Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 269 (2025) 1389–1397

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Conference on Computer Science and Computational Intelligence 2025

Optimization subproblem importance analysis based on machine learning prediction in a three stage-export container scheduling

Aditya Saputra[a,b] , Ivan Kristianto Singgih[b,c,d,e,*]

[a]PT Insera Sena, Sidoarjo, Indonesia
[b]Study Program of Industrial Engineering, University of Surabaya, Surabaya, Indonesia
[c]The Indonesian Researcher Association in South Korea (APIK), Seoul, 07342, South Korea
[d]Kolaborasi Riset dan Inovasi Industri Kecerdasan Artifisial (KORIKA), Jakarta, Indonesia
[e]Indonesia Artificial Intelligence Society, Jakarta 12930, Indonesia

Abstract

A traditional way to optimize a complex optimization problem is to divide the problem into several subproblems and solve each subproblem separately or another problem that consists of some subproblems. Despite many attempts to define and solve various optimization problems in the container terminal logistics field, how to measure the importance of each subproblem is often ignored in many studies and remains a difficult issue. Most studies directly propose methods to solve a specific subproblem after stating the importance of the specific subproblem, with or without simply considering the effect of other subproblems as input. The advancement of machine learning techniques allows a new paradigm for understanding the importance of such optimization subproblems. In this study, a scheduling problem for export containers in a terminal is considered. The case considers scheduling subproblems on subsequent processing stages on yard cranes, internal trucks, and quay cranes. With the input of each stage's processing time information (mean and standard deviation values) and the selected scheduling rule for each stage, the makespan of all containers' processing is predicted. The numerical experiments show that the scheduling rules for quay cranes and internal trucks have the most significant impact on system performance. These finding challenges conventional approaches by revealing that not all subproblems contribute equally to system optimization. The proposed machine learning framework enables terminal operators to adapt their optimization focus to address high-impact areas, reducing computational complexity while providing a data-driven methodology for understanding interdependencies between operational components.

Keywords: regression machine learning; rule; container terminal; scheduling

* Corresponding author. Tel.: +62-31-298-1392.
E-mail address: ivanksinggih@staff.ubaya.ac.id

10.1016/j.procs.2025.09.080

1. Introduction

Recent review on container terminal logistics focused solely on defining various optimization problems and solving them. Even though various integrated problems, starting from the gate operation management up to the seaside operation, have been studied extensively [1,2], none of them could identify which subproblem is more important than the others and how much the effect would be when we optimize a certain subproblem. All reviews would list types of integrated problems simultaneously, and researchers would refer to the type of integrated problem they prefer to focus on, instead of dealing with important problems after properly observing the effect of optimizing each subproblem. The failure to identify important problems causes difficulties in optimizing the whole system. In other words, there is no way to understand whether the important integrated problems have been successfully defined or not. Such a condition is troublesome, considering that improperly defined integrated problems would cause the generation of only a local optimal solution for the whole problem, considering that several subproblems would be solved independently and have contradictory decisions between the defined problems. In addition to that, solving integrated problems is often conducted under the assumption of subsequent processes [3], which limits the ability to optimize the whole system properly.

Machine learning opens opportunities to understand various behaviors of many systems. Studies conduct predictions that include big data are conducted widely in various fields, e.g., production scheduling [4], vehicle routing problem [5,6], hub location problem [7], ship berthing problem [8], container storage [9], etc. Despite the fast growth in the number of studies, studies that applied operations research and machine learning techniques only dealt with predefined (partial) optimization problems. Such a situation limits the whole system in obtaining global optimal solutions and can be resolved using machine learning techniques for understanding the whole system. The benefits of machine learning for understanding complete optimization problem characteristics are illustrated in Figure 1. As shown in Figure 1, combining traditional operations research methods with artificial intelligence opens new possibilities in solving optimization problems. In the simplest way, a specific subproblem (partial problem) can be solved using the combination of operations research and artificial intelligence. The subproblems include single machine scheduling, routing, or allocation problems. The new framework discussed in this study is the integrated problem understanding and solving. Utilizing machine learning techniques to observe the behavior of the whole system allows understanding relationships between subproblems and identifying subproblems that have a greater effect on the whole system's performance. Given that any system's behavior would be highly influenced by its data characteristics, the importance of subproblems could continuously change at different times depending on the situation of the system. Different from the first approach, the performance of the whole system is properly assessed and optimized, instead of only optimizing a certain subproblem without any observation of its influence to the whole system's performance.

