Proceedings of MARTEC 2022
The International Conference on Marine Technology
21-22 December 2022, BUET, Dhaka, Bangladesh

A STEP TOWARDS IMO GREENHOUSE GAS REDUCTION GOAL: EFFECTIVENESS OF MACHINE LEARNING BASED CO_2 EMISSION PREDICTION MODEL

Ishrar Israil Monisha[1], Nafisa Mehtaj [2], Zobair Ibn Awal [3]

Department of Naval Architecture and Marine Engineering
Bangladesh University of Engineering and Technology (BUET),
Dhaka-1000, Bangladesh.
[1]E-mail: ishrarmonisha@name.buet.ac.bd
[2]E-mail: nafisamehtaj@gmail.com
[3]E-mail: zobair@name.buet.ac.bd

ABSTRACT

Ships are the world's most economical means of freight transportation, and day by day, it is expanding quickly. The increase in ship transportation activities has resulted in a significant concern about CO_2 emissions. International Maritime Organization has agreed to set a goal of reducing the maritime sector's total gas emissions by at least 50% by 2050. In this regard, a CO_2 emission prediction model followed by an emission inventory can play a vital role in decision-making to optimize the ship's speed, draft, trim, and other influencing parameters under Ship Energy Efficiency Management Plan to decrease carbon emissions during operation. Machine learning, a branch of the data science approach, can be utilized to create effective emission-prediction models. In this research, two machine-learning models have been developed using actual voyage data collected from the noon reports of ships in Bangladesh. The models have been trained with the ship's speed, engine rpm, wind force, and sea condition during voyages. The models' performances have been assessed employing the Coefficient of Determination (R^2) and Root Mean Square Error (RMSE). The prediction accuracies for the K Nearest Neighbor Regression model and the Light Gradient Boosted Machine Regression model are 84% and 81% with RMSE of 5.12 and 5.53, respectively.

Keywords: *Machine learning, CO_2 emission prediction, Maritime transportation.*

1. INTRODUCTION

The maritime transportation sector is responsible for over 80% of the global merchandise exchange [1]. With the increase in shipping transportation activities, the possibility of environmental pollution due to greenhouse gas (GHG) emissions from ship operations exists concurrently. As per the fourth GHG study of the International Maritime Organization (IMO), the total shipping emitted 962 million tonnes of CO_2 in 2012, increasing by 9.3% to 1,056 million tonnes in 2018 [2]. It demonstrates that CO_2 emissions from maritime transportation are rising steadily, which is influencing global emissions in a significant manner. So, any actions taken to lessen GHG emissions should concentrate primarily on CO_2. "IMO Initial Strategy" was announced in 2018 as an emission reduction policy by IMO to decrease carbon emissions by 70% and yearly GHG emissions by a minimum of 50% by 2050 relative to the 2008 baseline [3]. From the analysis, it is evident that a CO_2 emission inventory can show decision-makers the way forward in developing and assessing the execution of applicable regulations to achieve the IMO's goal regarding the emission reduction strategy [4]. Emission inventories contain essential information on the existing condition of the functional area and represent the potential to understand the impacts of the conducted activities. Therefore, assessing the CO_2 emissions from ships by generating a comprehensive emission inventory is vital, which is currently limited in Bangladesh.

Emission inventory to estimate the volume of pollutants released into the atmospheric environment has been subjected to several analyses. The top-down and bottom-up processes are the conventional techniques for developing ship emission inventories. Top-down approaches have been used by Goldsworthy and Goldsworthy [5], Gusti and Semin [6], etc. As this approach is fuel-based, it performs by using highly integrated data on fuel consumption by ship type, gross tonnage, engine type, navigational phase, and emission factors [7]. Bottom-up approaches have been implemented, including in MOPSEA by Vangheluwe et al. [8], EMS by Denier van der Gon and Hulskotte [9], etc. This approach uses individual vessel activity data and technical specifications, including the operation time, engine power, load factor, emission factors of engines in all

navigational phases, specific fuel consumption, gross tonnage based on vessel type, etc., to estimate the emissions by location.

