IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 10, October, 2026, pp. 2785-2795
Journal homepage: http://iieta.org/journals/jesa

# Leakage-Controlled Machine Learning for Forecasting Vessel Turnaround Time at Tanger Med Port

Amine Bouhlal[1*], Rachida Ait Abdelouahid[1], Abdelaziz Marzak[2], Olivier Debauche[3]

[1] LIAS, Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Casablanca 20670, Morocco
[2] LTIM, Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Casablanca 20670, Morocco
[3] DEAL, Gembloux Agro-Bio Tech, University of Liège, Gembloux 5030, Belgium

Corresponding Author Email: am.bouhlal@gmail.com

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/jesa.591003

**Received:** 14 July 2026
**Revised:** 5 October 2026
**Accepted:** 23 October 2026
**Available online:** 31 October 2026

***Keywords:***
*feature engineering, information leakage, port operations forecasting, tabular machine learning, Tanger Med, vessel turnaround time*

**ABSTRACT**

Tanger Med, on the Strait of Gibraltar, is a major transshipment port. Forecasting vessel turnaround time (VTT), the interval between port entry and departure, supports berth planning and congestion management. We present a reproducible tabular machine learning pipeline built on seven years of harmonized Vessel Situation Report (VSR) exports (2019-2025; 96,901 visits). Twenty predictors cover vessel particulars, calendar effects, operational context, and availability-time historical indicators computed only from calls already departed at prediction time. Eleven regression models were compared under a temporal split (train 2019-2022, validation 2023, test 2024-2025), with model selection on validation data only. CatBoost achieved the best validation mean absolute error (MAE = 3.63 h) and, on the held-out test years, MAE = 4.62 h, root mean squared error (RMSE) = 11.06 h and coefficient of determination ($R^2$) = 0.41 for stays up to 60 days (all calls: MAE = 5.05 h, $R^2$ = 0.06). An ablation study shows that the historical indicators raise test $R^2$ from 0.30 to 0.41, and SHapley Additive exPlanations (SHAP) computed on the final model confirm them as the strongest predictors. Errors are lowest for ferries (MAE = 0.71 h) and highest for bulk carriers (MAE = 26.76 h, n = 27).

## 1. INTRODUCTION

Tanger Med, located on the Strait of Gibraltar in northern Morocco, is one of the busiest container ports in Africa [1] and a major transshipment hub linking Europe, Africa and the Americas. Its operational efficiency has a direct impact on regional and global supply chains. A central issue at any high-throughput port is to predict the vessel turnaround time (VTT), from port entry to port departure. A reliable VTT prediction supports several operational decisions, including berth allocation, rostering of pilots, tugs and stevedoring gangs, congestion management, and the estimated arrival of inbound vessels waiting at anchorage. When VTT is under- or over-estimated, vessels wait idle at anchorage, effective port capacity is reduced, and demurrage costs increase. Port-stay and berth-occupancy forecasting has an established operations-research tradition based on deterministic and stochastic scheduling models, surveyed by Bierwirth and Meisel [2] and by Steenken et al. [3]. These models generally assume that service-time distributions are known or estimated from aggregate statistics. The complementary, data-driven task addressed in this paper is to predict the service time itself (VTT) from vessel- and call-level operational data, which is a key input to these scheduling models.

Two difficulties arise in operational forecasting of this kind. The first is that port operations data are not directly machine-readable: they are exported as spreadsheets whose headers, column names, and categorical vocabularies change from year to year, so that a data harmonization problem must be solved before any model can be trained. The second, more subtle difficulty is temporal information leakage in historical aggregate features. A common shortcut for building "historical efficiency" features is to sort calls by arrival time and take an expanding or rolling mean of the target with a simple index shift. However, a vessel that arrived before the current call may not yet have departed when the current vessel arrives; its VTT is then not observable at prediction time, and using it introduces information that would not exist in a live deployment. This is a specific case of the leakage risks documented in the forecasting-evaluation literature [4].

This paper makes four contributions: (1) a leakage-controlled VTT forecasting study for Tanger Med based on seven years of real operational data; (2) an availability-time construction of historical efficiency features that excludes any call whose outcome was not observable at prediction time; (3) a leakage-controlled evaluation protocol in which models are selected on a validation year distinct from the training and test years; and (4) a comparison of eleven regression algorithms (linear, instance-based, kernel and tree-ensemble) under this protocol, together with an interpretability analysis and an operational prediction module. The remainder of the paper is organized as follows. Section 2 reviews related work on port-stay prediction and tabular machine learning. Section 3 describes the data sources and their harmonization, the target

definition, feature engineering, a possible real-time extension based on the Automatic Identification System (AIS), and the modeling protocol. Section 4 presents descriptive statistics, the model comparison, the per-vessel-type error analysis, the ablation study, the comparison on equal training subsamples, and the temporal drift analysis. Section 5 examines feature importance with SHapley Additive exPlanations (SHAP). Section 6 describes the operational prediction module. Section 7 discusses the results and the limitations of the study, and Section 8 concludes.

## 2. RELATED WORK

### 2.1 Berth allocation and port service-time modeling

There is a large body of literature on the operations research of container terminals. Steenken et al. [3] classify and review operations research in container terminal operations, and Bierwirth and Meisel [2] survey berth allocation and quay crane scheduling models in which vessel service time is treated as an exogenous parameter rather than as the output of a prediction model. Yazar Okur and Tuna [5] study the schedule reliability of global shipping lines and show that reliability, in addition to average transit time, matters for carrier performance and shippers' decisions. This view is in line with our forecasting approach, which reports per-class errors in addition to overall accuracy. Independent research on port cities of the Strait of Gibraltar area also identifies Tanger Med as a major container terminal in Africa [1].

A second empirical stream predicts vessel arrival times and port stay directly from operational data. Early work at the port of Cagliari trained artificial neural networks (ANN) to predict vessel arrival times for workforce planning [6], and later used classification trees to predict late arrivals [7]. At the Port of Rotterdam, support vector machines and neural networks based on AIS and weather data were used to predict the estimated time of arrival (ETA) of containerships up to several days ahead [8]. AIS-based ETA estimation was also developed for German port calls [9] and for real-time Mediterranean traffic streams [10]. More recently, machine-learned arrival-time predictions were used for robust berth scheduling at a European port [11], and deep learning ETA models were proposed for a Moroccan bulk port [12]. A systematic review of machine learning in port operations [13] covers these applications; studies that directly predict the total VTT from operational data remain rare. Table 1 positions our study within this literature.

