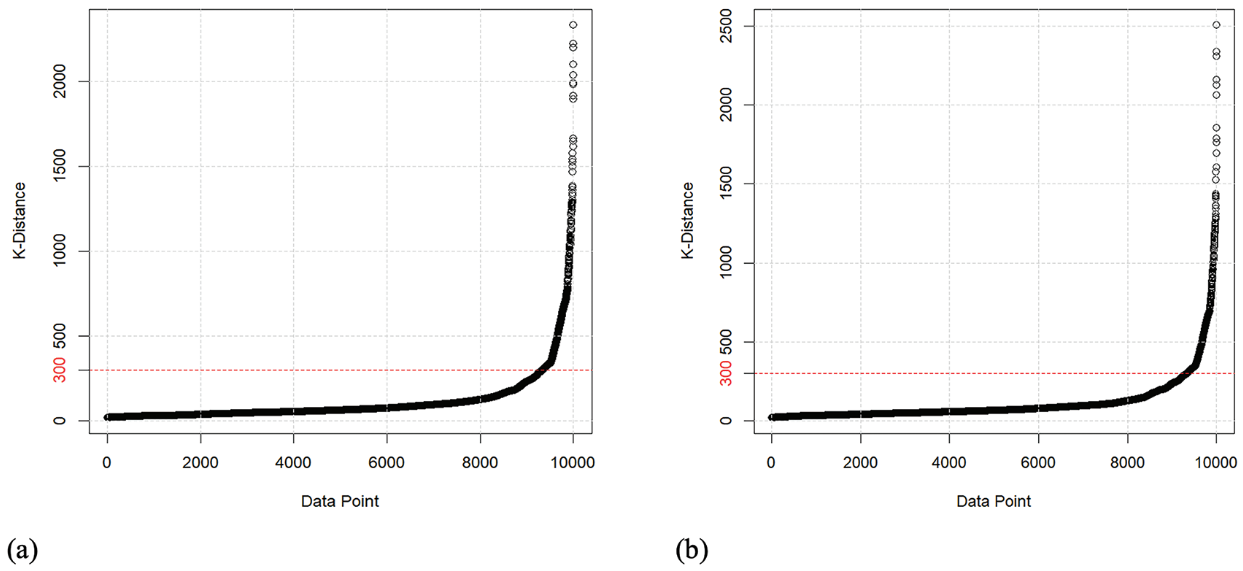


Figure 4. K-distance plot without weather (a) and with weather (b)

