UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

Master Thesis

on the topic of

Modelling and optimization of ship's fuel consumption using Random Forest Regression (RFR)

Submitted to the Faculty of Engineering
of University Duisburg Essen

by

Hibatul Wafi
3021919

Betreuer:	M. T. Muhammad Fakhruriza Pradana
1. Gutachter:	Prof. Dr.-Ing. B. Noche
2. Gutachter:	Prof. Dr. Ucker
Studiengang:	ISE General Mechanical Engineering
Studiensemester:	Summer semester 2023
Datum:	04.05.2023

Contents

List of Tables

List of Figures

1 Introduction

The research on efficient ship operation is a direction that is being actively pursued by marine industry stakeholders as efficient ship operations equates to increase in profitability. One of the determining factors is the reduction of Fuel Oil Consumption (FOC). FOC takes up considerable portion in ship's operating cost. This is clearly indicated through findings made by Ronen [1] and Stopford [2]. The former mentioned that FOC consumption of a large ship potentially constitute to 75% of the total ship operating cost while the latter noted that FOC makes up to two-thirds of vessel voyage cost and over one-quarter of vessel's overall cost.

With that, maritime industry stakeholder actively searches for inexpensive approach to reduce FOC. As such, they investigate ways to optimise operational measures as technical solutions are expensive [3]. The operational measures include the inclusion of weather/environmental routing, speed optimisation, trim optimisation and virtual (just-in/time) arrival policy [3]. It is noted by Beşikçi et al. [4] that lowering ship speed will have the greatest impact in fuel economy, reducing the ship speed by $2-3$ Knots could halve the operating cost of shipping company [2, 5]. Beşikçi et al. further elaborated that the main cause of this is the nonlinear relationship between ship speed and fuel consumption. Ronen [1, 6] and Wang et al. [7] approximated that fuel consumption can be derived through third order function of the ship speed.

Due to volatility and ever-increasing bunker fuel price, developing a model that could accurately predict ship speed would be beneficial to forecast the ship's FOC. The model could potentially help maritime industry stakeholder make decisions at the most opportune moment. Data driven i.e., machine learning approaches have been attempted by several authors in different literatures to model fuel consumption and reported good results in its predictive performance [4, 8–11]. However, powerful machine learning models are usually unintuitive making it difficult to interpret its decisions [12]. This brings us to Random Forest, a powerful model that offers partial interpretability in their decisions. With this consideration, modelling using Random Forest will be the focus of this thesis.

1.1 Research Objective, Contributions and Boundary

This thesis aims to predict the ship's speed captured by Automatic Identification System (AIS) using random forest regressor. In this study, this speed shall be designated as the ship's Speed Over Ground (SOG). The modelling uses fused hourly data from AIS information of Hammershus Ro-Ro ferry and local meteorological weather data in region of travel. Subsequently, the ship actual speed, which is designated as speed through water (STW), shall be derived from the predicted SOG to enable estimation for fuel consumption over different journey periods. The modelling is performed in `Python` programming language using machine learning packages `sklearn` offered by Pedregosa et al. [13].

Using this approach, we shall raise the following research questions (RQs), namely:

- **RQ1**. Is it feasible to fuse AIS data and meteorological data to accurately predict the ship's SOG ?
- **RQ2**. During modelling, which parameters have the greatest impact in increasing the model's predictive performance ?
- **RQ3**. During evaluation, what are the performance measures that should be considered to help us gain the most information out of the model's behaviour ?

To answer the research questions, the following research boundaries are set:

- Random forest has the capability to solve both classification and regression problem. Because the target variable, SOG, is continuous, we will only adopt the regression algorithm of random forest.
- The focus of this work is a detailed study on the performance and possible optimisation configuration of random forest as predictor for the target variable. As such, we will not perform exhaustive comparison study between different machine learning models.
- The estimation for fuel consumption shall be done using simple formulation by Ronen [1, 6]. This thesis will not consider the more comprehensive method such as the method proposed by Kim et al. [14] as the focus of this work is to estimate the SOG using random forest-
- The Hammershus Ro-Ro ship sails between port of Køge, Rønne, Ystad and Sassnitz. However, we will only consider the journey between port of Køge, Rønne and Ystad as part of the data for the voyage between port of Rønne and Sassnitz are missing.

The use of AIS data provides the following contributions as indicated by Rakke [15]:

- Avoid expenses of purchasing (possibly) unaffordable ship information from online database and shipping companies.
- Independent of commercial parties, as information are available in public domain.

Additionally, this work will provide the following contribution:

- Robust modelling approach that requires minimal data pre-processing and minimal model configuration.

The remaining sections of this study is organized as follows. Section 2 include literature review pertaining to relevant past and present research. Theory to random forest will also be discussed in Section 2.2.2. Section 3 discuss the methodology used to develop random forest model used for SOG prediction. The discussion comprises analysis of training data, feature selection and reduction and selection of tree tuning parameter. In Section 4, the model will be evaluated using appropriate performance metrics and their effectiveness will be compared. The summary of this study and reflections of the research process will be discussed in Section 5.

2 Theoretical Background

The literature review in Section 2.1 presents past and present research on utilisation of machine learning method to achieve energy efficient operation. The concept of modelling ship operation using machine learning model is explained in Section 2.1.1. Performance of random forest as predictor in previous research are discussed in Section 2.1.3. Brief conclusion of the literature review is presented in Section 2.1.4.

2.1 Literature Review

2.1.1 Modelling Approach for Ship Operation

The work by Yan et al. [16] provides a thorough review of the different attempts that have been made by different authors to predict different parameters of ship's operation, this includes ship's fuel consumption. Per definition by Haranen et al. [17], the modelling of ship operation is categorised into White Box Model (WBM), Black Box Model (BBM) and Grey Box Model (GBM). Machine learning approach is categorised as BBM, BBM approach is defined as purely data driven approach requiring no prior knowledge about the ship operation. The literature review by Yan et al. [16] indicated that about 42% of the research utilised BBM model based on machine learning approach.

Yan et al. [16] elaborated further in their work, that BBMs in general have a good fitting ability for unseen data. BBMs based on machine learning model are able to generalise better compared to BBMs based on statistical modelling. With increasing amount of data, better generalisation performance and handling noisy of data should be expected in a BBM model. However, for this same reason, the quality of BBM model is highly dependent on data quantity and quality. BBM model are also generally complex making it challenging to analyse and explain. Shipping industry experts also are having difficulty accepting models that violate the domain knowledge.

From the work of Yan et al. [16], it can be concluded that model accuracy and appropriateness is a significant factor that should be considered when modelling. The model should obey shipping domain knowledge and an intuitive model will help shipping experts analyse its accuracy. For this reason, tree-based model will be considered as it is known for its intuitiveness and interpretability [18]. Breiman et al. [19] even claimed that random forest is the "most interpretable" and "most accurate". With that,Section 2.1.3 will focus on predictive performance of random forest against different machine learning model on different data types and data source.

2.1.2 Use of AIS Data for Scientific Research

Apart from its intended use as collision avoidance system AIS data have seen potential usage in the field of scientific research. In the third Green House Gas (GHG) study by Smith et al. [20], uses AIS to estimate global shipping emission inventories. Rakke [15] proposed a methodology termed ECAIS to calculate ship emissions based on the fuel consumption from AIS data. Through Holtrop-Mennen approximation and literature approximation, the ship's power propulsion can be determined which is subsequently used to predict specific fuel consumption. Wen et al. [21] attempted to minimise the Energy Efficiency Operational Indicator (EEOI) using green routing. Recent research by Kim et al. [22] used publicly accessible AIS data, ship static data and environmental data to estimate EEOI without requiring the actual FOC. The study used big data technology as public data are of large capacity. Generally, the study using AIS data is done to achieve independence from the need to use commercial database. The detail of AIS data will be discussed in Section 2.3

2.1.3 Prediction Performance of Random Forest

Majority of the BBM approach based on ML is dominated by ANN [16]. However, there are literatures that considered decision tree-based modelling approach to predict fuel consumption. Some example of decision tree based modelling include Decision Tree (DT), Random Forest (RF) and Extra Tree (ET). Soner et al. [23] implemented tree-based model, which include bagging, random forest (RF), and bootstrap. In their work, they used data captured from onboard sensors of a ferry to predict speed through water and fuel consumption per hour. From the test dataset, the random forest model described root mean square error (RMSE) of 0.34 Knots during its prediction of Speed Through Water (STW). Yan et al. [24] used random forest (RF) model to minimise fuel consumption for a voyage of a dry bulk ship. The model use ship operational data and sea and weather data from noon report and EMCWF. The prediction performance report from this literature reported mean absolute percentage error (MAPE) of 7.91%.

The research by Gkerekos et al. [9] highlighted the performance of different machine learning models to predict ship's fuel consumption per day using both noon data and automated data logging and monitoring (ADLM) system from a bulk carrier. This research concludes that tree based model displayed good prediction performance on both noon data and sensor based data. Using default parameters, RF model obtained R^2 score of 87.55% and 96.26% for noon-data datasets and sensor-based data respectively. It is also noted that it that the data from a 3-month period in ADLM system would be sufficient to create a model with better performance than the model generated by noon data from a collection period of 2.5 year. This literature also concluded that automatic sensor-based data have the potential to increase the model accuracy score, R^2, by $5-7\%$ across different machine learning models.

Li et al. [3] performed more extensive research on the effects of data fusions between

meteorological data, ship voyage data and AIS data on different machine learning models to predict the ship's FOC. This research highlighted the advantage of fusing meteorological data and ship voyage data. The evaluation on different model performance indicated that RF are among preferable model candidate that could be used in commercial scale due to its good prediction capability and robustness against different datasets. The findings in this research reported that R^2 score are above 96% when deployed on the best datasets and achieved R^2 score in range between 74% – 90% over test data. This literature also exhibited the robustness of RF, as it attained the lowest standard deviation at 0.015 of the R^2 score when evaluated against random splits of datasets.

Abebe et al. [10] used different approach in their research by predicting the ship's Speed Over Ground (SOG) instead of FOC. In this work, AIS data and noon-report weather data from 14 tracks and 62 ships are used for the SOG prediction. The observation showed that RF model achieved RMSE of 0.25 knots, while using 489 seconds for training. Decision tree achieved RMSE of 0.36 knots, taking up 52 seconds for training. This shows that RFR outperforms DTR at cost of computational power.

2.1.4 Conclusion of Literature Review

This literature review described the capability of Random Forest Regressor to predict fuel consumptions and ship speed, irrespective of data source and type of data used. Promising results from different performance measures across different literatures indicated the capability random forest model as predictor. As such, this thesis aims to find optimisation possibilities to extract maximum prediction performance from random forest. Due to the nonlinear, third order function estimate of fuel consumption [1, 6]. Accurate prediction of ship speed is paramount to ensure optimal ship operation resulting in increase of profitability.

2.2 Tree-based model

Random forest belongs to the family of tree-based model and its functional principle stems from decision tree. Decision tree is a non-parametric model that can perform both classification and regression tasks for discrete variable and continuous variable. It is a powerful algorithm, capable of fitting complex datasets. Tree-based model requires very little to no data pre-processing [12, 25]. To grasp the concept of random forest, The principle working of decision tree will be introduced in depth in Section 2.2.1. It is then followed by Section 2.2.2 which presents the principle function behind random forest. Brief explanation for Extra-Trees, method introduced to further improvise random forest, will be presented in Section 2.2.3.

2.2.1 Decision Tree

Decision tree is a white box model[1] [12]. In machine learning sense, this means that the model is intuitive, and the structure of the model is interpretable. Thus, the structure of the model can be analysed in detail. To train Decision Trees, `Scikit-Learn` [13] uses the *Classification and Regression Tree* (CART) algorithm [18]. Partition space shown by Figure 1 are used to illustrate the decision of CART algorithm. This process can be alternatively represented by the binary tree of Figure 2, observation that satisfies the condition are assigned to the left branch and the opposite is assigned to the right branch. The binary tree representation can be especially helpful when multiple input variables are involved, as the responses can be represented by a single tree [25].

