Defence Science Journal, Vol. 74, No. 4, July 2024, pp. 583-591, DOI : 10.14429/dsj.74.19336

A Novel and Efficient LR-LSTM AIS Route Data Prediction for Longer Range

Nitish Raj[#,*] and Prabhat Kumar[$]

[#]*Weapon and Electronics System Engineering Establishment, New Delhi - 110 066, India*

[$]*National Institute of Technology, Patna - 800 001, India*

[*]*E-mail: raj.nitp@gmail.com*

ABSTRACT

The growth of technology has enabled different industries to generate an excessive amount of data- one such industry being the maritime sector. Sophisticated sensory systems installed on various vessels generate information at a very large scale which can further be used in optimizing operational efficiency, improving safety standards, and aiding in the decision-making process. Researchers have henceforth identified statistical learning methods and machine learning techniques as potent tools for excavating useful insights from this copious amount of data available. This research evaluates how these algorithms work by focusing exclusively on the analysis of sensory data collected from vessels within the maritime domain. A comparison study has been conducted between statistical learning methods (which includes regression analysis, and time series analysis) vis-a-vis machine learning approaches. The major objective of this study was to determine the most effective method for detecting anomalies while simplifying marine operations and optimizing vessel behavior. The scope of the conducted analysis is restricted to the prediction of the next trajectory points. Accurate prediction of vessel positions plays a crucial role in maritime operations, enabling efficient route planning, collision avoidance, and maritime traffic management. In this article, the authors propose a combination model that combines the benefits of Linear Regression (LR) and Long Short-Term Memory (LSTM) techniques to anticipate vessel positions based on Automatic Identification System (AIS) data. The proposed model takes advantage of the interpretability of LR and the temporal dependencies collected by LSTM to capture temporal dependencies, which improve prediction accuracy and reveal the underlying links between vessel features and future positions.

Keywords: Machine learning; Ship trajectory prediction; Neural network; AIS

1. INTRODUCTION

The maritime domain is essential to international trade, logistics, and transportation. Advanced sensing systems onboard ships produce a large amount of data about navigation, performance, the environment, and other operational factors. Insights regarding vessel behavior, operational effectiveness, safety, and decision-making in the marine sector may be gained from the analysis of this sensory data.

For data analysis and prediction, conventional statistical learning approaches are frequently utilized in a variety of industries. Regression analysis, time series analysis, and clustering are three techniques that have proved successful in removing significant patterns and relationships from structured data. Machine learning techniques, on the other hand, have become increasingly popular in the extraction of knowledge from complicated and unstructured data as a result of recent improvements in computer power and the accessibility of enormous amounts of data.

Analyzing sensory data from ships using machine learning algorithms is becoming increasingly popular in the marine sector. Finding hidden patterns, spotting abnormalities, and optimising various elements of marine operations are all achievable with the aid of machine learning models, such as models for deep learning, supervised learning, and unsupervised learning. By utilizing these cutting-edge algorithms, marine stakeholders may improve safety protocols, reduce downtime, and increase fuel efficiency.

This research paper’s goal is to investigate the uses and carry out a comparison study to evaluate the efficacy of statistical land machine-learning methodologies in the maritime sector, with an emphasis on sensory data from vessels.

The ability to harness the potential of data-driven decision-making and create effective systems that improve ship operations and maritime safety depends on maritime practitioners being aware of the advantages and disadvantages of statistical learning and machine learning methodologies. This project intends to contribute to the continuing initiatives in the marine industry to harness data analytics for enhanced operational outcomes by bridging the gap between conventional statistical approaches and cutting-edge machine-learning techniques. Traditional LR models have been widely used for vessel position prediction by leveraging spatial and temporal features. However, these models often struggle to capture complex temporal dependencies present in AIS data. To address this limitation, the authors propose a combined LR-

Received : 10 August 2023, Revised : 01 December 2023
Accepted : 05 December 2023, Online published : 24 July 2024

LSTM based model approach that integrates the benefits of both techniques.

1.1 Principal Aims of Research

This study paper's primary objective is to examine and contrast the marine applications of statistical and machine learning approaches, with an emphasis on vessel sensory data mostly from the AIS dataset, and to come to a conclusion with better prediction outcomes. These are the main aims of this research project, which it has achieved:

1.1.1 To Assess the Effectiveness of Statistical Learning Methods

This objective aims to evaluate the performance and suitability of statistical learning techniques, such as regression analysis and time series analysis in analyzing and predicting vessel behavior based on sensory data. By applying these methods, the research aims to determine their ability to uncover patterns, relationships, and trends in the maritime context.

1.1.2 Evaluation of Efficacy of Machine Learning Algorithms

This objective focuses on assessing the performance and suitability of various machine learning methods, such as deep learning, unsupervised learning, and supervised learning, in analysing a vessel's sensory data. The research aims to explore the capabilities of these algorithms in tasks such as vessel behavior prediction, anomaly detection, and optimization in maritime operations.

1.1.3 Comparison of Statistical and Machine Learning Approaches

This objective aims to provide a comparative analysis of statistical learning methods and machine learning algorithms in the maritime domain. By evaluating their performance, strengths, and limitations, the research aims to highlight which approach is more suitable for different types of analysis tasks in the maritime industry.

