Data analytics and throughput forecasting in port management systems against disruptions: a case study of Busan Port

Truong Ngoc Cuong1 · Le Ngoc Bao Long2 · Hwan-Seong Kim1 · Sam-Sang You2

Accepted: 2 October 2022 / Published online: 23 October 2022
© The Author(s), under exclusive licence to Springer Nature Limited 2022

Abstract
Forecasting cargo throughput is an essential albeit challenging task in ensuring efficient seaport management. In this study, data analytics is employed to analyze the nonlinear dynamic behaviors, as well as disruptions in port throughputs. Further, nonlinear analytical methods, including the Lyapunov exponent (LE), information entropy, Hurst exponent, and wavelet decomposition, are employed to explore the complex dynamic behavior of port throughput under supply chain disruptions. By employing the discrete wavelet transform (DWT) and the long short-term memory (LSTM) network, we develop a novel hybrid model of port throughput forecasting. DWT is employed to decompose the original data into a finite set of frequency components, so that the various hidden features of cargo throughput can be extracted via different modes, such as the trend, residual, and seasonal components. Thereafter, each component, obtained from the DWT spectra, is predicted via a machine learning model. Additionally, hypothesis testing, model evaluation, and statistical significance tests are employed to comprehensively evaluate the introduced forecasting models. Regarding prediction accuracy and efficiency, our extensive simulation results confirm the superiority of the hybrid strategy over five benchmarked models. Finally, by employing business forecasting software, we show that the robust hybrid strategy achieves accurate predictions of port throughputs against market disruptions. Our findings can help decision-makers understand disruption mechanisms in port systems, thus enabling them to successfully achieve their business goals.

Keywords Time-series data · Port throughput · Nonlinear dynamic analysis · Discrete wavelet transform · Machine learning · Hybrid forecasting model
1 Introduction

With the advent of globalization, ocean transportation, as well as port management, are assuming an essential role in international trade (Notteboom 2016). Cargo throughput is the commonest indicator in evaluating port performance, representing a benchmarking reference for other ports (Justice et al. 2016). Moreover, cargo throughput reflects the economic growth and rigor of a port (Du et al. 2019). However, exploring the underlying mechanisms of cargo throughput development against exogenous and endogenous disruptions is challenging (Gou and Lam 2019). Numerous tools for investigating the dynamic properties of port throughput have been employed in earlier research. These have included nonlinear control theory (Cuong et al. 2021); data envelopment analysis (Tetteh et al. 2016); decomposition-ensemble methodology (Xie et al. 2017) and stochastic modeling (Saini et al. 2017). Although each approach exhibits unique advantages, they do not describe all the typical features of the cargo throughput system in isolation.

The instability of system is induced by multiple shocks, although only a few studies have considered this (Justice et al. 2016). The nonlinear system theory avails an efficient strategy to describe the complex dynamic behavior of a real system (Slotine and Li 1991; Morris and Pratt 2003; Cuong et al. 2021). However, a few empirical studies only have explored the dynamic behavior of cargo throughput systems through nonlinear techniques. The approach employed here can elucidate the underlying mechanisms of cargo throughput systems, helping port policymakers determine appropriate strategies (Zhang and Lam 2017). Additionally, we aim to fill the research gap between the theoretical and practical applications of nonlinear analysis tools, such as the Lyapunov exponent (LE), information entropy, Hurst exponent, and wavelet decomposition, for investigating the dynamic properties of cargo throughput systems under disruptions. Based on the tools discussed here, port decision-makers can gain further insights into nonlinear phenomena, such as periodicity, stability, and chaotic behaviors.

Moreover, port throughput forecasting is crucial to the management, construction, and improvement of port infrastructures, and accurate business forecasting can optimize resource utilization and efficiency, while avoiding investment mistakes (Du et al. 2019; Twrdy and Batista 2016). Conversely, low prediction accuracy, owing to a bias, can cause significant financial losses (Xie et al. 2017). Therefore, a highly effective forecasting method is required for efficient port operations. Additionally, predicting future trends and events leads to informed business decisions.

Autoregressive integrated moving average (ARIMA) forecasting techniques have been widely utilized among practitioners (as a benchmark). Models that have been employed in logistics and supply chain management (Schulze and Prinz 2009) include exponential smoothing (Gosasang et al. 2011), Grey forecasting (Liu and Chen 2006) and deep learning (Yang and Chang 2020; Shankar et al. 2021; Zuo et al. 2021). However, port throughput data reveal complex nonlinear dynamic behaviors under the various disruptions. In such cases, however, Grey forecasting or linear regression do not provide high accuracy, in ever-changing and noisy datasets (Punia et al. 2020).
Recently, deep machine learning has exhibited the promise of AI applications in many forecasting fields, such as multichannel retailers (Ferreira et al. 2016; Punia et al. 2020), financial markets (Fischer and Krauss 2018), and insurance big data analysis (Lin et al. 2017). The advantage of deep learning is in its capacity to create a hybrid function set comprising numerous nonlinear transformations and valuable expressions that can produce additional abstractions, followed by potential benefits (Punia et al. 2020; Hochreiter and Schmidhuber 1997). The convolutional neural network (CNN) is superior to existing methods for image recognition, while the recurrent neural network (RNN) exhibits excellent performance in speech recognition and natural language processing. RNN stores long sequential information in hidden memory for proper processing, representation, and storage (Memarzadeh and Keynia 2020). RNN employing recurrent links between hidden layers has been proposed to process history-related issues in the field of artificial intelligence (AI). Unfortunately, RNN cannot learn long-term historical data, causing an issue that was resolved by Hochreiter and Schmidhuber (1997) via the long short-term memory (LSTM) network, an extension of the RNN algorithm. LSTM can learn the long- and short-term effects of preceding data. LSTM networks with deep architectures have achieved great success in different applications involving large datasets (Zhang et al. 2021; Hochreiter and Schmidhuber 1997).

High-performing powerful algorithms enhance the accuracy of port throughput forecasting during significant economic changes that affect maritime transportation activities, e.g., the 2009 financial crisis or the coronavirus pandemic (COVID-19) (Cullinane and Haralambides 2021). By following existing data-driven solutions to issues of production and operations management (Lin et al. 2017; Fischer and Krauss 2018; Johannesen et al. 2019; Punia et al. 2020; Islam and Amin 2020; Sipper and Moore 2021), this study presents a novel hybrid strategy for addressing complex throughput forecasting scenarios. The proposed algorithm is based on deep learning and employs the LSTM methodology with wavelet decomposition. The effectiveness of our method is assessed via a set of evaluation indices and statistical significance tests. Through effective forecasting, port management can gain valuable insights, reduce costs, increase profitability, learn from mistakes, predict throughput trends, and attain sustainable competitive advantage.

In summary, our novel data analytics methodology is based on time-series analysis, entropy analysis, LE, Hurst exponent, and wavelet decomposition. It can be applied to forecasting issues involving time-series data with market uncertainties caused by multiple disruptive factors. For policymakers, this novel method can offer insights to the system dynamics of port throughput and help them explore the underlying mechanisms of business uncertainties caused by maritime transportation disruptions. Furthermore, our schemes can help port authorities predict the productivity of the port management system, subsequently helping managers to make appropriate operational decisions. Additionally, this study presents an advanced machine learning-based method for improving the accuracies and efficiencies of predictions.

The remainder of this paper is organized as follows: Sect. 2 presents the analysis of port throughput via novel data analytics. Section 3 describes the hybrid forecasting method, and Sect. 4 presents the empirical simulation results, as well as the
statistical analysis. Finally, Sect. 5 draws the main conclusions of the study, highlighting the most significant findings.

2 Data analytics for port throughput

2.1 Cargo throughput and maritime traffic volume of Busan Port

To many analysts, the market dynamics of the twenty-first century point out to accelerating globalization and the advent of cross-border economic frictions. In this situation, the hub of global trade and economy is shifting to East Asia, around China. Approximately 50% of global container flows take place in this region, thereby directly and indirectly increasing competition for hub port positions among Asian ports (Lam and Notteboom 2014). Such competition takes place in various fronts, ranging from the production of the port service to attracting transshipment cargo (Dirzka and Acciaro 2021).