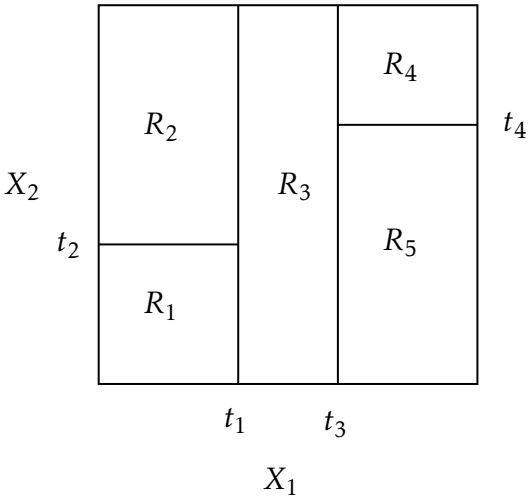

Figure 1: Example of partition space [25]

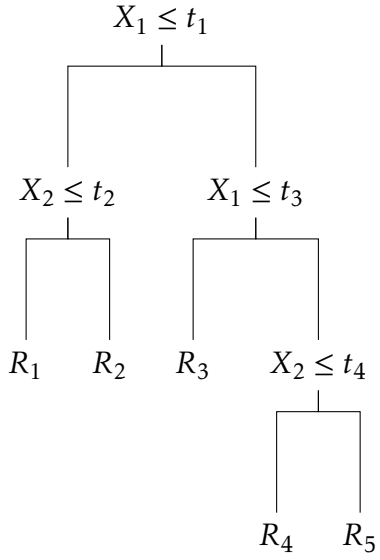

Figure 2: Example of partition tree [25]

Now, we need to understand the principle of selection for the feature k_t and threshold t_k. We shall first start with the principle of selection of the threshold t_k; Assuming a case with single feature k and response y, with m data points. The algorithm starts by looking for possible thresholds. This is determined by calculating the splitting value.[2]. Then, the mean of the response y of partition space S_1 and S_2 is calculated as seen in Figure 3.

This step is then followed by calculating the sum of squared error (SSE) of each data points in partition space S_1 and S_2 and dividing it by the number of data points m_{s_1} and m_{s_2} respectively to obtain the MSE. Subsequently, the MSE from the respective partition space S_1 and S_2 is summed. The process is then recursively repeated until

[1]This is not to be interchanged with the definition described by Haranen et al. [17] regarding modelling of ship operation.

[2]For example, suppose there are data points at $k = [0.2, 0.4]$, then the splitting value is the value in between, i.e., $t_k = 0.3$

a threshold t_k that produce minimum sum of MSE is determined. This algorithm is defined by the following cost function $J(k, t_k)$, with $\hat{y}_{S_i}$, being the mean of the response, y_{S_i}, in partition space S_i. [12, 26]:

$$\text{MSE}_{S_i} = \frac{1}{m_{S_i}} \text{SSE}_{S_i} \quad \textbf{where} \quad i = (1, 2) \tag{2.1}$$

$$J(k, t_k) = \frac{1}{m_{S_1}} \text{SSE}_{S_1} + \frac{1}{m_{S_2}} \text{SSE}_{S_2} \begin{cases} \text{SSE}_{S_i} = \sum_{i \in S_i} (\hat{y}_{S_i} - y_{S_i})^2 \\ \hat{y}_{S_i} = \frac{1}{m_{S_i}} \sum_{i \in S_i} y \end{cases} \tag{2.2}$$

Once complete, then the partition space is further split into two more regions and this process is recursively continued until a stopping rule is applied. The stopping rule are either when the tree reaches the maximum depth, (This is controlled by the parameter `max_depth` in `Scikit-Learn`), or when it cannot find a split that can further reduce MSE. This best split also corresponds to the best possible fit to the predicted value.

Same principle is also applied when multiple features are present. Consider there are k_t features, then for each respective features $k_1, k_2, \ldots, k_t$, The MSE for each of the features is calculated using the cost function $J(k, t_k)$. The feature that can ***minimise*** the cost function will be selected as the root of the tree. The tree is then grown further by recursively repeating this process [12, 25].

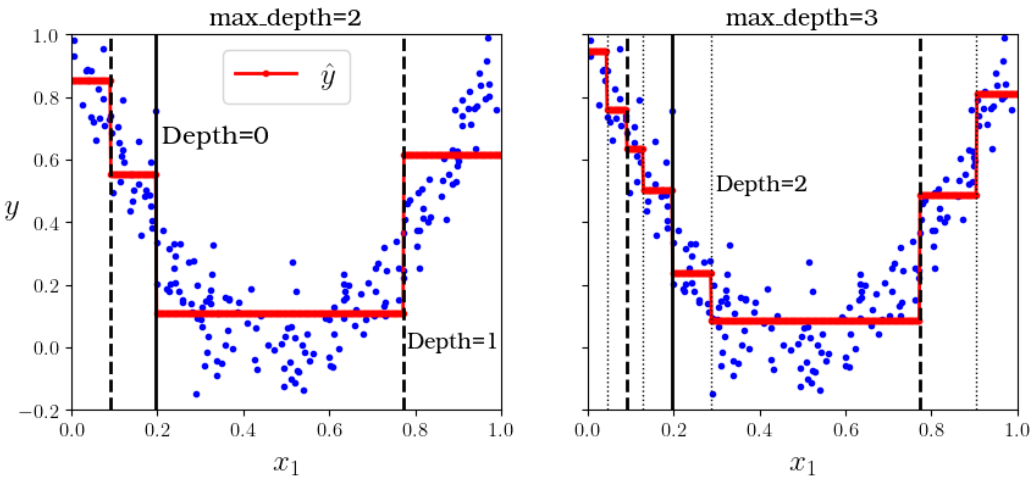

Figure 3: Prediction of two Decision tree regression models [12]

While powerful, decision tree suffers from overfitting when the model is unconstrained. Decision tree makes very few assumptions regarding the training data. Therefore, it will adapt to the training data and fitting it very closely [12]. Additionally, an individual tree tends to be unstable, when the data is altered, a completely different set of splits might be found [25, 26]. Therefore, it is necessary to regularise i.e., restrict the decision tree's freedom during the training. Overfitting could be reduced by controlling how deep the tree can grow through the `max_depth` parameter.

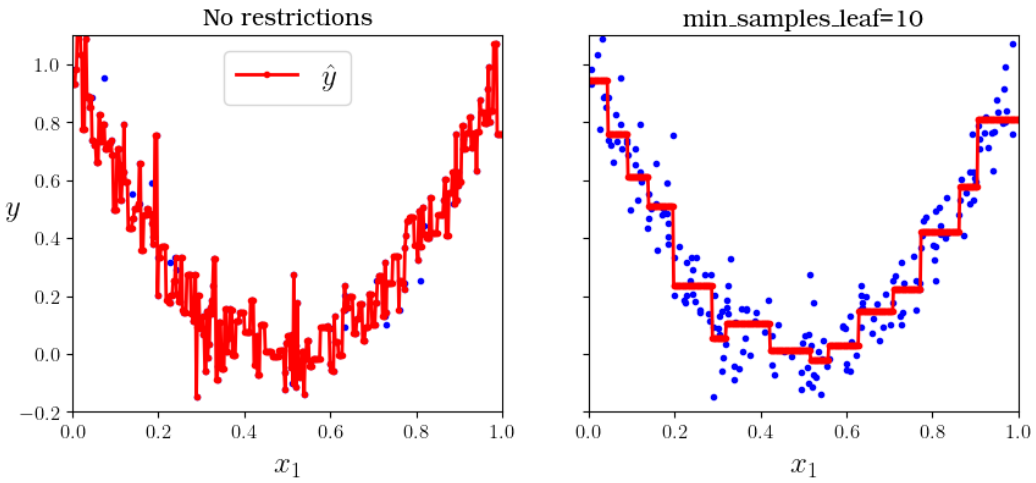

Figure 4: Regularising a Decision Tree regressor [12]

Additionally, setting the amount of minimum number of samples a leaf node has, through `min_samples_leaf` can alleviate overfitting as well, as shown in Figure 4. However, to address the fundamental drawbacks of decision tree, we shall look into random forest.

2.2.2 Random Forest

To understand random forest, the concept of ensemble method shall first be understood. Ensemble is defined as group of predictors such as classifier or regressor. Predictions are aggregated across multiple predictors, for regression task, the prediction is the average across the predictors. This principle is applied to random forest, a group of decision trees is trained on different random set of training data. For regression task, this means the prediction value is the average of the prediction across the decision trees. Such ensemble of decision trees is called ***Random Forest*** [19,25,27].

Ensemble methods achieve the best performance when the predictors are as independent to one another. In statistical sense, this can be achieved by reducing correlation among the trees. This can be realised by adding randomness during tree construction process. For this purpose, random forest utilises *bagging* [28] method (short for *bootstrap aggregating*) during the training process. First, bootstrap sample is created, this means that a sample of the dataset is randomly selected and allowed to appear more than once. This sampling technique is referred to as sampling *with* replacement. Once predictors are trained, then the prediction of the new instance is *aggregated* across the predictors. [12,25,26]

To add further randomness, random forest involves random selection of input features k that are considered to split the tree. This means that the feature k that will be used to split the tree is selected from this random subset of feature. The selection for the best feature to be used as the root of the tree and its subsequent node, as well as the stopping rule for the tree's growth is similar to that of decision tree. [12,25,26]

These measures introduced in random forest address the tendency of decision tree to overfit. In fact, the instability of decision tree mentioned in Section 2.2.1 is exploited in random forest to gain randomness during construction of the tree. Experience from Hastie et al. [25] shown that random forest requires minimal parameter tuning to achieve satisfactory performance while Kuhn et al. [26] reported that tuning parameter does not have a drastic effect on performance.

However, what random forest gains in predictive performance, loses in interpretability. Random forest is considered as Black Box Model (BBM) [12, 29]. [3] The randomness means that it is challenging to analyse and describe the decisions made during the selection of the samples and during the selection of the input features. Nevertheless, the interpretability of a single tree in a random forest still holds. As it is still possible to traverse through the tree to reach the predicted value.

2.2.3 Extra-Trees (Extremely Randomised Trees)

Additionally, extra-trees (Extremely Randomised Trees) is introduced by Geurts et al. [30] to further randomise random forest. The key difference lies on how each split is selected; in extra-trees each tree split is selected in random instead of searching for the best split. This technique saves computational power, as searching for best split is one of the tasks that takes up most computational power [12]. Extra-trees also do not bootstrap the samples, which mean it samples *without* replacement.

2.3 Introduction to AIS data

Automatic Identification System (AIS) is an automated tracking system onboard ships to automatically transmit information about the ship to other ships and coastal authorities. As part of the revised new chapter V of SOLAS[4] regulation. In 2000, International Maritime Organization (IMO) requires installation of AIS class A equipment on all ships of 300 gross tonnage and upward engaged on international voyages, cargo ships of 500 gross tonnages and upwards not engaged on international voyages and all passenger ships irrespective of size. This requirement is then made compulsory to all ships by 2004. [15, 31]

AIS uses Very High Frequency (VHF) with special protocol for communication system for information exchange between the ships. This information will be received by either ships directly, buoys, land based station and satellites. The information transmitted by AIS is distinguished into three different types. **Static information** which is entered into the AIS on installation. **Dynamic information**, which is automatically updated from the ship's sensors connected to AIS and **voyage-related**

[3] Again, not to be interchanged with the definition described by Haranen et al. [17] regarding modelling of ship operation.

[4] International convention for the Safety of Lives at Sea

information, which might need to be manually entered and updated during the voyage. The structure of the AIS data that is relevant to this thesis is summarised in Table 1 [31].