1.1.4 To Provide Recommendations for Integration

Based on the research findings, this objective aims to provide practical recommendations for integrating statistical and machine learning approaches into the maritime industry. These recommendations will include guidelines for selecting appropriate techniques, data pre-processing steps, model training, and validation procedures, and considerations for the implementation of data-driven decision-making systems in the maritime domain.

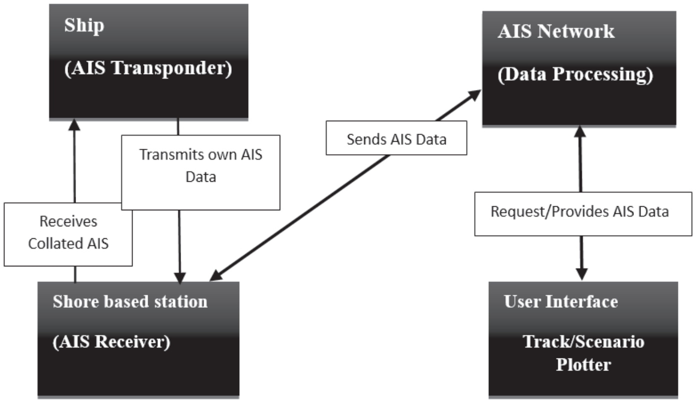

Figure 1. AIS communication architecture.

By achieving these objectives, this research paper aims to contribute to the existing knowledge base by evaluating the comparative effectiveness of statistical and machine learning methods in the analysis of vessels' sensory data. The outcomes of this study can aid maritime stakeholders, including ship operators, port authorities, and maritime agencies, to improve operational efficiency, safety, and decision-making in the maritime industry.

1.2 Glimpse of Important Sensory System Fitted on Board Ship/Vessels: Automatic Identification System (AIS)

The Automatic Identification System (AIS) is a technology used in the maritime domain for real-time vessel tracking, collision avoidance, and enhancing situational awareness. It provides a means for vessels to exchange vital information, allowing for efficient and safe navigation on the seas. AIS aims to improve maritime safety by enabling vessels to electronically exchange essential data, such as vessel identification, position, course, speed, and navigational status as depicted in communication architecture Fig. 1. This information enhances situational awareness, facilitates collision avoidance, and enables efficient traffic management.

1.2.1 Transponder Installation

AIS functionality is embedded within AIS transponder, a device installed on vessels. These transponders continuously collect and transmit vessel-specific information, making it readily available to nearby vessels and shore-based stations as shown in Fig. 2.

1.2.2 VHF Data Link

AIS transponders utilize a Very High Frequency (VHF) data link to broadcast their information over radio waves. This communication link allows vessels to exchange data with other vessels and shore-based AIS stations within range.

1.2.3 Data Collection and Validation

AIS data collected from vessels undergoes a rigorous process of validation and verification to ensure accuracy and reliability. Various parameters, including vessel identification, position, and speed, are carefully scrutinized to minimize errors ordiscrepancies.

1.2.4 Shore-Based Stations

Shore-based AIS stations act as receiving points, intercepting AIS transmissions from nearby vessels. These stations process the received data and forward it to the AIS network for further analysis and dissemination.

1.2.5 AIS Network

The AIS network serves as a centralized system that collects, aggregates, and stores AIS data received from shore-

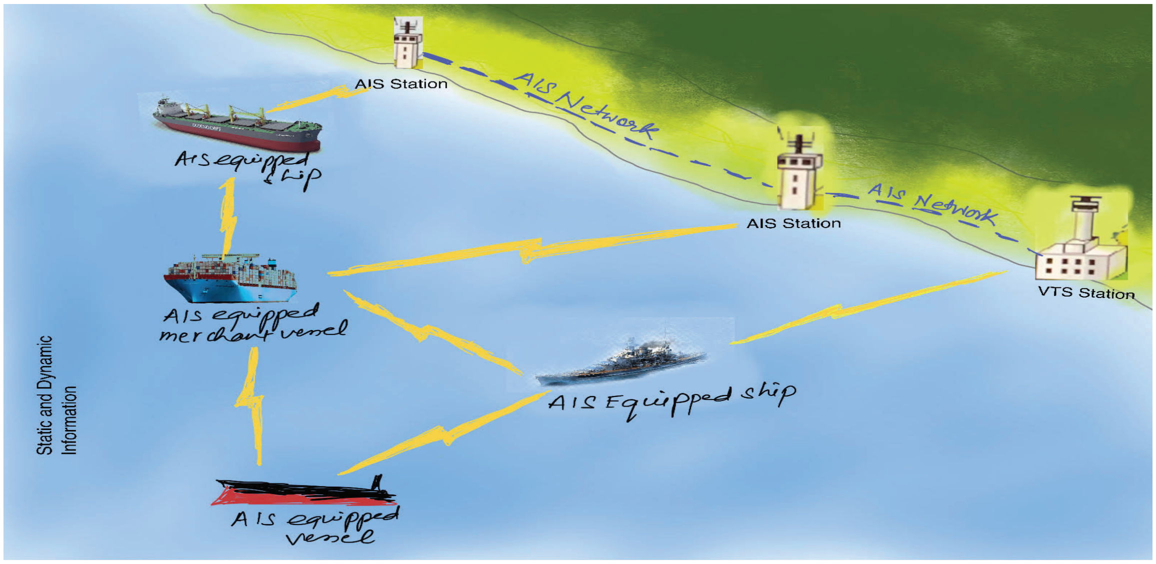

Figure 2. Link exchange capabilities of AIS.

based stations. It acts as a hub for processing and distributing the information to authorized users.