Busan Port is located at the mouth of the Naktong River in South Korea. The port is the sixth-busiest port in the world and the largest transshipment center in Northeast Asia (Lee et al. 2021). Data obtained from the port authority show that the port handled 118.74 million tons of transshipment in March 2020, a record high since Busan New Port began operations in 2006. Cargo throughput and ship traffic of the port (during the period of our research, i.e., 2001–2021) are presented in Table 1. Data analytics is vital in getting meaningful insights to time series data. Nonlinear time-series analysis is employed here to systematically explore observed data (typically univariate) and the dynamic behavior of the port management system. The throughput trend was investigated via LE, information entropy, Hurst exponent with fractal dimension, wavelet decomposition, and statistical significance tests.

2.2 Determination of dynamic behavior via LEs

Graphical methods and quantifiers are commonly employed within the port and logistics research community to identify existing dynamic behaviors, whether a system is stable, periodic, quasiperiodic, or chaotic (Sprott 1995; Hwarng and Xie 2008). Regarding nonlinear time-series analysis, a more accurate alternative, involving the calculation of some quantifiers, can be proposed (Hwarng and Xie 2008). Data analytics is the science of analyzing raw data to offer informed conclusions. Some well-known measures include LE, information entropy, fractal dimension, and capacity dimension (Sprott 1995).

The Lyapunov spectrum facilitates the qualitative and quantitative characterizations of nonlinear features (Wolf et al. 1985). Employing the collected data, LE is obtained by measuring the average exponential rates of divergence or convergence of nearby orbits in phase space. Even though the Lyapunov spectrum can be determined via different approaches, LE is defined in a manner that is most relevant to the spectral calculations from a time series. For a continuous system in $n$-dimensional
| Year | Total                                      | Entry into port | Departure                                      |
|------|--------------------------------------------|-----------------|-----------------------------------------------|
|      | Number of ships | Cargo throughput | Number of ships | Cargo throughput | Number of ships | Cargo throughput |
| 2001 | 83,547 | 541.14 | 41,779 | 270.40 | 41,768 | 270.74 |
| 2002 | 92,587 | 583.36 | 46,321 | 291.36 | 46,266 | 291.99 |
| 2003 | 94,533 | 626.94 | 47,241 | 313.28 | 47,292 | 313.66 |
| 2004 | 97,329 | 660.36 | 48,671 | 330.07 | 48,658 | 330.29 |
| 2005 | 96,711 | 709.06 | 48,343 | 354.35 | 48,368 | 354.70 |
| 2006 | 100,787 | 728.16 | 50,385 | 363.60 | 50,402 | 364.56 |
| 2007 | 102,852 | 791.10 | 51,395 | 395.13 | 51,457 | 395.97 |
| 2008 | 115,931 | 832.56 | 57,979 | 416.34 | 57,952 | 416.22 |
| 2009 | 100,113 | 852.33 | 50,012 | 425.54 | 50,101 | 426.80 |
| 2010 | 104,995 | 936.42 | 52,484 | 467.81 | 52,511 | 468.61 |
| 2011 | 100,875 | 1027.15 | 50,447 | 513.50 | 50,428 | 513.65 |
| 2012 | 100,845 | 1062.23 | 50,437 | 534.40 | 50,408 | 527.83 |
| 2013 | 99,249 | 1128.25 | 49,588 | 567.88 | 49,661 | 560.37 |
| 2014 | 95,378 | 1107.03 | 47,718 | 557.17 | 47,660 | 549.86 |
| 2015 | 98,087 | 1247.88 | 49,047 | 627.93 | 49,040 | 619.94 |
| 2016 | 100,197 | 1324.57 | 50,089 | 666.04 | 50,108 | 658.53 |
| 2017 | 99,687 | 1332.26 | 49,842 | 669.14 | 49,845 | 663.12 |
| 2018 | 94,816 | 1345.18 | 47,345 | 676.84 | 47,471 | 668.34 |
| 2019 | 93,701 | 1361.34 | 46,834 | 684.98 | 46,867 | 676.36 |
| 2020 | 89,018 | 1324.71 | 44,430 | 665.69 | 44,588 | 659.02 |
| 2021 | 46,704 | 706.16 | 23,186 | 354.58 | 23,518 | 351.58 |

Table 1 Cargo throughput and ship traffic at Busan Port (cargo throughput in million tons, including exports, imports and transshipment) Source Korean Shipping Port Logistics Information System, 2021
phase space, an infinitesimal $n$-sphere of initial conditions becomes an $n$-ellipsoid in the long term. Thus, the $i$th LE index is defined as follows:

$$\lambda_i = \lim_{t \to \infty} \frac{1}{t} \log_2 \frac{p_i(t)}{p_i(0)},$$  

(1)

where $p_i(t)$ is the length of the ellipsoidal principal axis at time $t$ and the $\lambda_i$s are typically ordered from the largest to smallest (Wolf et al. 1985). A system is considered to be chaotic if a minimum of one LE or the largest LE (LLE) is positive; otherwise, its dynamic behavior is considered to be stable, periodic, or quasiperiodic (Hwarng and Xie 2008). Chaotic behaviors tend to be very complex in a random manner, although they change according to deterministic rules.

Figure 1 shows the LLE of Busan’s cargo throughput. The application of Wolf et al. (1985) theory reveals that LLE exhibited sequential negative and positive values throughout the development of Busan Port since in 2001. This means that the system exhibits an unstable, complex behavior, tending to be chaotic. These behaviors are caused by transportation disruptions, such as seasonal demand variations and external shocks, particularly the 2009 financial crisis and the COVID-19 pandemic. Moreover, test results showed that LLE fluctuated with a large amplitude at several periods in the time series. Fluctuations during these periods are much larger than those of an average period of the entire system. For example, the value of LLE was $-2.8$ at $t=40$, compared with the average range of the entire system, which was approximately $[-1 2]$. Owing to the many uncertainties about the drastic changes in LLE, the existence of large fluctuations requires further clarification. This study offers alternative algorithms, which are more robust than the LE scheme, for identifying specific dynamic behaviors (Clément and Laurens 2011). Furthermore, other alternatives must be explored to verify the dynamic behavior of port throughput under the impact of disruptions.

2.3 Resolving complexity via entropy analysis

Along with LEs, entropic measurement offers a deeper elucidation of the nonlinear features of systems (Jahanshahi et al. 2019). Information entropy is a concept in information theory, employed to measure the complexity and randomness of a time series. The entropy of a dynamic system is determined by the predictability of its dynamic behavior. For example, more-complex systems are less predictable.
than less complex ones. The Kolmogorov–Sinai entropy is a complicated scheme for computing entropy to measure complexity and randomness from a finite time series. Several methods, such as approximate entropy and sample entropy, have been introduced to address this limitation. However, these approaches have suffered from weak principles (Ye et al. 2020). Contrarily, the permutation entropy (PE, $H_{NPE}$) represents a robust tool that characterizes the complexity and randomness of a nonlinear time series (Bandt and Pompe, 2002). Our data reveals that the complexity measure is only a function of the probabilities of different states (the computational steps for measuring $H_{NPE}$ are summarized in Appendix 1).

Additionally, the measured entropy value, $H_{NPE}$, is always between 0 and 1. Smaller entropy offers more periodicity and regularity in the time series. However, the larger the entropy value, the more irregular and random the time series would be. More specifically, if the time series is white noise, the entropy will be 1. Therefore, PE can be applied to the estimation of the complexity and dynamic changes in the time series. The measures of the complexity in maritime supply chains relate to their manageability and controllability. Thus, complexity reduction might enable port authorities to simplify their strategic goals for business operations against disruptions.