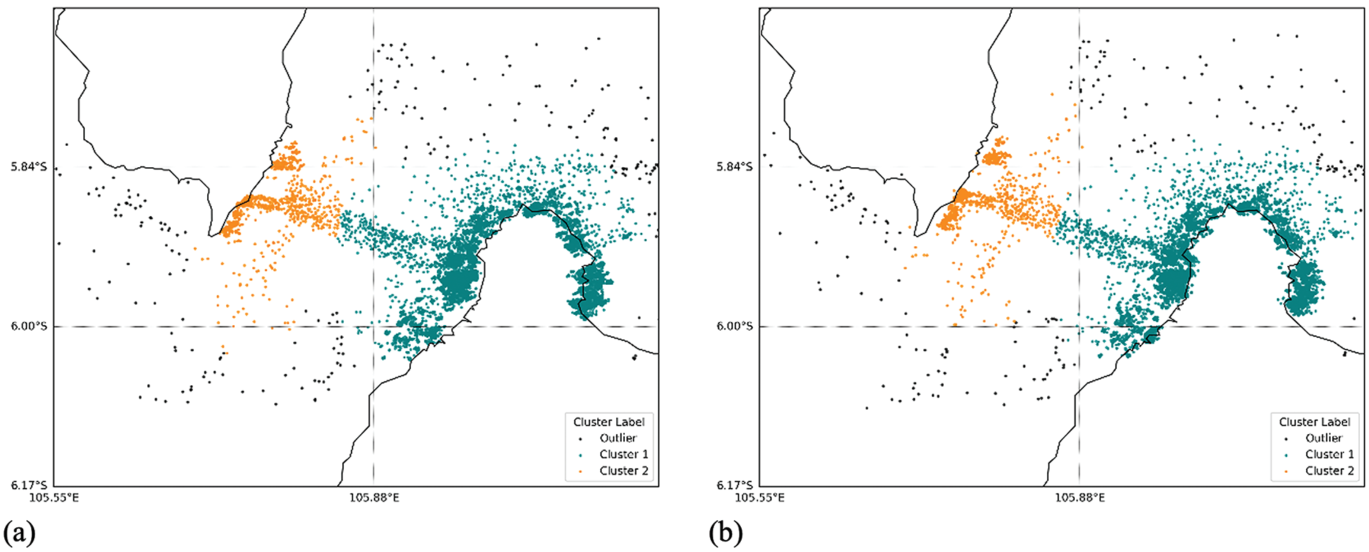


Figure 5. Clustering results without weather (a) and with weather (b)

representation of the uncertainty of the ship's motion.

Table 9 shows the relatively low AD values, in the Loglogistic distribution, both in MSE latitude and MSE longitude, for the models without weather and with weather. This value indicates that the MSE latitude and MSE longitude in the models with and without weather follow the Loglogistic distribution. The Loglogistic distribution shows the lowest AD value, but the selection of the distribution is still further analyzed through visualization, using probability plots to ensure the distribution matches the actual data. Based on the probability plot of each distribution, the MSE latitude and longitude are estimated to follow the Weibull distribution because the points on the probability plot follow a line. This indicates that the Weibull distribution is suitable to represent the residual data. The probability plot of the Weibull distribution against the residual latitude and longitude is visualized in Figure 6.

Figure 6 shows the residual distribution of MSE latitude and MSE longitude, that follows the Weibull distribution. The data points mostly follow a line, especially in the middle of the curve, which indicates a good fit of the Weibull distribution to the residual data. Some deviations are seen in the tail of the data, indicating a poor fit to extreme data. Overall, the Weibull distribution is effective in describing the residual pattern of the prediction model.The visualization in Figure 7 shows a comparison of trajectory

Table 8. *CCZ cluster characteristics*

Model	Cluster	Amount of Data	Coordinate		Silhouette Coefficient
			Minimum	Maximum	
No Weather	1	8,549	6°2' S, 105°50' E	5°48' S, 106°10' E	0.5425
	2	1,087	6°2' S, 105°42' E	5°47' S, 105°53' E	
With Weather	1	8,540	6°2' S, 105°51' E	5°49' S, 106°10' E	**0.5489**
	2	1,100	6°0' S, 105°42' E	5°46' S, 105°53' E	

Table 9. *Goodness of FitBi-GRU Residuals*

No	Distribution	Weatherless model		Model with weather	
		AD MSE Latitude	AD MSE Longitude	AD MSE Latitude	AD MSE Longitude
1	Weibull	40,078,015	29,672,837	7,021,465	6,214,237
2	Lognormal	15,057,068	27,259,079	20,361,333	6,507,489
3	Exponential	69,989,553	93,338,939	143,392,115	477,650,399
4	Normal	191,116,802	189,148,761	166,552,149	200,805,814
5	Logistic	194,946,908	159,954,120	170,933,386	207,814,098
6	Loglogistics	**4,403,143**	**10,092,395**	12,350,652	**2,919,226**