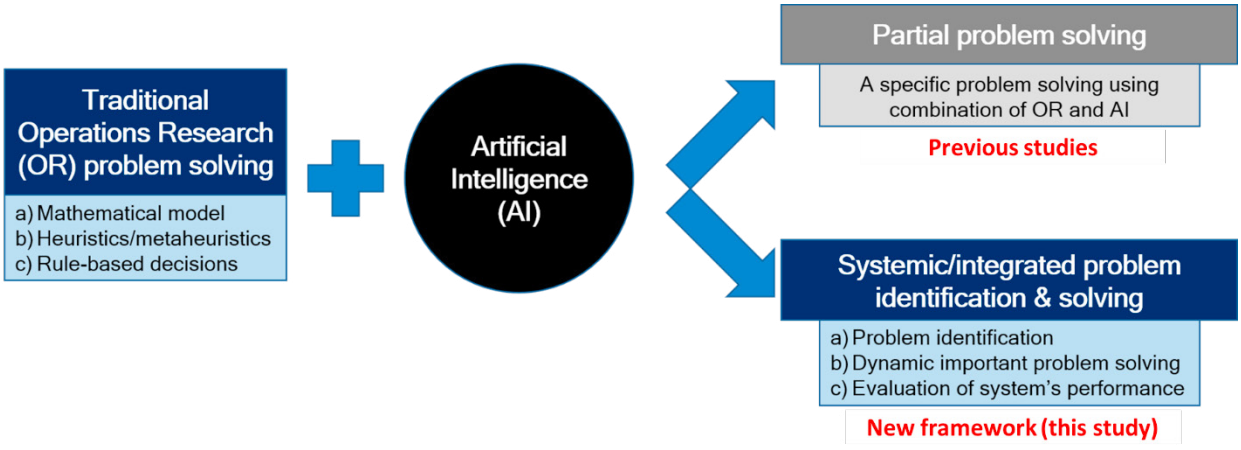

Fig. 1. Benefits of implementing machine learning to understand optimization problems.

Details on the complete framework for utilizing machine learning are shown in Figure 2. Machine learning models are trained to understand good or bad behaviors of the system and are responsible as an alarm system to inform decision makers when any potential performance reduction in the near future is detected (1), as conducted in [4]. At this stage, all existing subproblems are listed and clearly defined with all possible solution methods for solving each subproblem (2a). The whole system operates simultaneously with a hypothesis on the importance of each subproblem (2b). The results of each subproblem solving (2c) become input data for machine learning models that evaluate how strong the influences of each subproblem are on the whole system's performance (2d). The results are then used to identify most important subproblems that could be used to focus the next optimization phase on such set of subproblems. This strategy would not only reduce the computational time, but also ensure effective subproblem solving that significantly improves the performance of the whole system.

