



A STEP TOWARDS IMO GREENHOUSE GAS REDUCTION GOAL: EFFECTIVENESS OF MACHINE LEARNING BASED CO₂ EMISSION PREDICTION MODEL

I. I. Monisha^{1*}, N. Mehtaj² and Z. I. Awal³

^{1*}Department of Naval Architecture and Marine Engineering, BUET, Dhaka-1000, ishrmonisha@name.buet.ac.bd

²Department of Naval Architecture and Marine Engineering, BUET, Dhaka-1000, nafisamehtaj@gmail.com

³Department of Naval Architecture and Marine Engineering, BUET, Dhaka-1000, 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 CO₂ 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 CO₂ 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 (R²) 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, CO₂ emission prediction, maritime transportation, ship's energy efficiency.

NOMENCLATURE

C_F	carbon emission factor	Ω	regularization function
L^P	minkowski distance	L	training loss function
t	tree number	r	residual
F	space with tree structure	R^2	coefficient of determination
f_t	tree with leaf score	RMSE	root mean square error

1. Introduction

The maritime transportation sector is responsible for over 80% of the global merchandise exchange (World Investment Report, 2021). 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), total shipping emitted 962 million tonnes of CO₂ in 2012, increasing by 9.3% to 1,056 million tonnes in 2018 (IMO 4th GHG Study, 2020). It demonstrates that CO₂ 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 CO₂. "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 (Initial IMO GHG Strategy, 2018). From the analysis, it is evident that a CO₂ 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 (Alvarez, 2021). 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 CO₂ 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 (2015), Gusti and Semin (2016), 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 (Ay *et al.*, 2022). Bottom-up approaches have been implemented, including in MOPSEA by Vangheluwe *et al.* (2007), EMS by Denier van der Gon and Hulskotte (2010), 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 CO₂ emission can be estimated and the operators can know the ships' emission inventories, they can ensure voyages with less CO₂ emission. Machine learning being an advanced section of data-science methods, the researchers have leveraged its advantages to perform CO₂ emission forecasting. Lepore *et al.* (2017) predicted CO₂ 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.* (2018) utilizing five machine learning algorithms based on engine power, shaft speed, and emission of gaseous pollutants, including CO₂. To date, multiple studies have been conducted in the diverse frameworks in Bangladesh implementing machine learning, including flood damage analysis (Ganguly *et al.*, 2019), atmospheric particulate matter concentration prediction (Shahriar *et al.*, 2020), etc.

CO₂ 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 CO₂ 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. Fig. 1 is a visual representation of the established methodology for CO₂ 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 CO₂ 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.

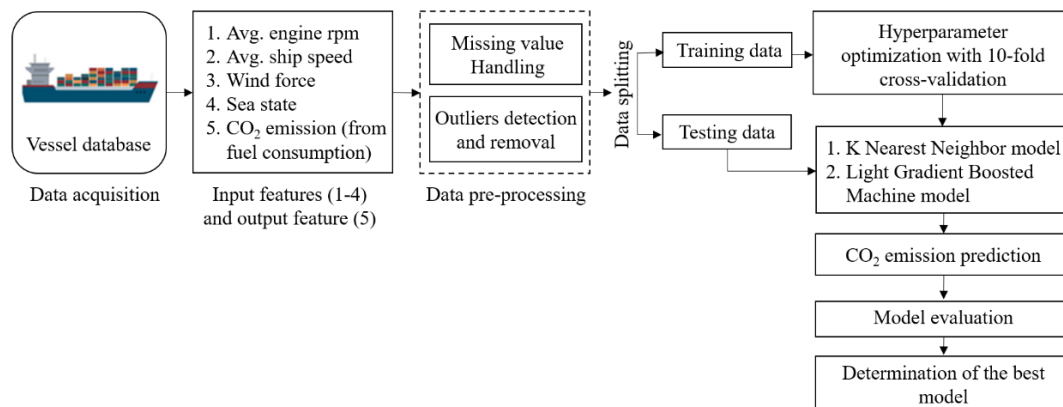


Fig. 1: Proposed methodology for CO₂ emission prediction

2.2 Data acquisition and pre-processing

Two years of operational data of four bulk carriers of Bangladesh have been collected from 823 noon reports. 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 CO₂ emission the most have been taken, which include the average engine rpm, ship speed, fuel consumption, wind force, and sea state. In this study, CO₂ emission has been determined following the rule of IMO (IMO Guidelines for Voluntary Use of EEOI, 2009), which has been stated in Eq. (1).

