

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

# Maritime Transport Research

journal homepage: [www.elsevier.com/locate/martra](http://www.elsevier.com/locate/martra)

Full length article

## Understanding and predicting quay crane breakdowns using explainable AI

Robert Klar<sup>a,b</sup> , Anders Andersson<sup>c</sup> , Vangelis Angelakis<sup>a</sup> <sup>a</sup> Department of Science and Technology (ITN), Linköping University (LiU), Campus Norrköping Luntgatan 2, 601 74 Norrköping, Sweden<sup>b</sup> Department of Traffic Analysis and Logistics (TAL), Swedish National Road and Transport Research Institute (VTI), Olaus Magnus väg 35, 583 30, Linköping, Sweden<sup>c</sup> Department of Vehicle Systems and Driving Simulation (FSK), Swedish National Road and Transport Research Institute (VTI), Olaus Magnus väg 35, 583 30, Linköping, Sweden

### ARTICLE INFO

#### Keywords:

Quay cranes  
 Container terminal operations  
 Breakdown prediction  
 Predictive maintenance  
 Machine learning  
 Explainable artificial intelligence (XAI)  
 Port performance

### ABSTRACT

Quay cranes (QCs) play a vital role in ship-to-shore operations, enabling the seamless transfer of cargo between sea and land. However, increasing trade volumes require faster and more cost-effective container handling, exerting significant pressure on QCs and leading to greater wear on critical components such as wires, hoists, and rope clamps. While operations research has explored maintenance scheduling to improve terminal performance, comparatively little work has examined how machine learning can exploit the growing volume of QC monitoring and operational data to predict breakdowns before they occur. This study contributes to this area by integrating terminal operations data, QC monitoring logs, and meteorological observations into a unified analytical framework. We employ explainable artificial intelligence (XAI), using both global and local SHapley Additive exPlanations (SHAP) to identify the operational and environmental factors most strongly associated with QC failures and to illustrate concrete, instance-level examples of how specific conditions contribute towards breakdowns. In parallel, we develop a robust machine learning pipeline built around nested cross-validation to assess the predictive capability of multiple classifiers for forecasting QC breakdowns. Our XAI analysis reveals that breakdown risk is closely linked to QC working time, the distribution of moves across simultaneously operating QCs, hoist overload and trolley alignment warnings, and adverse weather conditions. Among the evaluated models, LightGBM achieved the highest predictive accuracy, reaching up to 83% in identifying breakdown-prone scenarios. These findings demonstrate the feasibility and value of data-driven predictive maintenance for QCs, providing insights that support safer, more reliable, and more efficient terminal operations.

### 1. Introduction

Global container throughput is increasing rapidly. It is projected to reach 978 million twenty-foot equivalent units (TEUs) in 2025, which is a 28.85% increase compared to 759 million TEUs in 2020 and a 334.67% increase compared to 225 million TEUs in 2000 (Kim, 2024). Container port terminals are thus often referred to as co-drivers of globalization (Nikolaou and Dimitriou, 2021).

In this context, quay cranes (QCs), which transfer containers between vessels and terminal trucks, are essential for the seamless integration of marine and landside port terminal operations (Dragović et al., 2025; Kim, 2024). These cranes play a key role in

\* Correspondence to: Olaus Magnus väg 35, 583 30, Linköping, Sweden.

E-mail address: [robert.klar@vti.se](mailto:robert.klar@vti.se) (R. Klar).

<https://doi.org/10.1016/j.martra.2026.100152>

Received 18 February 2026; Received in revised form 2 April 2026; Accepted 28 April 2026

Available online 12 May 2026

2666-822X/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

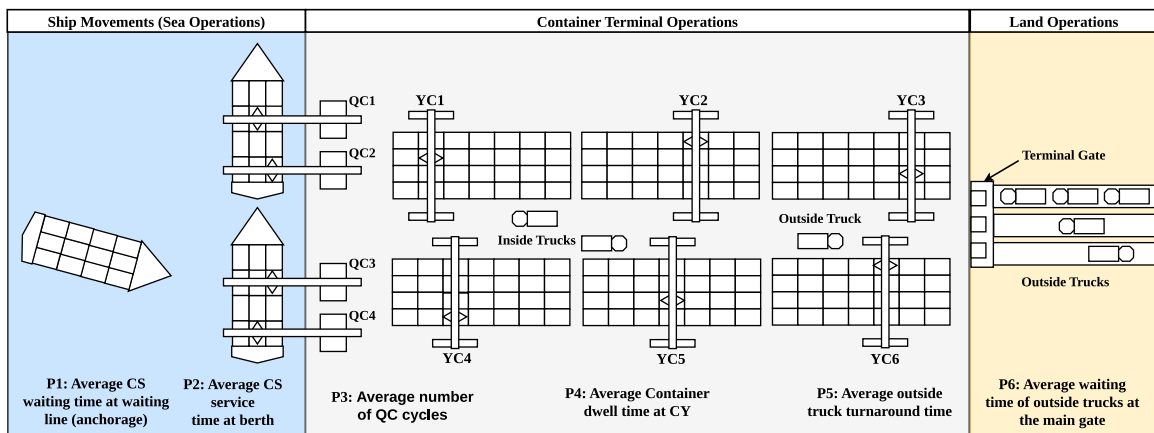


Fig. 1. Schematic overview of a container terminal (CT), adapted from Klar et al. (2024a), in conjunction with six port performance indicators (Notteboom et al., 2021).

the intricate chain of port operations by enabling ship-to-shore and shore-to-ship operations, thereby facilitating the loading and unloading of containers to and from container ships (CS) (Notteboom et al., 2021).

The steady growth of maritime transport, coupled with increasing trade volumes and ship sizes, requires the efficient and ideally uninterrupted operation of port facilities and equipment (Kizilay and Eliyi, 2021). In addition, ports are under increasing pressure to improve their profitability, environmental friendliness, energy performance, and operational efficiency due to growing global sustainability efforts (Klar et al., 2023). This is particularly true for QCs since a significant number of handling operations are required for ship-to-shore operations (Liu and Ge, 2018).

As a result, port managers are striving to complete vessel operations as swiftly as possible to increase resource utilization, increase vessel turnover, reduce vessel waiting times, and improve customer satisfaction (Merkel et al., 2022; Cahyono et al., 2022) — all of which are critical for competitiveness in the maritime industry (Abou Kasm and Diabat, 2020).

However, this drive for efficiency leads to increased wear and tear on port equipment (Notteboom et al., 2021). Among all equipment, QCs are the most error-prone, as they bridge maritime and landside operations and are directly impacted by container throughput, often becoming the bottleneck in terminal operations (Carlo et al., 2015). Their vulnerability stems from the complexity of tasks, high workload, limited availability due to high investment and maintenance costs, and safety concerns, especially when multiple QCs operate in parallel (Klar et al., 2023). Consequently, preventive maintenance activities such as component replacement, equipment checks, cleaning, and lubrication must be performed quarterly, monthly, or even weekly (Li et al., 2019).

Fig. 1 presents a schematic of a container terminal (CT), illustrating three types of handling equipment: QCs, which transfer containers between a CS and internal trucks; yard cranes (YCs), which transfer containers between internal or external trucks and storage block in the container yard (CY); and internal trucks, which deliver containers between QCs and YCs. The figure also highlights six key terminal performance indicators (KPIs), with QC performance (P3 in Fig. 1) recognized as a critical KPI in port operations.

This metric is frequently identified as a bottleneck due to limitations in average QC movements per hour and is directly affected by QC downtime. It is especially crucial for maritime shipping companies, as it directly influences vessel service time at the port (Notteboom et al., 2021). Furthermore, port operations are highly interconnected (Klar et al., 2024a), meaning QC downtime disrupts both ship-to-shore and shore-to-ship operations, affecting marine and landside activities alike.

In response, recent research has treated QC maintenance (and scheduling) jointly with various other port operations using multi-objective optimization approaches. These include (1) integrated scheduling of various handling equipment in automated CTs, including QCs, YCs, and automated guided vehicles (AGVs), while accounting for QC failures (Li et al., 2024), (2) a bi-objective optimization model of integrated berth allocation and QC assignment (Xu et al., 2025) with preventive QC maintenance activities (Li et al., 2022), and (3) solving an integrated berth allocation, QC assignment, and truck deployment problem considering QC maintenance (Wang et al., 2024).

Building on the classification from Li et al. (2019), current research addresses QC disruptions from three main perspectives: the specific types of QC faults, the underlying causes of these faults, and the rescheduling of QC tasks following fault occurrences. While most existing literature emphasizes the scheduling aspect, there is limited research leveraging machine learning (ML) to understand and predict fault causes.

Existing ML-based studies on QC breakdown prediction are typically focused on specific components, such as the electric control system (Gothandapani et al., 2024; Putra et al., 2024) and the hoist system (Jalal et al., 2023; Mukherjee et al., 2024), and often lack integration with terminal operations or weather data. The need for more comprehensive ML-driven research into QC downtime is also highlighted by Filom et al. (2022) and Gothandapani et al. (2024), who emphasize the importance of applying ML in port asset management, particularly for QCs. Specifically, the need for data-driven retrofit prediction models to support timely and effective maintenance interventions is highlighted (Gothandapani et al., 2024).

In response to these gaps, this study proposes a predictive framework that leverages the growing volume of data generated by modern QCs, which have evolved from primarily mechanical systems into highly computerized, sensor-rich assets (Tao et al., 2024). Using supervised ML techniques and QC monitoring data from a medium-sized Swedish port, the framework integrates not only QC sensor data but also historical terminal operations and weather data to predict the causes of QC downtime.

The core intention of this work is to enable early identification of potential QC breakdowns through ML classifiers, thereby supporting a shift from reactive to preventive maintenance strategies. To ensure relevance and precision, the analysis focuses exclusively on QCs operating at the critical ship-to-shore link, rather than on broader CT operations. This targeted approach allows us to underline the 15 most QC-relevant features (see Fig. 4), which are directly tied to the operational and environmental conditions affecting QC performance.

The paper is organized as follows. Section 2 provides the necessary background on predictive maintenance in the context of Industry 4.0 (2.1) and terminal operations (2.2), identifies research gaps in QC maintenance (2.3), and introduces explainable artificial intelligence (2.4) before outlining the contributions of this study (2.5). Section 3 outlines the methodology, including data collection and feature engineering (3.1), feature contribution assessment (3.2), the ML classifiers used (3.3), and the implementation of the ML pipeline using nested cross-validation (3.4). The results are presented in two parts: Section 4 explores global feature importance and instance-level examples of how individual conditions contribute to breakdowns, while Section 5 evaluates the performance of the selected classifiers. Section 6 discusses the findings and their potential policy implications. Finally, Section 7 summarizes the key insights and outlines directions for future research.

## 2. Identified research gaps and contributions of this study

This section begins by highlighting the importance of predictive maintenance within the context of Industry 4.0 (2.1), with a particular focus on terminal operations (2.2). It then reviews related work on QC breakdown prediction and maintenance (2.3), followed by an overview of explainable artificial intelligence in predictive maintenance (2.4). These subsections collectively chart the area identifying clear research gaps and support the contribution of this study (2.5).

### 2.1. Predicting downtime in industry 4.0

The emergence of Industry 4.0 has significantly transformed manufacturing operations by integrating advanced technologies such as big data analytics, ML and artificial intelligence (AI), Internet of Things (IoT), and cloud computing into production processes, thereby enabling more efficient and resilient industrial systems. A crucial aspect of this transformation is the accurate prediction of machinery failure or downtime, which remains one of the primary sources of revenue loss and operational disruption in modern manufacturing environments (Vuttipittayamongkol and Arreeras, 2022). These environments are rich in diverse data sources, such as industrial IoT sensors on the machinery and infrastructure, open-source datasets, and historical operational data. These have come with a promise of significantly improvement in anticipating and timely responding to machine failures.

This promise is underpinned on effectively fusing these heterogeneous data streams, to derive comprehensive and accurate predictive insights, enhancing model robustness and reliability (Huang et al., 2020).

The integration of ML-based fault prediction not only improves operational foresight and reduces maintenance costs, enabling predictive maintenance, but also supports strategic planning and decision-making processes, reinforcing the resilience and adaptability of manufacturing systems in the face of dynamic market conditions (Kashpruk et al., 2023). In this context, latency-aware resource allocation in edge computing environments becomes increasingly relevant, particularly in industrial and port settings where low-latency predictive maintenance is critical (Ahmadvand and Foroutan, 2025).

