Journal of The Korea Convergence Society
Vol. 13. No. 3, pp. 33-41, 2022

e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2022.13.03.033

Linear interpolation and Machine Learning Methods for Gas Leakage Prediction Base on Multi-source Data Integration

Khongorzul Dashdondov[1], Kyuri Jo[2], Mi-Hye Kim[3*]
[1]Post-Doctor, Department of Computer Engineering, Chungbuk National University
[2]Assistant Professor, Department of Computer Engineering, Chungbuk National University
[3]Professor, Department of Computer Engineering, Chungbuk National University

다중소스 데이터 융합 기반의 가스 누출 예측을 위한 선형 보간 및 머신러닝 기법

홍고르출[1], 조겨리[2], 김미혜[3*]
[1]충북대학교 컴퓨터공학과 박사, [2]충북대학교 컴퓨터공학과 조교수, [2]충북대학교 컴퓨터공학과 교수

Abstract In this article, we proposed to predict natural gas (NG) leakage levels through feature selection based on a factor analysis (FA) of the integrating the Korean Meteorological Agency data and natural gas leakage data for considering complex factors. The paper has been divided into three modules. First, we filled missing data based on the linear interpolation method on the integrated data set, and selected essential features using FA with OrdinalEncoder (OE)-based normalization. The dataset is labeled by K-means clustering. The final module uses four algorithms, K-nearest neighbors (KNN), decision tree (DT), random forest (RF), Naive Bayes (NB), to predict gas leakage levels. The proposed method is evaluated by the accuracy, area under the ROC curve (AUC), and mean standard error (MSE). The test results indicate that the OrdinalEncoder-Factor analysis (OE-F)-based classification method has improved successfully. Moreover, OE-F-based KNN (OE-F-KNN) showed the best performance by giving 95.20% accuracy, an AUC of 96.13%, and an MSE of 0.031.

Key Words : Natural Gas, Leak prediction, Linear Interpolation, K-nearest neighbors, Convergence

요 약 본 논문에서는 다중 요인을 고려한 천연 가스 누출 정도 예측을 위해 관련 요인을 포함하는 기상청 자료와 천연가스 누출 자료를 통합하고, 요인 분석을 기반으로 중요 특성을 선택하는 머신러닝 기법을 제안한다. 제안된 기법은 3단계 절차로 구성되어 있다. 먼저, 통합 데이터 셋에 대해 선형 보간법을 수행하여 결측 데이터를 보완하는 전처리를 수행한다. 머신러닝 모델 학습 최적화를 위해 OrdinalEncoder(OE) 기반 정규화와 함께 요인 분석을 사용하여 필수 특징을 선택하며, 데이터 셋은 k-평균 클러스터링으로 레이블을 지정한다. 최종적으로 K-최근접 이웃, DT(Decision Tree), RF(Random Forest), NB(Naive Bayes)의 네 가지 알고리즘을 사용하여 가스 누출 수준을 예측한다. 제안된 방법은 정확도, AUC, 평균 표준 오차(MSE)로 평가되었으며, 테스트 결과 OE-F 전처리를 수행한 경우 기존 기법에 비해 성공적으로 개선되었음을 보였다. 또한 OE-F 기반 KNN(OE-F-KNN)은 95.20%의 정확도, 96.13%의 AUC, 0.031의 MSE로 비교 알고리즘 중 최고 성능을 보였다.

주제어 : 천연 가스, 누출 예측, 선형 보간, K-최근접 이웃, 융합

*This research was financially supported by the Ministry of Trade, Industry, and Energy (MOTIE) of Korea under the "Regional Specialized Industry Development Program" (R&D, P0002072) supervised by the Korea Institute for Advancement of Technology (KIAT).

*Corresponding Author : Mi-Hye Kim(mhkim@cbnu.ac.kr)

Received December 28, 2021 Revised February 9, 2022
Accepted March 20, 2022 Published March 28, 2022

1. Introduction

A gas leak has had a severe impression on environmental pollution and health. Therefore, it is vital to predict the risk level related to gas leakage. The researchers have studied a pilot project of this mapping to explore the first step in understanding the effects of NG leaks in [1, 2]. In the past, research on gas leaks and accidents focused on accident responses through measurements and monitoring using physical sensors. In this study, we measured the NG mass using an OmniCube sensor, a technical specification for a sensor from the Korean company UPO[3].

