IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 October 2023, accepted 3 December 2023, date of publication 7 December 2023,
date of current version 18 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340240

RESEARCH ARTICLE

NDAMA: A Novel Deep Autoencoder and Multivariate Analysis Approach for IoT-Based Methane Gas Leakage Detection

KHONGORZUL DASHDONDOV[1], MI-HYE KIM[2], AND KYURI JO[2]
[1]Department of Computer Engineering, College of IT Convergence, Gachon University, Seongnam 13120, Republic of Korea
[2]Department of Computer Engineering, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju, Chungcheongbuk-do 28644, Republic of Korea

Corresponding author: Kyuri Jo (kyurijo@chungbuk.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government, Ministry of Science and ICT (MSIT) under Grant NRF-RS-2023-00252745 and Grant NRF-RS-2023-00217022.

ABSTRACT Natural gas is widely used for domestic and industrial purposes, and whether it is being leaked into the air cannot be directly known. The current problem is that gas leakage is not only economically harmful but also detrimental to health. Therefore, much research has been done on gas damage and leakage risks, but research on predicting gas leakages is just beginning. 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 using accuracy, F1-score, mean square error (MSE), mean intersection over union (mIoU), and area under the ROC curve (AUC). The presented method in this study outperformed all compared methods. Moreover, the deep autoencoder and ordinal encoder-based XGBoost (DA-MA-XGBoost) showed the best performance by giving 99.51% accuracy, an F1-score of 99.53%, an MSE of 0.003, mIoU of 99.40 and an AUC of 99.62%.

INDEX TERMS Deep autoencoder, multivariate analysis, methane gas, risk detection, XGBoost.

I. INTRODUCTION

Early gas leakage prediction makes preventing future economic losses possible. In addition, natural gas leakage can exacerbate adverse health effects, such as hypertension, pulmonary disease, pneumonia, asthma, and other respiratory diseases. Therefore, gas leak detection is essential for gas-intensive countries. Thus far, we found that more re-search is necessary to predict gas leakage. Although there are studies on the harmful effects of gas leaks, more research is necessary for predicting gas leaks [1], [2]. Deep learning (DL) has recently made internet of things (IoT)-based multivariate time-series analysis possible because of its potent feature extraction and representation learning capabilities. Nevertheless, some existing time-series analysis studies have included unsupervised DL-based techniques [3]. To close this knowledge gap, we examine unsupervised learning-based risk detection and clustering for IoT time series within a unified framework [4], [5].

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 preprocessing, data labeling, and predictive analysis. The data preprocessing module removes outliers using the deep autoencoder (DAE) reconstruction error (RE) [6] and normalizes the data using OrdinalEncoder (OE) transformation techniques. The data labeling module selects only natural gas (NG) CH4 [7], [8] data from the

The associate editor coordinating the review of this manuscript and approving it for publication was Santosh Kumar.

preprocessed data, divides it into groups using the k-means clustering algorithm, and classifies the data according to that group. Afterward, the predictive analysis module builds a model that predicts gas loss using machine learning algorithms on the available data.

There are few related works that define natural gas leak emission levels. Methane levels ranged between approximately 1800 and 2600 parts-per-billion (ppb) throughout, which was consistent with the wind direction in [9]. Additionally, [10] identified the primary as discriminated small leaks <6 L min−1 from medium leaks (6−40 L min−1) and a high bin (>40 L min−1) for the estimated leak level.

We used a data survey from the Los Gatos Research CH4 analyzer's high-sensitivity mobile and portable survey [11]. There was more than a sequence of differences in the sensitivity of a device used to measure CH4 levels. Therefore, this CH4 analyzer was susceptible to only a rare parts-per billion (ppb) withdrawal from the background, and LDCs frequently use handheld sensors with parts-per-million (ppm) level sensitivities [12]. In this study, CH4 (ppm) is the target feature, and OE methods are used for the real number to be a labeling feature for the data preprocessing part [13]. The measurement and estimated leak flow rate levels of CH4 are shown in Table 1.

TABLE 1. CH4 leak detection measurement and estimated rate range.

Methane level (ppb) [9]	Detection measurement range of ch4 (ppm) [10]	Estimated leak flow rate (g min) [11]
Low (<1800 ppb)	Low (<4.5 ppm)	Low (<1.6 g min^{-1})
Medium (1800~2600 ppb)	Medium (4.5 Ppm ~ $9x10^4$ ppm)	Medium (1.6~26 g min^{-1})
High (>2600 ppb)	High (> $9x10^4$ ppm)	High (>26 g min^{-1}).

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 preprocessing.

The main contribution of this paper is the following novelty:

- A novel deep learning-based method is proposed to predict gas leakages by removing outliers with DAE model.
- The proposed method is evaluated on a 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 can widely use for comparative research.
- The proposed method is compared with baseline models based on extreme gradient boost (XGB), K-nearest neighbors (KNN), decision tree (DT), random forest (RF), and naive Bayes (NB) algorithms and shows improved performance.

