POLISH MARITIME RESEARCH 1 (109) 2021 Vol. 28; pp. 83-96
10.2478/pomr-2021-0008

# MODELLING SHIPS MAIN AND AUXILIARY ENGINE POWERS WITH REGRESSION-BASED MACHINE LEARNING ALGORITHMS

**Fatih Okumuş** *
**Araks Ekmekçioğlu**
**Selin Soner Kara**
Yildiz Technical University, Istanbul, Turkey

* Corresponding author: *hfatihokumus@gmail.com (F.Okumus)*

## ABSTRACT

*Based on data from seven different ship types, this paper provides mathematical relationships that allow us to estimate the main and auxiliary engine power of new ships. With these mathematical relationships we can estimate the power of the engine based on the ship's length (L), gross tonnage (GT) and age. We developed these approaches using simple linear regression, polynomial regression, K-nearest neighbours (KNN) regression and gradient boosting machine (GBM) regression algorithms. The relationships presented here have a practical application: during the pre-parametric design of new ships, our mathematical relationships can be used to estimate the power of the engines so that more environmentally friendly ships may be built. In addition, with the machine learning methodology, the prediction of the main engine (ME) and auxiliary engine (AE) powers used in the numerical calculation of ship-based emissions provides data for researchers working on emission calculations. We conclude that the GBM regression algorithm provides more accurate solutions to estimate the main and auxiliary engine power of a ship than other algorithms used in the study.*

**Keywords:** : machine learning, regression, ship emissions, engine power, prediction

## INTRODUCTION

Machine learning is a system that investigates the work and construction of algorithms that can make predictions by making inferences using mathematical and statistical methods from the available data. In machine learning, which is a sub-discipline of artificial intelligence, the algorithms work by building a model to make predictions from sample inputs with the help of computers and software.

The effects of ship-based exhaust emissions include reduction in the air quality, especially in a country's inland waters, the straits, and port areas. These emissions and greenhouse gases are also among the major factors causing global climate change. Two approaches stand out in the literature to estimate emissions from ships. One is the top-down approach, which is based on the fuel consumed by the ship, and the other is the bottom-up approach, which uses the ship's main and auxiliary machinery forces, based on the manoeuvre, cruise, and port activities.

Looking at the research on the application of machine learning in the maritime industry in the literature, Ekmekçioğlu et al. [1] calculated the exhaust emissions from ships arriving at four major ports of Turkey for a year using real numerical data such as the main engine power and speed, auxiliary engine power, and the duration of stay in port. In his study, Trozzi [2] proposed a function based on the ship type and gross tonnage in calculation of the ship's main engine power. He used non-linear regression for ship-based emission calculation. He also proposed the estimated average vessel ratios of the auxiliary engines / main engines by ship type. Yan et al. [3] proposed a two-stage fuel consumption prediction and fuel reduction model for a dry bulk ship. In the first stage, they created a fuel consumption prediction model that takes into account the ship's sailing speed, cargo weight, sea and weather conditions by using the random forest regression. In the second stage, they developed a speed optimisation model based on the prediction model proposed in the first stage. They concluded that the proposed model could reduce the ship's fuel consumption by 2–7% and this reduction would also lead to lower $CO_2$ emissions. Huang et al. [4] calculated ship exhaust emissions using the activity-based STEAM (Ship Traffic Emissions Assessment Model) method. They used machine learning (80% training set, 20% test set) and the polynomial regression method to calculate the value of the unknown main engine power according to the ship's dimensions. Tran [5] emphasised the effect of fuel consumption on $CO_2$ emissions and used fuzzy clustering to examine the effect of loading a bulk carrier, which he took as a case study, on the fuel consumption. He concluded that the novel methodology showed that machine learning could be used to make decisions for the optimum loading of the ship, in the study where parameters such as wind speed, wave height, ship speed, distance travelled, and shaft speed were analysed.

In their studies, Yan et al. [6] applied big data analysis by considering environmental factors to optimise the engine speeds of inland ships. They proposed a distributed parallel k-means analysis for clustering environmental factors into multiple groups and a model to optimise ships' energy efficiency. They conducted a case study to verify their method on the Yangtze River, and concluded that the method they developed could increase ships' energy conservation and emission reduction. Cepowski [7] used the ship's speed and deadweight or TEU capacity properties to estimate the total machine power of bulk carriers and container vessels. Requia et al. [8] estimated and compared $PM_{2.5}$ components with ordinary kriging (OK) interpolation, hybrid interpolation and machine learning (forest-based regression) methods. They concluded that the forest model offers the best performance because the $R^2$ value is higher than 0.7 for most of the particle components. They stated that their results may be useful for more accurate prediction of $PM_{2.5}$ components in the air. Uyanık et al. [9] performed the fuel consumption optimisation of a container ship with machine learning using multiple linear regression, ridge and lasso regression, support vector regression, tree-based algorithms and boosting algorithms. They compared the prediction models in their studies and they found that parameters such as the main engine rpm, cylinder values, scavenge air and shaft indicators are highly correlated with fuel consumption, and stated that they found the most accurate estimate with multiple regression and ridge regression. Barua et al. [10] explored international freight transportation management through machine learning. They discussed how it is applied in the fields of maritime transport, air cargo and intermodal transport using different machine learning methods such as demand forecasting, operation and asset maintenance, vehicle trajectory and on-time performance prediction. They proposed four directions for future research. Peng et al. [11] estimated the energy consumption of ships in China's Jingtang port and discussed their strategies to reduce energy consumption and proposed prediction models. They used the gradient boosting regression, random forest regression, BP network, linear regression and K-nearest neighbour regression machine learning models and analysed 15 features that have an impact on ships' energy consumption as input. They concluded that net tonnage, deadweight tonnage, actual weight and efficiency of facilities are the four most important features to predict the energy consumption of the ships. Jeong et al. [12] made predictions of time for shipbuilding production processes using machine learning technology. In their study, they analysed data with the R and Phthon programs, they created prediction models and confirmed these models using criteria such as the mean absolute percent error and root mean squared logarithmic error. Gkerekos et al. [13] investigated the effectiveness of different multiple regression algorithms to estimate ships' main engine fuel oil consumption. They considered the noon reports and automatic data logging and monitoring systems for data collection. They compared machine learning regression algorithms such as linear regression, decision tree regressors, random forest regressors, extra trees regressors, support vector regressors, K-nearest neighbours, artificial neural networks and ensemble methods, and stated that the best performance was shown by extra trees regressors and random forest regressors. Jonquais and Krempl [14] used machine learning to make predictions about shipping times between South East Asia and North America. By using the random forest algorithm and creating four models to produce estimates, they created a tool that gives superior results over traditional methods. Bodunov et al. [15] estimated a destination and an estimated time of arrival (ETA) for maritime traffic using a machine learning method using geo-spatial data, random forest, gradient boosting decision trees, XGBoost trees and extremely randomised trees models for destination prediction; they used feed forward neural networks for arrival time estimation. They achieved 97% accuracy in the destination estimate and 90% accuracy in the ETA estimate. In their study, Yuan and Nian [16] emphasised the importance of improving ship energy efficiency and reducing ship emissions, and they developed a Gaussian process metamodel to predict ships' fuel consumption in different scenarios, taking into account the operating and weather conditions such as speed, trim, wind and wave effects. With the case study, they demonstrated the accuracy and effectiveness of using the Gaussian process metamodel for the prediction of ships' energy consumption. Farag and Ölçer [17] stated that fuel consumption is a very important tool in reducing greenhouse gas emissions. They developed an estimation model for the fuel consumption of ships using artificial neural network and multiple regression techniques. Finally, they used the model they developed to estimate the fuel savings that one ship can make during a voyage. Bui-Duy and Vu-Thi-Minh [18] created a deep-based fuel consumption model for the shipping route selection of container ships in Asia. They offered an idea that helped choose the optimal route to minimise fuel costs. They stated that the model, which has five input variables, namely average velocity, sailing time, ship capacity, wind speed and wind direction, has an accuracy of close to 95%. Hao Cui et al. [19] proposed a new machine-learning-based ship design optimisation approach. They used a multi-objective particle swarm optimisation method, multi-agent system and CAE software to build an optimisation system. They conducted a dry cargo vessel design optimisation as a case study to evaluate the conformity of the method they created to the real world. Peker et al. [20] created a model that can predict the heating and cooling load of houses by using machine learning algorithms with a data set with eight input and two output values. They used and compared machine learning algorithms such as support vector machine regression, linear regression, random forest regression and nearest neighbour regression, and concluded that the best predictive success was achieved by the random forest regression algorithm. In this study, the ship length, gross tonnage, and age data were weak in predicting the ship's main power. With KNN regression, the main engine power can be successfully estimated, but the most successful algorithm was the GBM algorithm. Similarly, linear and polynomial regression is not sufficient for predicting auxiliary machine power. While KNN regression received a pass grade, the GBM regression algorithm predicted quite successfully.

