Predicting Port Congestion at Busan Port Using Machine Learning Algorithms and Temporal Variables

1st Sang-Hyun Ha
2: *AI Grand ICT Research Center*
3: *Dong-Eui University*
4: Busan, Republic of Korea
5: shha@deu.ac.kr

2nd Ki-Hwan Kim
2: *AI Grand ICT Research Center*
3: *Dong-Eui University*
4: Busan, Republic of Korea
5: 15541@deu.ac.kr

3rd Young-Jin Kang
2: *AI Grand ICT Research Center*
3: *Dong-Eui University*
4: Busan, Republic of Korea
5: 15073@deu.ac.kr

4th Ji Yeon Kim
2: *Department of Artificial Intelligence*
3: *Dong-Eui University*
4: Busan, Republic of Korea
5: jnsp0907@naver.com

6th Seok-Chan Jeong
2: D*ept. of e-Business, AI Grand ICT Research Center*
3: *Dong-Eui University*
4: Busan, Republic of Korea
5: scjeong@deu.ac.kr

***Abstract*— Ports are vital infrastructures for global logistics and trade, with South Korea handling approximately 99.7% of its import and export cargo via maritime routes. Busan Port, the largest in the country and the seventh-largest container port worldwide, plays a crucial role in both national logistics and regional economic growth. However, congestion in port operations leads to delays, increased costs, and reduced efficiency. This study employs actual data from Busan Port (February to September 2024) to analyze the impact of temporal variables—such as day of the week, time of day, and monthly factors—on port congestion. Machine learning algorithms, including Random Forest, XGBoost, and LightGBM, were utilized to predict congestion levels using a refined dataset of approximately 1.4 million samples. Results indicate that Extra Trees and CatBoost classifiers achieved high accuracy (0.9654) and AUC (0.9952), while Extreme Gradient Boosting reached an AUC of 0.9989, demonstrating exceptional performance. These findings highlight the effectiveness of machine learning in capturing the nonlinear dynamics of port operations. Future research will aim to enhance model performance by expanding data collection and integrating external factors such as weather conditions and cargo characteristics. This study underscores the potential of machine learning tools in optimizing port management and improving logistical efficiency.**

Keywords—Port Congestion Prediction, Busan Port, XGBoost, LightGBM, Temporal Variables

I. INTRODUCTION

Ports serve as essential infrastructure for logistics and trade, playing a crucial role in the global economic system. In South Korea, approximately 99.7% of domestic import and export cargo is handled via maritime routes, making the shipping port logistics industry highly influential on the national economy [1]. Busan Port is the largest port in the country and ranks as the world's seventh-largest container port, with about 30% of shipping port workers concentrated in the region. The efficient operation of Busan Port is directly linked not only to the smooth functioning of the national logistics network but also to the revitalization of the regional economy.

However, congestion arising during port operations negatively impacts the entire logistics system. Congestion leads to issues such as logistics delays, increased operational costs, and longer waiting times, which undermine the efficiency and economic effectiveness of port operations. To address these issues, a systematic approach utilizing data-driven predictive tools is necessary. Existing studies primarily analyze congestion levels based on current data or focus on short-term predictions, with relatively limited development of medium- to long-term prediction models.

This study analyzes the impact of temporal variables (day of the week, time of day, and monthly factors) on port congestion at Busan Port using actual data from February to September 2024. Furthermore, congestion was predicted using machine learning algorithms (Random Forest [2], XGBoost [3], LightGBM [4]). Specifically, approximately 2.18 million raw data points were refined and about 1.4 million data points were utilized for training. While handling real data, the diversity and size of the dataset were somewhat limited for applying AI models; nevertheless, the study confirmed the potential of AI-based predictions.

II. RELATED WORKS

Port congestion prediction is a significant research topic for optimizing port operations and logistics networks. Related studies are primarily categorized into three approaches: utilization of temporal variables, integration of external factors, and data-driven machine learning modeling.

First, temporal variables are major determinants of port congestion, and research leveraging these variables has been actively conducted. W. S. Kang et al. analyzed temporal data using Random Forest and XGBoost models, demonstrating high predictive performance [5]. However, they highlighted limitations in achieving more precise predictions due to insufficient data segmentation and the lack of multivariate combination analysis.

Second, external factors such as weather conditions and cargo characteristics are other important elements influencing port congestion [6-7].