$$\text{CO}_2 \text{ emission} = \text{Fuel Consumption} \times C_F \tag{1}$$

Here, the carbon emission factor (C_F) varies depending on the fuel type. C_F value for marine diesel oil/marine gas oil (MDO/MGO) is 3.206, and for heavy fuel oil (HFO) is 3.114 (tonnes-CO₂/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 Fig. 2. The samples located below [25th percentile - 1.5 (75th percentile - 25th percentile)] or above [75th percentile + 1.5 (75th percentile - 25th percentile)] (Zhang *et al.*, 2019) 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.

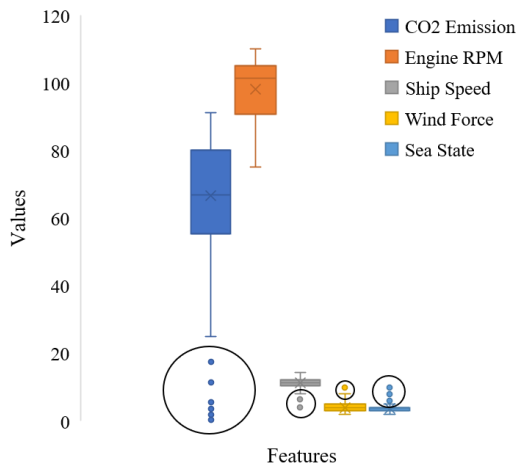


Fig. 2: Outliers detection in the statistics of data distribution

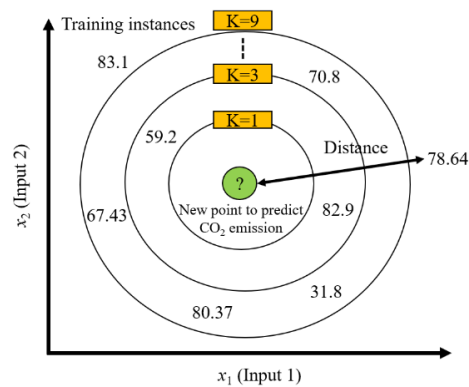


Fig. 3: K Nearest Neighbor Regression (Tran, 2019)

2.3 Machine learning algorithms

2.3.1 K Nearest Neighbor Regression

The first model in this research has been developed using the K Nearest Neighbor Regression (KNNR) algorithm. As the title implies, the algorithm utilizes k nearest data points to estimate the continuous output for a new data sample. This algorithm differs from other methods in that it conducts an action on the dataset while distinguishing rather than training the data using the system’s previously obtained dataset. This method performs a distance-based calculation to identify the k nearest neighbors to a new data point (x_q), where k is a user-selected parameter, as represented in Fig. 3. The figure is produced utilizing the emission values of the current study.

After defining the value of k, all points along with the point x_q are considered in an n-dimensional space. The distances of x_q from all other points are then calculated. The spacing between x_q and any other points (x_j) can be calculated using the Minkowski distance (L^P) as stated in Eq. (2), where $P = 1$ denotes Manhattan distance and $P = 2$ denotes Euclidean distance.

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

The k points (neighbors) with the smallest distances are selected after sorting the distances of all points. The final output for x_q is estimated using the weighted mean of its k closest neighbors.

2.3.2 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 Fig. 4, provides an efficient and effective implementation of the gradient-boosting algorithm based on decision trees. In the gradient-boosting algorithm, the weak learners are consecutively merged in such a manner that every new learner fits the residuals from the preceding stage, improving the model. The final model combines the outcomes of each step to get a strong learner. The decision tree acts as a weak learner in this algorithm. LGBMR was created to speed up the training. It adds dynamic feature extraction to expand the gradient-boosting algorithm. Significantly, adding a new tree does not change the model’s already-existing trees, and the added one fits the current model’s residuals. This strategy of tree construction enables to reduce the errors at each subsequent stage.