## AIM OF RESEARCH

Previous studies were examined according to their methods, inputs, outputs and R$^2$ values and the similarities and differences between this study and other articles were revealed. The comparison with previous studies is shown in Table 1.

*Tab. 1. Comparison with previous studies.*

| Study | Method | Inputs | Outputs | R$^2$ |
|---|---|---|---|---|
| This study | Linear regression | GRT, length, age | ME power | 0.68 |
| This study | Linear regression | ME Power, length, age | AE power | 0.68 |
| This study | Polynomial regression | GRT, length, age | ME power | 0.8 |
| This study | Polynomial regression | ME power, length, age | AE power | 0.69 |
| This study | K-nearest neighbour regression | GRT, length, age | ME power | 0.86 |
| This study | K-nearest neighbour regression | ME power, length, age | AE power | 0.74 |
| This study | Gradient boosting regression | GRT, length, age | ME power | 0.95 |
| This study | Gradient boosting regression | ME power, length, age | AE power | 0.93 |
| Yan et al. [3] | Random forest regression | Sailing speed, cargo weight, weather conditions | Fuel consumption | 0.72 |
| Huang et al. [4] | Polynomial regression | Cargo ships, length, breadth | ME power | 0.91 |
| Huang et al. [4] | Polynomial regression | Tankers, length, breadth | ME power | 0.87 |
| Requia et al. [8] | Forest model | 25 predictors representing land use | PM$_{2.5}$ emission | 0.93 |
| Peng et al. [11] | Random forest regression | 15 features consisting of inherent properties of container ships and external features of ports | Ship energy consumption | 0.94 |
| Peng et al. [11] | Linear regression | 15 features consisting of inherent properties of container ships and external features of ports | Ship energy consumption | 0.77 |
| Peng et al. [11] | K-nearest neighbour regression | 15 features consisting of inherent properties of container ships and external features of ports | Ship energy consumption | 0.62 |
| Peng et al. [11] | Gradient boosting regression | 15 features consisting of inherent properties of container ships and external features of ports | Ship energy consumption | 0.91 |
| Gkerekos et al. [13] | Random forest regression | Load conditions, weather conditions, speed, sailing distance, draft | Ship ME fuel consumption | 0.87 |
| Gkerekos et al. [13] | K-nearest neighbour regression | Load conditions, weather conditions, speed, sailing distance, draft | Ship ME fuel consumption | 0.78 |
| Gkerekos et al. [13] | Boosting | Load conditions, weather conditions, speed, sailing distance, draft | Ship ME fuel consumption | 0.90 |
| Jonquais and Krempl [14] | Random forest regression | Carrier, shipper, route | Shipping times for departure | 0.88 |
| Jonquais and Krempl [14] | Neural networks model | Carrier, shipper, route | Shipping times for departure | 0.85 |
| Farag and Ölçer [17] | Artificial neural network | Speed, depth, wind speed, wave parameters, swell parameters, sea current | Brake power | 0.96 |
| Farag and Ölçer [17] | Artificial neural network | Speed, depth, wind speed, wave parameters, swell parameters, sea current | Fuel consumption | 0.89 |

In this study, we use different machine learning methods and comparisons in order to estimate the main and auxiliary engine powers of the ships, which are necessary for numerical calculation of the emissions of exhaust gas originating from the maritime sector.

## MATERIALS AND METHOD

### MODEL VALIDATION

The accuracy of the model's predictions is calculated by comparing the actual power values of the main and auxiliary engine with the corresponding predicted values. Ten-fold cross-validation was applied to check the model performance. The dataset was randomly divided into 10 parts, train the model on 9 partitions and predict the properties of the remaining set. This process was repeated 10 times for each section. The prediction ability of the model is then evaluated as the average performance of the model in all repetitions. The root mean squared error (*RMSE*), mean absolute error (*MAE*), and R-squared ($R^2$) were used to assess the performance of the developed regression models.

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

As shown above, $y_i$ and $\hat{y}_i$ respectively represent the actual power values and estimated power values. Since the aim of training the model is to reduce the difference between these two values as much as possible, the model with a small *RMSE* value was accepted as superior.

The *MAE* measures the average magnitude of errors in a series of estimates, regardless of their direction. It is the average of the absolute differences between the estimate and the actual observation that all individual differences have equal weight on the test sample. Its analytical expression is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{2}$$

The $R^2$ correlation coefficient is used to evaluate the performance of the models and is given as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \overline{y}_i)} \tag{3}$$

$\overline{y}_i$ represents the mean value of $y_i$. It is a measure showing how close each data point is to the regression line with the $R^2$ value. It is always positive and between 0 and 1.

### DATA SET

In this study, data containing information from 4037 different ships were used. The dataset includes the ship type, gross tonnage, year of manufacture, length, and the main and auxiliary engine power for each ship. While 80% of these data of these ships are used to train the model, 20% of them are used for testing. Samples were taken from seven different ship types: chemical tanker, container, general cargo, LPG tanker, oil product tanker, Ro-Ro ship, and search and rescue ship. The gross tonnage of the ships varies between 74 and 162960. The oldest ship was produced in 1925, while the newest ship was built in 2018. The lengths of the ships were kept in a wide range from 18.25 m to 368 m. The main machine power and auxiliary machine power to be estimated vary in the ranges of 147–72240 kW and 37–9600 kW, respectively. Table 2 provides statistical data on the ships.

*Tab. 2. Statistical data of the data set*

| | Minimum | $1^{St}$. Qu. | Median | Mean | $3^{rd}$. Qu | Maximum |
|---|---|---|---|---|---|---|
| Gross tonnage | 74 | 3505 | 9927 | 21654 | 29982 | 162960 |
| Length | 18.28 | 106.00 | 141.00 | 154.86 | 189.99 | 368.00 |
| ME power | 147 | 1920 | 5400 | 8839 | 10500 | 72240 |
| AE power | 37 | 253 | 500 | 738 | 910 | 9600 |

### DETERMINING THE INDEPENDENT VARIABLES

The separation of resistance components in terms of the scale effect and its first use in model–ship extrapolation was introduced by Froude. In this method, which is today called the Froude hypothesis, total resistance is divided into friction and residual resistance; the friction resistance is assumed to be equal to one equivalent plate resistance in the same area as the ship's wet area, and the difference between the total resistance and friction resistance is defined as the residual resistance. There are various methods for calculating the ship's total resistance and resistance components. CFD (computational fluid dynamics), panel methods, other numerical techniques, model experiments, empirical and statistical approaches are the main methods used in calculation. It is an undeniable fact that the total resistance of ships has improved over time with the research and development studies of researchers on these methods.