AIS is also further differentiated by its equipment class. The classification is based on the reporting interval and the type of information that is conveyed. **Class A** autonomously report their position within 2-10 seconds interval, depending on the state of ship's movement. The reporting interval is less frequent at 3 minutes, When the ship is at anchor or moored and moving slower than 3 knots. Class A AIS is also capable of sending safety related information, meteorological and hydrological data, electronic broadcast to mariners and marine safety messages. **Class B** reports at longer interval and at a lower power. They can only receive safety related messages, not send them. [15, 31]

2.3.1 Current correction

As indicated in Table 1, the speed shown in AIS is the speed over ground (SOG). However, for calculation of ship's fuel consumption, the actual speed i.e. speed through water (STW) is required. This can be achieved by correcting the SOG for the current speed, in consideration of the research by Zhou et al. [32] which shows the impact of current on ship's SOG. This correction is performed by considering the current speed V_c and the direction of the current γ *with respect to True North*. In principle, STW will be greater than SOG, when the current is moving against the current as the ship tries to compensate for the current to maintain the SOG. Similarly, the STW will be greater than the SOG when the current is moving in the same direction of the ship movement.

To calculate the correction, this study will adopt the methodology proposed by Kim et al. [22] and Yang et al. [33]. The x and y component of SOG can be obtained through vector decomposition using the ship's heading angle α *with respect to True North*. Similar vector decomposition is also performed for current speed V_{current}, it is resolved with current direction γ *with respect to True North*:

$$V^x_{\text{SOG}} = V_{\text{SOG}} \cdot \sin(\alpha) \tag{2.3}$$

$$V^y_{\text{SOG}} = V_{\text{SOG}} \cdot \cos(\alpha) \tag{2.4}$$

$$V^x_{\text{current}} = V_{\text{current}} \cdot \sin(\gamma) \tag{2.5}$$

$$V^y_{\text{current}} = V_{\text{current}} \cdot \cos(\gamma) \tag{2.6}$$

Then the resulting equation to determine STW, including the current compensation, is given by:

$$V^x_{\text{STW}} = V^x_{\text{SOG}} - V^x_{\text{current}} \tag{2.7}$$

$$V^y_{\text{STW}} = V^y_{\text{SOG}} - V^y_{\text{current}} \tag{2.8}$$

$$V_{\text{STW}} = \sqrt{(V^x_{\text{STW}})^2 + (V^y_{\text{STW}})^2} \tag{2.9}$$

Information Item	**Description**
Static	
MMSI	MMSI number of vessel
Callsign	Callsign of vessel
Name	Name of the vessel
IMO	IMO number of the vessel
Length	Length of vessel
Width	Width of vessel
Ship Type	Describes the AIS ship type of this vessel
Dynamic	
Ship's position	Automatically updated from position sensor connected to AIS. Longitude and Latitude.
Position time stamp in UTC	Automatically updated from ship's main position sensor. Format: DD/MM/YYYY HH:MM:SS
Course over Ground (COG)	***If available***, automatically updated from ship's main position sensor connected to AIS.
Speed Over Ground (SOG)	***If available***, automatically updated from the position sensor connected to AIS.
Heading	Automatically updated from the ship's heading sensor connected to AIS
Navigational status	Navigational status information has to be manually entered by the Officer on Watch (OOW) and changed as necessary. For example : "*underway by engines*","*engaged in fishing*","*at anchor*".
Rate of Turn (ROT)	***If available***, Automatically updated from the ship's ROT sensor or derived from the gyro.
Voyage Related	
Ship's draught	To be manually entered at the start of the voyage using the maximum draft for the voyage and amended as required
(Hazardous) Cargo Type	Type of cargo from AIS message.
Destination and ETA	To be manually entered at the start of the voyage and kept up to date as necessary.

Table 1: Structure of AIS data [31]

2.4 Calculation of Fuel Oil Consumption

3 Research Methodology

In this chapter the methodology used to develop random forest model will be discussed. The details of fusion between AIS data, ECMWF and CMEMS data source used for training the model will be presented in Section 3.1. Suitable methodology application during data pre-processing will be described in Section 3.2. The selection for appropriate, domain knowledge based, feature selection will be explained in Section 3.2.2. The selection of the most optimal model hyperparameter for different tree-based model will be explained in Section 3.3.2. Different performance metrics is used to validate the model's generalisation capability, The underlying principle of the metrics is elaborated in Section 3.3.1. Summary of methodology application in this study is summarised in Section 3.3.3 and visually represented in figure Figure 5.

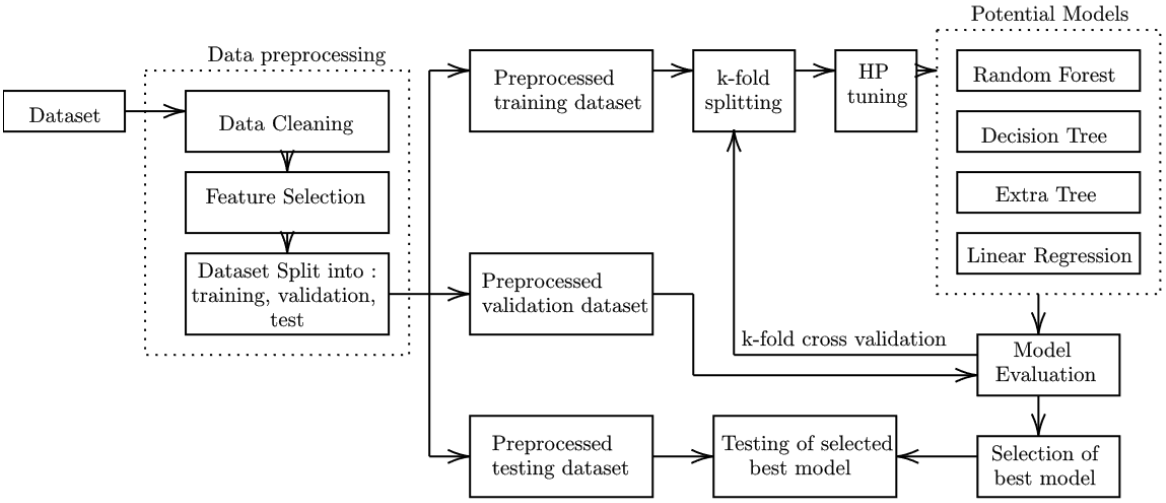

Figure 5: Scheme of proposed methodology

3.1 Data Acquisition

For the purpose of model training, 2021 AIS data from Ro-Ro ferry ship Hammershus is collected. The shore-based AIS data is made available by Danish Maritime Authority which tracked her journey between ports of Køge, Rønne, Ystad and Sassnitz and structured according to Table 1. The AIS data is fused with weather data from ECMWF[5] with temporal resolution of 1 hour at granularity of 0.25° (longitude) x 0.25° (latitude), data from ECMWF provides information for wind, waves and seawater temperature. The information for current is obtained from CMEMS[6] with temporal resolution of 3 hours at granularity of 0.25° (longitude) x 0.25° (latitude).

[5]European Centre for Medium-Range Weather Forecast
[6]Copernicus Marine Environment Monitoring Service

The resulting fusion resulted in dataset with temporal resolution of 1 hour. Some information static information from the AIS data which only indicated the ship's identity are excluded. This includes ship's MMSI, Callsign, Name, IMO and Navigational Status. Additionally, information of the ship's Rate of Turn (ROT) is not available in this case. The weather information is synchronised so that the wind, waves, seawater temperature and sea current belongs to the same weather grid with same temporal resolutions.

The features (1) wind direction, (2) swell direction, (3) and wind wave direction are oriented to true north. However, to reflect the actual direction of weather effects that are acting on the ship, these features are converted to true direction; where true direction is defined as the direction of weather effect with respect to the bow of the ship. The value ranges between 0° and 180°. Subsequently, through vector decomposition, the northward and eastward wind velocity is converted to absolute wind speed and wind direction *with respect to True North*,φ:

$$V_{\text{wind}} = \sqrt{(V^N_{\text{wind}})^2 + (V^E_{\text{wind}})^2} \tag{3.1}$$

$$\varphi = \begin{cases} 360 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} > 0 \quad \wedge \quad V^N_{\text{wind}} < 0 \\ 180 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} < 0 \quad \wedge \quad V^N_{\text{wind}} > 0 \\ 270 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} > 0 \quad \wedge \quad V^N_{\text{wind}} > 0 \\ \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{otherwise} \end{cases} \tag{3.2}$$

Similarly, information of Northward and Eastward current Velocity is converted to absolute current speed and current direction *with respect to True North* γ.

$$V_{\text{current}} = \sqrt{(V^N_{\text{current}})^2 + (V^E_{\text{current}})^2} \tag{3.3}$$

$$\gamma = \begin{cases} 360 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} < 0 \quad \wedge \quad V^N_{\text{current}} > 0 \\ 180 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} > 0 \quad \wedge \quad V^N_{\text{current}} < 0 \\ 270 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} < 0 \quad \wedge \quad V^N_{\text{current}} < 0 \\ \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{otherwise} \end{cases} \tag{3.4}$$

This conversion is performed as the information of current speed and current direction, γ, is necessary to perform the correction formula shown in Equation (2.7) and Equation (2.8). However, for training purpose, this feature will not be considered. Instead, the true current direction and true wind direction will be considered. The initial structure have 27 features, 9 AIS features and 18 weather features. The structure of the initial dataset i.e. before data preprocessing and feature selection, is summarised in Table 2

Feature	Feature Name
AIS data	
Position Time Stamp [DD/MM/YYYY HH:MM:SS]	`Time`
Latitude [°]	`LAT`
Longitude [°]	`LON`
Width [m]	`width`
Length [m]	`length`
SOG [Knots]	`sog`
COG [m/s]	`cog`
Heading [°]	`heading`
Draught [m]	`draught`
Weather Data (0.5° Granularity)	
Wind Speed [m/s]	`windspeed`
True North Wind Direction, φ [°]	`truenorthcurrentdir`
Air Temperature Above Oceans [K]	`oceantemperature`
Maximum Wave Height [m]	`waveheight`
Swell Period [s]	`swellperiod`
Wind Wave Period [s]	`windwaveperiod`
Wave Period [s]	`waveperiod`
Sea Surface Temperature [K]	`surftemp`
Combined Wind Wave Swell Height [m]	`windwaveswellheight`
Swell Height [m]	`swellheight`
Wind Wave Height [m]	`windwaveheight`
Current Speed [m/s]	`curspeed`
True North Current Direction γ [°]	`truenorthcurrentdir`
True Wind Direction [°]	`truewinddir`
True Current Direction [°]	`truecurrentdir`
True Swell Direction [°]	`trueswelldir`
True Wind Wave Direction [°]	`truewindwavedir`
True Wave Direction [°]	`truewavedir`

