Journal of Advanced Marine Engineering and Technology, Vol. 49, No. 5, pp. 333~345, 2025
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2025.49.5.333

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

Research on the prediction of compression pressure of 2-stroke marine diesel engine with data-driven models and analysis based on the SHAP

Min-Ho Park[1,2] · Jae-Jung Hur[3] · Byeong-Deok Yea[4] · Jae-Hyuk Choi[3] · Won-Ju Lee[2,†]

(Received October 3, 2025 ; Revised October 20, 2025 ; Accepted October 29, 2025)

Abstract: In marine diesel engines, the compression pressure in the cylinder is important because it affects combustion performance. However, as the sensor measuring compression pressure is installed in relation to the combustion chamber, its performance may be decreased, reliability may be difficult to guarantee, and there is a possibility of failure. Therefore, it is necessary to develop a model to predict compression pressure. In this study, engine data from the training ship was acquired, preprocessed, and then trained on GBDT-based CatBoost and LightGBM models. The results of the two models were confirmed using five performance metrics and scatter and line plots. LightGBM showed better performance than CatBoost with the train and test sets for all performance metrics, but the difference was minimal. Analysis of the two trained models using SHAP revealed that the 'M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM' variables had the greatest impact on prediction of cylinder compression pressure.

Keywords: Prediction, Compression pressure, 2-stroke marine diesel engine, Data-driven models, SHAP

1. Introduction

Maritime transport of global trade moves over 80% of goods traded worldwide by volume **[1]-[3]**. Goods traded worldwide include crude oil, liquefied natural gas (LNG), coal, iron ore, grain, vehicles, chemicals, and containers. Ships are used to transport these goods, and types of ships include oil tankers, LNG carriers, bulk carriers, car carriers, chemical ships, and container ships **[4]**. A 2-stroke marine diesel engine (main engine) is installed on the ship to propel the ship. **Figure 1** shows a system schematic of the main engine (M/E).

The M/E is powered by the explosion of a mixture of pressurized air and atomized fuel injected into the cylinder. The explosive force generated inside the cylinder causes the piston to reciprocate, which in turn rotates the crankshaft and the propeller, propelling the ship **[5]**.

The exhaust gas generated by the explosion of a mixture of pressurized air and atomized fuel flows through the exhaust manifold toward the turbine of the turbocharger (T/C) as the exhaust valve opens due to pressurized hydraulic oil. High temperature and high pressure exhaust gas with heat energy rotates the turbine wheel of the T/C after passing through nozzle ring. Accordingly, the compressor wheel connected to the turbine wheel with common shaft rotates and sucks in outside air through the filter silencer. The diffuser located between the compressor and the compressor casing increases the pressure of the intake air. The pressurized air passes through the air cooler where it is cooled and its volumetric efficiency is improved before it is sent to the scavenge air receiver.

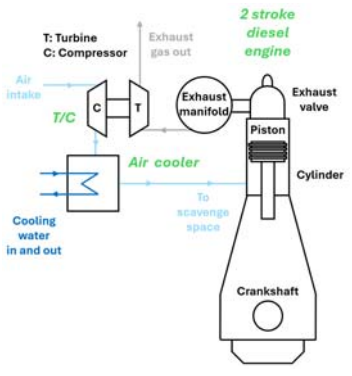

Figure 1: A system schematic of the M/E

† Corresponding Author (ORCID: http://orcid.org/0000-0001-8380-8969): Professor, Division of Marine System Engineering, National Korea Maritime & Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49315, Korea, E-mail: skywonju@kmou.ac.kr, Tel: +82-51-410-4262
1 Division of Marine Engineering, National Korea Maritime & Ocean Engineering University, E-mail: mhpark@g.kmou.ac.kr
2 Interdisciplinary Major of Maritime AI Convergence, National Korea Maritime & Ocean University
3 Division of Marine System Engineering, National Korea Maritime & Ocean University
4 Division of Maritime AI & Cybersecurity, Korea Maritime & Ocean University, E-mail: byea@kmou.ac.kr

Pressurized air enters the each cylinder from the scavenge air receiver through the scavenging port of each cylinder **[6]**. The pressurized air entering the cylinder is compressed by the reciprocating motion of the piston, and the fuel injected into the high-temperature, high-pressure environment is mixed with the pressurized air and explodes. This cycle repeats until the M/E stops, resulting in a reciprocating rotational motion.

The piston ring made of a special material is fitted into the ring groove of the piston and comes into contact with the cylinder liner. To prevent direct friction between the piston ring and cylinder liner, lubrication oil is sprayed through holes in the cylinder liner to lubricate and seal the space between the piston ring and cylinder liner. However, if damage occurs to the piston ring groove, piston ring, or cylinder liner, a gap will form in the compression space, reducing the compression pressure and combustion efficiency. In severe cases, blow-by occurs, where exhaust gas leaks between the piston and cylinder liner. Therefore, continuous monitoring of the compression pressure of the M/E is important to check the sealing status of the compression space.

Academia also recognizes the importance of in-cylinder pressure in marine diesel engines and has been conducting various studies. Tang et al. developed a real-time two-stroke marine diesel engine model with fast computing speed of mean value model and in-cylinder pressure prediction capability of zero dimensional model **[7]**. Hountalas examined the effect of compression failure through the reduction of compression ratio **[8]**. Patil *et al.* used artificial neural networks (ANN) to estimate the in-cylinder pressure of a marine engine using engine speed, power, and crank angle as inputs **[9]**. Shen *et al.* developed a mean value engine model with in-cylinder pressure trace predictive capability and a novel compressor model for the 2-stroke diesel engine **[10]**. Patil *et al.* used in-cylinder pressure signals and multiple linear and polynomial regression models to estimate the severity of the faults for marine engines **[11]**. Galliakis used ANN to predict peak cylinder pressure of 4-stroke marine diesel engine using engine speed, torque, lambda, and specific fuel consumption as inputs **[12]**. Asimakopoulos *et al.* used support vector regression, random forest regression, and XGBoost to predict maximum pressure **[13]**. Then, the condition of the piston ring was determined based on the difference between the actual value and model's predicted value. Tsitsilonis and Theotokatos used instantaneous crankshaft torque to predict pressure variations in all engine cylinders **[14]**.

Sensors that measure compression pressure in marine diesel engines are continuously exposed to high-temperature and high-pressure environments due to the explosive gases generated within the cylinder. Accordingly, it is practically difficult to continuously maintain optimal performance and make reliable measurements. However, recent advancements in artificial intelligence (AI) have made it possible to develop virtual sensors that can compensate for these shortcomings. Therefore, in this study, we conducted a prediction of compression pressure using an AI model based on an engine dataset acquired from an actual ship. Furthermore, the influence of variables was analyzed using SHAP on the trained models.

The remainder of this paper is organized as follows. In Section 2, the experimental subject and data acquisition process were described. In Section 3, data preprocessing and data analysis using correlation heatmap were performed. In Section 4, the theory behind CatBoost, LightGBM, and SHAP is explained. In Section 5, modeling of CatBoost and LightGBM and five performance metrics were described. In Section 6, the prediction results of CatBoost and LightGBM models and their interpretation by SHAP were analyzed. In Section 7, this research process has been concluded.

2. Data Acquisition

2.1 Experimental Subject

The training ship was used as an experimental subject to develop an AI model to predict compression pressure. The main specifications of the training ship are shown in **Table 1**.

The training ship is equipped with the MAN 6S40ME-B9.5 with an output of 6,618 kW, and the dataset consisting of sensor values is automatically downloaded by the ACONIS system.