To investigate the stability and disruption of port throughput, the PE method is employed to clarify the dynamic behaviors of throughput trends. The values of training and test samples in the dataset were 197 and 50 data points, respectively, while the sample size, N, was 247. The embedding dimensions and scale factor were set at $m=4$ and $s=12$, respectively. Figure 2 shows that port throughput can be described by employing the basic properties of information entropy. First, the entropy value is zero at $t=0$ because the variable is not random. Next, $H_{NPE}(m)$ increases to 0.78 at $t=5$ and remains, mostly sideways, for a few subsequent time series before increasing rapidly. From $t=38$, PE reaches a value of 1 and maintains that pattern until the end of the period. The test results are employed to describe the historical development of the port, operated by the Busan Port Authority (BPA). Further, internal factors, such as operating procedures, management policies, and constantly changing infrastructure, cause significant fluctuation in cargo throughput trends. Additionally, external shocks, particularly disruptions, such as the 2009 financial crisis and Covid-19, directly disturb the management of port business and operations. Thus, information entropy can be a measure of the degree of systematic ordering or randomness of the time series.

**Fig. 2** Temporal variations of the entropy of cargo traffic (Busan Port)
Next, $H_{NPE}$ is employed to investigate the cargo throughput trend for the preceding three years (2018–2020). The test results of the short-term performance are shown in Fig. 3. At the beginning of the test period, the entropy attained low values yearly. Toward the end of the period, it increased significantly and approached 1 in 2020. However, these values were relatively low in 2018 and 2019. Especially, a high entropy value was obtained in 2020 because the port encountered supply chain disruptions in its cargo handling activities, as well as shipping services, owing to the impact of COVID-19; these affected the dynamic behavior of port operations. External shocks have affected cargo throughput, accounting for increasing irregular behaviors. Thus, the entropy value is ~ 1. Concurrently, historical data revealed that the port management system achieved smooth operations during the earlier two years (2018–2019).

### 2.4 Determining the predictability via HE and fractal dimension

The port management system encounters complex, dynamic, and increasingly challenging disruptive perturbations, such as time delays, regulatory changes, adverse weather, and dynamic interactions between port systems. By employing the introduced approaches, port authorities can identify the quantitative behaviors of nonlinearity, complexity, and chaos of the time series. HE is a robust and efficient tool in exploring the underlying mechanisms, having been applied to time series analysis in finance and logistics research (Corazza and Malliaris 2002; Tzouras et al. 2015; Yao et al. 2014).

HE, proposed by Hurst (1951), is employed in fractal analysis and it is a measure of long-term memory of time-series data. The method can be employed to determine the duration of the impacts of external shocks. The advantages of HE include the simplicity of the basic assumptions, as well as the increased stability of its validity and robustness for numerous types of time series (Corazza and Malliaris 2002). HE ($H$) is determined via the asymptotic dynamic behavior of the rescaled range as a function of the period (Feder 1988). Employing the collected data, $H$ is calculated via rescaled range analysis (R/S analysis). Regarding the time-series data, $X = X_1, X_2, ... X_N$, the R/S analysis steps are presented in Appendix 2.

The value of $H$ ranges between 0 and 1, measuring the degree of the deviation of a given time series from a random walk ($H=0.5$), with higher values indicating a smoother trend, less volatility, and less roughness. A time series can be classified into the following categories: $H=0.5$, representing a random distribution that

![Fig. 3](image)

**Fig. 3** Comparison of the entropies of Busan Port within 2018–2020
is indistinguishable from noise or disruption; $0 < H < 0.5$, which means that the system exhibits mean-reverting and anti-persistent series behaviors (more chaos); and $0.5 < H < 1.0$, which indicates that the system tends to a persistent series (less chaotic or trending). Values near 0 and 1 indicate that the mean-reversion process and trending pattern (the trend will continue) were strong, respectively. The $H$ estimator is a rapid tool for investigating whether the time series is random walking, mean-reverting, or trending.

The complexity of time-series signals was investigated by fractal analysis. Regarding a self-similar time series, $H$ directly relates to the pattern fractal dimension ($D$), where $1 < D < 2$, such that $D = 2 - H$ (Liaw and Chiu 2009). $D$ is a relative measure of the number of elemental units employed to fabricate a pattern. Notably, the rough pattern is represented by a higher $D$, with a fractal measure ranging from 1 (the dimension along the straight line) to 2 (the dimension of the plane in which the pattern lies). Further, the fractal dimension can also be regarded as an informational measure, which can be employed to directly observe the time series and investigate its nonlinear features. Assuming $D = 1.5$ (or $H = 0.5$) obeys a random walk, with any $D$ value other than 1.5, time-series data can be inferred as the absence of randomness.

Figure 4 presents $H$ for a sliding window exhibiting a data length of 147. Overall, it is significantly higher than 0.5 (or $D < 1.5$) when analyzing the cargo volume of Busan Port, revealing the distinct long-term memory feature in the loading and unloading operations. Results reveal the following information: (1) port throughput is affected by the shipping market and external disruptions, as well as previous and present throughput levels. (2) $H$ is higher than 0.5 under prolonged time series conditions, with many environmental shocks, such as the financial crisis or the pandemic, indicating that the trending pattern was strong on the throughput volume and

![Fig. 4](image-url)  

**Fig. 4** $H$ over a sliding window of 145 observations (from May 2009 to July 2021). **a** R/S ratio; **b** $H$
that the trend will persist (Tzouras et al. 2015; Ye et al. 2020). If the actual cargo volume is lower than the average at a certain point, the throughput will tend to be lower at the subsequent time. Conversely, if the cargo throughput is higher than the average, the cargo volume will tend to increase over time. (3) the port authority could combine the previous interevent time distribution with the present situation and employ analytical and statistical tools to improve management efficiency towards increasing port throughput.

2.5 Port productivity analysis using wavelet decomposition

As demonstrated, the cargo throughput of Busan Port is highly volatile, presenting chaotic behavior. As such, forecasting is particularly challenging, for unstable systems are affected by uncertainties and external shocks (Xie et al. 2017). Such predictions cannot be accurate even if they are based on actual numbers. Further, most data are commonly messy or incomplete.

Thus, the wavelet transform is introduced next, to improve forecasting accuracy. So far, this technique has presented a very efficient time-series analysis tool, not very common within the data science community. Wavelet transforms decompose a signal into several sub-signals of different frequency bands. Previous signal-processing techniques, such as the Fourier transform (FT) and short time FT (STFT), exhibit major drawbacks regarding time information and resolution. The utilization of a fully scalable modulated window in the wavelet analysis can solve the above drawbacks. The window or mother wavelet is shifted at every position of the signal, making it possible to calculate the spectrum at each position (Rajaee et al. 2011). There are two basic types of wavelet transformations: the continuous wavelet transform (CWT) and discrete wavelet transform (DWT). CWT is defined in terms of the dilations and translations of a mother wavelet function that can be expressed, as follows (Memarzadeh and Keynia 2020):

\[ W_{\psi}f(a, b) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(t) \overline{\psi} \left( \frac{t-b}{a} \right) dt \]  

(2)

where the transformed signal is a function of two variables (‘a’ and ‘b’), which represent the scale and translation parameters, respectively; \( \overline{\psi} \) denotes the complex conjugate of ‘t’ (Mallat 1989); \( f(t) \) is the input signal; \( \psi(t) \) is the transforming function, which is called the mother wavelet. Since CWT produces \( N^2 \) coefficients from a dataset of length, \( N \), unnecessary information is hidden within the coefficients (Rajaee et al. 2011). To overcome this limitation, DWT calculates the wavelet coefficients on discrete dyadic scales and time positions, as follows:

\[ \psi_{m,n}(t) = \frac{1}{\sqrt{a_0^n}} \psi \left( \frac{t - b_0 a_0^m}{a_0^m} \right) \]  

(3)

where \( m \) and \( n \) are the integers that regulate the wavelet dilation and translation, respectively; \( b_0 \) is the location parameter, which must be greater than zero (> 0); \( a_0 \) is a specified fixed dilation step, which must be greater than one (> 1). The ideal
choices of $a_0$ and $b_0$ depend on the wavelet functions. The dyadic wavelet can be written in more compact notation, as follows:

$$\psi_{m,n}(t) = 2^{-\frac{m}{2}} \psi(2^{-m}t - n)$$  \hspace{1cm} (4)