The OmniCube, a major product, is a transmitter that measures, transmits, and stores voltage and current from sensors. It supports various physical quantity measurement sensors (vibration/ noise/ stress/ wind speed & direction) as Korea Certification (KC) certified products. It provides an IoT-based remote monitoring system, and predictive diagnosis real-time data integrated analysis solution. OmniCube-mini, an explosion-proof data logger that can be applied to explosion-proof areas where flammable or combustible materials are likely to cause fire or explosion, is a sensor-embedded product and is a multi-purpose transmitter with high precision and durability. It is a product that can be used without a separate gateway using dual antennas. When NG is released into the air, it creates enormous problems for air pollution and the environment. By integrating the data measured by OmniCube with the raw meteorological data, we can determine the level of gas leakage. This will expand the study of the level of flatulence and its impact on the future development of predicting the risk of chronic diseases.

The related work of this study [2,4] used OE normalization for the target feature of NG and k-means clustering to label CH4 for the data pre-processing part. In recent years, machine-based methods have been widely used in environmental engineering to predict natural gas leakage [5]. They used factor analysis to reduce the size of the high-dimensional features and achieve the cluster analysis properties [6,7].

Standard machine-learning procedures are distributed into supervised, semi-supervised, and unsupervised learning. The data we operate is unsupervised data without a label. This research aimed to compare the KNN, DT, RF, and NB [6, 8-9] classification procedures for feature selection using FA and OE normalization for risk prediction of NG leaks. The model calculated accuracy, MSE, and ROC curve.

The main contribution of this study is data implemented in the Korean company testing process. Also, we filled missing data using the linear interpolation method in the pre-processing data section. Then the OE-F-based machine learning (ML) method is used to predict risk detection of NG leakage in the experimental integrated data without labeling. Therefore, our approach is appropriate for the early prediction of NG leaks in the air. Accordingly, this research Identified the relationship between the gas and the environmental elements supposed by the gas to predict the level of gas leakage risk without immediately determining the gas leak data.

An outline of the article is as follows. Section 2 provides a proposed methods on linear interpolation and feature-selection based on factorial analysis for gas leak detections. The evaluation metrics is explained in Section 3. Section 4 presents the experimental dataset and methods used for comparison, and the results of comparative experiments. Finally, conclusions are generated in Section 5.