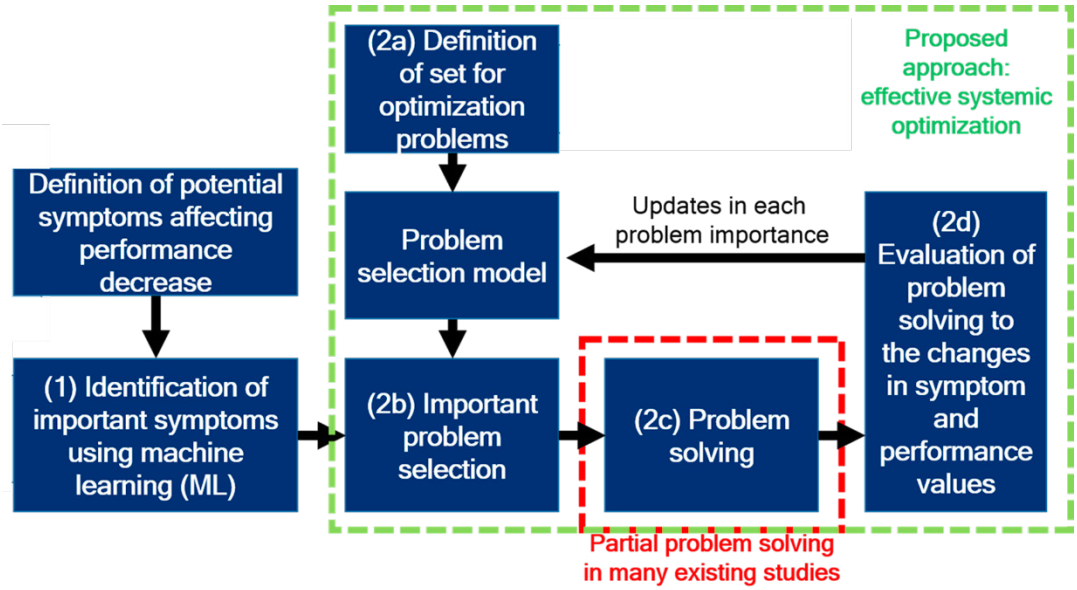

Fig. 2. Framework for continuously evaluating the importance of subproblems solved by using various solution methods.

The rest of the study is as follows. Section 2 explains the problem discussed in this study, which is the three-stage export container scheduling. Section 3 presents the proposed machine learning model. Section 4 shows the numerical experiments. Section 5 concludes the study and identifies potential topics for future studies.

2. Problem Definition

The three-stage export container scheduling is a scheduling problem that includes the container transportation using the yard cranes, internal trucks, and quay cranes. The system can be identified as a flow shop. The movement is illustrated in Figure 3 [10]. The optimization problem is defined as follows:

- Input (parameters):
  1. A set of export containers to transport
  2. Processing time required by the yard crane to process each container
  3. Positioning time required by the yard crane to move to the starting point of each container after completing the processing of any previous container
  4. Processing time required by the internal truck to process each container
  5. Processing time required by the quay crane to process each container

- Output (decision variables):
  1. Scheduling rule applied on the yard crane (first-come-first-served or shortest processing time)
  2. Scheduling rule applied on the internal truck (first-come-first-served or shortest processing time)
  3. Scheduling rule applied on the quay crane (first-come-first-served or shortest processing time)
- Constraints:
  1. The yard crane can only process at most one container at the same time.
  2. The internal truck can only process at most one container at the same time.
  3. The quay crane can only process at most one container at the same time.
  4. The export container processing by the internal truck can only be performed after the container processing by the yard crane is completed.
  5. The export container processing by the quay crane can only be performed after the container processing by the internal truck is completed.
  6. The container must be processed as long as the required processing time by each of the yard crane, the internal truck, and the quay crane.
- Objective:
  Minimizing the makespan (latest completion time for all export containers)