**Table 1.** Empirical vessel arrival and port-stay prediction studies

| Study | Port/Setting | Data | Model(s) | Target | Reported error |
|---|---|---|---|---|---|
| Fancello et al. [6] | Cagliari, Italy (transshipment) | Terminal call records | ANN | Vessel arrival time | Arrival-time deviation (h) |
| Pani et al. [7] | Cagliari, Italy | Historical call records | Decision trees (CART) | Late-arrival class | Classification accuracy |
| Parolas [8] | Rotterdam, Netherlands | AIS + weather | SVM, ANN | Containership ETA | MAE (h), horizon-dependent |
| Jahn and Scheidweiler [9] | German ports | AIS trajectories | ANN | ETA | ETA deviation |
| Bodunov et al. [10] | Mediterranean AIS stream | Streaming AIS | ML ensemble | Destination + ETA | ETA error (min) |
| El Mekkaoui et al. [12] | Moroccan bulk port | AIS voyage data | Deep learning | Vessel ETA | MAE (h) |
| Kolley et al. [11] | Hamburg, Germany (container) | Port-call + AIS data | KNN, ANN, RF | Arrival time for berth scheduling | Interval-based/MAE |
| This study | Tanger Med, Morocco | 96,901 calls, 2019-2025, 20 tabular features | CatBoost (+10 benchmarks) | Total turnaround time (VTT) | MAE 4.62 h, $R^2$ 0.41 (2024-2025) |

Note: AIS, Automatic Identification System; ANN, artificial neural network; CART, classification and regression trees; ETA, estimated time of arrival; KNN, k-nearest neighbors; MAE, mean absolute error; ML, machine learning; $R^2$, coefficient of determination; RF, Random Forest; SVM, support vector machine; VTT, vessel turnaround time.

The errors reported in Table 1 are not directly comparable between studies: targets differ (ETA deviation, lateness class, total turnaround time), as do horizons, fleets and operating contexts. Our study differs from most previous work in its sample size (96,901 calls over seven years at one port), in its target (total turnaround time rather than arrival time) and in its leakage-controlled methodology (availability-time historical features, strict temporal splitting and evaluation-set integrity), which, to our knowledge, the studies in Table 1 do not report.

### 2.2 Tabular machine learning and tree ensembles

The predictive modeling in this work builds on the tree-ensemble literature. Random forests (RF), reviewed by Scornet and Hooker [14], and Extremely Randomized Trees, which remain competitive in recent applied comparisons [15], reduce the variance of a single decision tree by introducing randomness in split selection, while the gradient-boosted tree implementations XGBoost [16], LightGBM [17] and CatBoost [18] iteratively fit residuals and remain the most widely used methods on structured, heterogeneous tabular data. Natural gradient boosting (NGBoost) [19] extends gradient boosting to probabilistic, distributional outputs instead of point estimates. Grinsztajn et al. [20] show on medium-sized real-world tabular benchmarks that tree-based ensembles still outperform deep learning on this data modality, a pattern also observed in our model comparison. The models are implemented with the scikit-learn library, which is widely used as a basis for large-scale automated machine learning benchmarks [21].

### 2.3 Interpretability and evaluation methodology

Tree ensembles can be interpreted post hoc with permutation importance or with SHAP, introduced by Lundberg and Lee [22] as a game-theoretic attribution

framework with an efficient exact algorithm for tree-based models. In this study, SHAP values are computed directly on the final model (Section 5). On evaluation, Cerqueira et al. [4] compare performance-estimation methods for time series and show that naive cross-validation overstates accuracy relative to out-of-sample, chronologically ordered evaluation; this motivates both the temporal train/validation/test protocol and the availability-time feature design adopted here. The first version of this study filtered outliers with the boxplot fence rule [23]; Section 3.2 explains why it was replaced by a fixed scope rule.

### 2.4 Positioning of this study

Relative to this literature, our study combines (i) a multi-year, harmonized, real operational dataset from a major North African transshipment hub, (ii) an explicit and verifiable leakage-control mechanism for historical aggregate features at the level of individual timestamps rather than arrival order, and (iii) a broad comparison of eleven regression models under a temporally disjoint validation and test protocol, with model selection locked before the test set is scored.

## 3. MATERIALS AND METHODS

Figure 1 summarizes the methodology. The historical indicators of the core-contribution phase are computed only from vessels that had already departed before each new arrival, so that no future information reaches the model. AIS data already exist at Tanger Med but were not used to produce the results of this paper (Section 3.1.1).

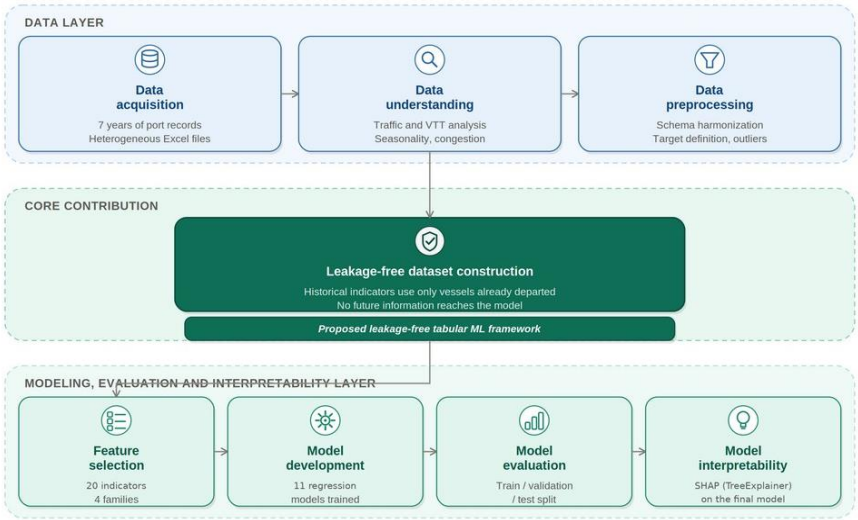


**Figure 1.** Research methodology: data layer, leakage-free dataset construction (core contribution) and modeling, evaluation and interpretability layer

**Table 2.** Data sources by year

| Year | File | Format | Key Schema Characteristics |
|---|---|---|---|
| 2019 | VSR_A2019 | .xls | 30 columns, French vessel types, header hidden on row 3 |
| 2020 | VSR_A2020 | .xls | 30 columns, French types, single GENERAL sheet |
| 2021 | vsr2021 | .xls | 30 columns, French types, single ALL sheet |
| 2022 | vsR_A2022 | .xlsx | 40 columns, port of registry added |
| 2023 | VSR_2023 | .xls | 51 columns, IMO number, richest schema |
| 2024 | VSR_A_2024 | .xls | 35 columns, no booking number field |
| 2025 | Reports_2025 | .xlsx | 25 columns, slimmer schema, header on row 0 |

Note: VSR, Vessel Situation Report; IMO, International Maritime Organization.