Table 1: Main specifications of the training ship

Subject	Value	Subject	Value
IMO No.	9807279	Breadth extreme (m)	19
Name	HANNARA	Year built	2019
Vessel type	Training ship	Engine type	MAN 6S40ME-B9.5
Flag	South Korea	Power	6,618 kW at 146 rpm
Gross tonnage (t)	9,196	Propeller′s blades	4
Summer deadweight (t)	3,671	Propeller′s diameter (m)	4
Length overall (m)	133		

Table 2: Sailing routes of the training ship.

Voyage destination	Date for acquiring data
Incheon	2022.05.25–30
Ulleungdo	2022.06.25–27

The training ship sailed to Incheon and Ulleungdo as shown in **Table 2**, and the dataset was collected during that period.

2.2 Acquisition of Data

The datasets stored by the ACONIS system were in Microsoft Access format with a 10-second cycle. The dataset in Microsoft Access format was converted to comma separated values (CSV) format for data preprocessing.

3. Data Analysis

3.1 Data Preprocessing

Excluding the 2022.05.27 and 2022.05.30 Incheon datasets, which have a majority of 0 values in the dataset, only the 2022.05.25, 2022.05.26, and 2022.05.29 Incheon datasets were used. These datasets were named Incheon1, Incheon2, and Incheon3, respectively. The rows and columns of these datasets were (8640, 192), (8640, 192), and (8436, 192), respectively. The dataset stored by the ACONIS system consisted of 192 sensor parameters. The Incheon datasets were acquired under various driving conditions, with M/E loads ranging from 0% to 81.8%.

Excluding the 2022.06.26 Ulleungdo dataset, which have a majority of 0 values in the dataset, only the 2022.06.25 and 2022.06.27 Ulleungdo datasets were used. These datasets were named Ulleungdo 1 and Ulleungdo 2, respectively. The rows and columns of these datasets were (8429, 192) and (8640, 192), respectively. The Ulleungdo datasets were acquired under various driving conditions, with M/E loads ranging from 0% to 69.7%.

As the goal of this study was to predict compression pressure, 29 related variables were extracted from 192 variables, as shown in **Table 3**. Accordingly, the sizes of the Incheon1, Incheon2, Incheon3, Ulleungdo 1, and Ulleungdo 2 datasets became (8640, 29), (8640, 29), (8436, 29), (8429, 29), and (8640, 29).

The graphs are shown in **Figure 2** with the number of rows in each dataset on the x-axis and "average cylinder compression pressure" and "M/E rpm" on the y-axis. Referring to **Figure 2**, it can be seen that “average cylinder compression pressure” (blue color) has a significant correlation with “M/E rpm” (red color).

However, referring to **Figure 2**, there are considerable sections where “M/E rpm” is 0. Therefore, in this study, we decided to predict the compression pressure after filtering the sections where the telegraph is dead slow or higher. The rpm values for each telegraph of the training ship are as shown in **Table 4**.

Table 3: 29 variables selected for predicting compression pressure.

No.	Variable name	No.	Variable name
1	M/E A/C C.W INLET PRESS	16	M/E NO.5 CYL EXH GAS OUTLET TEMP
2	M/E A/C C.W INLET TEMP	17	M/E NO.6 CYL EXH GAS OUTLET TEMP
3	M/E A/C C.W OUTLET TEMP	18	M/E CYL EXH GAS OUTLET TEMP MEAN
4	M/E SCAV AIR RECEIVER IN PRESS	19	M/E T/C EXH GAS INLET TEMP
5	M/E NO.1 SCAV AIR FIRE DET. TEMP	20	M/E T/C EXH GAS OUTLET TEMP
6	M/E NO.2 SCAV AIR FIRE DET. TEMP	21	M/E RPM
7	M/E NO.3 SCAV AIR FIRE DET. TEMP	22	M/E T/C RPM
8	M/E NO.4 SCAV AIR FIRE DET. TEMP	23	CYLINDER COMPRESSION PRESS CYL 1
9	M/E NO.5 SCAV AIR FIRE DET. TEMP	24	CYLINDER COMPRESSION PRESS CYL 2
10	M/E NO.6 SCAV AIR FIRE DET. TEMP	25	CYLINDER COMPRESSION PRESS CYL 3
11	M/E SCAV AIR RECEIVER TEMP	26	CYLINDER COMPRESSION PRESS CYL 4
12	M/E NO.1 CYL EXH GAS OUTLET TEMP	27	CYLINDER COMPRESSION PRESS CYL 5
13	M/E NO.2 CYL EXH GAS OUTLET TEMP	28	CYLINDER COMPRESSION PRESS CYL 6
14	M/E NO.3 CYL EXH GAS OUTLET TEMP	29	AVERAGE CYLINDER COMPR. PRESS
15	M/E NO.4 CYL EXH GAS OUTLET TEMP		

Referring to **Table 4**, we filtered only those with "M/E rpm" greater than 73 (Dead Slow) and variables related to cylinder compression pressure are less than 250 bar. When the ‘M/E rpm’ was 0, the values of variables related to cylinder compression pressure was 250 bar. However, in some cases, it showed 250

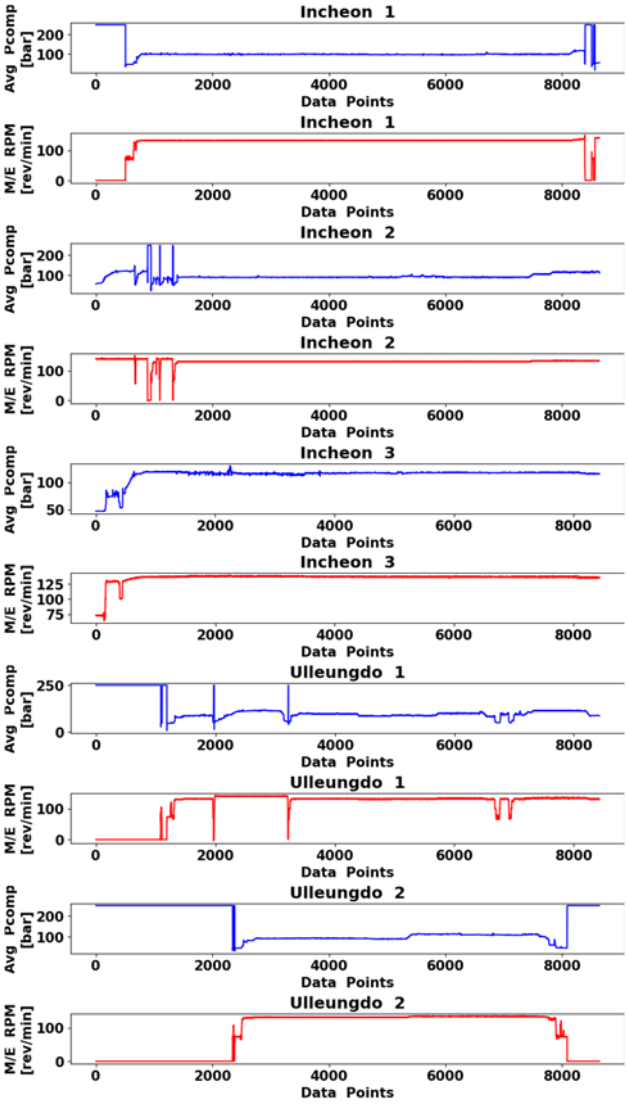

Figure 2: Line plots of average compression pressure and M/E RPM for Incheon 1, 2, and 3 and Ulleungdo 1 and 2

Table 4: Harbor speed table for the training ship

Telegraph	RPM
Navigation Full	141
Full	130
Half	126
Slow	100
Dead Slow	73

bars even when an error occurred in the sensor, so values less than 250 were filtered out. As a result, the sizes of the Incheon1, Incheon2, Incheon3, Ulleungdo 1, and Ulleungdo 2 datasets became (7928, 29), (8148, 29), (7709, 29), (7148, 29), and (5597, 29). Line plots of "average cylinder compression pressure" and "M/E rpm" for the filtered dataset are shown in **Figure 3**.