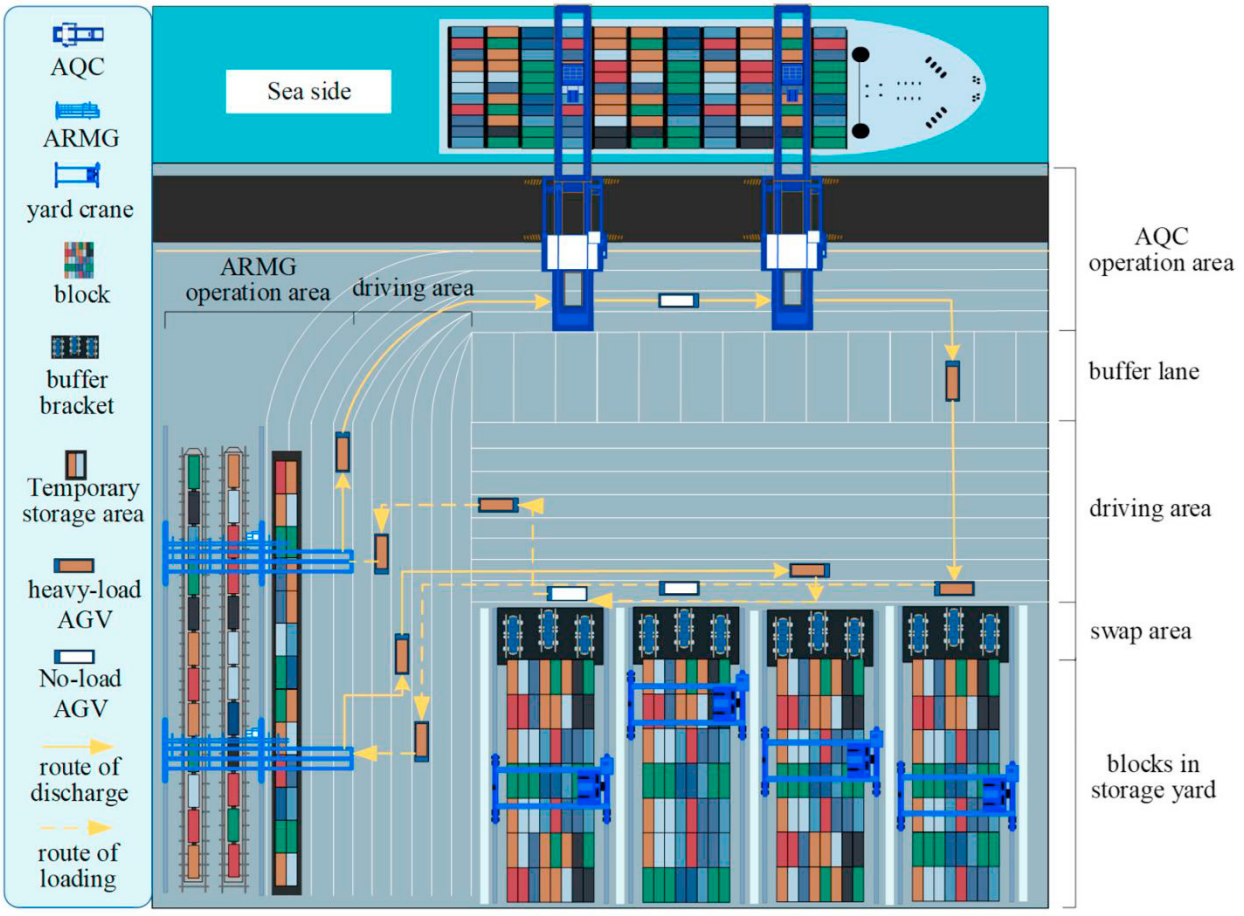

Fig. 3. Export container movement in a container terminal (from yard cranes at the bottom to the quay cranes at the top of the figure).

3. Proposed Regression Machine Learning Model

Data considered for experiments in this study are listed in Table 1. The first part of the input data represents the problem itself, which consists of the mean and standard deviation values for processing times of the YC, the truck, the QC, and the ratio of the mean values (presented for each pair of equipment). The second part of the input data represents the optimization rule applied to each equipment (YC, truck, and QC). The observed rules are First-Come-First-Served (FCFS) and Shortest Processing Time (SPT). Considering the problem-related features and the optimization decision features simultaneously is important to assess the effect of the optimization decisions on specific problem characteristics. The output for the prediction is the makespan, which is the latest completion time at the QC for any export container. Considering the numerical data type for the predicted value, regression machine learning models are used [5], instead of classification models that are required when class-type targets are predicted [4] or clustering techniques that are required when no target data exist. When discussing any optimization problem, dealing with numerical values allows observing more details on the system's behavior when compared with the class-type target values, even though slightly more computational time might be required when dealing with numerical target data.

Table 1. Data for this study.

Data name	Type	Description
YC_mean_time	Input	Mean value of YC processing time for a single container
YC_stdev_time	Input	Standard deviation value of YC processing time for a single container
truck_mean_time	Input	Mean value of truck processing time for a single container
truck_stdev_time	Input	Standard deviation value of truck processing time for a single container
QC_mean_time	Input	Mean value of QC processing time for a single container
QC_stdev_time	Input	Standard deviation value of QC processing time for a single container
YC_per_truck_ratio	Input	Comparison between the YC and the truck mean processing times (YC/truck)
YC_per_QC_ratio	Input	Comparison between the YC and the QC mean processing times (YC/truck)
truck_per_QC_ratio	Input	Comparison between the truck and the QC mean processing times (YC/truck)
YC_rule	Input	YC scheduling rule (0: First-Come-First-Served/FCFS, 1: Shortest Processing Time/SPT; one-hot encoding)
truck_rule	Input	Truck scheduling rule (0: FCFS, 1: SPT; one-hot encoding)
QC_rule	Input	QC scheduling rule (0: FCFS, 1: Shortest Processing SPT; one-hot encoding)
makespan	Output	The latest completion time at the QC of any export container

The proposed regression machine learning models are models that have been proven to outperform other methods for prediction purposes or models that have been widely considered by many studies: (1) linear regression (LR) [11], (2) support vector regression (SVR) [12,13], (3) random forest (RF) [11,14], (4) K-nearest neighbors (KNN) [15,16], (5) decision tree (DT) [17], and (6) gradient boosting (GB) [12].

4. Numerical Experiments

An Excel-based simulation is used to produce 500 instances that include the input and output data. In each instance, random scheduling rules (among the FCFS and SPT) are selected for each of the YC, truck, and QC. Range of the values for each feature is presented in Table 2.