Ships and AIS stations communicate with each other after establishing links for communication using some dedicated channels. The same communication capabilities are described in Reference 2.

1.2.6 Collision Avoidance and Situational Awareness

AIS plays a crucial role in collision avoidance by enabling vessels to track nearby ships, monitor their movements, and identify potential collision risks. It enhances situational awareness by providing real-time information on vessel positions and intentions, facilitating informed navigation decisions.

1.2.7 Regulatory Compliance

AIS is a mandatory system for certain vessels, including commercial ships over a certain size, passenger vessels, and certain types of fishing vessels. Compliance with AIS regulations ensures transparency and contributes to overall maritime safety.

AIS datasets can be obtained from both commercial sources, such as Marine Traffic and VT Explorer, as well as non-commercial sources, including MarineCadastre[1] and Sailwx. It is crucial to ensure data quality because it has a direct impact on how well data mining and machine learning algorithms function. Recently, Tu[1], *et al.* conducted an assessment to evaluate the quality of AIS data, focusing on four key aspects: position precision, timestamp resolution, data completeness, and identification of erroneous records. According to their evaluation, among the non-commercial service providers, the AIS dataset provided by MarineCadastre.gov stood out as the most comprehensive and complete. As a result, for this research, historical AIS data has been obtained from MarineCadastre.gov due to its reliability and richness of information.

1.3 Prediction Problem of Vessel Trajectory

To generate future geographical coordinates (i.e., latitude (LAT) and longitude (LON)) of a vessel's trajectory by using a mathematical approach, the problem is defined as follows:

Let D be a set of historical ship data points, where each data point consists of a timestamp t_i and corresponding latitude LAT_i and longitude LON_i coordinates. The goal is to predict the future latitude and longitude coordinates of a ship at a given future time stamp t_{future}. The formulated problem is a mathematical model that takes the historical ship data D as input and generates the predicted latitude and longitude coordinates LAT_{future} LON_{future} as output at the future time stamp t_{future}. The problem can be modeled mathematically as:

$$(LAT_{future}, LON_{future}) = f(D, t_{future})$$

where, f is the predictive function that converts predicted latitude and longitude from the ship's previous data D and the future timestamp. Depending on the machine learning or mathematical modelling technique selected, the precise form of the prediction function f may change. Depending on the complexity and peculiarities of the ship data as well as the required level of prediction accuracy, it may entail a variety of techniques, including regression models, time series analysis, or deep learning models. The study utilized the Shift model and the Multiple Linear Regressor, LSTM, as baseline predictive models. The LSTM+RWA model was introduced to achieve faster and more accurate results.

2 RELATED WORK

Statistical learning techniques use statistical tools to evaluate data, spot trends, and anticipate the future. These methods include regression analysis, time series analysis, clustering, and hypothesis testing[3]. Regression analysis allows for the modelling and prediction of relationships between variables, while time series analysis focuses on analyzing sequential data to identify trends and patterns over time. Clustering techniques group similar data points together based on their characteristics or behavior[4].

Algorithms are used in machine learning approaches so that computers may learn from data and make predictions or judgments without having to explicitly program them. In contrast to supervised learning algorithms, which use labeled data to train models for prediction tasks, unsupervised learning algorithms identify patterns and structures in unlabelled data. Using neural networks with numerous layers, deep learning, a subtype of machine learning, extracts complex features and produces high-level abstractions[5].

Numerous uses for statistical learning techniques can be explored in the maritime sector. Based on sensory data, regression analysis has been used to model and predict vessel behavior, including speed, fuel consumption, and emissions[6]. Techniques for time series analysis have been used to predict equipment failures, identify trends in sea conditions, and find abnormalities in vessel performance[7]. Fleet management and optimization have benefited from the use of clustering algorithms to categorize ships according to their features[8].

The maritime sector has seen intriguing applications for machine learning technology. Supervised learning algorithms have been used to predict vessel behavior using historical data, including arrival time, course, and fuel use. The detection of anomalous behavior or occurrences in ship operations, such as engine failures or security breaches, has been done using unsupervised learning algorithms[9]. To automatically distinguish nautical objects like ships, buoys, or navigational markers, deep learning techniques have been employed to extract attributes from sensory input[10].

"Numerous research studies have looked into the analysis of vessel sensory data in the marine environment. Researchers have developed regression models to anticipate vessel speed and fuel consumption based on sensory data to enhance trip planning and fuel economy. Time series analysis has been used to identify behavioural anomalies, such as sudden shifts in a vessel's performance or deviations from predictable patterns.". Machine learning algorithms have also been used to categorize different types of vessels, find marine abnormalities, and improve route planning and cargo distribution[11].