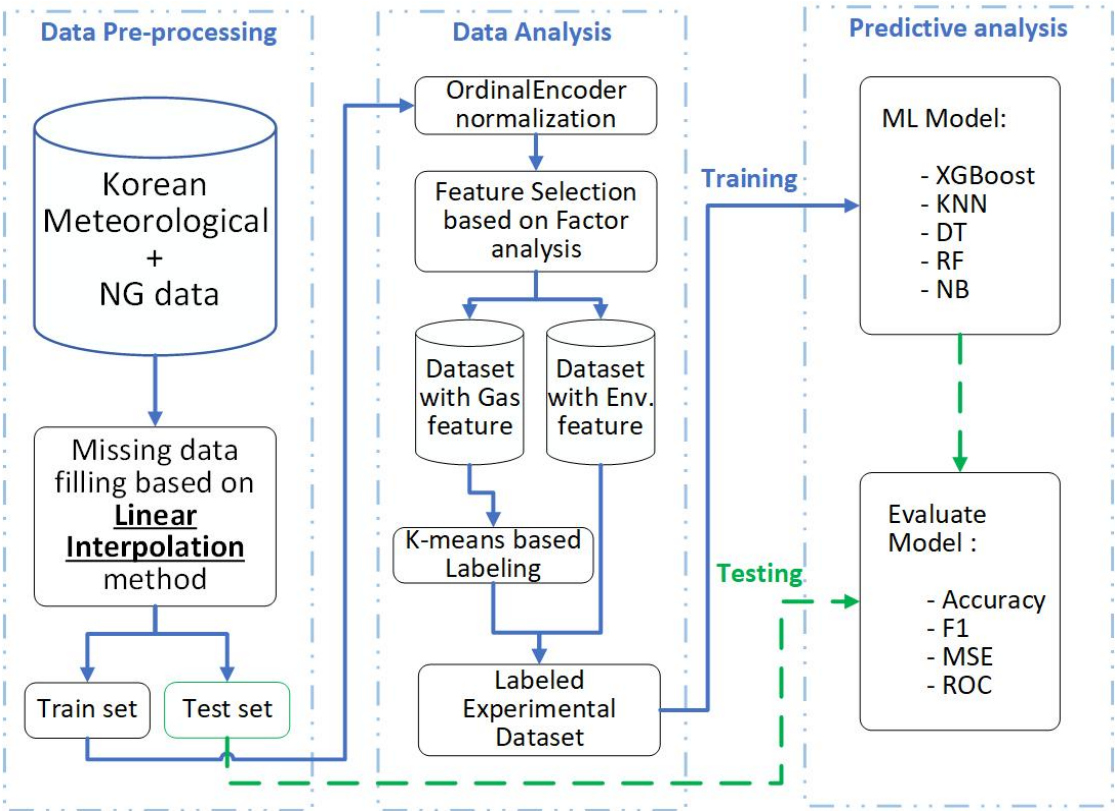

Fig 1. The general system architecture of the proposed method

2. Methodology

In this paper, we used device based methane gas data. Initially, we used date of "08/01/2020-08/31/2020" for UPO company testing data. This data collected from the location of GPS_LAT of 34.840702, and GPS_LONG of 127.675005. Then, we integrated UPO data with Korean meteorolog ical data [4]. Then filling a row of missing values using linear interpolation method.

Fig. 1 shows the architecture of the proposed approach. In this method, initially, we filled missing data using the linear interpolation method. Then we normalized data by the OE technique [10]. Behind the normalization, we selected essential features using the DT classifier. After that, we also decided on the feature selection (FS) method based on FA.+

We also divided data into two parts: gas and environment. We used unsupervised k-means clustering algorithms for labeling into CH4 gas data in the gas dataset. After labeling, we merged both data environment and gas by the labeled dataset and trained machine learning models for predictive analysis in this labeled dataset. After the training, we evaluated the prediction power of the models by accuracy measurements.

2.1 Linear Interpolation

An incomplete dataset can cause bias due to systematic differences between observed and unobserved data. These data usually contain missing values due to machine failure, changes in the siting of monitors, and human error. Thus,

this study uses the linear interpolation method (imputation technique) and substitution of the mean value for the replacement of missing values in the environmental data set [8-9,11]. Linear Interpolation means to estimate a missing value by connecting dots in a straight line in increasing order. In short, It estimates the unknown value in the same increasing order from previous deals. To perform a linear interpolation, we used Eq. (1).

$$y = y_1 + (x - x_1)\frac{y_2 - y_1}{x_2 - x_1} \quad (1)$$

where x is the independent variable and y is the value we want to look up [12].

2.2 Feature importance

We applied a DT classifier in the machine-lea rning method to select essential features in NG. If the importance score is more significant than zero, we will select features in this step. Our sug gested DT-based important features interpretabil ity prediction system of the integrated dataset h as described in Fig. 2. For the this dataset, "Vapo r_pressure", "Ground_temp_ 0cm,", "Humidy", "U nderground_ temp_20cm", "Dew_point_temp", "T otal_cloud_ volum", "Underground_Humidy_UPO ", "Underground_ temp_5cm", "Underground_tem p_10cm", "Under ground_Temp_UPO", "Wind_dir ection", "Under ground_temp_30cm", "Lower_mi ddle_layer_cloud", and "Wind_speed" were maint ained as most useful features with importance sc ores of 0.34188, 0.1898, 0.11804, 0.09892, 0.0749 9, 0.0495, 0.02636, 0.02461, 0.02175, 0.01948, 0.0 1732, 0.00976, 0.0039, and 0.00369 to predict DT among the environment values shown Fig. 2.