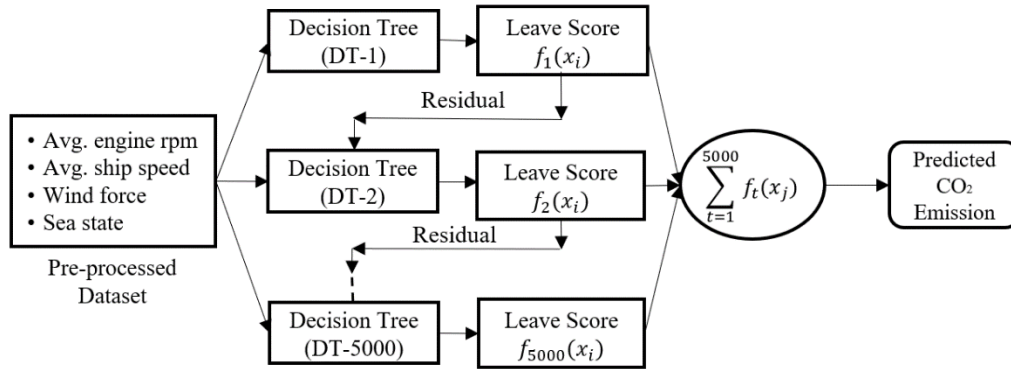


Fig. 4: Light Gradient Boosted Machine Regression (Li et al., 2018)

Predictions of all the trees are added to get the estimation as stated in Eq. (3).

$$\hat{y}_i = \sum_{t=1}^t [\underset{f_t}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, \hat{y}_i^{(t)}) + \Omega(f_t)] (x_i) \tag{3}$$

Here, $f_t \in F$, t is the tree number, F is a space containing all possible tree structures, f_t 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 Eq. (4) in the case of a squared error loss function, where f_t is obtained by fitting the residual r .

$$L(y, \hat{y}^{(t-1)} + f_t(x)) = [y - \hat{y}^{(t-1)} - f_t(x)]^2 = [r - f_t(x)]^2 \tag{4}$$

2.4 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
K Nearest Neighbor Regression (KNNR)	n_neighbors	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21	9
	weights	uniform, distance	distance
	metric	euclidean, manhattan, minkowski	minkowski
Light Gradient Boosted Machine Regression (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

Among the hyperparameters of the KNNR model mentioned in Table 1, n_neighbors denote neighbors' number to utilize in the queries for k nearest neighbors. In prediction, weight is used as a function where its value 'uniform' refers to the equal weight of all points in every neighborhood and distance' refers to weighing the points based on the inverse of distances among them. Besides, metrics are applied to measure the distance of a new data point from any other point; its possible types are: 'minkowski', 'manhattan', and 'euclidean'.

From the LGBMR model's hyperparameters, n_estimators denote the boosted trees' number to fit. max_depth refers to the maximum tree depth for base learners. num_leaves are the maximum tree leaves for base learners. learning_rate regulates how much each model contributes to the prediction. boosting_type includes a number of different boosting algorithms, whose possible values are: gbdt, dart, rf, and goss. gbdt means conventional Gradient Boosting Decision Tree, dart means Dropouts meet Multiple Additive Regression Trees, rf is Random Forest, and goss refers to Gradient-based One-Side Sampling.

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 Eq. (5). RMSE, as expressed in Eq. (6), is the square root of the average squared distance between expected and predicted results.

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

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

Here, n is the total data samples in the testing set, y_i and \hat{y}_i represent actual and predicted CO₂ emission, respectively, and \bar{y} ($= \frac{1}{n} \sum_{i=1}^n y_i$) is the mean value of y .

3.2 Model comparison and validation

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 R^2 and RMSE.