### 2.2. Predicting operations' disruptions in port terminals

Port operations, although critical in global supply chains, remain vulnerable to equipment failures and operational disruptions that can cascade into accidents, significant delays, and financial losses. Predicting such disruptions, particularly those involving terminal equipment like QCs, straddle carriers, and AGVs, is essential for improving safety, resilience, minimizing maintenance costs, and ensuring throughput efficiency (Knatz et al., 2022). Disruptions in this context can be caused by mechanical wear and tear, software or control system failures, environmental conditions (e.g., wind, salt corrosion), operator errors, or inadequate maintenance schedules (Lam and Su, 2015). The challenge becomes even more complex in the era of Industry 4.0, where many ports operate with semi-automated control systems integrating physical machinery with digital infrastructure, such as IoT sensors, predictive analytics, and machine-to-machine communication (Klar et al., 2023). While these technologies offer opportunities for real-time monitoring and proactive maintenance, they also introduce new vulnerabilities and call for sophisticated predictive models (Jbair et al., 2022). As ports evolve into cyber-physical systems (CPS), robust analytical and forecasting approaches are essential to ensure system reliability and operational continuity.

**Table 1**  
Comparison of relevant papers on predicting and preventing QC downtime and their contributions.

Reference	Objectives	Applied methodology	Conclusions
Gothandapani et al. (2024)	Analyze the performance degradation of a number of QCs and assess the potential of retrofitting.	Prospective longitudinal panel study over 16 years to analyze performance of five QCs using KPIs.	Electrical control system is identified as the primary source of performance degradation.
Putra et al. (2024)	Early failure detection of QCs using IoT-based measurement data.	Failure mode and effects analysis based on IoT-based temperature and vibration data from the hoist motor.	Most failures occurred in the electrical power supply system and the hoisting system.
Crespo Del Castillo et al. (2024)	Development of a data-driven asset health index methodology for evaluating QC condition.	The proposed asset health index methodology consists of six steps incorporating dynamic operational data.	Degradation of QCs is influenced by operational, environmental, mechanical, electrical, and managerial factors.
Awasthi et al. (2024)	Develop a deep learning model for detecting and predicting errors in QC operations.	Long short-term memory model using synthetic minority oversampling for binary classification.	Proposed model is accurate and precise in detecting errors, but recall is limited (data spans one month).
Jalal et al. (2023)	Estimate the reliability, availability, and maintainability of QCs in CTs using data from 53 QCs over one year.	Component analysis including criticality ranking, Stochastic petri net modeling, and Monte Carlo simulation.	The mean operational availability of QCs was found to be 97%, with main host identified as the most critical component.
Mukherjee et al. (2024)	Development of an unsupervised method for detecting discordant vibration patterns in QC motors using irregular time-series data from IoT sensors.	Load-based clustering of the data and anomaly (discord) detection using one-class support vector machines for each cluster.	Their proposed discord detection framework effectively flags anomalous vibration patterns in QC motors, even under irregular and unlabeled conditions.
This study	Understanding and predicting QC breakdowns using explainable AI, integrating terminal operations, QC monitoring, and weather data.	SHAP-based feature importance and local instance-specific breakdown explanations; performance evaluation across multiple classifiers using nested cross-validation.	Identifies key operational, control-system, and environmental factors; achieves up to 83% prediction accuracy and F1 scores using LightGBM.

### 2.3. Related QC breakdown prediction work and research gaps

As the primary equipment enabling container transfer between sea and land modes of transport, QCs are essential to efficient port operations. However, their limited availability and frequent use makes downtime particularly disruptive (Li et al., 2022). This has prompted a range of research efforts aimed at understanding, predicting, and mitigating such events.

These efforts can be broadly categorized into four methodological approaches: (1) Data-driven diagnostics, which leverage sensor data and operational logs to detect early signs of failure or assess asset health (Putra et al., 2024; Crespo Del Castillo et al., 2024; Gothandapani et al., 2024); (2) Reliability modeling and simulation, which use probabilistic and stochastic frameworks to estimate downtime and system performance (Jalal et al., 2023); (3) ML-based anomaly detection and error prediction, including supervised and unsupervised models for identifying abnormal patterns in sensor data (Awasthi et al., 2024; Mukherjee et al., 2024); and (4) Operations research, which integrates proactively scheduled maintenance windows with optimized QC allocation and scheduling strategies (Li et al., 2022; Khalilpoor et al., 2025).

Table 1 provides an overview of related work on understanding and predicting QC downtime, as well as its prevention, based on the first three methodological approaches: data-driven diagnostics, reliability modeling and simulation, and ML-based prediction. These approaches are most relevant to the aim of this study, which is to understand and predict QC downtime using ML.

A recurring theme across these studies is the identification of critical subsystems that contribute disproportionately to downtime. Notably, the electrical control system has been identified as a primary source of degradation and failure. For instance, Gothandapani et al. (2024) identify it as the primary cause of long-term performance decline, and Putra et al. (2024) report frequent failures in the electrical power supply and hoisting mechanisms based on IoT sensor data. Several studies have also identified the hoist system, particularly the hoist motor and its associated components, as key causes of downtime. Jalal et al. (2023) rank hoist-related components as the most critical based on stochastic modeling, and Mukherjee et al. (2024) demonstrate that discordant vibration patterns from the hoist motor are strong indicators of emerging faults.

While these related works have contributed significantly to component-level insights, particularly regarding the hoist motor and electrical control systems, there is a lack of studies that systematically exploit all error event messages and warnings captured in the QC monitoring system and integrate them with terminal operations data and historical weather data. Gothandapani et al. (2024) also highlight this gap, emphasizing the need for a data-driven retrofit prediction model to guide timely interventions and extend the QC lifecycle. Furthermore, most studies are limited by data constraints, e.g., a one-month duration in Awasthi et al. (2024) or the absence of labeled data (Mukherjee et al., 2024), and do not consider contextual factors such as terminal operations or weather conditions. These limitations underscore the need for a more comprehensive and integrated approach towards understanding and predicting QC downtime.

### 2.4. Explainable artificial intelligence for predictive maintenance

Explainable artificial intelligence (XAI) is a key requirement for modern predictive maintenance systems, where model transparency is essential for operational trust, regulatory compliance, and actionable decision support (Dereci and Tuzkaya, 2024). The

**Table 2**  
Summary of input and output variables for the QC breakdown prediction problem.

Category	Example variable	Unit/Resolution	Data source	Role
<b>Operational Features</b> ( $D_{ops}$ )	Operation duration	Minutes	Terminal operations records (Yilport Gävle)	Input
	Number of QC moves	QC moves per crane & vessel		
<b>Monitoring Data</b> ( $D_{mon}$ )	Hoist overload warnings	Count per operation; event duration in minutes	QC monitoring system (Yilport Gävle)	Input
	Trolley alignment errors	Count per operation; event duration in minutes		
<b>Weather Variables</b> ( $D_{weather}$ )	Temperature	°C (15-min interval)	Swedish Meteorological and Hydrological Institute	Input
	Humidity	% (15-min interval)		
<b>Output Variable</b>	QC downtime occurrence	Binary (1 = downtime, 0 = normal)	Derived from breakdown causing events (see Alg 1)	Output

need for XAI arises from the black-box nature of many machine learning models, which makes it difficult for humans to understand how decisions are made (Abdulrashid et al., 2024). XAI seeks to clarify how AI models produce their outputs, typically through post-hoc methods that interpret an already-trained model at either a global or local level (Cummins et al., 2024).

XAI approaches are often categorized into model-agnostic and model-specific techniques, both of which offer pathways to enhance user trust while maintaining high-performing predictive maintenance systems (Cummins et al., 2024). Applying XAI on top of predictive frameworks reveals deeper insights into how key explanatory variables influence model outputs, thereby supporting more informed and practical maintenance decisions (Ghosh and De, 2024).

While XAI has gained momentum in predictive maintenance across various industries (Cummins et al., 2024), its application in the maritime sector, particularly container terminals, remains limited. Existing work has primarily focused on understanding or predicting import container dwell times (Lee et al., 2024) or forecasting fuel consumption, leaving the potential of XAI for equipment-related tasks, such as quay crane breakdown prediction, largely unexplored (Ghosh and De, 2024).

### 2.5. Contribution of this study

This study addresses the identified research gap by proposing a data-driven prediction model to understand QC breakdowns and guide timely interventions to mitigate its impact. Unlike previous studies that focus on individual QC components, this work analyzes the entire QC system in its role as the ship-to-shore link, which is critical to terminal operations. Specifically, we leverage historical sensor data, terminal operations logs, and weather conditions to understand and predict QC downtime. The absence of XAI in QC breakdown prediction represents a clear research gap addressed in this work. The key contributions of this study are:

- 1. Real-world data integration:** We utilize extensive data from a medium-sized Swedish port, combining QC monitoring, terminal operations, and weather conditions to support a shift from reactive to preventive maintenance.
- 2. Model development and evaluation:** We propose a supervised ML pipeline that includes data aggregation, feature engineering, feature selection, and performance evaluation using nested cross-validation.
- 3. Interpretability through XAI:** We apply XAI methods to identify key factors, such as QC usage intensity, mechanical warnings, and adverse weather conditions, contributing to downtime.
- 4. Actionable insights:** Based on the most influential features, we recommend proactive maintenance strategies to enhance the resilience and efficiency of port operations.

Beyond these practical contributions, this study advances scientific research on QC reliability and predictive maintenance in several ways. First, to the best of our knowledge, this is the first study to integrate QC monitoring logs, terminal operations data, and weather observations into a unified predictive framework for QC downtime. Second, it introduces explainable AI (SHAP) to this field, addressing a methodological gap in previous studies that typically analyze isolated subsystems or short-duration datasets. Lastly, the use of interpretable feature attribution and local explanations provides new insights into the environmental, operational, and mechanical factors contributing to QC downtime, thereby advancing the theoretical understanding of downtime occurrence.

### 3. Methodology

This section presents the methodology employed in this study, which consists of a comprehensive ML pipeline encompassing all stages from data aggregation to model evaluation. The pipeline is organized into three main components: (1) data aggregation and preprocessing, which integrates operational, monitoring, and weather data, and formulates the prediction problem (3.1); (2) XAI-based feature contribution assessment using SHAP, which identifies key features driving breakdowns and enhances interpretability (3.2); and (3) model training and performance evaluation using multiple classifiers 3.3 with nested cross-validation to ensure robust and unbiased performance estimates (3.4, 3.5, 3.6). Fig. 2 provides an overview of the full pipeline.

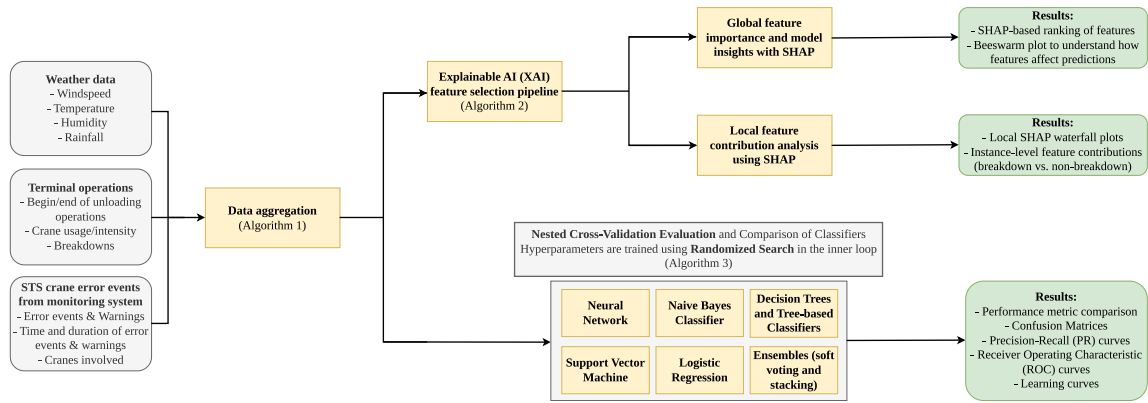


Fig. 2. Overview of the methodological framework combining XAI-based interpretability and predictive ML evaluation.

### 3.1. Data aggregation and feature construction

The data used in this study comprises three distinct sources: two proprietary datasets from Yilport Gävle, the terminal operator of the port of Gävle, and one open dataset. An overview of these datasets, including example variables, units, data sources, and role within the prediction model, is provided in Table 2.

The first,  $D_{\text{ops}}$ , contains terminal operations data for each vessel call, including start and end times for berthing and QC usage. It also includes QC usage in terms of time and container handling across three QCs. This allows for an assessment of each individual crane, as well as a correlation to the other QCs if they are working in parallel.

The second,  $D_{\text{mon}}$ , is a log of error events and warnings captured by the QC monitoring system, detailing the type, duration, and location of each event. While not all recorded events result in operational disruptions, this study investigates whether they can serve as indicators of QC downtime.

The third dataset,  $D_{\text{weather}}$ , provides historical weather data, such as temperature, humidity, rainfall, and wind speed, obtained from the Swedish Meteorological and Hydrological Institute (SMHI) (Swedish Meteorological and Hydrological Institute (SMHI), 2025).