Also, the number of ship gas emissions is estimated to be around 450, but the vast majority of these are at a level that can be neglected in terms of both quantity and impact. However, carbon dioxide ($CO_2$), carbon monoxide (CO), nitrogen oxides (NOx), sulphur oxides (SOx), and particulate matter (PM) are the most common gas emissions and have the greatest impact on both human health and the ecosystem. MARPOL (International Convention for the Prevention of Pollution from Ships) is reducing the limits of these harmful emissions to ever more demanding levels. Although various internal combustion engine technologies have been developed to overcome these difficult constraints, it is difficult to do so unless there are efficient ships. For this reason, it would not be a correct approach to ignore the developments that occur over time while estimating the main engine power of the ships.

Also, the NOx emission factors used in the calculations vary according to the shipbuilding year. As a matter of fact, when looking at the results of the relative influence of the model created with the GBM algorithm, it is seen that the age of the ship has an effect amounting to 21.86%. Fig. 1 contains the result of the relative influence of the independent variables.

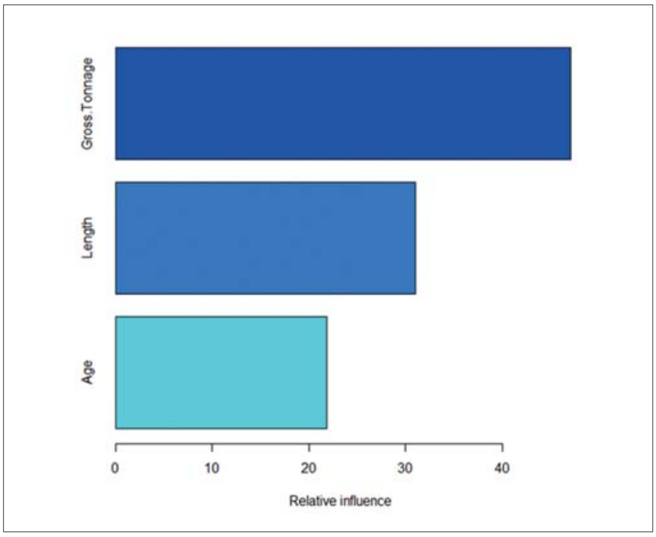

*Fig. 1. Relative influence of the independent variables*

The admiralty coefficient formula is one effective empirical expression that can be used to predict the power curves of ships and is expressed as in Eq. (4). Ships with a similar hull form, speed and displacement have the same admiralty coefficient.

$$P_E = \frac{\Delta^{2/3}.V^3}{C} \tag{4}$$

*Tab. 3. Admiralty coefficient for different ship types [21]*

| Ship type | Admiralty constant |
|---|---|
| General cargo ship | 400±600 |
| Bulker and tanker | 600±750 |
| Reefer | 550±700 |
| Feeder ship | 350±500 |
| Warship | 150 |

In Eq. (4), Δ, V, PE and C are the displacement, velocity, effective power and admiralty coefficient respectively. Table 3 contains the admiralty coefficient suggested by Schneekluth and Bertram for different ships [21]. When Table 3 is examined, the effect of the ship type on the power can be seen clearly. After the effective efficiency is calculated, the main engine power can be calculated using the efficiency of the gear box, the mechanical efficiency of the shaft line, efficiency of the hull, rotation relative efficiency, and open water efficiency of the propeller.

It is seen that two important features of ships have emerged in order to predict the main engine power in ships. The change in the admiralty coefficient of different ship types indicates that ships have different power requirements in relation to their job description. In the study, an independent variable representing the ship type was thus needed to estimate the main engine power. For this reason, it was investigated whether the gross tonnage can represent the ship type and, for this, the gross tonnage length curves were examined depending on the ship type. In Fig. 2, the gross tonnage length distributions of different types of ships in the data set are given. In addition, the curves where the gross tonnage changes depending on the length for the same ship type are shown in Fig. 3, using the available data.

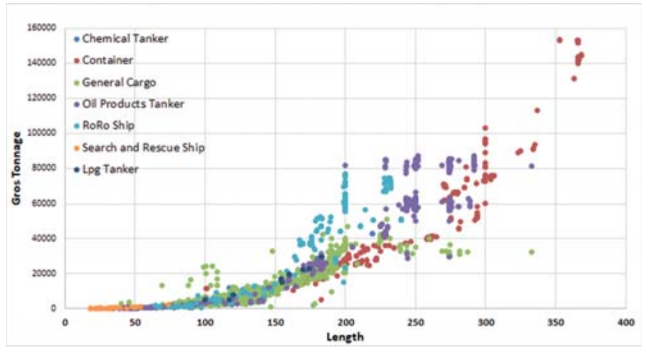

*Fig. 2. Gross tonnage and length distributions of different types of ships*

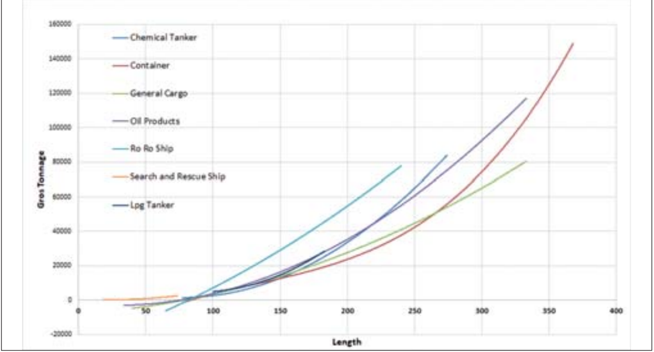

*Fig. 3. Gross tonnage and length curves of different types of ships*

Fig. 3 shows that the gross tonnage value shows different trends for different ship types. There are also supporting empirical statements showing that the gross tonnage and length values of the ships in the data set used differ according to the ship types. Similar to the admiralty coefficient, the empirical statement in Eq. (5) can be used to estimate the gross tonnage. Here, the gross tonnage expression is given as a function of CN (cubic number). The symbol k indicates the coefficient, which varies according to the ship type. In Eq. (6), the explicit expression of CN is given and the symbols $L_{pp}$, B, and D represent the values of the length between the perpendiculars, beam and depth respectively.

$$GT = k.\ CN \tag{5}$$

$$CN = L_{pp}.\ B.\ D \tag{6}$$

When the friction resistance affecting the ships is examined, it is seen that it basically depends on the friction coefficient, the density of the fluid it is in, the wet surface area and the square of the speed. Among these variables, the square of the wet surface area and velocity is directly related to the ship's design parameters. The wet surface area, defined as the area of the surface of the ship in contact with water, is one of the important parameters of the resistance and power calculation. Considering that the wet surface area is also a function of the ship's length, this length is also used as an independent variable.

It would not be right to think that the power of the ship's auxiliary engines is in a linear relationship depending on the ship's main engine power. However, the ship is not completely independent from the main engine power. While calculating the power of auxiliary machinery, many variables such as crew needs and the power requirements of the control systems should be taken into consideration. In this study, the main engine power and ship length were used as an indicator of the size and power needs of the ship to estimate the auxiliary engine power. In addition, the gross tonnage was used to symbolise the special needs of the ship type.

## LINEAR REGRESSION

Linear regression is a method used to model the connection between one or more independent variables and a dependent variable. The main purpose of linear regression is to obtain the function of the relationship between parameters. Creating an appropriate model in the learning process signifies choosing the most appropriate parameters for the hypothesis function by using the training set. The hypothesis function may depend on one or more parameters. Provided that a model based on a single parameter is constructed, it is named as single regression; if it is constructed with two or more parameters, it is named as multiple regression. Single linear regression is formulated as in Eq. (7).