Table 2: Structure of fused dataset

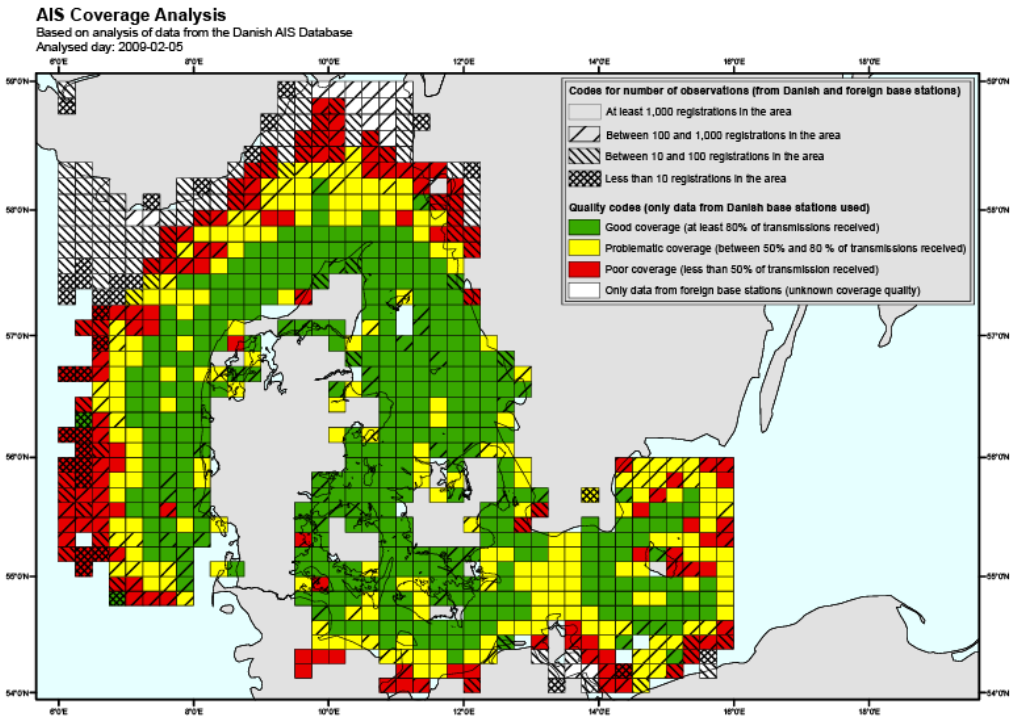

Figure 6: Shore based AIS Coverage based on data from AIS database [34]

3.2 Data Preprocessing

This section presents the steps taken to during data preprocessing. The dataset will be first subjected to data cleaning which include identification of anomalies and missing values, the steps are explained in Section 3.2.1. Boundary condition is then applied to ensure that the model represent operating condition at steady state. Using domain knowledge, appropriate features are selected and discarded to ensure the model obeys shipping domain knowledge. This dataset is to be split into training, validation and test dataset. These steps will be further elaborated in Section 3.2.2.

3.2.1 Data Cleaning

The journey between the port of Køge, Rønne, Ystad and Sassnitz is plotted using `QGIS`[7]. The plot of the journey is shown in Figure 7, it can be seen, that the journey between Rønne and Sassnitz is not represented completely. As in this information is missing due to poor coverage in the area between Sassnitz and Rønne. This is shown by the plot shown in Figure 6. Therefore, the data plot for the journey between Sassnitz and Rønne will be excluded. Basic threshold of decimal degrees of 55.04° N for latitude is applied, this threshold will exclude the journey between Sassnitz and Rønne.

In its initial state, the dataset contains 7453 data points which described the journey of the ship in one year. The initial data points represented all navigational status of

[7]https://qgis.org/en/site/, `QGIS` is a free and open source geographic information system

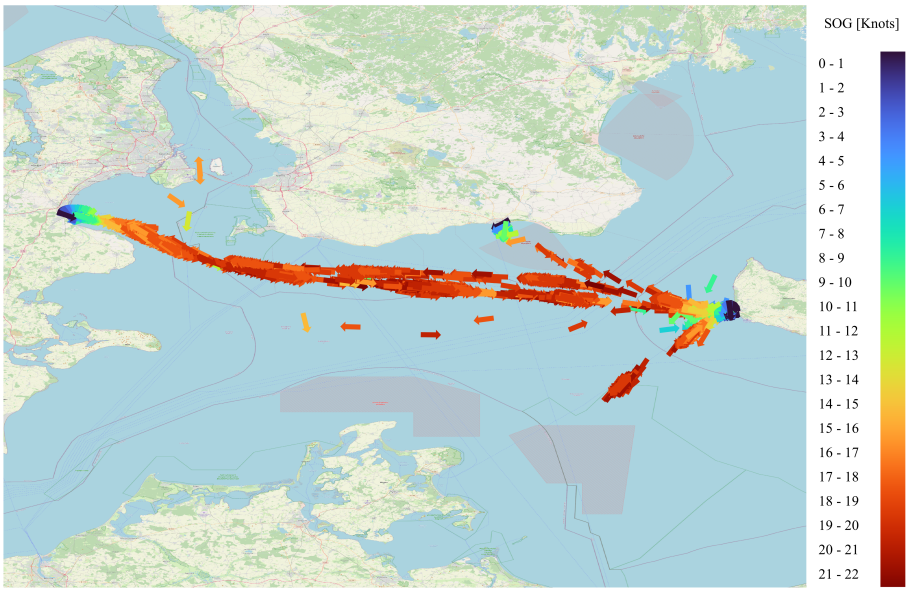

Figure 7: Journey of the ship in a year

the ship, which include “mooring”, “anchoring” and “underway using engine”. This is clearly observed in the histogram for the SOG distribution in figure BLALA. To ensure that the dataset represents the actual operating condition of ship in steady state, a threshold of 5 knots is applied. SOG can vary due to changing sea state, but it can also be reduced by the ship’s operator around the port when it departs from port of origin or arriving at port of arrival. Any data points with SOG less than 5 knots will be discarded which is considered as manoeuvring [10]. After applying the SOG threshold, the amount of data points significantly decrease from 7453 data points to 3506 data points. This indicated that about half of the total data points represented the ship’s stationary behaviour.

From preliminary analysis, possible source of error is identified for data points representing current speed. In range of current speed between 0.01 and 0.03 [m/s], noticeable peak in data points is observed. This peak attributed to missing information on northward and eastward current speed in some data points from the provided dataset. This resulted in single random error value for current speed which resulted in the peak observed in the histogram.

To address the missing values, the missing values for eastward current and northward current are imputed using `KNNImputer` feature from `Scikit-Learn`. Each sample’s missing values are imputed using the mean of nearest neighbour found in training dataset [13]. Once the missing values of northward and southward current are imputed, the current speed for the missing values will be recalculated.

The imputing approach using k-nearest neighbour is also applied to other weather features that contained missing values i.e. `NaN` values. Imputing missing values is necessary as modelling package by `Scikit-learn` cannot handle missing values.

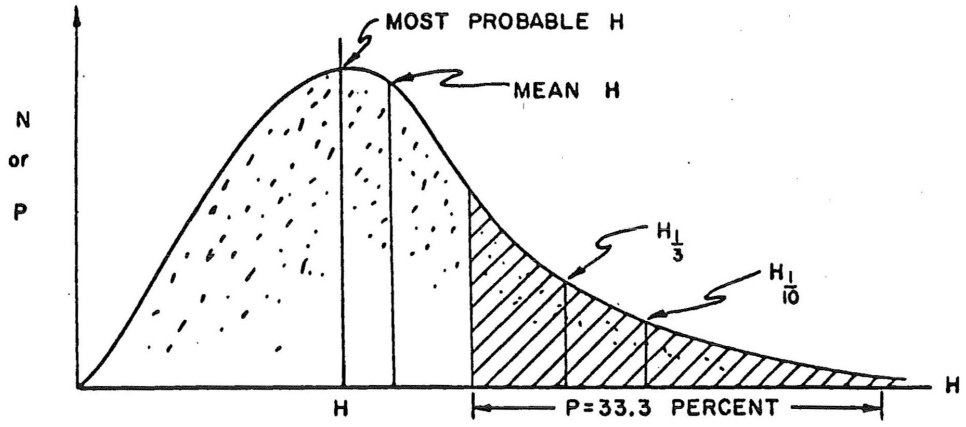

Figure 8: Statistical distribution of wave heights [36]

Imputing strategy using k-nearest neighbour is considered as it should reflect the weather conditions within the region of missing values.

3.2.2 Feature Selection

To select appropriate features for the model, correlation between the features is first studied. Feature selection is necessary to simplify the model and subsequently save computing cost during training. Selection of features is based on statistical approach of High Correlation Filter proposed by Abebe et al. [10]. This approach considers pairs of features with correlation features higher than 0.7 as one entity. However, the selection of highly correlated features must not violate natural state of matter. Therefore, in addition to statistical approach, the scientific reasoning behind the correlations will be considered and prioritised over the statistical approach.

From AIS data, the information on (1) time, (2) latitude, (3) longitude, (4) width, and (5) length are not included for training. As time, latitude and longitude have no impact on the ship. While the width and length is properties from the ship that remain constant.

The features (1) combined wind wave swell height, (2) swell height, maximum wave height (3) and wind wave height are physically correlated. In sea wave theory, wind wave swell height is also known as significant wave height $H_{1/3}$. It is defined as the mean of the highest one-third of waves in the wave record [35].

The distribution of wave heights can be represented by probability density function. Hence, the term "highest one-third of waves" here means the region of wave heights that belong in the upper one-third of a probability density function, this is illustrated in Figure 8. From this distribution, the relation between significant wave height $H_{1/3}$, the highest ten percent of waves H_{10} and average wave height $\overline{H}$ can be summarised as follows [35, 36]:

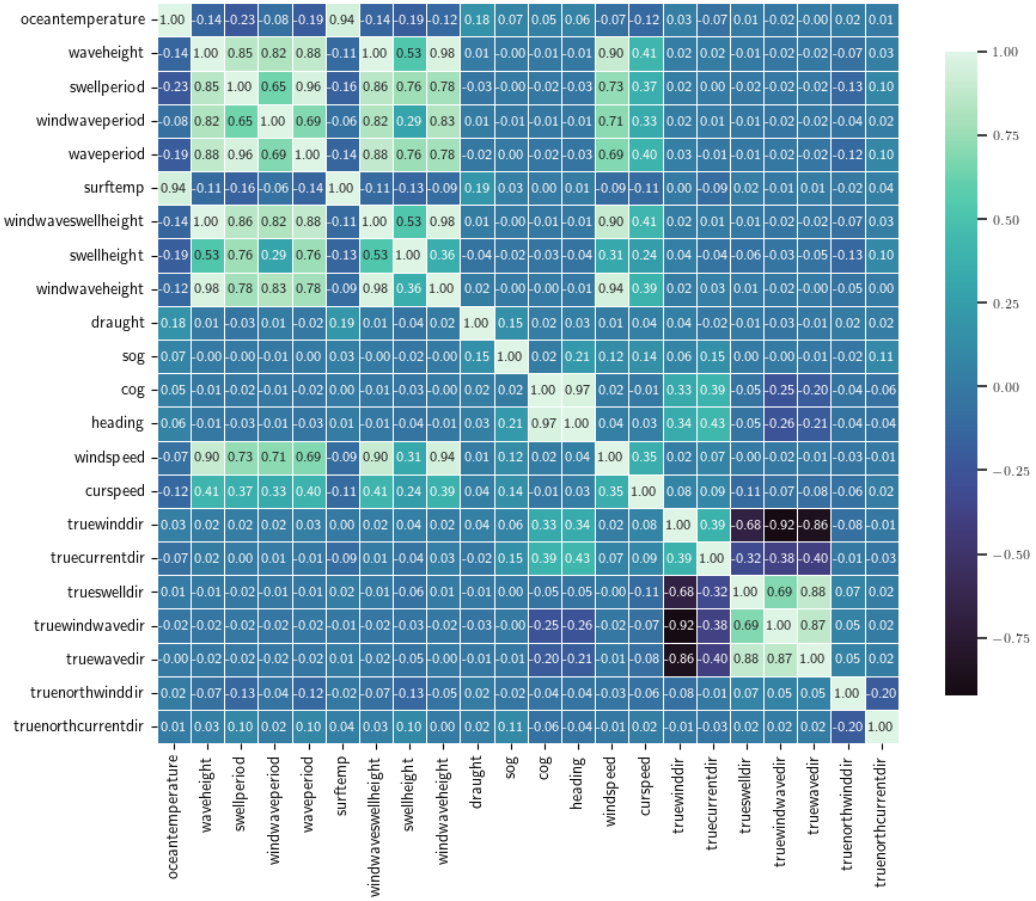

Figure 9: Correlation Heat Map

$$\overline{H} = 0.625 \cdot H_{1/3} \tag{3.5}$$

$$H_{10} = 2.03 \cdot \overline{H} = 1.27 \cdot H_{1/3} \tag{3.6}$$

$$H_{\max} = 2 \cdot H_{1/3} \tag{3.7}$$

Additionally, Bitner-Gregersen [37] described the relation between the significant wave height, wind wave height and swell height through following equation:

$$H_{1/3} = \sqrt{(H_{\text{swell}})^2 + (H_{\text{windwave}})^2} \tag{3.8}$$

From here, it is clear that significant wave height should be retained for modelling, as it holds critical information regarding wave properties. The features swell height, wind wave height and maximum wave height will be dropped as it can be defined through correlations defined in Equation (3.5),Equation (3.6),Equation (3.7) and Equation (3.8). This decision is also statistically supported through the high correlation filter method. As shown in Figure 17, high correlation are observed between these features.

