ICIC Express Letters
Volume **18**, Number **3**, March **2024**

ICIC International ©2024 ISSN 1881-803X
pp. **255–262**

A ROBUST EXPLAINABLE FEATURE SELECTION METHOD BASED ON ORDER STATISTICS

Dohee Kim[1], Sangjae Lee[1], Sunghyun Sim[2] and Hyerim Bae[1,*]

[1]Major in Industrial Data Science and Engineering
Department of Industrial Engineering
Pusan National University
2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea
{ kimdohee; selflsj }@pusan.ac.kr; *Corresponding author: hrbae@pusan.ac.kr

[2]Major in Industrial Management and Big Data Engineering
Division of Industrial Convergence System Engineering
Dong-Eui University
176, Eomgwang ro, Gaya Dong 24, Busanjin-gu, Busan 47340, Korea
ssh@deu.ac.kr

Received June 2023; accepted August 2023

Abstract. *Recently, Feature Selection (FS) methods have been extensively researched using eXplainable Artificial Intelligence (XAI). Among these methods, SHapley Additive exPlanations (SHAP) is a representative approach. SHAP evaluates the impact of feature subset on predicted values based on game theory. Consequently, the feature importance can vary when iterated, and change depending on the prediction model used. In this paper, we propose a robust FS method based on order statistics. We determine feature ranking through two approaches: Feature Importance (FI) derived from the model-specific using different prediction models and model-agnostic using iterative experiments. Finally, we construct feature subsets based on the sum of these rankings and criteria. Through experiments, we validate the robustness and efficiency of our approach. By utilizing this approach, we can identify suitable feature subsets without the need to explore various FS methodologies, regardless of the prediction model and data dimension.*
Keywords: Feature selection, Time-series forecasting, Order statistics, XAI, SHAP

1. **Introduction.** FS is considered an important factor in Time-Series Forecasting (TSF) problems, but it comes with various constraints [1]. High-dimensional data provides opportunities to capture and reflect external volatility, patterns, and trends. However, as the dimensionality increases, the required space for data processing grows exponentially, leading to a gradual deterioration in data quality. This phenomenon is known as the 'curse of dimensionality' [2]. FS aims to reduce dimensionality by selecting a subset from the original set of features based on noise, relevance, and redundancy in high-dimensional data. It strives to identify and choose meaningful features to address the problem at hand. By reducing the dimensionality and emphasizing important features, FS improves the efficiency of data analysis and prediction tasks, enhancing the understanding and prediction of the given problem.

FS is commonly divided into three types: filter method, wrapper method, and embedded method and the procedure of FS is illustrated in Figure 1 [3]. FS is conducted through five stages, and the performance varies depending on the decisions made at each stage. In the first stage, the starting point of FS is identified, and the search direction is selected from options such as forward, backward, or random. In the second stage, the search strategy is determined, which can be randomized, exponential, or sequential. Depending on the chosen strategy, it is possible to encounter NP-hard problems. The third stage