The remaining article is outlined as follows. Section II provides a detailed survey of related work. The proposed method is explained in Section III. Section IV presents the experimental dataset, the methods used for comparison, the evaluation metrics, and the comparative experiment results. Section V concludes this research. Appendix shows feature descriptions.

II. RELATED WORKS

Researchers have studied a pilot project of this mapping approach to explore the first step in understanding the effects of NG leaks. Some researchers have presented an automatic encoder-based anomaly detection method for wind turbine condition monitoring to implement preventive maintenance programs [14], [15]. The data used in our research are unlabeled, and the reconstruction-based architecture [16] of autoencoder (AE) and unsupervised approaches have recently made considerable strides in this endeavor. Similarly, some other researchers proposed an unsupervised learning-based leakage detection method that learns the characteristics of normal operating conditions by reconstructing input data and detects tube leakage by calculating its reconstruction error [17]. Additionally, this paper aims to achieve the best gas leak detection results by optimizing hyperparameter settings [18].

Recently, various studies have utilized deep learning (DL) and machine learning (ML) techniques to address the challenge of detecting gas leaks in industrial control systems. One such approach [19] introduces a lightweight architecture called the Long Short-Term Memory Variational Autoencoder (LW-LSTM-VAE) and its combination with a one-class support vector machine (SVM) [20] for anomaly detection for this purpose. Real-time detection of leaks in natural gas gathering pipelines is critical to ensuring the safe transportation of energy from production sources. Since leak samples are rare in real pipelines, modeling healthy data is a prerequisite for reliable leak detection, and one of the main solutions is multivariate time series (MTS) [23] feature analysis. According to the authors [24], reconstruction and prediction-based one-dimensional convolutional LSTM-AE are proposed to enhance feature learning for MTS data. Based on the learned features, a multimodel decision scheme with one one-class support vector machine (OCSVM) is developed to deal with leak detection under multiple operation conditions (MOC).

Moreover, Yang et al. [21] present a new design model to increase the accuracy of pipeline leak detection. They combined a one-dimensional convolutional neural network (1DCNN) for adaptive data feature extraction and an improved particle-particle optimization algorithm called variable amplitude PSO (VAPSO) to optimize support vector machine (SVM) parameters. Miao et al. [22] discussed the important issue of pipeline leakage due to corrosion, which adversely affects the safety and reliability of oil and natural gas transportation. The authors proposed a new method for the general identification of semi-controlled domains using

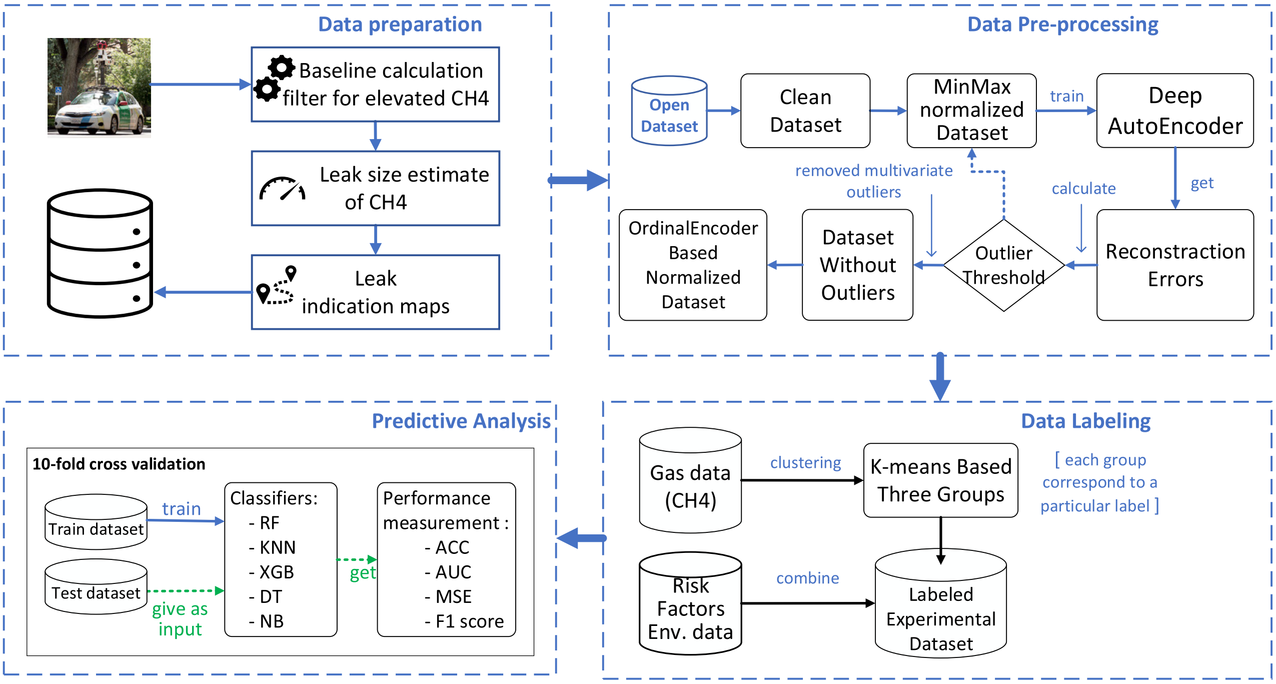


FIGURE 1. Architecture of the proposed approach.

laser optical sensing technology, and the experimental results identified the potential risk of missing with a recognition accuracy of more than 95%.