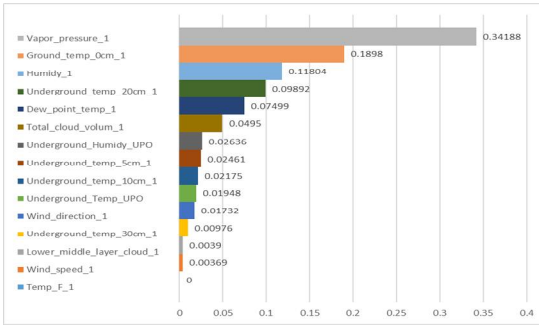

Fig. 2. The feature importance of the DT in the dataset

2.3 Ordinal Encoder

The normalization technique organizes a database to minimize duplicate and redundancy data. We encode categorical variables as an integer array. The input of this transformer is identical to the integer or a string array and represents a value obtained according to the category (discrete) characteristics. This section converts features into ordinal integers. As a result, one integer column (0 to n-1) appears in one element, and n is the number of categories. We implemented OE normalization for all selected components [2]. Fig. 3 shows plots of CH4 with and without OE.

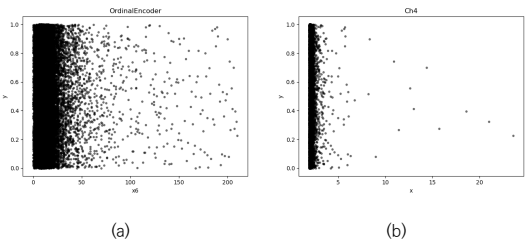

Fig 3. Normalization of NG data with and without OE. (a) with OE, (b) without OE.

2.4 Factor analysis

The FA is used to reduce the number of variables to a few factors. This method calculates and scores the most common variance of all variables.

In this study, we used the most common approach for principal component analysis for the extracted factor from the data set. PCA extracts the maximum variance and sets them into the first factor [5]. Then the variance explained by the first attribute is calculated, and the maximum variance of the second attribute is calculated. This process moves to the last element. We also analyzed factors for the dependent variable, CH4, with other environmental elements. Table 1 shows the communality extraction, means, and standard deviations of the features. The number of cases has N=2565.

The model shows the correlation between the variables because KMO factor = 0,786, sample sizes are compatible, and significance is equal to 0.001 (alpha<0.005), rejected the null hypothesis. Here, we can perform factor analysis.

Table 1 shows the communalities value for each feature and extracted by PCA. The outcomes have displayed the percentage of features explained by the coefficient for the given variable. In other words, we obtained the result indicating 95.4% of the variation in Underground_Temp_UPO.

The results suggested the best job of explaining variation in 99.1% of Vapor_pressure, Dew_point_temp, and Underground_temp_20cm; 98.5% of Underground_temp_10cm and CH4; 96.2% of Humidy, 96.0% of Ground_temp_0cm. In this case, the model is better for some variables than for others. The model explains above features are the best for CH4 features.

In additionally, not worse for other features such as Underground_temp_30cm, Underground _Humidy_UPO, Wind_speed, Wind_direction, Total_cloud_volum, Lower_middle_layer_cloud, and Underground_temp_5cm. This model extracted by PCA.

Table 1. Descriptive statistics and Communality extraction of Factor model (Initial = 1, Extraction Method: Principal Component Analysis)

Features	Mean	Std. Deviation	Communality Extraction
Underground_Temp_UPO	49.64	34.14	.954
Underground_Humidy_UPO	32.95	14.23	.809
Wind_speed	172.09	119.11	.849
Wind_direction	68.59	38.33	.808
Humidy	151.76	61.66	.962
Vapor_pressure	242.93	94.40	.991
Dew_point_temp	163.13	58.78	.991
Total_cloud_volum	35.54	21.76	.869
Lower_middle_layer_cloud	24.09	20.06	.865
Ground_temp_0cm	264.92	217.25	.960
Underground_temp_5cm	154.48	91.77	.930
Underground_temp_10cm	133.26	72.55	.985
Underground_temp_20cm	118.91	65.44	.991
Underground_temp_30cm	105.98	60.02	.954
CH4	32.47	32.97	.985

Table 2 shows a rotated component matrix of the suggested factor model. We also extracted the PCA method and rotated Varimax with the Kaiser normalization method, which converged in 10 iterations. In this case, we extracted six components. The rotated component matrix is called the load, which is the main output for the principal component analysis, which then includes calculating the relationships between each variable and the calculated components.

