Defence Science Journal, Vol. 75, No. 2, March 2025, pp. 250-256, DOI : 10.14429/dsj.20465

Predicting Vessel Speed Using Ensemble MLR and RF Models on AIS Data

Nitish Raj[#,*], Prabhat Kumar[$] and Divyanshi Goyal[^]

[#]*Weapons and Electronics System Engineering Establishment, Delhi - 110 066, India*
[$]*National Institute of Technology, Patna - 801 503, India*
[^]*University School of Automation and Robotics, GGSIPU, Delhi - 110 032, India*
[*]*E-mail: raj.nitp@gmail.com*

ABSTRACT

Defence agencies like the DRDO, Indian Navy, Indian Army, Indian Air Force (IAF), and many other organizations require ongoing oversight of their applications and devices to detect anomalous behaviour. The purpose of this research is to track the ship's speed to facilitate navigation. AIS data plays a substantial role in this investigation, as it provides critical information such as the speed, course, and position of a vessel. Monitoring vessel speed is essential to ensure smooth functioning, enabling collision avoidance, route optimisation, and accurate analysis of the ship's schedule for transportation. In this study, an ensemble approach using Multiple Linear Regression (MLR) and Random Forest (RF) models to predict vessel speed more accurately is developed. The proposed ensemble model outperformed existing methods, showing significant improvements in prediction accuracy and robustness. This enhanced performance also aids in the precise Estimation of the Time of Arrival (ETA) of the vessel, contributing to more efficient operation procedures and environmentally friendly practices. Early oversight of vessel speed ensures maritime navigation and safety, promoting reliable and optimized routes for vessels.

Keywords: MMSI; Multiple linear regression; AIS; Vessel speed prediction

1. INTRODUCTION

The International Maritime Organisation (IMO) is responsible for developing the AIS (Automatic Identification System), a technical standard for the maritime industry. It is a technique for tracking and identifying marine traffic that sends messages via Very High-Frequency (VHF) marine spectrum using transponders[1]. After numerous security regulations came into existence after September 11, 2001, it was observed that the automatic identification system should be fitted with marine vessels[2]. It helps exchange the ship's actual location, speed, and relative direction with other ships. AIS was first employed in Vessel Traffic Service (VTS) and radar augmentation, but it was later found useful in research[3]. AIS data is prioritised above sonar, radar data, or CCTV because it enables path planning, facilitating the avoidance of ship collisions. Additionally, real-time anomaly detection using AIS data allows for the identification of potential security and navigation safety concerns[4]. It is a transponder system, where all the vessels can broadcast their details to each other.

This probe uses machine learning technology to predict a vessel's speed. Machine learning is one of the rising technologies in the field of artificial intelligence. The model is trained on the training dataset to find patterns among the data points. For the speed prediction, an ensemble model with regression and a random forest model are used because speed is a continuous value. AIS data has multiple features; that's why this probe employs the Multiple Linear Regression method.

Despite the advancements in machine learning models for maritime applications, several challenges remain for future researchers. One significant challenge is handling data quality issues in AIS, such as missing or noisy data, which can degrade the model's predictive performance. Additionally, incorporating real-time data streams to enhance prediction accuracy and responsiveness in dynamic maritime environments is still a developing area. Another challenge is the integration of diverse data sources (e.g., weather conditions, ocean currents) to improve the robustness of speed prediction models. Addressing the trade-off between prediction accuracy and computational efficiency is also crucial for deploying these models in real-time operational settings.

2. LITERATURE REVIEW

The prediction of vessel speed using Automatic Identification System (AIS) data has garnered significant attention in recent years, driven by the potential to enhance maritime safety, efficiency, and sustainability. Machine learning approaches have emerged as a dominant paradigm in this field, demonstrating their effectiveness in accurately predicting vessel Speed Over Ground (SOG).