In general, the body of the potential of statistical and machine learning methods. In analyzing vessels' sensory data in the maritime domain. However, there is still a need for further research to compare the effectiveness of different methods, address challenges related to data quality and scalability, and develop practical implementations that can be readily adopted by the maritime industry.

3 METHODOLOGY

3.1 Information Gathering and Pre-Processing

Data collection and pre-processing are the first stages of the methodology used in this investigation. The information used in this study came from sources[12] including public databases.

To ensure the accuracy and reliability of the data that was gathered, a thorough process of validation and verification was carried out. Once the data was acquired, it underwent a meticulous pre-processing phase.

This involved cleaning and organizing the data to eliminate any inconsistencies, errors, or missing values. Data cleaning techniques, such as outlier detection and removal, were applied to enhance the quality and integrity of the dataset.

Algorithm 1: AttributeReduction

```
Require: ais data: AIS dataset
Ensure: ais data: AIS dataset with selected columns
ais data ← ais data[['MMSI', 'BaseDateTime', 'LAT', 'LON'
                    'SOG','COG','Heading','Length ',  'Width ']]
return ais data
```

Algorithm 2: BaselineAISCleaning

```
Require: ais data: AIS dataset using AttributeReduction
Ensure: ais data: Cleaned AIS dataset
ais data ← remove duplicate entries from ais data
ais data ← remove rows with missing values from ais data
ais data$^{'LAT'}$ ← convert ais data$^{'LAT'}$ to numeric format
ais data$^{'LON'}$ ← convert ais data$^{'LON'}$ to numeric format
ais data ← filter out erroneous values or outliers
ais data ← ais data$^{(ais\ data\ [\ 'LAT'}$ ≥ −90) & (ais data$^{'LAT'}$ ≤
  90)]
ais data ← ais data$^{(ais\ data\ [\ 'LON'}$ ≥ −180) & (ais data$^{'LON'}$ ≤
  180)]
Additional cleaning steps...
...
return ais data
```

Algorithm 3: AdvanceAISCleaning

```
Require: df : AIS dataset using BaselineAISCleaning
Ensure: df : Cleaned AIS dataset
df ← validate Speed Over Ground (SOG) and Course Over
Ground(COG) values
df ← df$^{(df\ ['SOG'}$ ≥ 0)&(df$^{'SOG'}$ ≤ 30)]
df ← df$^{(df\ ['COG'}$ ≥ 0)&(df$^{'COG'}$ ≤ 360)]
df ← validate Heading values
df ← df$^{(df\ ['Heading'}$ ≥ 0)&(df$^{'Heading'}$ ≤ 360)]
Validate vessel names and call signs
df ← df where vessel names and call signs are not missing
Validate vessel types
Cross-reference with AIS standard or relevant documentation to
check validity
Validate status values
Remove or flag records with unrecognized or unexpected status
values
Validate vessel dimensions
df ← df$^{(df\ ['Length'}$ ≥ 0)&(df$^{'Width'}$ ≥ 0)]
Validate cargo information
Cross-reference with relevant cargo code list or documentation to
check validity
Validate transceiver class
Remove or flag records with invalid or unexpected transceiver
classvalues
return df
```

The Attribute Reduction algorithm as described in Algorithm 1 is designed to reduce the number of features from the dataset. The baseline AIS cleaning algorithm is also designed to pre-process and remove erroneous data from thedataset. using the designed algorithm described in Algorithm 2.

Furthermore, an advanced AIS cleaning algorithm as described in Algorithm 3 is written for pre-processing and cleaning of data from the dataset. This includes the validity of SOG and COG values. Usually, a vessel of tanker type attains speed in the range of 0 to 30 knots and the valid range of course over ground is 0 to 360 degrees. Also heading corresponds to 0° to 360°.

The vessel hull form is validated by checking its dimensions. Overall, the validation of collected data and the pre-processing step were vital in ensuring the dataset's quality, dependability, and consistency. Detailed and thorough attention to requisite parameters set up a strong foundation for further analysis and interpretation, which eventually boosted the reliability and validity of the results.

Furthermore, data normalization techniques were not employed to standardize the variables because this step is particularly important when dealing with data from different sources with varying scales and units of measurement.

Standardization facilitated meaningful comparisons and accurate interpretation of results. In the case of the study, AIS was the only sensory data chosen to get data. So, normalisation was not required.

4. APPROACHES

4.1 Baseline Model

The baseline model in this code involves a simple approach of shifting the latitude and longitude values by specific time intervals to create predictions for future coordinates. The shifting technique allows us to compare these predicted values with the actual latitude and longitude coordinates to compute the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

For each specified future interval (e.g., 5, 10, 20, and 30 min), the latitude and longitude columns are shifted by that interval to obtain the predicted values. The NaN rows resulting from the shifting process are then removed from the dataset. The RMSE and MAPE are calculated by comparing the shifted latitude and longitude values with the original ones. These calculations are performed separately for each future interval.