### 3.1 Data sources and harmonization

Seven Excel workbooks covering 2019-2025 were obtained as official Vessel Situation Reports (VSR). Each annual export uses a different schema: header rows are placed differently, column names change across years (e.g., “Booking #.” vs. “BRN”, “Length” vs. “LOA”, “ATAP” vs. “Port Entry”), and vessel types are recorded in French (e.g., “Navires porte-conteneurs”, “Vraquier”, “Navire de type ferry”). Table 2 summarizes the source files.

An ingestion routine detects the header row automatically by selecting the candidate row with the largest number of recognized column names, and maps more than 30 raw column spellings onto 15 canonical fields. French vessel-type labels are harmonized into eight English analytical categories (Container, Ferry, General Cargo, roll-on/roll-off (Ro-Ro)/Car Carrier, Tanker, Bulk Carrier, Barge, Other). Because a port call can involve several berths, all berth movements of a visit are collapsed into one visit-level record, keyed to the first berth of the call (earliest berthing timestamp). The merged raw table contains 105,579 berth-movement rows over the seven years.

Visit-record construction: the visit key combines the booking reference, the vessel name and the port-entry timestamp. In total, 18,116 raw rows (17.2%) have no booking reference and are keyed by vessel name and entry timestamp only; International Maritime Organization (IMO) numbers appear only in the 2023 export and are therefore not used in the key. Of the visits, 95,093 consist of a single berth-movement row, while the most complex calls span up to 30 rows. Because the entry timestamp is part of the key, repeated

calls of the same vessel, as well as different vessels sharing a name, remain separate visits; 2,177 vessel names have more than one entry date. Only 100 visits span a calendar-year boundary; they are assigned to their year of port entry.

3.1.1 Towards a real-time extension

AIS vessel tracking is standard equipment on commercial ships and is already received by major ports such as Tanger Med for navigational safety [24, 25]. The pipeline of this paper does not use AIS: it is built entirely from the batch VSR exports described above, which are compiled only after each call is complete. A natural next step, not evaluated here, is to connect the existing AIS feed to the same modeling pipeline, so that the historical indicators of Section 3.3.1 are updated continuously instead of once per batch export, giving earlier and more frequent VTT predictions. Since Tanger Med already manages berth allocation and call records digitally through its Port Community System, this platform is the natural place to connect a live AIS feed, mainly through a data connection and the same leakage-safe rules used in this study. The other steps of the pipeline (Figure 1) would remain unchanged.

### 3.2 Target definition and cleaning

The VTT is defined as the elapsed time between port entry and port departure for a visit:

$$\text{VTT (h)} = \text{Port Departure} - \text{Port Entry} \quad (1)$$

The source reports also provide a Call Time field in days; the recomputed VTT agrees with Call Time × 24 for all visits, which confirms the derivation of the target. Records without entry or departure times are removed after visit-level aggregation, and stays shorter than 15 minutes are discarded as data-entry errors. The first version of this study filtered upper-tail outliers with a per-vessel-type Tukey fence fitted on the training years and applied to all splits. Applying such a fence to 2024-2025 requires the realized duration of each call, which is unknown at prediction time; this filter was therefore removed. Instead, stays longer than 60 days are excluded by a fixed rule, set a priori and not derived from any data split: such stays (vessels under repair, long lay-ups, detentions) are not typical commercial calls, and the longest stay in the data exceeds one year (8,927 h). Because this cap also depends on the realized duration, it defines the scope of the study rather than a prediction-time filter, and the headline test metrics are also reported without it (Section 4.2). After the collapse of berth movements, there are 98,713 visits, of which 96,914 are retained after cleaning. The VTT distribution is strongly right-skewed (median 3.9 h, mean 10.6 h, maximum 744 h), which motivates a log(1+x) transform of the target before modeling.

### 3.3 Feature engineering

Eighteen raw fields are engineered into 20 model-ready features in four families, summarized in Table 3, which also states when each feature becomes available relative to the prediction moment. Predictions are made at the vessel's port entry. A twenty-first feature used in the first version of this study, the number of berth movements in a visit, was removed: it was the realized count of berth-movement rows per visit, known only after the call, and therefore constituted target leakage.

Port congestion is measured as the number of vessels present in port at each new arrival, computed with a sweep-line algorithm over entry and departure events of complexity O(n log n).

Feature availability at prediction time: none of the 20 features requires information generated after the prediction moment. The terminal feature is the first berth of the call, i.e., the assignment in force when the vessel first berths; it serves as a proxy for the berth assignment planned at entry, and intra-call reassignments are not used. At the container terminals (TC1 to TC4), which are linear quays where several vessels berth at the same time, the terminal is fixed in advance by the operator and the liner service, and a change of position along the quay does not change it. At terminals with distinct berths, the first berth may differ from the planned one if the vessel is reassigned before berthing; this case is discussed in Section 7.3. Congestion counts only the vessels present at arrival, and the three historical features use only calls that had already ended before the arrival, as explained below. Table 3 summarizes this availability by family.

**Table 3.** Feature families and availability at prediction time (20 features)

| Family | N | Features | Rationale | Available at Prediction Time |
|---|---|---|---|---|
| Vessel particulars | 9 | vessel_type, gt, loa, beam, draught, capacity, size_volume, loa_beam_ratio, gt_per_loa | Physical constraints on handling complexity | Before entry (static particulars) |
| Temporal | 6 | arr_hour, arr_dow, is_weekend, month, season, is_holiday | Scheduling and seasonality patterns | Before entry (from planned arrival) |
| Operational | 2 | terminal, congestion | Direct operational determinants of service time | At entry: first-berth assignment; vessels present at arrival |
| Historical/availability-time | 3 | hist_vtt_type, hist_vtt_berth, roll20_vtt_type | Past performance as the strongest predictor | At arrival, from already-departed calls only |
| Removed (first version) | 1 | n_berth_movements | Realized count of berth movements | After departure (target leakage) |

Note: gt, gross tonnage; loa, length overall; arr_dow, arrival day of week; vtt, vessel turnaround time; roll20, rolling mean over the last 20 departed calls. In deployment, temporal features are taken from the planned arrival; in the evaluation, they are computed from the recorded entry timestamp.

3.3.1 Availability-time historical features

A standard approach to historical efficiency features sorts calls by arrival time and applies an expanding mean with a simple index shift. This is not sound in an operational setting, because a vessel that arrived earlier may still occupy its berth when a new vessel arrives; its VTT has not yet been observed. Our approach instead sorts the historical calls of each group (by berth or by vessel type) by departure time, builds a cumulative sum of VTT over that order and, for each new arrival, uses a binary search (searchsorted) over departure

timestamps to find exactly how many previous calls of the group had already departed. The mean VTT of these completed calls is used as the historical efficiency feature; groups without qualifying history receive a training-year median as a fallback.

### 3.4 Modeling protocol

Temporal split: because the objective is to forecast future calls, the data are partitioned by calendar year rather than by random cross-validation, which would let information from future periods influence model fitting or selection (Table 4). The 13 visits whose port entry falls in December 2018 (with departure in early 2019) are not assigned to any of the three windows.

