ICIC Express Letters ICIC International ©2020 ISSN 2185-2766
Part B: Applications pp. 1-8
Volume **11,** Number **9**, September **2020**

Ensemble method for feature selection in time-series prediction based on image similarity

Hyunjae Lee[1], Dohee Kim[2], Sangjae Lee[2], Hanbyeol Park[2],
Hyerim Bae[2]*, and Keonwoo Choi[3]

[1] Department of Industrial Engineering, Pusan National University, Busan 46241, Korea;
hyunjea414@pusan.ac.kr (H.L.);

[2] Major of Industrial Data Science and Engineering, Department of Industrial Engineering, Pusan National University,Busan 46241, Korea; kimdohee@pusan.ac.kr (D.K.); selflsj@pusan.ac.kr (S.L.);

pb104@pusan.ac.kr (H.P.); hrbae@pusan.ac.kr (H.B.);

[3] Shipping Big Data Analysis Center, Korea Maritime Institute, Busan, 49111, Korea;
ak8102@kmi.re.kr(K.C.)

* Correspondence: hrbae@pusan.ac.kr; Tel.: +82-51-510-2733

Received XXX 2022; accepted XXX 2022

ABSTRACT. This paper proposes the implementation of an ensemble feature selection method based on image similarity in time-series data. It is important to find the optimal lag of a feature while analyzing time-series data as the presence of lags in time-series data increases significantly the number of features to be considered. Feature selection plays an important role in distinguishing which features are important and finding features that have a meaningful effect on improving prediction performance. Feature selection based on only numerical data has limitations with respect to predicting rapidly changing time-series data. In this work, we present the availability of our ensemble feature selection method where multiple subsets from feature selection methods based on image similarity are combined into a single subset. We show that our proposed method exhibits better performance than a single feature selection technique based on only image similarity or number.

Keywords: Ensemble method, Feature selection, Time-series, Lag, Image similarity

1. Introduction. Preemptive decision-making based on accurate predictions of the future is critical for shipping companies [1]. Appropriate decisions using timely and accurate container freight forecasting can result in significantly reduced costs for shippers and carriers [2]. Owing to the COVID-19 pandemic, the container freight index has risen significantly, making it more difficult to predict [3]. Therefore, it is necessary to develop an accurate and reliable container freight index model. For this reason, this study focuses on the Shanghai Containerized Freight Index (SCFI) which reflects the fluctuation of 13 spot freight rates on the Shanghai export container transport market [4]. SCFI, one of the most important container freight indexes to evaluate container shipping in the maritime industry, is generally considered a solid indicator in terms of supply and demand balance, container shipping, world trade movement, and the shipping industry [5]. Nevertheless, SCFI is a solid indicator, that rose by about 800 percent during the COVID-19 period, so it is necessary to accurately predict it and prepare preemptive strategies based on accurate prediction.

Owing to redundant features and the curse of dimensionality, the prediction performance of time-series models can decrease in a high-dimensional dataset. To solve this problem, we present a feature-selection method. Feature selection is the process of

identifying and selecting the relevant features from the original dataset so that a model can focus on distinguishing features that are useful for prediction. Feature selection has three advantages: (i) improvement of prediction performance, (ii) selection of relevant features, and (iii) reduction of time required for analysis and learning [7]. Typically, the lag with the highest cross-correlation efficiency is selected as the optimal lag between a target and its features [9]. However, in real-world situations, temporal variations in data do not represent simple regularities, making them difficult to analyze and predict accurately. Only correlations and their combinations do not explain the characteristics of real data [10]. The convolutional neural network (CNN) model filters out noise from image-based data and extracts more relevant features, allowing learning algorithms to focus on important features and perform better future predictions [11]. Graph shape-based feature selection methods do not select features only with numerical data values but find related features even if they rise rapidly through trends or have different units. We propose a feature selection method that ensembles mean squared error (MSE), dynamic time warping (DTW), cosine similarity, and correlation. Finally, we compared each single feature selection subset to verify its performance.