Table 2. Range of each feature in the generated data.

Data name	Range of values [min, max]
YC_mean_time	[2.8, 16]
YC_stdev_time	[1.48, 7.7)
truck_mean_time	[3.4, 14.5]
truck_stdev_time	[1.89, 7.73]
QC_mean_time	[2.9, 15]
QC_stdev_time	[1.34, 7.3]
YC_per_truck_ratio	[0.28, 3.4]
YC_per_QC_ratio	[0.33, 3.63]
truck_per_QC_ratio	[0.41, 2.4]
YC_rule	0 or 1
truck_rule	0 or 1
QC_rule	0 or 1

Correlations between features are presented in Figure 4. It is shown that the input features are not highly correlated with each other, except the ratio values, which are derived from other input features. Even though the correlations between the input features are low, some significant correlations could be observed between the input and output features, which would be observed further below.

The experiments are conducted on Google Colab and based on sklearn, pandas, numpy, and seaborn Python libraries. The following experiment setting is used: (1) train-test split ratio of 80%:20%, and (2) StandardScaler normalization. For all of the models, whenever possible, the same random seed (that equals to 42) is used. Basic hyperparameter settings are used for each method. The results are shown in Table 3. Models with R^2 values equal to more than 70% (LR, RF, and GB) are considered suitable for predicting the behavior of the system well. Among those models, LR has the least Root Mean Squared Error (RMSE) values and MSE values, while GB has the least Mean Absolute Error (MAE) values.

The main output of this study is the rank of the feature importances. Such a rank would help decision makers to understand which aspect of the optimization problem they should be focusing on. Table 4 presents the feature importance results from the linear regression model for predicting makespan in the observed three-stage export container scheduling problem. The non-high correlation coefficients (<0.8) between features indicates the absence of significant multicollinearity among the variables. This ensures that all input features can be utilized without introducing information redundancy, which is essential for maintaining the stability and accuracy of the regression model. Therefore, the correlation matrix supports both the feature selection process and the validity of using a regression-based predictive approach.

The analysis reveals that the scheduling rules have the most significant impact on system performance: QC_rule and truck_rule. The next important features are the mean processing times of the truck and the QC, which are highly related to the first two important features. When the features on the scheduling rules are compared with each other, it is concluded that optimizing QC and truck operations is much more important than optimizing the YC operation. In other words, having a complex optimization effort for the YC operation would not minimize the makespan significantly without dealing with the QC and truck optimization issues. Such an insight could also suggest setting weights on the objective factors related to the QC and truck higher than the YC's when an integrated problem is solved.