All data used in this study was collected over the full calendar years of 2023 and 2024, during which 347 vessel calls occurred. Depending on vessel size and cargo volume, each vessel call could involve up to three QCs operating in parallel, resulting in varying levels of work intensity. For each call, it is known whether the QCs operated without interruption or experienced a breakdown, based on the set of breakdown-causing events specified ( $E_{\text{crit}}$ ). This enables post-hoc identification of contributing operational or environmental factors and forms the basis of the binary classification problem investigated in this study.

To merge data streams with differing timestamps, the terminal operation window (defined by the start and end times of each vessel call) serves as the common temporal reference. Within this window, QC monitoring events are filtered by crane ID and timestamp, then summarized by count and duration. Weather observations from SMHI's 15-minute records are aggregated into mean, minimum, and maximum values. These monitoring and meteorological features are combined with operational metrics, such as QC moves, utilization, and loading/discharging time.

The objective of this study is to identify which operational characteristics, error events, and weather conditions contribute to disruptions. To achieve this, we develop a predictive model that integrates these three complementary datasets.

Each resulting observation  $i$  corresponds to a single QC during a vessel call and is represented by a feature vector  $x_i \in \mathbb{R}^p$ , which can be decomposed as

$$x_i = [x_i^{\text{ops}}, x_i^{\text{mon}}, x_i^{\text{weather}}],$$

where the components refer to operational, monitoring, and environmental variables as exemplified in Table 2. Collectively, the dataset is defined as a feature matrix  $X \in \mathbb{R}^{n \times p}$ , where  $n$  denotes the number of crane-vessel observations and  $p$  the number of features.

Each observation is associated with a binary target variable  $y_i \in \{0, 1\}$ , where

$$y_i = \begin{cases} 1, & \text{if a breakdown occurred during the operation window,} \\ 0, & \text{otherwise.} \end{cases}$$

The predictive task is to learn a function that maps each feature vector to the probability of a breakdown:

$$\hat{y}_i = f(x_i) \approx \mathbb{P}(y_i = 1 | x_i), \quad f: \mathbb{R}^p \rightarrow [0, 1]. \quad (1)$$

It should be noted that both features and target are constructed over the same operational time window. Consequently, the proposed model is not intended to predict breakdowns in advance at a specific time point, but rather, to characterize and identify

conditions associated with breakdown occurrence at the level of the vessel call. While the framework is not strictly forward-looking, it remains essential for quantifying how well breakdown occurrences can be explained by observed operational, monitoring, and environmental factors. The use of nested cross-validation ensures that these relationships generalize beyond the training data, providing robust and unbiased estimates of predictive performance.

The full data aggregation and feature construction procedure is outlined in Algorithm 1. All features are derived exclusively from data within each vessel's operational window, ensuring that the aggregated features accurately represent the operational conditions within the time window associated with the breakdown outcome. Only events occurring within this window are included in the aggregated feature vector. On average, two QCs are involved in each vessel call, resulting in a dataset of 735 QC-specific observations, of which 230 correspond to breakdown cases. Since each vessel call may be represented multiple times, once for each assigned QC, this structure enables crane-specific analyses and interventions while capturing interactions among QCs operating simultaneously.

---

**Algorithm 1** Crane-Level Data Aggregation and Feature Engineering for QC Breakdown Prediction

---

```

1: Input: Vessel-level terminal operations dataset ( $D_{ops}$ ), QC monitoring event logs ( $D_{mon}$ ), Quarter-hourly weather observations ( $D_{weather}$ ), Set
   of downtime causing event types ( $E_{crit}$ ), such as emergency brakes
2: Output: Crane-level dataset with temporally aligned operational, event-based, and weather features, and breakdown labels
3: Step 1: Transform Vessel-Level to Crane-Level Records
4: for each vessel call in  $D_{ops}$  do
5:     for each QC assigned to the vessel do
6:         Create one crane-level record
7:         Compute crane-specific workload metrics (working hours, crane performance, utilization share)
8:     end for
9: end for
10: Step 2: Temporal Alignment of Monitoring Events
11: for each crane-level record with time window  $[t_{start}, t_{end}]$  do
12:     Filter  $D_{mon}$  by crane identifier and  $t \in [t_{start}, t_{end}]$ 
13:     Compute event-type counts for all observed event types
14:     Compute aggregated monitoring features:
15:         total number of events, number of unique event types
16: end for
17: Step 3: Derivation of Breakdown Targets
18: Define breakdown label:
19:  $y = 1$  if any event in  $E_{crit}$  occurs within  $[t_{start}, t_{end}]$ , else 0
20: Step 4: Weather Feature Aggregation
21: for each crane-level record with time window  $[t_{start}, t_{end}]$  do
22:     Filter  $D_{weather}$  to observations within  $[t_{start}, t_{end}]$ 
23:     Compute summary statistics (mean, max) for:
24:         temperature, humidity, rainfall, wind speed
25: end for
26: Step 5: Feature Consolidation
27: Concatenate operational features, monitoring features, and weather features into a single feature vector per crane-level record
28: Step 6: Leakage-Aware Feature Filtering (for ML)
29: Remove descriptive identifiers (vessel name, timestamps, crane ID)
30: Remove all features from  $E_{crit}$  used to derive target  $y$  to avoid data leakage
31: Step 7: Final Dataset Construction
32: Assemble all crane-level records into a structured dataset
33: Return: ML-ready dataset with features  $X$  and target  $y$ 

```

---

### 3.2. XAI-based feature contribution pipeline for understanding QC breakdowns

The goal of this pipeline is to identify which features are most insightful for understanding the occurrence of QC breakdowns during vessel handling, specifically during loading and unloading operations. To achieve this, we employ SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), a game-theoretic approach that explains model outputs by assigning an importance value to each feature for a given prediction.

To enhance interpretability and capture nonlinear relationships, we use tree-based ensemble models, in particular LightGBM (Ke et al., 2017), which also achieves the highest predictive performance in our evaluation (see Table 4). Tree-based models are particularly well suited for SHAP-based analysis due to the availability of efficient exact algorithms (TreeSHAP) (Lundberg et al., 2020), enabling consistent and computationally efficient attribution of feature contributions.

This SHAP-based framework allows us to analyze how input features influence model predictions at both the global and local levels. Global feature importance is assessed by aggregating SHAP values across all observations, while local explanations provide instance-level insights into how specific operational or environmental conditions contribute to predicted breakdowns.

To formalize the computation of feature contributions, SHAP decomposes the prediction for each observation  $i$  as:

$$\hat{y}_i = f(x_i) = \phi_0 + \sum_{j=1}^p \phi_j^{(i)}, \quad (2)$$

where  $\phi_0$  denotes the baseline prediction (the expected model output across all observations) and  $\phi_j^{(i)}$  represents the contribution of feature  $j$  to the prediction for instance  $i$  (Lundberg and Lee, 2017).

Global feature importance is quantified as the average absolute SHAP value across all observations:

$$\text{Importance}_j = \frac{1}{n} \sum_{i=1}^n |\phi_j^{(i)}|, \quad (3)$$

providing a ranking of features according to their overall influence on model predictions.

To complement the SHAP analysis, we present boxplots of the most relevant features identified, stratified by breakdown occurrence. These visualizations provide an intuitive understanding of how feature distributions differ between classes and support the interpretation of the model-derived feature importance. To further quantify whether the observed differences are statistically significant, we apply the Mann–Whitney U test (Chicco et al., 2025), a non-parametric test that evaluates whether values from one group tend to be systematically higher or lower than those from another by comparing their ranks. This approach is well suited for small samples and non-normal distributions. Effect sizes are reported using rank-biserial correlations, and 95% confidence intervals are estimated via bootstrap resampling of the median difference. The complete pseudocode outlining the feature extraction and decision-making process of the LightGBM classifier is provided in Algorithm 2.

---

#### Algorithm 2 Feature Importance Analysis for QC Breakdowns using LightGBM and SHAP

---

- 1: **Input:** CSV dataset containing breakdown labels (binary), operational variables, QC monitoring events, and weather features.
  - 2: **Load and Prepare Data:**
  - 3:   Read dataset as DataFrame (see Alg 1)
  - 4:   Define target:  $y \leftarrow$  breakdown column
  - 5:   Define feature matrix:  $X \leftarrow$  dataset excluding target and non-informative columns
  - 6: **Low Variance Filter:**
  - 7:   Remove features with variance below threshold (e.g.,  $< 0.01$ )
  - 8: **Model Training:**
  - 9:   Train class-balanced LightGBM classifier on feature matrix  $X$
  - 10: **SHAP-Based Explainability Analysis:**
  - 11:   Compute SHAP values using TreeExplainer (see Lundberg et al. (2020))
  - 12:   Compute global feature importance using mean absolute SHAP values
  - 13:   Generate SHAP summary bar plot (global importance ranking)
  - 14:   Generate SHAP beeswarm plot (feature impact and direction)
  - 15: **Local Explanation:**
  - 16:   Select representative breakdown and non-breakdown examples
  - 17:   Generate SHAP waterfall plots explaining individual predictions
  - 18: **Feature Distribution Analysis:**
  - 19:   Select top features based on SHAP importance ranking
  - 20:   Generate boxplots comparing breakdown vs. non-breakdown distributions
  - 21: **Output:**
  - 22:   Ranked list of important features
  - 23:   Global and local SHAP explanation plots
  - 24:   Feature distribution visualizations supporting interpretability
- 

### 3.3. Overview of applied machine learning classifiers

The classifiers used, tuned, and compared in our nested, cross-validation-based performance evaluation are presented below.

1. **Logistic Regression:** A linear classification model that estimates probabilities using a sigmoid function (Sarker, 2021).
2. **Gaussian Naive Bayes (GaussianNB):** A probabilistic classifier based on Bayes' theorem, assuming features follow a normal distribution (Sarker, 2021).
3. **Neural Networks:** The Multilayer Perceptron is a feedforward neural network that learns non-linear relationships through layered processing and weight optimization (Sarker, 2021).
4. **Support Vector Machine (SVM):** Constructs optimal hyperplanes to separate data classes and can be extended to non-linear problems using kernel functions (Sarker, 2021).
5. **Decision Tree:** A versatile, non-parametric supervised learning method used for classification and regression by sorting instances from a root node to leaf nodes based on specific attribute tests (Sarker, 2021).
6. **Random Forest:** An ensemble method that builds multiple decision trees in parallel and aggregates their outputs to improve accuracy and reduce overfitting (Sarker, 2021).

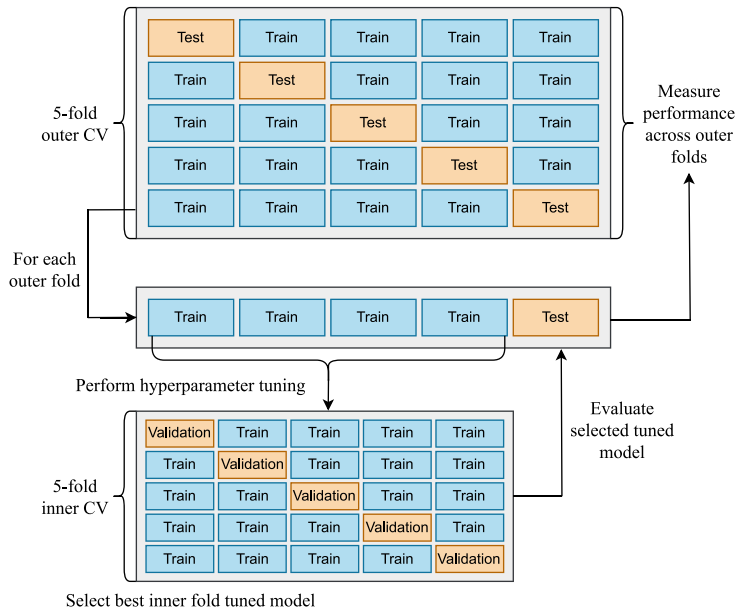


Fig. 3. Overview of nested cross validation.

- Gradient Boosting Machine (GBM):** Builds decision trees sequentially, with each tree trained to correct the residual errors of the previous ensemble (Natekin and Knoll, 2013).
- LightGBM:** A highly efficient implementation of GBM, optimized for scalability and performance in high-dimensional and large-scale datasets (Ke et al., 2017). While traditional implementations can be computationally expensive due to evaluating splits across many samples and features, LightGBM introduces techniques to mitigate these costs. Gradient-based one-side sampling retains instances with large gradients (more informative) while randomly sampling from those with small gradients, reweighting them to preserve an unbiased estimation of information gain. Exclusive feature bundling exploits feature sparsity by combining mutually exclusive features into compact bundles, thereby reducing effective dimensionality with minimal information loss. In combination with histogram-based splitting and a leaf-wise tree growth strategy, these design choices substantially improve computational efficiency and make LightGBM particularly well-suited for high-dimensional, sparse datasets.