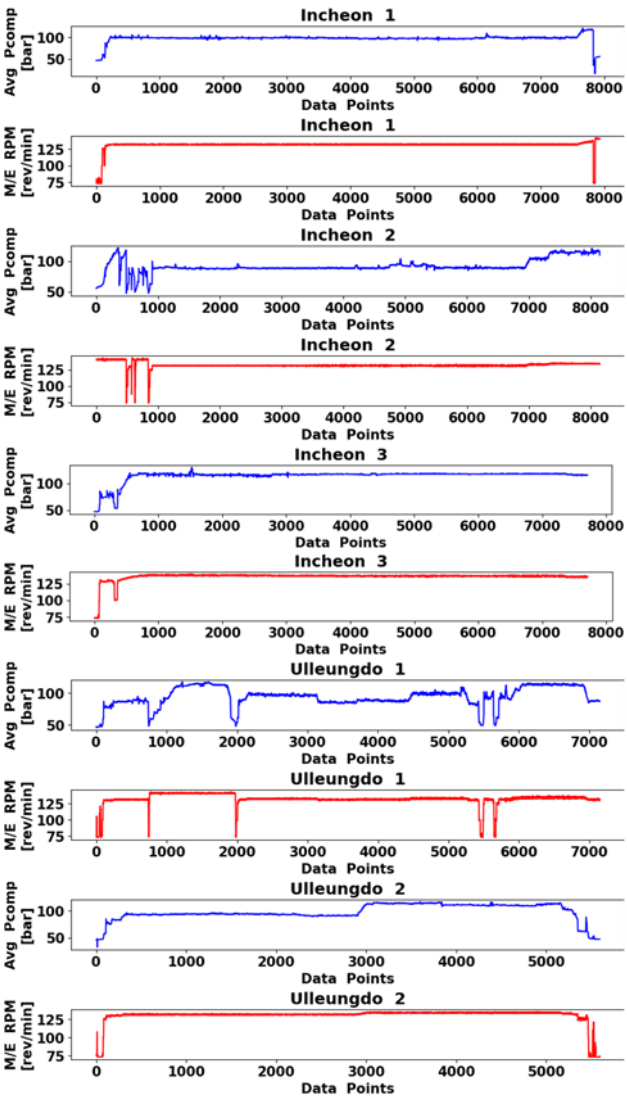

Figure 3: Line plots of average compression pressure and M/E RPM for Incheon 1, 2, and 3 and Ulleungdo 1 and 2, where M/E RPM is 73 or more and cylinder compression pressure is less than 250 bar

3.2 Correlation Heatmap

For ease of data processing, the datasets of Incheon1, Incheon2, and Incheon3 were combined into the Incheon dataset, and the datasets of Ulleungdo 1 and Ulleungdo 2 were combined into the Ulleungdo dataset. Accordingly, the Incheon and Ulleungdo datasets have sizes of (23785, 29) and (12745, 29), respectively.

A correlation heatmap was created to check the correlation between the seven output variables related to cylinder compression pressure and the remaining variables. In the correlation heatmap, plus one means positive strong correlation; as it approaches zero, it becomes low correlation. Minus one implies negative strong correlation **[15]**. **Figure 4** shows a correlation heatmap based on

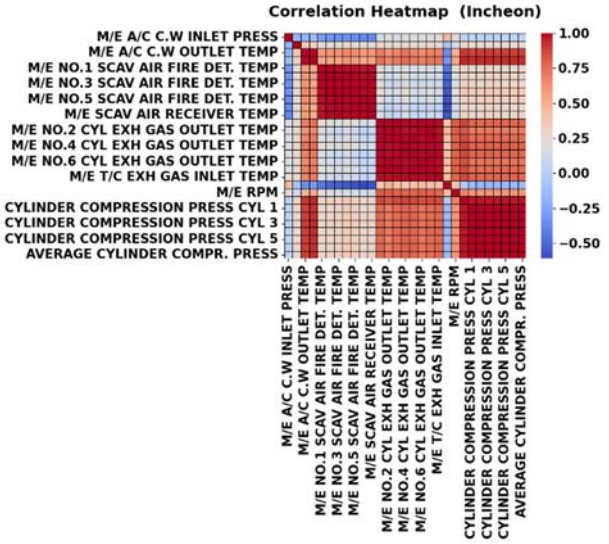

Figure 4: Correlation heatmap for Incheon dataset

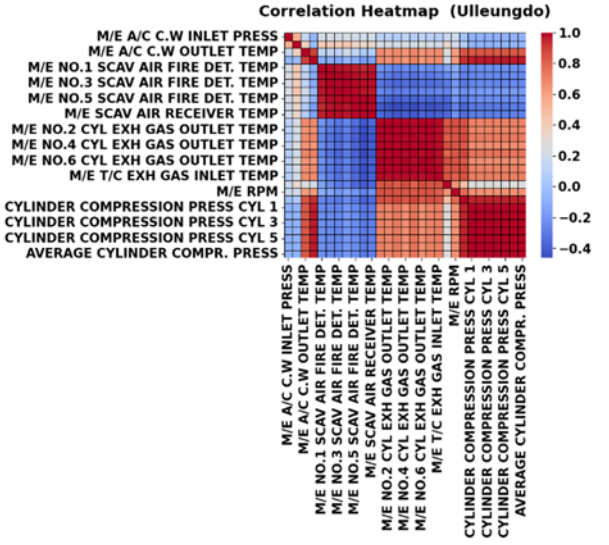

Figure 5: Correlation heatmap for Ulleungdo dataset

the correlation matrix for the Incheon dataset. Red and blue represent positive and negative values, respectively. Although some variables are not shown due to space constraints, the variables that are close to 0 or have negative values in the correlation heatmap are “M/E A/C C.W INLET PRESS”, “M/E A/C C.W INLET TEMP”, “M/E NO.1-6 SCAV AIR FIRE DET. TEMP”, and “M/E T/C EXH GAS OUTLET TEMP”.

In **Figure 5**, the values of the aforementioned variables are also shown to be close to 0 or have negative values in the correlation heatmap. Therefore, these variables were decided to be removed.