Training Label	
SOG [Knots]	sog
Training Features	
COG [m/s]	cog
Heading [°]	heading
Draught [m]	draught
Wind Speed [m/s]	windspeed
Air Temperature Above Oceans [K]	oceantemperature
Maximum Wave Height [m]	waveheight
Wave Period [s]	waveperiod
Sea Surface Temperature [K]	surftemp
Combined Wind Wave Swell Height [m]	windwaveswellheight
Current Speed [m/s]	curspeed
True Wind Direction [°]	truewinddir
True Current Direction [°]	truecurrentdir
True Wave Direction [°]	truewavedir

Table 3: Structure of fused dataset

From Figure 17, high correlation is observed between wave period, swell period and wind wave period. Bitner-Gregersen further elaborated that the state of the sea can be described through the significant height $H_{1/3}$ and spectral peak T_p with help of Torsethaugen peak [38]. Hence, the features swell period and wind wave period are discarded as it only distinguish whether the sea is dominated by swell or by wind. The feature wave period will still be retained. Consequently, the features true wind wave direction and true swell direction will be discarded as the features that explained the magnitude of these features are discarded.

Statistically, the heading and COG are highly correlated, but both features are retained as it explain two different parameters of the ship. Course Over Ground reflects the ship course heading while heading represented the actual heading of the ship at a particular point of time. Same principle also apply between air temperature above ocean and sea surface temperature. Air temperature above oceans represents the temperature of wind while sea surface temperature represents current temperature of current.

From feature selection, 5 features from AIS data are discarded while 11 features are removed from the weather data. To predict the ship speed, The SOG will be selected as the label to train the model. The remaining attributes will be selected as training features. This is summarised in Table 3.

3.3 Modelling

In this section, the modelling of ship speed through SOG using selected features will be performed using tree-based regressor model. The tree-based regressor model con-

sidered are decision tree regressor, random forest regressor and extra-tree regressor. In addition, the tree-based models are compared against multiple linear regressor to as benchmark. The methodology to develop the best model is divided into several steps.

For training, the dataset is split into training, validation and test dataset in ratio of 73:18:9. Journey data from the month of June is arbitrarily selected as test dataset. The remaining dataset will be split into training and validation dataset in 80:20 ratio. The explanation of training process and selection of the best model is broken down into several sections. In Section 3.3.2, the tuning parameter of `Scikit-Learn` will be studied extensively as suitable tuning could result in improved model performance.

Appropriate statistical performance measures are applied to each model; the performance measures selected will help to evaluate how well a model is able to make generalisation on validation and test dataset. The evaluation will be cross validated in form of k-folding. The details on evaluation methodology used in this thesis will be discussed in Section 3.3.1.

3.3.1 Performance Metrics for Model Validation

To gain sensible estimate of model performance and how precise a model is, the model will be cross validated by means of k-folding. K-fold cross validation split the training set into k subsets which is called *folds*, then the model will be trained k times using k-1 subsets and remaining one for validation, this process is illustrated in Figure 10. For each iteration, each model is evaluated using different performance metrics such as (1) Coefficient of Determination (R^2), (2) Explained Variance (EV), (3) Mean Absolute Error (MAE), (4) Root Mean Square (RMSE) and (5) Median Absolute Deviation (MAD). The results from each iteration is then averaged, where the information on model precision can be gained from the standard deviation. Performing k-fold cross validation checks model robustness against different datasets. The properties of each performance metric will be discussed in the following sections.

3.3.1.1 Coefficient of Determination (R^2)

The coefficient of determination R^2 gives a measure on prediction quality, R^2 quantifies the ability of the regression model to approximate the actual values. R^2 is defined by Equation (3.9), where y represents true target output, $\hat{y}$ represents the predictor output and $\overline{y}$ represents the mean. R^2 score range between 0 and 1, higher values i.e. $R^2 \to 1$ indicate better model fit and score of 1 indicate perfect prediction.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_\mathrm{i} - \hat{y}_\mathrm{i})^2}{\sum_{i=1}^{n}(y_\mathrm{i} - \overline{y}_\mathrm{i})^2} \quad \textbf{where} \quad \overline{y} = \frac{1}{n}\sum_{1}^{n} y_\mathrm{i} \tag{3.9}$$

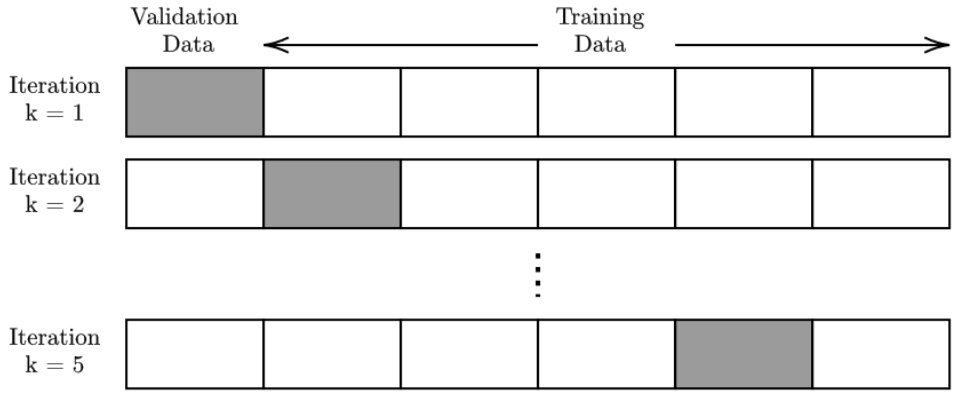

Figure 10: Visual illustration of k-folding, Grey shaded box represents the validation data while white box represents the training data

3.3.1.2 Explained Variance (EV)

Explained variance indicate how well a model can capture variance from a dataset. It is defined by Equation (3.10), where σ_x represents standard deviation of parameter x. EV score range between 0 and 1, where the best score of $EV = 1$ can be obtained if $\sigma^2_{(y-\hat{y})} \to 0$.

$$EV(y, \hat{y}) = 1 - \frac{\sigma^2_{(y-\hat{y})}}{\sigma^2_y} \tag{3.10}$$

3.3.1.3 Mean Absolute Error (MAE)

MAE indicated the expected value of absolute (L^1 norm) error, and it can be calculated by:

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_\mathrm{i} - \hat{y}_\mathrm{i}| \tag{3.11}$$

3.3.1.4 Root Mean Square Error (RMSE)

The RMSE describe the expected value of quadratic error. RMSE place large penalty on large deviation between true and estimated values and for this reason, it can be used to as a metric to indicate model performance against outliers. Ideal score is observed when RMSE $\to 0$. RMSE can be considered as absolute measure of model fitness. Omitting the root term, RMSE becomes MSE, which is the loss function of Equation (2.2) that is used to determine the most optimal split in a regression decision tree.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.12}$$

3.3.1.5 Median Absolute Deviation (MAD)

MAD is a performance metrics that considers the median of the absolute errors. It is robust to outlier as it only consider median performance

$$MAD(y, \hat{y}) = \text{median}(|y_1 - \hat{y}_1|, \ldots, |y_i - \hat{y}_i|) \tag{3.13}$$

3.3.2 Model Hyperparameter Optimisation

The subject of parameter tuning was briefly discussed in Section 2.2.1. In Section 2.2.1 parameter tuning was applied to decision tree regressor to avoid overfitting by changing the minimum amount of samples a leaf node has. This example implies that altering model hyperparameter will affect the model performance. However, the optimisation of the hyperparameter cannot be performed *a priori* and as such iterative process will be performed until best hyperparameter value is found.

`Scikit-Learn` offers `GridSearchCV` and `RandomizedSearchCV` to help search for the most optimal hyperparameter. Both solutions operate with similar principle: The selected hyperparameters to be tuned with its value range is evaluated using cross validation to evaluate the best possible combination between the selected hyperparameters. The difference between `GridSearchCV` and `RandomizedSearchCV` lies in how it searches for the best value for the selected hyperparameters: `GridSearchCV` involves construction of grids containing all possible combinations of hyperparameter value in specified range.`RandomizedSearchCV` randomly samples hyperparameter values.

The exhaustive nature of `GridSearchCV` means that it is computationally costly to perform, especially when there are multiple hyperparameters to be considered and value search space is large. `RandomizedSearchCV` gives more control to computing budget by setting the number of iteration and usually produces more accurate results than `GridSearchCV` approach. [12, 39].

For this reason, the `RandomizedSearchCV` will be employed to search for best possible hyperparameter. However, the limitation of *a priori* knowledge of hyperparameter value still exists. In spite of `RandomizedSearchCV` ability to control the computational budget, it is still takes considerable time to obtain the best hyperparameter value. The computational budget may be spent on searches in unpromising search space. With that, initial exploration on the effect of each hyperparameter on model performance will be performed to give better overview on which search space that should be considered during hyperparameter optimisation. In the next subsections, the effect of tunable hyperparameter of tree-based model from `Scikit-Learn` will

be explored to give baseline numbers for the search space. RMSE is used as performance metrics as the hyperparameter parameter optimisation done in this thesis aims to reduce the error during prediction.

3.3.2.1 Number of features

Defined with default value as `max_features=None` in `Scikit-Learn`. This hyperparameter controls the number of features to be considered when looking for the best split, the default `None` option means it will consider all features. This parameter tuning is available for Decision Tree Regressor, Random Forest Regressor and Extra-Tree Regressor. Initial exploration indicated Random Forest Regressor and Extra Tree Regressor benefit from considering more features, Decision Tree Regressor requires further fine-tuning to optimise the model as the default `None` means it will consider all features when searching for best split.

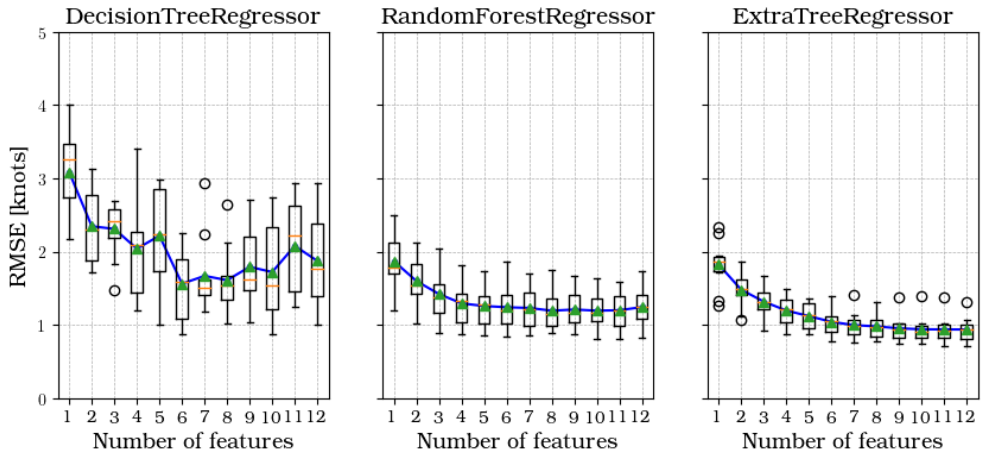

Figure 11: Hyperparameter tuning of `max_features`

3.3.2.2 Number of sample in a leaf node

Defined with default value as `min_samples_leaf=1` in `Scikit-Learn`. This parameter controls number of samples required to be at leaf node, where split point will be considered if the leaf contains at least `min_samples_leaf=n` training samples in each left and right branch. As shown in Figure 4, tuning this hyperparameter to higher values helps to smoothen the model and avoid overfitting. However, this may lead to underfitting as the model is unable to capture the trend within the data. This is supported by the findings shown in Figure 12, the DTR benefits from regularisation at certain breakeven point, in this case, it is found to be at `min_samples_leaf=4`. But after this breakeven point, the model's performance degrades. It is also observed that RFR and ETR does not benefit from any form of regularisation.