Besides determination, predicting carbon emissions based on influencing factors is also an essential topic. Before the operation, if CO_2 emission can be estimated and the operators can know the ships' emission inventories, they can ensure voyages with less CO_2 emission. Machine learning being an advanced section of data-science methods, the researchers have leveraged its advantages to perform CO_2 emission forecasting. Lepore et al. [10] predicted CO_2 emissions of a Ro-Pax cruise ship by implementing Multiple Linear Regression, LASSO Regression, Random Forest Regression, Principal Component Regression, etc. An emission inventory model to determine the gaseous emissions from two cargo ships was created by Fletcher et al. [11] utilizing five machine learning algorithms based on engine power, shaft speed, and emission of gaseous pollutants, including CO_2. To date, multiple studies have been conducted in the diverse frameworks in Bangladesh implementing machine learning, including flood damage analysis [12], atmospheric particulate matter concentration prediction [13], etc.

CO_2 emission prediction models based on machine learning can lead to productive emission inventories, which are, till now, an unexplored area in Bangladesh. This paper aims to develop an effective machine-learning model through a comparative analysis to predict CO_2 emissions from ships in Bangladesh.

2. METHODOLOGY

2.1 Research Framework

The current research comprises data acquisition, data pre-processing, application of machine learning algorithms, hyperparameters optimization, and model evaluation. Figure 1 is a visual representation of the established methodology for CO_2 emission prediction of the current study.

Relevant data were collected from the ships' noon reports to develop an efficient model to perform prediction on the CO_2 emission from the ships of Bangladesh. The data has been analyzed and pre-processed to feed two machine-learning models based on the algorithms named K Nearest Neighbor Regression and Light Gradient Boosted Machine Regression. Iterations have been performed by optimizing the hyperparameters of these models to come up with better accuracy. Eventually, the developed models were evaluated according to their accuracy to identify the most effective one.

Data Acquisition and Pre-processing

Two years of operational data of four bulk carriers of Bangladesh have been collected from 823 noon reports.

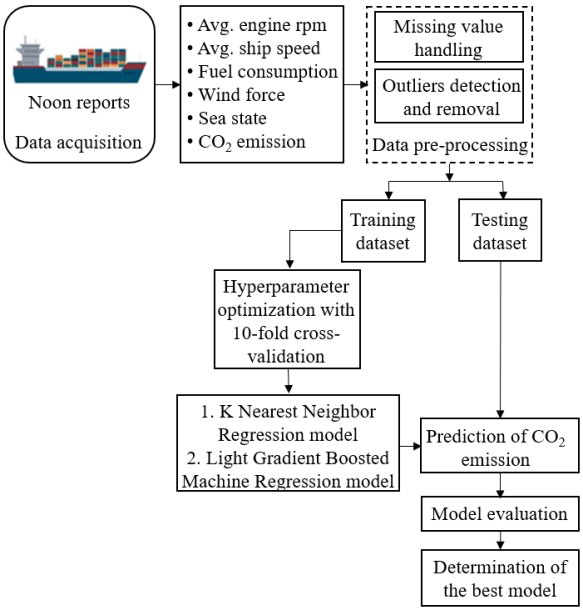

Figure 1: Proposed methodology for CO_2 emission prediction

The noon reports contain names of the vessels, position, voyage no., average ship speed, distance to destination, estimated time of arrival, average engine rpm, amount of fresh water, fuel oil, lube oil, etc., remaining on the board, total cargo carried, fuel consumption, weather and sea condition variables, and many more. From the noon report, features influencing CO_2 emission the most have been taken, which include the average engine rpm, ship speed, fuel consumption, wind force, and sea state. In this study, CO_2 emission has been determined following the rule of IMO [14], which has been stated in Equation 1.

$$CO_2 \text{ emission} = \text{Fuel Consumption} \times C_F \quad (1)$$

Here, the carbon emission factor (C_F) varies depending on the fuel type. C_F value for marine diesel/gas oil (MDO/MGO) is 3.206, and for heavy fuel oil (HFO) is 3.114 (tonnes-CO_2/tonnes-Fuel).

Samples with at least one or more missing features have been removed from the dataset. The dataset also contains outliers which negatively affect the model training process and result in lower accuracy. The statistics of the entire data distribution have been illustrated in the form of box plots in Figure 2. The samples located below [25th percentile - 1.5 (75th percentile - 25th percentile)] or above [75th percentile + 1.5 (75th percentile - 25th percentile)] [15] of the box plots have been identified as outliers (marked by black circles) and removed from the

dataset. Performing data pre-processing leads to the dataset containing a total of 397 samples.