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (7)$$

In Eq. (7), $y$ refers to the value of the dependent variable, $x$ refers to the value of the independent variable, $\beta_0$ is the population's $y$ intercept, $\beta_1$ the slope of the population regression line and $\varepsilon$ a random error term. Similarly, multiple linear regression is expressed as in Eq. (8).

$$y = \beta_0 + \beta_1 x + .... + \beta_k x_k + \varepsilon \quad (8)$$

As distinct from Eq. (7), $k$ represents the number of independent values. In multiple linear regression analysis, the contribution of some of the modelled independent variables to the model may be insignificant. Therefore, it is necessary to identify the independent variables that will explain the dependent variable in the most appropriate way, and remove the insignificant variables from the model. This process is called "variable selection".

Various methods have been developed for independent variable selection. These can be examined as three main groups.

- Forward selection
- Backward elimination
- Standard stepwise regression

For selection of the variables for main engine and auxiliary engine power estimates, the forward selection, backward elimination and standard stepwise selection methods have been applied to determine the contribution of our variables to the model.

According to the simple correlation matrix between the ME power dependent variable and the other independent variables for the ME, the highest correlation coefficient was found. L and GT were the highest independent correlations with ME power. The degree of significance of all independent variables p was examined separately. Then, while the L and GT variables are constant, it is necessary to find the independent variable with the highest partial correlation with ME power. For this reason, the age variable is a candidate to enter the model. Since the significance level of the L, GT and age variables is $p < 0.05$, these variables are included in the model.

According to the simple linear correlation matrix between the AE power dependent variable and other independent variables for the AE, the highest correlation coefficient was found. The independent variables with the highest correlation with AE power were ME power, L, and GT, respectively. The degree of significance of all independent variables p was examined separately. The ME power, L and age variables are included in the model because their p value is less than 0.05. The GT variable was removed from the model because its p value is greater than 0.05.

Within the scope of this study, Model.ME.1 was created to estimate the ship's main engine power. While constructing the model, the length, gross tonnage and age of the ship were used as independent variables. Moreover, the ship's auxiliary engine power was estimated by linear regression, using the main engine power, length and age. The model thus created was named as Model.AE.1. Table 4 contains the errors from the linear model's train and test sets.

*Tab. 4. Error values of the linear model*

| | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | *RMSE* | *R*$^2$ | *MAE* | *RMSE* | *R*$^2$ | *MAE* |
| Model. ME.1 | 6592.29 | 0.688 | 4396.62 | 6257.2 | 0.684 | 4143.66 |
| Model. AE.1 | 448.99 | 0.650 | 251.847 | 430.77 | 0.679 | 246.48 |

## POLYNOMIAL REGRESSION

Independent variables are not continuously required to be in a linear relationship with the dependent variable. As a consequence, the predictive power of the linear model will weaken. In such circumstances, polynomial regression is used. For multiple exponents of the argument, the polynomial model is created as in Eq. (9).

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 +.... + \beta_p x^p + \varepsilon \quad (9)$$

In the equation, expression $p$ refers to the polynomial degree. Polynomial regression can be applied as single or multiple regression as in linear regression.

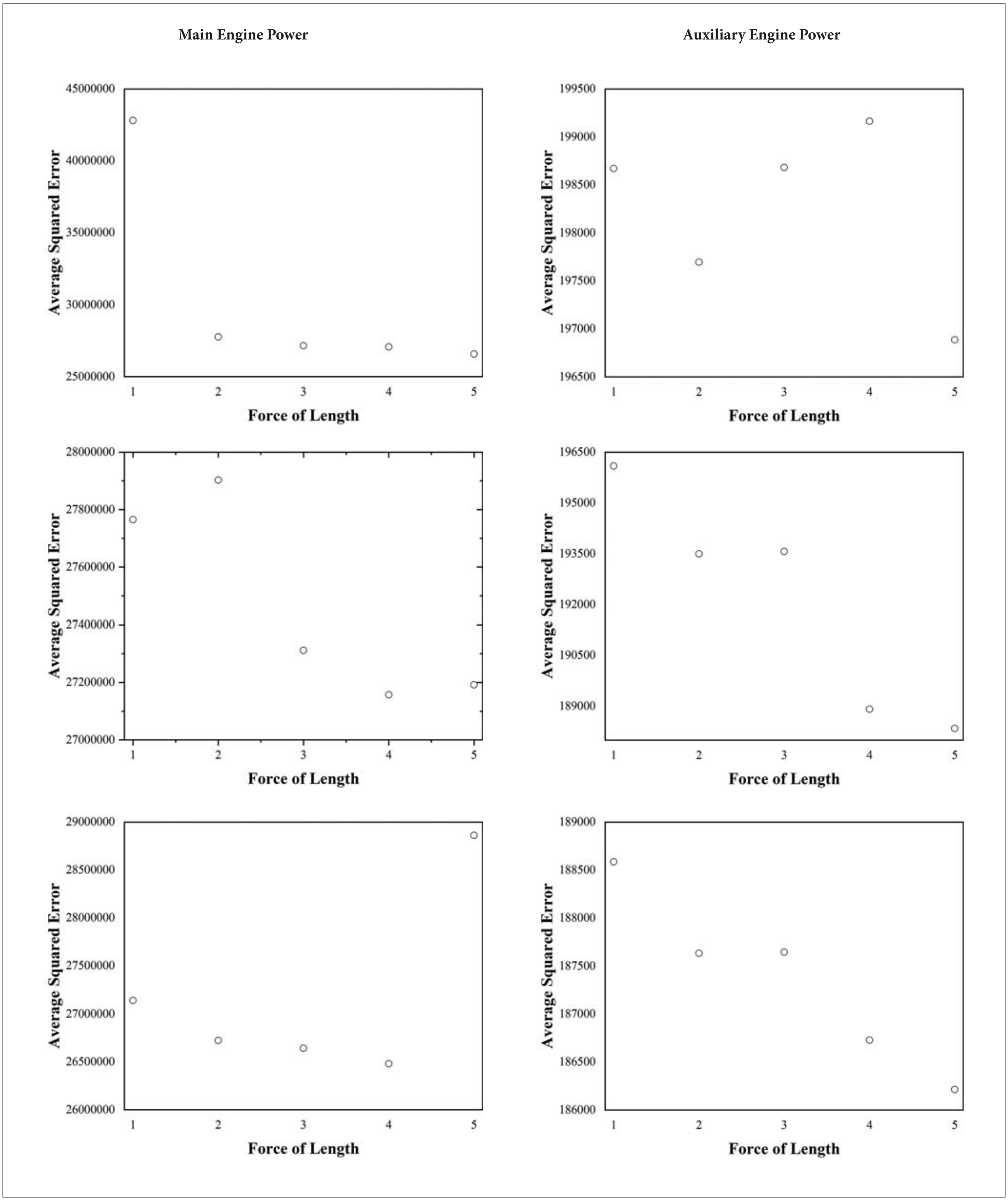

*Fig. 4. Forces of independent variables*

In this part of the study, the answer to the question of which polynomial levels should be created using the data in the whole data set without any test–train separation was sought. Polynomial forces between 1 and 5 were investigated for each predictor in Model.ME.1. Average squared error values were examined for each polynomial force and, according to the results, the forces of the estimators of the final model were decided.

*i*, *j*, and *k* represent the polynomial degrees of the independent variables (length, gross tonnage, and age), respectively. When Fig. 4 is examined, a 2nd degree polynomial is suitable for length, 4th degree for gross tonnage and 2nd degree for age. With reference to these results, Model.ME.2 was created to estimate the ship's main engine power. Similar steps were used to estimate the auxiliary engine power, for which Model.AE.2 was created. As a consequence of the applied operations, the force of the main engine power was 5, and the force of the length and age was 4 and 5, respectively. Fig. 4 shows the average square error obtained for various forces of the independent variables. Table 5 contains the errors from the polynomial model's train and test sets.