For a discrete time series, $x_t$, the dyadic wavelet transform becomes

$$W_{m,n}(t) = m^{-\frac{m}{2}} \sum_{i=0}^{N-1} \psi(2^{-m}t - n)x_i$$  \hspace{1cm} (5)

where $W_{m,n}$ is the wavelet coefficient for the discrete wavelet of scale $a=2^m$ and location $b=2^{mn}$. This formula considers a finite time series, $x_t$, $t=0, 1, 2, \ldots, N-1$, and $N$ is an integer power of 2, $N=2^M$, and $n$ is the time-translation parameter. This gives the ranges of $m$ and $n$ as $0 < n < 2^{M-m-1}$ and $1 < m < M$, respectively. Generally, the time-series patterns of the port throughput exhibit nonlinear and dynamic properties. The trend of the cargo throughput is severely impacted by several factors, such as market changes, shipping costs, and disruptive factors. Next, DWT can be employed to describe the asymmetric nature of this time series. Put differently, DWT can extract the different characteristics of the dataset, such as the low- and high-frequency components.

Here, the Haar wavelet was employed as the wavelet filter bank. This wavelet divides the input data, $f$, into three distinct levels ($a_2$, $d_1$, and $d_2$), where $a$ and $d$ represent the approximation and detail, respectively ($f=a_2+d_2+d_1$). The discontinuous wavelet transform, which was obtained by denoising the 1D Level 2 for the Busan port system, is shown in Fig. 5. Following the Haar wavelet of the cargo throughput trend, a significant difference was observed between the Level 1 and 2 results. Although the corresponding Level 2 of the low-frequency component ($a_2$) represents the long-term trend of the port throughput, shown in Fig. 5b, the high-frequency components ($d_1$ and $d_2$) represent the short-term volatility trends. These trends exhibit a high frequency and periodic volatility (Fig. 5c, d).

Figure 6 shows the wavelet power spectra, illustrating the significant advantages of wavelet analysis over spectral analysis. The horizontal and vertical axes represent the time dimension and period, respectively. The color code indicates the range of the power spectra from blue (low power) to red (high power). It can be observed that the high power regions were distributed on periods of 27 to 45 across the entire vertical axis, beginning from 2002. The high-energy regions were widely and cyclically scattered, indicating a strong movement of cargo throughput over the entire developmental periods. This strong volatility resulted from the port development process, owing to the influences of many uncertain factors. The periodicity of the high-energy regions was due to seasonal demand, represented by the high-frequency component ($d_1$ and $d_2$) (González-Concepción et al. 2012).
3 Novel forecasting scheme employing the DWT–LSTM hybrid model

Conventional management mechanisms are generally designed to ensure stability, efficiency, and predictable performance. The cargo throughput of Busan Port is a complex dynamic system, open to environmental influences and susceptible to changing circumstances. These disruptions greatly impact short- and long-term development strategies. Therefore, a robust forecasting method, which decision-makers can rely on to improve operational efficiency, must be developed to guarantee the sustainability and profitability of the port. Machine learning algorithms can identify data patterns, learning from available data or observations. LSTM networks have recently demonstrated high effectiveness in time-series forecasting (Fischer and Krauss 2018; Punia et al. 2020). This machine learning scheme works well on linear and nonlinear time series behaviors (Punia et al. 2020). In this study, the cargo throughput data contain several patterns over volatile periods of international
shipping, suggesting the LSTM strategy as a suitable alternative. In addition to the deep learning approach, wavelet transform can effectively convert cargo throughput datasets into low- and high-frequency series, and the decomposed series can be employed to forecast cargo throughput via the LSTM network (Chang et al. 2019). The extant experiments have demonstrated that the wavelet transform can efficiently support time-series forecasting (Schlüter and Deuschle 2010). In this study, a novel hybrid strategy, comprising DWT and the LSTM network, is proposed for cargo throughput forecasting.

### 3.1 LSTM model

LSTM networks represent a class of RNNs. RNN evaluates the relationship between the data entered in the past and future; it can excellently facilitate the analysis of continuous arrays of data. RNN is trained employing a backpropagation algorithm, after which it learns via the gradient descent method (Medsker and Jain 1999). Contrary to normal neural networks, RNN exhibits a recurrent learning structure and circulates input data according to the layers. This structure
avails it with the memory of past information and updates it with the latest data. Figure 7 shows the memory cell of the LSTM structure.

LSTM networks comprise input, several hidden, and output layers. The number of neurons in the input layer is equal to the number of explanatory variables (the feature space). The number of neurons in the output layer reflects the output space. The two neurons in this study determine whether a stock outperforms the cross-sectional median at $t+1$ or not. The main characteristics of the LSTM networks are contained in the hidden layer(s) consisting of the so-called memory cells (Graves and Schmidhuber 2005). Each memory cell exhibits the following three gates that maintain and adjust its cell state, $s_t$: a forget gate ($f_t$), an input gate ($i_t$), and an output gate ($o_t$). At every time step, $t$, each of the three gates is presented with the input, $x_t$ (one element of the input sequence), and output, $h_{t-1}$, of the memory cells at the previous time step, $t-1$. Thereafter, the gates act as filters, each fulfilling one of the purposes:

- Forget gate defines the information to extract from the cell state.
- Input gate specifies the information to add to the cell state.
- Output gate specifies the information from the cell state that is utilized as output.

The formulas below are vectorized; they describe the update of the memory cells in the LSTM layer at every time step, $t$. The following notation is employed:

- $x_t$ is the input vector at time step, $t$.
- $W_{f,x}, W_{i,h}, W_{\hat{s},x}, W_{\tilde{s},h}, W_{i,x}, W_{i,h}, W_{o,x},$ and $W_{o,h}$ are weight matrices.
- $b_f, b_s, b_i$ and $b_o$ are bias vectors.
- $f_t, i_t$, and $o_t$ are the vectors of the activation values of the respective gates.
- $s_t$ and $\hat{s}_t$ are the vectors of the cell states and candidate values.
- $h_t$ is a vector for the layer output.
During a forward pass, the cell states, \(s_t\), and outputs, \(h_t\), of the layer at a time step, \(t\), are calculated, as follows: in the first step, the LSTM layer determines the information that must be extracted from its previous cell states, \(s_{t-1}\). Therefore, the activation values, \(f_t\), of the forget gates at \(t\) are computed employing the current input, \(x_t\); outputs, \(h_{t-1}\), of the memory cells at the previous \((t-1)\); and bias terms, \(b_f\), of the forget gates. The sigmoid function scales all the activation values in the range between 0 (completely forget) and 1 (completely remember):

\[
f_t = \text{sigmoid}(W_f x_t + W_{f,h} h_{t-1} + b_f)
\]

(6)

In the second step, the LSTM layer determines the information that must be added to the network cell states (\(s_t\)). This procedure comprises two operations. First, the candidate values (\(\tilde{s}_t\)), which could be added to the cell states, are computed. Second, the activation values \(i_t\) of the input gates are calculated as follows:

\[
\tilde{s}_t = \tanh(W_{i,x} x_t + W_{i,h} h_{t-1} + b_i)
\]

(7)

\[
i_t = \text{sigmoid}(W_{i,x} x_t + W_{i,h} h_{t-1} + b_i)
\]

(8)

In the third step, the new cell state (\(s_t\)) is calculated from the results of the two preceding steps, as follows:

\[
s_t = f_t \circ s_{t-1} + i_t \circ \tilde{s}_t
\]

(9)

where the symbol \((\circ)\) denotes the Hadamard (element-wise) product. In the final step, the outputs, \(h_t\), of the memory cells are derived, as follows:

\[
o_t = \text{sigmoid}(W_{o,x} x_t + W_{o,h} h_t + b_o)
\]

(10)

\[
h_t = o_t \ast \tanh(s_t)
\]

(11)

When processing an input sequence, its features must be presented timestep by timestep to the LSTM network. Thus, the input at each timestep, \(t\) (in this study, one single standardized return), is processed by the network. The final output of the entire sequence is returned immediately after the last element of the sequence is processed.

