Deep AutoEncoder-based Framework for the Classification of Natural Gas Leaks Grade using Multivariate Outlier Detection

Khongorzul Dashdondov
Department of Computer Engineering, Chungbuk National University, Cheongju, Korea
khongorzul63@gmail.com

Mi-Hye Kim
Department of Computer Engineering, Chungbuk National University, Cheongju, Korea
mhkim@cbnu.ac.kr

Kyuri Jo†
Department of Computer Engineering, Chungbuk National University, Cheongju, Korea
kyurijo@chungbuk.ac.kr

ABSTRACT

Natural gas is widely used for domestic and industrial purposes, and we cannot actually directly know that it is being leaked into the air. The current problem is that gas leakage is not only economically harmful but also detrimental to health. Therefore, a lot of research has been done on the risk of gas damage and leakages, but research on predicting gas leakage is just being done. In this study, we propose a method based on deep learning to predict gas leakage from environmental data. Our proposed method has successfully improved the performance of machine learning classification algorithms by efficiently preparing training data using a deep autoencoder model. The proposed method was evaluated on an open dataset containing natural gas and environmental information and compared with extreme gradient boost (XGBoost), K-nearest neighbors (KNN), decision tree (DT), random forest (RF), and Naive Bayes (NB) algorithms. The proposed method is evaluated by the accuracy, F1-score, mean standard error (MSE), and area under the ROC curve (AUC). As a result, the presented method in this study outperformed all compared methods. Moreover, Deep Autoencoder and OrdinalEncoder-based XGBoost (DAE-OE-XGBoost) showed the best performance by giving 99.193% accuracy, an F1-score of 99.38%, an MSE of 0.004, and an AUC of 99.53%.

KEYWORDS

Gas leak prediction, Deep autoencoder, Outlier remove, K-means; XGBoost

ACM Reference format:

Khongorzul Dashdondov, Mi-Hye Kim, and Kyuri Jo 2022. Deep AutoEncoder-based Framework for the Classification of Natural Gas Leaks Grade using Multivariate Outlier Detection. In *Proceedings of ACM KDD conference (URBCOMP'22). ACM,* Washington, DC, *USA, 6 pages.* https://doi.org/

† The corresponding author.

1 Introduction

Predicting gas leakage early makes it possible to prevent future economic losses. In addition, the natural gas leakage can exacerbate adverse health effects, such as hypertension, pulmonary, exacerbates pneumonia, asthma, and other respiratory diseases. Therefore, gas leak detection is essential for gas-intensive countries. So far, we have found that very little research has been done to predict gas leakage. Although there are studies on the harmful effects of gas leaks [1-3], not enough research has been done to predict gas leaks.

This study proposes a novel method based on the deep learning method that predicts gas loss by combining gas data with environmental data. The proposed method consists of three main modules: data pre-processing, data labeling, and predictive analysis. The data pre-processing module removes outlier using deep autoencoder reconstruction error and normalizes the data using OE and LN transformation techniques. The data labeling module selects only natural gas (NG) CH4 data from the data pre-processed data, divides it into groups using the K-means clustering algorithm, and classifies the data according to that group. Afterward, the predictive analysis module then builds a model that predicts gas loss using machine learning algorithms on the available data.

In other words, models created according to the proposed method improve the prediction results better than constructing a predictive model using machine learning algorithms on the data without pre-processing the data.

The main contribution of this paper is the following novelty:

- We have proposed a novel method based on deep learning to predict gas leakages by removing outliers with deep autoencoder model.
- We evaluated the proposed method on real data open dataset and can be used to compare the results in other research works. In addition, the study was implemented using actual open data that had not previously been used with the ML algorithm, which future researchers will widely use for comparative research.
- We compared the proposed method with baseline models based on extreme gradient boost (XGBoost), K-nearest neighbors (KNN), decision tree (DT), random forest (RF), and Naive Bayes (NB) algorithms showed improved performance.

An outline of the article is as follows. Section 2 provides a detailed survey of related work. The proposed method is explained in Section 3. Section 4 presents the experimental dataset, the methods used for comparison, the evaluation metrics, and the results of comparative experiments. Finally, conclusions are generated in Section 5.