*Tab. 5. Error values of the polynomial model*

| | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | *RMSE* | *R*² | *MAE* | *RMSE* | *R*² | *MAE* |
| Model. ME.2 | 5174.01 | 0.807 | 3112.51 | 5006.42 | 0.800 | 2955.65 |
| Model. AE.2 | 431.59 | 0.676 | 232.28 | 421.22 | 0.691 | 238.55 |

## K-NEAREST NEIGHBOURS – REGRESSION

The K-nearest neighbours regression method is a simple algorithm that stores all available states and predicts the numerical target based on distance similarity. KNN was first used as a nonparametric technique in statistical prediction and pattern recognition in the early 1970s.

Contrary to alternative supervised learning algorithms, KNN does not have a training stage. With KNN, principally the closest points to the new point are searched. K represents the number of the closest neighbours of the unknown point. We select the amount K of the algorithm (usually an odd number) to estimate the results.

The KNN algorithm is predicted by the majority vote of its neighbours. The closest neighbours are found with a distance function. Eq. (10), (11), and (12) contain distance functions that are frequently used for regression.

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2} \tag{10}$$

Manhattan $$\sum_{i=1}^{k}|x_i-y_i| \tag{11}$$

Minkowski $$\left[\sum_{i=1}^{k}(|x_i-y_i|)^q\right]^{1/q} \tag{12}$$

The three distance functions above can only be used in continuous variables. To choose the most suitable value for K, the data should first be examined. In general, a large K value is more sensitive as it reduces overall noise, although no guarantee is granted. Cross-validation is another way to retrospectively determine a good K value, using an independent dataset to validate the value.

In this part of the study, the number of neighbours was determined. Model.ME.3 was designed to estimate the ship's main engine power and Model.AE.3 to estimate the auxiliary engine power. The arguments used to estimate the outputs were not changed. To determine the number of neighbours, numbers between 1 and 10 were examined and determined according to the RMSE values. Fig. 5 shows the RMSE values of the neighbour numbers.

When Fig. 5 is examined, the minimum error value for Model.ME.3 is obtained when the number of neighbours is 1. On the other hand, for Model.AE.3, the neighbour number should be 4. The Euclidean distance was used for both models. After determining the number of neighbours, the RMSE, $R^2$ and MAE errors were calculated for the test and train sets. Table 6 contains these error values.

*Tab. 6. Error values of the KNN model*

| | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | *RMSE* | *R*² | *MAE* | *RMSE* | *R*² | *MAE* |
| Model. ME.3 | 119.46 | 0.999 | 11.36 | 4245.57 | 0.856 | 1372.19 |
| Model. AE.3 | 350.22 | 0.787 | 173.18 | 385.99 | 0.739 | 220.57 |

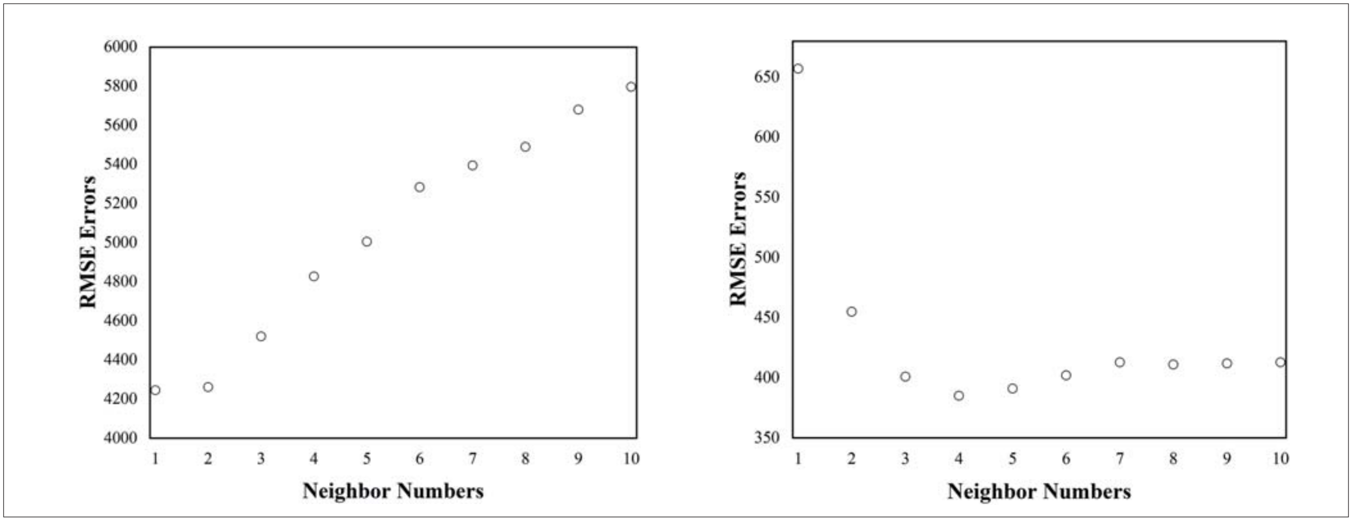

*Fig. 5. RMSE values of neighbour numbers*

## GRADIENT BOOSTING MACHINE (GBM)

The gradient boosting machine (GBM) is a nonparametric regression technique that combines a regression tree with the gradient boosting algorithm. Unlike the regression method, which basically produces a single best model, the GBM model adaptively combines multiple classification and regression tree models using the gradient boosting technique to optimise performance. That is, unlike standard regression methods that produce a single predictive model, it fits many simple models and combines them in prediction, thereby increasing the predictive performance. In addition, it does not need any assumptions about the functional relationship between dependent and independent variables. GBM uses the gradient boost algorithm from Boost algorithms.

This method requires the most training time. Besides, a considerable amount of parameters need to be determined from the outset. Initially, Model.ME.4 was designed to estimate the ship's main engine power, and Model.AE.4 was created to estimate the power of the auxiliary engine. Interaction depth, n.trees, shrinkage and n.minobsinnode variables were determined by tuning. The interaction depth 1 through 7 in 2 increments, n.trees between 1000 and 10,000 with 1000 increments, the shrinkage value as 0.01 or 0.1, and the n. minobsinnode value between 10 and 20 were searched. The optimum values of n.trees = 0000, interaction depth = 7, shrinkage value = 0.01 and n.minobsinnode = 10 were obtained for Model.ME.4. The final values used for Model.AE.4 were n.trees = 2000, interaction depth = 7, shrinkage = 0.01 and n.minobsinnode = 11. Fig. 6 shows the effect of these variables on the RMSE for the main engine and Fig. 7 shows the effect of these variables on the RMSE for the auxiliary engine.

The error rates for the final models created after the tuning process are listed in Table 7.