**Table 4.** Temporal data split

| Split | Years | N Visits | Purpose |
|---|---|---|---|
| Train | 2019-2022 | 46,674 | Model fitting |
| Validation | 2023 | 16,633 | Model selection only |
| Test | 2024-2025 | 33,594 | One-shot generalization estimate |

Preprocessing: the target is modeled in log(1+x) space; predictions are inverted before scoring, with a safety clip at log(1+5000) to bound extreme extrapolations. Tree-based models use median imputation and ordinal encoding of categorical features. Scale-sensitive models (linear regression, k-nearest neighbors (KNN) and support vector regression (SVR)) use median imputation, standardization and one-hot encoding of categorical features.

Models: eleven regression models are compared, namely two baselines (linear regression, decision tree); five tree ensembles (RF, Extra Trees, XGBoost, LightGBM, CatBoost); three other tabular learners (KNN, SVR with a radial basis function (RBF) kernel, NGBoost); and a two-level stacking ensemble in which XGBoost, RF and LightGBM base learners feed a ridge-regression meta-learner. The stacking ensemble uses two-fold cross-validation on two contiguous, non-shuffled chronological blocks of the training years. The out-of-fold predictions of the earlier block therefore come from base learners fitted on the later block; this look-ahead stays within 2019-2022 and affects only the meta-learner weights, while no validation or test data reach the meta-learner. For computational tractability, SVR was trained on a random 3,000-visit subsample and KNN and NGBoost on a random 15,000-visit subsample; these three models are treated as computational baselines.

Reproducibility: a single random seed (42) is used, and all hyperparameters are fixed a priori rather than tuned (CatBoost: 700 iterations, learning rate 0.05, depth 8; RF and Extra Trees: 200 trees, maximum depth 20; full per-model settings are provided with the code), so that no search-induced selection pressure is applied to the validation year. Software: Python 3.12, scikit-learn 1.6.1, XGBoost 3.3.0, LightGBM 4.6.0, CatBoost 1.2.10, NGBoost 0.5.11, pandas 2.2.2 and NumPy 2.0.2.

Evaluation: all metrics are computed on the original hour scale after inverting the log transform: mean absolute error (MAE, the primary and most operationally interpretable metric), root mean squared error (RMSE) and coefficient of determination ($R^2$). Following established guidance on forecast model selection [26], models are selected on validation MAE only; the test set is scored once, after the best model is locked, and test metrics play no role in model or hyperparameter selection.

## 4. RESULTS

### 4.1 Descriptive statistics

Tanger Med's traffic mix is dominated by ferries, which account for 51.0% of all visits in the clean dataset of the first version of this study (Table 5), followed by containers (21.6%) and general cargo (10.7%). VTT differs by more than an order of magnitude across vessel types (Table 5, Figure 2): ferries and Ro-Ro/car carriers have median stays of about 2.5-2.8 h, reflecting scheduled short-turnaround operations, while tankers and bulk carriers exceed 30 h on average, reflecting slower, less frequent and more variable cargo handling. This bimodality explains the strong right skew of the pooled VTT distribution and motivates the log-space target transform. Table 5 and these shares are computed on that dataset (N = 95,688); the corrected pipeline retains about 1,200 additional long-stay visits, which slightly raises upper-tail means and standard deviations without altering the qualitative picture (corrected pooled statistics: median 3.9 h, mean 10.6 h, maximum 744 h).

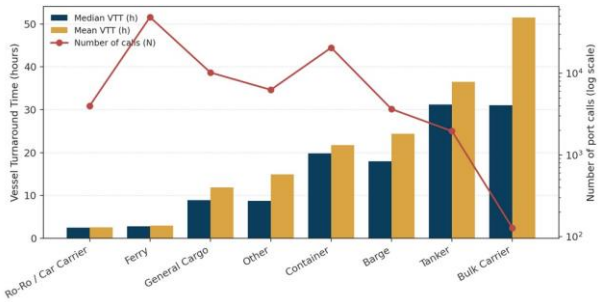


**Figure 2.** Median and mean vessel turnaround time (VTT) by vessel type (bars, left axis) and call volume (line, right axis, log scale), first-version dataset (N = 95,688)

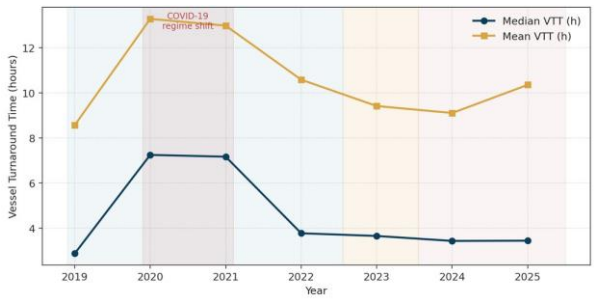


**Figure 3.** Yearly median and mean vessel turnaround time (VTT), 2019-2025, with the train, validation and test windows and the COVID-19 period shaded

The data also show a marked change of operating regime: the median VTT rises from 2.87 h in 2019 to 7.25 h and 7.17 h in 2020-2021, then returns to 3.4-3.8 h in 2022-2025 (Figure 3), consistent with the disruption of global shipping schedules during the COVID-19 pandemic. Because the affected years fall inside the training window, the models are exposed to this regime change during fitting rather than only at evaluation.

### 4.2 Model comparison

Table 6 reports validation and test metrics for the eleven models, sorted by validation MAE, the model-selection criterion (Figure 4).

**Table 5.** Vessel turnaround time (VTT) by vessel type, first-version clean dataset (N = 95,688)

| Vessel Type | N | Median (h) | Mean (h) | Std. Dev. (h) |
|---|---|---|---|---|
| Ferry | 48,834 | 2.80 | 2.97 | 0.92 |
| Container | 20,624 | 19.85 | 21.75 | 9.97 |
| General Cargo | 10,223 | 8.93 | 11.84 | 7.33 |
| Other | 6,280 | 8.77 | 14.91 | 16.78 |
| Ro-Ro/Car Carrier | 3,993 | 2.53 | 2.61 | 0.50 |
| Barge | 3,647 | 17.95 | 24.42 | 18.99 |
| Tanker | 1,959 | 31.17 | 36.46 | 24.72 |
| Bulk Carrier | 128 | 31.08 | 51.47 | 49.80 |

Note: Ro-Ro, roll-on/roll-off.