2 Related works

Researchers have studied a pilot project of this mapping approach to explore the first step in understanding the effects of NG leaks [3]. Zhu et al. [4] proposed a regression-based deep belief network (DBN) model to predict the amount and rate of valve gas leakage in a natural gas pipeline system. Several studies used statistical and machine learning-based methods to detect gas leakage in residential and construction environments in water and natural gas networks [5-6]. In [7], advantages of the statistical shape analysis (SSA) method have been presented in comparison to principal component analysis (PCA), discrete wavelet transforms (DWT), and polynomial curve fitting (PCF) algorithm for improvement of detection selectivity. Also, Song et al. [8] present a gas leak detection method for galvanized steel pipes based on acoustic emission. A machine-based approach to environmental engineering has been widely used to predict natural gas leaks. Our previous research [9] used an ordinal-encoder (OE) normalization and k-means clustering for the data preprocessing section. However, we improved the performance of our previous study by using a deep autoencoder-based outlier removal process. Classification methods are trained on data being normally distributed. In very rare cases, learning from data with outlier errors reduces the ability to predict other standard distributed data. Therefore, in this study, we aimed to show that we can improve the power of the model by first removing the outlier values during training and then training the model on the most normally distributed data.

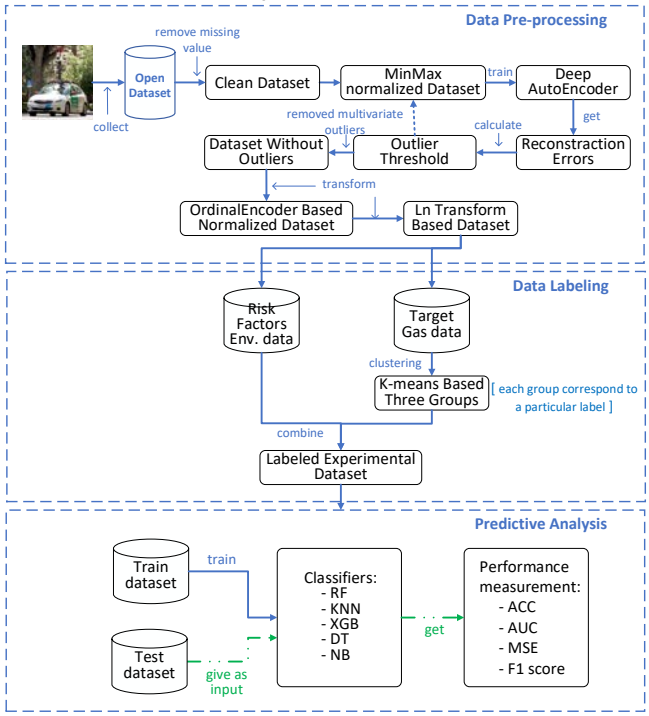

Figure 1. General architecture of the proposed method.

Autoencoder is widely used for reducing data dimensions by learning data representation [10-12]. The authors of [13] used a clustering algorithm and reconstruction error from the deep autoencoder model to detect outliers in unsupervised mode. Another usage of the autoencoder is that remove image denoising [14-18] and time-series data.

3 Methodology

The proposed approach has three modules: data preprocessing, data labeling, and predictive analysis. The general architecture of the proposed method is presented in Figure 1. The first module uses deep autoencoder, OrdinalEncoder, and Ln transformation techniques. As a result of this module, normalized clean data is passed to the k-means algorithm in the next module for data labeling. After that, several machine learning algorithms are trained using the prepared experimental data.

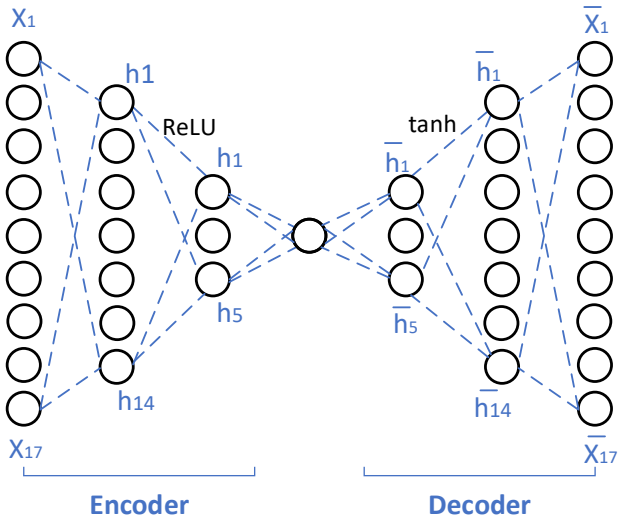

Figure 2: Structure of the proposed Deep AutoEncoder method.