*Tab. 7. Error values of the GBM model*

| | Train | | | Test | | |
|---|---|---|---|---|---|---|
| | *RMSE* | *R*$^2$ | *MAE* | *RMSE* | *R*$^2$ | *MAE* |
| Model. ME.4 | 1135.29 | 0.990 | 714.19 | 2562.01 | 0.947 | 1246.5 |
| Model. AE.4 | 201.11 | 0.931 | 118.42 | 248.20 | 0.926 | 118.05 |

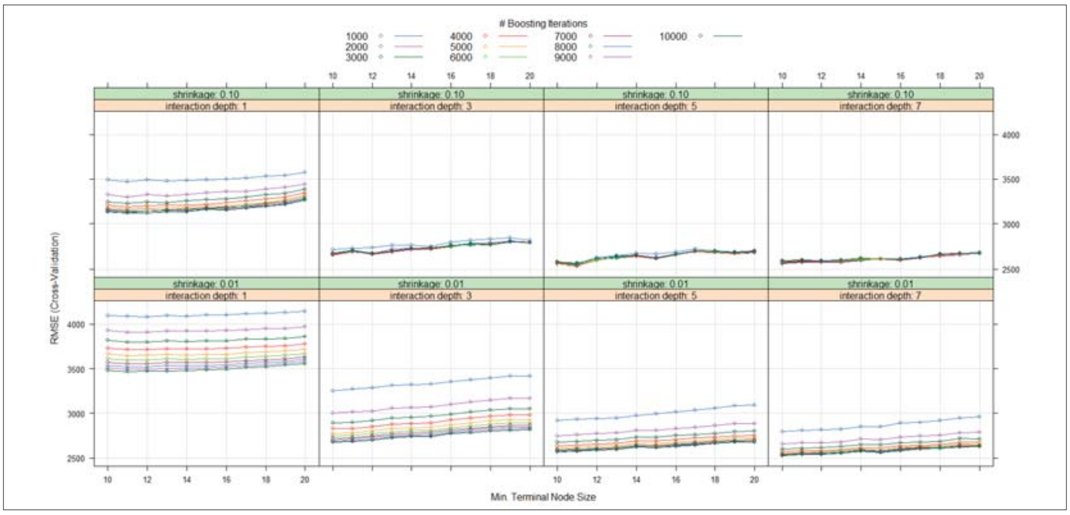

*Fig. 6. Effects of variables on RMSE for ME*

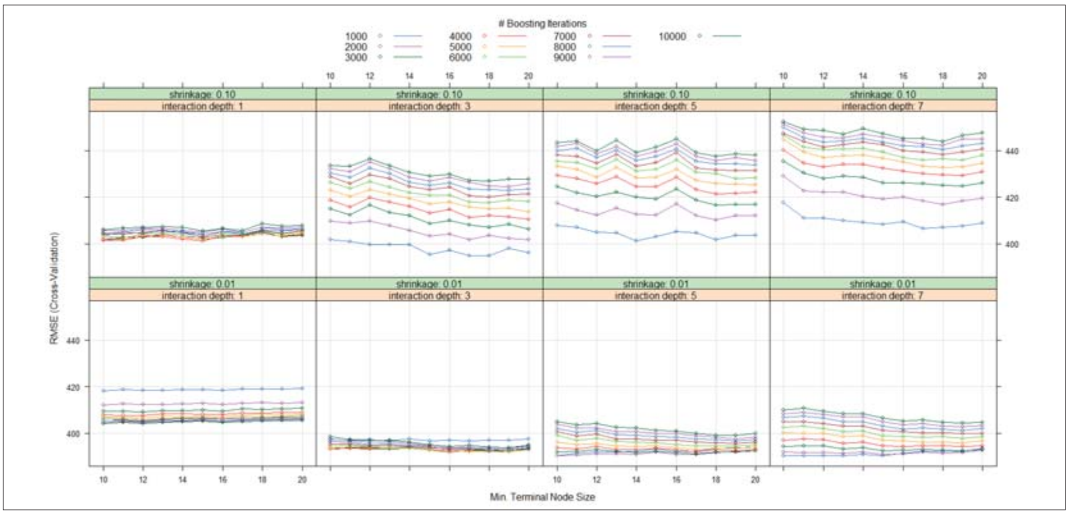

*Fig. 7. Effects of variables on RMSE for AE*

## RESULTS AND DISCUSSION

Based on the length, gross tonnage, and age data from 4037 different ships, this study estimated the main and auxiliary engine power values. As a predictor, four different regression models, linear, polynomial, KNN and GBM, were studied. The models were trained on 80% of the data set and tested in 20%. The performance of the models was evaluated with ten-fold cross-validation and the RMSE, MAE and $R^2$ errors were calculated and interpreted.

In Fig. 8, a comparison chart of the coefficients of determination ($R^2$) of the regression algorithms is given. The fact that the coefficient of determination is close to 1 indicates that the success of the algorithm is high. As a result of the study, the best regression algorithm for main engine power prediction is the gradient boosting machine with an $R^2$ value of 0.947. Among the models created for estimating the auxiliary motor power, the best performing model was again the gradient boosting machine and its $R^2$ value is 0.926.

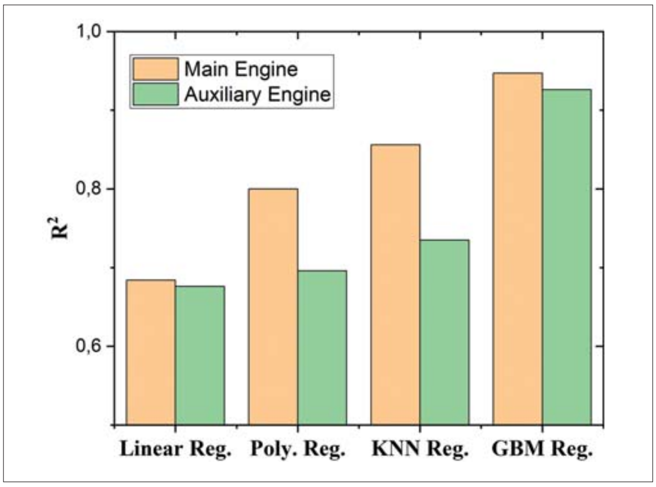

*Fig. 8. Coefficients of determination of four models*

In Fig. 9, the comparison chart of the mean absolute error (MAE) values of the regression algorithms is given. The fact that the mean absolute error value is close to 0 indicates that the success of the algorithm is high.

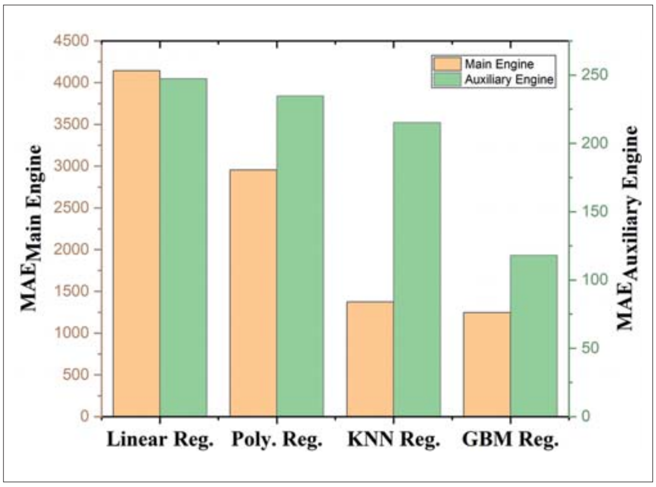

*Fig. 9. Mean absolute error values of four models*

The comparison graph of the root mean square error (RMSE) values of the regression algorithms is given in Fig. 10. Here too, it can be understood from the fact that the average square error value is close to 0 that the success of the algorithm is high.

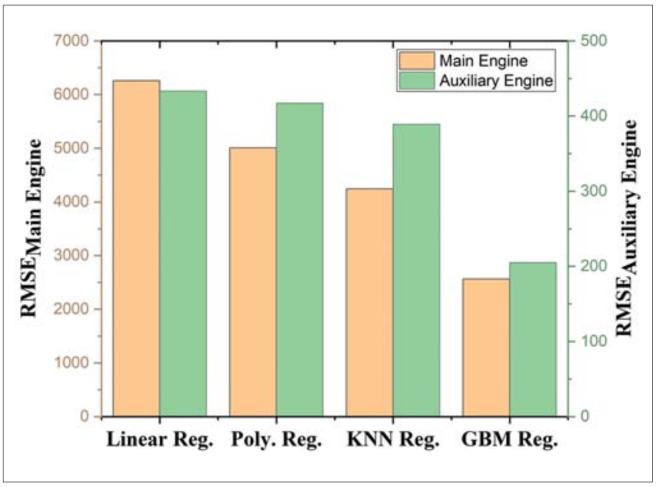

*Fig. 10. Root mean squared error values of four models*

The graphs showing the main and auxiliary engine power values estimated by the algorithms and the actual index values in the test data are given in Fig. 11.