### 3.2 DWT–LSTM hybrid model

To enhance the degree of the accuracy of predicting cargo throughput, a novel hybrid methodology is proposed, based on the DWT–LSTM model. The proposed approach can improve the existing forecasting method to cope with highly complex and chaotic trends. Figure 8 shows the overall process of the proposed methodology, which can be summarized in the following steps:

**Step (1).** Data decomposition: DWT decomposes the original cargo throughput into \(n\) components.
Step (2). Data characteristic analysis (DCA): all the decomposed components are thoroughly analyzed via the wavelet power spectra to determine the data characteristics.

Step (3). Individual predictions: based on DCA, one LSTM model was selected for the prediction of each component. Accordingly, the forecasting result of the corresponding component was obtained.

Step (4). Ensemble prediction: the predicted results of the components were combined with an aggregated cargo throughput.

Step (5). Performance evaluation: data analytics, as well as statistical analyses, were performed to assess the prediction results.

The details of this scheme are also described in Algorithm 1, Appendix 3.

4 Empirical test results

4.1 Experimental setup and evaluation of the forecasting models

The future developments of a port can be predicted from trends, patterns, and current and historical throughput analysis. The time series employed in this study comprises data collected from cargo volumes of Busan Port. Cargo throughput is expressed as million tons (January 2001 to July 2021), and it is also available in the Korean PORT-MIS. Our monthly data consisted of 247 data points of import, export, and transshipment cargoes (Table 2). Typically, when constructing a model, the sample dataset is
split into training and testing sets. Good data splitting will increase the robustness of the forecasting model and prevent overfitting or underfitting. In this study, the training and testing set utilized the same datasets for all forecasting methods. The training data were employed to teach an algorithm and generate forecasts for the data points, including 198 data points, within the testing set, covering the period from January 2001 to May 2017. The forecast evaluates the accuracy of the model by comparing the forecasted values with the observed values of the testing set (June 2017–July 2021); the more accurate the training data labels, the better the model will perform.

The following four criteria are frequently employed to evaluate forecasting models: the mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and root mean squared logarithmic error (RMSLE). MAE estimates the absolute difference between the actual and forecasted values. It can handle small actual values. The RMSE metric is obtained as the square root of MSE. This indicator is more sensitive to outliers than the mean absolute percentage error (MAPE). However, small actual values might produce misleadingly large MAPE values, whereas actual values close to zero can generate infinitely large MAPEs. RMSLE considers the relative error between the predicted and actual values, neglecting the scale of the data.

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2
\]  
(12)

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|
\]  
(13)

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}
\]  
(14)

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]  
(15)

| Year | Total ships | Entry into port | Departure |
|------|-------------|-----------------|-----------|
|      | Number of ships | Cargo volume | Number of ships | Cargo volume | Number of ships | Cargo volume |
| Sum  | 2,007,942 | 20,228.19 | 1,003,573 | 10,146.05 | 1,004,369 | 10,082.14 |
| Mean | 95,616.29 | 963.2473 | 47,789.19 | 483.1453 | 47,827.1 | 480.102 |
| SD   | 12,869.1 | 284.3913 | 6466.564 | 143.9262 | 6402.67 | 140.4722 |
| Minimum | 46,704 | 541.141 | 23,186 | 270.3985 | 23,518 | 270.7425 |
| Maximum | 115,931 | 1361.337 | 57,979 | 684.9822 | 57,952 | 676.3552 |

Table 2 Dataset statistics for cargo throughput (cargo volume columns in million tons)
\[ RMSLE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\log(\hat{y}_i + 1)^2 - \log(y_i + 1)^2)} \]  

(16)

where \( y_i \) and \( \hat{y}_i \) are the actual and forecasted values, respectively, and \( N \) is the sample size. A lower value indicates a more accurate model. To further compare performances, the Diebold–Mariano (DM) test was used to investigate the difference between the significance of forecasting performances (Diebold and Mariano 1995). Additionally, the Pesaran–Timmermann (PT) test was employed to evaluate the null hypotheses whereby the prediction is independently distributed for each forecasting method. The \( p \)-values of zero up to the fourth digit indicated that the null hypothesis can be rejected at any rational significance level (Pesaran and Timmermann 1995). Put differently, each machine learning method exhibits statistically significant predictive accuracy. In this study, the DM and PT tests were employed to evaluate the effectiveness of the proposed forecasting methods.

### 4.2 Forecasting results employing the hybrid algorithm.

Employing the hybrid prediction method, the original time series of cargo throughput was decomposed by the DWT technique. As discussed in Sect. 2.5, cargo throughput demonstrated highly complex dynamic behaviors. Figure 9 shows the managerial insights into the forecasting results that were obtained by our models. Figure 9a shows that the DWT–LSTM-based hybrid method produces better forecasting performance than single deep learning methods, such as RNN. The hybrid learning methods describe several details regarding the low-frequency (Fig. 9b) and high-frequency (Fig. 9c, d) components. Particularly, the test results indicated that the new hybrid approach produces the best performance: forecasting data accurately converged to their actual datasets. Although the RNN-based forecasting results showed that the model could not predict the dynamics of cargo throughput comprising high frequency components (\( d_1 \) and \( d_2 \)), it availed underfitting prediction and larger prediction errors, \( a_2 \), compared with the hybrid forecasting scheme. The average forecasting errors from each forecast are presented in Table 3. Based on all the performance indices, the hybrid forecasting strategy outperformed the existing benchmark methods.

### 4.3 Comparison of forecasting performance

To evaluate the effectiveness of our forecasting models, different tests were utilized, employing the hybrid model, plus other benchmark algorithms, including seasonal ARIMA (SARIMA), RNN, LSTM, and a gated recurrent unit (GRU). Many factors determine how much training data are required. More-complicated cases generally require more data than less-complex ones. Figure 10 shows the comparisons when the relative ratio of the training sample size is 80%. Notably, all the deep learning methods (RNN, LSTM, GRU, and the hybrid method) exhibited better forecasting patterns than the SARIMA model. The deep learning algorithms predicted the peak
throughputs caused by disruptions. Moreover, the proposed algorithm delivered the best performance over the others, whereby the time series data converged completely to the actual data.

GRU and LSTM were the second- and third-best algorithms, respectively. The average relative forecasting errors are presented in Table 4 for model comparison. In this experiment, different ratios (60%, 70%, and 80%) of the training dataset to the sample size were employed. Notably, the bold font in the table indicates the best forecasting performance among the models. The test results showed that DWT–LSTM was the most accurate scheme. The proposed hybrid method

![Figure 9](image_url)

**Fig. 9** Comparison of the models for forecasting the cargo throughput at Busan Port. a $a_2$, b $d_1$, and c $d_2$

| $ij$ | $a_1$ | $a_2$ | $d_1$ | $d_2$ |
|------|-------|-------|-------|-------|
| RNN  | MAE   | 11.9771 | 2.6229 | 3.4607 | 0.0240 | 0.0314 |
|      | MSE   | 0.8079  | 0.3075 | 0.8988 | 0.0029 | 0.0085 |
| Proposed | MAPE | 5.4544 | 1.9741 | 3.5333 | 0.0181 | 0.0415 |
|      | RMSLE | 0.0713  | 0.1508 | 0.2671 | 0.0696 | 0.0175 |
|      |      | 12.4843 | 2.6452 | 3.5333 | 1.1680 | 0.0173 |
|      |      | 0.2848  | 0.2534 | 0.5337 | 0.2062 | 0.0073 |

Table 3 Performance comparison of forecasting algorithms (the smaller, the better)
outperformed the benchmark methods based on all five performance criteria. Particularly, when the relative ratio of the training sample size was set to 70% and 80%, the proposed hybrid algorithm delivered the best forecasts of cargo throughput. GRU, LSTM and RNN delivered the second-, third-, and fourth-best performances, respectively; SARIMA delivered the worst performance. However, when the training sample size was set to 60%, LSTM outperformed GRU, delivering the
second-best performance. GRU was the third best, and the other ranking orders remained the same.