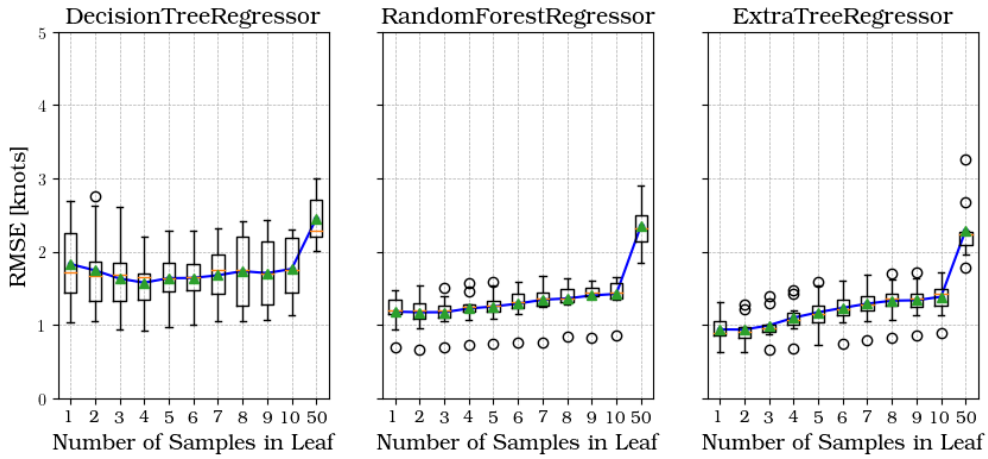

Figure 12: Hyperparameter tuning of `min_samples_leaf`

3.3.2.3 Depth of Tree

Defined with default value as `max_depth=None` in `Scikit-Learn`. This hyperparameter controls the growth of the tree. Leaving it at `max_depth=None` means the tree will grow until all leaves are pure i.e. until minimum MSE is obtained or when the number of samples is less than the minimum number of samples required to split an internal node. Similar to `min_samples_leaf`, DTR shows improvement until a certain breakeven point. RFR performance seems to stabilise at certain depth while ETR benefits from allowing full growth of the tree. The results are summarised in Figure 13

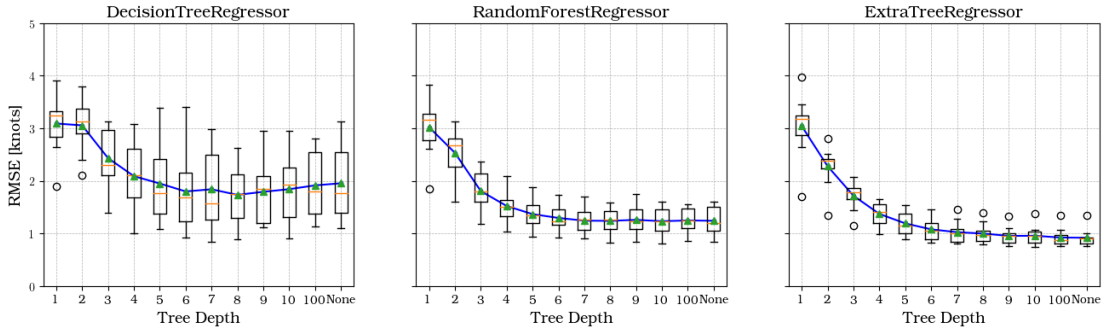

Figure 13: Hyperparameter tuning of `max_depth`

3.3.2.4 Number of Trees

Defined with default value as `n_estimators=100`. This hyperparameter controls the amount of trees i.e. predictors in a forest. Tuning of number of trees will have an effect on the training time and it is only available to RFR and ETR. The default value seems to yield satisfactory result, as the performance for both RFR and ETR stabilise after in this case stabilise after 100 trees, as seen in Figure 11.

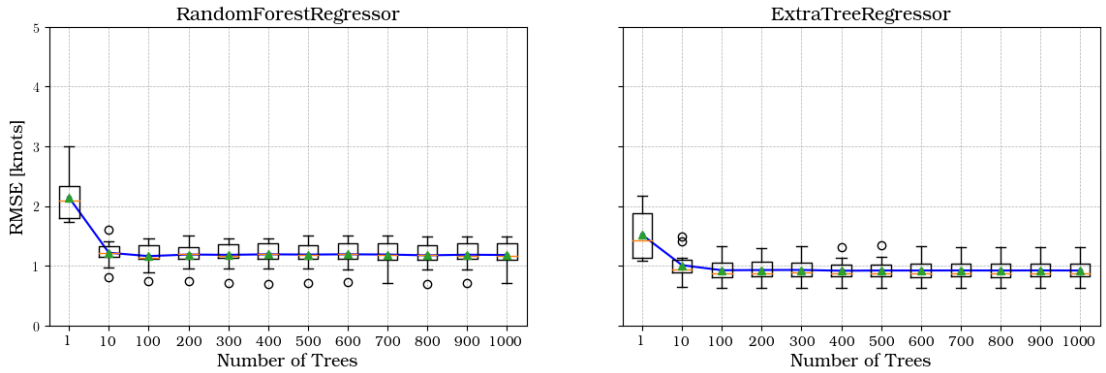

Figure 14: Hyperparameter tuning of `n_estimators`

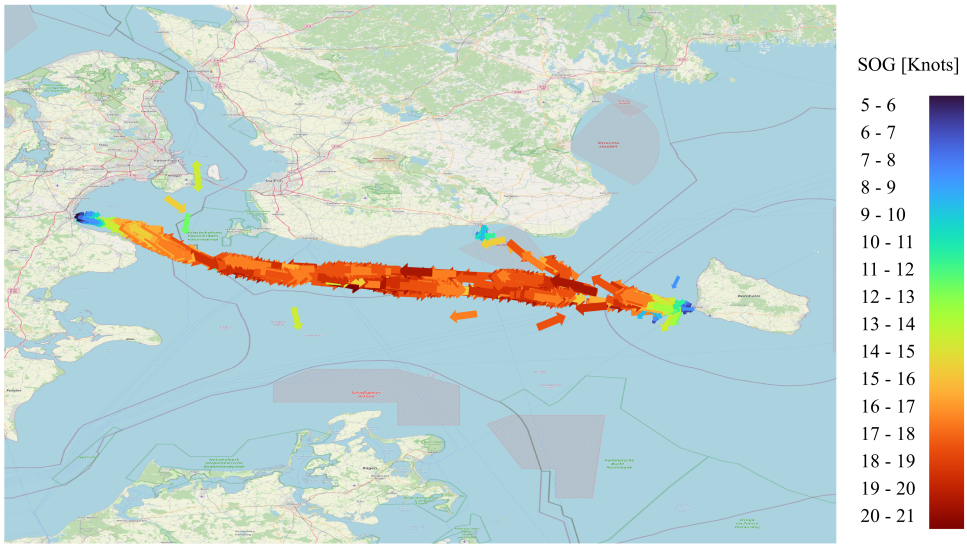

Figure 15: Journey of the ship in June

3.3.3 Methodology Application

- Two data sources are imported. `AIS_weather_H_ok2_copy.csv` and `AIS_weather_h_rename_copy.csv`. The information from the latter comma delimited file will be used for calculating the ship Speed Through Water (STW). The information required is the true north current direction. Which is obtained from the vector component of the Northward and Southward current.
- This dataframe will be merged with the main dataframe from the file `AIS_weather_H_ok2_copy.csv`.
- Omission of the journey data between Ronne and Sassnitz
- SOG threshold is applied to omit ship mooring and maneuvering to accurately represent the ship's steady state operation [4, 9, 10, 33]. This threshold is selected as 5 knots according to [10]

- The AIS data from June is filtered. This data will be used as validation data to check the model's performance.

3.4 Data Analysis

- The features are represented in a histogram plot. For the feature Current speed, anomaly is detected. Certain spike is detected around $0.01 - 0.03$ m/s. Reasons unknown. The data is retained, including the spike, until a definitive answer can be found.

- OPEN QUESTION : What is the necessity of feature standardization / normalization ? Normalization is required for ANN as model training requires the value between 0 and 1. But in case of RFR, there is no such requirement. Through testing, data standardization also does not seem to improve the model's performance.

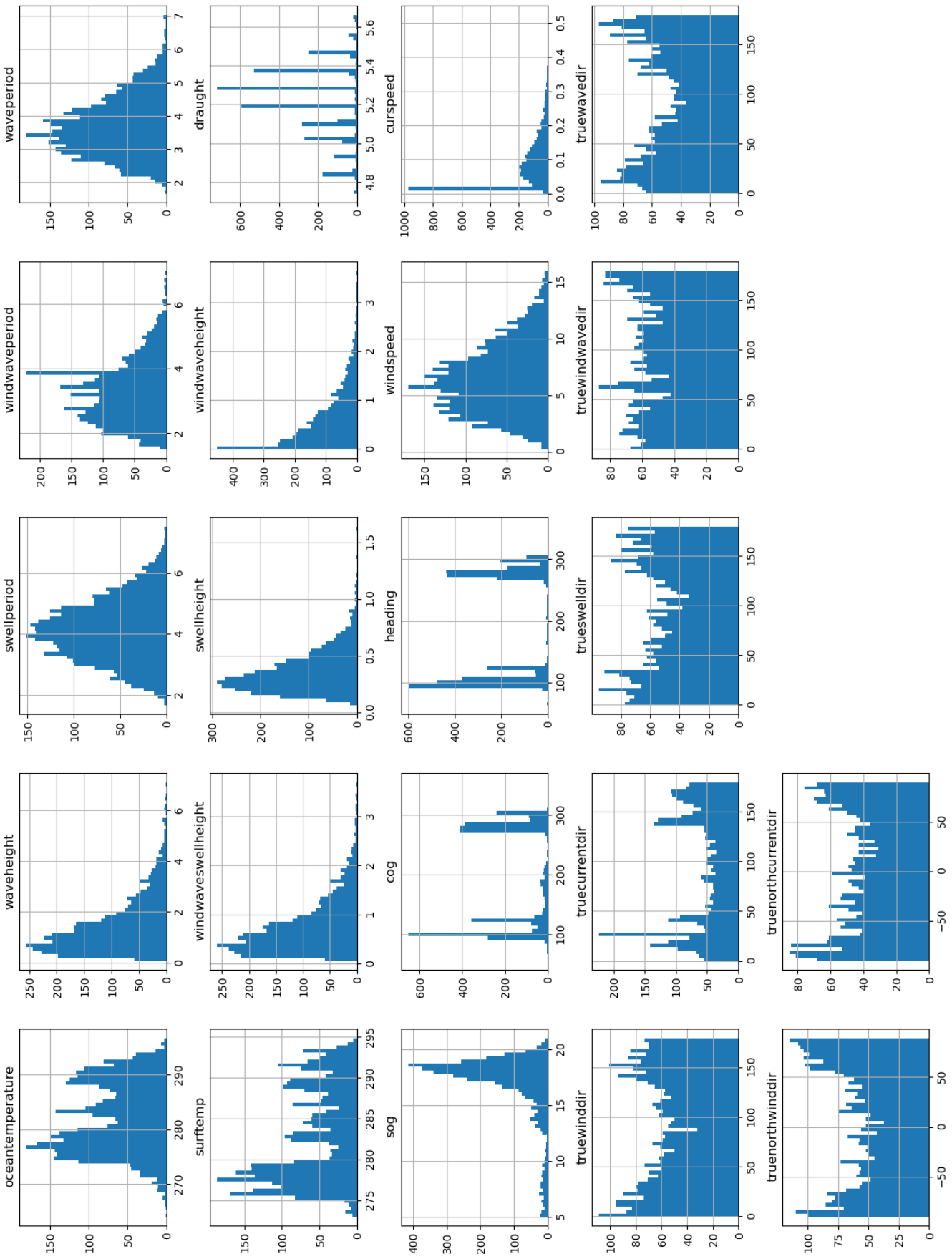

Figure 16: Histogram of the features

- The correlation of the features against SOG are determined. It is found that :
  - Draught
  - Course Over Ground (COG)
  - heading
  - Wind Speed
  - Current Speed
  - True Current direction