DOI: 10.24507/icicel.18.03.255

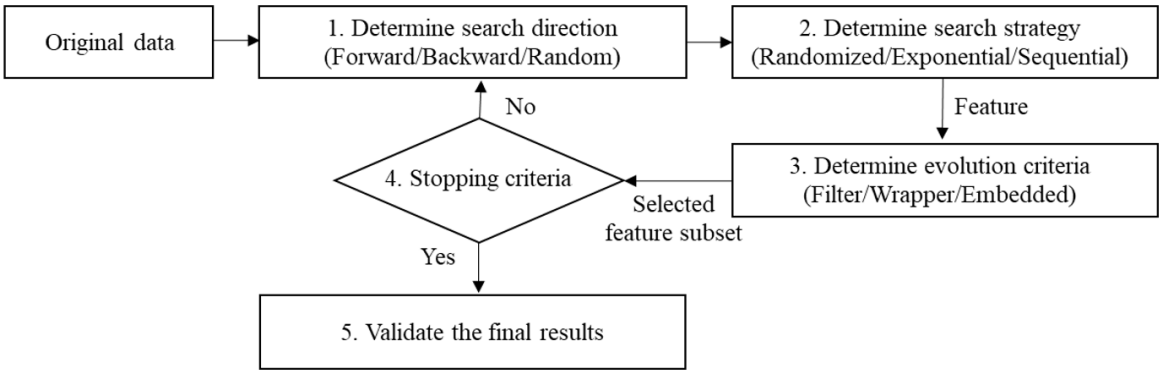


FIGURE 1. Five-stage of FS procedure

involves determining the evolution criteria, which includes filter, wrapper, embedded, and hybrid methods. Filter methods are independent of the learning algorithm and select features based on statistical evidence such as Pearson's correlation coefficient, ANOVA, and Chi-square. These methods are efficient and computationally faster but may find it challenging to identify subsets that have a significant impact on learning like prediction. Wrapper methods act as surrogate models to evaluate the predictive power of subsets of features. However, even variables that have a significant influence on the output may be eliminated, and they often involve a high computational cost due to iterative processes. Lastly, embedded methods perform feature selection during training, allowing for the selection of subsets that have an impact on specific models. Consequently, the choice of methodology depends on the type of model and desired prediction performance [2,3]. In the fourth stage, stopping criteria are specified, taking into consideration of factors such as overfitting and computational complexity. Finally, in the last stage, the validity of the results is verified using techniques such as cross-validation and confusion matrices for validation purposes.

Recently, there has been significant research in the FS aimed at achieving higher prediction or classification accuracy while reducing data dimensionality. Hybrid approaches and optimization methodologies have been extensively studied to achieve this goal. Thakkar and Lohiya [4] propose a novel filter-based FS method that utilizes the difference between the standard deviation and the mean and median values. Through this approach, they derive a reduced subset of features that exhibit high discernibility and deviation. The fusion of statistical measures allows for the incorporation of statistical significance. In addition, various efforts have been made to improve performance, such as ensemble methods that combine multiple filter methodologies [5], hybrid approaches that integrate filter and wrapper methods [6,7], and the application of metaheuristic algorithms [8,9].

On the other hand, there are also studies that focus on interpretability for model explanation and feature selection [10]. These studies utilize XAI methodologies for FS, aiming to provide explanations for all decisions made throughout the entire process of machine learning, particularly in the process of feature selection [3]. Many studies have found SHAP to be a suitable feature selection method, considering the limitations of other methods such as permutation importance [11], which is inconsistent and unable to calculate negative influences, and Local Interpretable Model-agnostic Explanations (LIME) [12], which is more suitable for single prediction explanations rather than providing global results. Therefore, due to these drawbacks, researchers have widely utilized SHAP as a feature selection method [10,13]. On the other hand, there are limitations to using SHAP for FS. SHAP is a representative post-hoc method that determines the influence of features on the results after the experiment without repetition [14]. Accordingly, it suffers from a critical drawback: the Shapley values fluctuate [15] due to various factors such as the prediction period and the model, making SHAP unreliable and unsuitable as an FS method.

In this study, based on previous research, SHAP is utilized to calculate *FI*, and then rankings are obtained using order statistics. Additionally, to address the limitations of SHAP, the rankings are further enhanced by training and summing multiple rankings to create a feature ranking criterion. The performance of FS is evaluated by comparing it with previous FS methodologies to determine how much it improves prediction accuracy.

The contributions of this study are as follows: 1) We propose a novel framework and prioritized criterion for utilizing SHAP as an explainable FS method; 2) We introduce a dual perspective approach to feature ranking by considering both model-specific and model-agnostic viewpoints. This comprehensive perspective enhances the robustness and reliability of the feature ranking process; 3) We present a Correction Feature Rank (CFR) for recalculating and rearranging redundant rankings, facilitating effective Feature Ranking (FR) based on order statistics; 4) Through extensive experimentation, our proposed method demonstrates improved predictive performance, even when the original data dimension is halved. It outperforms other existing feature selection methods, highlighting its efficacy in enhancing prediction accuracy. In addition to improved performance, our proposed method allows for the interpretation of feature importance.

The remainder of this paper is organized as follows. Section 2 presents a description of the proposed method. Section 3 provides the results derived from our experiments and compares them with the results of other FS methods. Finally, Section 4 presents the conclusions along with future research ideas.

2. **Proposed Method.** This section describes our methodology. An overall framework of the proposed method is shown in Figure 2.