These models were selected to align with the tabular nature of the input data, which lacks strong sequential patterns, and to balance model complexity given the limited sample size and high feature dimensionality. Tree-based machine learning models are used in many domains where interpretability is important (Lundberg et al., 2020).

In addition to the individual models, ensemble methods were employed to combine the predictive capabilities of the top-performing classifiers within the nested cross-validation framework. Specifically, both soft voting and stacking approaches were applied. Soft voting aggregates predicted class probabilities across base learners, while stacking combines their outputs using a meta-learner (Shaikh et al., 2024). The ensemble models were constructed using the best-performing classifiers identified in the outer cross-validation folds. These strategies aim to mitigate the limitations of individual models, reduce model variance, and improve predictive robustness and generalization performance, which is particularly important given the relatively small sample size of the dataset.

### 3.4. Performance evaluation using nested cross-validation

Following the comprehensive ML evaluation guidelines by Lones (2024) and Kapoor et al. (2024), which highlight the unreliability of single model evaluations, we adopt cross-validation for more robust performance assessment. This method repeatedly splits the data into folds, allowing models to be trained and tested on different subsets. We use 5 outer folds with stratified k-fold cross-validation to preserve class distribution and reduce bias.

To incorporate hyperparameter tuning, we follow the same authors' recommendations and apply nested cross-validation (Yates et al., 2023). Within each outer fold, the training data is further partitioned into 5 inner folds, where each fold is iteratively used for validation while the remaining folds are used for training. This ensures a strict separation between hyperparameter tuning and final model evaluation, thereby preventing overfitting. The procedure is illustrated in Fig. 3. Feature preprocessing and selection were performed exclusively within the training data of each inner fold, ensuring that no information from validation or test folds influenced model training. Formally, for each training fold  $D_{\text{train}}^{(k)}$ , a feature selection operator  $S(\cdot)$  was applied to obtain a subset

of features  $\mathcal{F}^{(k)} = S(\mathcal{D}_{\text{train}}^{(k)})$ . Feature selection was implemented using an embedded Lasso-based approach, where  $\ell_1$ -regularization shrinks less informative feature coefficients toward zero (Fonti and Belitser, 2017). The model was then trained using only  $\mathcal{F}^{(k)}$ , and the same feature transformation was applied to the corresponding validation and test folds. This ensures that feature selection is fully nested within the training process and does not incorporate information from held-out data.

Nested cross-validation is particularly valuable in our context, as it addresses a key limitation of the study: the restricted sample size (735 observations). By ensuring that each observation is used for both training and testing in different folds, and by preventing optimistic bias through the separation of data used for model tuning and final evaluation, nested cross-validation enhances the robustness and generalizability of the results, even with a limited sample size.

The nested cross-validation procedure used in this study is summarized in Algorithm 3, which outlines the outer evaluation loop and the inner hyperparameter-tuning loop.

---

### Algorithm 3 Robust Machine Learning Pipeline for QC Breakdown Prediction

---

```

1: Input: Aggregated crane-level dataset ( $N \approx 730$  samples)
2: Step 1: Load and Prepare Data
3: Load dataset generated by Algorithm 1
4: Remove descriptive and leakage-prone features
5: Separate features  $X$  and target labels  $y$ 
6: Step 2: Define Models and Hyperparameter Spaces
7: Models = {Decision Tree, Random Forest, Logistic Regression, Neural Network, SVM, GaussianNB, LightGBM}
8: Define model-specific hyperparameter search spaces
9: Step 3: Nested Cross-Validation with Leakage Prevention
10: for each model in Models do
11:   Initialize outer stratified  $K$ -fold cross-validation ( $K = 5$ )
12:   for each outer train-test split do
13:     Construct preprocessing pipeline on outer training fold only:
14:     (i) Feature scaling (if required by model)
15:     (ii) Lasso-based feature selection
16:     Initialize inner stratified  $K$ -fold cross-validation for hyperparameter tuning
17:     Train and tune the model using inner folds
18:     Select best hyperparameters based on mean inner-fold performance
19:     Retrain model on full outer training fold using best hyperparameters
20:     Evaluate final model on held-out outer test fold
21:     Store performance metrics (F1-score, ROC-AUC, Precision, Recall, Accuracy)
22:   end for
23:   Compute mean and standard deviation of metrics across outer folds
24: end for
25: Step 4: Ensemble Learning
26: Select top-performing base models (e.g., LightGBM, Random Forest, Logistic Regression)
27: (a) Voting Ensemble
28: Combine predicted probabilities of selected models via soft voting
29: Evaluate ensemble predictions using outer test folds
30: (b) Stacking Ensemble
31: Use base model predictions as meta-features
32: Train meta-learner (e.g., Logistic Regression) on inner folds
33: Generate final predictions on outer test folds
34: Evaluate stacked model performance
35: Step 5: Performance Evaluation across Classifiers
36: Compute performance metrics (Accuracy, F1-score, Precision, Recall, ROC-AUC) to evaluate and compare classifiers
37: Generate Confusion Matrix, Precision-Recall Curve, ROC Curve, Calibration Curve, and Learning Curve for analysis
38: Output: Model performance evaluation results, Various plots to evaluate and compare model performance

```

---

### 3.5. Hyperparameter tuning using random search

Random search is a widely used hyperparameter optimization method that is less affected by the curse of dimensionality than grid search. Instead of exhaustively evaluating all parameter combinations, it samples hyperparameter values independently from predefined distributions, allowing more efficient exploration of the search space (Bischl et al., 2023).

Hyperparameter tuning was performed using randomized search within a nested cross-validation framework. The outer loop consisted of 5-fold stratified cross-validation to estimate generalization performance, while the inner loop employed 5-fold stratified cross-validation for hyperparameter optimization. For each classifier, a model-specific search space was defined, comprising key structural and regularization hyperparameters such as tree depth (e.g., 3–15), number of estimators (100–300), regularization strength (e.g.,  $C = 0.001$ –10), learning rate, and network architecture. The search spaces consisted of discrete candidate values tailored to the relatively small dataset and model complexity.

Hyperparameter optimization was conducted independently for each classifier. The number of sampled configurations was determined dynamically based on the size of the search space, with a minimum of 10 and a maximum of 50 sampled combinations per model. Model selection within the inner loop was based on mean cross-validated accuracy, reflecting the objective of maintaining a balanced overall error rate while avoiding excessive false positive predictions that would lead to unnecessary operational interruptions, which are particularly costly in this application.

In addition, ROC–AUC was considered as a complementary metric specifically for ensemble selection, as it provides a threshold-independent measure of class discrimination. While accuracy was used during hyperparameter tuning to control the overall error rate and limit costly false positive breakdown predictions, ROC–AUC enables the identification of ensembles with strong ranking performance and improved sensitivity to breakdown events across varying decision thresholds.

### 3.6. Model comparison and evaluation metrics

To evaluate classification performance, this study follows the guidelines by Tharwat [Tharwat \(2021\)](#). The binary classification task aims to predict whether a breakdown risk is absent or likely, based on QC monitoring, terminal operations, and weather data. Evaluation is based on the confusion matrix, which includes true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Key metrics include accuracy (overall correctness), precision (correctness of positive predictions), and recall (ability to identify actual positives). Since precision and recall often conflict, the F1-score is used to balance them via their harmonic mean.

Given the operational relevance of both missed detections and false alarms, multiple complementary metrics are reported to provide a comprehensive assessment of model performance. Precision–Recall (PR) curves visualize the trade-off between precision and recall across different classification thresholds, while Receiver Operating Characteristic (ROC) curves plot the true positive rate against the false positive rate. Both curves offer insight into model behavior under varying conditions, with the area under the curve (AUC) serving as a comparative measure of overall performance ([Zhou, 2021](#)).

## 4. QC feature and model insights derived from XAI

This section presents the key features identified through our XAI-based feature selection pipeline (Algorithm 2) to explain QC breakdowns. To support an intuitive understanding of how these features influence the predictions, we provide both global and local SHAP visualizations. The global plots highlight overall feature importance patterns across the dataset, while the local plots illustrate how individual features shape specific instance-level outcomes.

All SHAP analyses are based on the LightGBM model, selected for two main reasons. First, as shown in Section 5, LightGBM achieves the best predictive performance among the evaluated classifiers. Second, LightGBM is widely recognized for its speed, effectiveness, flexibility, and ability to handle large-scale data, making it highly suitable for real-time predictive maintenance applications ([Bageci Das, 2025](#)). These properties ensure that the derived explanations align with both operational demands and practical deployment considerations.

### 4.1. Global feature importance with SHAP

To identify and assess the influence of the explanatory variables, we compute the average absolute SHAP value for each feature across all samples. This metric quantifies the contribution of each variable to predicting QC downtime. The 15 most influential features, ranked by their mean absolute SHAP values, are presented in [Fig. 4\(a\)](#).

The results indicate that three major feature groups dominate the prediction of breakdowns: (1) QC operational characteristics, particularly discharge and load operation duration and the number of required QC moves; (2) system error indicators, especially those related to the hoist and trolley assemblies; and (3) environmental factors, including wind speed and temperature. Increased vessel handling time, evidenced by longer operational periods and a greater number of QC moves, is consistently associated with a higher likelihood of downtime. This is often accompanied by an increase in recorded error events and warnings. The results also show that an imbalanced operational share between QCs working in parallel affects breakdown risk.

[Fig. 4\(b\)](#) provides more detail on how the identified 15 most relevant features affect predicting breakdowns. The *x*-axis illustrates the direction and magnitude of each feature's contribution, and the *y*-axis shows its overall relative importance. Most feature patterns, such as longer discharge and load durations, a higher QC time share, and hoist- or trolley-related warnings, are associated with an increased probability of breakdowns. Conversely, certain conditions, such as higher ambient temperatures, lower wind speeds, and higher humidity, are linked to a reduced risk of breakdowns. Notably, an overall increase in moves or a higher average number of moves per hour across all QCs does not inherently elevate breakdown risk; however, QCs that carry a disproportionate share of the workload face a significantly higher risk of breakdown.

Overall, the SHAP analysis reveals that the risk of QC downtime is influenced by the interaction between operational workload, control-system disturbances, mechanical load conditions, and environmental stressors. Operational intensity features, such as total moves across QCs, QC time share, QC move share, QC discharge time, and QC load time, emerge as dominant contributors. This indicates that sustained workload pressure and uneven task allocation are key predictors of increased downtime risk. Control and safety system indicators, including gantry safe-op stops, trolley alignment warnings, ship profile validation faults, and positioning errors, further emphasize the importance of stable control systems, as frequent anomalies are associated with a higher likelihood of disruption. Mechanical load indicators, such as hoist overload and excessive load imbalance events, reflect physical stress conditions

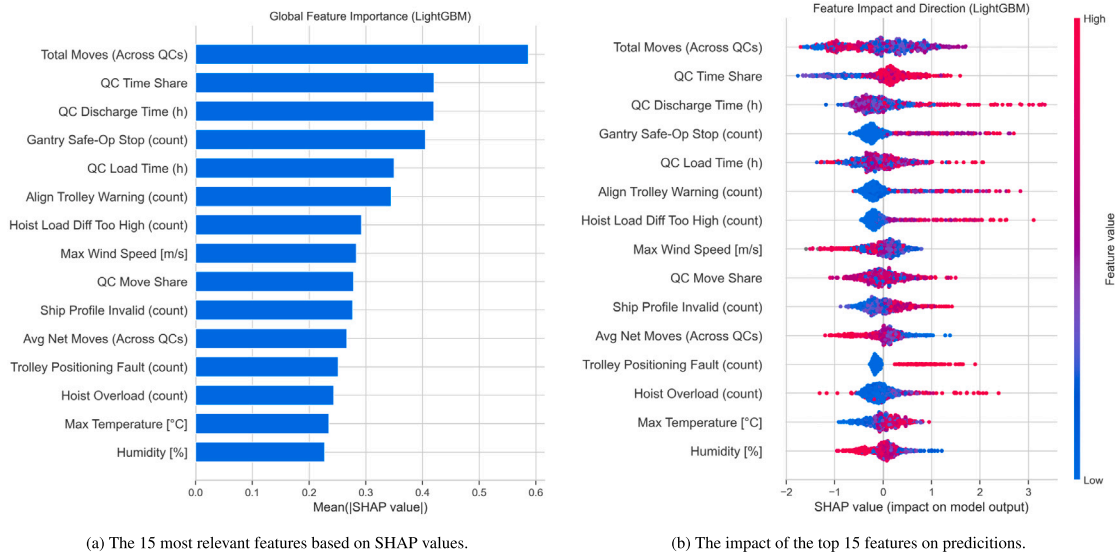


Fig. 4. Understanding relevant features and their magnitude (Subfigure 4(a)) and how these affect predictions (Subfigure 4(b)) using SHAP (Lundberg et al., 2020).