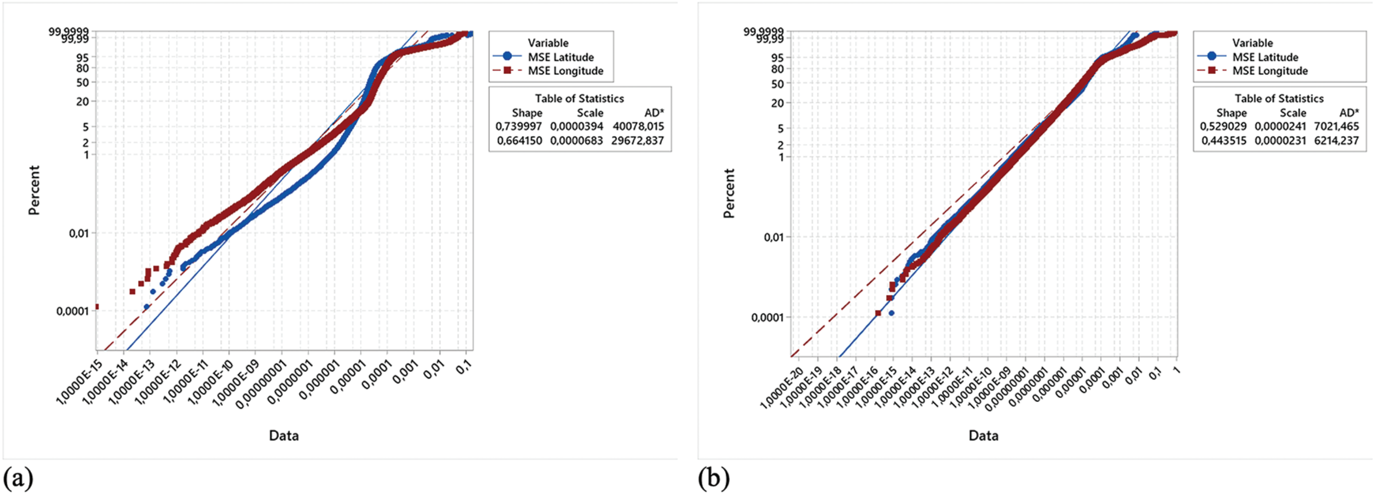


Figure 6. *Probability plot model without weather (a) and with weather (b)*

prediction and simulation results for two different ships with and without weather variables. The Weibull distribution produces realistic path variations, indicating uncertainty in unpredictable environmental conditions.

3.4.2. Chances of a Ship Encounter

The probability of ship encounter is obtained from the analysis of trajectory simulations generated through MCS. The distance between ships at each trajectory point is calculated. Ship encounters are likely to occur if the distance between ships is less than 1 nautical mile (1,852 km). The frequency of occurrence is calculated from all simulations to estimate the probability of encounter in each CCZ as shown in Table 10.

Table 10 shows the encounter probability in CCZ for models with and without weather data. In cluster 1, the encounter probability with weather data is 0.41653, which is higher than the probability without weather data, at 0.41189, indicating the contribution of weather data to the accuracy of risk estimation. In cluster 2, the encounter probability with weather data is 0.83199, which is higher than the probability without weather data 0.71645, indicating that weather information provides a more realistic picture of the risk of ship interaction in the field.

4. Conclusion

This study develops a Bi-GRU-based trajectory prediction model, incorporating AIS and weather data to improve

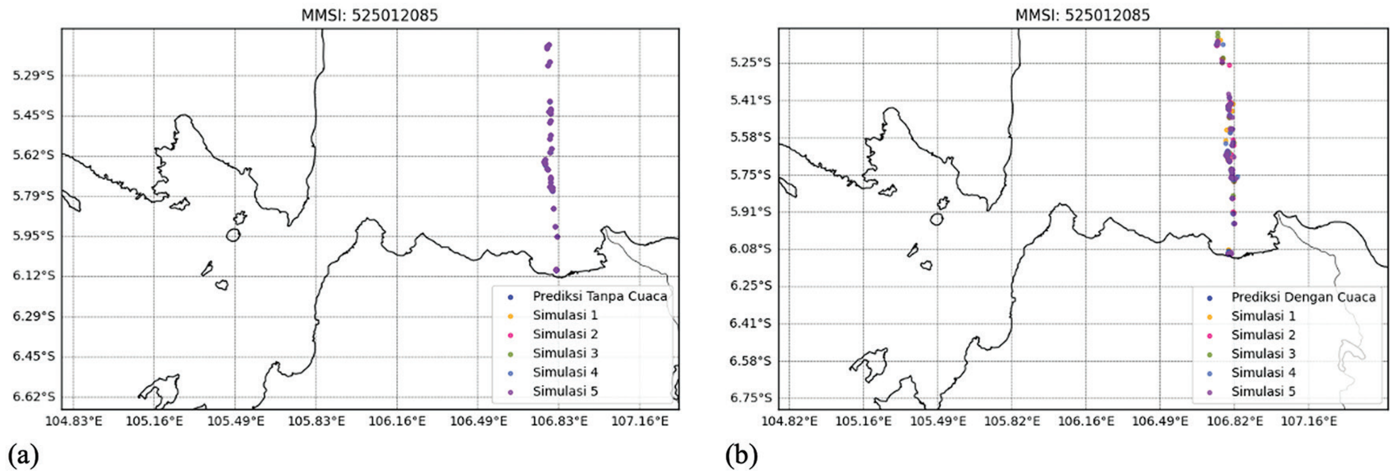


Figure 7. *Comparison of model simulations without weather (a) and with weather (b)*