In the literature related SCFI prediction, there are several forecasting SCFI studies including the COVID-19 period. Tengrongre Wang (2021) forecasted container freight indexes including SCFI with ARMA model which is often used for long-term tracking data research. Kaan Koyuncu & Leyla Tavacıoğlu (2021) observed that the SARIMA model provides comparatively better results in forecasting SCFI than the existing freight rate forecasting models while performing short-term forecasts on a monthly rate. Enna Hirata & Takuma Matsuda (2022) compared the LSTM model and a SARIMA model for forecasting the SCFI and showed that the LSTM model based on deep learning outperforms SARIMA models in most of the datasets. Chih-Hsuan Wang & Ying-Ting Lu (2022) collected representative features and predicted SCFI by applying machine learning(random forest, xgboost) and deep learning(DNN, RNN). In this study, it is also proved that deep learning showed better prediction performance than machine learning. After the COVID-19 occurrence, the need for predicting SCFI research has emerged due to the increase in volatility, but no studies have explained and utilized the external variables. The contributions of this study are summarized as follows:

- We selected variables among external variables that affect SCFI prediction in rapidly changing time-series data by ensemble feature selection method based on image similarity.
- We improved the prediction accuracy of SCFI by applying LSTM. Our proposed method shows better prediction performance over most prediction periods compared previous method(RNN).
- We considered lags for each feature while considering the characteristics of time series by determining variables that can help in predicting data with significant relationships.

The remainder of this paper is organized as follows: Section 2 presents several algorithms used in our proposed method. Section 3 presents the proposed method based on an ensemble. Section 4 presents a description of the used dataset and results of our proposed method compared with single feature selection. Finally, section 5 presents conclusions.

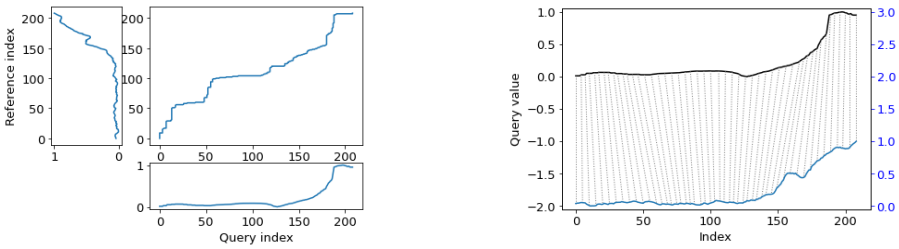
2. Background. In this section, we describe the algorithms used in our proposed method. It contains a definition and overview of the MSE, DTW, cosine similarity, and correlation.

2.1. Mean Squared Error (MSE). The MSE is widely used to assess image quality and distortion because it is easy to calculate and does not require significant computational complexity [12]. The MSE is also a full reference of the image pixels and is averaged by the

image size. Sara et.al.(2019) defined the MSE between two images, where *g(x, y)* is a feature and $\hat{g}(x, y)$ is the target [13]. Where *N* is horizontal pixels and *M* is vertical pixels.

$$\text{MSE} = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}[\hat{g}(x,y) - g(x,y)]^2 \quad (1)$$

2.2. Dynamic Time Warping (DTW). Research on the indexing of time series is mainly conducted by Euclidean distance measurement method. DTW, one of the methods of indexing time series, is a stronger distance measure for time series by matching similar shapes between time series and is widely used in various fields including science, medicine, and finance [14]. Some studies have demonstrated that DTW is better for indexing than the Euclidean distance [15, 16]. The advantage of DTW is that it can be applied to time series with different time sequences. DTW represents the distance based on the path, which means that the smaller the distance, the more similar the pattern [14]. We utilized these DTW characteristics for feature selection to select relevant features with a small distance between the SCFI and features. Figure 1 shows an example of the application of DTW.



(a) (b)

FIGURE 1. AN EXAMPLE OF DTW
(a) A warping matrix (b) The alignment result

2.3. Cosine Similarity. Cosine similarity is a simple and effective way to measure the angle between two normalized vectors as a dot product. Khan et al. (2019) defined cosine similarity between two vectors X and Y in the N dimension [17].

$$\text{Cosine Similarity}(\text{X},\text{Y}) = \frac{X \bullet Y}{\|X\|\,\|Y\|} = \frac{\sum_{i=1}^{N} X_i \times Y_i}{\sqrt{\sum_{i=1}^{N} {X_i}^2}\sqrt{\sum_{i=1}^{N} {Y_i}^2}} \quad (2)$$

Img2vec model provided by Keras was used as a preprocessing method to obtain a vector. This library uses the ResNet50 model, which is pre-trained on ImageNet, to generate image vectors[18]. We measure the cosine similarity between the generated image vectors.