Correlation heatmaps for the Incheon and Ulleungdo datasets, excluding variables with low correlation, are shown in **Figures 6** and **7**. In **Figure 6**, the correlation heatmap shows that the negative

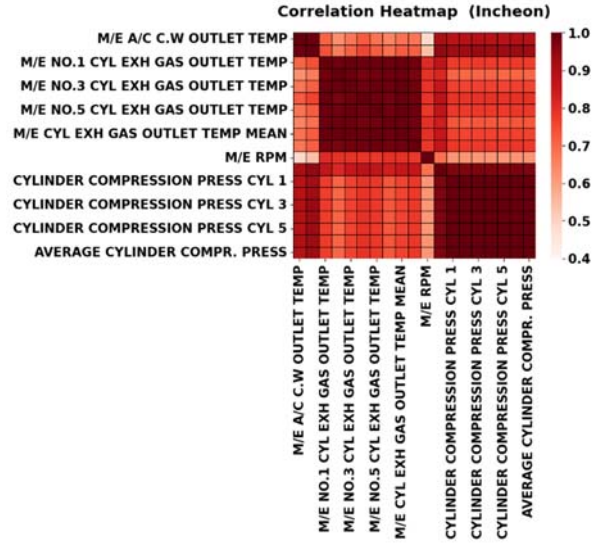

Figure 6: Correlation heatmap for Incheon dataset excluding variables with low correlation

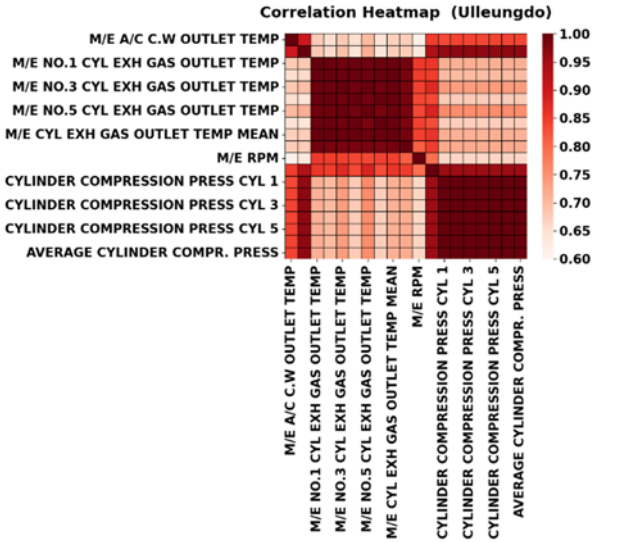

Figure 7: Correlation heatmap for Ulleungdo dataset excluding variables with low correlation

portion of the value range of the variables has removed. In **Figure 6**, the range of values of variables in the correlation heatmap was between 0.4 and 1.

In **Figure 7**, the range of values of variables in the correlation heatmap was between 0.6 and 1.

4. Theory

4.1 CatBoost

CatBoost is a machine learning ensemble technique based on gradient boosted decision trees (GBDT) **[16]**. CatBoost introduces two critical algorithmic advances of ordered boosting

enabling a permutation-driven alternative to the classic algorithm and innovative algorithm for processing categorical features **[17]**. CatBoost has the differences from other GBDT-based models such as focusing on optimizing decision trees for categorical variables, building symmetric (balanced) trees, a faster prediction time in large datasets compared to other GBDT models, and providing model analysis tools such as feature importance and feature analysis charts **[18]**.

4.2 LightGBM

LightGBM uses Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS excludes a significant proportion of data instances with small gradients and uses only the rest to estimate the information gain. EFB bundles mutually exclusive features to reduce the number of features **[19]**. LightGBM has the advantages of fast learning speed, high accuracy, GPU learning, enabling the handling of large datasets, and reducing the memory occupied when running **[18]**.

4.3 SHAP

SHAP (SHapley Additive exPlanations) assigns each feature an importance value for a particular prediction **[20]**. SHAP measures how much each feature contributes to a model's prediction and interprets the prediction as the aggregate of the Shapley values corresponding to all input features:

$$g(x') = \varphi_0 + \sum_{M}^{j=1} \varphi_j$$

where $g(x')$ is the value of the model, φ_0 is the constant that explains the model, and φ_j is the imputed value of each feature **[21]**.

5. Modeling

The following steps were taken to create the dataset for training and testing the CatBoost and LightGBM models. In **Table 3**, variables corresponding to 3, 4, 11–19, 21, and 22 were selected as input variables, and variables corresponding to 23-29 were selected as output variables. Accordingly, the Incheon and Ulleungdo datasets were divided into the train set (X_train_incheon, y_train_incheon, X_train_ulleungdo, y_train_ulleungdo) and the test set (X_test_incheon, y_test_incheon, X_test_ulleungdo, y_test_ulleungdo) in a ratio of 7:3. We combined the train and test sets, which were separated by voyage, to create a unified train set (X_train, y_train) and test set (X_test, y_test). The X_train had the shape of (25570, 12), the y_train had the shape of (25570, 7), the X_test had the shape of (10960, 12), and the y_test had the shape of (10960, 7).

Table 5: Parameters used to build CatBoost and LightGBM models

CatBoost	
Parameter	**Value**
learning_rate	0.03
depth	6
l2_leaf_reg	3
iterations	1000
LightGBM	
Parameter	**Value**
learning_rate	0.1
max_depth	-1
min_child_weight	0.001
colsample_bytree	1.0
subsample	1.0
n_estimators	100

CatBoostRegressor was imported from the CatBoost library, version 1.2.8, and LGBMRegressor was imported from the LightGBM library, version 4.6.0.

Table 5 shows the default parameters that were used to build each model.

CatBoostRegressor's 'loss_function' used 'MultiRMSE', and LGBMRegressor's 'eval_metric' used 'RMSE'. As LGBMRegressor cannot predict multi-output variables, Scikit-learn's 'MultiOutputRegressor' was used as a wrapper.

The five-performance metrics were used to evaluate the model's performance such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and R2 score. The formulas for these performance metrics are shown below.

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

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

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

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},\ where\ \ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

6. Results and Discussion

6.1 Model Prediction

CatBoost and LightGBM models were trained with train set, and the evaluation results on the train set and test set are as shown in **Tables 6** and **7**, respectively.

Referring to **Table 6**, it shows that LightGBM outperforms CatBoost on all performance metrics for the train set. However, there was no significant difference in the performance metrics for the two models, and CatBoost also performed as well as LightGBM.

Referring to **Table 7**, it shows that LightGBM outperforms CatBoost on all performance metrics for the test set. However, there was no significant difference in the performance metrics for the two models, and CatBoost also performed as well as LightGBM.

Figures 8 and **9** are scatter plots with the x-axis representing actual values of test set and the y-axis representing predicted values of test set. The line crossing the diagonal is the reference line, and if the data points are concentrated on this line, it

Table 6: Evaluation results for the train set of CatBoost and LightGBM models

CatBoost		LightGBM	
MAE	0.3362	MAE	0.2897
MSE	0.2428	MSE	0.2231
RMSE	0.4927	RMSE	0.4723
MAPE	0.0035	MAPE	0.0031
R2	0.9989	R2	0.999

Table 7: Evaluation results for the test set of CatBoost and LightGBM models

CatBoost		LightGBM	
MAE	0.3543	MAE	0.3123
MSE	0.3171	MSE	0.263
RMSE	0.5632	RMSE	0.5128
MAPE	0.0037	MAPE	0.0033
R2	0.9986	R2	0.9988