**Table 6.** Model comparison sorted by validation MAE

| Model | Val. MAE (h) | Val. $R^2$ | Test MAE (h) | Test RMSE (h) | Test $R^2$ |
|---|---|---|---|---|---|
| CatBoost | 3.63 | 0.392 | 4.62 | 11.06 | 0.408 |
| Extra Trees | 3.65 | 0.393 | 4.52 | 10.84 | 0.432 |
| Stacking (XGB+RF+LGBM) | 3.68 | 0.392 | 4.42 | 11.04 | 0.410 |
| RF | 3.69 | 0.401 | 4.43 | 10.94 | 0.421 |
| NGBoost | 3.74 | 0.388 | 4.37 | 10.82 | 0.434 |
| XGBoost | 3.76 | 0.405 | 4.60 | 10.96 | 0.418 |
| LightGBM | 3.76 | 0.399 | 4.44 | 10.77 | 0.439 |
| KNN | 3.93 | 0.366 | 4.08 | 10.75 | 0.441 |
| Decision Tree | 4.17 | 0.318 | 5.12 | 11.87 | 0.318 |
| SVR (RBF) | 4.47 | 0.315 | 4.80 | 11.55 | 0.354 |
| Linear Regression | 4.53 | −11.40 | 5.53 | 11.76 | 0.331 |

Note: MAE, mean absolute error; RMSE, root mean squared error; $R^2$, coefficient of determination; KNN, k-nearest neighbors; SVR, support vector regression; RBF, radial basis function; XGB, XGBoost; RF, Random Forest; LGBM, LightGBM. Test metrics under the 60-day cap. SVR was trained on 3,000 visits, KNN and NGBoost on 15,000 visits (computational baselines).

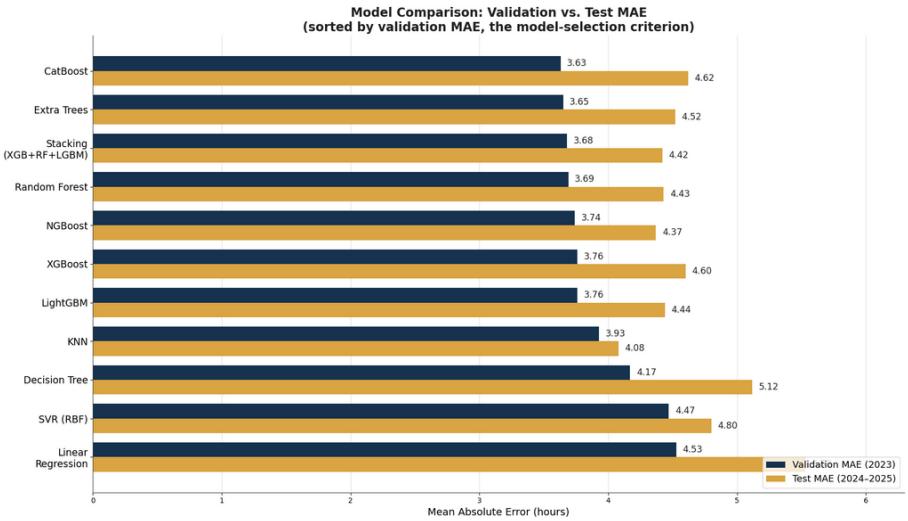


**Figure 4.** Validation mean absolute error (MAE, 2023) and test MAE (2024-2025) for the eleven models, sorted by validation MAE

The leading tree ensembles are separated by only about 0.13 h of validation MAE (3.63-3.76 h), which indicates that, once features are of good quality, the choice of algorithm matters less for this task. CatBoost has the best validation MAE (3.63 h) and obtains, on the test years, MAE = 4.62 h, RMSE = 11.06 h and $R^2$ = 0.41 under the 60-day cap (Figure 5). For completeness, the same pipeline was run with the cap disabled, keeping every visit in all splits; the same CatBoost configuration then gives test MAE = 5.05 h, RMSE = 39.81 h and $R^2$ = 0.06. A few multi-month stays dominate the squared-error metrics, whereas MAE is far less affected and easier to interpret; the 60-day setting is therefore reported as primary. KNN has the lowest raw test MAE (4.08 h) but ranks only eighth on validation; selecting it would amount to model selection on the test set, which the protocol forbids. Linear regression reaches a strongly negative validation $R^2$ (−11.4), although its test $R^2$ is positive; this value suggests that VTT is a highly nonlinear, interaction-heavy function of the features. The stacking ensemble slightly outperforms its base learners on validation but does not surpass CatBoost. Compared with the first version of this study (Extra Trees, test MAE = 3.52 h, $R^2$ = 0.68), the corrected results are more modest; the

difference reflects information that the earlier pipeline used, but that is not available at prediction time.

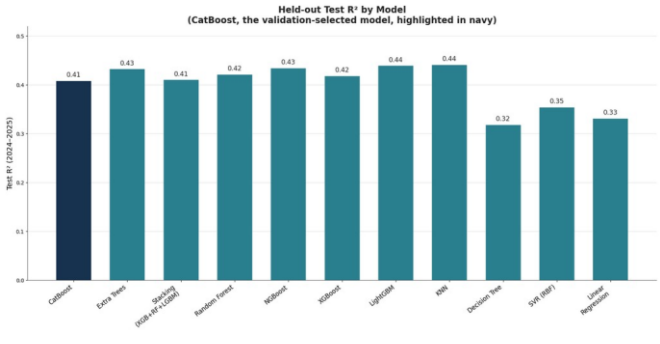


**Figure 5.** Held-out test coefficient of determination ($R^2$) by model (60-day cap); CatBoost, the validation-selected model, is highlighted

### 4.3 Per-vessel-type error analysis

Table 7 breaks down the test error of the selected CatBoost model by vessel type (Figure 6).

Errors below one hour for ferries, which account for 41% of test-period calls, indicate an accuracy that could be useful for scheduling this class, although this partly reflects the large sample and the tightly scheduled operations of ferries rather than the vessel type itself. Errors increase for lower-volume, higher-variance types, and median absolute errors are well below the class means for long-stay categories (General Cargo: median 2.2 h vs. mean 6.2 h), which indicates that a minority of difficult calls dominates those means. The bulk-carrier MAE (26.76 h) is based on only 27 test calls and is not statistically reliable. The growth of error within the test period is discussed in Section 4.6.