2.4. Pearson Correlation Coefficient (Correlation). The Pearson correlation coefficient is a metric of the linear correlation between two variables *X*(Independent variable) and *Y*(Dependent variable) in statistics, and it is widely used in various applications to analyze linear correlations. The equation for the Pearson correlation coefficient is as follows [19]:

$$\text{Pearson Correlation Coefficient}(\text{X},\text{Y}) = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} \quad (3)$$

3. Proposed method. This section describes our proposed method based on the ensemble method. The ensemble is contained using the image similarity measurement method described in Section 2. In the feature-selection method based on an ensemble, multiple feature subsets considering lags from each feature selection are combined to create an optimal

subset of features with an optimal lag that improves prediction performance [9].

Our proposed method consists of two steps for creating the best feature subset with an optimal lag. The first step involves creating each subset using each feature-selection method. Each subset is created according to the feature-selection methods presented previously based on the criterion for each method in the state of considering lag of 8-weeks ago, 4-weeks ago, no lag, 4-weeks later, and 8-weeks later. From the perspective of cosine similarity and correlation, we only selected features with a value higher than 0.8, which represents it is similar. Among the features considering all lags, if there are many values of more than 0.8, the lag of the highest value was selected. Consistent with the cosine similarity and correlation, feature selection using the DTW and MSE also needs to set a criterion. As mentioned earlier, DTW and MSE are similar because the values are smaller. We selected features within the top 20%. If there are different lags for one feature, the lag with the smallest value for each feature was finally selected.

In the second step, we focused as much as possible on using the features of the subsets derived from each method. First, we used all features selected in each method. If the same feature is selected in the three image similarity measurement methods and two lags out of the three are the same, the lag with more is selected. However, in this case, if all the lags are different, the lag with the highest positive correlation coefficient is selected. Figure 2 describes the framework of our ensemble method.

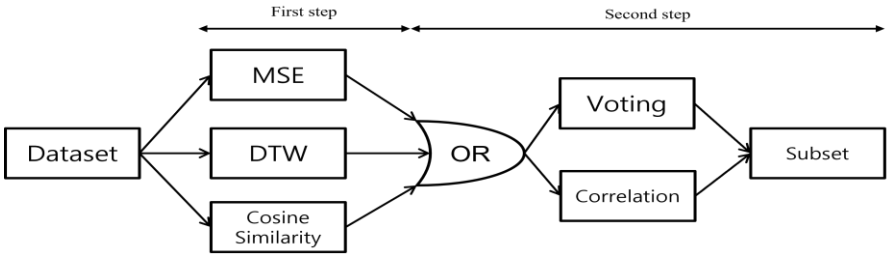


FIGURE 2. A framework of the ensemble method

4. Experiment. In this section, we present a description of our dataset and the results of our proposed method based on ensemble feature selection. We performed short-term predictions as 1 week & 4 weeks, mid-term prediction as 12 weeks, and long-term prediction as 24 weeks.

4.1. Dataset. The dataset was provided by the Korea Maritime Institute (KMI), which focuses on research and development in the shipping sector. In Table 1, 18 variables, including the SCFI, External Variables, and Port Congestion Index, were used as data. The target to be predicted is the SCFI, and the feature used to predict it has a total of 18 variables, including the previous SCFI. The data used to build the model were from January 1, 2018, to December 31, 2021. The number of data points used was 209 (weekly).

TABLE 1. Description of dataset

	Group name	Variable name	Unit
Target	SCFI	SCFI (Comprehensive)	Index
Feature	External Variables	Average Earnings	$/day
		Bunker Price	$/Tone
		Total Container ships Number	No
		Total Container ships TEU	TEU
		Newbuilding Prices (1650/1850 TEU)	$m
		Newbuilding Prices	$m

		(13000/14000 TEU)	
		Newbuilding Prices (3500/4000 TEU)	$m
		Newbuilding Prices (13000/13500 TEU)	$m
		5 Year Finance based on Libor	$m
	Port Congestion Index	Port congestion index (Comprehensive)	Million TEU
		Port congestion index - East coast North America	Million TEU
		Port congestion index - West coast North America	Million TEU
		Port congestion index - The United Kingdom	Million TEU
		Port congestion index - Mediterranean/Black Sea	Million TEU
		Port congestion index - East Asia	Million TEU
		Port congestion index - South East Asia	Million TEU
		Port congestion index - China	Million TEU

4.2. Result of subsets configured through each feature selection. Table 2 shows a subset created of each single feature selection and the proposed method presented in section 3.