Moreover, for the empirical evaluation, statistical significance tests were conducted, employing the DM and PT tests. The monthly predictions of the training set at different ratios (60%, 70%, and 80%) are shown in Tables 5, 6, and 7, respectively. The test results show that the proposed method always outperforms the other benchmark methods, and the performance differences are statistically significant at the 95% confidence level. Panels A in Tables 5, 6, and 7 present the \( p \)-values for the null hypothesis that the paired methods produce equal performances. The \( p \)-values present the confidence level that method \( i \) exhibits an inferior prediction accuracy to that of method \( j \). All hypotheses were rejected at the 95% significance level, indicating that the proposed method is superior to all the benchmark methods at a 95% confidence level.

### Table 4 Forecasting model evaluation with different ratios of training sample sizes (the smaller, the better with bold indicating the top performer)

| Relative ratio | Performance indicator | MAE   | MSE   | RMSE  | MAPE  | RMSLE |
|----------------|-----------------------|-------|-------|-------|-------|-------|
| 60%            | SARIMA                | 0.3523| 1.0532| 1.4577| 1.3306| 0.3747|
|                | RNN                   | 0.1358| 0.9376| 1.5508| 0.9334| 0.7447|
|                | LSTM                  | 0.0631| 0.1221| 0.2907| 0.2491| 0.2731|
|                | GRU                   | 0.0785| 0.2218| 0.3158| 0.2140| 0.2907|
|                | DWT-LSTM              | 0.0112| 0.0168| 0.0221| 0.0286| 0.0303|
| 70%            | SARIMA                | 0.2216| 0.4100| 0.4862| 0.4224| 0.5219|
|                | RNN                   | 0.0839| 0.1843| 0.1052| 0.1227| 0.9535|
|                | LSTM                  | 0.0573| 0.0942| 0.1091| 0.1328| 0.1030|
|                | GRU                   | 0.04281| 0.1066| 0.1003| 0.1207| 0.1005|
|                | DWT-LSTM              | 0.0066| 0.0097| 0.0113| 0.0119| 0.0101|
| 80%            | SARIMA                | 0.1144| 0.5217| 0.7875| 0.6024| 0.2247|
|                | RNN                   | 0.0131| 0.0694| 0.1037| 0.0352| 0.0418|
|                | LSTM                  | 0.0232| 0.2011| 0.0934| 0.0521| 0.0442|
|                | GRU                   | 0.0291| 0.3442| 0.2297| 0.1147| 0.3751|
|                | DWT-LSTM              | 0.0007| 0.0011| 0.0109| 0.0114| 0.0231|

### Table 5 Statistical significance tests for port throughput when the relative ratio of the training sample size is 60% (panel A: DM test, panel B: PT test)

\[ i \neq j = \begin{cases} 
  \text{A: DM test (} p \text{-value)} \\
  \text{B: PT test (} p \text{-value)} 
\end{cases} \]

| i j | Proposed | SARIMA | RNN | LSTM | GRU | Method | Results |
|-----|----------|--------|-----|------|-----|--------|---------|
| Proposed | –        | 0.0206 | 0.0003| 0.0020| 0.0001| Proposed | 0.0000 |
| SARIMA | –        | 0.0033 | 0.0006| 0.0007|       | SARIMA  | 0.0000 |
| RNN   | –        | 0.0010 | 0.0000|       |       | RNN     | 0.0000 |
| LSTM  | –        | 0.0000 |       |       |       | LSTM    | 0.0000 |
| GRU   | –        |        |       |       |       | GRU     | 0.0000 |
ones regarding their forecasting performance. Panel B of each table presents the $p$-values of the PT test for the null hypothesis that the actual data and forecast values are independently distributed. It can be concluded with a high degree of certainty that the predictions from the forecasting models are independently distributed. All test results demonstrated that the proposed hybrid algorithm outperforms the other algorithms in forecasting cargo throughputs in a dynamic environment.

### 4.4 Managerial insights and discussion

The collection and analysis of data can be expensive and time-consuming. The data analytics techniques presented here (LE, information entropy, HE, wavelet decomposition, statistical significance with DM, and the PT evaluation) are comprehensive and effective in analyzing nonlinear dynamical behaviors, as well as identifying business disruptions. Researchers have recently improved many business forecasting techniques, but have invested comparatively little in assessing prediction accuracies and efficiencies. However, biased forecasts can increase logistics costs (transportation, carrying, or warehousing costs), thus reducing profit margins.

Policymakers can employ different forecasting models to arrive at informed predictions. This study involves in-depth research on forecasting port throughput, showing at the same time our methods’ functionality and success. Forecasting algorithms,
based on the hybrid deep learning method and four benchmark models, were presented to predict the cargo throughput of Busan Port. Holistically, the five metrics, as well as two statistical tests (DM and PT), demonstrated that the hybrid predictive approach outperforms the other benchmark algorithms with regard to all characteristics, ensuring higher forecasting accuracies and efficiencies. Our study was designed to confirm the superiority of deep learning models, such as GRU, LSTM, and RNN, over the traditional SARIMA models.

Moreover, for larger training sample sizes (70% and 80%), GRU outperformed LSTM, whereas LSTM was well suited for smaller training sample sizes. Indeed, for a training sample size of 60%, LSTM demonstrated a higher forecasting accuracy than GRU. It was also highlighted that the hybrid method outperforms LSTM and GRU in forecasting the time series. From the foregoing performance analyses, the following deep insights into cargo throughput can be mentioned: DWT represents a crucial pre-step for filtering the time series of cargo throughput data. The wavelet transform via deep learning outperformed the corresponding deep learning method in the absence of data decomposition. Particularly, the cargo throughput data exhibited highly complex and nonlinear behaviors in the presence of disruptions, including low- and high-frequency components. Although the traditional algorithms exhibited limitations in time series forecasting, the DWT methodology was efficient for decomposing the time-series data into different frequencies by filtering the data as a pre-treatment process. Thereafter, the prediction mechanisms based on deep learning models enhanced significantly the prediction accuracy and efficiency of each component to achieve the forecasting goals. The benefits of effective forecasting include improved business performances, gaining competitive advantage, offering rapid response to cyclical or seasonal change, identifying disruptions, and facilitating port strategic planning. Additionally, policymakers can exploit the novel forecasting model to keep their customers happy by delivering services and products while achieving strategic goals with profits in volatile markets.

5 Conclusions

We introduced the data analytics and forecasting algorithms for cargo throughput forecasts at the port of Busan. First, time-series data analysis was conducted, employing nonlinear analysis tools. Specifically, dynamic analysis tools for exploring the underlying mechanisms of maritime logistics under the impact of market volatility included LE, information entropy, HE, and wavelet decomposition. According to the system behavior revealed by the nonlinear paradigms, the cargo throughput dynamics of the port exhibited highly complex and chaotic behaviors against uncertainties and external disruptions. These drastic throughput changes were mainly caused by market disruptions, such as seasonal demand, the 2009 global financial crisis, and the ongoing COVID-19 pandemic. The results can help port managers understand the nature of complexities and the dynamics of port operations.

Next, a robust forecasting method was proposed, employing the wavelet transform and deep learning methods. The proposed approach was compared with other benchmark methods against supply chain disruptions. Three different ratios (60%,
70%, and 80%) were employed for the training set samples. For a forecasting benchmark, different common methods, including SARIMA, RNN, LSTM, and GRU, were employed in the tests. Further, five performance metrics, comprising MSE, MAE, MAPE, RMSE, and RMSLE, were employed to evaluate the algorithms. The DM and PT statistical significance tests were used to confirm the key empirical findings. The evaluation results indicated that the proposed method outperforms all other benchmarks that are widely employed in the data science community. Our hybrid strategy combined the DWT and LSTM methods (DWT–LSTM), enhancing the degree of prediction accuracy and efficiency by simply denoising cargo throughput data. This method was optimized by combining the advantages of DWT and LSTM to obtain the target outcomes. Another finding is that predictions employing GRU are more accurate than those employing LSTM in cases involving large training datasets. On the other hand, LSTM was superior to GRU for smaller training datasets.