  Have relatively stronger correlation to SOG compared to other features, albeit the correlation is a weak one
- The correlation between the features is displayed using the following the heat map. From the heat map it can be observed that between these features:
  - Waveheight and wind wave swell height
  - Waveheight and wind wave height
  - Windwaveswellheight and wave period

  Have a strong correlation between each other.
- Open topic:
  - Feature reduction is possible, [10] suggested high feature correlation filter, the filter suggest that two features which has a high correlation (> 90%) is to be combined into a single feature. But the author is unsure whether this combination is physically sensible. Hence, this filter is yet to be applied for feature reduction.
  - Some of these features can be connected through wave equations, but the author has not found an equation which could relate these features.
- The random forest regressor could not function when `NaN` values are present. With that, the missing values are filled in using the `imputer` function. The missing values are filled in by means of `KNN`.

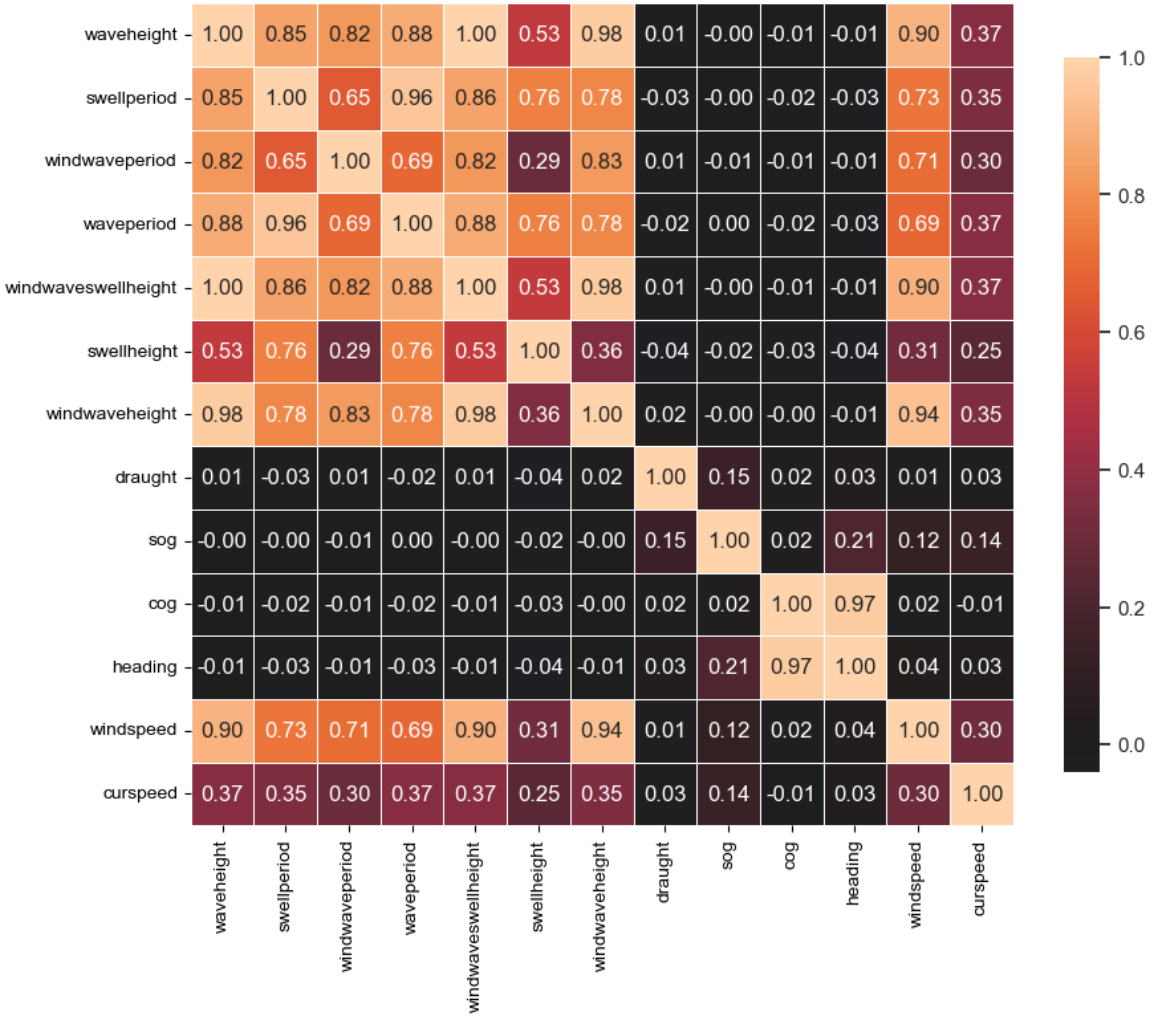

Figure 17: Correlation Heat Map

3.5 Modelling

- The data is split into 80:20 ratio. But considering the validation data, it is split into approximately 73:18:9.
- The model is then trained using Random Forest Regression (RFR). Additional training is also performed using Decision Tree Regressor (DTR). DTR model performance will be used as a benchmark as it is also a tree-based modelling method with similar methodology to RFR.
- The computational time of DTR is significantly faster than RFR Model Evaluation

3.6 Predicting STW

- The ship's Speed Through Water STW can be calculated using vector component of the SOG and current speed. The direction used will be according to True North. [33, 40]
- SOG represents the speed of the ship with reference to the ground, while the STW represent the ship's speed with reference to water.
- SOG also can be termed by the ship's speed that is captured by the GPS, and does not consider any effect of the current
- This means that the ship's STW will be greater than the ship's SOG when there is current moving against the ship's movement direction and vice versa
- The vector decomposition can be defined from the following equations, which is based on the equation by [33]:
  - The ship's SOG V_g can be decomposed into V_g^x and V_g^y, which represents the x and y components of the SOG respectively using the ship's course heading (COG) β *with respect to True North*:

$$V_g^x = V_g \sin(\beta) \tag{3.14}$$

$$V_g^y = V_g \cos(\beta) \tag{3.15}$$

  - To consider the effect of sea current. The current speed V_c will also be decomposed to x and y components respectively using the current direction γ *with respect to True North*:

$$V_c^x = V_g \sin(\gamma) \tag{3.16}$$

$$V_c^y = V_g \cos(\gamma) \tag{3.17}$$

  - from here the ship' STW V_{wx} and V_{wy} component can be found from the following equation:

$$V_w^x = V_g^x - V_c^x \tag{3.18}$$

$$V_w^y = V_g^y - V_c^y \tag{3.19}$$

  - The magnitude of the STW can be readily obtained from the following vector synthesis

$$V_w = \sqrt{(V_w^x)^2 + (V_w^y)^2} \tag{3.20}$$

- This principle is applied to the following Python script. 3.16

```
1
2         # Convert SOG from [Knots] to [m/s]
3
4         dfprog["vgms"] = dfprog["sog_pred"]/1.9438
5
6         # Convert the angles from [Degrees] to [Radians]
7
8         rad_gamma = np.deg2rad(dfprog["gamma"])
9         rad_cog = np.deg2rad(dfprog["cog"])
10
11        # Decomposition in x-component
12
13        dfprog["vgx"] = dfprog["vgms"] * np.sin(rad_cog)
14        dfprog["vcx"] = dfprog["curspeed"] * np.sin(rad_gamma)
15        dfprog["stw_x"] = (dfprog["vgx"] - dfprog["vcx"])
16
17        # Decomposition in y-component
18
19        dfprog["vgy"] = dfprog["vgms"] * np.cos(rad_cog)
20        dfprog["vcy"] = dfprog["curspeed"] * np.cos(rad_gamma)
21        dfprog["stw_y"] = (dfprog["vgy"] - dfprog["vcy"])
22
23        # Vector synthesis and reconversion to [Knots] from [m/s]
24
25        dfprog["vwms_p"] = np.sqrt(dfprog["stw_x"]**2 + dfprog["stw_y"]**2)
26        dfprog["stw_pred"] = dfprog["vwms_p"]*1.9438
27
28
```

4 Result and Discussion

The result of the research is discussed in this chapter. This comprises model validation and how different statistical metrics are used to analyze the model's performance.

4.1 Model Evaluation

The model are tested against four metrics, namely:

- R^2 : Indicate model fit. Best Score = 1
- Explained Variance `EV` : Indicate amount of variance in model. Best Score = 1
- Mean Absolute Error `MAE` : Indicate how much error a model makes in its prediction. Best Score = 0
- Root Mean Square Error `RMSE` : Same as MAE, more sensitive to outlier. Best Score = 0
- Median Absolute Error `MAD` : Check robustness against outlier. Best Score = 1

The result is summarized in the following table

Model	RFR	DTR	LR
R^2	0.9328181446941499	0.8526085810220092	1
EV	0.932872958708872	0.8526260247615258	2
MAE	0.5546347329650284	0.8108982427834758	3
RMSE	0.7095480848510665	1.5566896535262504	4
MAD	0.38484635910000087	0.5475717149999983	5

Table 4: Model performance

Model	RFR	DTR	LR
R^2	0.9328181446941499	0.8526085810220092	1
EV	0.932872958708872	0.8526260247615258	2
MAE	0.5546347329650284	0.8108982427834758	3
RMSE	0.7095480848510665	1.5566896535262504	4
MAD	0.38484635910000087	0.5475717149999983	5

Table 5: Model performance

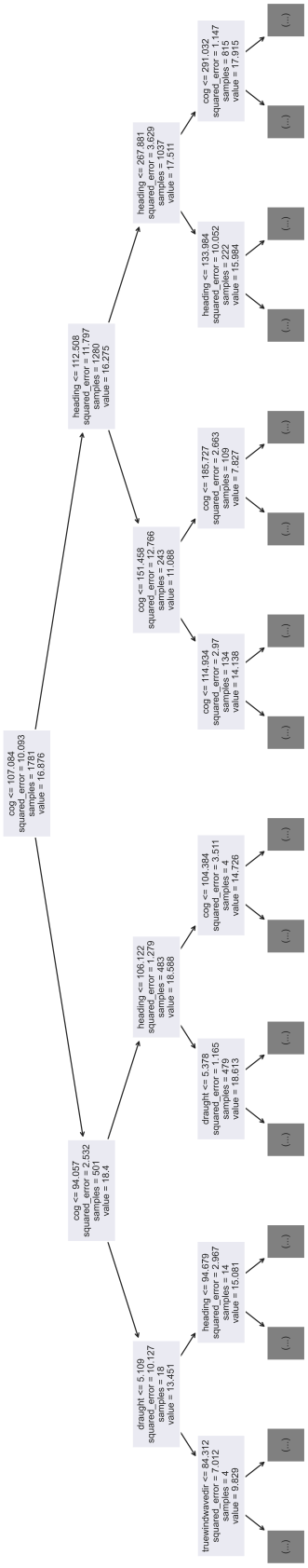

Figure 18: Correlation Heat Map

5 Summary and Outlook

In this chapter the summary of this research will be discussed. This section includes reflections of the research process and presents any possible suggestions and recommendations in this line of research. This chapter concludes this thesis.