TABLE 2. A subset of feature selection.

Group name	Variable name	MSE	DTW	Cosine Similarity	Correlation	**Proposed Method**
SCFI	SCFI (Comprehensive)	√	√	√	√	**√**
External Variables	Average Earnings	√ (-4 weeks)	√ (+8 weeks)	√ (-8 weeks)	√ (-8 weeks)	**√ (-8 weeks)**
	Bunker Price					
	Total Container ships Number			√ (+8 weeks)		**√ (+8 weeks)**
	Total Container ships TEU			√ (+4 weeks)		**√ (+4 weeks)**
	Newbuilding Prices (1650/1850 TEU)			√ (+4 weeks)		**√ (+4 weeks)**
	Newbuilding Prices (13000/14000 TEU)	√ (-8 weeks)		√ (Original)	√ (-8 weeks)	**√ (-8 weeks)**
	Newbuilding Prices (3500/4000 TEU)	√ (+8 weeks)	√ (Original)	√ (+8 weeks)	√ (-8 weeks)	**√ (+8 weeks)**
	Newbuilding Prices (13000/13500 TEU)	√ (-8 weeks)	√ (-8 weeks)	√ (+8 weeks)	√ (-8 weeks)	**√ (-8 weeks)**
	5 Year Finance based on Libor			√ (-4 weeks)		**√ (-4 weeks)**
Port Congestion Index	Port congestion index (Comprehensive)		√ (+8 weeks)			**√ (+8 weeks)**
	Port congestion index - East coast North America				√ (-8 weeks)	
	Port congestion index - West coast North America			√ (-8 weeks)	√ (+4 weeks)	**√ (-8 weeks)**
	Port congestion index - United Kingdom				√ (-4 weeks)	
	Port congestion index -Mediterranean/Black Sea					
	Port congestion index - East Asia					
	Port congestion index - South East Asia					

	Port congestion index - China					
Total number of features used		5	5	10	8	**11**

4.3. Result of the experiment. The model that was used in the experiment is the LSTM which is one of the represented deep learning models. LSTM is suitable for time series data analysis because it can store, discard (forget) or add important information for prediction at the current time sequence and send it to the next time sequence [20]. It consists of a cell state, input gate, output gate, and forget gate. The interaction of three gates and the cell state regulates the flow of information over the time interval. The LSTM is illustrated in Figure 3. To assess the performance of the model, the root mean squared error (RMSE) and the mean absolute percentage error (MAPE) are used. The equation for the RMSE and MAPE are as follows, where X is the SCFI from prediction model at the particular point, and Y is the real SCFI value at the same point [21]:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \qquad MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{Y_i - X_i}{Y_i}\right| \tag{4}$$

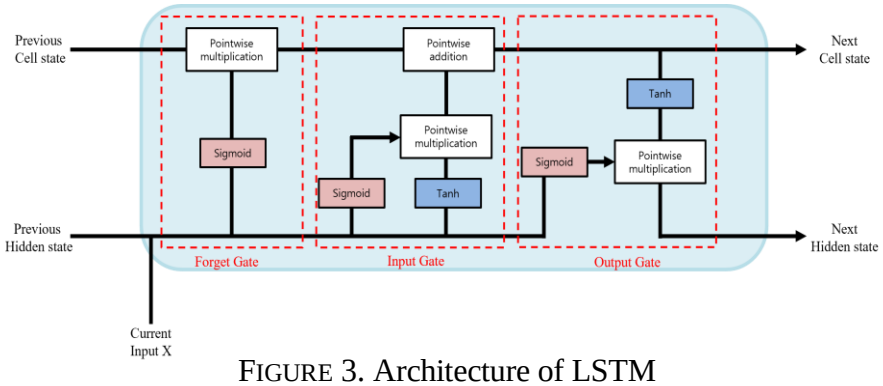


FIGURE 3. Architecture of LSTM

The results of the experiment are shown in Table 3, and the plots showing the results of the proposed method are shown in Figure 4. The proposed method shows better prediction performance over most prediction periods than RNNs. For the subset we create with the ensemble method, LSTM shows about 20% improved prediction performance than RNN.

TABLE 3. Result of the experiment.