Table 3

Descriptive and inferential statistics comparing operational and environmental variables for breakdown and non-breakdown observations.

Feature	Group	Descriptive Statistics			Inferential Statistics		
		Median	Mean	Standard deviation	p-value	Effect size	95% CI
Align Trolley Warning (count)	Breakdown	1.000	3.693	9.268	0.000	-0.175	[0.000, 1.000]
	No Breakdown	0.000	2.132	6.490			
QC Load Time (h)	Breakdown	10.767	10.229	5.077	0.001	-0.183	[0.600, 3.217]
	No Breakdown	8.467	8.709	4.949			
Hoist Overload (count)	Breakdown	3.000	8.299	13.129	0.002	-0.164	[1.000, 3.000]
	No Breakdown	1.000	5.075	9.818			
QC Discharge Time (h)	Breakdown	6.785	7.369	4.856	0.014	-0.134	[0.100, 2.133]
	No Breakdown	5.667	6.261	4.167			
QC Time Share	Breakdown	0.926	0.809	0.241	0.049	-0.108	[-0.010, 0.068]
	No Breakdown	0.887	0.765	0.254			
Humidity [%]	Breakdown	81.370	77.843	13.863	0.160	0.077	[-4.514, 0.655]
	No Breakdown	82.600	78.976	15.065			
Max Wind Speed [m/s]	Breakdown	8.300	8.700	3.154	0.339	-0.052	[-0.300, 1.100]
	No Breakdown	8.000	8.462	3.315			

directly linked to protective shutdown mechanisms. Environmental factors, such as wind speed, temperature, and humidity, also contribute to QC downtime by affecting crane dynamics, mechanical stress, and electronic reliability.

These findings collectively indicate three dominant pathways used by the model when predicting QC breakdowns: operational overuse and workload imbalance, control-system instability, and mechanical overload. Environmental conditions act as amplifying factors. These insights provide a basis for predictive maintenance that is both interpretable and actionable. This enables proactive workload management, early anomaly detection in control systems, and adaptive operational planning under adverse environmental conditions.

We also acknowledge that several operational variables in our dataset exhibit correlation (e.g., QC load time, discharge time, number of moves, and crane time share). In such cases, SHAP may distribute importance across correlated features in a non-unique manner, meaning that the magnitude of individual feature attributions should be interpreted at the level of broader feature groups.

Finally, we emphasize that SHAP explains model-inferred predictive structures and does not establish causality. Consequently, the identified patterns highlight conditions associated with breakdowns for operational decision-making, but should not be interpreted as evidence of causal effects without further domain-specific validation.

To further illustrate the relationships identified through the SHAP analysis, boxplots stratified by breakdown occurrence were generated (Fig. 5). Jittered data points were overlaid to show the distribution. Points beyond 1.5 interquartile range (IQR) were excluded from the overlay for clarity. Six representative examples are shown: terminal-operation features (Subfigures 5(a) and 5(d)), warnings and error events captured by the QC monitoring system (Subfigures 5(b) and 5(e)), and weather-related influences

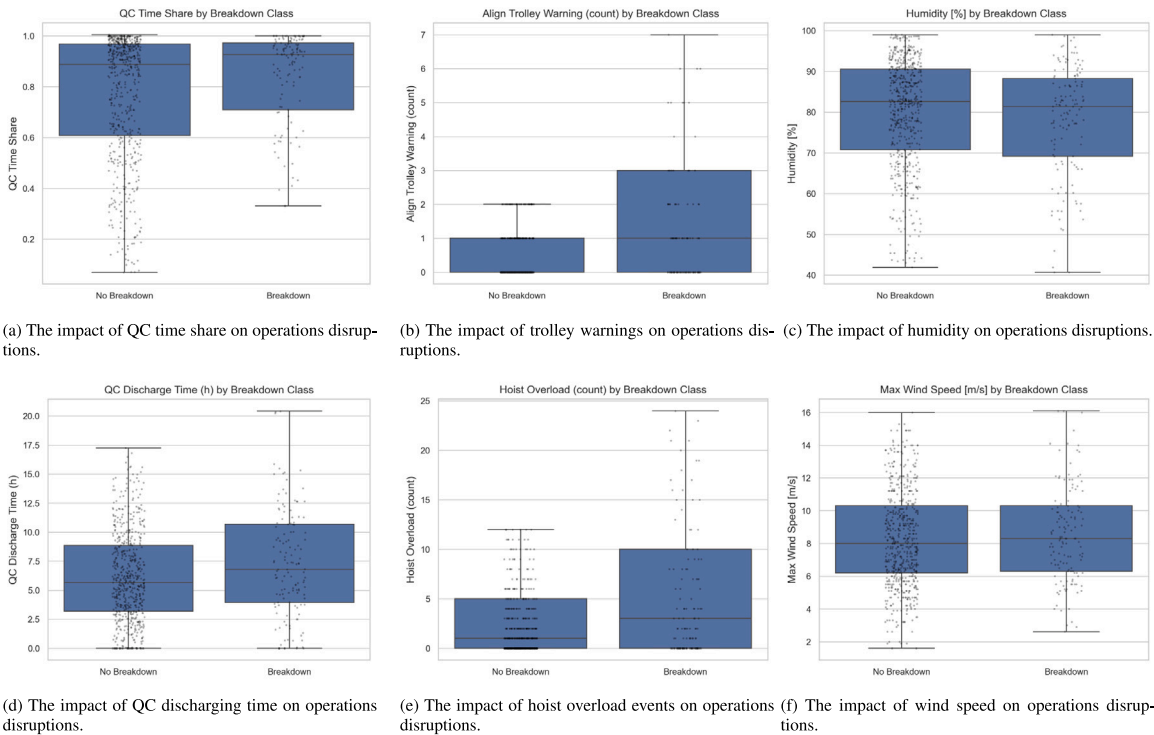


Fig. 5. Visualizing the impact of selected features from terminal operations data (Subfigures 5(a) & 5(d)), QC monitoring system records (Subfigures 5(b) & 5(e)) and weather data (Subfigures 5(c) & 5(f)) on QC breakdowns.

(Subfigures 5(c) and 5(f)). These plots confirm that the likelihood of breakdowns increases with higher operational demand, as well as with more frequent hoist overload and trolley alignment warnings. Conversely, weather impacts, such as lower temperatures and stronger winds, only mildly increase the risk of downtime on their own.

These observations are consistent with the statistical tests reported in Table 3, which presents the Mann–Whitney U test results for the selected features. Both the operational and monitoring variables shown in Fig. 5 are statistically significant ( $p < 0.05$ ), with effect sizes indicating that their distributions differ systematically between breakdown and non-breakdown cases. By contrast, the weather variables do not exhibit statistically significant differences between the two groups, suggesting that weather conditions act as potential amplifying factors rather than direct drivers of breakdowns.

Taken together, these findings reinforce that when multiple QCs operate in parallel, the risk of breakdown increases substantially if one crane carries a disproportionately large share of the workload. This pattern is also reflected in QC Time Share being the second most influential global feature in Fig. 4.

#### 4.2. Local SHAP explanations

To further visualize the decision-making process behind our feature importance analysis for assessing QC breakdowns using LightGBM and SHAP, we present two extreme cases: one clear non-breakdown case (see Fig. 6(a)) and one breakdown prediction case (see Fig. 6(b)). A SHAP waterfall plot visualizes the predicted SHAP values of a single sample by displaying the contribution of all features (Ponce-Bobadilla et al., 2024). These plots illustrate how the SHAP values (evidence) of each feature shift the model’s prediction from the baseline (mean) expected value toward the final predicted value, highlighting the positive or negative influence of each feature.

While Fig. 4 shows the global contribution of predictors across all samples, Fig. 6 illustrates how individual features drive specific predictions. For example, in Fig. 6(a), the QC in the displayed instance was used only 18.5% of the vessel operation time, resulting in a relatively short discharging duration (2.4 h) and short overall usage time (5.817 h). Moreover, no warnings occurred in the QC monitoring system during that period, collectively pushing the prediction toward a non-breakdown outcome.

In contrast, the QC in the breakdown case shown in Fig. 6(b) was utilized extensively, with more than 20 h spent discharging containers and a total usage time exceeding 45 h. This intensive usage coincided with multiple warnings captured by the QC monitoring system, including several hoist load-related warnings and ISU main contactor warnings, which may interfere with the crane’s main electrical circuit and its ability to operate normally. These factors jointly shift the model’s prediction toward a breakdown.

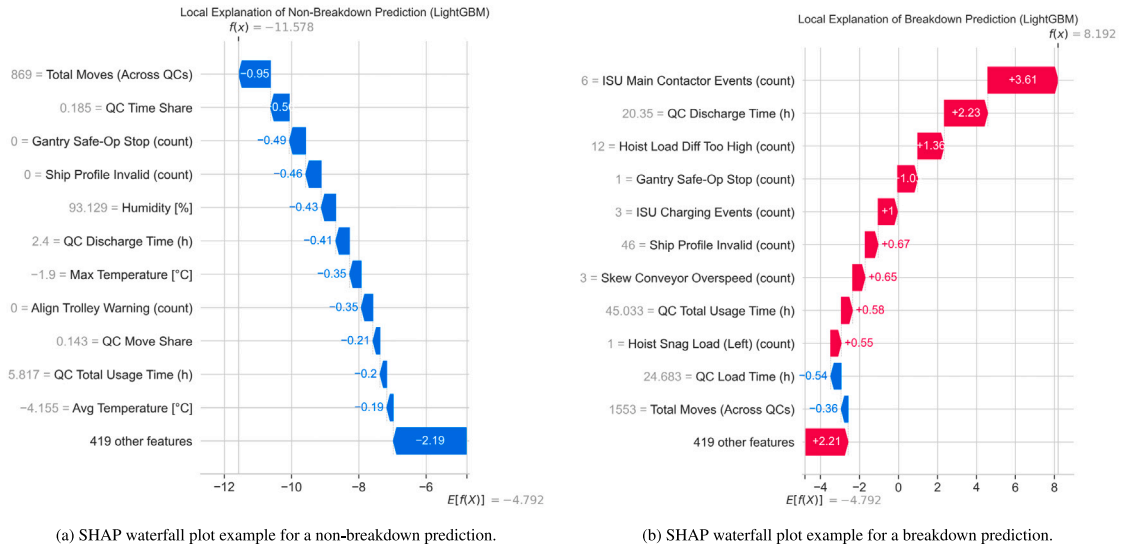


Fig. 6. SHAP waterfall plots visualizing how specific features contribute to either a non-breakdown (Fig. 6(a)) or a breakdown (Fig. 6(b)) prediction.

Table 4  
Performance Comparison of ML Models Using Nested Cross-Validation (Test Scores).

Model	Mean performance evaluation across folds					Standard deviation				
	Accuracy	F1-Score	Precision	Recall	ROC-AUC	Accuracy	F1-Score	Precision	Recall	ROC-AUC
Dummy - Majority Class	0.69	0.56	0.47	0.69	0.50	0.00	0.00	0.00	0.00	0.00
Dummy - Random Guessing	0.54	0.55	0.55	0.54	0.51	0.03	0.02	0.02	0.03	0.03
Decision Tree	0.73	0.66	0.70	0.73	0.63	0.05	0.08	0.08	0.05	0.08
Random Forest	0.78	0.76	0.78	0.78	0.78	0.02	0.02	0.02	0.02	0.02
Logistic Regression	0.79	0.78	0.78	0.78	0.79	0.02	0.03	0.03	0.02	0.04
Support Vector Machine	0.79	0.78	0.79	0.79	0.73	0.02	0.02	0.02	0.02	0.02
Neural Network	0.72	0.72	0.72	0.72	0.68	0.03	0.03	0.02	0.03	0.02
GaussianNB	0.78	0.76	0.78	0.78	0.77	0.02	0.02	0.02	0.02	0.03
LightGBM	0.83	0.81	0.85	0.83	0.80	0.01	0.02	0.02	0.01	0.04
Ensemble (Stacking)	0.83	0.81	0.84	0.83	0.82	0.02	0.03	0.01	0.02	0.03
Ensemble (Voting)	0.83	0.82	0.83	0.83	0.82	0.02	0.02	0.02	0.02	0.02

### 5. Predictive performance evaluation using nested cross-validation

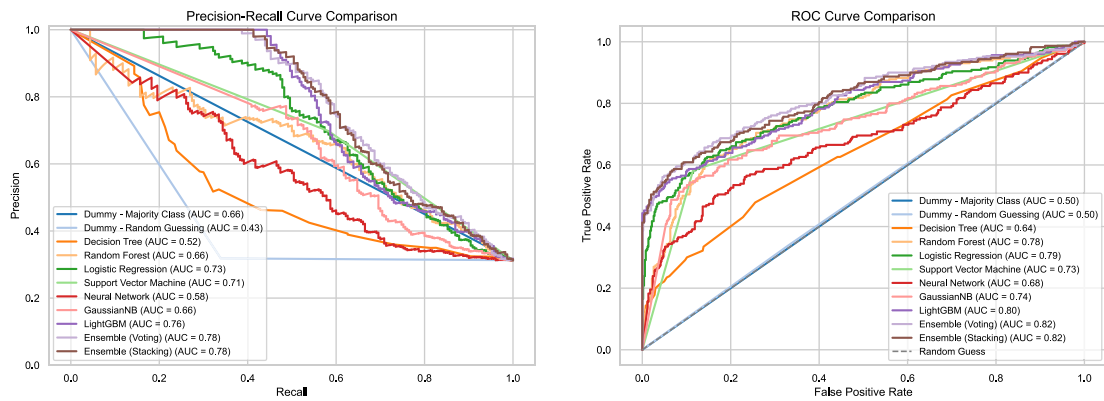
This section presents the results from our performance evaluation of the classifiers described in Section 3.3 using nested cross-validation (3.4), which classify QC operations during ship handling as either prone to breakdowns or not. This enables a systematic comparison and benchmarking of classifier performance across a range of evaluation metrics (Section 5.1). In addition, Section 5.2 provides insights into how the classifiers scale with additional data and how well they generalize.