In Table 2, moderate to strong correlations be tween Underground_temp_10cm, Underground_ temp_20cm, Underground_temp_30cm, Under gr ound_temp_5cm, and component 1 (here Under ground_Humidy_UPO shows a positive correlatio n with component 1); Ground_temp_0cm, Under ground_Temp_UPO, Humidy, and component 2; Dew_point_temp, Vapor_pressure, and compone nt 3 are shown.

Furthermore, Lower_middle_layer_cloud, Total _ cloud_volum, and component 4; Wind_speed, Wind_direction, and component 5; CH4 and component 6 were found, with other features and each component showing very low correlations.

Table 2. The Rotated Component Matrix

Factors	Components					
	1	2	3	4	5	6
Underground_temp_10cm	.926					
Underground_temp_20cm	.926					
Underground_temp_30cm	.880					
Underground_temp_5cm1	.771					
Underground_Humidy_UPO	-.50					
Ground_temp_0cm		.968				
Underground_Temp_UPO		.941				
Humidy		-.75				
Dew_point_temp			.961			
Vapor_pressure			.957			
Lower_middle_layer_cloud				.770		
Total_cloud_volum				.768		
Wind_speed					.860	
Wind_direction					.821	
CH4						.897

2.5 K-means Clustering

In this subsection, we will define one of the standard ML algorithms for K-means clustering used for the class of CH4. The main theory distributes the character to the nearest class n values into k subgroups. We split into three classes for gas leakage by low, medium, and high.

Fig. 4 illustrates k-means clustering results by the histogram for NG. The cluster values have low-1190, medium-695, and high-680. From the histogram of the clusters, it can see that the imbalanced data with CH4.

3. Evaluation Metrics

The performance evaluation of this paper was completed using accuracy, AUC, and MSE. Precision is a fraction of the true positive (TP) predictions among all positive predictions (PP), and recall measures what proportion of actual positives (AP) were identified correctly. We can find precisions and recall as follows [4,13]:

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{PP} \tag{2}$$

and

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{AP} \tag{3}$$

The accuracy measures the degree for the nearness of the calculated value to its actual weight. Accuracy is a proportion of correct classification among all the classes, expressed as Eq. (4).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} = \frac{TP+TN}{AP+AN} \tag{4}$$

Besides, another evaluation metric used the MSE [14] for the predicted leaks relative to actual values.

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(X(i,j)-Y(i,j))^2 \tag{5}$$

with m and n mentioned the number of observations, where m is the number of data cases and n is serves to predict NG. There, X and Y are the concrete and predicted values for the i- and j- th data attributes, respectively.

4. Experimental Study

We built the dataset is integrated from the Korean Meteorological Agency and natural gas leakage data from the UPO company test data.

The data was collected by UPO on a trial basis from August 1 to August 31, 2020, in Jeollanamdo province. Initially, we integrated UPO [3] data with Korean meteorological data [15] and then filled a row of missing values using the linear interpolation method. First we have a total of 2565 records, 5 features include location of GPS_LAT of 34.840702, and GPS_LONG of 127.675005. Then we added location environment weather data from the Korean meteorological web application. Therefore

features increased 5 to 16 include the NG attribute. Afterward, we selected features according to the OE-F model of NG level detection, which was, in this case, seven features.