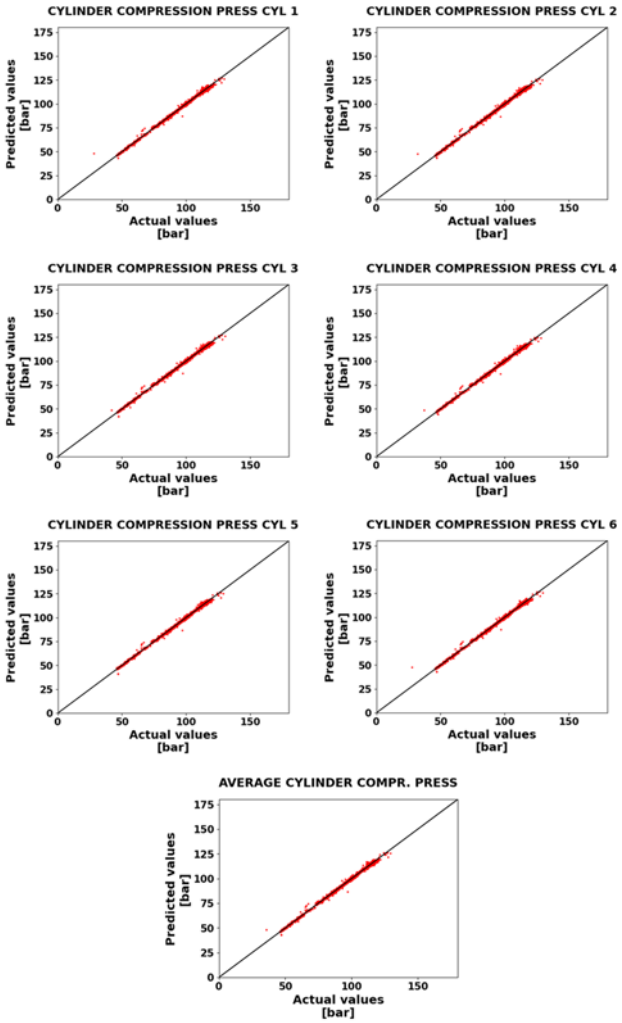

Figure 8: Scatter plot of CatBoost's predictions for the test set.

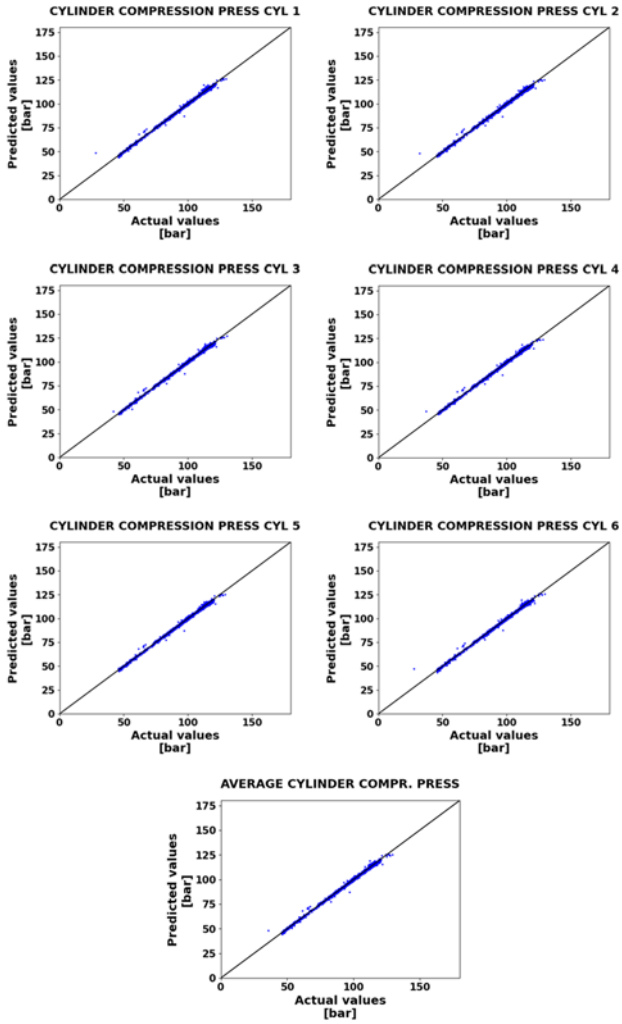

Figure 9: Scatter plot of LightGBM's predictions for the test set

indicates that the model's prediction is excellent. Referring to **Figure 8**, most of the data points, except for some, are concentrated on the diagonal line.

Referring to **Figure 9**, the predicted results of the LightGBM model are also excellent, like the CatBoost model. Most of the blue data points are concentrated along the diagonal line.

Figures 10 and **11** show the predictions of CatBoost and LightGBM for the test set as line plots. The x-axis represents data points, and the y-axis represents the values of variables associated with cylinder compression pressure. The red and blue line plots represent the actual values of test set for CatBoost and LightGBM, and the gray lines represent the predicted values of CatBoost and LightGBM.

Referring to **Figure 10**, CatBoost's predictions for the cylinder compression pressure of the test set almost overlap the actual values. The red line that juts out downwards near the 6000th data point on the x-axis appears equally for all cylinder compression pressure variables. This phenomenon occurs similarly to other data points. Therefore, the compression pressure of all cylinders changes similarly as the M/E status changes.

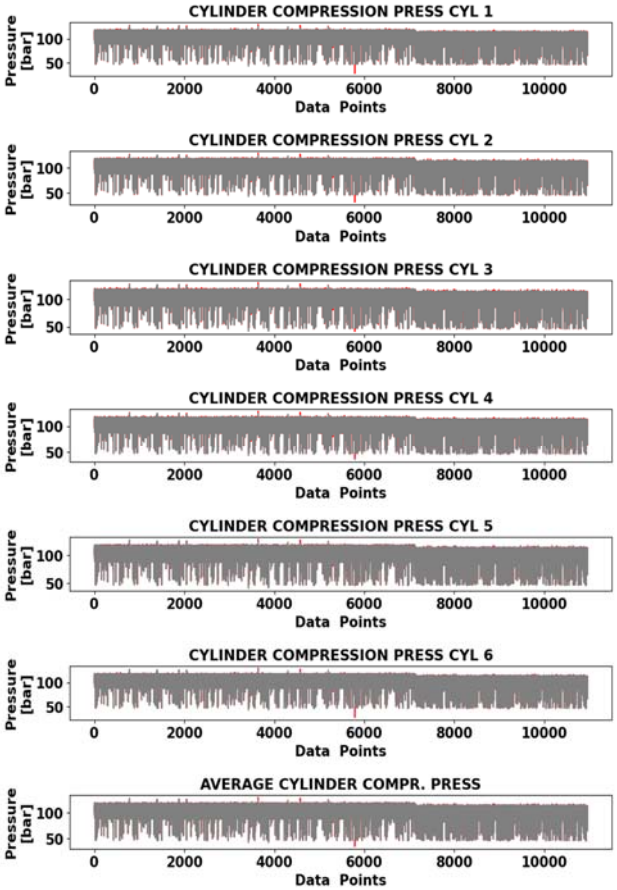

Figure 10: Line plot of CatBoost's predictions for the test set.

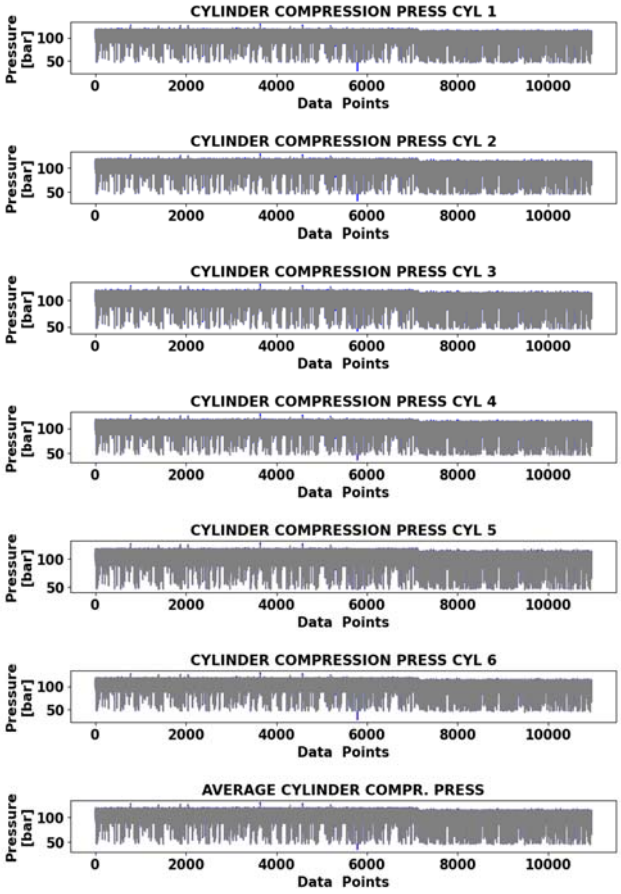

Figure 11: Line plot of LightGBM's predictions for the test set.