#### 5.1. Performance comparison across classifiers

Two baseline dummy classifiers are used to evaluate the performance of our trained models, the majority class classifier, which always chooses the most common class, i.e. no breakdown in the case of this study, and the random classifier, which randomly chooses whether input features lead to disruption or not in our binary classification setting.

The comparison of the seven selected classifiers in Table 4, ranging from simple to more complex models, shows that all methods are able to capture meaningful patterns in the data, as they consistently outperform the baseline models (majority class and random guessing) across all evaluation metrics.

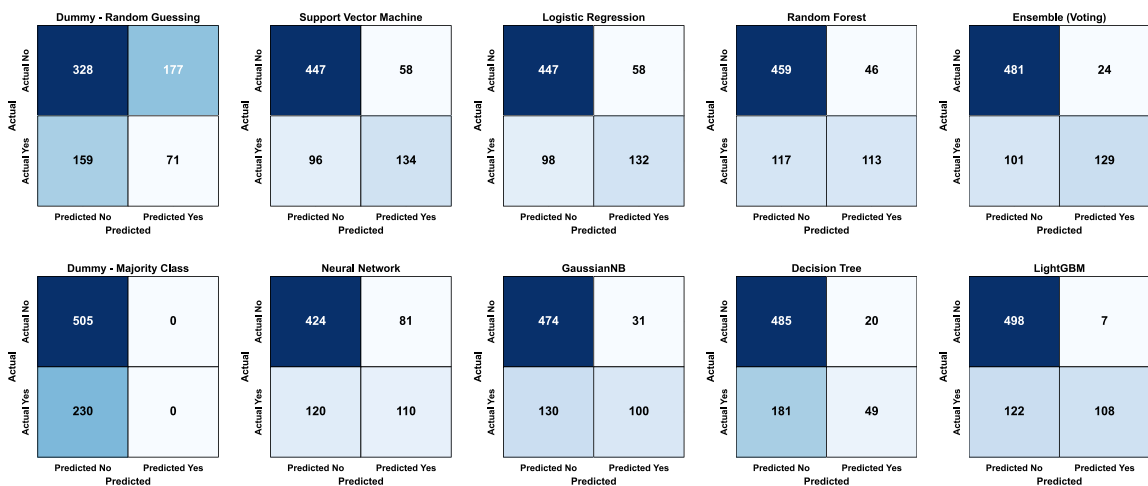
Among the individual classifiers, LightGBM achieves the best overall performance, with a mean accuracy of 83% across folds, which motivates its use in the previous XAI-based feature contribution analysis (Section 4). Logistic Regression is the second-best performing model, achieving an accuracy of 79%, while SVM and Random Forest also demonstrate competitive performance, with accuracies close to 80%.



(a) Precision-Recall (PR) curve comparison.

(b) Receiver Operating Characteristic (ROC) curve comparison.

**Fig. 7.** Performance evaluation across the utilized classifiers using PR and ROC curves.



**Fig. 8.** Confusion matrices aggregated across outer folds for baseline models, classifiers, and a soft voting ensemble evaluated using nested 5-fold cross-validation.

In addition to the individual models, ensemble approaches combining LightGBM, Random Forest, and Logistic Regression provide marginal improvements in predictive performance, particularly in terms of ROC–AUC (for both stacking and voting) and F1-score (for stacking).

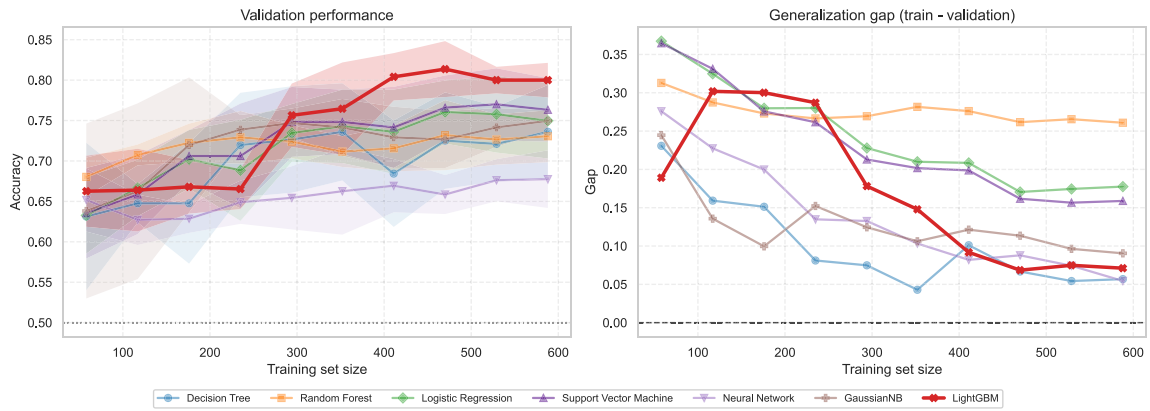
The visual comparison of the seven different ML classifiers and the two dummy classifiers in Fig. 7(a) reveals that LightGBM and Logistic Regression have the highest performance, as they are predominantly represented in the upper right corner of the plot. The larger area under the curve indicates both high recall and high precision, which is also reflected in the higher AUC values of these classifiers. High precision is achieved by minimizing false positives (i.e., predicting no breakdown when there is one), while high recall is achieved by minimizing false negatives (i.e., predicting a breakdown when there is none).

The superior performance of LightGBM and Logistic Regression is also evident in Fig. 7(b), as their ROC curves are closest to the top-left corner, indicating that these models best distinguish between the positive (breakdown) and negative (non-breakdown) classes. This is further supported by the AUC–ROC values of these classifiers, where higher values indicate better overall classification performance. Combining these models using soft voting and stacking ensembles further improves predictive performance.

The comparatively weaker performance of GaussianNB and Random Forest in Fig. 7(a) can be attributed to their limitations in probability estimation under feature dependence and class imbalance, whereas LightGBM and Logistic Regression provide better-calibrated probabilities, which is particularly important for precision–recall evaluation.

The comparison of the confusion matrices in Fig. 8 and Table 4 shows that all of our seven classifiers have better performance compared to our two baseline classifiers. However, it is also noticeable that some of the models, especially the Decision Tree model, predict most disruption-free operations as such, but tend to predict the absence of disruptions in cases where they are not, which is also known as a type 2 error, falsely indicating that no breakdown will occur.

The comparison among the classifiers further reveals that Logistic Regression and SVM are the most capable of detecting breakdowns, whereas LightGBM, which has the highest predictive accuracy across classes, has a very low type 1 error, corresponding



**Fig. 9.** The left panel shows validation performance as a function of training set size, with shaded regions indicating  $\pm$  standard deviation across 5-fold cross-validation. The right panel shows the generalization gap (training minus validation performance), where larger values indicate stronger overfitting.

to few false positive breakdown predictions. The Voting ensemble represents a favorable trade-off, maintaining relatively low rates of both type 1 and type 2 errors.

### 5.2. Learning curves

Learning curves illustrate the evolution of model performance as a function of training data size. They are constructed using the best-performing model configurations identified through nested cross-validation and subsequently retrained on the full dataset. This setup enables an assessment of how predictive performance scales with increasing data while keeping the model configuration fixed. An overview of model performance is presented in Fig. 9. Solid lines denote validation performance, while dashed lines represent training performance. Shaded regions indicate  $\pm$  standard deviation across 5-fold cross-validation, providing an estimate of variability across data splits.

The performance of the retrained models is slightly lower than the nested cross-validation estimates. This is expected, as nested cross-validation optimizes hyperparameters independently within each training fold, allowing models to adapt to fold-specific data characteristics and yielding mildly optimistic performance estimates. In contrast, the final retrained model uses a single hyperparameter configuration selected based on average cross-validation performance, which may not be globally optimal and therefore results in slightly reduced performance.

Despite this difference, the relative ranking of models remains stable. LightGBM achieves the strongest generalization performance, with a validation accuracy of 0.80 and a small generalization gap of 0.07, indicating a favorable bias-variance trade-off. In contrast, Random Forest exhibits pronounced overfitting, achieving near-perfect training accuracy (0.99) but a substantially lower validation accuracy (0.74), resulting in a generalization gap of 0.26. Logistic Regression and SVM show similar patterns, with training accuracies above 0.92 but larger gaps (0.18 and 0.16, respectively), indicating moderate overfitting. Simpler models, such as Decision Tree, GaussianNB, as well as the Neural Network, exhibit smaller gaps, but have lower overall predictive performance.

Overall, these results confirm the robustness of the nested cross-validation procedure, which provides reliable and low-variance performance estimates across data splits. The learning curves further indicate that LightGBM offers the most favorable trade-off between predictive accuracy and generalization, while also suggesting that additional training data may further improve performance.

These findings are consistent with the limited dataset size (735 samples, including 230 breakdown cases), which increases the variance of high-capacity models and exacerbates overfitting effects, as observed for Random Forest. In contrast, gradient-boosted models such as LightGBM are better suited to this data regime due to their inherent regularization mechanisms, enabling more stable generalization under limited data conditions.

## 6. Discussion

In recent years, to keep pace with the rapid growth of container transportation, major ports worldwide have deployed large, highly computerized QCs. These cranes are designed to maximize handling efficiency, reduce vessel turnaround times, and meet increasing throughput demands (Tao et al., 2024). As the primary interface for ship-to-shore interface, QCs are indispensable to terminal performance. However, due to their high operational intensity, mechanical complexity, and increasing levels of computerization and automation, QCs are also among the most error-prone assets in port environments (Klar et al., 2023). Ensuring disruption-free QC operations is therefore essential for maintaining port productivity, achieving environmental targets, and enhancing sustainability (Lim et al., 2019).

This study addresses this challenge by integrating QC monitoring data, terminal operations data, and historical weather observations to predict QC breakdowns using ML. By combining these heterogeneous data sources, we explore the factors that increase downtime risk and assess the predictive potential of different classifiers.

### 6.1. Assessment of key findings

Our findings, derived from SHAP-based feature importance analysis using the best-performing classifier, LightGBM, and supported by global and local SHAP explanations, reveal clear and consistent patterns characterizing the drivers of QC breakdowns. Together, these results demonstrate that XAI provides a rigorous and transparent analytical framework for identifying the operational, mechanical, and environmental conditions under which disruptions are most likely to occur.

The SHAP analysis identifies three dominant categories of features that consistently contribute to elevated downtime risk:

1. **Operational intensity:** Higher workload levels, reflected in longer discharge and load durations, increased move counts, and disproportionately high QC time and move shares, are strongly associated with increased downtime risk. These patterns suggest that sustained operational pressure and uneven workload distribution may accelerate mechanical wear and increase system strain.
2. **Control-system and hoist-related warnings:** System-generated events, including hoist overload conditions, excessive load differentials, trolley alignment warnings, and safety-related gantry stops, emerge as strong predictors of downtime. These indicators likely reflect underlying mechanical stress, sensor misalignment, or control-system instability, and may serve as early warning signals preceding forced operational stoppages.
3. **Environmental stressors:** Although environmental conditions are less impactful than operational intensity, factors such as elevated wind speeds, low temperatures, and low humidity levels appear to further increase the risk of downtime. These factors can affect crane stability, mechanical stress, and electronic reliability, exacerbating existing operational and mechanical vulnerabilities.

Taken together, these findings suggest a coherent explanatory framework in which QC downtime risk emerges from the interaction between operational workload, control-system stability, and environmental exposure. This integrated understanding provides a strong foundation for predictive maintenance and operational optimization, supporting proactive workload balancing, early detection of control-system anomalies, and adaptive operational planning under adverse environmental conditions.