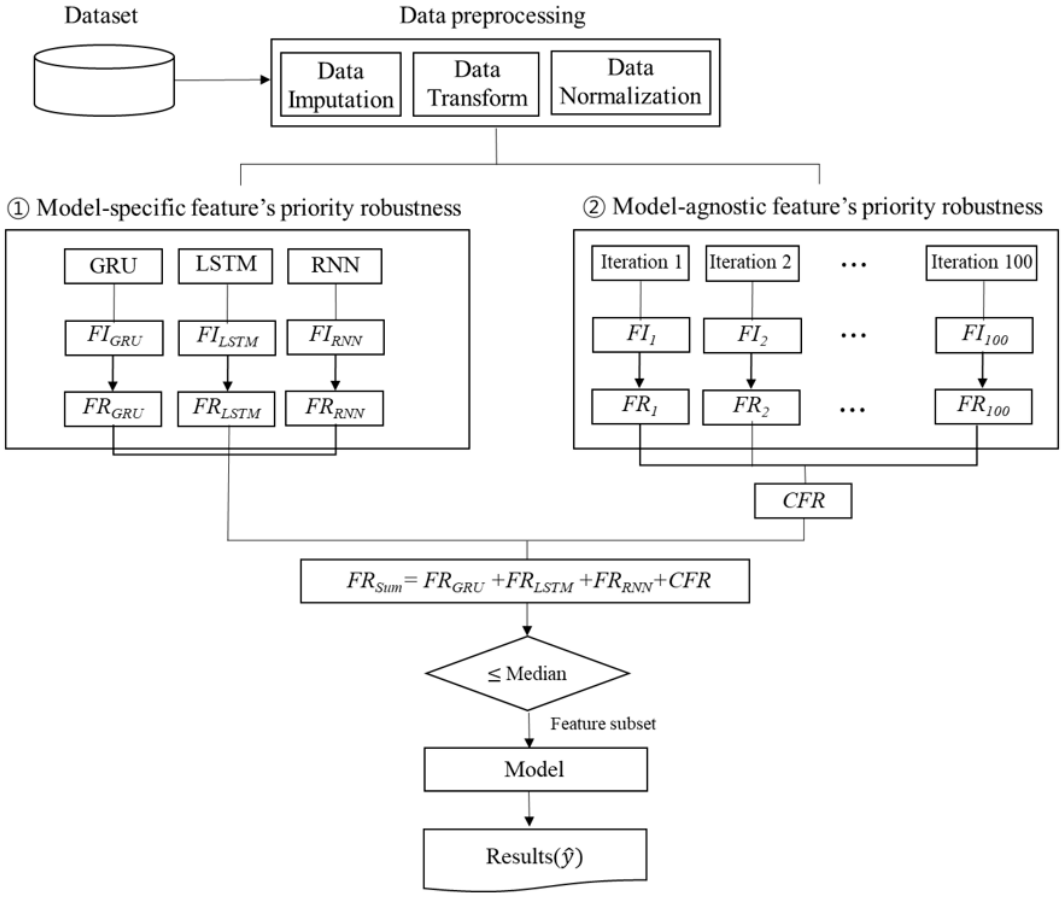


FIGURE 2. Overall framework of the proposed method

First, we conducted data pre-processing including handling missing data, ensuring consistent time units, and performing normalization. Then, we approached the problem from two perspectives. In the first perspective, we employed representative deep learning-based time series prediction models such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) to perform predictions. We utilized the post-hoc analysis to apply SHAP and calculated the *FI* values and rankings for each model. In the second perspective, we utilized eXtreme Gradient Boosting (XGBoost) and SHAP to randomize the samples of the data, allowing for repeated experiments. Through these iterations, we obtained *FI* values and rankings for each iteration and combined them to derive a *CFR*. Subsequently, we combined the rankings from the three deep learning-based models and *CFR* using ranking summation.

Based on this ranking, we determined a feature subset with dimensions less than or equal to half of the original data dimensions. This subset was then used as the input for the GRU model to derive the final output.