L. Potgieter et al., study analyzes data from 2011 to 2018 to assess weather- and system-related port congestion risks at the Cape Town Container Terminal[6]. Using a mixed-methods approach, it incorporates qualitative insights from nine interviews and quantitative time-series data to evaluate congestion frequency and its time impact. Findings classify both weather- and system-related congestion as major risks,

urging improved mitigation strategies for operational sustainability.

L. Vukic and K. H. Lai focused on analyzing the risk of weather- and system-related port congestion at the Cape Town Container Terminal (CTCT) between 2011 and 2018 [7]. The study employed a mixed-method research approach comprising qualitative data (interviews and emails) and quantitative data (time series data analyzed using Excel), leading to the development of risk profiles and heatmaps. Key attributes analyzed included weather data (strong winds, fog, swells, etc.), system data (TOS errors, power outages, maintenance, etc.), congestion frequency, and the time impact. The study primarily relied on traditional statistical methods and visualization techniques for data processing, without utilizing AI-based techniques. In conclusion, this study systematically analyzed the risks at CTCT using quantitative and qualitative approaches and delivered meaningful results within the scope of traditional methodologies.

Weather conditions can cause operational delays or alter the speed of operations, while the type and quantity of cargo significantly impact the efficiency of terminal operations. Some studies have shown that integrating these external factors into data enhances the performance of predictive models. Nonetheless, the processes of collecting and integrating external data remain challenging.

Third, data-driven machine learning models have established themselves as essential tools in port congestion prediction. T. N. Cuong et al. analyzed the cargo throughput data of Busan Port over a period of approximately 20 years, from January 2001 to July 2021[8]. The data used in this study were collected from the Korean PORT-MIS and consisted of monthly data with a total of 247 data points. It included metrics such as port entry, departure, and transshipment volumes, recorded in million tons. The data exhibited nonlinear and complex dynamic characteristics influenced by seasonal variations and long-term trends, which were effectively reflected in the analysis and forecasting. W. Peng et al. proposed a high-frequency container port congestion measurement method based on AIS (Automatic Identification System) data[9]. They analyzed the movement information of 3,957 container ships from March to April 2017. The berth and anchorage areas. S. E. Mekkaoui et al., in their systematic literature review on the application of machine learning techniques in port operations, highlighted that while Artificial Neural Networks (ANN) are predominantly utilized, research in the field of Reinforcement Learning (RL) remains relatively scarce, emphasizing the need for further studies in this area[10].

This study combined machine learning models with temporal domains to predict port congestion and explored the potential integration of external factors by referencing previous studies.

III. DATASETS

A. Datasets

In this study, an analysis was conducted to predict port congestion based on actual data from Busan Port. The collected dataset comprises approximately 966,276 records, as shown in Table 1, with a total of 32 attributes. These attributes include extensive information related to ship handling operations and terminal activities.

TABLE I. DATASETS INFORMATIONS

Column	Description
TERMINAL_CODE	Terminal Code
TERMINAL_SHIP_YEAR	Terminal Year
TERMINAL_SHIP_VOYAGE_NO	Terminal Voyage
TERMINAL_SHIP_NAME	Terminal Ship Name
SHIPPING_CODE	Shipping Company Code
SHIPPING_VOYAGE_NO	Shipping Voyage Number
SHIPPING_ROUTE_CODE	Shipping Route Code
BERTH_CODE	Berth Code
ETB	Estimated Berthing Time
CCT	Cargo Cut-off Time
ETD	Estimated Departure Time
ALONGSIDE	Berthing Direction
DISCHARGING_COUNT	Discharging Quantity
LOAD_COUNT	Loading Quantity
SHIFT_COUNT	Transshipment Quantity
VVD_YEAR	Shipping Company Year
REG_DT	Registration Date and Time
ATB	Actual Berthing Time
ATD	Actual Departure Time
COMMENCE_TIME	Operation Time
DISCHARGE_COMPLETED	Discharging Completed
DISCHARGE_REMAIN	Remaining Discharging
DISCHARGE_TOTAL	Total Discharging
LOADING_COMPLETED	Loading Completed
LOADING_REMAIN	Remaining Loading
LOADING_TOTAL	Total Loading
CALL_INDEX	Call Index
TERMINAL_CODE	Terminal Code
VESSEL_STATUS	Vessel Status
INOUT_STATUS	Cargo Movement Status
VESSEL_STATUS_STR	Vessel Congestion Level
INOUT_STATUS_STR	Cargo Movement Congestion

Table 2 categorizes port area congestion levels, detailing the primary characteristics associated with each grade. In the Normal category, traffic for both vehicles and ships flows smoothly, and no additional measures are required. The Warning level is characterized by a slight increase in vehicle waiting times; however, congestion remains not severe, necessitating ongoing situation monitoring. At the Primary level, vehicle waiting times are significantly extended, requiring adjustments to work schedules and the implementation of preventive measures. The Danger category indicates intensified vehicle congestion that adversely affects port operations, thereby necessitating urgent responses and enhanced traffic management. Finally, the Critical level denotes severe waiting times for both vehicles and ships, demanding immediate action, the allocation of additional resources, and the deployment of robust countermeasures.