Referring to **Figure 11**, the predicted values of the LightGBM model also appear to overlap with most of the actual values.

6.2 Interpretation based on the SHAP

To better understand the prediction results of machine learning models, various eXplainable AI (XAI)-related techniques are being developed. Among various XAI techniques, this study used SHAP to analyze the variables affecting the predictions of CatBoost and LightGBM models.

The explainer variables were set using SHAP's KernelExplainer, the trained CatBoost and LightGBM models were used as the values of ‘model’ argument, the test set (X_test) was used as the values of ‘data’ argument, and the 'identity' was used as the value of ‘link’ argument.

The 'shap_values' function was used to train the explainer variables, and the 'X' argument used 1000 data points sampled at a constant random_state from the test set (X_test), and the 'nsamples' argument was set to 100.

SHAP provides vertical bar charts and vertical scatter plots that can be used to analyze the influence of variables through the 'summary_plot' function for explainer variables trained by the

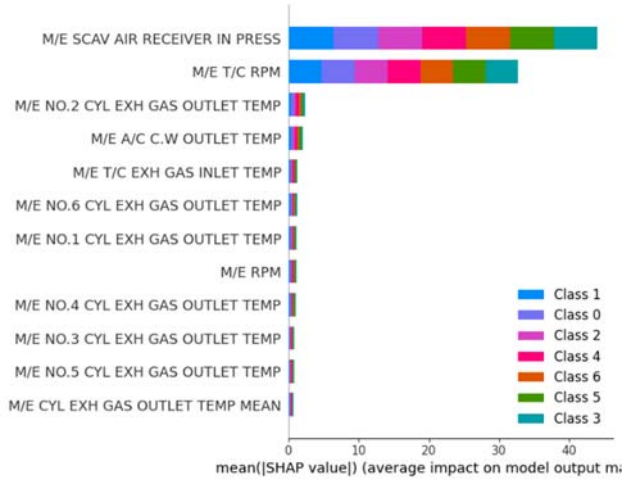

Figure 12: Vertical bar chart of CatBoost's explainer for the test set

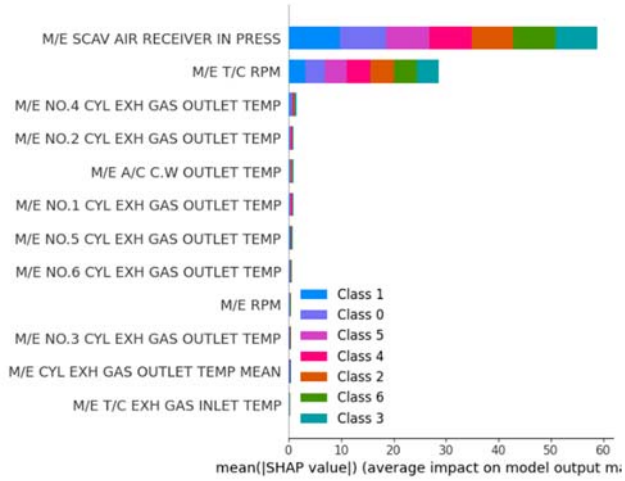

Figure 13: Vertical bar chart of LightGBM's explainer for the test set

'shap_values' function.

Figures 12 and **13** are vertical bar charts that help to analyze the influence of input variables on output variables. The x-axis represents the average absolute SHAP values, and the y-axis represents input variables. The bar of each input variable is composed of bar of output variables, and the length of the each bar is determined according to the size of the average absolute SHAP value of each output variable. As a result, the bars of the output variables were accumulated and sorted in descending order along the y-axis based on the total bar length.

Referring to **Figure 12**, the input variables 'M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM' have the greatest influence on the output variables. The two input variables have similar average absolute SHAP values between all output variables.

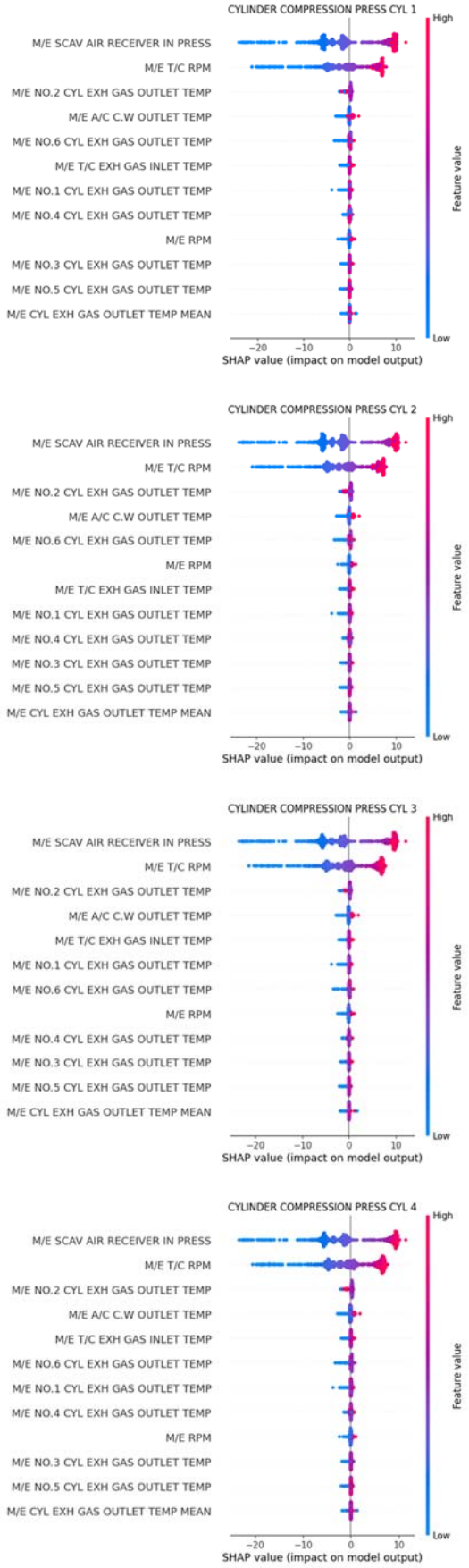

Figure 14: Vertical scatter plot 1 of CatBoost's explainer for the test set

Referring to **Figure 13**, the input variables 'M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM' also have the greatest influence on the output variables. For the LightGBM model, the 'M/E SCAV AIR RECEIVER IN PRESS' input variable has an overwhelming influence on the output variable compared to other variables. Among the remaining input variables, 'M/E A/C C.W OUTLET TEMP' and 'M/E NO.2 CYL EXH GAS OUTLET TEMP' had a high impact on the output variables in both CatBoost and LightGBM models, but their impact was minimal compared to the 'M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM' input variables.