The results for each MMSI (Maritime Mobile Service Identity) are stored in a dictionary, where each MMSI has its own RMSE and MAPE values for different future intervals. To ensure consistency in the analysis, the lists of RMSE and MAPE values are padded with NaN values for MMSIs with fewer data points.

The averages of the related values across all MMSIs and future intervals are then used to calculate the average RMSE and MAPE values. This indicates the baseline model's overall performance in terms of how well it predicts latitude and longitude coordinates.

4.2 CVM Model

The constant velocity method takes a step further by incorporating the vessel's velocity into the prediction process. This approach assumes that a vessel keeps its speed constant during a certain time interval. The CVM more precisely estimates the vessel's future position by taking into account its velocity and heading.

The data is pre-processed before the CVM is applied by figuring out how much time has passed between adjacent data points. Then, using the latitude, longitude, and time difference, the velocity is calculated. A rolling mean is used to calculate the average velocity across the given period. By changing the latitude and longitude data and altering them by the average velocity and heading, the CVM forecasts the next position under the constant velocity assumption.

This method is based on the assumption of linear motion of the vessel. Vessels change their position over a certain period using constant velocity and it is used to predict the vessel location in future time. The code first pre-processes the data by sorting it based on the timestamp and converting it to date-time format. The time difference between subsequent data points is then calculated, excluding any zero values.

The methodology then uses latitude, longitude, and time difference to calculate the object's velocity. A rolling mean is then used to obtain the average velocity for the given period. This average speed is taken to be constant throughout the period.

Using the constant velocity assumption, the code predicts the next position by shifting the latitude and longitude values and adding the product of the average velocity, time interval, and the cosine and sine of the shifted heading values, respectively. These predicted latitude and longitude values are stored in separate columns.

The formulated methodology iterates over unique MMSI values at different time intervals to perform predictions for each combination. These predictions are stored in a data frame. Finally, the Root Mean Square Error (RMSE) and

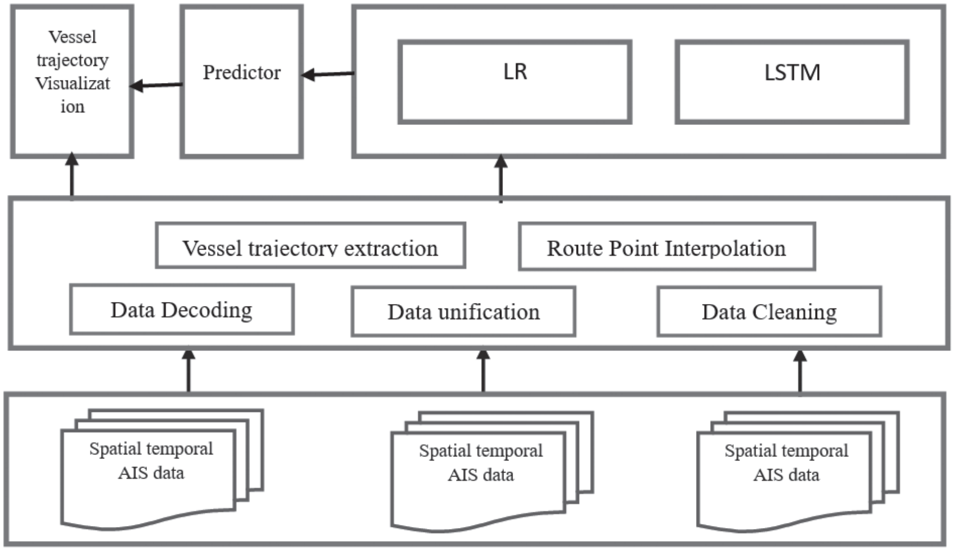
Vessel trajectory Visualizat ion
Predictor
LR
LSTM
Vessel trajectory extraction
Route Point Interpolation
Data Decoding
Data unification
Data Cleaning
Spatial temporal AIS data
Spatial temporal AIS data
Spatial temporal AIS data

Figure 3. Framework of theproposed model for experimentation.

Mean Absolute Percentage Error (MAPE) for each period are calculated by the code by comparing the actual latitude and longitude data with the projected geographical coordinates. The average RMSE and MAPE are printed for each time interval.

In summary, the constant velocity method estimates future positions based on the assumption that the object's velocity remains constant over a given time interval. It provides a straightforward approach for predicting vessel trajectory positions but does not take into account variations in velocity and unpredictable or complex vessel movement patterns.

4.3 Proposed Framework

This paper's approach involves a two-step process. The framework that was followed is enunciated in Fig 3. First, LR was employed to model the linear relationships between vessel attributes (such as latitude, and longitude) and the target positions. LR provides interpretable coefficients, thus enabling us to understand the impact of each attribute on the predicted positions. However, LR may fail to capture non-linear dependencies and temporal dynamics. To overcome this, an LSTM component is incorporated that learns the temporal patterns and dependencies present in the AIS data. Recurrent neural networks of the LSTM are made to recognize long-term dependencies in sequential data input.