In recent years, oil and natural gas pipelines have been planned to be equipped with leak detection systems to monitor operations and detect leaks. Although leak detection methods used today cannot prevent leaks, they can play an important role in limiting the impact of leaks. Many different leak detection methods have been developed and tested [25]. In [26], these methods were studied, their strengths and limitations were analyzed, and the reliability of leak detection methods was analyzed. Zhu et al. [27] 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.

In [28], the advantages of the statistical shape analysis (SSA) method were presented in comparison to principal component analysis (PCA), discrete wavelet transforms (DWT), and polynomial curve fitting (PCF) algorithms for improving detection selectivity [29]. Additionally, Song et al. [30] presented a gas leak detection method for galvanized steel pipes based on acoustic emissions. Machine-based approaches to environmental engineering have been widely used to predict natural gas leaks. Our previous research used OE normalization and k-means clustering for data preprocessing [31]. However, we improved the performance of our previous study by using a DAE-based outlier removal process. Classification methods are trained on normally distributed data. In very rare cases, learning from data with outlier errors reduces the ability to predict other standard distributed data. Therefore, in this study, we 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.

Autoencoders are widely used for reducing data dimensions [32] by learning data [33] representations [34] and fault detection in acid gas removal units [35]. The authors of [36] used a clustering algorithm and reconstruction error from the deep autoencoder model to detect outliers in an unsupervised mode. Another autoencoder usage is to remove image denoising and time-series data. In the following, we introduce the proposed model with an open urban dataset that can be easily extended to the case of dataset batches.

III. METHODOLOGY

The proposed approach has four modules: data-create processing, data preprocessing, data labeling, and predictive analysis. The general architecture of the proposed method is presented in Fig. 1. The first module describes the research procedure by creating a CH4 dataset from an IoT-based mobile vehicle. As shown in the figure, the re-search procedure and the data processing steps have several stages, as mentioned in the related literature [1]. The second module uses DAE and OE transformation techniques. As a result of this module, normalized clean data are 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.

A. DATA PREPARATION

This work uses vehicles based on IoT technology to detect gas leakage rates using machine learning using data from a survey of outdoor street CH4 leakage [1]. These mobile surveys frequently combine data processing algorithms to provide a variety of data outputs, including maps showing the locations of leaks and estimates of their magnitude. Based on the time stamp from the vehicle position, this module first examined the atmospheric CH4 concentrations. The CH4 concentration [37] is filtered for each set of increased values following the baseline computation. Subsequently, the CH4 emission rate is determined. Afterward, the primary location of the natural gas leak was identified, and the highest CH4 concentration was measured for real-time stamp indication. Finally, the data collected for the visualization of the combined roadways were examined for leak indicators. Weller et al. [38] investigated the relationship between leakage indicators and the occurrence of actual leaks in their study, as well as possible biases in leak size estimations. Their analysis determined emissions by calculating the mean natural logarithm of the maximum excess CH4 during peaks with 20 or more observations. Additionally, they conducted Monte Carlo simulations to assess the variability of emissions based on the estimated emission levels derived from all observed leakage events. In this process, they randomly selected a specific number of detections (observed peaks) from a pool of all confirmed peak detections for each given number of detections. They then calculated the average percentage difference between the simulated and reference emission levels for each verified peak, yielding numerical values [39]. Environmental characteristics include longitude and latitude features [40].

B. DATA PREPROCESSING

We use a DAE to clean our data. The AE 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 as close to the original input as possible [41]. The structure of the AE consists of an encoder and a decoder. The encoder compresses input data by reducing the data dimension, while the decoder reconstructs the compressed data into output. Thus, the number of input neurons equals the number of output neurons in the AE [42]. The RE [43] of the AE is the difference between the input and its reconstructed output.

Fig. 2 shows the structure of the proposed AE model in this study. First, it projects input X to a lower dimension that works in the encoder; then, it reconstructs output X' from the low-dimensional projection in the decoder. Sequentially, the proposed AE has five hidden layers with 17, 14, 5, 1, 5, 14, and 17 nodes. Moreover, hidden layers in the encoder use the "ReLU" [44], [45] activation function and hidden layers in the decoder use the "tanh" [46] activation function. In summary, 17 features after min-max normalization are used to train the AE, where the activation functions for the encoder and decoder are the rectified linear unit (ReLU) and hyperbolic tangent (tanh), respectively. In other words, the AE learning process compresses the input into a lower-dimensional space called the latent space and decompresses the compressed data into output that closely matches the original data [47]. Then, it calculates the difference between the input and reconstructed output and changes the network weights to reduce this difference.