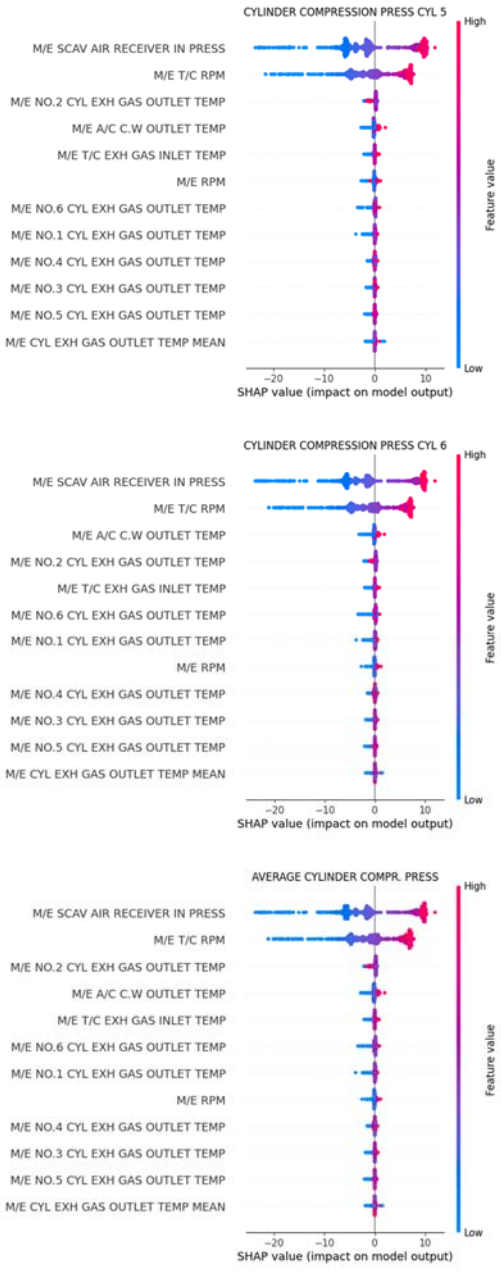

Figure 15: Vertical scatter plot 2 of CatBoost's explainer for the test set

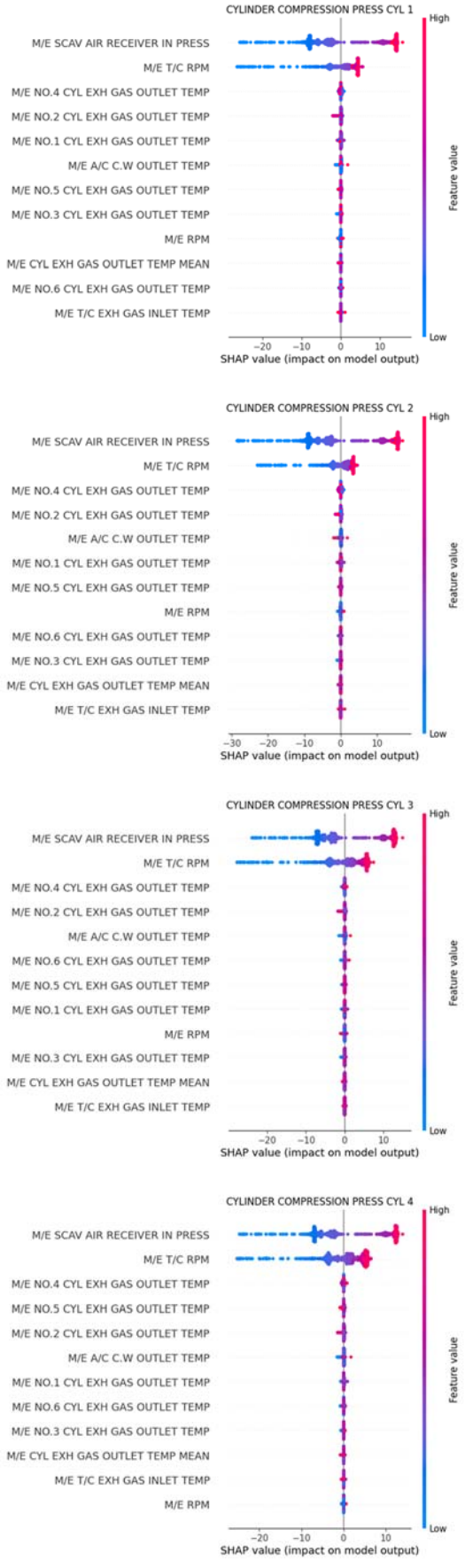

Figure 16: Vertical scatter plot 1 of LightGBM's explainer for the test set

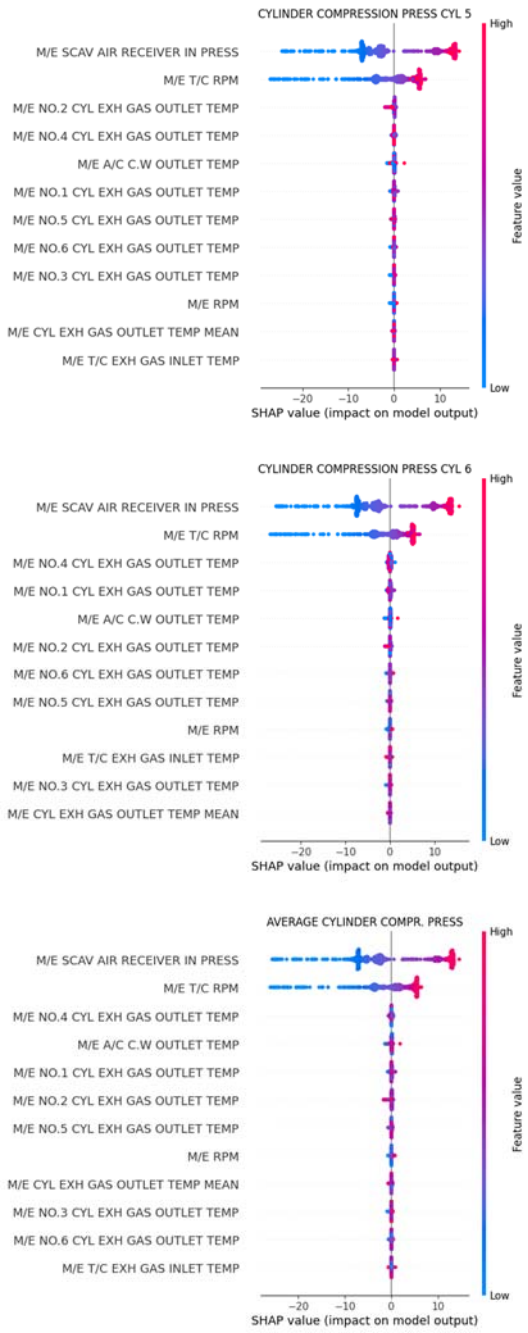

Figure 17: Vertical scatter plot 2 of LightGBM's explainer for the test set

Figures 14-17 are vertical scatter plots that help to analyze the influence of input variables on output variables. The x-axis represents the SHAP values, and the y-axis represents input variables. The scattered points for each input variable are the data points used in SHAP analysis. The color of the scattered points is red when the value of the input variable is high, and blue when it is low. The vertical positioning of scattered points at a specific SHAP value indicates that there are many scattered points corresponding to that SHAP value. A negative SHAP value means that the predicted value has decreased, and a positive SHAP value means that the predicted value has increased. Therefore, if the scattered points progress from red to blue from right to left in the same way as the color bar, the size of the input variables is highly correlated with the size of the predicted values.

Referring to **Figures 14-15**, the input variables 'M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM' have the greatest influence on all output variables for CatBoost model. The colors of the scattered points of these two variables change from right to left, from red to purple and then to blue. This is consistent with the direction from high to low of the color bar, meaning that the higher the value of the corresponding input variables, the higher the predicted value of the output variable, and the lower it is, the lower it decreases.