Abebe[5], *et al.* explored various machine learning models, including linear regression, Support Vector Machines, and Artificial Neural Networks, to predict ship speed based on AIS and weather data. Their study emphasized the importance of feature engineering and selection in improving prediction accuracy. Burger[6], *et al.* on the other hand, proposed an extended linear regression model that incorporated a

Received : 06 August 2024, Revised : 25 November 2024
Accepted : 04 December 2024, Online published : 24 March 2025

priori AIS information for vessel trajectory prediction. By leveraging historical AIS data and considering vessel-specific attributes like type, draft, and navigational status, their model achieved improved accuracy compared to traditional linear regression models. This study underscored the significance of incorporating domain-specific knowledge and historical data for enhancing prediction performance.

While most research focused on traditional machine learning techniques, deep learning approaches have also gained traction in vessel speed prediction.

In addition to deep learning, simpler machine learning models have also proven effective in certain scenarios. Moreover, research has extended to vessel speed prediction in inland waterways, where factors like river flow and channel characteristics play a crucial role.

In addition to predicting vessel speed, AIS data has been used for other maritime applications, such as anomaly detection[7] and route optimisation[8]. This highlights the versatility and value of AIS data in the maritime domain and its potential for further research and applications. This study aligns with the research by Raj and Kumar[9], which highlights the efficacy of combining Linear Regression (LR) and Long Short-Term Memory (LSTM) techniques for predicting vessel positions based on AIS data to enhance maritime operations.

The research landscape in vessel speed prediction using AIS data is continuously evolving, with new methodologies and techniques being developed. As the field progresses, it is expected that more sophisticated models will be developed, incorporating a broader range of features and utilizing advanced machine learning algorithms to achieve even higher accuracy and reliability in the prediction of vessel speed.

Rao[10], *et al.* analyse AIS data to predict vessel speeds in the Arctic, specifically without relying on ice condition data. The study employs decision tree-based methods, including Random Forest, XGBoost, and LightGBM, achieving a mean absolute error (MAE) of around 3.5 knots. The authors also highlight that visibility metrics, such as daylight data, and the distance from landmarks contribute significantly to enhancing the predictive capabilities of their models.

Recent advancements in machine learning, including models like Linear Regression, SVM, ANN, and deep learning methods like RNNs, have significantly improved vessel speed prediction using AIS data.

3. METHODOLOGY

In the maritime sector, accurately predicting vessel speed is crucial for efficient route planning, collision avoidance, and maritime traffic management. The availability of vast amounts of data from sophisticated sensory systems installed on vessels, viz the automatic identification system, presents an opportunity to enhance these predictive capabilities. AIS data provides information about the vessel at any instant of time but there may be a chance of transponder system failure which might not be able to transmit the data, so there is a need for some backup or the model which helps to obtain the vessel information if there is any failure or mishap. This study focuses on developing an ensemble model that leverages both Multiple Linear Regression and Random Forest algorithms to achieve this goal.

3.1 PROBLEM FORMULATION

The prediction problem can be mathematically formulated as follows in Eqn. (1):

$$SOG_i = f(x_i) + \in \tag{1}$$

where:

- SOG_i is the speed over the ground of vessel i.
- X_i=(X_{1i}, X_{2i}, X_{3i}+...X_{ki}) represents the vector of predictor variables for vessel i. These variables include latitude, longitude, Course Over Ground (COG), and rate of turn.
- $f(X_i)$ is the mapping function which maps the predictor variables to the values of the predicted speed variable. This study, $f(X_i)$ is represented by an ensemble model combining MLR and RF.
- $\in$ is the term for error capturing the speed variation which is not explained by the model.

The ensemble model is designed to capture both the linear relationships through MLR and the complex, non-linear interactions through RF. The MLR component is formulated in Eqn. (2) as:

$$SOG_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki} + \in_i \tag{2}$$

where, α_0 is the intercept, and (α_1, $\alpha_{2, \ldots}$, α_k) are the coefficients corresponding to the predictor variables.

The RF component builds numerous decisions trees and averages their outputs to improve prediction accuracy. The importance of each predictor variable is assessed through the feature Gini scores derived from the RF model, which indicates the contribution of each feature to the prediction accuracy. By integrating the linear interpretability of MLR with the robust, non-linear predictive power of RF, the ensemble model aims to provide accurate and reliable speed predictions. This approach not only predicts vessel speed with high accuracy but also offers insights into the contributions of different features, ultimately improving maritime operations, enhancing traffic management, and promoting eco-friendly practices through informed decision-making.