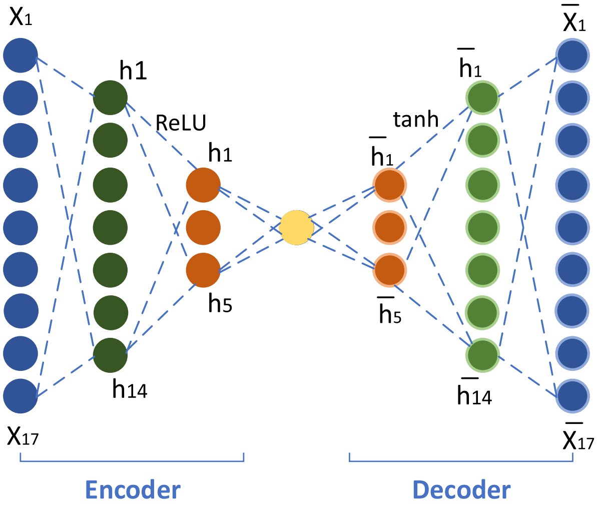


FIGURE 2. Structure of the proposed DAE method architecture of the proposed approach.

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

$$RE = \frac{1}{n}\sum\nolimits_{i=1}^{n} \left\| x_i - x_i' \right\|_2^2 \tag{1}$$

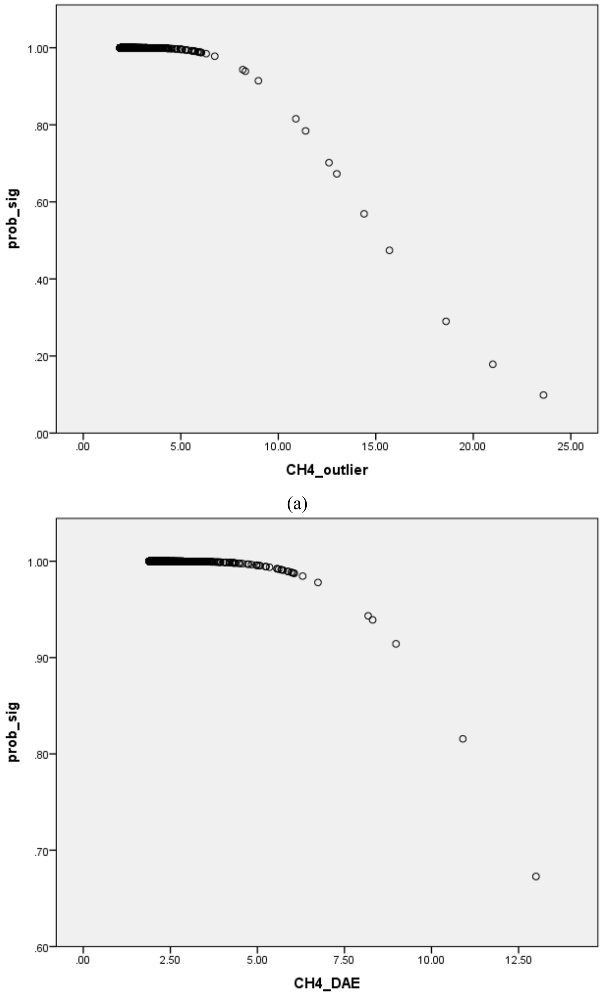
where n is the number of records, x is the original input, and x' is the reconstructed input. First, the RE of the training dataset was calculated using the DAE model. The mean and standard deviation of these REs were then used to estimate a threshold for splitting the training dataset, which can be described as follows.

$$Threshold = \frac{1}{k}\sum_{i=1}^{n} RE + \sqrt{\frac{1}{k}\sum_{i=1}^{n}\left[RE_i - \frac{1}{k}\sum_{i=1}^{n} RE_i\right]} \tag{2}$$

where k is the number of instances in the training dataset and RE_i is the reconstruction error of the i th training instance. Consequently, two different training datasets were prepared, and the RE-based threshold was estimated for further analysis. Subsequently, a threshold was used to select an appropriate hypertension prediction model from the DAE models that were trained on the two prepared datasets. For the DAE model, the learning rate was configured to minimize the mean squared error set to 0.001, and the Adamax optimizer was employed [48]. The batch size was set to 32 and the number of epochs was specified to 1000. The performance of the DAE model was compared under different threshold values for the CAR_SPEED feature in Table 2. This table helps to understand how the DAE model performs with varying threshold settings. It provides information about mean values, variability, and statistical significance of the model's performance. The MSE between predicted and actual values

TABLE 2. Comparative results of threshold values for DAE model.