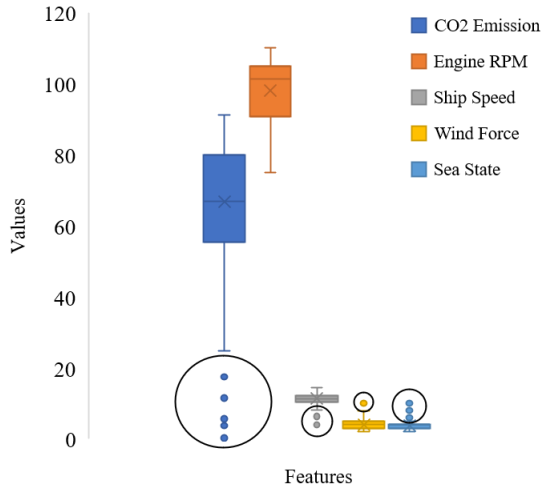

Figure 2: Outliers detection in the statistics of data distribution

2.2 Machine Learning Algorithms

2.2.1 K Nearest Neighbor Regression

The first model in this research has been developed using the K Nearest Neighbor Regression (KNNR) algorithm, where k signifies the number of closest data samples of a new point x_q. In this method, all the distance between x_q and other points x_j are measured using Minkowski distance (L^P) stated in Equation 2 where P = 1 denotes Manhattan distance and P = 2 denotes Euclidean distance. The value of k is selected after multiple iterations to estimate the weighted average values of the k nearest data points to predict the continuous output for the query point, as represented in Figure 3, utilizing the emission values of the current study.

$$L^P(x_j, x_q) = \left(\sum |x_{j,i} - x_{q,i}|^P\right)^{1/P} \quad (2)$$

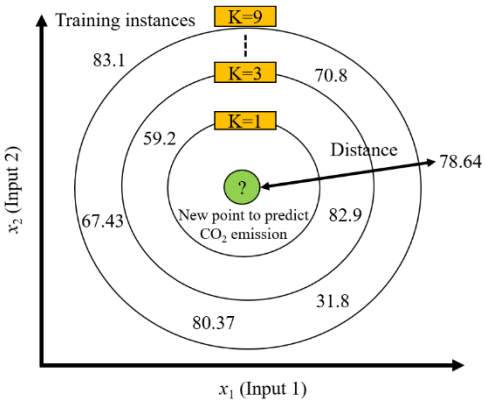

Figure 3: K Nearest Neighbor Regression [16]

2.2.2 Light Gradient Boosted Machine Regression

Light Gradient Boosted Machine Regression (LGBMR) is the second algorithm utilized in this study. LGBMR, as presented based on the parameters of the current study in Figure 4, provides an efficient and effective implementation of the gradient-boosting algorithm based on Decision Trees (DT).

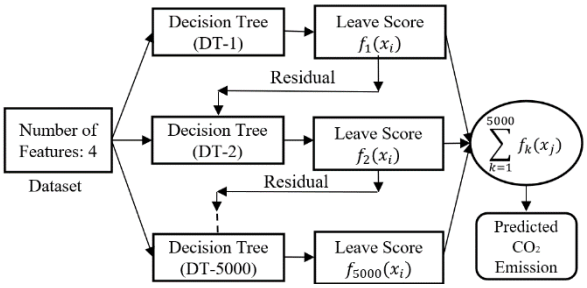

Figure 4: Light Gradient Boosted Machine Regression [17]

The gradient boosting algorithm combines the weak learners consecutively to get a strong learner so that every new learner fits the residuals from the preceding stage, improving the model. LGBMR was created to speed up the training. It adds dynamic feature extraction to expand the gradient-boosting algorithm. Predictions of all the trees are added to get the estimation as stated in Equation 3.

$$\hat{y}_i = \sum_{i=1}^{k} \left[\underset{f_k}{argmin} \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(k)}) + \Omega(f_k)\right](x_i) \quad (3)$$