4.3.1 Proposed Model (LR-LSTM+RWA)

4.3.1.1 Hyper Parameters

For a good fit of the model, the following hyper parameters were used which improvise their significance in the training and evaluation of the models. Hyper parameters are a set of parameters that are chosen before the beginning of the learning process and have an impact on how the models perform and behave. Understanding and selecting appropriate hyper parameters is crucial for achieving optimal model performance. Hyper parameters used in this approach are described here.

- **Epochs**: The number of epochs determines how many times the training dataset is supplied into the neural network during training. The full training dataset is iterated over during each epoch, which also includes forward and backward propagation, weight adjustments, and other operations. A larger number of epochs may enable the model to discover more intricate patterns in the data. However, using too many epochs could lead to overfitting, which would make the model overly reliant on the training data and underperform on unseen data. A holdout dataset should be used for testing and validation to ascertain the ideal number of epochs. In the proposed model, both the LSTM models for latitude and longitude are trained for 10 epochs.

- **Batch Size**: The batch size is the number of samples that are processed before the model's weights are updated during training. It affects the consistency and pace of the training process. In the code, the batch size is not explicitly specified, and the default batch size is used in the training process. The batch size should be chosen based on the available computational resources and the size of the dataset. Smaller batch sizes may offer more precise updates but slower convergence, whereas larger batch sizes may speed up training but necessitate more memory.

- **Test Size**: The test size is the proportion of the dataset that is used for testing the trained models. In the code, a test size of 0.2 is employed, meaning that 20% of the data is set aside for testing and the other 80% is used for training. The dataset size and the preferred trade-off between training and testing should be taken into consideration when choosing the test size.

- **Random State**: The random state is an integer used to seed the random number generator for reproducibility. To guarantee that the data is split consistently each time the code is run, the suggested model uses a random state of 42 in the train-test split function, allowing for consistent evaluation and comparison. Changing the random state value will result in a different train-test split. It is important to set a random state when sharing or comparing results to ensure consistency.

- **Number of Layers**: The neural network architecture's depth or complexity is indicated by the neural network's layer count. The number of layers is determined by the complexity of the problem and the available data. Deeper models with more layers can potentially capture more intricate patterns but may also be prone to overfitting. Both of the LSTM models for latitude and longitude in the suggested model are made up of a single LSTM layer followed by a dense layer. In the suggested model, there are 50 units in the LSTM layer.

- **Optimizer**: During the training phase, the optimizer is in charge of changing the model's weights following the estimated gradients. In the formulated methodology, the Adam optimizer is used for both LSTM models. The Adam optimizer is a popular choice due to its adaptive learning rate and momentum.

- **Activation Function**: The neural network gains non-linearity from the activation function, which enables it to learn intricate patterns. The LSTM layers in the suggested model use the ReLU (Rectified Linear Unit) activation function.

After extensively configuring the hyper parameters of the model, a range of results is obtained. Through this rigorous process, the performance of the model across different parameter values was observed and evaluated. After careful consideration and analysis, a set of parameter values that best align with the model accuracy was determined. This comprehensive parameter tuning process and having established the suitability of the chosen parameter values enables the making of informed decisions and drawing meaningful conclusions from the model's predictions.

4.3.1.2 Architecture

An architecture has been proposed to handle the sequential processing of latitudes and longitudes, as illustrated

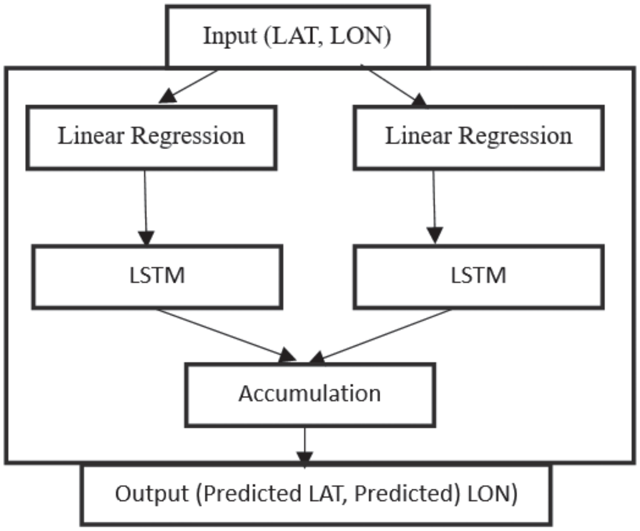

Figure 4. The architecture of theproposed LR+LSTM model.

in Fig 4. The model when executed without the RWA algorithm performs poorly. However, with RWA algorithm described in Algorithm 4 outperforms.

Through the analysis of this dataset, gaps between consecutive AIS data points are identified, signifying areas where additional waypoints need to be interpolated to capture the continuous movement of vessels. Applying the RWA algorithm involves grouping the AIS data based on the unique vessel identifier, known as the MMSI (Maritime Mobile Service Identity) number. Within each group, the data is chronologically sorted by the base date-time to ensure a sequential representation of vessel movements. Subsequently, each group is iterated through, examining consecutive data points to determine time intervals between them. For each identified time interval exceeding a predefined threshold, the number of intermediate steps required to evenly distribute waypoints within the interval is calculated.