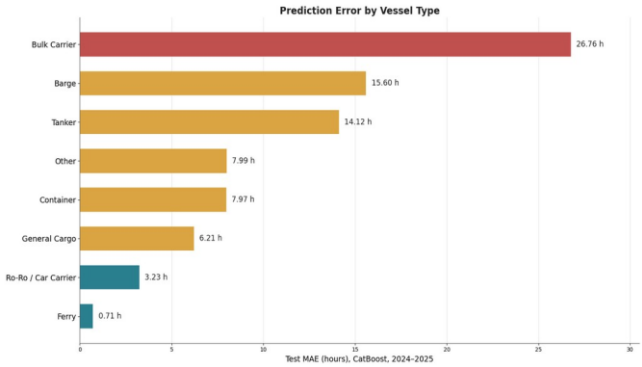


**Figure 6.** Test mean absolute error (MAE) by vessel type, CatBoost, 2024-2025

**Table 7.** Test error by vessel type, CatBoost, 2024-2025

| Vessel Type | N (test) | MAE (h) | RMSE (h) | Median AE (h) |
|---|---|---|---|---|
| Ferry | 13,940 | 0.71 | 2.00 | 0.46 |
| Other | 6,017 | 7.99 | 15.41 | 5.15 |
| Container | 4,930 | 7.97 | 11.62 | 5.79 |
| Ro-Ro/Car Carrier | 4,632 | 3.23 | 6.89 | 2.45 |
| General Cargo | 2,148 | 6.21 | 17.25 | 2.18 |
| Barge | 1,322 | 15.60 | 24.67 | 9.48 |
| Tanker | 578 | 14.12 | 20.78 | 9.41 |
| Bulk Carrier | 27 | 26.76 | 40.93 | 12.94 |

Note: MAE, mean absolute error; RMSE, root mean squared error; Median AE, median absolute error; Ro-Ro, roll-on/roll-off.

**Table 8.** Ablation over nested feature sets, selected model (CatBoost)

| Feature Set | N feat. | Val. MAE (h) | Val. $R^2$ | Test MAE (h) | Test $R^2$ |
|---|---|---|---|---|---|
| Vessel + temporal | 15 | 3.84 | 0.364 | 6.52 | 0.264 |
| + operational | 17 | 3.94 | 0.341 | 6.03 | 0.304 |
| + historical (full set) | 20 | 3.63 | 0.395 | 4.61 | 0.407 |

Note: MAE, mean absolute error; $R^2$, coefficient of determination.

### 4.4 Ablation study

To isolate the contribution of each feature family, the CatBoost configuration is retrained on three nested feature sets: (i) vessel particulars and temporal features only (15 features), (ii) plus terminal and congestion (17), and (iii) plus the three availability-time historical features (20). Table 8 shows the results. The two operational features raise test $R^2$ from 0.26 to 0.30, but the main improvement comes from the availability-time historical features, which raise test $R^2$ from 0.30 to 0.41 and reduce test MAE from 6.03 h to 4.61 h (the full-set run gives 4.61 h, against 4.62 h for the separate run of Table 6). In all three sets, validation MAE is much lower than test MAE, and the historical features matter most for the test horizon, where static vessel and calendar information generalizes poorly. This interventional evidence is consistent with the SHAP results of Section 5.

### 4.5 Comparison on equal training subsamples

Because SVR, KNN and NGBoost were trained on smaller subsamples in the main comparison, all eleven models were retrained on the same random subset of 3,000 training visits to remove the effect of training-set size. The leading tree ensembles then become almost indistinguishable: Extra Trees (validation MAE = 3.78 h), NGBoost (3.80 h) and the stacking ensemble (3.81 h) are close to CatBoost (3.82 h), with XGBoost (3.85 h) and RF (3.86 h) close behind. The ordering of the weaker families is unchanged (SVR 4.47 h, linear regression 4.53 h, single decision tree 4.93 h). The ranking of model families is therefore not an artefact of unequal training-set sizes, and CatBoost's advantage in Table 6 reflects its use of the full 46,674-visit training set.

### 4.6 Temporal drift within the test period

Both test years are complete calendar years (first entry on 1 January 2024, last entry on 31 December 2025), and monthly call counts are of similar magnitude in both years, so the difference between 2024 and 2025 cannot be attributed to partial coverage. Comparing the same calendar months also controls for seasonality. Table 9 reports the test MAE of the selected model for each month.

**Table 9.** Test mean absolute error (MAE) by calendar month, 2024 vs. 2025, selected model

| Month | N (2024) | MAE 2024 (h) | N (2025) | MAE 2025 (h) |
|---|---|---|---|---|
| January | 1,371 | 3.51 | 1,296 | 3.82 |
| February | 1,253 | 4.17 | 1,241 | 4.07 |
| March | 1,171 | 4.54 | 1,271 | 5.10 |
| April | 1,333 | 3.59 | 1,283 | 4.49 |
| May | 1,430 | 3.24 | 1,395 | 6.69 |
| June | 1,493 | 3.09 | 1,355 | 6.17 |
| July | 1,799 | 2.49 | 1,587 | 5.59 |
| August | 1,844 | 2.78 | 1,657 | 5.28 |
| September | 1,448 | 2.74 | 1,406 | 6.69 |
| October | 1,453 | 3.40 | 1,371 | 7.38 |
| November | 1,288 | 3.73 | 1,270 | 7.88 |
| December | 1,320 | 4.18 | 1,259 | 7.71 |

The gap widens with the forecast horizon: early-2025 months are very close to 2024 (January 3.51 vs. 3.82 h; February 4.17 vs. 4.07 h), while from May 2025 onward the error roughly doubles relative to the same month of 2024 (May 3.24 vs. 6.69 h; October 3.40 vs. 7.38 h; November 3.73 vs. 7.88 h). Since monthly sample sizes are comparable and the calendar month is held fixed, neither sample size nor seasonality explains this increase; it indicates that the operating regime is not stable in 2025. The availability-time historical aggregates follow this change only partially, and the mapping from these aggregates to VTT, fitted on 2019-2022, becomes less representative as the horizon extends. Periodic retraining is therefore a prerequisite for sustained accuracy (Section 7.3).

## 5. INTERPRETABILITY ANALYSIS

Feature attribution is computed with SHAP using TreeExplainer directly on the selected CatBoost model; no surrogate model is used. Table 10 and Figure 7 report the mean absolute SHAP value of each feature on the test set.

The three availability-time historical features, namely the 20-call rolling vessel-type average (mean |SHAP| = 0.216), the historical berth efficiency (0.211) and the historical vessel-type efficiency (0.100), account for more attribution mass (0.527) than the 17 remaining features. This agrees with the ablation of Section 4.4, which was obtained independently by removing these features and retraining the model. It is also consistent with port planners' intuition that the best signal for how long an upcoming call will take is how long comparable recent calls at the same berth, or of the same vessel type, have taken. The largest residuals are concentrated in the bulk-carrier class, which has high intrinsic VTT variance and the smallest sample of any vessel type in the dataset.