References

[1] D. Ronen. The effect of oil price on containership speed and fleet size. *Journal of the Operational Research Society*, 62(1):211–216, 2011. doi:10.1057/jors.2009.169. 5, 6, 10

[2] Stopford. The organization of the shipping market. page 47, 2009. 5

[3] Xiaohe Li, Yuquan Du, Yanyu Chen, Son Nguyen, Wei Zhang, Alessandro Schönborn, and Zhuo Sun. Data fusion and machine learning for ship fuel efficiency modeling: Part i – voyage report data and meteorological data. *Communications in Transportation Research*, 2:100074, 2022. doi:10.1016/j.commtr.2022.100074. 5, 9

[4] E. Bal Beşikçi, O. Arslan, O. Turan, and A. I. Ölçer. An artificial neural network based decision support system for energy efficient ship operations. *Computers & Operations Research*, 66:393–401, 2016. doi:10.1016/j.cor.2015.04.004. 5, 31

[5] N. Wijnolst, Tor Wergeland, and Kai Levander. *Shipping Innovation*. IOS Press, 2009. 5

[6] David Ronen. The effect of oil price on the optimal speed of ships. *The Journal of the Operational Research Society*, 33(11):1035, 1982. doi:10.2307/2581518. 5, 6, 10

[7] Shuaian Wang and Qiang Meng. Sailing speed optimization for container ships in a liner shipping network. *Transportation Research Part E: Logistics and Transportation Review*, 48(3):701–714, 2012. URL: https://www.sciencedirect.com/science/article/pii/S1366554511001554, doi:10.1016/j.tre.2011.12.003. 5

[8] Miyeon Jeon, Yoojeong Noh, Yongwoo Shin, O-Kaung Lim, Inwon Lee, and Daeseung Cho. Prediction of ship fuel consumption by using an artificial neural network. *Journal of Mechanical Science and Technology*, 32(12):5785–5796, 2018. URL: https://link.springer.com/article/10.1007/s12206-018-1126-4, doi:10.1007/s12206-018-1126-4. 5

[9] Christos Gkerekos, Iraklis Lazakis, and Gerasimos Theotokatos. Machine learning models for predicting ship main engine fuel oil consumption: A comparative study. *Ocean Engineering*, 188:106282, 2019. doi:10.1016/j.oceaneng.2019.106282. 5, 9, 31

[10] Misganaw Abebe, Yongwoo Shin, Yoojeong Noh, Sangbong Lee, and Inwon Lee. Machine learning approaches for ship speed prediction towards energy efficient shipping. *Applied Sciences*, 10(7):2325, 2020. doi:10.3390/app10072325. 5, 10, 22, 23, 31, 34

[11] Young-Rong Kim, Min Jung, and Jun-Bum Park. Development of a fuel consumption prediction model based on machine learning using ship in-service data. *Journal of Marine Science and Engineering*, 9(2):137, 2021. doi:10.3390/jmse9020137. 5

[12] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems / Aurélien Géron*. O'Reilly, Sebastopol, CA, second edition edition, 2019. 4, 5, 10, 11, 12, 13, 14, 28

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL: http://jmlr.org/papers/v12/pedregosa11a.html. 5, 11, 22

[14] Ki-Su Kim and Myung-Il Roh. Iso 15016:2015-based method for estimating the fuel oil consumption of a ship. *Journal of Marine Science and Engineering*, 8(10):791, 2020. doi:10.3390/jmse8100791. 6

[15] Stian Glomvik Rakke. *Ship emissions calculation from AIS*. PhD thesis, NTNU. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2410741. 6, 9, 14, 15

[16] Ran Yan, Shuaian Wang, and Harilaos N. Psaraftis. Data analytics for fuel consumption management in maritime transportation: Status and perspectives. *Transportation Research Part E: Logistics and Transportation Review*, 155:102489, 2021. URL: https://www.sciencedirect.com/science/article/pii/S1366554521002519, doi:10.1016/j.tre.2021.102489. 8, 9

[17] Michael Haranen, Pekka Pakkanen, Risto Kariranta, and Jouni Salo. White, grey and black-box modelling in ship performance evaluation. 2016. URL: https://www.researchgate.net/publication/301355727_White_Grey_and_Black-Box_Modelling_in_Ship_Performance_Evaluation. 8, 11, 14

[18] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. Routledge, 2017. URL: https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman, doi:10.1201/9781315139470. 8, 11

[19] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. URL: https://link.springer.com/article/10.1023/a:1010933404324, doi:10.1023/A:1010933404324. 8, 13

[20] T. W. P. Smith, J. P. Jalkanen, B. A. Anderson, J. J. Corbett, J. Faber, S. Hanayama, E. O'Keeffe, S. Parker, L. Johansson, L. Aldous, C. Raucci, M. Traut, S. Ettinger, D. Nelissen, D. S. Lee, S. Ng, A. Agrawal, J. J. Winebrake, M. Hoen, S. Chesworth, and A. Pandey. Third imo greenhouse gas study 2014. 2015. URL: https://research.manchester.ac.uk/en/publications/third-imo-greenhouse-gas-study-2014. 9

[21] Yuanqiao Wen, Xiaoqiao Geng, Lichuan Wu, Tsz Leung Yip, Liang Huang, and Dingyong Wu. Green routing design in short seas. *International Journal of Shipping and Transport Logistics*, 9(3):371, 2017. doi:10.1504/IJSTL.2017.083474. 9

[22] Seong-Hoon Kim, Myung-Il Roh, Min-Jae Oh, Sung-Woo Park, and In-Il Kim. Estimation of ship operational efficiency from ais data using big data technology. *International Journal of Naval Architecture and Ocean Engineering*, 12:440–454, 2020. URL: https://www.sciencedirect.com/science/article/pii/S2092678220300091, doi:10.1016/j.ijnaoe.2020.03.007. 9, 15

[23] Omer Soner, Emre Akyuz, and Metin Celik. Use of tree based methods in ship performance monitoring under operating conditions. *Ocean Engineering*, 166:302–310, 2018. URL: https://www.sciencedirect.com/science/article/pii/S0029801818314446, doi:10.1016/j.oceaneng.2018.07.061. 9

[24] Ran Yan, Shuaian Wang, and Yuquan Du. Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship. *Transportation Research Part E: Logistics and Transportation Review*, 138:101930, 2020. doi:10.1016/j.tre.2020.101930. 9

[25] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: Data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman*. Springer series in statistics. Springer, New York, 2nd ed. edition, 2009. doi:10.1007/b94608. 4, 10, 11, 12, 13, 14

[26] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer, New York, 2013. 12, 13, 14

[27] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pages 278–282 vol.1, 1995. doi:10.1109/ICDAR.1995.598994. 13

[28] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. URL: https://link.springer.com/article/10.1007/bf00058655, doi:10.1007/BF00058655. 13

[29] Michael Affenzeller, Bogdan Burlacu, Viktoria Dorfer, Sebastian Dorl, Gerhard Halmerbauer, Tilman Königswieser, Michael Kommenda, Julia Vetter, and Stephan Winkler. White box vs. black box modeling: On the performance of

deep learning, random forests, and symbolic regression in solving regression problems. pages 288–295. Springer, Cham, 2020. URL: https://link.springer.com/chapter/10.1007/978-3-030-45093-9_35, doi:10.1007/978-3-030-45093-9{\textunderscore}35. 14

[30] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. URL: https://link.springer.com/article/10.1007/s10994-006-6226-1, doi:10.1007/s10994-006-6226-1. 14

[31] International Maritime Organization. Revised guidelines for the onboard operational use of shipborne automatic identification systems (ais), 25/06/2023. URL: https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx. 3, 14, 15, 16

[32] Yang Zhou, Winnie Daamen, Tiedo Vellinga, and Serge Hoogendoorn. *AIS data analysis for the impacts of wind and current on ship behavior in straight waterways*. 2017. 15

[33] Liqian Yang, Gang Chen, Jinlou Zhao, and Niels Gorm Malý Rytter. Ship speed optimization considering ocean currents to enhance environmental sustainability in maritime shipping. *Sustainability*, 12(9):3649, 2020. doi:10.3390/su12093649. 15, 31, 36

[34] Danish Maritime Authority. Safety at sea, navigational information, ais data, 25/06/2023. URL: https://dma.dk/safety-at-sea/navigational-information/ais-data. 4, 21

[35] Leo H. Holthuijsen. *Waves in oceanic and coastal waters*. Cambridge University Press, Cambridge, 2007. 23

[36] Charles L. Bretschneider. *Generation of waves by wind. State of the art*. 1965. URL: https://apps.dtic.mil/sti/citations/ad0612006. 4, 23

[37] Elzbieta M. Bitner-Gregersen. Joint probabilistic description for combined seas. In *24th International Conference on Offshore Mechanics and Arctic Engineering: Volume 2*, pages 169–180. ASMEDC, 2005. doi:10.1115/OMAE2005-67382. 24

[38] K. Torsethaugen, S. Haver, and S. Norway. Simplified double peak spectral model for ocean waves. 2004. URL: https://www.semanticscholar.org/paper/Simplified-Double-Peak-Spectral-Model-For-Ocean-Torsethaugen-Haver/0f1b1509791d441861ff6c2940dd13b1f939f149. 25

[39] J. Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 2012. URL: https://www.semanticscholar.org/paper/

Random-Search-for-Hyper-Parameter-Optimization-Bergstra-Bengio/188e247506ad992b8bc62d6c74789e89891a984f. 28

[40] Yang Zhou, Winnie Daamen, Tiedo Vellinga, and Serge P. Hoogendoorn. Impacts of wind and current on ship behavior in ports and waterways: A quantitative analysis based on ais data. *Ocean Engineering*, 213:107774, 2020. doi:10.1016/j.oceaneng.2020.107774. 36

Declaration in lieu of oath

I hereby solemnly declare that I have independently completed this work or, in the case of group work, the part of the work that I have marked accordingly. I have not made use of the unauthorised assistance of third parties. Furthermore, I have used only the stated sources or aids and I have referenced all statements (particularly quotations) that I have adopted from the sources I have used verbatim or in essence.
I declare that the version of the work I have submitted in digital form is identical to the printed copies submitted.
I am aware that, in the case of an examination offence, the relevant assessment will be marked as 'insufficient' (5.0). In addition, an examination offence may be punishable as an administrative offence (Ordnungswidrigkeit) with a fine of up to €50,000. In cases of multiple or otherwise serious examination offences, I may also be removed from the register of students.
I am aware that the examiner and/or the Examination Board may use relevant software or other electronic aids in order to establish an examination offence has occurred
I solemnly declare that I have made the previous statements to the best of my knowledge and belief and that these statements are true and I have not concealed anything.
I am aware of the potential punishments for a false declaration in lieu of oath and in particular of the penalties set out in Sections 156 and 161 of the German Criminal Code (Strafgesetzbuch; StGB), which I have been specifically referred to.

Section 156 False declaration in lieu of an oath
Whoever falsely makes a declaration in lieu of an oath before an authority which is competent to administer such declarations or falsely testifies whilst referring to such a declaration incurs a penalty of imprisonment for a term not exceeding three years or a fine.

Section 161 Negligent false oath; negligent false declaration in lieu of oath
(1) Whoever commits one of the offences referred to in Sections 154 to 156 by negligence incurs a penalty of imprisonment for a term not exceeding one year or a fine. (2) No penalty is incurred if the offender corrects the false statement in time. The provisions of Section 158 (2) and (3) apply accordingly.

Place,date	Signature