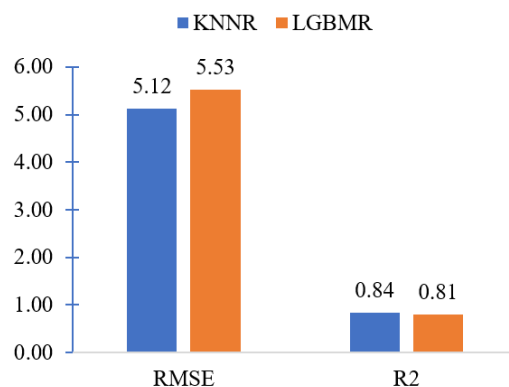


Fig. 5: Model comparison based on RMSE and R^2

Fig. 5 depicts the comparison results of the R^2 and RMSE of the two models. From the results presented, it can be observed that the KNNR model has provided the R^2 of 0.84 and RMSE of 5.12, whereas LGBMR has also yielded comparable R^2 and RMSE, which are 0.81 and 5.53, correspondingly. The relatively high R^2 and low RMSE

values of the KNNR model initially indicate that it can predict CO₂ emissions in a comparatively accurate manner under different navigational conditions.

Fig. 6 and Fig. 7 visualize the comparison between the actual and predicted CO₂ emissions for KNNR and LGBMR models, respectively. As expected from the results of R² and RMSE, the KNNR model has initially 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.

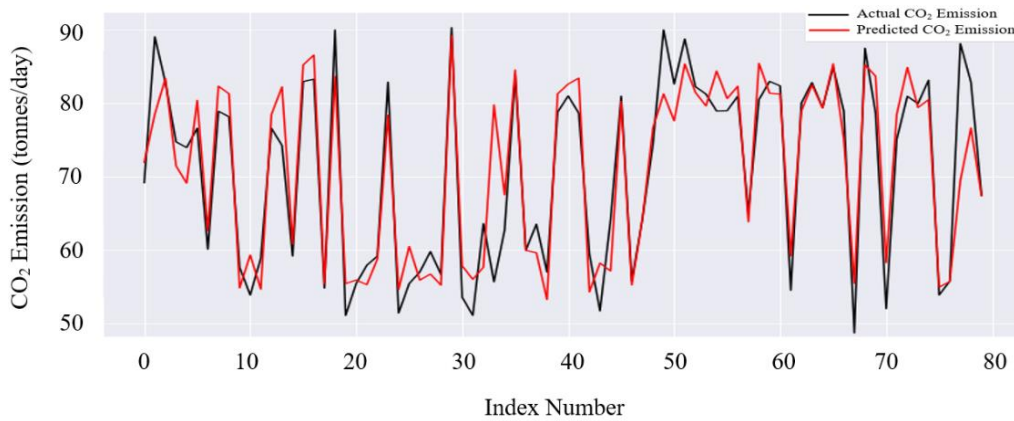


Fig. 6: Actual and forecasted CO₂ emission comparison by K Nearest Neighbor Regression (KNNR) model

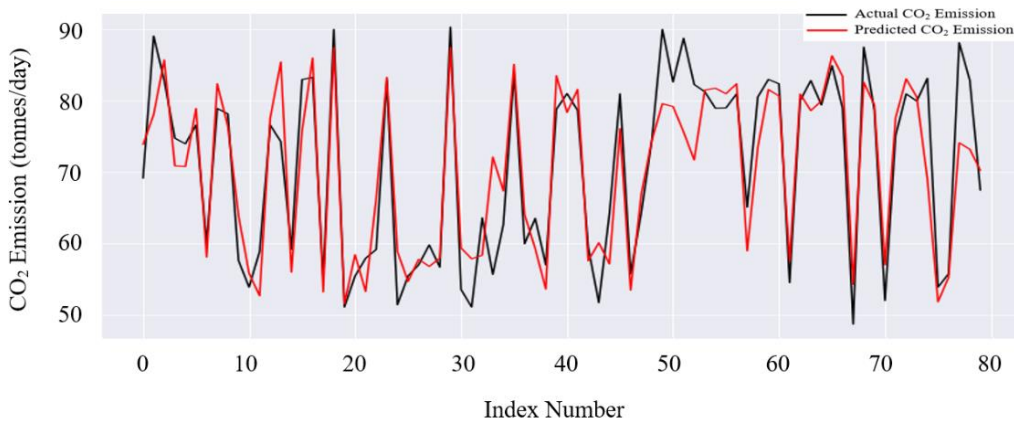


Fig. 7: Actual and forecasted CO₂ emission comparison by Light Gradient Boosted Machine Regression (LGBMR) model

An existing ocean-going bulk carrier of Bangladesh having the dimension of- 182 m (Length) x 32.26 m (Breadth) x 11.92 m (Draft) has been considered to perform the validation check of both models. In this case, implementing the developed KNNR and LGBMR models, the CO₂ emission of this ship has been predicted for five data samples of each input feature. The results of these predictions during the validation have been displayed in Fig. 8 and Fig. 9. These figures demonstrate a comparison between actual CO₂ emission and emission values predicted by KNNR and LGBMR models, respectively.