3.1 Data pre-processing

We use Deep Autoencoder to clean our data. The autoencoder is an unsupervised artificial neural network that learns how to efficiently compress and encode data and then reconstruct the data from the reduced encoded representation to a representation close to the original input as possible [15]. The structure of AE consists of encoder and decoder parts. The encoder part compresses input data by reducing data dimension, while the decoder part reconstructs the compressed data into output. Thus, the number of input neurons equals the number of output neurons in AE. Reconstruction error of autoencoder is a difference between input and its reconstructed output. Figure 2 shows the structure of the proposed autoencoder model in this study. Firstly, it projects input X to a lower dimension that works in the encoder part; then, it reconstructs output X' from the low dimensional projection in the decoder part. Sequentially, the proposed autoencoder has five hidden layers with 17, 14, 5, 1, 5, 14, and 17 nodes. Moreover, hidden layers in the encoder part use the "ReLU" activation function, and hidden layers in the decoder part use the "tanh" activation function. In summary, 17 features after the min-max normalization are used to train the

autoencoder where the activation functions for encoder and decoder are rectified linear unit (ReLU) and hyperbolic tangent (tanh), respectively. In other words, the learning process of AE is that it compresses the input into a lower-dimensional space called latent space and uncompressed back the compressed data into the output that closely matches the original data. Then, it calculates a difference between the input and reconstructed output and changes the network weights to reduce this difference.

First, we trained the deep autoencoder model on the whole dataset. Then we calculated reconstruction errors of them by the mean of the squared difference between input and output described in expression (1):

$$RE = \frac{1}{m}\sum_{i=1}^{m}\left\|x_i - x_i^{'}\right\|_2^2 \quad (1)$$

where *m* is the number of records, x is the original input, and x' is the reconstructed input.

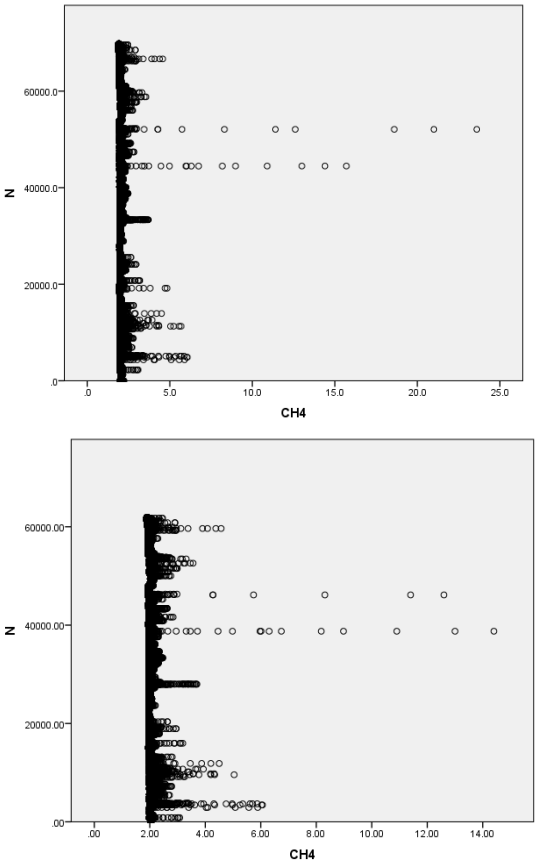

Figure 3: Plots of CH4 data with and without outliers.

Fig. 4 shows data with and without outliers from the dataset by a number of values. Fig. 4a shows the original dataset with outliers. Fig 4b shows based on the DAE method of the dataset without outliers. After that, the outlier threshold value is estimated by summing up the average reconstruction error and standard deviation. Then, if the reconstruction error of data exceeds the threshold value, this data will be removed from the dataset. This module's last step is to normalize outlier removed data using OE and Ln transformation technique. The advantage of normalization in machine learning is that their 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 [9]. Figure 4 shows plots of CH4 with and without OE.