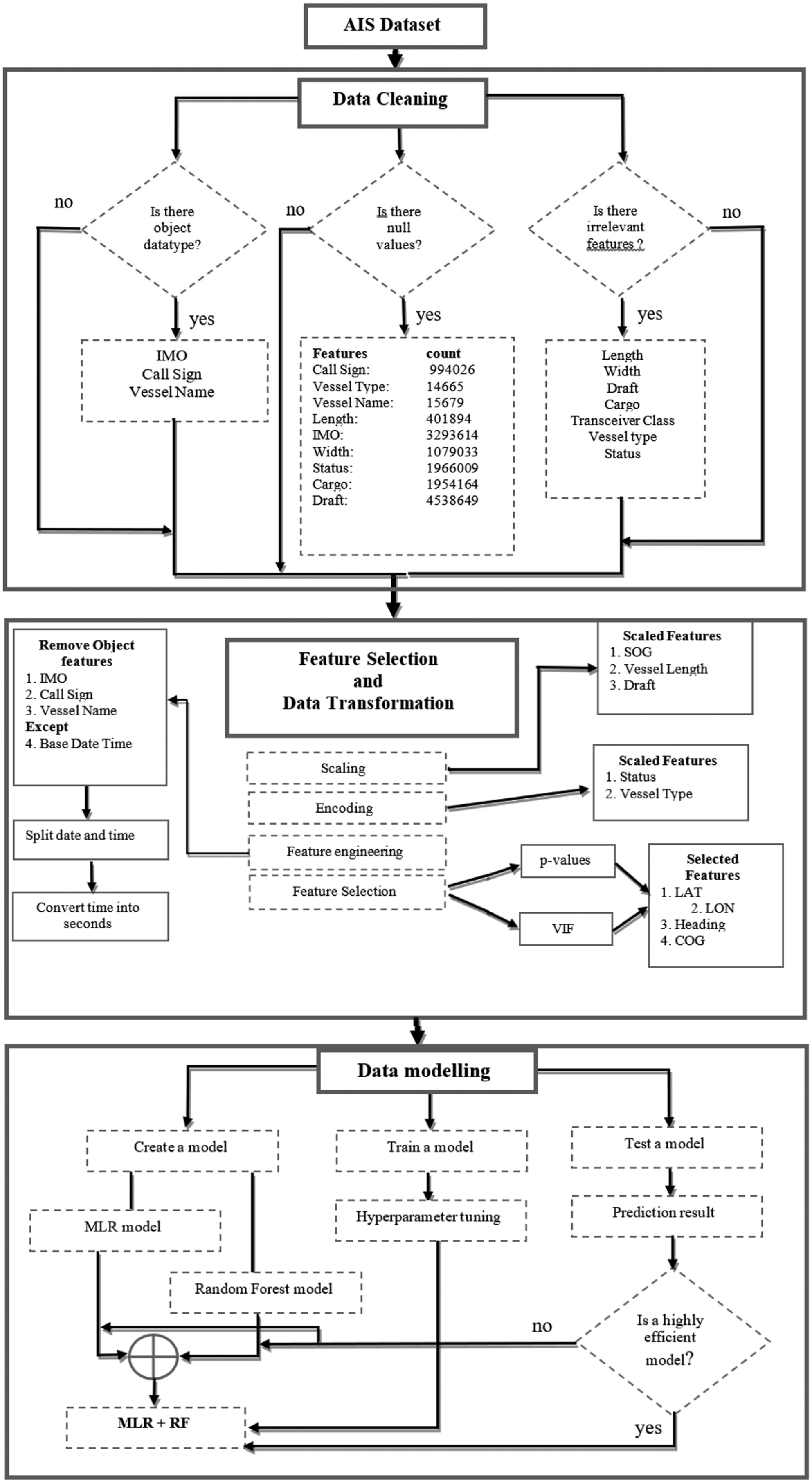
3.2 Data Gathering and Preprocessing

3.2.1 Data Source

The main dataset for this research was obtained from the Marine Cadastre website[11,] the data files are Automatic Identification System repository hosted on the site. AIS is a system that automatically tracks and exchanges comprehensive information on vessels, such as their identification, geographic position, COG, timestamp, navigational status and SOG. This data is essential for VTS and serves as a comprehensive basis for maritime research and predictive modelling.