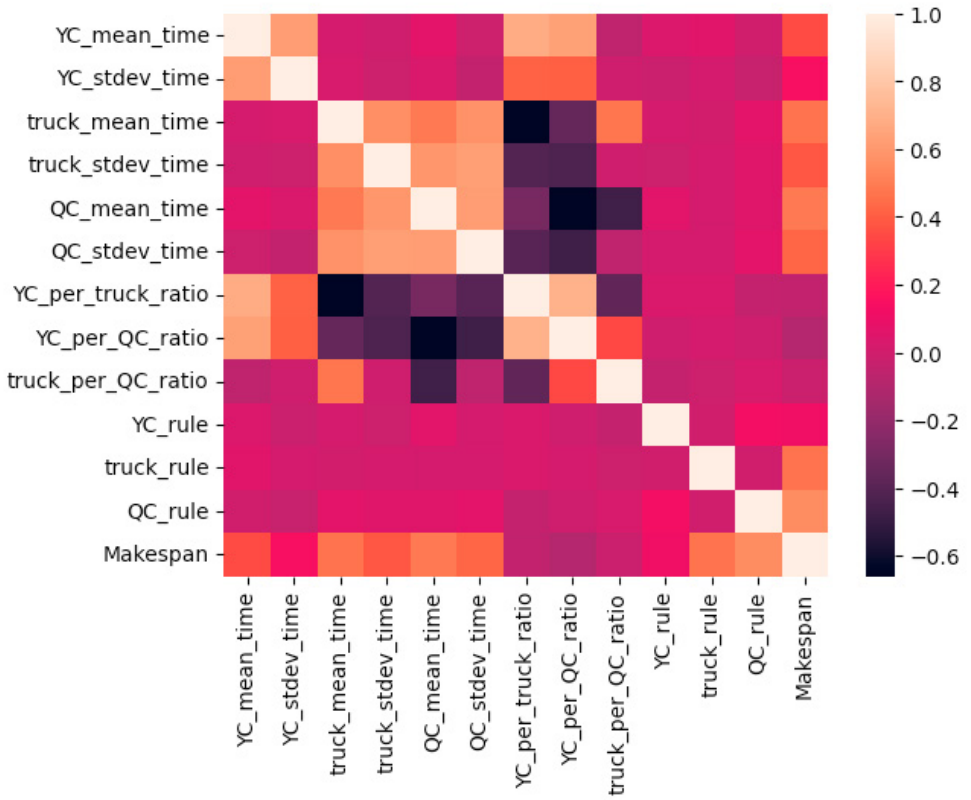

Fig. 4. Correlations between features

These findings suggest that terminal operators should focus on highly impactful parts of the system rather than attempting to simultaneously optimize all operational aspects. Having this information, decision makers would optimize their logistics system more effectively and efficiently. After identifying the most significant feature, e.g., the mathematical models n QC optimization rule, the decision makers need to test various optimization methods to solve the QC scheduling problem and observe the quality of the produced objective values for the whole system when implementing each method. The solution methods those must to be tested, include exact methods [8], mathematical model solvers [3], metaheuristics/algorithms [5], and simulation [3].

Table 3. Experiment result.

Regression models	R^2	Root Mean Squared Error (RMSE)	Mean Squared Error (MSE)	Mean Absolute Error (MAE)
LR	**85.2%**	**15.9***	**253.4***	13.1
SVR	27.8%	35.1	1233.4	27.4
RF	**79.4%**	18.8	352.0	15.0
KNN	44.1%	30.9	956.6	24.6
DT	50.5%	29.1	846.0	21.4
GB	**84.6%**	16.2	263.7	**12.9***

As shown in Table 3, the regression models exhibit varying levels of predictive accuracy. Among them, Linear Regression, Gradient Boosting, and Random Forest demonstrate notably stronger performance, particularly with R-

squared values exceeding the 70% threshold. This suggests that these models are better suited to capture the underlying relationships within the dataset, offering a more reliable basis for decision-making in predictive tasks.

Table 4. Feature importances.

Rank	Feature	Importance (coefficients of features)
1	QC_rule	**25.12**
2	truck_rule	22.00
3	truck_mean_time	18.61
4	QC_mean_time	17.20
5	YC_per_truck_ratio	12.75
6	YC_mean_time	4.22
7	YC_per_QC_ratio	3.92
8	QC_stdev_time	3.86
9	YC_stdev_time	2.30
10	truck_per_QC_ratio	1.44
11	YC_rule	0.28
12	truck_stdev_time	0.24

Table 4 indicates a clear dominance of QC_rule in influencing the model's predictions. Its prominence among the features reflects the significant role of scheduling decisions in determining output behavior. This reinforces the practical insight that operational strategies beyond merely quantitative inputs can substantially shape predictive outcomes.

This study differs from previous research by not only addressing a specific scheduling subproblem, but also by evaluating the relative contribution of each subcomponent using feature importance derived from machine learning models. While many prior studies in container terminal operations focus on optimizing individual subsystems independently, this approach considers the combined effect of multiple scheduling decisions across stages. The results indicate that scheduling rules for quay cranes and internal trucks are among the most influential in shaping overall system performance. This perspective supports more focused and data-driven optimization strategies, without disregarding the role of other components.

5. Conclusion

This study demonstrates the value of machine learning techniques in understanding and prioritizing optimization efforts in container terminal operations. The findings indicate that scheduling rules for quay cranes and internal trucks, along with their mean processing times, are the most critical factors affecting the overall system performance. This challenges conventional approaches that treat all subproblems with equal importance and suggests a more targeted optimization strategy.

The proposed framework enables terminal operators to adapt their optimization focus based on changing operational conditions, reduces computational complexity by prioritizing high-impact areas, and provides a data-driven methodology for understanding interdependencies between operational components. Future research should (1) consider real data, e.g., the ones collected using IoT technologies or the ones obtained from the design of real systems, (2) expand this approach to include other operational aspects, and (3) explore dynamic feature importance analysis that adapts to real-time operational conditions. Characteristics of the real data would affect the definition of the input features. Some new features might be introduced following the nature of the data, e.g., the probability distribution of the features, lower and upper quartiles of the feature values, or possibly some new features that are calculated based on some basic ones. Such strategies might help ensure a good performance of the regression models for dealing with the specific real data. When considering different data, e.g., the real data, the importance order of the features might

be different, which would require the decision makers to perform different optimization decisions (e.g., solving different scheduling problems, reducing certain operation parameters, if possible, etc.).