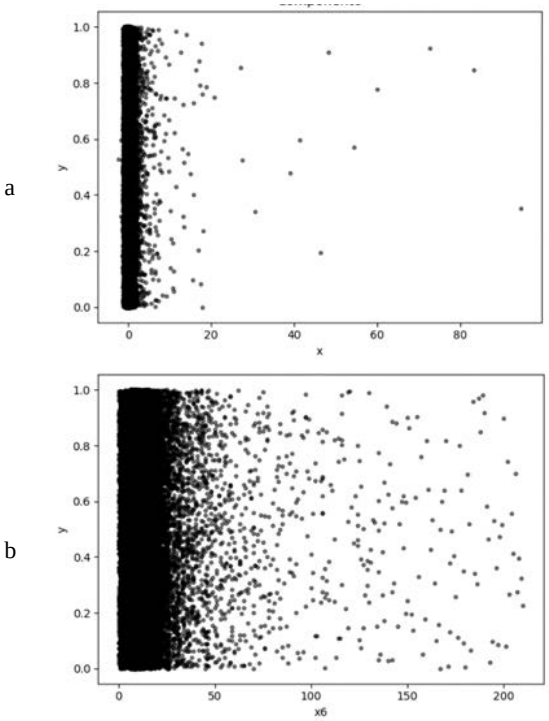

Figure 4: Plots of CH4 data with OE normalization data for NG: (a) CH4 and (b) normalization OE of CH4.

3.2 Data labeling

This module selects the CH4 feature, which is the value of methane from the preprocessed dataset for data labeling. The first open dataset has no label. Therefore, we used the simple and commonly used k-means algorithm to make the label for our outlier removed dataset. K-means is a multi-variable clustering method developed by MacQueen in 1967 [19]. The basic idea is to divide the samples into k subgroups of n samples in the most comparable class. First, all samples belong to group k. The method then calculates the Euclidean norm between the samples and the core point of each cluster. It computes the Euclidean norm until the allocation of all samples is no longer changed. We assign the class label as low, medium, or high based on the result of the k-means algorithm. Finally, we combined the class labels with the outlier removed

dataset except for the CH4 feature because the CH4 feature is used to determine class labels.

3.3 Predictive analysis

We trained machine learning-based RF, KNN, XGBoost, DT, and NB algorithms on our experimental dataset. We split the training dataset base 70% for training, and 30% for the testing.

NB: The Naïve Bayes is a probability-based classification algorithm [20]. It computes the probability for each class label and selects the class label with has the highest probability. It calculates the probability by considering all features separately; it is called conditional independence.

KNN: The k-nearest neighbor algorithm is used for classification purposes [20]. First, a user defines the value of the k parameter which is the number of nearest samples used to predict. Then all distances between test data and the training dataset are calculated and sorted by descending order. Finally, the top k number of instances from the ordered dataset is used to predict the class label. The majority voted class label will be assigned to the output label.

DT: The decision tree classifier is an interpretable label and a commonly used algorithm [21]. It builds a model to predict the target variable via decision rules trained from the data.

RF: The random forest is a type of ensemble algorithm [23]. It consists of several decision tree classifiers trained in different sub-samples of the whole dataset. For the prediction, the majority voted class label of these decision trees will be chosen as output.

XGBoost: XGBoost uses a method called CART (Classification and Regression) in which all leaves are related to the final score of a model, unlike the decision-making tree that only considers the result values of leaf nodes [22]. While a common decision-making tree is interested in how well the classification has been done, CART enables to even compare superiority among models that retain identical classification results.

4 Experimental Study

4.1 Dataset

We used the open gas leak dataset from [23]. Natural gas (NG) masses were measured using a Picarro CH4 sensor and a Google Street view machine [1]. This refers to gas sensors that are resistant to fire and wired and wireless transmitters that can be used in high-sensitivity facilities. In addition, the vehicle used is an IoT-based remote monitoring system, with a dual-antenna diagnostic solution used for real-time data aggregation analysis. Further, we present a list of environmental and gas features in raw data properties of NG found in mobile-device-based methane gas research [1, 2]. The environmental features are DasTemp, OutletValve, GPS_ABS_LAT, GPS_ABS_LONG, WS_WIND_LAT, EtalonTemp, WarmBoxTemp, WIND_N, WS_WIND_LON, WS_SIN_HEADING, WIND_E, WIND_DIR_SDEV, WS_COS_HEADING, CavityTemp, CAR_SPEED, CavityPressure, and gas feature is CH4. Initially, we removed a row of missing values and features unrelated to gas leaks, after which there were a total of 69,824 records from 78,771 records originally and 17 features from 33. After removing outliers from 69828 records, the 61148 records remained. Table 1 shows the number of records in each class for the testing and training process.