An interpolation method was used to calculate the latitude and longitude values for these intermediate waypoints by taking into account the start and end coordinates of the interval. This process takes into account the non-linear nature of the vessel’s path, considering variables such as the vessel’s course or heading.

Once the intermediate waypoints are determined, timestamps based on the original data’s temporal sequence are assigned. As a result, a consistent timeline for trajectory prediction can be maintained in the enhanced dataset.

The resulting augmented data offers a larger and more accurate dataset for trajectory analysis and prediction since it combines the original AIS data points with the recently created waypoints.

The availability of an extensive and varied dataset for training trajectory prediction models is assured through the inclusion of the RWA method in the data augmentation pipeline. This method improves the precision and dependability of trajectory forecasting while reducing the restrictions brought on by a lack of AIS data.

Overall, the suggested methodology for data augmentation, which makes use of the RWA algorithm, provides a reliable method for producing intermediate route waypoints. By offering an expanded dataset that captured the continuous movement patterns of vessels with more granularity and precision, it considerably aided in trajectory prediction job optimization. The research proceeded to develop LR+LSTM models with customized hyper parameters.

```
Algorithm 4: Route Waypoint Augmentation
Require: ais_data: AIS dataset, interval seconds: desired time interval in seconds
Ensure: augmented data: augmented AIS dataset
augmented data ? empty list
grouped data ? group ais data by ’MMSI’
for each mmsi, group in grouped data do
    group ? sort group by ’BaseDateTime’
    group ? reset index of group
    for i ? 0 to length(group) - 1 do
        current point ? get element at index i from group
        next point ? get element at index i + 1 from group
        start time ? convert current point[‘BaseDateTime’] to datetime
        end time ? convert next point[‘BaseDateTime’] to datetime
        time diff ? compute time difference in seconds: end time - start time
        num steps ? round down(time diff /interval seconds)
        if num steps > 0 then
            start lat, start lon ? current point[‘LAT’], currentpoint[‘LON’]
            end lat, end lon ? next point[‘LAT’], next point[‘LON’]
            positions ? interpolatePositions(start lat, start lon, end lat, end lon, num steps)
            time step ? create timedelta with seconds = interval seconds
            current time ? start time + time step
            for each position in positions do
                augmented data.append({‘MMSI’: mmsi, ’BaseDateTime’: current time, ’LAT’: position[0], ’LON’: position[1]})
                current time ? current time + time step
            end for
        end if
    end for
end for
augmented df ? create DataFrame from augmented data
return concatenate ais data and augmented df
```

The hyper parameters were initialized with the selected values during training and iteratively changed them using the appropriate optimization techniques. To achieve accurate performance estimation and avoid overfitting. By analyzing the results, the most promising models based on their performance metrics were selected. This comprehensive exploration and tuning of hyper parameters laid the groundwork for enhancing the models’ capabilities and advancing the research objectives.

5 RESULTS AND ANALYSIS

In this study, the performance of four different methods, including a baseline algorithm, the CVM method, a simple LR+LSTM model, and the approach of LR+LSTM+RWA, were evaluated for performing future trajectory prediction for vessels at various time intervals using historical AIS data. Based on the evaluation results, the LR+LSTM+RWA model exhibited the lowest RMSE values for both LAT and LON predictions, indicating better accuracy compared to the baseline algorithm and the CVM method. The simple LR+LSTM

Table 1. Comparison of performance of various prediction models

Algorithm	Future Interval (min.)	Avg RMSE (LAT)	Avg RMSE (LON)	Avg MAPE (LAT)	Avg MAPE (LON)
Baseline (Shift +1)	5	0.01304	0.01734	0.00037	0.00014
	10	0.02178	0.02780	0.00071	0.00026
	20	0.03368	0.04148	0.00114	0.00042
	30	0.04488	0.05544	0.00153	0.00057
CVM	5	0.0282	0.0413	0.0243	0.0079
	10	0.0324	0.0444	0.0463	0.0146
	20	0.0385	0.0491	0.0686	0.0215
	30	0.0621	0.0693	0.1359	0.0422
LR+LSTM	5	12.66264	44.10535	0.40509	0.46834
	10	12.46050	44.11329	0.39898	0.46910
	20	12.52241	44.13069	0.40302	0.46902
	30	12.55558	43.95859	0.40225	0.46730
LR+LSTM+RWA	5	0.01203	0.01246	0.00038	0.00012
	10	0.01192	0.01229	0.00038	0.00011
	20	0.01204	0.01260	0.00038	0.00011
	30	0.01221	0.01234	0.00039	0.00012

without the RWA algorithm exhibits poor performance even in the baseline model. However, after introducing route waypoint augmentation, the LR+LSTM model drastically improved with the same hyper parameters that were used in simple LR+LSTM. Additionally, the LR+LSTM+RWA model achieved relatively low MAPE values, demonstrating its effectiveness in capturing the patterns and trends in the AIS data.