	ORIGINAL CAR_SPEED FEATURE	DATASET WITH VARYING THRESHOLDS		
		High 0.0495	Medium 0.0493	Low 0.0409
N	69835	61255	61213	61148
Mean	4.174	3.852	3.862	3.877
Std. Dev.	5.503	4.5885	4.5929	4.6002
MSE	0.021	0.0186	0.0185	0.0184
t-statistic	200.40	207.75	207.68	208.39
p-value	0.0001			
95% CI	4.13-4.21	3.85-3.89	3.83-3.89	3.84-3.91

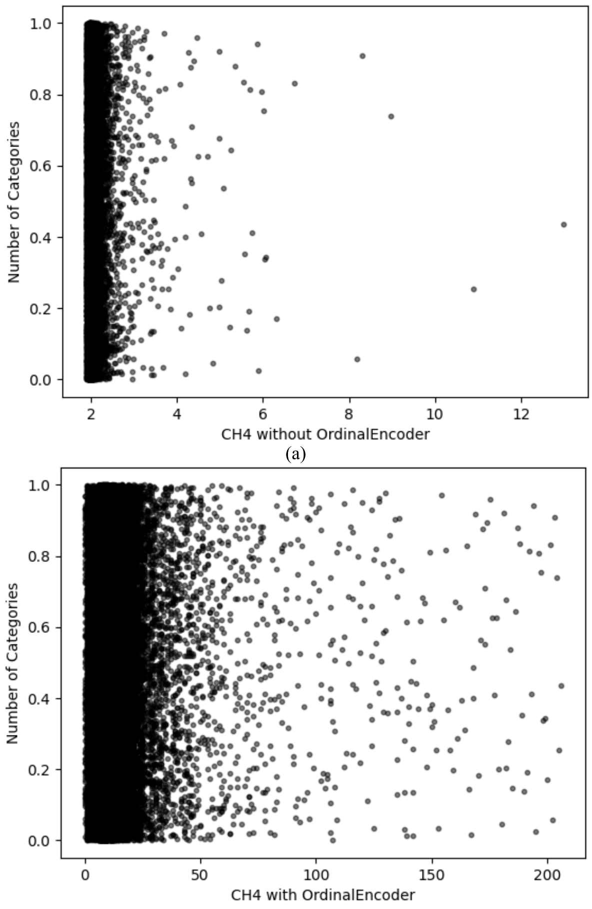


(a)

(b)

FIGURE 3. Plots of CH4 data (a) with and (b) without outliers by significant probability.

was measured to be 0.021 for the Original Dataset, 0.0186 for the high threshold, 0.0185 for the medium threshold, and 0.0184 for the low threshold. Lower MSE values indicate better performance of the model. The t-test and p-values suggest significant differences between the original dataset and each threshold category. Mean differences and confidence intervals were also calculated to provide more insight into the impact of thresholding on the CAR_SPEED feature. The mean value difference between the original dataset and each



(a)

(b)

FIGURE 4. Plots of CH4 data with and without OE transformation: (a) without OE and (b) with OE of CH4.

threshold category was calculated, which showed that the mean difference for the original dataset was 4.174. For the high threshold, it was 3.852; for the medium threshold, it was 3.862; and for the low threshold, it was 3.862.

Based on our analysis, we added the mean and standard deviation, which we named "low," to calculate a threshold value for comparison. We calculated the 75th percentile as "high" and the 50th percentile as "medium" using this threshold value and compared the results in Table 2. Our findings revealed that a lower threshold value of 0.0409 led to better outcomes for the selected data.

Fig. 3 shows data with and without outliers from the dataset by several values. Fig. 3 (a) shows the original dataset with outliers. Fig. 3 (b) shows a plotted dataset without outliers based on the DAE method. After that, the outlier threshold value is estimated by summing the average reconstruction error and standard deviation. Then, if the reconstruction error of the data exceeds the threshold value, these data will be removed from the dataset. In this figure, N means the number of cases.

This module's last step is to normalize outlier-removed data using the OE transformation technique. We encode

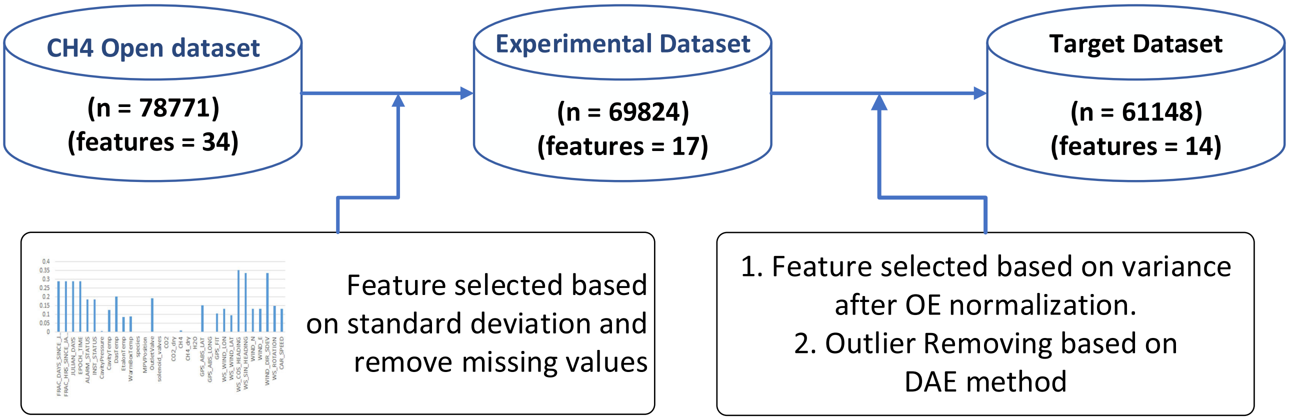


FIGURE 5. Target dataset based on the CH4 open dataset.