Residual analysis plays an important role in verifying the regression model. The residues are the difference between the estimated value and the actual value. Graphs representing the deviation of the estimated value from the actual value are shown in Fig. 12.

Fig. 12 shows that the linear and polynomial regression algorithms with a high error rate now move away from the zero line. On the other hand, the low error rates of the GBM and KNN algorithms are somewhat closer to the zero line.

## CONCLUSION

In this study, regression-based algorithms are used to estimate ships' main and auxiliary machine powers. Four different regression algorithms, linear, polynomial, KNN, and GBM, have been designed. Each method requires data pre-processing, data distribution determination, regression and performance evaluation steps, which are important stages of machine learning. K-cross- validation validity, a hyperparameter frequently used in the literature, was used to compare the performance results of the machine learning methods. For KNN regression, the optimum neighbour numbers were searched from one to ten. In addition, as GBM regression, for the interaction depth, n.trees, shrinkage and n.minobsinnode parameters tuning was performed for four, ten, two, and ten different parameters, respectively. In the study of 4037 ship samples, the algorithm that can best estimate the power of both machines compared to $R^2$, *RMSE* and *MAE* was found to be the gradient boosting machine. Although this method provides good results, the number of parameters to be determined from the outset and the training

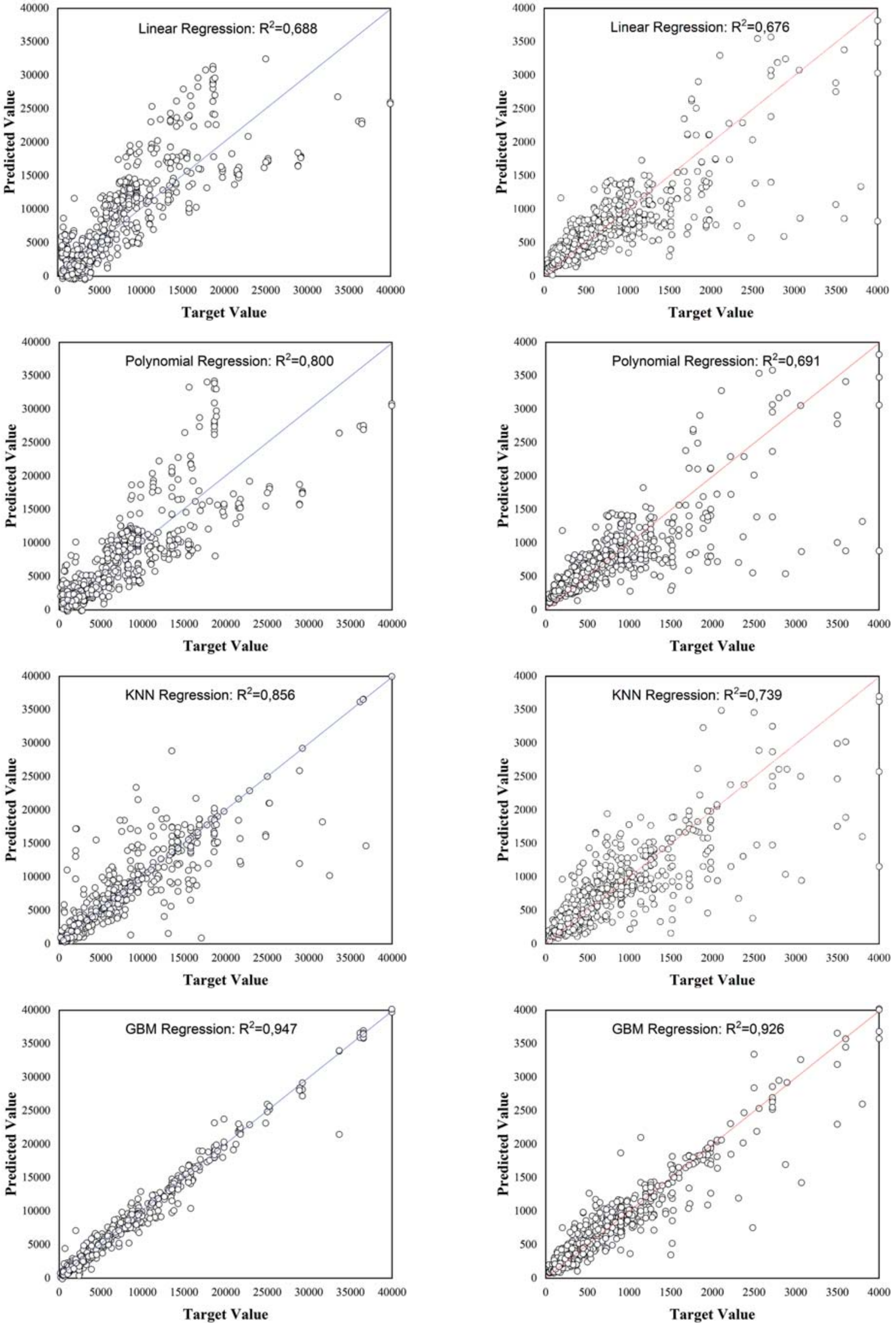

*Fig. 11. Difference between target values and forecast values*

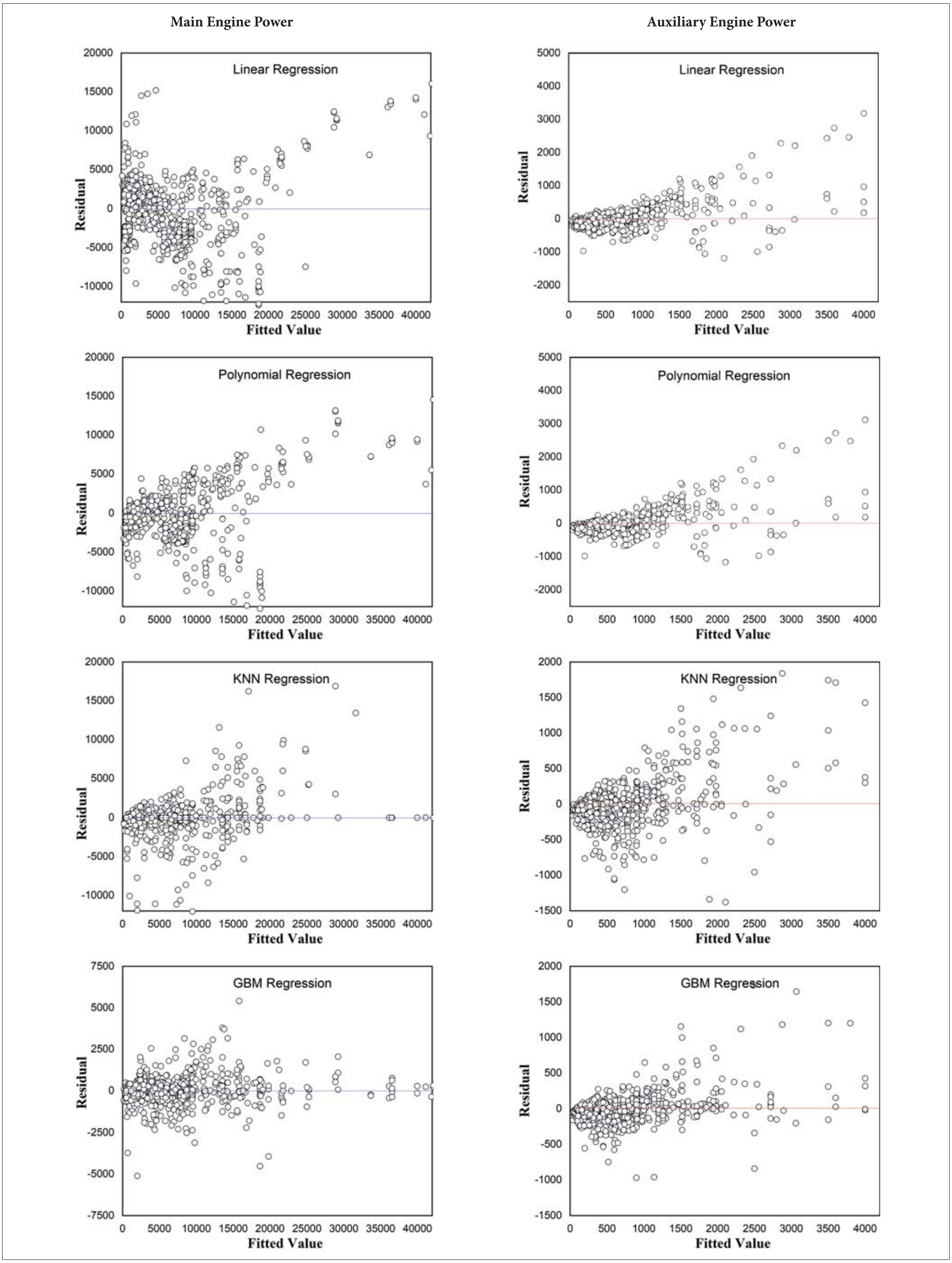

*Fig. 12. Residuals*

time proved to be more important as negative aspects of the method. However, the linear and polynomial regressions were not able to adapt to the data set. As a result, the GBM algorithm for estimating ships' main and auxiliary machine powers is quite suitable. It showed good results in estimating both the main power and the auxiliary machine power. The basis for this method's effectiveness is that the predictions are made in order, not independently.

## REFERENCES

1. A. Ekmekçioğlu, K. Ünlügençoğlu, and U. B. Çelebi, 'Ship emission estimation for Izmir and Mersin international ports – Turkey', *Journal of Thermal Engineering*, vol. 5, no. 6, pp. 184–195, 2019, doi: 10.18186/thermal.654319.

2. C. Trozzi, 'Emission estimate methodology for maritime navigation', *Co-leader of the Combustion & Industry Expert Panel*, 2010.

3. R. Yan, S. Wang, and Y. 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*, vol. 138, no. July 2019, p. 101930, 2020, doi: 10.1016/j.tre.2020.101930.

4. L. Huang, Y. Wen, Y. Zhang, C. Zhou, F. Zhang, and T. Yang, 'Dynamic calculation of ship exhaust emissions based on real-time AIS data', *Transportation Research Part D: Transport and Environment*, vol. 80, no. August 2019, p. 102277, 2020, doi: 10.1016/j.trd.2020.102277.

5. T. A. Tran, 'Effect of ship loading on marine diesel engine fuel consumption for bulk carriers based on the fuzzy clustering method', *Ocean Engineering*, vol. 207, no. January 2019, p. 107383, 2020, doi: 10.1016/j.oceaneng.2020.107383.

6. X. Yan, K. Wang, Y. Yuan, X. Jiang, and R. R. Negenborn, 'Energy-efficient shipping: An application of big data analysis for optimizing engine speed of inland ships considering multiple environmental factors', *Ocean Engineering*, vol. 169, no. August, pp. 457–468, 2018, doi: 10.1016/j.oceaneng.2018.08.050.