Table 10*. Chances of ship encounter*

Model	Cluster	Critical distance	Trajectory combination	Chance of encounter
No Weather	1	2,012,405	4,885,683	0.41189
	2	932,639	1,301,751	0.71645
With Weather	1	2,078,429	4,989,908	0.41653
	2	924,579	1,111,275	0.83199

maritime navigation in the Sunda Strait. The model is complemented by DBSCAN clustering for identifying CCZs and MCS for estimating ship encounter probabilities. The findings provide valuable insights into ship characteristics, trajectory prediction accuracy, and encounter risk assessment.

1. The Sunda Strait is a high-density maritime corridor with both domestic and international shipping activities. Bulk carriers and container ships dominate vessel traffic, reflecting the strait's importance in cargo transportation. Flag state analysis reveals that Indonesia and Panama are the most prevalent, indicating a mix of national and global shipping interests. The intersection of domestic routes (Merak-Bakauheni) and international shipping lanes contributes to navigational complexities in the strait.

2. Integrating weather data into the Bi-GRU trajectory prediction model significantly reduces errors in ship position forecasting. Quantitative improvements in prediction accuracy:

a. Latitude error reduction: 50% (from 0.0054° to 0.0027°).

b. Longitude error reduction: 43% (from 0.0107° to 0.0061°).

The use of sequential learning (Bi-GRU) allows the model to capture ship movement patterns more effectively than traditional methods. The enhanced trajectory prediction is crucial for route optimization, encounter avoidance, and fuel efficiency in congested maritime zones.

3. DBSCAN clustering successfully identifies two CCZs in the Sunda Strait, highlighting high-risk navigational areas. Cluster quality improved with weather data integration, as indicated by an increase in the silhouette coefficient from 0.5425 to 0.5489. MCS estimates ship encounter probability:

a. Without weather data:

● Cluster 1: 0.41189

● Cluster 2: 0.71645

b. With weather data:

● Cluster 1: 0.41653

● Cluster 2: 0.83199

The increase in encounter probability after weather data integration suggests that weather conditions significantly impact ship navigation risks. The results support the need for real-time weather data incorporation in maritime traffic management systems to enhance navigational safety.

This study presents a data-driven framework for predicting ship trajectories and encounter risks, offering valuable support for maritime traffic monitoring and risk assessment. By integrating weather-aware clustering, the model enhances the identification of navigational hazards, contributing to safer maritime operations and proactive encounter prevention. The findings highlight the importance of real-time weather data incorporation in maritime navigation systems to optimize route planning and improve vessel safety. Future research should focus on real-time AIS-weather data streaming to enable dynamic risk assessments and adaptive navigation strategies. Additionally, integrating ship maneuverability modeling—considering factors such as braking capabilities and response times—can further enhance the accuracy of encounter risk predictions and strengthen maritime safety measures.

Footnotes

Authorship Contributions

Concept design: I. D. Ratih, and M. R. Habibi, Data Collection or Processing: I. D. Ratih, and K. C. S. Arum, Analysis or Interpretation: I. D. Ratih, M. R. Habibi, and K. C. S. Arum, Literature Review: I. D. Ratih, M. R. Habibi, and K. C. S. Arum, Writing, Reviewing and Editing: I. D. Ratih, M. R. Habibi, and K. C. S. Arum.

Funding: The authors did not receive any financial support for the research, authorship and/or publication of this article.

REFERENCES

[1] K. Spyrou-Sioula, et al. “AIS-enabled weather routing for cargo loss prevention,” *Journal of Marine Science and Engineering,* vol. 10, 1755, Nov 2022.

[2] X. Han, C. Armenakis, and M. Jadidi, “Modeling vessel behaviours by clustering ais data using optimized DBSCAN,” *Sustainability*, vol. 13, 8162, Jul 2021.

[3] S. Vukša, P. Vidan, M. Bukljaš, and S. Pavić, “Research on ship collision probability model based on Monte Carlo simulation and Bi-LSTM,” *Journal of Marine Science and Engineering,* vol. 10, Aug 2022.