TABLE II. DATASETS INFORMATIONS

Order	Level	Description
1	Normal	Low gate congestion and smooth conditions
2	Warning	Attention needed due to possible increase in gate congestion
3	Primary	Mid-level gate congestion with high chance of worsening, preventive measures needed
4	Danger	High gate congestion requiring immediate action
5	Critical	Very severe gate congestion threatening to paralyze port operations

B. Datasets Anylisis

The Pearson Correlation Coefficient (PCC) is a statistical measure that quantitatively represents the degree of linear correlation between two variables. This coefficient is exclusively applicable to numerical data and is calculated

based on the Cauchy-Schwarz inequality. The value of the Pearson Correlation Coefficient ranges between -1 and +1, indicating both the strength and direction of the linear relationship between the variables. A coefficient close to +1 signifies a perfect positive linear correlation, meaning that as one variable increases, the other variable also increases proportionally. Conversely, a coefficient near -1 indicates a perfect negative linear correlation, where an increase in one variable corresponds to a proportional decrease in the other. When the coefficient approaches 0, it suggests the absence of a linear correlation between the two variables. Generally, values between -1 and -0.3 or between 0.3 and 1 are considered to represent strong correlations, implying a significant linear relationship between the variables. Due to these characteristics, the Pearson Correlation Coefficient is widely utilized in research to analyze and interpret the relationships between variables.

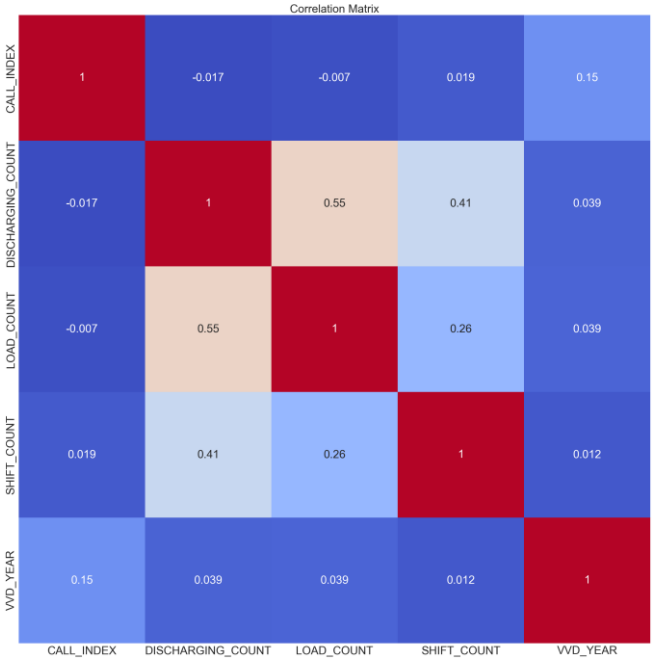

Fig. 1. Results of Correlation Analysis of Berth Planning History

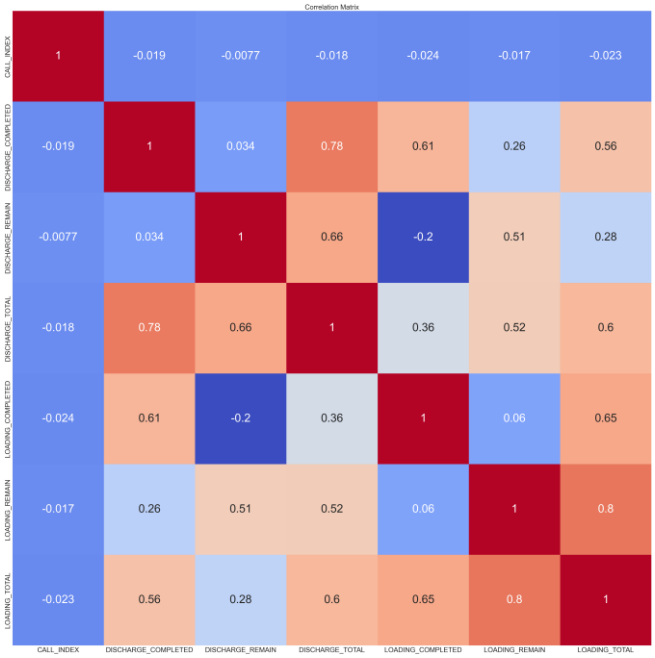

Fig. 2. Results of Correlation Analysis of Current Operations and Historical Data