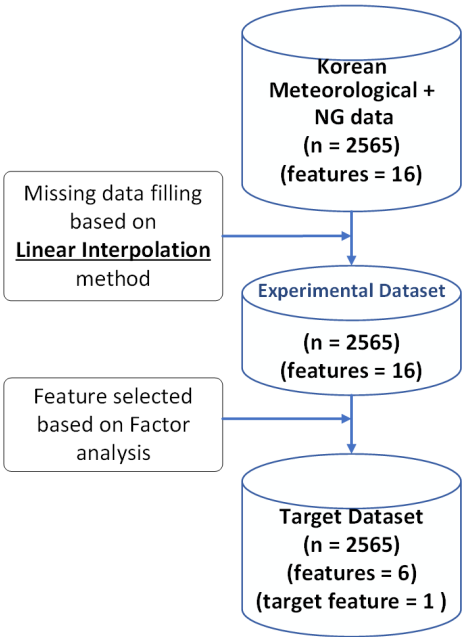

Fig. 5. The experimental dataset preparation procedure of the NG dataset

Fig. 5 shows the procedure used to create the target dataset. Regarding the default settings of the training (70%) and testing (30%) sets, we defined three class labels, low, medium, and high, for the target features of an experimental dataset.

Table 3 describes the descriptive statistics for the target variables.

Table 3. Descriptive statistics of levels for target dataset

Class	Total	Train 70%	Test 30%
Low	1190	843	347
Medium	695	481	214
High	680	471	209
Total	**2565**	**1796**	**770**

We compared the proposed OF-F model with KNN, DT, RF, and NB. According to the compared algorithms, we selected the best values of the input parameters by changing these values until the performance increased.

Table 4. Evaluation performance of the compared algorithms on the target dataset (%)

Algorithms	Accuracy	AUC	MSE
OE—F-KNN	**95.20**	**96.13**	**0.031**
OE-F-DT	89.22	94.88	0.045
OE-F-RF	93.45	95.59	0.033
OE-F-NB	81.17	89.09	0.078
KNN	94.58	92.40	0.029
DT	93.31	95.18	0.011
RF	93.31	95.18	0.011
NB	80.66	82.49	0.077

The accuracy, AUC, and MSE measurements of the performance results are shown in Table 4. The OE-F-KNN algorithm had the highest accuracy of 95.20%, AUC of 96.13, and MSE of 0.031 than other algorithms such as DT, RF, and NB. Table 4 shows that they were reduced from 15 dimensions to 6 dimensions by factor analysis, the accuracy of gas leak detection increased in OE-F-based KNN, RF, and NB algorithms. Then indicates that the proposed factor analysis-based feature reduction method is suitable for predicting gas leak detection.

We provided multi-class ROC curves for each compared model in the experimental dataset in Fig. 6. As noted above, we proposed to find better model performance to predict KNN for this dataset.

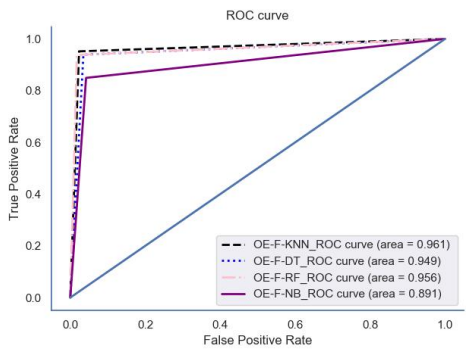

Fig 6. Multi-class ROC curves by the compared OE-F algorithms for all class

5. Conclusion

This paper presents predicted natural gas leakage levels determined using the linear interpolation method for the filling missing values and feature selection based on a factor analysis of actual integrated data. We conducted training with OrdinalEncoder normalization on an experimental dataset and then classified gas data using the k-means clustering method. The investigation revealed various component factors associated with gas leakages. The OE-F method analysis found that gas leaks were closely related to temperature, humidity, and other wind factors. According to the test results, the proposed OE-F-KNN algorithm has accuracy, MSE, and AUC outcomes of 95.2%, 96.13%, and 0.031, respectively. The results here demonstrate that the proposed method is suitable for the early prediction of NG losses in the air. The system was implemented using SPSS and Python and tested its performance on actual open data.

REFERENCES

[1] Ministry of Public Safety and Security. (2019) 2019th Yearbook of Disaster, Ministry of Public Safety and Security; Ministry of Public Safety and Security: Sejong, Korea.