Here, $f_k \in F$, k is the tree number, F is a space containing all possible tree structures, f_k is one of the trees with the leaf score, and Ω is the regularization function. L is the training loss function which can be expressed as stated in Equation 4 in the case of a squared error loss function, where f_k is obtained by fitting the residual r.

$$L\left(y, \hat{y}^{(k-1)} + f_k(x)\right) = \left[y - \hat{y}^{(k-1)} - f_k(x)\right]^2 = \left[r - f_k(x)\right]^2 \quad (4)$$

2.3 Hyperparameter Optimization

Both machine learning algorithms used in the study utilize a variety of hyperparameters (model configuration variables). The random search method incorporating repeated 10-fold cross-validation has been employed in the model-building stage. This task aims to randomly choose hyperparameters from the provided set and generate the best possible combination. This method has been widely employed in different research areas and identified as resistant to overfitting. Table 1 represents the provided set of the hyperparameters to be tested and the obtained optimal values for each model.

Table 1. Hyperparameter optimization results

Algorithm	Hyperparameter (s)	Provided Values	Optimal Values
KNNR	n_neighbors	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21	9
	weights	uniform, distance	distance
	metric	euclidean, manhattan, minkowski	minkowski
LGBMR	n_estimators	10, 50, 100, 500, 1000, 5000	5000
	max_depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	7
	learning_rate	0.00001, 0.001, 0.01, 0.1, 1	0.01
	boosting_type	gbdt, dart, goss, rf	goss
	num_leaves	10, 20, 30, 40, 50, 60, 70, 80, 90	20

3. RESULTS AND DISCUSSION

3.1 Model Evaluation

The efficacy of the prediction models has been examined utilizing two evaluation metrics: Coefficient of Determination (R^2) and Root Mean Square Error (RMSE). R^2 denoting the fitness of data with the model, has been implemented to quantify the models' prediction accuracy. The range of R^2 is between 0 and 1, where values near 1 reflect the higher effectiveness of the prediction model. R^2 has been calculated using Equation 5. RMSE, as expressed in Equation 6, is the square root of the average squared distance between expected and predicted results.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

3.2 Model Comparison

In this research, 317 data entries have been randomly determined as the training dataset to comprehend the characteristics of the data samples for the prediction models. The rest 80 samples have been used to test the models. Each of the two models being trained using the optimal hyperparameters has been evaluated on the testing dataset, with each evaluation metric described in Section 3.1.

Figure 5 depicts the comparison results of the RMSE and R^2 of the two models. From the results presented, it can be deduced that the best-performing model is KNNR achieving R^2 of 0.84 and RMSE of 5.12. LGBMR has also yielded comparable R^2 and RMSE, which are 0.81 and 5.53, correspondingly. The low RMSE of the KNNR model means that it has a better fitting performance and can predict CO_2 emissions in a comparatively accurate manner under different navigational conditions.

Figures 6 and 7 visualize the comparison between the actual and predicted CO_2 emissions for KNNR and LGBMR models, respectively. As expected, the KNNR model has outperformed the LGBMR model and has fit the actual values with greater accuracy.

Here, the better performance of the KNNR model is due to the dataset size and the outliers handling. KNNR is a comparatively slow learning algorithm that gathers all data samples before making decisions at execution time. In addition, outliers affect the algorithm significantly as it gets all the information from the input rather than from an algorithm that tries to generalize data. Hence, KNNR has worked well with the small dataset of this study from which the outliers have been removed.

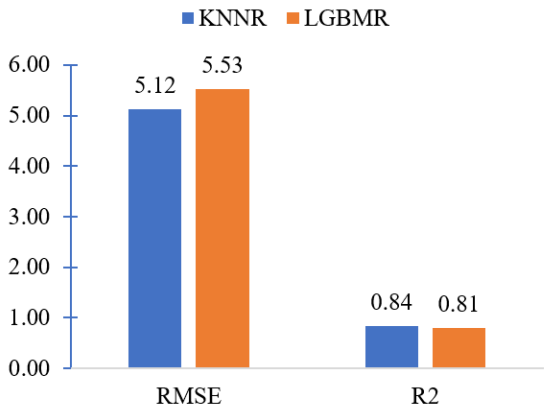

Figure 5: Model comparison based on RMSE and R^2

4. CONCLUSION

This paper presents a comparison-based study of two machine learning models to forecast CO_2 emissions. The developed models can predict the CO_2 emissions from oceangoing ships of Bangladesh using the ships' actual operational data obtained from noon reports. From Section 3.2, it can be seen the predicted result and actual value are in good agree-