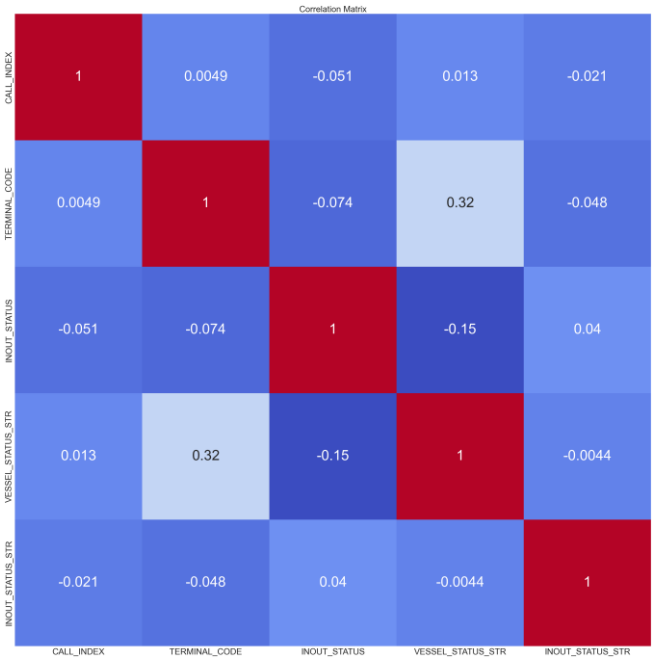

Fig. 3. Results of Correlation Analysis of Terminal Congestion History

C. *Validation Methods*

In this study, the classification performance metrics used as evaluation indicators include Accuracy, defined by Equation (1). Accuracy measures the proportion of correctly classified samples out of the total number of samples. Area Under the Curve (AUC) evaluates the model's classification performance by calculating the area under the Receiver Operating Characteristic (ROC) curve. Additionally, Recall, as defined by Equation (2), represents the proportion of correctly predicted positive samples out of all actual positive samples. Precision, defined by Equation (3), indicates the proportion of actual positive samples among those predicted as positive by the model. Furthermore, the F1 Score, according to Equation (4), is the harmonic mean of Recall and Precision, which is particularly important in the context of imbalanced datasets. Kappa, defined by Equation (5), measures the agreement between the classification results and random chance, while the Matthews Correlation Coefficient (MCC), as described by Equation (6), assesses classification performance by taking into account both positive and negative classes. These metrics provide a comprehensive set of criteria for evaluating the classification performance of the model from multiple perspectives.

$$Accuracy = \frac{(Number\ of\ correct\ predictions)}{(Total\ number\ of\ predictions)} \quad (1)$$

$$Recall = \frac{(True\ Positives)}{(True\ Positives + False\ Negatives)} \quad (2)$$

$$Precision = \frac{(True\ Positives)}{(True\ Positives + False\ Positives)} \quad (3)$$

$$F1\ Score = 2 \times \frac{(Precision * Recall)}{(Precision + Recall)} \quad (4)$$

$$Kappa = \frac{(Observed - Expected)}{(1 - Expected)} \quad (5)$$

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP+FP) * (TP+FN) * (TN+FP) * (TN+FN)}} \quad (6)$$

IV. EXPERIEMENTS RESULTS

Various classification models were employed as outlined in Table 3. Decision Tree-Based Models partition the data through a hierarchical structure, with splits at each node determined by the characteristics of the data. Ensemble Models aim to overcome the limitations of individual models by combining multiple models to reduce errors and enhance performance. Linear Models are favored for their high computational efficiency and relatively straightforward interpretability of results. Geometric Models focus on identifying the optimal boundaries that separate the data, thereby improving classification accuracy. Probabilistic Models assume independence among features and utilize conditional probabilities to perform classification. By leveraging this diverse set of models, the study aims to comprehensively evaluate and enhance the predictive performance for port congestion levels.