Table 1. The number of records in each class

Class	Total	Train 70%	Test 30%
Low	20381	14255	6126
Medium	20508	14379	6129
High	20259	14169	6090
Total	**61148**	**42803**	**18345**

4.2 Evaluation Metrics

The performance evaluation of this paper was completed using accuracy, AUC, F1-score, and MSE. We can find precisions and recall as follows [4]:

$$Precision = \frac{TP}{(TP+FP)} \text{ and } Recall = \frac{TP}{(TP+FN)} \quad (2)$$

The F1 score is the harmonic mean of precision and recall as follows:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (3)$$

We have studied the multi-class case, and there the average of the F1-score of each class label with weighting depending on the average parameter as Eq. (3).

The accuracy is a measure of the degree of the nearness of the calculated value to its actual value. Accuracy is the sum of true positive fraction and true negative fraction among all the test data as Eq. (4).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

In addition, one of our evaluated metrics is the mean squared error (MSE) for the predicted leaks relative to actual values was used:

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

with m and n being the number of observations, which m is the number of data and n is predicting NG. The X and Y are the actual and predicted values for the i, j - th data point, respectively.

4.3 Performance evaluation

Data preprocessing and predictive analysis modules were implemented in Python using the sklearn library [24]. The data labeling module was performed in SPSS 23.0.

First, we measured the performance of baseline models to compare them with our proposed method. We trained baseline models on the raw dataset directly using machine learning algorithms shown in Figure 1. Also, OE-based baseline models are trained on the dataset without removing outliers. Table 2 shows the compared performances of the baseline model and the proposed method. As a result, we can see that OE-based data normalization can improve the performance of models that were trained on raw datasets. Moreover, the combination of deep autoencoder-based outlier removal and OE-based data normalization in the proposed methods outperformed all compared baselines. The accuracy, F1-score, MSE, and ROC curve measurements of the performance results are

shown in Table 2, where the highest values of evaluation scores are marked in bold. The KNN model showed the best accuracy of 98.57%, and it improved to 98.62% when using OE-based normalization on the baseline model. The XGBoost algorithm gave the best result from all the compared models, with an accuracy rate of 99.193%, an F1-score of 99.38, an MSE of 0.004, and an ROC of 99.53%. The DAE-OE-RF model achieved the second-best accuracy rate of 99%.013, an F1-score of 99.23%, an MSE of 0.005, and an AUC of 99.41%. The DAE-OE-NB model showed lower results compared to the other proposed predictive models of the evaluation metrics.

Table 2. Evaluation results of the compared algorithms on the experimental dataset (%).

	Classifier Algorithms	Accuracy	AUC	MSE	f1-score
Proposed method	**DAE-OE-RF**	**99.013**	**99.41**	**0.005**	**99.23**
	DAE-OE-KNN	98.872	99.16	0.007	98.88
	DAE-OE-XGB	**99.193**	**99.53**	**0.004**	**99.38**
	DAE-OE-DT	98.158	9912	0.007	98.82
	DAE-OE-NB	85.849	94.18	0.05	92.44
Baseline models	OE-RF	98.506	99.11	0.007	98.64
	OE-KNN	98.621	98.97	0.009	98.64
	OE-XGB	98.735	99.28	0.006	98.9
	OE-DT	97.241	98.59	0.01	98.2
	OE-NB	78.74	89.79	0.08	86.19
	RF	92.25	91.34	0.05	91.71
	KNN	98.573	98.88	0.006	98.96
	XGB	92.258	91.34	0.005	91.71
	DT	92.258	91.34	0.005	91.71
	NB	48.671	85.42	0.18	78.78

We provided multi-class ROC curves for each compared model in the experimental dataset in Figure 5. As noted above, we proposed to find better model performance to predict XGBoost and RF for this dataset.