[4] D. Inazu, T. Ikeya, T. Iseki, and T. Waseda, “Extracting clearer tsunami currents from shipborne Automatic Identification System data using ship yaw and equation of ship response,” *Earth Planets Space*, vol. 72, 41, Dec. 2020.

[5] Y. Suo, W. Chen, C. Claramunt, and S. Yang, “A Ship Trajectory Prediction Framework Based on a Recurrent Neural Network,” *Sensors*, vol. 20, no. 18, p. 5133, Sep. 2020, doi: 10.3390/s20185133.

[6] W. Wang, W. Xiong, X. Ouyang, and L. Chen, “TPTrans: Vessel Trajectory Prediction Model Based on Transformer Using AIS Data,” *IJGI*, vol. 13, no. 11, p. 400, Nov. 2024, doi: 10.3390/ijgi13110400.

[7] D. Nguyen and R. Fablet, “TrAISformer -- A Transformer Network with Sparse Augmented Data Representation and Cross Entropy Loss for AIS-based Vessel Trajectory Prediction,” 2021, doi: 10.48550/ARXIV.2109.03958.

[8] C. Wang, H. Ren, and H. Li, “Vessel trajectory prediction based on AIS data and bidirectional GRU,” in *Proc. - Int. Conf. Comput. Vis., Image Deep Learn., CVIDL*, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 260–264. doi: 10.1109/CVIDL51233.2020.00-89.

[9] U. Azmi, Z. N. Hadi, and S. Soraya, “ARDL METHOD: Forecasting Data Curah Hujan Harian NTB,” *Varian*, vol. 3, no. 2, pp. 73–82, May 2020, doi: 10.30812/varian.v3i2.627.

[10] A. Daranda and G. Dzemyda, “Navigation decision support: Discover of vessel traffic anomaly according to the historic marine data,” *Int. J. Comput. Commun. Control*, vol. 15, no. 3, 2020, doi: 10.15837/IJCCC.2020.3.3864.

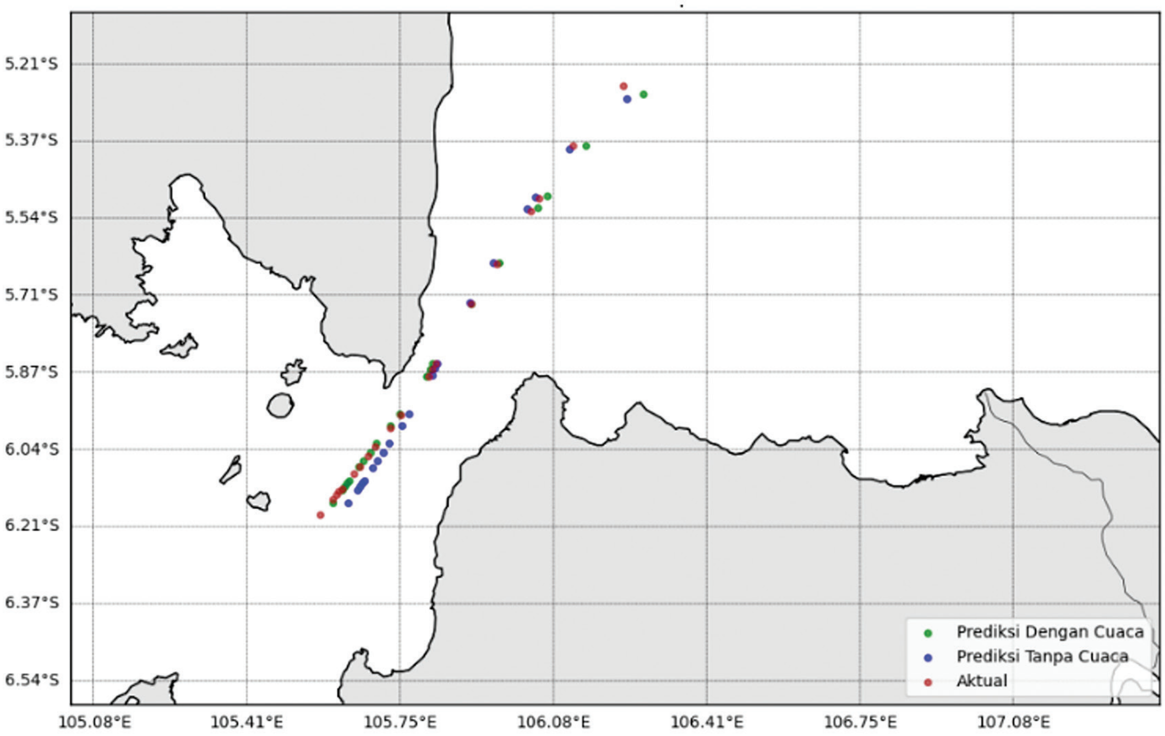
[11] A. C. A. Bima, P. Susanti, and M. Y. Asyhari, “Implementation of Forecasting with the Monte Carlo Simulation Method to Predict Supply and Demand for Psychotropic Drug Products,” *Brilliance*, vol. 3, no. 2, pp. 441–448, Jan. 2024, doi: 10.47709/brilliance.v3i2.3515.