Prediction period	Subset	Proposed method(LSTM)		RNN	
		RMSE	MAPE	RMSE	MAPE
1 week	MSE	1166.705	30.822	1247.236	35.169
	DTW	1263.25	32.489	2084.002	54.238
	Cosine similarity	571.524	14.978	1458.291	35.418
	Correlation	1225.578	27.745	1307.393	34.372
	Ensemble method	**280.262**	**7.86**	**942.325**	**24.536**
4 weeks	MSE	832.383	18.824	1037.784	31.239
	DTW	581.432	13.729	2489.446	64.995
	Cosine similarity	750.545	16.109	1816.732	40.475
	Correlation	527.881	12.964	947.199	26.322
	Ensemble method	**509.672**	**13.276**	**612.486**	**15.248**
12 weeks	MSE	2264.018	53.469	3091.222	79.974
	DTW	2134.867	54.278	3004.639	73.46
	Cosine similarity	2194.517	55.624	2185.468	58.833

	Correlation	2704.516	59.958	2014.491	51.254
	Ensemble method	**1876.258**	**45.21**	**1717.702**	**41.881**
24 weeks	MSE	2483.673	62.771	1865.811	50.016
	DTW	1926.935	49.039	1904.194	51.884
	Cosine similarity	**1236.007**	**30.869**	2037.731	57.326
	Correlation	2109.019	53.692	1626.801	45.851
	Ensemble method	1243.406	31.145	**1478.981**	**38.983**

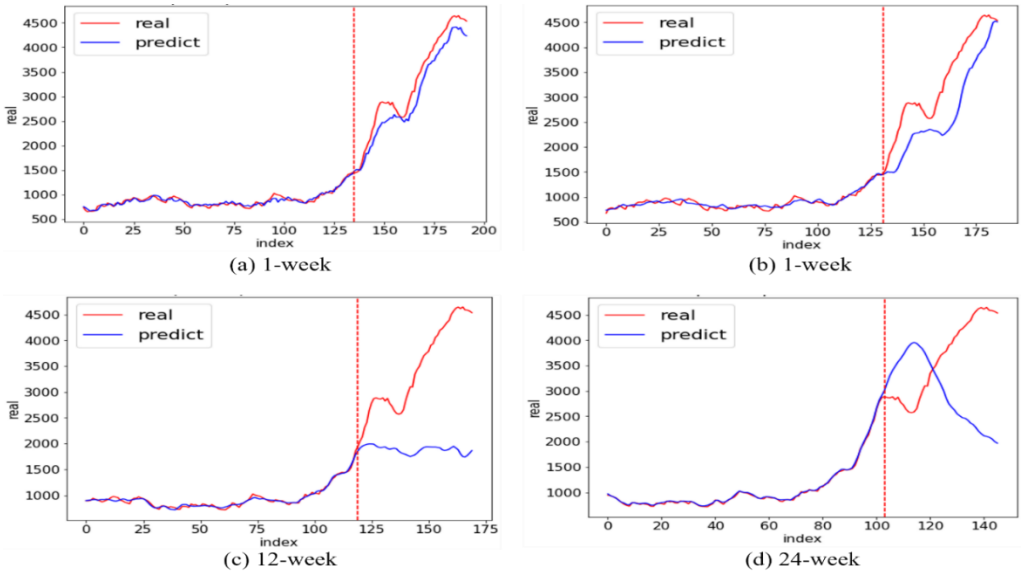


FIGURE 4. Graph visualization of prediction results each period

5. Conclusion. In this study, we propose an ensemble method for feature selection based on image similarity for making the preemptive decision. We introduce the idea of feature selection based on image similarity to find relevant features that can improve the prediction performance, even though lags are applied. We experimented with short-term and mid-long-term predictions using the Shanghai Containerized Freight Index(SCFI), which has risen significantly during the COVID-19 pandemic. We showed that our proposed method outperformed other single-feature selection methods with respect to short-term predictions corresponding to the 1-week and 4-week predictions. In addition, the mid-term prediction corresponding to the 12-week prediction was better than the other methods, but further analysis of the graph confirmed that the prediction graph did not follow the actual graph well. Our proposed method was not the best method for predicting long-term corresponding to the 24-week, but it was the next best. All methods, as well as the proposed method, show poor performance in mid- to long-term predictions. This is because the amount of data for mid- to long-term predictions was not sufficient.

In conclusion, we propose a method to show good prediction performance overall. Various image-based feature selection methods are combined to sufficiently reflect external influences that are not reflected in each feature selection method, while reducing dimensions. Through this, it was possible to provide explanatory power and better prediction performance for the influence of each shape. In future works, we will also verify that our proposed method performs well when applied to other types of time series data.