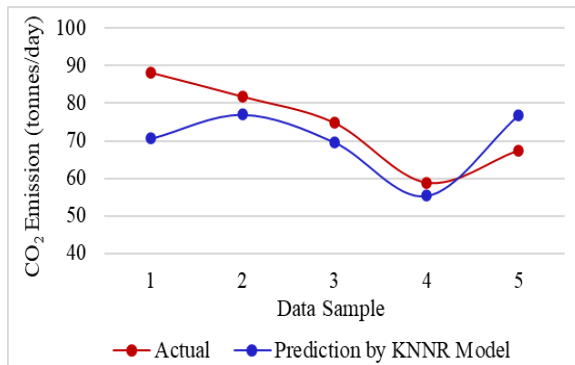


Fig. 8: Validation of K Nearest Neighbor Regression (KNNR) model

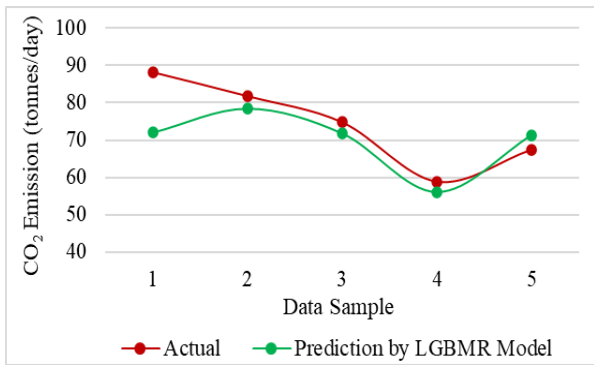


Fig. 9: Validation of Light Gradient Boosted Machine Regression (LGBMR) model

Figs. 8 and 9 state that the KNNR and LGBMR models have performed quite closely. The LGBMR model has worked well in all data samples (1, 2, 3, 4, and 5) when compared to the performance of the KNNR model. However, Figs. 5, 6 and 7 have initially illustrated that the KNNR model has outperformed the LGBMR model. Therefore, the validation outcomes of the KNNR model should have been better than that of the LGBMR model.

The reason for such performance of the KNNR model in the validation process is its training score. The training score of the KNNR and LGBMR models have been found as 0.99 and 0.88, respectively. In machine learning, the training score denotes how the model is generalized or fitted in the training data. If the model makes predictions on the training data very well, overfitting occurs. It negatively affects the model's performance when it faces new, unseen data and results in reducing the model's generalization ability. The very high training score of the KNNR model indicates overfitting here, although the model has provided relatively high R^2 and low error.

In addition, the validation performance of the LGBMR model is comparatively well than the KNNR model. In section 2.3.2, it has been stated that the LGBMR algorithm combines the weak learners consecutively to improve the model where the decision tree acts as a weak learner. This algorithm selects the leaf having the maximum delta loss to grow. That's why it attains a lower penalty for a loss (bad prediction) as the leaf is fixed. LGBMR can also effectively prevent over-fitting by limiting the depth of its tree. Moreover, it divides continuous feature values into discrete bins to speed up the training process. Due to all of these reasons, LGBMR has performed more robustly during validation with regard to moderately well R^2 and small RMSE.

4. Conclusion

This paper presents a comparison-based study of two machine learning models to forecast CO₂ emissions. The developed models can predict the CO₂ emissions from ocean-going ships of Bangladesh using the ships' actual operational data obtained from noon reports. From section 3.2, it can be seen that the predicted result and actual value are in good agreement for both models. Although the model using the KNNR algorithm has outperformed the one with the LGBMR algorithm in terms of R^2 and RMSE, KNNR's performance accuracy during the validation process is lower than that of the LGBMR model due to the overfitting characteristic.