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 [31]. Fig. 4 shows plots of CH4 with and without OE by number of categories.

C. DATA LABELING

This module selects the CH4 feature, which is the value of methane from the preprocessed dataset for data labeling. As the first open dataset had no labels, we used the k-means algorithm, a multivariable clustering method developed by MacQueen in 1967 [49]. It divides the samples into k subgroups of n samples in the most comparable class. The Euclidean norm measures the distance between data points and each cluster's core point (centroid). Based on the result of the k-means algorithm, we assigned the class label [1], [2] as low, medium, or high. When we use OE transformation, there is no imbalance problem when we divide labels by the k-means clustering algorithms. Additionally, this formed label range is close to the theoretical range. This is another advantage of using OE transformation in our research. Finally, we combine the class labels with the outlier removed dataset except for the CH4 feature because the CH4 feature is used to determine class labels [50].

D. PREDICTIVE ANALYSIS

We train machine learning-based RF, KNN, XGB, DT, and NB algorithms [49] on our experimental dataset. We split the training dataset base 70% for training and 30% for testing.

NB: Naive Bayes is a probability-based classification algorithm. It computes the probability for each class label and selects the class label with the highest probability and calculates the probability by considering all features separately. It is called conditional independence.

KNN: The k-nearest neighbor algorithm is used for classification purposes. First, a user defines the value of the k parameter, which is the number of nearest samples used for prediction. Then, all distances between the test data and the training dataset are calculated and sorted in descending order. Finally, the top k instances from the ordered dataset are 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. 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. It consists of several decision tree classifiers trained in different subsamples of the whole dataset. For prediction, the majority-voted class label of these decision trees is chosen as the output.

XGB: XGBoost [51] 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 [52]. While a common decision-making tree is interested in how well the classification performed, CART enables the comparison of superiority among models that retain identical classification results.

IV. EXPERIMENTAL STUDY

A. DATASET

In this study, the open gas leak dataset is used [53]. 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 [54], [55]. Furthermore, 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]. Among 78,811 case data in total, "Date" and "Time" were used as indices to eliminate duplicate and missing values. Fig. 5 shows the procedure for creating the target dataset. 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 34. After removing outliers from 69,828 records, 61,148 records remained. Fig. 5 expresses the standard deviation of 32-dimensional data among the total 34-dimensional data excluding "Date" and "Time". Data for "Frac_Days_Days_Jan1", "Frac_Hrs_Since_Jan1", "Julian_Days", and "Epoch_Time" where the standard deviation was "0" were removed. The descriptions of features in Table 7 are shown in Appendix.

In the process, "Alarm_Status", "Inst_Status", "Cavity Pressure", "Gps_Abs_Lat", "Gps_Abs_Lon" and "Ws_Rotation" of which the deviation of feature data had dis-appeared additionally were eliminated additionally and, thus, the final 14 features, where the gas feature is "CH4", and environmental features are "Car_Speed", "Cavitytemp", "Ws_Wind_Lat", "Warmboxtemp", "Ws_Cos_Heading", "Ws_Sin_Heading", "Wind_E", "Wind_Dir_Sdev", "DasTemp", "Wind_N", "Ws_Wind_Lon", "EtalonTemp" and "OutletValve", have been used for data analysis in the next stage. Table 3 shows the number of records in each class for the testing and training process.

TABLE 3. 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

B. EVALUATION METRIC

The performance evaluation was completed using accuracy, AUC, F1-score, mean Intersection over Union (mIoU), and MSE [47]. We can find the precision and recall as follows where TP is true positive, FP is false positive, and FN is false negative:

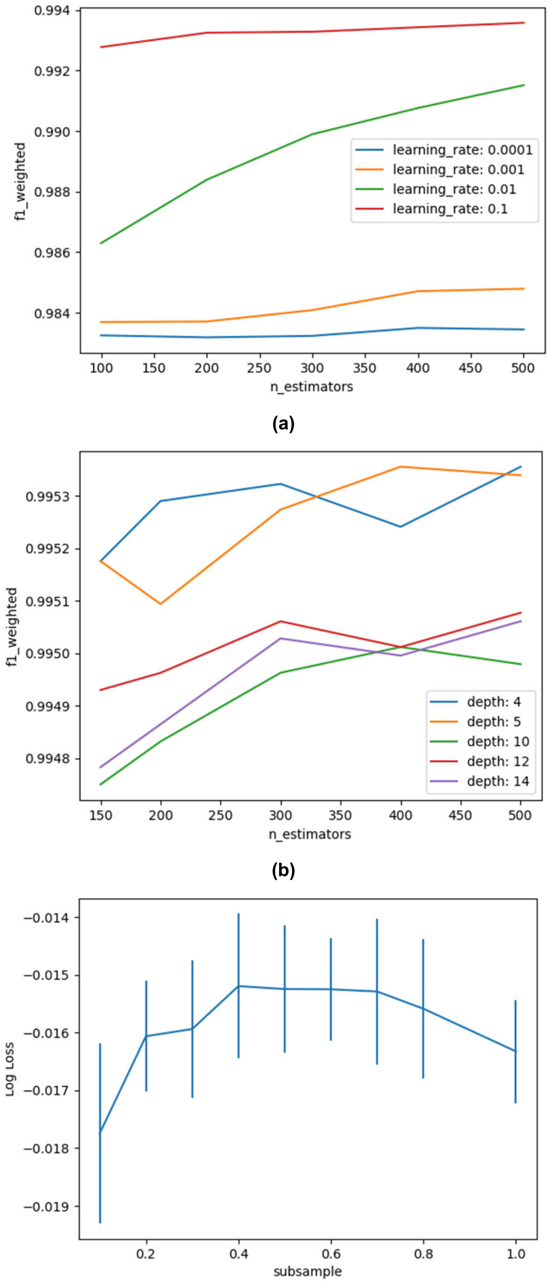
$$Precision = \frac{TP}{(TP+FP)}, Recall = \frac{TP}{(TP+FN)} \tag{3}$$

Precision and recall are important metrics for evaluating classification models. Precision measures the accuracy of positive predictions, while recall measures the model's ability to identify all positive instances. These metrics are useful for imbalanced datasets or when the cost of false positives and negatives varies.

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

$$F1 = \frac{2 \cdot Precision \cdot Recall}{(Precision + Recall)} \tag{4}$$

We studied the multiclass case, and the average of the F1-score of each class label with weighting depends on the average parameter, as shown in (4).



(a)

(b)

(c)

FIGURE 6. Comparison charts of the datasets. (a) Learning rate and max_depth; (b) depth and n_estimators; (c) accuracy score and max_depth Target dataset based on the CH4 open dataset.

Accuracy is a measure of the degree of closeness of the calculated value to its actual value. Accuracy is the sum of the true positive (TP) fraction and true negative (TN) fraction among all the test data, as shown in (5).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

The AUC (Area Under the Curve) is a crucial metric for multiple classification models as shown in (6). It's calculated by finding the area under the ROC (Receiver Operating Characteristic) curve with false positive rate (FPR) and true

TABLE 4. Evaluation results of the compared algorithms on the experimental dataset (%).

Methods	Classifier Algorithms	Accuracy	AUC	MSE	F1-score	mIoU score
Outlier removing method	DAE-OE-RF	99.01	99.41	0.005	99.23	99.07
	DAE-OE-KNN	98.87	99.16	0.007	98.88	98.87
	DAE-OE-XGB	99.19	99.53	0.004	99.38	99.18
	DAE-OE-DT	98.16	9912	0.007	98.82	98.47
	DAE-OE-NB	85.85	94.18	0.05	92.44	88.79
Normalization method	OE-RF	98.51	99.11	0.007	98.64	99.08
	OE-KNN	98.62	98.97	0.009	98.64	98.97
	OE-XGB	98.74	99.28	0.006	98.9	99.31
	OE-DT	97.24	98.59	0.01	98.2	98.39
	OE-NB	78.74	89.79	0.08	86.19	89.52
Baseline models	RF	87.57	89.03	0.082	87.20	87.57
	KNN	64.36	68.14	0.234	64.13	64.35
	XGB	87.57	89.03	0.082	87.20	87.57
	DT	87.57	89.03	0.083	87.20	87.57
	NB	81.12	91.23	0.075	89.29	86.96
Proposed method **k=10-fold**	**NDAMA-RF**	**99.29**	**99.45**	**0.005**	**99.36**	**99.17**
	NDAMA-KNN	98.96	99.25	0.007	99.00	98.99
	NDAMA-XGB	**99.51**	**99.62**	**0.003**	**99.53**	**99.40**
	NDAMA-DT	98.65	99.16	0.008	98.86	98.53
	NDAMA-NB	94.70	94.03	0.051	93.27	88.52

positive rate (TPR). A higher AUC value indicates better model performance, making distinguishing between positive and negative instances easier. It's beneficial for imbalanced data sets or when the cost of false positives and negatives varies.

$$AUC = \sum_{i=1}^{n} \frac{(FPR_i + FPR_{i+1}) \cdot (TPR_{i+1} - TPR_i)}{2} \quad (6)$$

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

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{i=0}^{n-1} [X(i,j) - Y(i,j)]^2 \quad (7)$$

with m and n being the number of observations, where m is the number of data points and n is predicting NG. X and Y are the actual and predicted values for the i, jth data point, respectively.