References

[1] Nasution, N.K.G., Jin, X., & Singgih, I.K. (2022). "Classifying games in container terminal logistics field: A systematic review." *Entertainment Computing* **40**: 100465.

[2] Raeesi, R., Sahebjamnia, N., & Mansouri, S.A. (2023). "The synergistic effect of operational research and big data analytics in greening container terminal operations: A review and future directions." *European Journal of Operational Research* **310**: 943–973.

[3] Li, W., Cai, L., He, L., & Guo, W. (2024). "Scheduling techniques for addressing uncertainties in container ports: A systematic literature review." *Applied Soft Computing* **162**: 111820.

[4] Singgih, I.K. (2021). "Production Flow Analysis in a Semiconductor Fab Using Machine Learning Techniques." *Processes* **9**: 407.

[5] Singgih, I.K., & Singgih, M.L. (2024). "Regression Machine Learning Models for the Short-Time Prediction of Genetic Algorithm Results in a Vehicle Routing Problem." *World Electric Vehicle Journal* **15**: 308.

[6] Bai, R., Chen, X., Chen, Z.-L., Cui, T., Gong, S., He, W., et al. (2023). "Analytics and machine learning in vehicle routing research." *International Journal of Production Research* **61**: 4–30.

[7] Li, M., Wandelt, S., Cai, K., & Sun, X. (2023). "Machine learning augmented approaches for hub location problems." *Computers & Operations Research* **154**: 106188.

[8] Korekane, S., Nishi, T., Tierney, K., & Liu, Z. (2024). "Neural network assisted branch and bound algorithm for dynamic berth allocation problems." *European Journal of Operational Research* **319**: 531–542.

[9] Saleh, M.A.M., Hicham, A., Maâti, M.L.B., Taha, H., & Mohammed, Y.M.A. (2024). "Development of a sustainable strategy model for predicting optimal container stacking locations in container yards using artificial intelligence and cubic data." *Kuwait Journal of Science* **51**: 100174.

[10] Yang, Y., He, S., & Sun, S. (2023). "Research on the Cooperative Scheduling of ARMGs and AGVs in a Sea–Rail Automated Container Terminal under the Rail-in-Port Model." *Journal of Marine Science and Engineering* **11**: 557.

[11] Okwir, S., Amouzgar, K., & Ng, A.H.C. (2025). "Exploring prediction accuracy for optimal taxi times in airport operations using various machine learning models." *Journal of Air Transport Management* **122**: 102684.

[12] Wang, F.-K., & Mamo, T. (2020). "Gradient boosted regression model for the degradation analysis of prismatic cells." *Computers & Industrial Engineering* **144**: 106494.

[13] Salazar-Rojas, T., Cejudo-Ruiz, F.R., & Calvo-Brenes, G. (2022). "Comparison between machine linear regression (MLR) and support vector machine (SVM) as model generators for heavy metal assessment captured in biomonitors and road dust." *Environmental Pollution* **314**: 120227.

[14] Abreu, L.R., Maciel, I.S.F., Alves, J.S., Braga, L.C., & Pontes, H.L.J. (2023). "A decision tree model for the prediction of the stay time of ships in Brazilian ports." *Engineering Applications of Artificial Intelligence* **117**: 105634.

[15] Rahman Mahin, M.P., Shahriar, M., Das, R.R., Roy, A., & Reza, A.W. (2025). "Enhancing Sustainable Supply Chain Forecasting Using Machine Learning for Sales Prediction." *Procedia Computer Science* **252**: 470–479.

[16] Arunadevi, M., Rani, M., Sibinraj, R., Chandru, M.K., & Durga Prasad, C. (2023). "Comparison of k-nearest Neighbor & Artificial Neural Network prediction in the mechanical properties of aluminum alloys." Materials Today: Proceedings.

[17] Ekiz, B., Baygul, O., Yalcintan, H., & Ozcan, M. (2020). "Comparison of the decision tree, artificial neural network and multiple regression methods for prediction of carcass tissues composition of goat kids." *Meat Science* **161**: 108011.