To summarize, our study contributes to the theories and practices of data analytics and business forecasting. Our first contribution is the proposal of powerful tools for investigating nonlinear datasets and complex dynamics. This can help policymakers gain greater insight into the dynamic behavior of seaport operations as well as explore disruption mechanisms. Second, our study has presented a robust and flexible forecasting method, which can be employed to predict complex maritime logistics patterns for achieving business goals. Therefore, these findings, reached by employing the proposed methods, can help decision-makers to prepare better for future port operations against maritime disruptions.

Appendix 1

For a time series length, the \( m \)-dimensional vector at the time \( i \) can be given by

\[
X^m_i = \{x(i), x(i + \tau), ..., x(i + (m-1)\tau)\}, i = 1, 2, ..., N - (m - 1)\tau
\]

(17)

where \( X^m_i \) denotes data in new time series, \( m \) is the embedding dimension and \( \tau \) is the time delay. The \( X^m_i \) has a permutation \( \pi_{r_0, r_1, ..., r_{m-1}} \), when \( X^m_i \) satisfies the following condition:

\[
x(t + r_0\tau) \leq x(t + r_1\tau) \leq ... \leq x(t + r_{m-1}\tau)
\]

(18)

where \( 0 \leq r_i \leq m - 1 \) and \( r_i \neq r_j \). It can be obtained from Eq. (18) that an \( m \)-tuple vector has \( m! \) possible distributions. Moreover, the calculation of frequency of distribution is given as

\[
P(i) = \frac{\text{Num}\{X^m_i\}}{N - (m - 1)\tau}
\]

(19)

where \( \text{Num}\{X^m_i\} \) represents the number of \( X^m_i \), and it is consistent with the type \( \pi \). Then, the PE with \( m \)-dimensions can be described as
It is noted that when $P(i) = \frac{1}{m}$, $H_{PE}$ is reaching the maximum value $\ln (m!)$. The PE measure can be normalized through $\ln (m!)$ as

$$H_{NPE}(m) = \frac{H_{PE}(m)}{\ln (m!)} \quad (21)$$

### Appendix 2

The Hurst exponent analysis steps are given as follows:

**Determine the mean**

$$\mu = \frac{1}{N} \sum_{i=1}^{N} X_i \quad (22)$$

**Evaluate the adjusted mean series $Y$,**

$$Y_t = X_t - \mu, \quad t = 1, 2, ... N \quad (23)$$

**Calculate the cumulative deviation $Z$,**

$$Z_t = \sum_{i=1}^{N} Y_i, \quad t = 1, 2, ... N \quad (24)$$

**Determine the range series $R$,**

$$R_t = \max(Z_1, Z_2, ... Z_t) - \min(Z_1, Z_2, ... Z_t), \quad t = 1, 2, ... N \quad (25)$$

**Compute the standard deviation $S$,**

$$S_t = \sqrt{\frac{1}{t} \sum_{i=1}^{t} (X_i - u)^2}, \quad t = 1, 2, ... N, \quad (26)$$

where $u$ is the mean value from $X_1$ to $X_t$. Next, the rescaled range can be given as

$$(R/S)_t = \frac{R_t}{S_t}, \quad t = 1, 2, ... N \quad (27)$$

Note that $(R/S)_t$ is the averaged over the regions $[X_{t'}, X_t]$, $[X_{t+1'}, X_{2t}]$ until $[X_{(m-1)t+1'}, X_{mt}]$, where $m = \lfloor N/t \rfloor$. To use all data for calculations, a value of $t$ is chosen such that it is divisible by $N$. The relationship between $(R/S)$, and the Hurst exponent ($H$) is calculated by
\[(R/S)_t = c_a t^H\] (28)

where \(c_a\) is a constant. To estimate the Hurst exponent, \((R/S)\) versus \(t\) in log–log axes can be plotted. The slope of the curve is estimated using the ordinary least squares (OLS) scheme.

### Appendix 3

Algorithm 1: Forecasting based on hybrid algorithm

**Input:** Data input \(x_t\), training data length \(L_y\), A LSTM model with \(P\) layers, weight \(W\), training epoch \(n\) dataset to be used

**Output:** Predicted throughput \(\hat{y}\) and error \(e_{LSTM}\)

1. Decompose \(x_t\) into \(a_2 + d_2 + d_1\) ▶ DWT technique
2. \(x_t = \sum_j x_j; \quad x_1 = a_2; \quad x_2 = d_2; \quad x_3 = d_1\)
3. Set \(n = 0, y(n) = y_0\) ▶ training data
4. for \((j = 1 \text{ to } 3)\) do
5. \(t \leftarrow t + 1\)
6. \(f_t = \text{sigmoid}(W_{f,t} x_t + W_{f,b} h_{t-1} + b_f)\) ▶ LSTM input
7. \(\tilde{s}_t = \text{tanh}(W_{s,t} x_t + W_{s,b} h_{t-1} + b_s)\) ▶ input gate
8. \(i_t = \text{sigmoid}(W_{i,t} x_t + W_{i,b} h_{t-1} + b_i)\) ▶ forget gate
9. \(s_j = f_j s_{j-1} + i_j \tilde{s}_j\) ▶ cell
10. \(o_j = \text{sigmoid}(W_{o,t} x_t + W_{o,b} h_{t-1} + b_o)\) ▶ output gate
11. \(h_j = o_j \ast \text{tanh}(s_j)\) ▶ LSTM output
12. \(v_j = h_j(y_j)\)
13. end
14. Prediction results:
15. \(E = \text{AP} (v^1, v^{i-1}, v^{-2}, ..., v^{i-M})\) ▶ average pooling
16. Compute \(P_k = \{P_1, ..., P_K\} \leftarrow \text{soft max}(E)\)
17. Find \(Idx \leftarrow \text{Support}(\text{max}(P_k))\) ▶ index of highest probability
18. \(\hat{y}_j = Idx_j\) ▶ predicted output of each component
19. \(\hat{y} = \sum_j \hat{y}_j\) ▶ invert DWT
20. \(e_{LSTM} = x_t - \hat{y}\) ▶ predicted error
21. End procedure

**Acknowledgements** This research was supported by Korea Institute of Marine Science & Technology Promotion (KIMST), funded by the Ministry of Oceans and Fisheries, Korea (20220573). The authors are grateful to helpful MEL reviews.
References

Bandt, C., and B. Pompe. 2002. Permutation entropy: A natural complexity measure for time series. Physical Review Letters 88 (17): 174102.

Barzegar, R., M.T. Aalami, and J. Adamowski. 2020. Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stochastic Environmental Research and Risk Assessment 34 (2): 415–433.

Chang, Z., Y. Zhang, and W. Chen. 2019. Electricity price prediction based on hybrid model of Adam optimized LSTM neural network and wavelet transform. Energy 187: 115804.

Clément, A., and S. Laurens. 2011. An alternative to the Lyapunov exponent as a damage sensitive feature. Smart Materials and Structures 20 (2): 025017.

Corazza, M., and A.T.G. Malliaris. 2002. Multi-fractality in foreign currency markets. Multinational Finance Journal 6 (2): 65–98.

Cuong, T.N., H.S. Kim, X. Xu, and S.S. You. 2021. Container throughput analysis and seaport operations management using nonlinear control synthesis. Applied Mathematical Modelling 100: 320–341.

Cullinane, A., and S. Laurens. 2011. An alternative to the Lyapunov exponent as a damage sensitive feature. Smart Materials and Structures 20 (2): 025017.

Dirzka, C., and M. Acciaro. 2021. Global shipping network dynamics during the COVID-19 pandemic’s initial phases. Journal of Transport Geography 99: 103265.