3.2.2 Data Collection

For this study, a subset of historical AIS data was extracted from the Marine Cadastre repository. The selection criteria focused on a specific geographical region and period relevant to the research objectives. The data was available in various formats, including CSV, JSON, and XML, and was downloaded in CSV format for ease of processing and analysis. Each day's data was stored in separate CSV files. For this investigation, the data from December 31, 2023, was utilised. The raw dataset initially comprised approximately 7.3 million



Features	count
Call Sign:	994026
Vessel Type:	14665
Vessel Name:	15679
Length:	401894
IMO:	3293614
Width:	1079033
Status:	1966009
Cargo:	1954164
Draft:	4538649

Figure 1. Framework for establishing the model.

records, which was later reduced to 900,000 records containing 2,551 unique MMSI (Maritime Mobile Service Identification) numbers.

3.2.3 Data Pre-Processing

Pre-processing of the raw AIS data involved several critical steps to ensure data quality and suitability for modelling. Figure 1 clearly states the steps, and the methodology adopted in this research.

3.2.3.1 Data Cleaning

The upper Block of Fig.1 shows the features of the AIS dataset that satisfies the condition of missing, duplicate and irrelevant values.

3.2.3.1.1 Handling Missing Values

Records with missing values in essential fields such as vessel identification, position, SOG, or timestamp were discarded to prevent inconsistencies in the analysis.

3.2.3.1.2 Outlier Detection and Removal

Outliers in the SOG were identified using statistical methods, such as the interquartile range (IQR) method. These outliers were then removed to ensure the accuracy of the model.

3.2.3.1.3 Duplicate Removal

Duplicate records, which can arise from multiple transmissions of the same data point, were removed to maintain data integrity.

3.2.3.1.4 Filtering Irrelevant Data

Records outside the specified geographical region or time were filtered out to focus the analysis on the relevant subset of data.

3.2.3.1.5 Null Value Handling

Entries where the length and breadth of the ship were recorded as zero were removed, as such values are not feasible.

3.2.3.2 Feature Selection and Engineering

The middle block of Fig.1 all the attributes contributing to the final research.

3.2.3.2.1 Selection of Relevant Features

Key features relevant to vessel speed prediction, including vessel type, length, draft, navigational status, and environmental conditions, were selected for the analysis. Features with object data types, except for the Base Date Time, were removed.

3.2.3.2.2 Feature Engineering

New features were derived from the existing data to enhance the model's predictive power. These included temporal features such as time of day, day of the week, and voyage phase (e.g., underway, anchored, moored). The BaseDateTime was split into separate date and time features, with the time converted into seconds for numerical analysis.

3.2.3.2.3 P-value Calculation

The significance of each feature was determined using p-value analysis, with a threshold of 0.05 to reject the null hypothesis, ensuring only relevant features were retained as shown in Table 1.

3.2.3.2.4 Variance Inflation Factor (VIF) Analysis

VIF was calculated to assess multicollinearity among features. Attributes with VIF values greater than 5 were considered to have high multicollinearity and were excluded from the analysis.

Table 1. P-values for predictor variables (MLR model)

Features	P_value	VIF
LAT	0.0	1.22
LON	0.02	1.15
COG	0.01	2.19
Heading	0.002	6.22
Status	0.4	11.7
Length	0.01	7.89
Width	0.02	8.58
Draft	0.31	0.03
Rate of turn	0.01	1.34

3.2.3.3 Data Transformation

The middle block of Fig.1 depicts the transformation steps of the AIS data pre-processing.

3.2.3.3.1 Scaling Numerical Features

Numerical features, such as vessel length and draft, were scaled to ensure uniform ranges, facilitating the convergence of the machine learning algorithms.