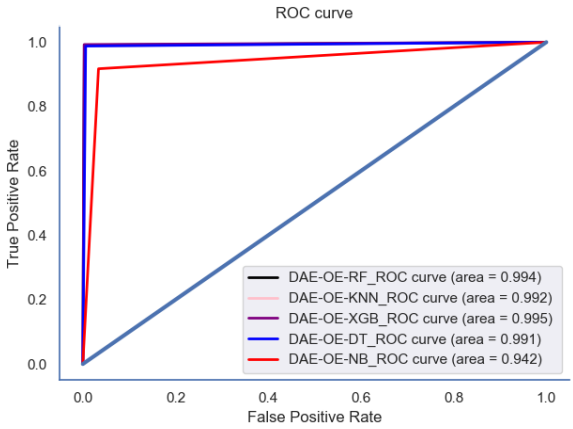

Figure 5: Receiver operating characteristic curves of the algorithms compared to the DAE-OE method.

Finally, we compared our proposed methods to show the effects of different modules by XGB, RF, and NB algorithms, as shown in Figure 6. As mentioned in section 3, module 1 is data processing, module 2 is data labeling, and module 3 is predictive analysis.

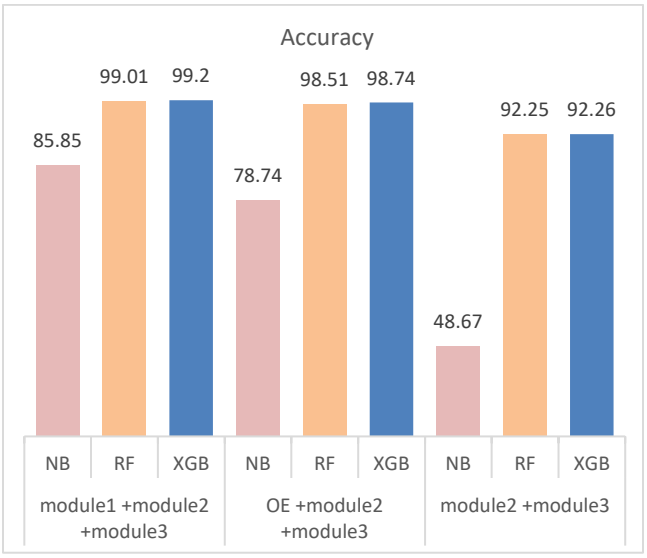

Figure 6: Comparison of the proposed modules and other guidelines.

5 Conclusion

This study proposed a method consisting of three modules to predict gas leakage. The preparation of efficient training data through data preprocessing and data labeling modules has dramatically improved the productive performance of machine learning algorithms. It is also possible to use this method to create gas leakage data levels for air assessments in Korea. In other words, we used a deep autoencoder model to distinguish highly distorted parts from the raw dataset, and the AE model fits the more commonly distributed majority dataset to reconstruct them with a minor error. Therefore, outliers can be distinguished by the AE model easily. The data were normalized using OE, and then Ln transformations, k-means clustering, and the experimental data were ready. The DAE-OE-XGB model had the best results from constructing a predictive model using RF, KNN, XGB, DT, and NB algorithms on the prepared experimental dataset. According to the test results, the proposed DAE-OE-XGBoost algorithm has accuracy, F1-score, MSE, and AUC outcomes of 99.193%, 99.38%, 0.004, and 99.53%, respectively.

ACKNOWLEDGMENTS

This research was financially supported by the Ministry of Trade, Industry, and Energy (MOTIE), Korea, under the “Regional Specialized Industry Development Program (R&D, P0002072)” supervised by the Korea Institute for Advancement of Technology (KIAT) and supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center

support program(IITP-2022-2020-0-01462) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).

REFERENCES

[1] Weller, Z. D., Yang, D. K., Fischer J. C., 2019. An open-source algorithm to detect natural gas leaks from mobile methane survey data. PLOS ONE 14, 2, e0212287. DOI:

[2] Joseph C. von Fischer, Daniel Cooley, et. al., 2017. Rapid, Vehicle-Based Identification of Location and Magnitude of Urban Natural Gas Pipeline Leaks. Environmental Science & Technology, 51, 7, 4091-4099. DOI: 10.1021/acs.est.6b06095.