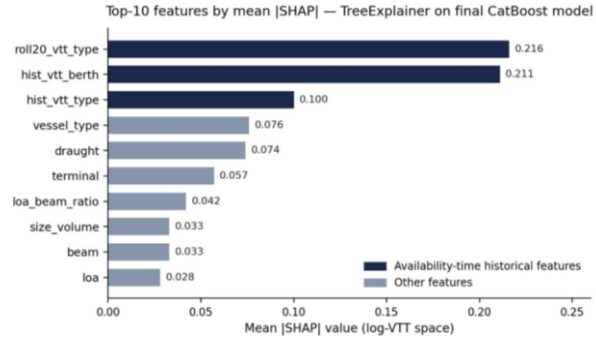


**Figure 7.** Top-10 features by mean absolute SHapley Additive exPlanations (SHAP) value, TreeExplainer on the final CatBoost model

**Table 10.** Top-ranked features by mean absolute SHapley Additive exPlanations (SHAP) value, TreeExplainer on the final CatBoost model

| Rank | Feature | Mean \|SHAP\| | Interpretation |
|---|---|---|---|
| 1 | roll20_vtt_type | 0.216 | Rolling 20-call vessel-type average |
| 2 | hist_vtt_berth | 0.211 | Historical berth efficiency |
| 3 | hist_vtt_type | 0.100 | Historical vessel-type efficiency |
| 4 | vessel_type | 0.076 | Categorical vessel class |
| 5 | draught | 0.074 | Vessel depth / handling complexity |
| 6 | terminal | 0.057 | Terminal assignment (first berth) |
| 7 | loa_beam_ratio | 0.042 | Hull shape factor |
| 8 | size_volume | 0.033 | Physical size proxy |
| 9 | beam | 0.033 | Vessel width |
| 10 | loa | 0.028 | Length overall |

## 6. OPERATIONAL PREDICTION MODULE

The selected model was wrapped in a lightweight prediction module, intended as a potential input to berth planning. Given the particulars of an incoming vessel and its expected arrival time, the module computes the congestion and availability-time features on the fly, using only calls completed before that arrival time, and returns a predicted VTT, an expected departure time and a historical error band equal to the test-set MAE of the vessel type. Table 11 shows illustrative, non-held-out predictions used to verify the input/output contract of the module.

**Table 11.** Illustrative example predictions from the operational module

| Vessel | Type | Berth | Planned Entry | Pred. VTT (h) | Historical Error Band |
|---|---|---|---|---|---|
| A | Container | TC1 | 2025-08-15 08:00 | 18.2 | ±8.0 h |
| B | Ferry | TF1 | 2025-08-15 10:00 | 2.4 | ±0.7 h |
| C | Tanker | T2 | 2025-08-15 14:00 | 13.3 | ±14.1 h |

Note: VTT, vessel turnaround time. The error band is the test-set mean absolute error of the vessel type (Table 7), not a prediction interval.

The historical error bands are set to the test-set MAE of each vessel type (Table 7); they have no coverage probability and are not prediction intervals. Since the model provides point estimates only, probabilistic extensions are left for future work (Section 8). Possible uses include berth-allocation sequencing, resource rostering, congestion early warning, and anchorage ETA management. Because the historical efficiency features are the dominant predictors (Sections 4.4 and 5) and drift is observed within the test period (Section 4.6), the accuracy of the module depends on regular recalibration, monthly or after material operational changes such as a new terminal, a berth reconfiguration or a shift in the vessel-type traffic mix.

## 7. DISCUSSION

### 7.1 Position relative to the literature

The dominance of tree ensembles over a linear baseline and the tight clustering of the top tree-based models are consistent with the large-scale empirical findings of Grinsztajn et al. [20]: tree-based methods remain highly competitive with, and often superior to, more complex model classes on real-world tabular data of moderate size. Relative to the container-terminal operations-research literature [2, 3], this work is complementary rather than competing: it provides a data-driven service-time predictor that could feed the service-time parameters assumed by berth-allocation optimization formulations, not replace them.

### 7.2 Why CatBoost performed best

CatBoost differs from other gradient-boosting implementations through its ordered-boosting scheme and its native, target-statistics-based handling of categorical features [18], which matters here because two of the main predictors (terminal and vessel type) are categorical. Its advantage is, however, modest and scale-dependent: in the comparison on equal subsamples (Section 4.5), Extra Trees narrowly outperforms CatBoost on a 3,000-visit training subsample (validation MAE 3.78 h vs. 3.82 h). CatBoost is thus the best model at the full scale of our dataset, but it is the tree-ensemble family as a whole, rather than any single member, that performs best.

### 7.3 Limitations

Several limitations apply. First, VTT as defined here measures total port occupancy, not productive berth or cargo-handling time. Second, the temporal features include only fixed-date Moroccan holidays and omit holidays that follow the Islamic calendar. Third, the dataset contains only 128 bulk-carrier calls (27 in the test period), so the error reported for this class is not statistically reliable. Fourth, KNN, SVR and NGBoost were trained on random subsamples for computational reasons; the comparison on equal subsamples (Section 4.5) mitigates, but does not remove, this comparability issue. Fifth, there is temporal drift within the test period: for the same calendar months, the test MAE is close to its 2024 level in early 2025 but almost doubles from May 2025 onward (Section 4.6); the model must therefore be retrained periodically, and the reported test metrics depend on the horizon. Sixth, RMSE and $R^2$ are highly sensitive to a few multi-month stays; the primary metrics are therefore reported under a fixed 60-day cap, and uncapped results are given in Section 4.2 for transparency. Seventh, at terminals with distinct berths, the first berth is used as a proxy for the planned assignment and may differ from it when a vessel is reassigned before berthing. Eighth, the study covers a single port: the availability-time feature construction is general, but the reported accuracy cannot be assumed to transfer to other ports.

## 8. CONCLUSIONS

This paper presents a reproducible, leakage-controlled tabular machine learning framework for forecasting VTT at Tanger Med from seven years of operational records. CatBoost, selected on a held-out validation year, reaches a test MAE of 4.62 h and an $R^2$ of 0.41 for stays up to 60 days in 2024-2025 (MAE of 5.05 h on all calls). These results are less accurate than those of our earlier pipeline (MAE of 3.52 h, $R^2$ of 0.68), which relied on information not available at prediction time. For ferries, the busiest and most schedule-critical class, the error is below one hour; demonstrating operational value would nevertheless require a scheduling or simulation experiment. The ablation and the SHAP analysis both identify the availability-time historical indicators as the strongest predictors. Their construction applies to any setting where aggregates are built from event histories that are still in progress.

Future work includes: addressing the temporal drift of Section 4.6 through scheduled retraining or online updates of the historical features; adding the planned number of berth shifts declared by the shipping agent before arrival, once these declarations are linked to the pipeline; extending the framework to other ports to study transfer and cross-port generalization; producing calibrated prediction intervals instead of heuristic error bands, for example with quantile gradient boosting, conformal prediction or the distributional output of NGBoost; implementing the real-time acquisition layer of Section 3.1.1, in which the availability-time rule of Section 3.3.1 is applied to individual AIS events instead of batch exports; coupling the predictive module with berth-allocation models such as those surveyed by Bierwirth and Meisel [2] in a scheduling experiment; and comparing deep-learning sequence architectures with the tree ensembles used here, in light of the tabular-data findings of Grinsztajn et al. [20].