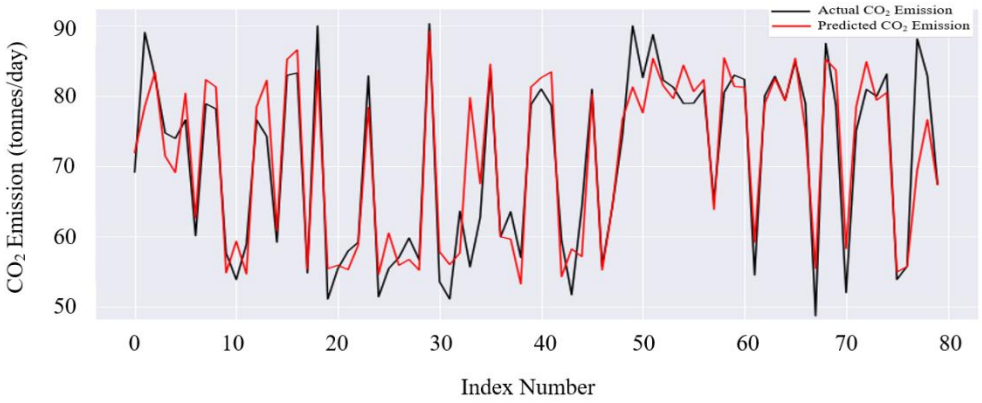

Figure 6: Actual and forecasted CO_2 emission comparison by K Nearest Neighbor Regression (KNNR) model

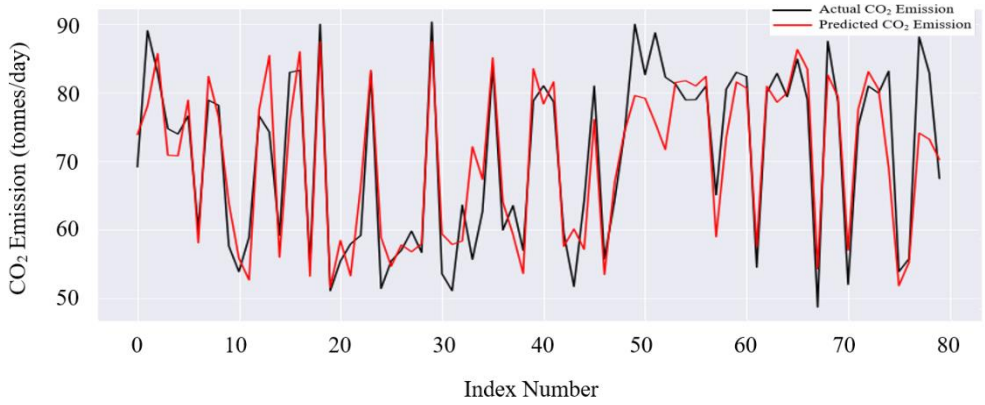

Figure 7: Actual and forecasted CO_2 emission comparison by Light Gradient Boosted Machine Regression (LGBMR) model

ment for both models, where the model using the KNNR algorithm outperformed the one with the LGBMR algorithm. Though in Bangladesh, there exist several machine-learning based researches on different areas, the task of predicting a ship's CO_2 emission implementing machine learning has not been performed here so far. Accordingly, for the first time in Bangladesh, the amount of CO_2 emitted from the ships has been forecasted based on machine learning algorithms in the current study. Due to the shortcoming of the ship's operational data availability in Bangladesh, complex machine learning algorithms could not be used in this study to get predicted outcomes with more accuracy. The operational data is not stored in a structured way in most of the ships in Bangladesh. Moreover, voyage data providing digital devices are not installed in several ships, which results in a lacking of an adequate amount of relevant data required for studies. Hence, the current study recommends connecting more workable digital devices having voyage data storage facilities with the ships in Bangladesh. In the future, this research can be upgraded by including more data, allowing it to develop advanced machine-learning models and make them flexible to use in a wide range of ships.