2.1. **FI.** We calculate the *FI* based on the Shapley value. This value represents the average marginal contribution of a feature value based on results across all possible combinations of features. When the feature dimension of N, the feature importance is depicted as shown in Figure 3, and each row is independent. The equation for Shapley value that the only additive method that satisfies the properties of local accuracy, Missingness, and consistency is given by Equation (1), where f is the model, x is the available variables, and x' are the selected variables. The quantity $f_x(z') - f_x(z' \setminus i)$ expresses, for every single prediction, the deviation of Shapley values (ϕ) from their mean: the combination of the i-th variable.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z' \setminus i)] \tag{1}$$

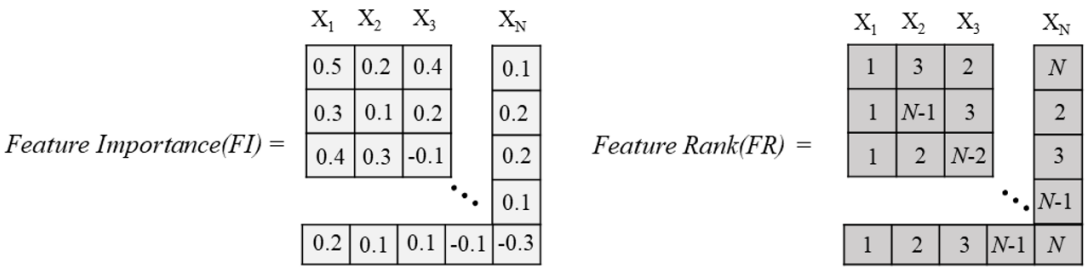


FIGURE 3. Description of *FI* and *FR*

2.2. **FR.** Based on the calculated *FI* values, we sort the feature in ascending order. We define each order as the *FR* as shown in Figure 3. In this study, we obtained a total of 103 *FR*s, including model-specific *FR* ($FR_{GRU}, FR_{LSTM}, FR_{RNN}$) and model-agnostic *FR* ($FR_1, FR_2, \ldots, FR_{100}$).

2.3. **CFR.** For the model-agnostic *FR*, which refers to the *FR* obtained through multiple samplings of the sample data using a single model and iterating through the process, both the *FI* values and the resulting *FR* can vary across different iterations of the experiment. To enhance the robustness of this ranking, we propose a three-step approach to derive the *CFR*.

Firstly, we set the *FR* as the mode value among the *FR*s obtained from each iteration of the experiment. Secondly, in cases where there are overlapping rankings, we recalculate the rank based on the probability of the actual experiment and select the rank with the smaller value. In Figure 4, it can be observed that in the example alongside the *CFR* procedure, the *FR* of X_1 and X_8 are both equal to 1. Although they are determined as

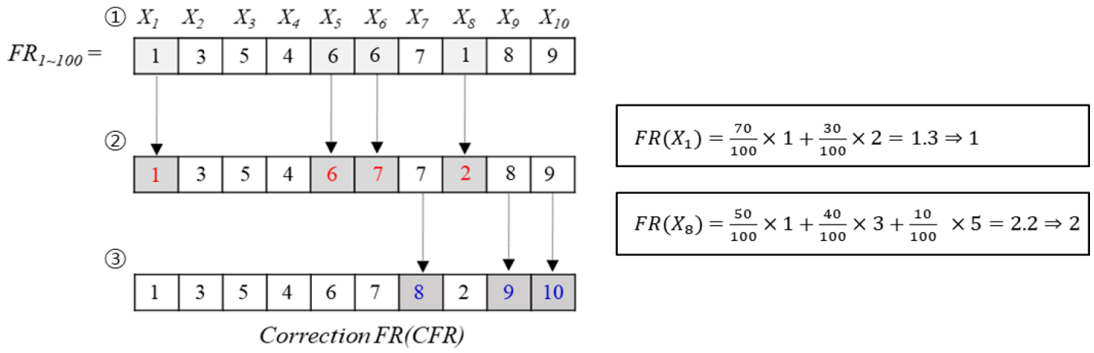


FIGURE 4. The procedure of *CFR*

the mode value, in the actual experiment, $FR(X_1)$ appeared 70 times as $FR(X_1) = 1$ and 30 times as $FR(X_1) = 2$. Therefore, the ranking probability value is re-calculated as 1.3. On the other hand, $FR(X_8)$ had the highest occurrence as $FR(X_8) = 1$, but it appeared 50 times as $FR(X_8) = 1$, 40 times as $FR(X_8) = 3$, and 10 times as $FR(X_8) = 5$, resulting in a higher variance with a ranking probability value of 2.2. In such a case, the ranking is changed to $FR(X_1)$ being ranked 1 and $FR(X_8)$ being ranked 2. Lastly, we shift the rank by the number filled in front and derived the final *CFR*.