TABLE III. CLASSIFICATION MODELS

Model	Category
Decision Tree Classifier (dt)	Decision Tree Based
Extra Trees Classifier (et)	Decision Tree Based
Random Forest Classifier (rf)	Ensemble Models
Gradient Boosting Classifier (gbc)	Ensemble Models
Extreme Gradient Boosting (xgboost)	Ensemble Models
CatBoost Classifier (catboost)	Ensemble Models
Ada Boost Classifier (ada)	Ensemble Models
Logistic Regression (lr)	Linear Models
Ridge Classifier	Linear Models
SVM - Linear Kernel (svm)	Geometric Models
Naive Bayes (nb)	Probabilistic Models
K Neighbors Classifier (knn)	Other Models
Dummy Classifier	Other Models
Linear Discriminant Analysis (lda)	Other Models
Quadratic Discriminant Analysis (qda)	Other Models

Decision tree classifiers utilize a tree structure to categorize data or predict values by applying decision rules based on features at each node. While this model is intuitive and easy to visualize, it is prone to overfitting. To address the issue of overfitting, Extra Trees introduce randomness to create trees more quickly and simply, whereas Random Forest employs an ensemble technique by combining multiple trees and averaging their predictions or determining the outcome through majority voting. Gradient Boosting applies a boosting technique where each tree sequentially corrects the errors of the previous ones, and Extreme Gradient Boosting (XGBoost) emphasizes efficiency and scalability for large and complex datasets, incorporating regularization and tuning options to prevent overfitting. CatBoost automatically handles categorical variables and offers rapid training speeds along with high accuracy. AdaBoost iteratively enhances weak learners by assigning greater weights to misclassified instances, making it applicable to various classification problems.

Logistic Regression and Ridge Classifier are linear classification models known for their simplicity and ability to reduce overfitting. Support Vector Machines (SVM) identify optimal decision boundaries, making them effective even in high-dimensional data spaces. Naive Bayes is a straightforward probabilistic classifier that assumes independence among features. K Nearest Neighbors (KNN) classifies data based on the proximity of neighboring instances, while the Dummy Classifier serves as a basic benchmark for comparison. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) utilize linear and quadratic decision boundaries, respectively, to maximize inter-class variance and minimize intra-class variance. By incorporating this diverse array of classification models, the study aims to thoroughly evaluate and enhance the predictive performance for port congestion levels.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
dt	Decision Tree Classifier	0.9904	0.9926	0.9904	0.9904	0.9904	0.9855	0.9855
et	Extra Trees Classifier	0.9654	0.9952	0.9654	0.9653	0.9652	0.9228	0.9231
rf	Random Forest Classifier	0.9640	0.9950	0.9640	0.9641	0.9635	0.9187	0.9197
catboost	CatBoost Classifier	0.9639	0.9952	0.9639	0.9638	0.9635	0.9189	0.9196
xgboost	Extreme Gradient Boosting	0.9445	0.9898	0.9445	0.9455	0.9433	0.8727	0.8761
gbc	Gradient Boosting Classifier	0.9008	0.0000	0.9008	0.9015	0.8976	0.7674	0.7738
lda	Linear Discriminant Analysis	0.7616	0.0000	0.7616	0.7423	0.7453	0.4254	0.4351
ridge	Ridge Classifier	0.7531	0.0000	0.7531	0.7317	0.7149	0.3601	0.3800
lr	Logistic Regression	0.6943	0.0000	0.6943	0.4820	0.5690	0.0000	0.0000
nb	Naive Bayes	0.6943	0.5514	0.6943	0.4820	0.5690	0.0000	0.0000
svm	SVM - Linear Kernel	0.6943	0.0000	0.6943	0.4820	0.5690	0.0000	0.0000
dummy	Dummy Classifier	0.6943	0.5000	0.6943	0.4820	0.5690	0.0000	0.0000
knn	K Neighbors Classifier	0.6257	0.5171	0.6257	0.4340	0.5124	-0.0001	-0.0003
ada	Ada Boost Classifier	0.2051	0.0000	0.2051	0.3512	0.1292	0.0053	0.0013
qda	Quadratic Discriminant Analysis	0.0502	0.0000	0.0502	0.7034	0.0524	0.0085	0.0166

Fig. 4. Results of Correlation Analysis of Current Operations and Historical Data

The Extra Trees Classifier (ET) and CatBoost Classifier demonstrate the highest performance among all models, exhibiting superior Accuracy, AUC, Recall, Precision, and F1 Scores. These models exhibit balanced performance and consistency across various evaluation metrics. Specifically, the Extra Trees model achieves an exceptionally high Accuracy of 0.9654 and an AUC of 0.9952. Additionally, XGBoost showcases outstanding performance with an AUC of 0.9989 and also delivers strong results across other metrics. The robust performance of these classifiers underscores their effectiveness in predicting port congestion levels, highlighting their suitability for applications requiring high precision and reliability.