These findings align closely with earlier research linking heavy loads and high-speed operations to mechanical degradation (Gothandapani et al., 2024; Crespo Del Castillo et al., 2024; Awasthi et al., 2024; Jalal et al., 2023; Mukherjee et al., 2024). Similarly, Crespo Del Castillo et al. (2024) highlights harsh weather as a contributing factor to breakdowns, which aligns with our SHAP analysis identifying weather conditions as relevant features.

At the same time, several factors highlighted in earlier studies, such as maintenance delays (Crespo Del Castillo et al., 2024), age-related degradation (Gothandapani et al., 2024), and human operator errors (Awasthi et al., 2024), are beyond the scope of the present dataset, but remain important considerations for comprehensive maintenance planning.

Our work thus complements existing literature on QC maintenance, which primarily addresses (post-) disruption scheduling aimed at maintaining operational efficiency and asset health monitoring. Building on our findings, we discuss proactive mitigation strategies and directions for future research to reduce the risk of QC downtime, improve operational efficiency, and lower emissions from port operations.

### 6.2. Proactive mitigation strategies

Understanding the factors associated with QC breakdowns enables targeted preventive interventions. As vessel turnaround time directly influences terminal efficiency, operational costs, and emissions (Iris and Lam, 2019), minimizing unplanned downtime is critical for both economic and environmental performance. QC breakdowns disrupt the entire port logistics chain, leading to extended vessel port stays, increased cycle times, congestion, and idling of ships and equipment, all of which contribute to elevated fuel consumption and emissions.

Based on the interpretability-driven local and global SHAP analysis, several actionable mitigation strategies are identified:

- **Early detection of critical system events through data-driven monitoring:** The SHAP analysis highlights the predictive importance of hoist-related faults, including overload conditions and differential load warnings, as well as trolley-related warnings such as misalignment, and safety-related gantry stop warnings. These indicators likely reflect developing mechanical stress or control system degradation. Continuous real-time monitoring of these signals can enable early intervention, preventing fault escalation and reducing the likelihood of unplanned downtime.
- **Environment-aware maintenance and operational planning:** Environmental factors, particularly low temperatures, low humidity, and high wind speeds, are mildly associated with an increased risk of downtime. These conditions can accelerate component degradation, affect sensor performance, and increase mechanical stress. Therefore, maintenance planning and inspection schedules should account for environmental exposure, especially during periods of intensive operation or severe weather conditions.

- **Load- and utilization-aware preventive maintenance and operational balancing:** Operational workload characteristics emerge as dominant predictors of downtime risk. These characteristics include the total number of QC moves, load and discharge durations, and the relative workload distribution across cranes, as reflected by QC time share and QC move share. Elevated move counts and prolonged load and discharge operations are associated with increased mechanical stress and component wear. Disproportionately high time and move shares for individual cranes, indicating workload imbalance across available QCs, are strongly associated with increased downtime risk. This suggests that uneven workload allocation can accelerate localized component degradation and increase the likelihood of failure. Therefore, preventive maintenance strategies should incorporate workload-based triggers that account for cumulative operational exposure and relative crane utilization. Additionally, operational planning should aim to distribute container moves and operational time evenly across cranes to reduce mechanical strain.

These findings suggest that QC downtime risk is driven by the interaction between operational workload, control-system stability, and environmental stressors. Effective mitigation strategies should therefore integrate these dimensions into maintenance planning and operational decision-making. While preventive maintenance interventions may introduce minor short-term operational interruptions, they are likely to significantly reduce the frequency and impact of unplanned breakdowns, thereby improving overall terminal efficiency, reducing energy consumption, and minimizing emissions.

### 6.3. Limitations and future research

Future research should focus on assessing the generalizability of these findings by expanding the dataset temporally and spatially. The current sample of 735 QC vessel handling instances, derived from 347 vessel calls, which represents two years of data from a mid-sized Swedish port, limits the model's generalizability. Enhanced studies with data from multiple port container terminals representing ports with different throughputs and differing environmental conditions would yield more generalizable findings.

This study integrates terminal operational records, QC monitoring data, and weather conditions to provide a comprehensive representation of the immediate operational environment. Several variables, such as vessel operation duration and the number of assigned cranes, implicitly capture aspects of terminal workload and resource allocation.

However, we acknowledge that additional terminal-level factors, such as the degree of port automation, workforce availability, and overall cross-terminal operational workload, may further influence QC utilization and maintenance conditions, thereby contributing to breakdown occurrences. Future work could enhance the explanatory power of data-driven predictive maintenance models by extending the proposed framework to incorporate these factors.

A limitation of the proposed approach is that features and target are derived from the same operational time window. As a result, the model should not be interpreted as a real-time predictive system for anticipating breakdowns prior to their occurrence. Instead, it provides insights into operational conditions associated with breakdown-prone vessel calls and enables the identification of key risk factors at the operational level. Despite this limitation, the results demonstrate that breakdown occurrences are systematically related to observable operational, monitoring, and environmental conditions, supporting the validity of the proposed modeling framework. Future work could extend this approach by introducing a temporal separation between features and target, enabling true, forward-looking prediction of breakdown events.

An additional limitation of the proposed approach is that multiple observations may originate from the same vessel call, as each assigned QC is represented as a separate sample. Since nested cross-validation is applied at the observation level, this may introduce dependencies between training and test sets, potentially leading to optimistic performance estimates. However, monitoring events and breakdowns are recorded at the individual crane level, and empirical observations indicate that breakdowns are often associated with uneven workload distribution across cranes (see Fig. 5). This supports the QC-level modeling approach, while capturing shared operational context through aggregated features. Future work could address this limitation by applying group-based cross-validation at the vessel-call level to explicitly account for dependencies between observations.

Another key limitation is the comprehensive feature matrix, which includes numerous QC monitoring events, warnings, terminal operating conditions, and weather conditions. Thus, it encompasses 1007 potential features. While this variety is valuable, it introduces challenges related to sparsity, multicollinearity, and interpretability. Future work should explore automated grouping of semantically related event types or embedding-based representations to streamline the feature space further.

An important direction for future research is treating novel fault types not captured in historical data. These faults pose significant challenges for predictive modeling and have substantial implications for port operations and maintenance planning.

The integration of terminal operations data, weather data, and QC monitoring data in this study reflects many of the port's operational characteristics and events. This further motivates the case for developing port digital twins, which can accumulate historical data to identify breakdown patterns and detect anomalies in real time. A digital twin can enhance situational awareness by continuously analyzing QC monitoring events, terminal operations, and weather conditions, enabling data-driven decision-making (Klar et al., 2023). This approach has the potential to elevate maintenance strategies from proactive to truly predictive.

To realize this potential, a digital twin would need to reach at least maturity level 3 (synchronization), or level 5 if autonomous interactions with QCs and other equipment are desired (Klar et al., 2024b). The LightGBM classifier used fits well with the real-time aspect of digital twins due to its fast training and prediction speeds, making it suitable for real-time predictive maintenance applications (Bagci Das, 2025).

Real-time digital twin-driven analytics also aligns with emerging smart port architectures, where edge computing and CPS enable predictive models to operate close to the equipment (Min, 2022). Processing data locally at the edge, closer to sensors and equipment,

reduces latency (Ahmadvand and Foroutan, 2025), enabling faster responses to anomalies and allowing for proactive reactions to mitigate the risk of QC downtime. CPS technologies further support this by integrating sensors, machine-to-machine communication, and automated control, facilitating autonomous decision-making in dynamic port environments (Min, 2022).

Another promising direction for future research is to quantify the potential cost and emissions savings enabled by proactive maintenance actions. For example, estimating CO<sub>2</sub> savings per hour of avoided downtime or calculating total cost savings per hour of disruption could provide valuable insights for port operators and policymakers. Such metrics would help justify investments in predictive maintenance systems and support sustainability goals in maritime logistics.

## 7. Conclusion

This study makes two key contributions to understanding and predicting QC breakdowns. First, it provides a comprehensive XAI analysis combining global SHAP feature attributions and local, instance-level explanations. This analysis uncovers how operational, mechanical, and environmental conditions shape downtime risk. This interpretability-driven component provides transparent, domain-specific insights into the mechanisms behind QC breakdowns.

Second, the study introduces a thorough ML pipeline that evaluates and compares multiple classifiers using nested cross-validation. This rigorous evaluation identifies LightGBM as the most robust model under the examined conditions. This best-performing classifier is thus used to generate global and local SHAP explanations. Together, these contributions form an integrated, data-driven framework for understanding and mitigating QC disruptions in terminal operations.

The results provide terminal operators and maintenance planners with actionable insights. By identifying high-risk operational conditions, such as sustained QC workload, uneven crane utilization, hoist overload events, and adverse weather conditions, this study supports the development of targeted preventive maintenance strategies. These strategies include early-warning triggers based on monitoring signals, interventions to balance workloads, and weather-aware operational planning. Implementing these proactive measures can reduce unplanned stoppages, lower operational costs, and decrease emissions associated with extended vessel turnaround times and idling equipment.

Despite these promising findings, the study is constrained by the size and scope of the dataset, which covers two years of crane-level operational records from a single medium-sized port. This limits the generalizability of the results to ports with different environmental conditions, or operational practices. Additionally, the analysis does not explicitly incorporate factors such as maintenance crew availability, vessel scheduling constraints, or human operator behavior, which may influence downtime patterns but fall outside the scope of the current dataset.

Future research should explore larger and more diverse datasets across multiple terminals to improve model robustness and generalizability. Incorporating temporal dynamics, such as time-series modeling of error events and environmental conditions, could enhance predictive accuracy. Further research could also integrate real-time sensor data, maintenance logs, and operational schedules to develop adaptive models for live decision support. Finally, quantifying the economic and environmental benefits of predictive maintenance would provide a more comprehensive assessment of the value of XAI-driven maintenance strategies in sustainable port operations.

## CRedit authorship contribution statement

**Robert Klar:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Anders Andersson:** Writing – review & editing, Supervision, Conceptualization. **Vangelis Angelakis:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Robert Klar reports financial support was provided by Swedish Transport Administration. Anders Andersson reports financial support was provided by Swedish Transport Administration. Vangelis Angelakis reports financial support was provided by Swedish Transport Administration. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work has been supported by Trafikverket Sweden as part of the Triple F (MODIG-TEK) project under Grant 2019.2.2.16. We would further like to thank Yilport Gävle for providing us with operational data and insights regarding their quay crane maintenance routines.