2.4. **Feature subset.** Finally, we calculate the rank sum of FR_{GRU}, FR_{LSTM}, FR_{RNN}, and *CFR* obtained from each experiment result to obtain the final FR_{Sum}. We determined a feature subset with dimensions less than or equal to half of the original data dimensions (= the median of rankings). This subset is used as the input for the GRU model to perform predictions. To evaluate the performance of the proposed model and compare it with other FS methods, we use Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) [16], and the equations are as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (\widehat{y}_t - y_t)^2}{n}} \tag{2}$$

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{y_t - \widehat{y}_t}{y_t} \right| \tag{3}$$

3. **Experiments.** In this section, we present our experimental setup and data descriptions. Additionally, we compare the proposed method with previous FS models.

3.1. **Experimental setup.** We conducted our experiments using two different datasets with distinct characteristics as indicated in Table 1.

TABLE 1. Description of experiment datasets

Notation	# of rows	# of features	Collected duration	Unit	Target variable	Data offer
DT1	213	8	2003.03.31∼2020.10.31	Monthly	Container throughput	Busan Port Authority (BPA)
DT2	273	33	2017.01.08∼2022.03.20	Weekly	Mandarin mean price	Jeju-Island

Our experiment was using Python with TensorFlow and PyTorch as the deep learning framework. Each dataset is divided into a training set of 70% and a testing set of 30%. The model is trained for a maximum of 1000 epochs with ADAM optimizer, 10^{-3} learning rate. The number of iterations for the model-agnostic perspective is 100. We compared

our prediction results with those of the existing FS method with Pearson Correlation Coefficient [3], Mean Square Error with data image [5], Dynamic Time Warping [5], Img2Vec with Cosine Similarity [5], and ensemble method [5] using the RMSE and MAPE metrics [16] for validation.

3.2. **Experimental results.** The SHAP *FI* plot for DT1 is shown in Figure 5. It shows that the Shapley value varies with each iteration, but the importance ranking does not change significantly. This means that we can further reduce volatility by selecting ranking, not importance value. The FR_{Sum} and feature subsets obtained for each dataset are presented in Figure 6. DT1 was constructed using a subset comprising features ranked up to 4 of a total of 8 features, while DT2 was constructed using a subset comprising features ranked up to 16 out of a total of 33 features (less than or equal to half of the original data feature dimensions).