3.2.3.3.2 Encoding Categorical Features

The categorical data, such as vessel type and navigational status, were transformed using methods like one-hot encoding and label encoding to turn them into a format that is compatible with machine learning models.

3.2.3.4 Data Splitting

The Lower Block of Fig.1 depicts the representation of the AIS dataset using a model.

3.2.3.4.1 Training and Testing Sets

The dataset that had undergone preprocessing was split into two subsets for training and testing. The training set was utilised for the development of the prediction model, whereas the testing set was set aside for model validation to evaluate performance measures such as mean absolute error, the R-squared value and mean squared error.

The rigorous preprocessing pipeline ensured a high-quality dataset, which was essential for training and evaluating multiple linear regression models aimed at predicting vessel speed. By adhering to these established preprocessing practices, the study aimed to produce robust and accurate predictive models, thereby contributing valuable insights to the field of maritime research.

3.2 Model Training and Evaluation

After preprocessing the AIS data, the next step involves training the ensemble Multiple Linear Regression and Random Forest (MLR+RF) model for predicting vessel speed. The training process involves fitting both the MLR and RF models to the training data and tuning their parameters to achieve the best performance.

3.2.1 Establishing the MLR+RF Model

The ensemble model is trained using the training subset of the pre-processed data. Here's how each component of the ensemble model is developed and their contributions to the final model:

3.2.1.1 Multiple Linear Regression Model

The MLR model's coefficients ($\alpha_1,\alpha_2,...,\alpha_k$) are estimated by minimizing the sum of squared errors between the observed Speeds Over Ground (SOG) and predicted SOG values. This optimization process ensures that the MLR model accurately captures the relationship between the predictor variables and the vessel's speed over the ground. The MLR model is trained using the following steps:

3.2.1.1.1 Data Splitting

The ratio of pre-processed data splitting for training and testing is 80:20.

3.2.1.1.2 Model Fitting

The MLR model is fitted to the training data, estimating the coefficients for each predictor variable.

3.2.1.1.3 Coefficient Analysis

The coefficients (α values) are analysed to understand the participation of each feature in the model's predictions in Table 2.

Table 2. Estimated Coefficients for Predictor Variables (MLR Model)

Feature	Coefficient (β)
Intercept (α_0)	2.45
Latitude	0.35
Longitude	0.15
COG	0.20
Rate of turn	0.25

The coefficients reveal the strength and direction of the link between each predictor variable and vessel speed.

Table 2 indicates latitude and longitude have positive coefficients, indicating that an increase in these features is associated with an increase in vessel speed. Conversely, COG and rate of turn have different directional impacts on vessel speed.

3.2.1.2 Random Forest Model

Random Forest (RF) model, as part of the ensemble approach, complements the MLR model by leveraging its ensemble of decision trees to improve prediction accuracy. During training, the RF model creates several decision trees and outputs the average forecast of each tree.

This method helps in reducing overfitting and increasing the model's robustness. The RF model is trained using the following steps:

3.2.1.2.1 Data Splitting

The pre-processed data is split into training and testing subsets of 80:20.

3.2.1.2.2 Model Fitting

The RF model is fitted to the training data, constructing multiple decision trees.

3.2.1.2.3 Feature Importance Analysis

The importance of each feature is assessed through feature Gini scores, which indicate how much each predictor variable contributes to the prediction accuracy and the same is indicated in Table 3.

Table 3. Feature gini scores for predictor variables (RF model)

Feature	Gini score
Latitude	0.30
Longitude	0.25
COG	0.20
Rate of turn	0.25

The attribute is based on the improvement in the splitting criterion that each feature provides for every tree in the forest, and Gini scores are computed. These scores aid in determining how each feature contributes to the predictions made by the model.