[2] D. Khongorzul, M. H. Kim & S. M. Lee. (2019). OrdinalEncoder based DNN for Natural Gas Leak Prediction. *J. Korea Convergence Society, 10(10),* 7-13.

[3] Available website: UPO company, http://www.upokorea.com/new/pdf/UPO_Catalogue.pdf

[4] D. Khongorzul & M. H. Song. (2022). Factorial Analysis for Gas Leakage Risk Predictions from a Vehicle-Based Methane Survey. *Applied Sciences 12(1),* 115.
DOI : 10.3390/app12010115

[5] Department for International Development. Live Data Page for Energy and Water Consumption. Available online: http://data.gov.uk/dataset/dfid-energy-and-water-consumption (accessed on 8 March 2021).

[6] USDT. Leak Detection Technology Study for PIPES Act; Tech. Rep.; U.S. Department of Transportation: Washington, DC, USA, 2007.

[7] M. Fagiani, S. Squartini, L. Gabrielli, M. Severini & F. Piazza. (2016). A statistical framework for automatic leakage detection in smart water and gas grids. *Energies, 9,* 665.
DOI : 10.3390/en9090665

[8] N. M. Noor, M. M. Al Bakri Abdullah, A. S. Yahaya & N. A. Ramli. (2015) Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. *Materials Science Forum, 803,* 278-281.

[9] C. M. Salgado, C. Azevedo, H. Proença & S. M. Vieira. (2016). Missing data. *Secondary analysis of electronic health records*, 143-162.

[10] Y. K. Kim & H. G. Sohn. (2018). *Disasters from 1948 to 2015 in Korea and power-law distribution. In Disaster Risk Management in the Republic of Korea*; pp. 77-97. Springer, Singapore.

[11] J. Peppanen, X. Zhang, S. Grijalva & M. J. Reno. (2016, September). Handling bad or missing smart meter data through advanced data imputation. In 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT) (pp. 1-5). IEEE.

[12] T. Kim, W. Ko & J. Kim. (2019). Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. *Applied Sciences, 9(1),* 204.

[13] D. Khongorzul, S. M. Lee, Y. K. Kim & M. H. Kim. (2019). Image Denoising Methods based on DAECNN for Medication Prescriptions. *Journal of the Korea Convergence Society, 10(5),* 17-26.
DOI : 10.15207/JKCS.2019.10.5.017.

[14] V. N. Vapnik. (1995). *The nature of statistical learning theory*. New York: Springer.

[15] Available website: Korean public data portal. https://www.data.go.kr/dataset/15000099/openapi.do

홍 고 르 출 (Khongorzul Dashdondov) [정회원]

· 2000년 12월 : 몽골국립대학교 수학과(이학사, 이학석사)
· 2013년 8월 : 충북대학교 전파통신공학과(공학박사)
· 2017년 3월 ~ 현재 : 충북대학교 컴퓨터공학과 연구원

· 관심분야 : Probability and Statistics, Queueing theory, Image processing, Machine Learning, Deep Learning
· E-Mail : khongorzul63@gmail.com

조 겨 리 (Kyuri Jo) [정회원]

· 2013년 2월 : 서울대학교 컴퓨터공학과 (공학사)
· 2018년 8월 : 서울대학교 전기컴퓨터공학부 (공학박사)
· 2019년 9월 ~ 현재 : 충북대학교 컴퓨터공학과 조교수

· 관심분야 : Machine Learning, Bioinformatics, Data Mining, Time-series Analysis, Network Biology
· E-Mail : kyurijo@chungbuk.ac.kr

김 미 혜 (Mi-Hye Kim) [정회원]

· 1992년 2월 : 충북대학교 수학과 (이학사)
· 1994년 2월 : 충북대학교 수학과 (이학석사)
· 2001년 2월 : 충북대학교 수학과 (이학박사)

· 2004년 9월 ~ 현재 : 충북대학교 컴퓨터공학과 교수
· 관심분야 : 빅데이터, 기능성 게임, 유비쿼터스 게임, 플랫폼, 퍼지측도 및 퍼지적분, 제스츄어 인식
· E-Mail : mhkim@cbnu.ac.kr