REFERENCES

[1] "World Investment Report 2021," United Nations Conference on Trade and Development (UNCTAD), 2021. Available: https://unctad.org/webflyer/world-investment-report-2021 [Last accessed 20 November 2022].

[2] "Fourth Greenhouse Gas Study," International Maritime Organization (IMO), 2020. Available: https://www.imo.org/en/OurWork/Environment/Pages/Fourth-IMO-Greenhouse-Gas-Study-2020.aspx [Last accessed 20 November 2022].

[3] "Initial IMO GHG Strategy," International Maritime Organization (IMO). Available: https://www.imo.org/en/MediaCentre/HotTopics/Pages/Reducing-greenhouse-gas-emissions from-ships.aspx [Last accessed 20 November 2022].

[4] Álvarez, P. S., "From maritime salvage to IMO 2020 strategy Two actions to protect the environment," *Marine Pollution Bulletin*, Vol. 170, pp. 1-12 (2021).

[5] Goldsworthy, L., and Goldsworthy, B., "Modelling of ship engine exhaust emissions in ports and extensive coastal waters based on terrestrial AIS data - An Australian case study," *Environmental Modelling & Software*, Vol. 63, pp. 45-60 (2015).

[6] Gusti, A. P., and Semin, "The effect of vessel speed on fuel consumption and exhaust gas emissions," *American Journal of Engineering and Applied Sciences*, Vol. 9, No. 4, pp. 1046-1053 (2016).

[7] Ay, C., Seyhan, A., and Beşikçi**,** E. B., "Quantifying ship-borne emissions in Istanbul Strait with bottom-up and machine-learning approaches," *Ocean Engineering*, Vol. 258, pp. 1-13 (2022).

[8] Vangheluwe, M., Mees, J., and Janssen, C., "Monitoring programme on air pollution from sea-going vessels (MOPSEA)," (2007).

[9] Denier van der Gon, H., and Hulskotte, J., "Methodologies for estimating shipping emissions in the Netherlands," (2010).

[10] Lepore, A., Reis, M. S., Palumbo, B., and Capezza, C., "A comparison of advanced regression techniques for predicting ship CO_2 emissions," *Quality and Reliability Engineering International*, Vol. 33, pp. 1281-1292 (2017).

[11] Fletcher, T., Garaniya, V., Chai, S., Abbassi, R., Yu, H., Van, C. T., Brown, R. J., and Khan, F., "An application of machine learning to shipping emission inventory," *International Journal of Maritime Engineering*, Vol. 160 (Part A4), pp. 381-395 (2018).

[12] Ganguly, K. K., Nahar, N., and Hossain, B. M. M., "A machine learning-based prediction and analysis of flood affected households: A case study of floods in Bangladesh," *International Journal of Disaster Risk Reduction*, Vol. 34, pp. 283-294 (2019).

[13] Shahriar, S. A., Kayes, I., Hasan, K., Salam, M. A., and Chowdhury, S., "Applicability of machine learning in modeling of atmospheric particle pollution in Bangladesh," *Air Quality, Atmosphere & Health*, Vol. 13, pp. 1247-1256 (2020).

[14] "Guidelines for Voluntary Use of the Ship's Energy Efficiency Operational Indicator (EEOI)," International Maritime Organization (IMO), 2009. Available: https://gmn.imo.org/wp-content/uploads/2017/05/Circ-684-EEOI-Guidelines.pdf. [Last accessed 20 November 2022].

[15] Zhang, Q., Sun, B., Wang, J., Liu, W., and Yu, F., "Development and validation of a novel cell-based assay for the detection of neutralizing antibodies of Aflibercept," *Frontiers in Drug, Chemistry and Clinical Research*, Vol. 2, pp. 1-6 (2019).

[16] Tran, H., "A survey of machine learning and data mining techniques used in multimedia system," ResearchGate [Preprint], September 2019. Available: https:// 10.13140/RG.2.2.20395.49446/1 [Last accessed 20 November 2022].

[17] Li, F., Zhang, L., Chen, B., Gao, D., Cheng, Y., Zhang, X., Yang, Y., Gao, K., Huang, Z., and Peng, J., "A light gradient boosting machine for remainning useful life estimation of aircraft," *21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3562-3567 (2018).