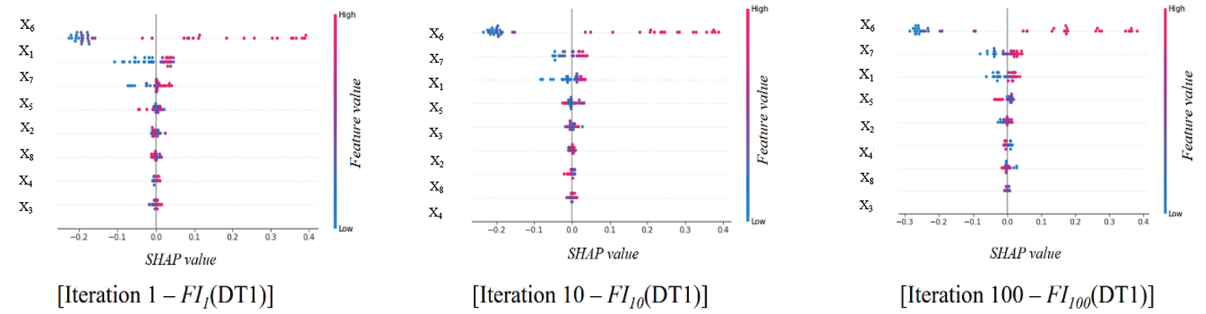


FIGURE 5. SHAP importance plot (FI) for DT1

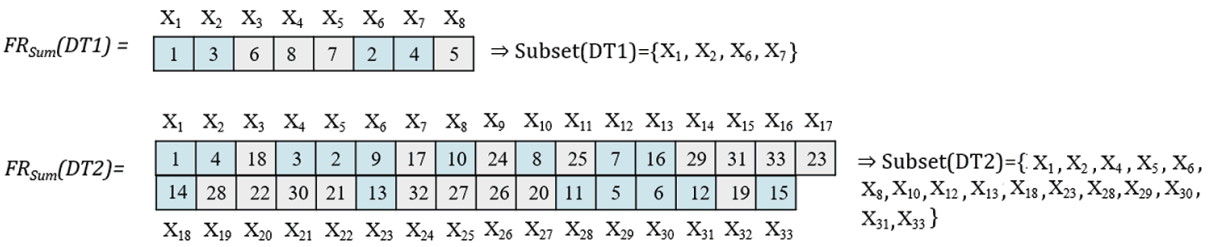


FIGURE 6. Final feature subset based on FR_{Sum}

The prediction results using GRU model with the final feature subset as input are presented in Table 2. Our proposed model shows a 40% improvement in performance compared to the case that not used FS (original data). Despite our model reducing the number of features by half compared to other FS methodologies, it achieved the best predictive performance. For DT2, where the target value contains zeros and MAPE calculation was not possible, it is evident that it achieved the best results based on RMSE. The predicted results of the proposed method can also be observed in Figure 7 through a comparison graph with the actual values.

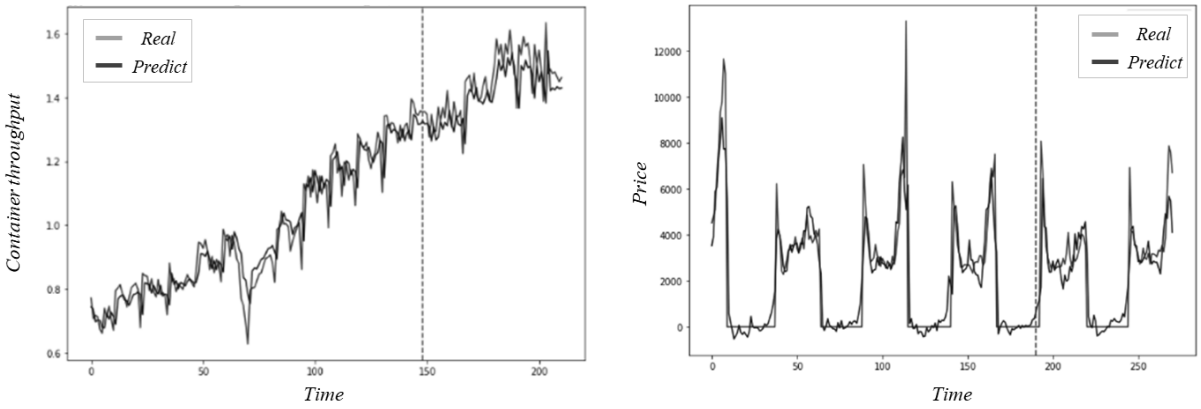
4. **Conclusions.** We present a new framework and prioritized criterion for incorporating SHAP as an FS method, leveraging its explainability. We establish feature rankings from two perspectives: model-specific and model-agnostic and derive the ranking sum using order statistics. Additionally, we propose *CFR*, a method for recalculating and rearranging redundant rankings, to enhance the effectiveness of feature ranking. Our proposed method demonstrates enhanced predictive performance, despite reducing the original data dimension by half. It also exhibits favorable performance compared to other previous FS methods, while providing insights into the influence of each selected feature. In our

Table 2. Prediction results of each dataset

Dataset	FS method	# of features	RMSE	MAPE
DT1	All (Not used FS)	8	115221	6.339
	Pearson Correlation Coefficient	7	115247	6.348
	Mean Square Error	3	104886	6.04
	Dynamic Time Warping	3	94758.6	5.074
	Img2Vec+Cosine Similarity	7	133084	7.369
	Ensemble method	3	137397	7.956
	Proposed method*	**4**	**76168.1**	**3.814**
DT2	All (Not used FS)	33	1206.56	–
	Pearson Correlation Coefficient	3	1338.39	–
	Mean Square Error	8	1337.64	–
	Dynamic Time Warping	8	1332.56	–
	Img2Vec+Cosine Similarity	32	1212.54	–
	Ensemble method	12	1319.46	–
	Proposed method**	**16**	**1173.21**	**–**