## References

- Abdulrashid, I., Zanjirani Farahani, R., Mammadov, S., Khalafalla, M., Chiang, W.-C., 2024. Explainable artificial intelligence in transport logistics: Risk analysis for road accidents. *Transp. Res. Part E: Logist. Transp. Rev.* 186, 103563. <http://dx.doi.org/10.1016/j.tre.2024.103563>.
- Abou Kasm, O., Diabat, A., 2020. Next-generation quay crane scheduling. *Transp. Res. Part C: Emerg. Technol.* 114, 694–715. <http://dx.doi.org/10.1016/j.trc.2020.02.015>.
- Ahmadvand, H., Foroutan, F., 2025. Latency and privacy-aware resource allocation in vehicular edge computing. arXiv preprint arXiv:2501.02804.
- Awasthi, A., Krpalkova, L., Walsh, J., 2024. Deep learning-based boolean, time series, error detection, and predictive analysis in container crane operations. *Algorithms* 17 (8), 333.
- Bagci Das, D., 2025. Real-time adaptable fault analysis of rotating machines based on Marine predator algorithm optimised LightGBM approach. *Nondestruct. Test. Eval.* 40 (3), 831–866.
- Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., Lindauer, M., 2023. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Min. Knowl. Discov.* 13 (2), 1484. <http://dx.doi.org/10.1002/widm.1484>.
- Cahyono, R.T., Kenaka, S.P., Jayawardhana, B., 2022. Simultaneous allocation and scheduling of quay cranes, yard cranes, and trucks in dynamical integrated container terminal operations. *IEEE Trans. Intell. Transp. Syst.* 23 (7), 8564–8578. <http://dx.doi.org/10.1109/ITITS.2021.3083598>.
- Carlo, H.J., Vis, I.F.A., Roodbergen, K.J., 2015. Seaside operations in container terminals: literature overview, trends, and research directions. *Flex. Serv. Manuf. J.* 27 (2–3), 224–262. <http://dx.doi.org/10.1007/s10696-013-9178-3>.
- Chicco, D., Sichenze, A., Jurman, G., 2025. A simple guide to the use of Student's t-test, Mann-Whitney U test, Chi-squared test, and Kruskal-Wallis test in biostatistics. *BioData Min.* 18 (1), 56.
- Crespo Del Castillo, A., Sasidharan, M., Nentwich, C., Merino, J., Kumar Parlikad, A., 2024. Data-Driven Asset Health Index – an application to evaluate Quay Cranes in container ports. *Marit. Policy & Manag.* 51 (8), 1805–1823. <http://dx.doi.org/10.1080/03088839.2023.2231449>.
- Cummins, L., Sommers, A., Ramezani, S.B., Mittal, S., Jabour, J., Seale, M., Rahimi, S., 2024. Explainable predictive maintenance: A survey of current methods, challenges and opportunities. *IEEE Access* 12, 57574–57602.
- Dereci, U., Tuzkaya, G., 2024. An explainable artificial intelligence model for predictive maintenance and spare parts optimization. *Supply Chain Anal.* 8, 100078.
- Dragović, B., Dragović, A., Dulebenets, M.A., 2025. The quay crane operation problem at marine container terminals: bibliometric analysis, emerging trends, and future research opportunities. *Transp. Res. Part E: Logist. Transp. Rev.* 201, 104266.
- Filom, S., Amiri, A.M., Razavi, S., 2022. Applications of machine learning methods in port operations – a systematic literature review. *Transp. Res. Part E: Logist. Transp. Rev.* 161, 102722. <http://dx.doi.org/10.1016/j.tre.2022.102722>.
- Fonti, V., Belitser, E., 2017. Feature Selection using LASSO. *VU Amst. Res. Pap. Bus. Anal.* (30), 1–25.
- Ghosh, I., De, A., 2024. Maritime fuel price prediction of European ports using least square boosting and facebook prophet: Additional insights from explainable artificial intelligence. *Transp. Res. Part E: Logist. Transp. Rev.* 189, 103686.
- Gothandapani, S.P.A., Ab Rahman, M.N., Hishamuddin, H., 2024. Quay crane performance improvement and lifecycle extension: Retrofit determination—a case study. *J. Kejuruter.* 36 (4), 1483–1493.
- Huang, M., Liu, Z., Tao, Y., 2020. Mechanical fault diagnosis and prediction in IoT based on multi-source sensing data fusion. *Simul. Model. Pr. Theory* 102, 101981. <http://dx.doi.org/10.1016/j.simpat.2019.101981>.
- Iris, Ç., Lam, J.S.L., 2019. A review of energy efficiency in ports: Operational strategies, technologies and energy management systems. *Renew. Sustain. Energy Rev.* 112, 170–182. <http://dx.doi.org/10.1016/j.rser.2019.04.069>.
- Jalal, M.R., Kader, A.S.A., Hamid, M.F.A., Kang, H.S., 2023. A stochastic Petri Net-based approach for operational performance estimation of quay cranes. *Qual. Reliab. Eng. Int.* 39 (5), 1660–1680. <http://dx.doi.org/10.1002/qre.3272>.
- Jbair, M., Ahmad, B., Maple, C., Harrison, R., 2022. Threat modelling for industrial cyber physical systems in the era of smart manufacturing. *Comput. Ind.* 137, 103611. <http://dx.doi.org/10.1016/j.compind.2022.103611>.
- Kapoor, S., Cantrell, E.M., Peng, K., Pham, T.H., Bail, C.A., Gundersen, O.E., Hofman, J.M., Hullman, J., Lones, M.A., Malik, M.M., Nanayakkara, P., Poldrack, R.A., Raji, I.D., Roberts, M., Salganik, M.J., Serra-Garcia, M., Stewart, B.M., Vandewiele, G., Narayanan, A., 2024. REFORMS: Consensus-based recommendations for machine-learning-based science. *Sci. Adv.* 10 (18), 3452. <http://dx.doi.org/10.1126/sciadv.adk3452>.
- Kashpruk, N., Piskor-Ignatowicz, C., Baranowski, J., 2023. Time series prediction in industry 4.0: A comprehensive review and prospects for future advancements. *Appl. Sci.* 13 (22), 12374. <http://dx.doi.org/10.3390/app132212374>.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* 30.
- Khalilpoor, S., Kamran, M.A., Babazadeh, R., Kia, R., 2025. Energy-Efficient model for integrated berth allocation and quay crane management. *Transp. Res. Interdiscip. Perspect.* 31, 101429. <http://dx.doi.org/10.1016/j.trip.2025.101429>.
- Kim, K.-H., 2024. Planning and Operation of Container Terminals. Elsevier.
- Kizilay, D., Eliyi, D.T., 2021. A comprehensive review of quay crane scheduling, yard operations and integrations thereof in container terminals. *Flex. Serv. Manuf. J.* 33 (1), 1–42. <http://dx.doi.org/10.1007/s10696-020-09385-5>.
- Klar, R., Andersson, A., Fredriksson, A., Angelakis, V., 2024a. Container Relocation and Retrieval Tradeoffs Minimizing Schedule Deviations and Relocations. *IEEE Open J. Intell. Transp. Syst.* 5, 360–379. <http://dx.doi.org/10.1109/OJITS.2024.3413197>.
- Klar, R., Arvidsson, N., Angelakis, V., 2024b. Digital twins' maturity: The need for interoperability. *IEEE Syst. J.* 18 (1), 713–724. <http://dx.doi.org/10.1109/JSYST.2023.3340422>.
- Klar, R., Fredriksson, A., Angelakis, V., 2023. Digital twins from smart city and supply chain twinning experience. *IEEE Access* 11, 71777–71799. <http://dx.doi.org/10.1109/ACCESS.2023.3295495>.
- Knatz, G., Notteboom, T., Pallis, A.A., 2022. Container terminal automation: revealing distinctive terminal characteristics and operating parameters. *Marit. Econ. Logist.* 24 (3), 537–565. <http://dx.doi.org/10.1057/s41278-022-00240-y>.
- Lam, J.S.L., Su, S., 2015. Disruption risks and mitigation strategies: an analysis of Asian ports. *Marit. Policy & Manag.* 42 (5), 415–435. <http://dx.doi.org/10.1080/03088839.2015.1016560>.
- Lee, Y., Park, K., Lee, H., Son, J., Kim, S., Bae, H., 2024. Identifying key factors influencing import container dwell time using eXplainable Artificial Intelligence. *Marit. Transp. Res.* 7, 100116.
- Li, Y., Chu, F., Zheng, F., Kacem, I., 2019. Integrated berth allocation and quay crane assignment with uncertain maintenance activities. In: 2019 International Conference on Industrial Engineering and Systems Management (IESM). IEEE, Shanghai, China, pp. 01–06. <http://dx.doi.org/10.1109/IESM45758.2019.8948115>.
- Li, Y., Chu, F., Zheng, F., Liu, M., 2022. A bi-objective optimization for integrated berth allocation and quay crane assignment with preventive maintenance activities. *IEEE Trans. Intell. Transp. Syst.* 23 (4), 2938–2955. <http://dx.doi.org/10.1109/ITITS.2020.3023701>.
- Li, T., Dong, Q., Sun, X., 2024. Integrated scheduling of handling equipment in automated container terminal considering quay crane faults. *Systems* 12 (11), 450. <http://dx.doi.org/10.3390/systems12110450>.

- Lim, S., Pettit, S., Abouarghoub, W., Beresford, A., 2019. Port sustainability and performance: A systematic literature review. *Transp. Res. Part D: Transp. Environ.* 72, 47–64. <http://dx.doi.org/10.1016/j.trd.2019.04.009>.
- Liu, D., Ge, Y.-E., 2018. Modeling assignment of quay cranes using queueing theory for minimizing CO<sub>2</sub> emission at a container terminal. *Transp. Res. Part D: Transp. Environ.* 61, 140–151. <http://dx.doi.org/10.1016/j.trd.2017.06.006>.
- Lones, M.A., 2024. Avoiding common machine learning pitfalls. *Patterns* 5 (10), 101046. <http://dx.doi.org/10.1016/j.patter.2024.101046>.
- Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2 (1), 56–67. <http://dx.doi.org/10.1038/s42256-019-0138-9>.
- Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. Curran Associates, Inc., pp. 4765–4774.
- Merkel, A., Kalantari, J., Mubder, A., 2022. Port call optimization and CO<sub>2</sub>-emissions savings—estimating feasible potential in tramp shipping. *Marit. Transp. Res.* 3, 100054.
- Min, H., 2022. Developing a smart port architecture and essential elements in the era of Industry 4.0. *Marit. Econ. Logist.* 24 (2), 189.
- Mukherjee, A., Sasidharan, M., Herrera, M., Parlikad, A.K., 2024. Unsupervised constrained discord detection in IoT-based online crane monitoring. *Adv. Eng. Inform.* 60, 102444.
- Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. *Front. Neurobot.* 7, <http://dx.doi.org/10.3389/fnbot.2013.00021>.
- Nikolaou, P., Dimitriou, L., 2021. Lessons to be learned from top-50 global container port terminals efficiencies: A multi-period DEA-tobit approach. *Marit. Transp. Res.* 2, 100032.
- Notteboom, T., Pallis, A., Rodrigue, J.-P., 2021. *Port Economics, Management and Policy*, first ed. Routledge, London, <http://dx.doi.org/10.4324/9780429318184>.
- Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S., Stodtmann, S., 2024. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin. Transl. Sci.* 17 (11), e70056.
- Putra, B.A., Putranto, L.M., Irnawan, R., 2024. Early failure detection of quayside container crane using the IoT based measurement data. In: 2024 International Conference on Technology and Policy in Energy and Electric Power (ICTPEP). IEEE, Bali, Indonesia, pp. 297–302. <http://dx.doi.org/10.1109/ICT-PEP63827.2024.10733496>, URL <https://ieeexplore.ieee.org/document/10733496/>.
- Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* 2 (3), 160. <http://dx.doi.org/10.1007/s42979-021-00592-x>.
- Shaikh, T.A., Rasool, T., Verma, P., Mir, W.A., 2024. A fundamental overview of ensemble deep learning models and applications: systematic literature and state of the art. *Ann. Oper. Res.* 1–77.
- Swedish Meteorological and Hydrological Institute (SMHI), 2025. Weather observation data. URL <https://www.smhi.se/data/hitta-data-for-en-plats/ladda-ner-vaerobeservationer/airtemperatureInstant>, (Accessed: June. 06, 2025).
- Tao, D., Yan, Y., Dong, D., Zhang, D. (Eds.), 2024. *Handbook of Port Machinery*. Springer Nature Singapore, Singapore, <http://dx.doi.org/10.1007/978-981-99-4848-2>.
- Tharwat, A., 2021. Classification assessment methods. *Appl. Comput. Inform.* 17 (1), 168–192. <http://dx.doi.org/10.1016/j.aci.2018.08.003>.
- Vuttipittayamongkol, P., Arreeras, T., 2022. Data-driven industrial machine failure detection in imbalanced environments. In: 2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, Kuala Lumpur, Malaysia, pp. 1224–1227. <http://dx.doi.org/10.1109/IEEM55944.2022.9989673>.
- Wang, T., Gao, G., Wang, K., Shi, J., 2024. Integrated operation models with quay crane maintenance in a container terminal. *Ocean & Coastal Management* 251, 107101. <http://dx.doi.org/10.1016/j.ocecoaman.2024.107101>.
- Xu, L.-S., Huang, T., Zhao, B.-W., Gong, Y.-J., Liu, J., 2025. Continuous berth allocation and time-variant quay crane assignment: Memetic algorithm with a heuristic decoding method. *IEEE Trans. Intell. Transp. Syst.* 26 (3), 3387–3401. <http://dx.doi.org/10.1109/TITS.2024.3517879>.
- Yates, L.A., Aandahl, Z., Richards, S.A., Brook, B.W., 2023. Cross validation for model selection: A review with examples from ecology. *Ecol. Monograph.* 93 (1), 1557. <http://dx.doi.org/10.1002/ecm.1557>.
- Zhou, Z.-H., 2021. *Machine Learning*. Springer, Singapore, <http://dx.doi.org/10.1007/978-981-15-1967-3>.

## Glossary

### Glossary of Abbreviations

*Abbreviation:* Definition

AGV: Automated Guided Vehicle

AI: Artificial Intelligence

CPS: Cyber-Physical System

CS: Container Ship

CT: Container Terminal

CY: Container Yard

FN: False Negative

FP: False Positive

GBM: Gradient Boosting Machine

GaussianNB: Gaussian Naive Bayes

IoT: Internet of Things

KPI: Key Performance Indicator

LightGBM: Light Gradient Boosting Machine

ML: Machine Learning

PR Curve: Precision-Recall Curve

QC: Quay Crane

ROC-AUC: Receiver Operating Characteristic-Area Under Curve

SHAP: SHapley Additive exPlanations

SMHI: Swedish Meteorological and Hydrological Institute

SVM: Support Vector Machine

TEU: Twenty-Foot Equivalent Unit

TN: True Negative

TP: True Positive

XAI: Explainable Artificial Intelligence

YC: Yard Crane