7. T. Cepowski, 'Regression formulas for the estimation of engine total power for tankers, container ships and bulk carriers on the basis of cargo capacity and design speed', *Polish Maritime Research*, vol. 26, no. 1, pp. 82–94, Mar. 2019, doi: 10.2478/pomr-2019-0010.

8. W. J. Requia, B. A. Coull, and P. Koutrakis, 'Evaluation of predictive capabilities of ordinary geostatistical interpolation, hybrid interpolation, and machine learning methods for estimating PM2.5 constituents over space', *Environmental Research*, vol. 175, no. April, pp. 421–433, 2019, doi: 10.1016/j.envres.2019.05.025.

9. T. Uyanık, Ç. Karatuğ, and Y. Arslanoğlu, 'Machine learning approach to ship fuel consumption: A case of container vessel', *Transportation Research Part D: Transport and Environment*, vol. 84, 2020, doi: 10.1016/j.trd.2020.102389.

10. L. Barua, B. Zou, and Y. Zhou, 'Machine learning for international freight transportation management: A comprehensive review', *Research in Transportation Business and Management*, no. July 2019, p. 100453, 2020, doi: 10.1016/j.rtbm.2020.100453.

11. Y. Peng, H. Liu, X. Li, J. Huang, and W. Wang, 'Machine learning method for energy consumption prediction of ships in port considering green ports', *Journal of Cleaner Production*, vol. 264, p. 121564, 2020, doi: 10.1016/j.jclepro.2020.121564.

12. J. H. Jeong, J. H. Woo, and J. G. Park, 'Machine learning methodology for management of shipbuilding master data', *International Journal of Naval Architecture and Ocean Engineering*, vol. 12, pp. 428–439, 2020, doi: 10.1016/j.ijnaoe.2020.03.005.

13. C. Gkerekos, I. Lazakis, and G. Theotokatos, 'Machine learning models for predicting ship main engine fuel oil consumption: A comparative study', *Ocean Engineering*, vol. 188, no. August, p. 106282, 2019, doi: 10.1016/j.oceaneng.2019.106282.

14. A. Jonquais and F. Krempl, 'Predicting Shipping Time with Machine Learning', 2019.

15. O. Bodunov, F. Schmidt, A. Martin, A. Brito, and C. Fetzer, 'Grand challenge: Real-time destination and ETA prediction for maritime traffic', *DEBS 2018 – Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems*, pp. 198–201, 2018, doi: 10.1145/3210284.3220502.

16. J. Yuan and V. Nian, 'Ship energy consumption prediction with Gaussian process metamodel', *Energy Procedia*, vol. 152, pp. 655–660, 2018, doi: 10.1016/j.egypro.2018.09.226.

17. Y. B. A. Farag and A. I. Ölçer, 'The development of a ship performance model in varying operating conditions based on ANN and regression techniques', *Ocean Engineering*, vol. 198, no. July 2019, 2020, doi: 10.1016/j.oceaneng.2020.106972.

18. L. Bui-Duy and N. Vu-Thi-Minh, 'Utilization of a deep learning-based fuel consumption model in choosing a liner shipping route for container ships in Asia', *Asian Journal of Shipping and Logistics*, 2020, doi: 10.1016/j.ajsl.2020.04.003.

19. H. Cui, O. Turan, and P. Sayer, 'Learning-based ship design optimization approach', *CAD Computer Aided Design*, vol. 44, no. 3, pp. 186–195, 2012, doi: 10.1016/j.cad.2011.06.011.

20. M. Peker, O. Özkaraca, and B. Kesimal, 'Modeling heating and cooling loads by regression-based machine learning techniques for energy-efficient building design', *International Journal of Informatics Technologies*, pp. 443–449, 2017, doi: 10.17671/gazibtd.310154.

21. V. Bertram and H. Schneekluth, *Ship Design for Efficiency and Economy*. Elsevier, 1998.

## CONTACT WITH THE AUTHORS

**Fatih Okumuş**
*e-mail: hfatihokumus@gmail.com*

Yildiz Technical University
Besiktas, 34000 Istanbul
Turkey

**Araks Ekmekçioğlu**
*e-mail: araks@yildiz.edu.tr*

Yildiz Technical University
Besiktas, 34000 Istanbul
Turkey

**Selin Soner Kara**
*e-mail: ssoner@yildiz.edu.tr*

Yildiz Technical University
Besiktas, 34000 Istanbul
Turkey