## DATA AVAILABILITY

The Vessel Situation Reports are operational records of the port authority and contain commercially sensitive information; they cannot be made public. The complete pipeline code (ingestion, cleaning, feature engineering, modeling protocol, ablation and drift analyses, and all hyperparameter settings) is available from the corresponding author on reasonable request; derived, aggregated data may be shared on request, subject to the approval of the port authority.

## REFERENCES

[1] Salhi, A., El Hasnaoui, Y., Pérez Cutillas, P., Heggy, E. (2023). Soil erosion and hydroclimatic hazards in major African port cities: The case study of Tangier. Scientific

Reports, 13: 13158. https://doi.org/10.1038/s41598-023-40135-3

[2] Bierwirth, C., Meisel, F. (2010). A survey of berth allocation and quay crane scheduling problems in container terminals. European Journal of Operational Research, 202(3): 615-627. https://doi.org/10.1016/j.ejor.2009.05.031

[3] Steenken, D., Voß, S., Stahlbock, R. (2004). Container terminal operation and operations research: A classification and literature review. OR Spectrum, 26(1): 3-49. https://doi.org/10.1007/s00291-003-0157-z

[4] Cerqueira, V., Torgo, L., Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning, 109(11): 1997-2028. https://doi.org/10.1007/s10994-020-05910-7

[5] Yazar Okur, İ.G., Tuna, O. (2022). Schedule reliability in liner shipping: A study on global shipping lines. Pomorstvo, 36(2): 389-400. https://doi.org/10.31217/p.36.2.22

[6] Fancello, G., Pani, C., Pisano, M., Serra, P., Zuddas, P., Fadda, P. (2011). Prediction of arrival times and human resources allocation for container terminal. Maritime Economics & Logistics, 13(2): 142-173. https://doi.org/10.1057/mel.2011.3

[7] Pani, C., Fadda, P., Fancello, G., Frigau, L., Mola, F. (2014). A data mining approach to forecast late arrivals in a transhipment container terminal. Transport, 29(2): 175-184. https://doi.org/10.3846/16484142.2014.930714

[8] Parolas, I. (2016). ETA prediction for containerships at the Port of Rotterdam using machine learning techniques. Master's thesis, Delft University of Technology, Delft, The Netherlands.

[9] Jahn, C., Scheidweiler, T. (2018). Port call optimization by estimating ships' time of arrival. In International Conference on Dynamics in Logistics, Bremen, Germany, pp. 172-177. https://doi.org/10.1007/978-3-319-74225-0_23

[10] Bodunov, O., Schmidt, F., Martin, A., Brito, A., Fetzer, C. (2018). Real-time destination and ETA prediction for maritime traffic. In Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems (DEBS '18), Hamilton, New Zealand, pp. 198-201. https://doi.org/10.1145/3210284.3220502

[11] Kolley, L., Rückert, N., Kastner, M., Jahn, C., Fischer, K. (2023). Robust berth scheduling using machine learning for vessel arrival time prediction. Flexible Services and Manufacturing Journal, 35(1): 29-69. https://doi.org/10.1007/s10696-022-09462-x

[12] El Mekkaoui, S., Benabbou, L., Berrado, A. (2020). Deep learning models for vessel's ETA prediction: Bulk ports perspective. Flexible Services and Manufacturing Journal, 35: 5-28. https://doi.org/10.1007/s10696-022-09471-w

[13] Filom, S., Amiri, A.M., Razavi, S. (2022). Applications of machine learning methods in port operations: A systematic literature review. Transportation Research Part E: Logistics and Transportation Review, 161: 102722. https://doi.org/10.1016/j.tre.2022.102722

[14] Scornet, E., Hooker, G. (2026). Theory of random forests. Annual Review of Statistics and Its Application, 13: 99-121. https://doi.org/10.1146/annurev-statistics-112723-034707

[15] Ghazwani, M., Begum, M.Y. (2023). Computational intelligence modeling of hyoscine drug solubility and solvent density in supercritical processing: Gradient boosting, extra trees, and random forest models. Scientific Reports, 13: 10046. https://doi.org/10.1038/s41598-023-37232-8

[16] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785

[17] Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, pp. 3149-3157. https://dl.acm.org/doi/10.5555/3294996.3295074.

[18] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montréal, Canada, pp. 6638-6648. https://dl.acm.org/doi/abs/10.5555/3327757.3327770.

[19] Duan, T., Avati, A., Ding, D.Y., et al. (2020). NGBoost: Natural gradient boosting for probabilistic prediction. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, pp. 2690-2700. https://dl.acm.org/doi/abs/10.5555/3524938.3525190.

[20] Grinsztajn, L., Oyallon, E., Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, pp. 507-520. https://dl.acm.org/doi/10.5555/3600270.3600307.

[21] Gijsbers, P., Bueno, M.L.P., Coors, S., et al. (2024). AMLB: An AutoML benchmark. Journal of Machine Learning Research, 25(1): 4978-5042. https://dl.acm.org/doi/abs/10.5555/3722577.3722678.

[22] Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, pp. 4765-4774. https://dl.acm.org/doi/10.5555/3295222.3295230.

[23] Mazarei, A., Sousa, R., Mendes-Moreira, J., Molchanov, S., Ferreira, H.M. (2024). Online boxplot derived outlier detection. International Journal of Data Science and Analytics, 19: 83-97. https://doi.org/10.1007/s41060-024-00559-0

[24] Ali, O., Ishak, M.K., Bhatti, M.K.L., Khan, I., Kim, K.I. (2022). A comprehensive review of Internet of Things: Technology stack, middlewares, and fog/edge computing interface. Sensors, 22(3): 995. https://doi.org/10.3390/s22030995

[25] Yang, D., Wu, L., Wang, S., Jia, H., Li, K.X. (2019). How big data enriches maritime research: A critical review of Automatic Identification System (AIS) data applications. Transport Reviews, 39(6): 755-773. https://doi.org/10.1080/01441647.2019.1649315

[26] Fildes, R., Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. Journal of Business Research, 68(8): 1692-1701. https://doi.org/10.1016/j.jbusres.2015.03.028

**NOMENCLATURE**

| | |
|---|---|
| AIS | Automatic Identification System |
| ANN | artificial neural network |
| ETA | estimated time of arrival |
| IMO | International Maritime Organization |
| KNN | k-nearest neighbors |
| MAE | mean absolute error, h |
| n, N | number of vessel visits (dimensionless) |
| $R^2$ | coefficient of determination (dimensionless) |
| RBF | radial basis function |
| RMSE | root mean squared error, h |
| Ro-Ro | roll-on/roll-off |
| SHAP | SHapley Additive exPlanations |
| SVR | support vector regression |
| VSR | Vessel Situation Report |
| VTT | vessel turnaround time, h |