Though in Bangladesh, there exist several machine-learning based researches on different areas, the task of predicting a ship's CO₂ emission implementing machine learning has not been performed here so far. Accordingly, for the first time in Bangladesh, the amount of CO₂ 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

- Álvarez, P. S. (2021): From maritime salvage to IMO 2020 strategy Two actions to protect the environment, *Marine Pollution Bulletin*, Vol. 170, pp. 1-12. <https://doi.org/10.1016/j.marpolbul.2021.112590>
- Ay, C., Seyhan, A., and Beşikçi, E. B. (2022): Quantifying ship-borne emissions in Istanbul Strait with bottom-up and machine-learning approaches, *Ocean Engineering*, Vol. 258, pp. 1-13. <https://doi.org/10.1016/j.oceaneng.2022.111864>
- Denier van der Gon, H., and Hulskotte, J. (2010): Methodologies for estimating shipping emissions in the Netherlands.
- Fletcher, T., Garaniya, V., Chai, S., Abbassi, R., Yu, H., Van, C. T., Brown, R. J., and Khan, F. (2018): An application of machine learning to shipping emission inventory, *International Journal of Maritime Engineering*, Vol. 160 (Part A4), pp. 381-395. <https://doi.org/10.5750/ijme.v160iA4.1073>
- Fourth Greenhouse Gas Study (2020), International Maritime Organization. Available: <https://www.imo.org/en/OurWork/Environment/Pages/Fourth-IMO-Greenhouse-Gas-Study-2020.aspx> [Last accessed 05 June 2023].
- Ganguly, K. K., Nahar, N., and Hossain, B. M. M. (2019): 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. <https://doi.org/10.1016/j.ijdrr.2018.12.002>
- Goldsworthy, L., and Goldsworthy, B. (2015): 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. <https://doi.org/10.1016/j.envsoft.2014.09.009>
- Guidelines for Voluntary Use of the Ship's Energy Efficiency Operational Indicator (EEOI) (2009), International Maritime Organization. Available: <https://gmn.imo.org/wp-content/uploads/2017/05/Circ-684-EEOI-Guidelines.pdf>. [Last accessed 05 June 2023].
- Gusti, A. P., and Semin (2016): 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. <https://doi.org/10.3844/ajeassp.2016.1046.1053>
- Initial IMO GHG Strategy (2018), International Maritime Organization. Available: <https://www.imo.org/en/MediaCentre/HotTopics/Pages/Reducing-greenhouse-gas-emissions-from-ships.aspx> [Last accessed 05 June 2023].
- Lepore, A., Reis, M. S., Palumbo, B., and Capezza, C. (2017): A comparison of advanced regression techniques for predicting ship CO₂ emissions, *Quality and Reliability Engineering International*, Vol. 33, pp. 1281-1292. <https://doi.org/10.1002/qre.2171>
- Li, F., Zhang, L., Chen, B., Gao, D., Cheng, Y., Zhang, X., Yang, Y., Gao, K., Huang, Z., and Peng, J. (2018): A light gradient boosting machine for remaining useful life estimation of aircraft, *Proceedings of 21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, Hawaii, USA, pp. 3562-3567. <https://doi.org/10.1109/ITSC.2018.8569801>
- Shahriar, S. A., Kayes, I., Hasan, K., Salam, M. A., and Chowdhury, S. (2020): Applicability of machine learning in modeling of atmospheric particle pollution in Bangladesh, *Air Quality, Atmosphere & Health*, Vol. 13, pp. 1247-1256. <https://doi.org/10.1007/s11869-020-00878-8>
- Tran, H. (2019): A survey of machine learning and data mining techniques used in multimedia system, *ResearchGate* [Preprint]. Available: <https://10.13140/RG.2.2.20395.49446/1> [Last accessed 05 June 2023].
- Vangheluwe, M., Mees, J., and Janssen, C. (2007): Monitoring programme on air pollution from sea-going vessels (MOPSEA).
- World Investment Report (2021), United Nations Conference on Trade and Development (UNCTAD). Available: <https://unctad.org/webflyer/world-investment-report-2021> [Last accessed 05 June 2023].
- Zhang, Q., Sun, B., Wang, J., Liu, W., and Yu, F. (2019): 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. <https://doi.org/10.15761/FDCCR.1000132>