Referring to **Figures 16-17**, the input variables 'M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM' also have the greatest influence on all output variables for LightGBM model. For the LightGBM model, the SHAP value of scattered points concentrated in the positive range of the 'M/E SCAV AIR RECEIVER IN PRESS' input variable was greater than that of CatBoost. The scattered points in the positive range of the 'M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM' variables of the CatBoost model are within a similar range, but in the case of LightGBM, the range is far apart.

7. Conclusion

In this study, the compression pressure in the combustion chamber was predicted based on engine data acquired from the training ship, and an analysis using SHAP was performed based on the trained model used for the prediction.

This study used data from 2022, when the vessel sailed the West and East Seas of South Korea. Since we did not conduct any additional experiments after 2022, we did not collect data from other years. However, considering seasonal and operational variability, future work using data from other years will be necessary. Furthermore, as this study was conducted only on a single vessel, future work on multiple vessels will be needed.

The data acquired from the training ship underwent a preprocessing process to train the model, and in the process, variables with high correlation were selected using a heatmap.

The models used for training were CatBoost and LightGBM, which are based on GBDT. The preprocessed dataset was split into the train set and a test set, and two models were trained using the train set. Five performance metrics were used to evaluate the

performance of the two models, and it was confirmed that they performed well on both the train and test sets. The LightGBM outperformed CatBoost in all performance metrics, but the difference between the two was minimal. The prediction results of the two models were confirmed by scatter and line plots, and most of the predicted values were found to be consistent with the actual values.

Finally, based on the trained model, analysis was performed using SHAP with vertical bar charts and vertical scatter plots. The analysis results showed that the input variables 'M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM' had the greatest influence on all output variables. However, the effects of the remaining input variables were minimal. Therefore, the two input variables ('M/E SCAV AIR RECEIVER IN PRESS' and 'M/E T/C RPM') are important when building models to predict variables related to cylinder compression pressure.

The compression pressure sensor in the engine cylinder is continuously affected by high-temperature and high-pressure exhaust gas, so the performance of the sensor continuously decreases. Therefore, the predictive model we developed can serve as an auxiliary to existing sensors and can replace them in emergencies.

This study contributed to the development of a virtual model for compression pressure, which is essential for the realization of autonomous ships where sensor reliability is crucial. Future works should also include predictions up to the maximum explosion pressure.

Acknowledgement

This research was supported by the Autonomous Ship Technology Development Program [20016140] funded by the Ministry of Trade, Industry, & Energy (MOTIE, Korea); and the project titled 'Fostering Talent in Advanced Ship Blue Tech (RS-2025-02221147)' funded by the Ministry of Oceans and Fisheries, Korea.

Author Contributions

Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing - review & editing, M. H. Park; Methodology, Supervision, J. J. Hur; Methodology, Data curation, B. D. Yea; Methodology, Supervision, J. H. Choi; Conceptualization, Supervision, Project administration, Funding acquisition, Writing - review & editing, W. J. Lee.

References

[1] UNCTAD, Shipping data: UNCTAD releases new seaborne trade statistics 2025. Available: https://unctad.org/news/shipping-data-unctad-releases-new-seaborne-trade-statistics.

[2] S. Piao, J. -J. Hur, J. H. Im, M. -K. Kang, and W. -J. Lee, "Improving efficiency of diesel engine fault detection based on multi-source data," Journal of Advanced Marine Engineering and Technology, vol. 49, no. 3, pp. 130-139, 2025.

[3] M. -H. Park, S. Yeo, J. -H. Kim, J. -H. Choi, W. -J. Lee, "Comprehensive review on recent progress in renewable and sustainable energy applications in shipping industry, and suggestions for future developments," Renewable and Sustainable Energy Review, vol. 225, 116152, 2026.

[4] M. -H. Park, J. -H. Choi, and W. -J. Lee, "Object detection for various types of vessels using the YOLO algorithm," Journal of Advanced Marine Engineering and Technology, vol. 48, no. 2, pp. 81-88, 2024.

[5] M. -H. Park, Q. D. Vuong, J. -J. Hur, B. -D. Yea, and W. -J. Lee, "Mapping predicted carbon dioxide emissions from ships using gradient-boosting-based models," Journal of Advanced Marine Engineering and Technology, vol. 48, no. 4, pp. 177-185, 2024.

[6] M. -A. Je, S. -H. Jung, T. Y. Jeong, S. C. Hwang, J. -S. Moon, "Exhaust Gas Recirculation (EGR) effect on two-stroke diesel engines under sailing condition," Journal of Advanced Marine Engineering and Technology, vol. 48, no. 1, pp. 1-8, 2024.

[7] Y. Tang, J. Zhang, H. Gan, B. Jia, and Y. Xia, "Development of a real-time two-stroke marine diesel engine model with in-cylinder pressure prediction capability," Applied Energy, vol. 194, pp. 55-70, 2017.

[8] D. T. Hountalas, "Prediction of marine diesel engine performance under fault conditions," Applied Thermal Engineering, vol. 20, no. 18, pp. 1753-1783, 2000.

[9] C. Patil, G. Theotokatos, and K. Milioulis, "In-cylinder pressure prediction for marine engines using machine learning," SNAME 8th International Symposium on Ship Operations, Management and Economics, 2023.

[10] H. Shen, J. Zhang, B. Yang, and B. Jia, "Development of a marine two-stroke diesel engine MVEM with in-cylinder pressure trace predictive capability and a novel compressor model," Journal of Marine Science and Engineering, vol. 8, no. 3, 2020.

[11] C. Patil, G. Theotokatos, and K. Tsitsilonis, "Data-driven model for marine engine fault diagnosis using in-cylinder pressure signals," Journal of Marine Engineering & Technology, vol. 24, pp. 70-82, 2024.

[12] I. Galliakis, “Prediction of peak cylinder pressure of a four-stroke marine diesel engine using neural networks”, Diploma Thesis, School of Naval Architecture and Marine Engineering, National Technical University of Athens, Greece, 2022.

[13] I. Asimakopoulos, L. D. Avendaño-Valencia, M. Lützen, N. G. M. Rytter, "Data-driven condition monitoring of two-stroke marine diesel engine piston rings with machine learning," Ships and Offshore Structures, vol. 19, no. 9, pp. 1241-1253, 2023

[14] K. M. Tsitsilonis and G. Theotokatos, "A novel method for in-cylinder pressure prediction using the engine instantaneous crankshaft torque," Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment, vol. 236, no. 1, 2021.

[15] M. H. Park, J. J. Hur, and W. J. Lee, "Prediction of oil-fired boiler emissions with ensemble methods considering variable combustion air conditions," Journal of Cleaner Production, vol. 375, 2022.

[16] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," Journal of Big Data, vol. 7, 94, 2020.

[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, “Catboost: Unbiased boosting with categorical features,” Advances in neural information processing systems, 31, 2018.

[18] M. -H. Park, J. -J. Hur, and W. -J. Lee, "Prediction of diesel generator performance and emissions using minimal sensor data and analysis of advanced machine learning techniques," Journal of Ocean Engineering and Science, vol. 10, pp. 150-168, 2025.

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, *et al*., “LightGBM: A highly efficient gradient boosting decision tree,” Advances in neural information processing systems, 30, 2017.

[20] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” Advances in neural information processing systems, 30, 2017.

[21] J. Zhang, X. Ma, J. Zhang, D. Sun, X. Zhou, C. Mi, and H. Wen, "Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model," Journal of Environmental Management, vol. 332, 2023.