All of these results were achieved after analysis and experimentation carried out on a large dataset from real-world scenarios. Almost 1048575 different MMSI numbers were taken in the dataset. Experiments were conducted using a comprehensive AIS dataset, comprising a diverse range of vessel types and geographical regions. The results demonstrate that the LR and LSTM model with RWA outperforms individual models in terms of prediction accuracy. The LR component contributes to interpretable insights, allowing us to identify the relative importance of vessel attributes in position prediction. On the other hand, the LSTM component captures non-linear patterns and long-term dependencies, enhancing the model's overall performance.

Overall, the results suggest that the approach of combined LR and LSTM with the RWA model shows a high degree of reliability in predicting AIS data for future time intervals, and further research can be conducted to enhance its performance and explore other advanced prediction techniques.

6 CONCLUSION

In summary, the study compared various approaches for predicting AIS data at various time points in the future. The baseline algorithm and CVM approach were outperformed by the LR+LSTM+RWA model, which also had the lowest RMSE values for LAT and LON forecasts.

Future research could look into ways to improve the LR+LSTM+RWA model by introducing new characteristics or utilizing different machine-learning methods. A thorough grasp of the model's robustness would also come from looking at how it performs in various settings or with different data distributions. Additionally, by benchmarking and developing the field of data prediction using the AIS dataset in the marine domain, comparative studies with other cutting-edge prediction approaches might be conducted.

REFERENCES

1. Tu, Enmei; Zhang, G.; Rachmawati, L.; Rajabally, E. & Huang, G.B. Exploiting AIS data for intelligent maritime navigation: A comprehensive survey from data to methodology. *IEEE Transact. Intell. Transport. Syst.*, 2017, **19**(5),1559-1582.
doi: 10.1109/TITS.2017.2724551
2. Bošnjak, Rino; Šimunović, Ljupko & Kavran, Zvonko. Automatic identification system in maritime traffic and error analysis. *Transact. Maritime Sci.*, 2012, **1**(2), 77-84.
doi: 10.7225/toms.v01.n02.002
3. Jim, Frost. Making predictions with regression analysis, n.d. https://statisticsbyjim.com/regression/predictions-regression (Accessed on 28 April 2023).
4. Jain, A.K.; Murty, M.N. & Flynn, P.J. Data clustering: A review. *ACM Comput. Surveys (CSUR)*, 1999, **31**(3), 264-323.
doi: 10.1145/331499.331504
5. Le. Cun, Y.; Bengio, Y. & Hinton, G. Deep learning. *Nature*, 2015, **521**(7553), 436-444.
6. Xie, Xianwei; Sun, B.; Li, X.; Olsson, T.; Maleki. N. & Ahlgren, F. Fuel Consumption prediction models based on machine learning and mathematical methods. *J. Marine Sci. Eng.*, 2023, **11**(4), 738.
doi: 10.3390/jmse11040738.
7. Zhang, B.; Hirayama, K.; Ren, H.; Wang, D. & Li, H. Ship anomalous behavior detection using clustering and deep recurrent neural network. *J. Marine Sci. Eng.,* 2023, **11**(4), 763.
doi: 10.3390/jmse11040763.
8. Shen, Sheng; Yang, H.; Yao, X.; Li, J.; Xu, G. & Sheng. M. Ship type classification by convolutional neural networks with auditory-like mechanisms. *Sensors*, 2020,**20**(1), 253.
doi: 10.3390/s20010253

9. Chalapathy, Raghavendra; Menon, Aditya Krishna & Chawla, Sanjay. Anomaly detection using one-class neural networks. *arXiv,* 2018, arXiv:1802.06360. doi: 10.48550/arXiv.1802.06360
10. Yasir, M.; Shanwei, L.; Mingming, X.; Hui, S.; Hossain, M.S.; Colak; A.T.; Wang, D.; Jianhua, W.; & Dang, K.B. Multi-scale ship target detection using SAR images based on improved Yolov 5. *Frontiers in Marine Sci.*, 2023, **9**, 1086140. doi: 10.3389/fmars.2022.1086140
11. Wei, Z.; Xie, X. & Zhang, X. Maritime anomaly detection based on a support vector machine. *Soft Comput.*, 2022, **26**, 11553–11566 . doi: 10.1007/s00500-022-07409-w
12. Vessel Traffic Data. Bureau of Ocean Energy Management (BOEM) and National Oceanic and Atmospheric Administration (NOAA). https://marinecadastre.gov/ais/. (Accessed 18 April 2023).

CONTRIBUTORS

Mr Nitish Raj obtained his BTech from NIT Patna and working as a Scientist at DRDO, posted at the Weapons and Electronics Systems Engineer Establishment, Ministry of Defence in New Delhi. In his current role, he holds the position of Senior System Manager of Naval Combat Systems. His research interests encompass System design & development, systems integration, and machine learning.
He contributed to the current work by coming up with the idea and designing the experiment, optimising the deep learning techniques used in the experiment, creating the programme, analysing the data, and writing the manuscript.

Mr Prabhat Kumar is a Professor in the Computer Science and Engineering Department at National Institute of Technology Patna, India. He is also the Professor-In-charge of the IT Services of the Institute and Chairman of Computer and IT Purchase Committee at NIT Patna.
He made contributions to the current study by giving the overall framework for conceptualizing, carrying out, and finishing the experiment.