3.2.1.3 Hyperparameters Tuning

Hyperparameter tuning is an essential aspect of model training. For the MLR model, hyperparameters such as the regularization parameter (if regularization is applied) are tuned to prevent overfitting and improve model performance. Hyperparameters for the RF model are optimized, including the number of trees, tree depth, and the least number of samples needed to split a node. The hyperparameters are routinely varied using methods like cross-validation to determine which values produce the best results. The choice of hyperparameters for the MLR and RF models was made after extensive experimentation and evaluation to ensure optimal model performance. For the RF model, the number of trees was set to 200, providing a balance between computational efficiency and prediction accuracy. A maximum depth of 20 was chosen to allow the model to learn intricate patterns without overfitting. The minimum samples required to split a node were set to 10, helping to prevent over-complexity in the model's structure. Additionally, the Gini criterion was used for evaluating splits, as it is well-suited for classification tasks in this context. The sqrt selection for maximum features ensures that only a subset of features is considered at each split, promoting diversity among the trees in the ensemble. Bootstrap sampling was enabled to further enhance the robustness of the model by training on different subsets of the data. These hyperparameters were systematically tuned through cross-validation to achieve a balance between bias and variance, ultimately leading to a more generalized model for vessel speed prediction.

Table 4. Optimised hyperparameters for RF model

Hyperparameter	Value
Number of trees	200
Maximum depth	20
Minimum samples split	10
Criterion	Gini
Maximum features	Sqrt
Bootstrap	True

Table 4 shown below is the tuned value used in the proposed RF segment of the ensemble model.

By carefully training and evaluating the ensemble MLR+RF model, and performing hyperparameter tuning, the study aims to develop a robust and accurate model for predicting vessel speed based on AIS data. This model can provide valuable insights for maritime operations and contribute to the optimization of vessel traffic management and eco-friendly practices. Combining the strengths of both MLR and RF models, the ensemble approach ensures higher prediction.

5. RESULTS AND DISCUSSION

An assessment was performed on the ensemble Multiple Linear Regression (MLR) and Random Forest (RF) models to predict vessel speed using the testing dataset. The model's accuracy was assessed using the mean absolute error, R-squared value (R^2) and mean squared error, as performance metrics. The results of these evaluations are presented in Table 5.

Table 5. Summary of model accuracy

Model	MAE	MSE	R^2
MLR+RF	0.60	1.00	0.89
DTR	0.90	1.50	0.70
RFR	0.70	1.00	0.80
SVR	0.75	1.10	0.78

Refer to Table 3, According to the MLR+RF model, the average absolute difference between the anticipated and actual vessel speeds is 0.60 knots. Additionally, the model has a Mean Absolute Error (MAE) of 0.60. The Mean Squared Error (MSE) value of 1.00 signifies that the average squared difference between the observed and predicted speeds of vessels is 1.00 square knots. The independent variables in the model account for 89% of the variability in vessel speed, as indicated by the R-squared value of 0.89. The performance metrics of the Decision Tree Regression (DTR) and Support Vector Regression (SVR) models were worse than those of the MLR+RF and Random Forest Regression (RFR) models.

6. CONCLUSION

The findings suggest that the MLR+RF ensemble model demonstrates a strong ability to accurately estimate vessel speed using the chosen features. The Random Forest component of the model significantly enhances the predictive accuracy and explanatory power, as evidenced by the lower MAE and higher R-squared value compared to the individual models.

The feature importance analysis reveals that vessel characteristics and navigational status are significant predictors of vessel speed. Environmental conditions like wind speed and wave height also play a role, albeit to a lesser extent. Temporal features such as time of day and day of the week have smaller but notable impacts on speed predictions.

Overall, the developed MLR+RF ensemble model serves as a valuable tool for predicting vessel speed and can aid in optimizing maritime operations, enhancing traffic management, and promoting eco-friendly practices. The superior performance of the ensemble model suggests that integrating more advanced machine learning techniques can further improve prediction accuracy and reliability. Hyperparameter tuning and feature selection optimization could also contribute to enhancing model performance.