*Proposed method's features (DT1): Container throughput, Manufacturing Production Index (MPI), Consumer Price Index (CPI), Export Price (EP)

**Proposed method's features (DT2): Price mean, Price min, Trade total, Shipment, Sweet-persimmon price, Redhyang price, Apple price, Orange price, woldong-satsuma price, Houseonju price, Orange import amount, House-satsuma shipment, woldong-satsuma shipment, Setoka shipment, Kara shipment, Beni-Madonna shipment



(a) DT1: Container throughput data (b) DT2: Mandarin price data

Figure 7. Prediction results with real data

future research, we will optimize the current feature selection criterion, which is currently set based on the median of ranking, using the Wilcoxon rank sum test. Additionally, we aim to enhance the robustness by applying our method to high-dimensional datasets in various industries. Furthermore, we anticipate expanding the applicability by combining it with TimeSHAP, which considers sequence information, and exploring its potential in the context of classification problems.

Acknowledgment. This work was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2023RIS-007) and supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2023-00218913).

REFERENCES

[1] T. Niu, J. Wang, H. Lu, W. Yang and P. Du, Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting, *Expert Systems with Applications*, vol.148, 113237, 2020.

[2] W. E. Marcílio and D. M. Eler, From explanations to feature selection: Assessing SHAP values as feature selection mechanism, *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020.

[3] B. Venkatesh and J. Anuradha, A review of feature selection and its methods, *Cybernetics and Information Technologies*, vol.19, no.1, pp.3-26, 2019.

[4] A. Thakkar and R. Lohiya, Fusion of statistical importance for feature selection in deep neural network-based intrusion detection system, *Information Fusion*, vol.90, pp.353-363, 2023.

[5] H. Lee, D. Kim, S. Lee, H. Park, H. Bae and K. Choi, Ensemble method for feature selection in time-series prediction based on image similarity, *ICIC Express Letters, Part B: Applications*, vol.14, no.7, pp.709-717, 2023.

[6] S.-X. Lv and L. Wang, Multivariate wind speed forecasting based on multi-objective feature selection approach and hybrid deep learning model, *Energy*, vol.263, 126100, 2023.

[7] A. M. Vommi and T. K. Battula, A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study, *Expert Systems with Applications*, vol.218, 119612, 2023.

[8] R. R. Mostafa et al., An improved gorilla troops optimizer for global optimization problems and feature selection, *Knowledge-Based Systems*, vol.269, 110462, 2023.

[9] E. H. Houssein et al., Boosted sooty tern optimization algorithm for global optimization and feature selection, *Expert Systems with Applications*, vol.213, 119015, 2023.

[10] M. Vijayan, S. S. Sridhar and D. Vijayalakshmi, A deep learning regression model for photonic crystal fiber sensor with XAI feature selection and analysis, *IEEE Transactions on NanoBioscience*, 2022.

[11] G. Hooker, L. Mentch and S. Zhou, Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance, *Statistics and Computing*, vol.31, pp.1-16, 2021.

[12] P. Nagaraj et al., A prediction and recommendation system for diabetes mellitus using XAI-based LIME explainer, *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2022.

[13] D. Fryer, I. Strümke and H. Nguyen, Shapley values for feature selection: The good, the bad, and the axioms, *IEEE Access*, vol.9, pp.144352-144360, 2021.

[14] C. Panati, S. Wagner and S. Brüggenwirth, Feature relevance evaluation using Grad-CAM, LIME and SHAP for deep learning SAR data classification, *2022 23rd International Radar Symposium (IRS)*, 2022.

[15] H. Xiao et al., Shapley-NAS: Discovering operation contribution for neural architecture search, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[16] H. Acikgoz, A novel approach based on integration of convolutional neural networks and deep feature selection for short-term solar radiation forecasting, *Applied Energy*, vol.305, 117912, 2022.