Du, P., J. Wang, W. Yang, and T. Niu. 2019. Container throughput forecasting using a novel hybrid learning method with error correction strategy. Knowledge-Based Systems 182: 104853.

Ferreira, K.J., B.H.A. Lee, and D. Simchi-Levi. 2016. Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management 18 (1): 69–88.

Feder, J. 1988. Fractals. New York: Plenum Press.

Fischer, T., and C. Krauss. 2018. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research 270 (2): 654–669.

González-Concepción, C., M.C. Gil-Fariña, and C. Pestano-Gabino. 2012. Using wavelets to understand the relationship between mortgages and Gross Domestic Product in Spain. Journal of Applied Mathematics. https://doi.org/10.1155/2012/917247.

Gou, X., and J.S.L. Lam. 2019. Risk analysis of marine cargoes and major port disruptions. Maritime Economics & Logistics 21 (4): 497–523.

Gosasang, V., W. Chandraprakaikul, and S. Kiattisin. 2011. A comparison of traditional and neural networks forecasting techniques for container throughput at Bangkok port. The Asian Journal of Shipping and Logistics 27 (3): 463–482.

Graves, A., and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18 (5–6): 602–610.

Hochreiter, S., and J. Schmidhuber. 1997. Long short-term memory. Neural Computation 9 (8): 1735–1780.

Hurst, H.E. 1951. Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers 116 (1): 770–799.

Hwarng, H.B., and N. Xie. 2008. Understanding supply chain dynamics: A chaos perspective. European Journal of Operational Research 184 (3): 1163–1178.

Islam, S., and S.H. Amin. 2020. Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques. Journal of Big Data 7 (1): 1–22.

Jahanshahi, H., A. Yousefpour, Z. Wei, R. Alcaraz, and S. Bekiros. 2019. A financial hyperchaotic system with coexisting attractors: Dynamic investigation, entropy analysis, control and synchronization. Chaos, Solitons & Fractals 126: 66–77.

Johannessen, N.J., M. Kolhe, and M. Goodwin. 2019. Relative evaluation of regression tools for urban area electrical energy demand forecasting. Journal of Cleaner Production 218: 555–564.

Justice, V., P. Bhaskar, H. Pateman, P. Cain, and S. Cahoon. 2016. US container port resilience in a complex and dynamic world. Maritime Policy & Management 43 (2): 179–191.

Lam, J.S.L., and T. Notteboom. 2014. The greening of ports: A comparison of port management tools used by leading ports in Asia and Europe. Transport Reviews 34 (2): 169–189.
Lee, Y., H. Song, and S. Jeong. 2021. Prioritizing environmental justice in the port hinterland policy: Case of Busan New Port. Research in Transportation Business & Management 41: 100672.
Liaw, S.S., and F.Y. Chiu. 2009. Fractal dimensions of time sequences. Physica A 338: 3100–3106.
Lin, W., Z. Wu, L. Lin, A. Wen, and J. Li. 2017. An ensemble random forest algorithm for insurance big data analysis. IEEE Access 5: 16568–16575.
Liu, Yan, and Yi-Mei. Chen. 2006. Application of grey system model in throughput forecasting of inland river port. Port & Waterway Engineering 4: 31.
Medsker, L., and L.C. Jain, eds. 1999. Recurrent neural networks: Design and applications. Boca Raton: CRC Press.
Memarzadeh, G., and F. Keynia. 2020. A new short-term wind speed forecasting method based on fine-tuned LSTM neural network and optimal input sets. Energy Conversion and Management 213: 112824.
Morris, S.A., and D. Pratt. 2003. Analysis of the Lotka-Volterra competition equations as a technological substitution model. Technological Forecasting and Social Change 70 (2): 103–133.
Noteboom, T. 2016. The adaptive capacity of container ports in an era of mega vessels: The case of upstream seaports Antwerp and Hamburg. Journal of Transport Geography 54: 295–309.
Pesaran, M.H., and A. Timmermann. 1995. Predictability of stock returns: Robustness and economic significance. The Journal of Finance 50 (4): 1201–1228.
Punia, S., K. Nikolopoulos, S.P. Singh, J.K. Madaan, and K. Litsiou. 2020. Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. International Journal of Production Research 58 (16): 4964–4979.
Rajaee, T., V. Nourani, M. Zounemat-Kermani, and O. Kisi. 2011. River suspended sediment load prediction: application of ANN and wavelet conjunction model. Journal of Hydrologic Engineering 16 (8): 613–627.
Saini, S., D. Roy, and R. de Koster. 2017. A stochastic model for the throughput analysis of passing dual yard cranes. Computers & Operations Research 87: 40–51.
Schlüter, S., and C. Deuschle. 2010. Using wavelets for time series forecasting: Does it pay off? (No. 04/2010). IWQW Discussion Papers.
Schulze, P.M., and A. Prinz. 2009. Forecasting container transhipment in Germany. Applied Economics 41 (22): 2809–2815.
Shankar, S., S. Punia, and P.V. Ilavarasan. 2021. Deep learning-based container throughput forecasting: A triple bottom line approach. Industrial Management & Data Systems. https://doi.org/10.1108/IMDS-12-2020-0704.
Sipper, M., and J.H. Moore 2021. Conservation machine learning: a case study of random forests. Scientific Reports 11 (1): 3629.
Slotine, J.J.E., and W. Li. 1991. Applied nonlinear control, vol. 199, No. 1. Englewood Cliffs, NJ: Prentice Hall.
Sprott, J.C. 1995. Chaos data Analyzer (Professional Version). Raleigh, NC: Physics Academic Software.
Tetteh, E.A., H.L. Yang, and F. Gomina Mama. 2016. Container ports throughput analysis: a comparative evaluation of China and five west African Countries’ seaports efficiencies. In International Journal of Engineering Research in Africa (Vol. 22, pp. 162–173). Trans Tech Publications Ltd.
Twrdy, E., and M. Batista. 2016. Modeling of container throughput in Northern Adriatic ports over the period 1990–2013. Journal of Transport Geography 52: 131–142.
Tzouras, S., C. Anagnostopoulou, and E. McCoy. 2015. Financial time series modeling using the Hurst exponent. Physica A: Statistical Mechanics and Its Applications 425: 50–68.
Wolf, A., J.B. Swift, H.L. Swinney, and J.A. Vastano. 1985. Determining Lyapunov exponents from a time series. Physica D: Nonlinear Phenomena 16 (3): 285–317.
Xie, G., N. Zhang, and S. Wang. 2017. Data characteristic analysis and model selection for container throughput forecasting within a decomposition-ensemble methodology. Transportation Research Part E: Logistics and Transportation Review 108: 160–178.
Yang, C.H., and P.Y. Chang. 2020. Forecasting the demand for container throughput using a mixed-precision neural architecture based on CNN–LSTM. Mathematics 8 (10): 1784.
Yao, C.Z., J.N. Lin, X.F. Liu, and X.Z. Zheng. 2014. Dynamic features analysis for the large-scale logistics system warehouse-out operation. Physica A: Statistical Mechanics and Its Applications 415: 31–42.
Ye, Y., Y. Zhang, Q. Wang, Z. Wang, Z. Teng, and H. Zhang. 2020. Fault diagnosis of high-speed train suspension systems using multiscale permutation entropy and linear local tangent space alignment. Mechanical Systems and Signal Processing 138: 106565.
Zhang, W., and J.S.L. Lam. 2017. An empirical analysis of maritime cluster evolution from the port development perspective – Cases of London and Hong Kong. *Transportation Research Part A: Policy and Practice* 105: 219–232.

Zhang, N., S.L. Shen, A. Zhou, and Y.F. Jin. 2021. Application of LSTM approach for modelling stress–strain behaviour of soil. *Applied Soft Computing* 100: 106959.

Zuo, Y., X. Fu, Z. Liu, and D. Huang. 2021. Short-term forecasts on individual accessibility in bus system based on neural network model. *Journal of Transport Geography* 93: 103075.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.