REFERENCES

1. Svanberg, M.; Santén, V.; Hörteborn, A.; Holm, H. & Finnsgård, C. AIS in maritime research. *Marine Policy*, 2019, **106**, 103520.
   doi: 10.1016/j.marpol.2019.103520.
2. Iphar, C.; Napoli, A.; Ray, C.; Alincourt, E. & Brosset, D. Risk analysis of falsified automatic identification system for the improvement of maritime security. *In* ESREL 2016, Taylor & Francis, 2016. pp. 606-613.
3. Adriel Kornelius, Y.; K Muyeba, M. & Leslie Hendric Spits Warnars, H. Ship movement analysis based on Automatic Identification System (AIS) data using convolutional neural network and multiple thread processing. *Int. J. Comput. Digital Syst.*, 2024, **16**(1), 1-10.
   doi:10.12785/ijcds/1601107.
4. Perez, H.M.; Chang, R.; Billings, R. & Kosub, T.L. Automatic Identification Systems (AIS) data use in marine vessel emission estimation. *In* 18th Annual International Emission Inventory Conference, U.S. Environmental Protection Agency, 2009.
5. Abebe, M.; Shin, Y.; Noh, Y.; Lee, S. & Lee, I. Machine learning approaches for ship speed prediction towards energy efficient shipping. *Appl. Sci.*, 2020, **10**(7), 2325.
   doi:10.3390/ app10072325.
6. Burger, C.N.; Kleynhans, W. & Grobler, T.L. Extended linear regression model for vessel trajectory prediction with a-priori AIS information. *Geo-spatial Info. Sc.*, 2022, **27**(1), 202-220.
   doi:10.1080/10095020. 2022.2072241.
7. Handayani, M.P.; Antariksa, G. & Lee, J. Anomaly detection in vessel sensors data with unsupervised learning technique. *In* International Conference on Electronics, Information, and Communication (ICEIC), Jeju, Korea (South), 2021, pp. 1-6,
   doi: 10.1109/ICEIC51217.2021.9369822.
8. Gao, D.W.; Zhu, Y.S.; Zhang, J.F.; He, Y.K.; Yan, K. & Yan, B.R. A novel MP-LSTM method for ship trajectory prediction based on AIS data. *Ocean Eng.*, 2021, **228**, 108956.
   doi:10.1016/j.oceaneng.2021.108956.
9. Raj, N. & Kumar, P. A novel & efficient LR LSTM AIS route data prediction for longer range. *Def. Sc J.,* 2024, **74**(4), 583-591.
   doi:10.14429/dsj.74.19336
10. Rao, P.S.; Kim, E.; Smestad, B.B.; Asbjørnslett, B.E. &

Bhattacharyya, A. Predicting vessel speed in the Arctic without knowing ice conditions using AIS data and decision trees. *Maritime Transport Res.*, 2021, **2**, 100024. doi: 10.1016/j.martra. 2021.100024.

11. Vessel traffic data. Bureau of Ocean Energy Management (BOEM) and National Oceanic and Atmospheric Administration (NOAA). https://marinecadastre.gov/ais/. [Accessed 23 May 2024].

CONTRIBUTORS

Mr Nitish Raj obtained his M.Tech (CSE) from IIT, Delhi and working as a Scientist at DRDO, posted at the Weapons and Electronics Systems Engineering Establishment, Ministry of Defence in New Delhi. In his current role, he holds the position of Senior Systems Manager of Naval Combat Systems. His research interests encompass system design & development, systems integration, and Machine Learning.
He contributed to the current work by producing the idea and designing the experiment, optimising the deep learning techniques used in the experiment, creating the programme, analysing the data, and finalising the manuscript.

Dr Prabhat Kumar obtained his PhD in Computer Science and working as a Professor in the Computer Science and Engineering Department at NIT Patna, India. He oversees IT Services and chairs the Computer and IT Purchase Committee at NIT Patna. His research focuses on wireless sensor networks, internet of things, cyber security, data science, software engineering, and e-governance.
He made contributions to the current study by assisting in the conceptualisation of the review, helping in the identification and contributing to the analysis and synthesis of findings.

Ms Divyanshi Goyal is currently a final-year Industrial Internet of Things (IIOT) student at the University School of Automation and Robotics GGSIPU, Shahdara, New Delhi.
In the current study she played a significant role in the acquisition and preprocessing of AIS data.