[12] M. Aslam, “A new goodness of fit test in the presence of uncertain parameters,” *Complex Intell. Syst.*, vol. 7, no. 1, pp. 359–365, Feb. 2021, doi: 10.1007/s40747-020-00214-8.

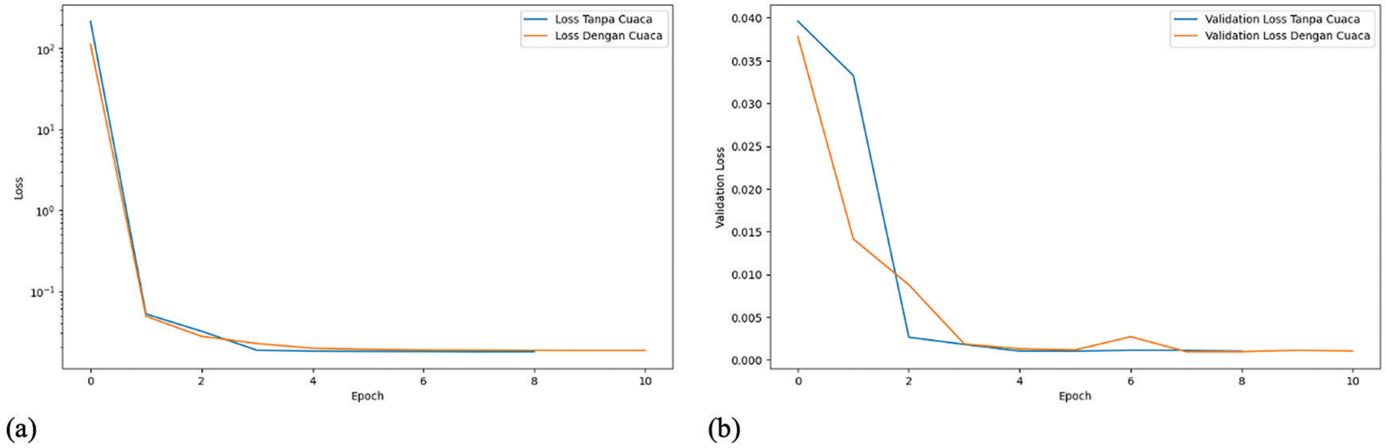
[13] M. Berlinger, S. Kolling, and J. Schneider, “A generalized Anderson–Darling test for the goodness-of-fit evaluation of the fracture strain distribution of acrylic glass,” *Glass Struct Eng*, vol. 6, no. 2, pp. 195–208, Jun. 2021, doi: 10.1007/s40940-021-00149-7.

[14] G. D. Ahadi and N. N. L. E. Zain, “Pemeriksaan Uji Kenormalan dengan Kolmogorov-Smirnov, Anderson-Darling dan Shapiro-Wilk,” *EMJ*, pp. 11–19, Jun. 2023, doi: 10.29303/emj.v6i1.131.

[15] A. Rizaldi, A. Muzwardi, E. Santoso, M. Iffan, and M. Fera, “The strategic development of maritime connectivity in the border area in Indonesia,” *JEECAR*, vol. 10, no. 4, pp. 701–711, Jun. 2023, doi: 10.15549/jeecar.v10i4.1378.



Supplementary Figure 1. *Comparison of Actual Ship Trajectories (Red) with Predicted Trajectories using Weather Factors (Green) and without Weather Factors (Blue)*



Supplementary Figure 2. *Loss (a) and Validation Loss (b) Bi-GRU Model*