[3] Ju, Y.-M., Lee, H.-S. and Oh, J.-C., 2018. Design and Implementation of Gas Leakage Alarm IoT System for Safety Helmet. 13, 6, 1411–1416. doi: 10.13067/JKIECS.2018.13.6.1411.

[4] Zhu, S.B., Li, Z.L., Zhang, S.M. and Zhang, H.F., 2019. Deep belief network-based internal valve leakage rate prediction approach. Measurement. 133, 182-192.

[5] Lee, Y.-H., 2020. A Study on the Damage Range According to Leakage Scenarios in Natural Gas Pipeline of LNG Fueled Ship. Journal of the Korean Society of Marine Environment and Safety. The Korean Society of Marine Environment and Safety, 26,4, 317–326. doi: 10.7837/kosomes.2020.26.4.317.

[6] Lee, J. Ah and Kim, M.-H., 2018. Gas Safety Monitoring App. Development Design for Gas Workers, Journal of the Korea Convergence Society, 9, 10, 61–67. doi: 10.15207/JKCS.2018.9.10.061.

[7] Krivetskiy, V.V., Andreev, M.D., Efitorov, A.O. and Gaskov, A.M., 2021. Statistical shape analysis pre-processing of temperature modulated metal oxide gas sensor response for machine learning improved selectivity of gases detection in real atmospheric conditions. Sensors and Actuators B: Chemical, 329, p.129187.

[8] Song, Y. and Li, S., 2021. Gas leak detection in galvanised steel pipe with internal flow noise using convolutional neural network. Process Safety and Environmental Protection, 146, 736-744.

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

[10] Lyudchik, O., 2016. Outlier detection using autoencoders (No. CERN-STUDENTS-Note-2016-079).

[11] Chen, J., Sathe, S., Aggarwal, C. and Turaga, D., 2017. Outlier detection with autoencoder ensembles. In Proceedings of the 2017 SIAM international conference on data mining, Society for Industrial and Applied Mathematics, 90-98.

[12] Kieu, T., Yang, B., and Jensen, C.S., 2018, June. Outlier detection for multidimensional time series using deep neural networks. In 2018 19th IEEE International Conference on Mobile Data Management (MDM), 125-134.

[13] Amarbayasgalan, T., Jargalsaikhan, B. and Ryu, K.H., 2018. Unsupervised novelty detection using deep autoencoders with density-based clustering. Applied Sciences, 8, 9, 1468.

[14] Khongorzul, D., Lee, S.M., Kim, Y.K., and Kim, M.-H., 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.

[15] Liou, C.; Cheng, W.; Liou, J.; Liou, D., 2014. Autoencoder for Words. Neurocomputing, 139, 84–96.

[16] Zhou, C, and Paffenroth, R.C., 2017. Anomaly detection with robust deep autoencoders. *In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 665-674.

[17] Lin, Y. and Wang, J., 2019. Probabilistic deep autoencoder for power system measurement outlier detection and reconstruction. IEEE Transactions on Smart Grid, 11, 2, 1796-1798.

[18] Spandonidis, C., Theodoropoulos, P., Giannopoulos, F., Galiatsatos, N. and Petsa, A., 2022. Evaluation of deep learning approaches for oil & gas pipeline leak detection using wireless sensor networks. Engineering Applications of Artificial Intelligence, 113, p.104890.

[19] Vapnik, V. N., 1995. The nature of statistical learning theory. Springer, New York.

[20] Pang-Ning Tan K V Steinbach Micheal. 2007. Introduction to data mining. India: Pearson Education.

[21] Anyanwu MN, Shiva SG, 2009. Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security.; 3, 3, 230–240.

[22] Chen, T.; Guestrin, C., 2016. Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining.* San Francisco, CA, USA, 785–794.

[23] Zachary D.W., Duck K.Y., Joseph C. von F., 2018. Instruction for Processing Mobile Methane Survey Data to Detect Natural Gas Leaks. Colorado State University. Available online: https://github.com/JVF-CSU/MobileMethaneSurveys/tree/master/Scripts/SampleRawData (accessed on) Oct 10, 2018

[24] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.