V. CONCLUTIONS

This study successfully predicted port congestion at Busan Port by integrating temporal variables with advanced machine learning algorithms. Utilizing actual data from February to September 2024, the research analyzed the influence of temporal factors—such as day of the week, time of day, and monthly variations—on congestion levels. The application of machine learning models, including Random Forest, XGBoost, and LightGBM, demonstrated the effectiveness of these techniques in forecasting port congestion accurately. Notably, the Extra Trees and CatBoost classifiers exhibited exceptionally high accuracy (0.9654) and AUC (0.9952), while Extreme Gradient Boosting achieved an outstanding AUC of 0.9989, underscoring the robustness of these models.

The findings indicate that machine learning algorithms are adept at capturing the nonlinear dynamics inherent in port operations, thereby providing reliable predictions even in the presence of data limitations and the exclusion of external factors. This highlights the potential of AI-based tools in enhancing the efficiency of port management and optimizing the national logistics network. The high performance of the Extra Trees and CatBoost models suggests that ensemble methods and those capable of handling categorical variables effectively are particularly well-suited for this application.

Despite the promising results, the study acknowledges certain limitations, primarily related to the diversity and size of the dataset. While approximately 9.6 million refined data points were utilized for training, expanding the dataset could further improve model accuracy and generalizability. Additionally, the integration of external factors such as weather conditions and cargo characteristics remain a challenge but is essential for achieving more comprehensive and precise predictions.

This study empirically demonstrates that machine learning models based on temporal patterns can be effectively utilized for predicting port congestion. Specifically, models based on XGBoost and LightGBM can serve as practical tools to enhance the efficiency of port operations by reducing waiting times and lowering operational costs. To address existing limitations, future research should focus on expanding the scope of data collection and incorporating external factors such as weather conditions and cargo characteristics to improve prediction accuracy.

In conclusion, this research highlights the potential of machine learning algorithms to enhance port operation efficiency and underscores the importance of developing data-driven decision-making tools. It is anticipated that the practical application of these findings will make substantial contributions to the optimization of port management practices.

ACKNOWLEDGMENT

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2025-RS-2020-II201791)

REFERENCES

[1] Kim, Doo-hwan, and Kangbae Lee. "Forecasting the Container volumes of Busan port USING LSTM." Journal of Korea Port Economic Association, vol. 36.2, 2020, pp. 53-62.

[2] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785-794.

[3] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y., "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems, vol. 30, 2017.

[4] Breiman, Leo, "Random forests," Machine learning 45, 2001, pp. 5-32.

[5] Kang, W. S., Song, T. H., Kim, Y. D., & Park, Y. S., "A study on seasonal variation in marine traffic congestion on major port and coastal routes," Journal of the Korean Society of Marine Environment & Safety, vol. *23*(1), 2017, pp. 1-8.

[6] Potgieter, Lilian, Leila L. Goedhals-Gerber, and Jan Havenga. "Risk profile of weather and system-related port congestion for the Cape Town container terminal," 2020. [Online]. Available: https://doi.org/10.25159/1998-8125/6149

[7] Vukić, Luka, and Kee-hung Lai, "Acute port congestion and emissions exceedances as an impact of COVID-19 outcome: the case of San Pedro Bay ports," Journal of Shipping and Trade 7.1, 2022.

[8] Cuong, Truong Ngoc, Hwan-Seong Kim, and Sam-Sang You. "Data analytics and throughput forecasting in port management systems against disruptions: a case study of Busan Port," Maritime Economics & Logistics 25, 2023, pp.61-89.

[9] Peng, W., Bai, X., Yang, D., Yuen, K. F., & Wu, J. "A deep learning approach for port congestion estimation and prediction," Maritime Policy & Management, vol. *50*(7), 2023, pp.835-860.

[10] Mekkaoui, Sara E., Loubna Benabbou, and Abdelaziz Berrado, "A systematic literature review of machine learning applications for port's operations," 2020 5th International conference on logistics operations management (GOL). IEEE, 2020.