Acknowledgment. This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C1102294) and supported by KMI (the Korea Maritime Institute), NRC(the National Research Council

for Economics, Humanities and Social Sciences)

REFERENCES

[1] Stopford; M. Maritime Economics 3e; Routledge: London; UK; 2008.

[2] Munim, Z.H., Schramm, HJ; Forecasting container shipping freight rates for the Far East – Northern Europe trade lane; *Marit Econ Logist* vol.19; no.1; pp.106–125; 2017.

[3] Ardelean, Adina, Lugovskyy, Volodymyr, Skiba, Alexandre, Terner, David; Fathoming Shipping Costs : An Exploration of Recent Literature, Data, and Patterns; Policy Research Working Paper;9992. World Bank, Washington, DC; © World Bank. https://openknowledge.worldbank.org/handle/10986/37276 License: CC BY 3.0 IGO.”; 2022

[4] Shanghai Shipping Exchange;Shanghai Export Containerized Freight Index (SCFI); Retrieved on September 21, 2021; http://en.sse.net.cn/; 2020

[5] Koyuncu, K. & Tavacıoğlu, L; “Forecasting Shanghai Containerized Freight Index by Using Time Series Models.”; Marine Science and Technology Bulletin; 10 (4), 426-434. 10.33714/masteb.1024663; 2021

[6] Builes-Jaramillo, Alejandro, et al.; "Nonlinear interactions between the Amazon River basin and the Tropical North Atlantic at interannual timescales." ;*Climate Dynamics* ;50.7 ;2951-2969; 2018.

[7] Hoque, Nazrul, Mihir Singh, and Dhruba K. Bhattacharyya; "EFS-MI: an ensemble feature selection method for classification."; *Complex & Intelligent Systems* ; 4.2; 105-118; 2018.

[8] Xuan, Zhou, et al.; "Forecasting performance comparison of two hybrid machine learning models for cooling load of a large-scale commercial building."; *Journal of Building Engineering*; 21; 64-73; 2019.

[9] An, Dawn, Nam H. Kim, and Joo-Ho Choi.; "Practical options for selecting data-driven or physics-based prognostics algorithms with reviews."; *Reliability Engineering & System Safety;* 133; 223-236; 2015.

[10] Livieris, I.E., Pintelas, E.G., & Pintelas, P.E; A CNN–LSTM Model for Gold Price Time-series Forecasting.; *Neural Comput. and Appl*; pp.1-10; 2020

[11] Zhou Wang and A. C. Bovik; A Universal Image Quality Index, in IEEE Signal Process; Lett; vol. 9; no. 3; pp. 81-84; 2022.

[12] Sara, U., Akter, M. and Uddin, M; Image Quality Assessment through FSIM, SSIM, MSE and PSNR, A Comparative Study; J. of Comput. and Commun; vol.7; no.3; pp.8-18; 2019.

[13] Liu, Chien-Liang, Wen-Hoar Hsaio, and Yao-Chung Tu.; "Time series classification with multivariate convolutional neural network."; *IEEE Transactions on Industrial Electronics*; 66.6; 4788-4797; 2018.

[14] Paparrizos, John, and Luis Gravano; "k-shape: Efficient and accurate clustering of time series."; *Proceedings of the 2015 ACM SIGMOD international conference on management of data*; 2015.

[15] A Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah; "Time-series clustering–a decade review."; *Information systems;* 53; 16-38; 2015.

[16] Khan, Ahmed Yar, et al.; "Malicious insider attack detection in IoTs using data analytics."; *IEEE Access* 8; 11743-11753; 2019.

[17] Jaredwinick; Image to Dense Vector Embedding*;* Available online: https://github.com/jaredwinick/img2vec-keras (accessed on 20 June 2022) 2019.

[18] Zhu, Hongwei, Xiaoming You, and Sheng Liu.; "Multiple ant colony optimization based on pearson correlation coefficient."; *IEEE Access* 7 ; 61628-61638; 2019.

[19] Hirata, E.; Matsuda; Forecasting Shanghai Container Freight Index: A Deep-Learning-Based Model Experiment.; *J. Mar. Sci. Eng.10*; 593; https://doi.org/10.3390/jmse10050593; 2022.

[20] Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman.; "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation."; *PeerJ Computer Science* 7; e623; 2021.