In addition, our models were evaluated using Jaccard loss, also known as the Intersection over Union (IoU) metric, a standard measure for assessing segmentation performance [56], [57]. As part of our experimental design, we computed the mIoU, which is a widely accepted metric for evaluating semantic segmentation models. The mIoU is determined by calculating the ratio of true positive pixels to the sum of true positives, false negatives, and false positives across all segmented pixels. This calculation is expressed by (8), underscoring the significance of accurately delineating target classes to achieve a high mIoU score as follows [58]:

$$mIoU = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{TP}{\sum_{j=1}^{N_c} FP + \sum_{j=1}^{N_c} FN - TP} \quad (8)$$

where N_c is the number of classes.

C. HYPERPARAMETER RESULTS

For better results, we tuned some XGBoost hyperparameters using the grid search infrastructure in scikit-learn [59] on the target dataset. In Fig. 6 (a), we show a plot of each learning rate as a series showing f1-weighted performance as the number of trees varied. In this figure, we show that the best result observed was a learning rate [60] of 0.1 with 500 trees. We can see that the expected general trend holds, where performance improves as the number of trees increases [5]. Next, in Fig 6 (b), we show a relationship between the number of trees in the model and the depth of each tree. We created a grid of 9 different n estimators' values (100 to 500) and 6 different max depth values (2, 4, 6, 8, 10, 12), and each combination was evaluated using 10-fold cross-validation. A total of 9 × 6 × 10 or 540 models were trained and evaluated. We can see that the best result was achieved with n estimators of 400 and a maximum depth of 5 in a f1-weighted score, but there was no significant difference from a maximum depth of 10.

Finally, we used the grid search capability built into scikit-learn to evaluate the effect of different subsample values from 0.1 to 1.0 on the target dataset. In Fig 6 (c), we show these mean and standard deviation log loss values to obtain a better understanding of how performance varies with the subsample value. We can see that 40% achieved the best mean performance, but we can also see that as the ratio increased, the variance in performance was almost the same.

D. PERFORMNCE EVALUATION

Data preprocessing and predictive analysis modules were implemented in Python using the sklearn library [53].

The data First, the performance of the baseline models is measured to compare them with our proposed method. We directly trained baseline models on the raw dataset using the machine learning algorithms shown in Fig. 1. Additionally, OE-based baseline models are trained on the dataset without removing outliers. All of these are machine learning algorithms that consider the same datasets and features/performance metrics. Therefore, DAE-OE-XGB is compared with other machine learning algorithms as well as with the proposed algorithm. By comparing the proposed algorithm with other machine learning algorithms, our model performed well when considering various features, such as accuracy, AUC, MSE, F1-score, and mIoU score. Table 4 shows the baseline models' and proposed methods' compared performances, where the highest values of evaluation scores are marked in bold. As a result, OE-based data normalization can improve the performance of models that were trained on raw datasets. Moreover, the combination of DAE-based outlier removal and OE-based data normalization in the proposed methods outperformed all compared baselines.

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 of all the compared models, with an accuracy rate of 99.193%, an F1-score of 99.38%, an MSE of 0.004, mIoU of 99.18% and an ROC of 99.53%. The DAE-OE-RF model achieved the second-best accuracy rate of 99.013%, F1-score of 99.23%, MSE of 0.005, mIoU of 99.07% and AUC of 99.41%. The DAE-OE-NB model exhibited lower results than the other proposed predictive models for the evaluation metrics. The NDAMA-XGBoost algorithm yielded the best results of all the models compared, with an accuracy rate of 99.51%, F1 score of 99.501, MSE of 0.003, mIoU of 99.40% and ROC of 99.618%.

TABLE 5. The Statistical significant of the overall mean accuracy, p-value, and log loss evaluations for gas leakage prediction NDAMA algorithms.

Classifier	Accuracy	p-value	Log loss	95% CI
NDAMA_XGB	99.51	4.21E-29	22.79	99.39-99.63
NDAMA_KNN	98.96	4.39E-29	39.30	98.87-99.00
NDAMA_DT	98.65	3.74E-29	44.99	98.57-98.75
NDAMA_RF	99.29	5.57E-29	23.58	99.20-99.37
NDAMA_NB	94.70	6.36E-29	240.69	94.58-94.83

TABLE 5 shows the results of ROC curve analysis for NDAMA methods. For all of the compared methods, accuracy (p-value<0.000001) was statistically significant. A low p-value (typically below a significance threshold like 0.05) may indicate that a model's performance is statistically significant [58]. Although the NDAMA improved the performance of single algorithms, the NDAMA_XGB outperformed the other NDAMA methods. In the case of XGB, accuracy was 87.57, and it has been improved to 99.51 (95% CI, 99.39-99.63) by using DAE based NDAMA_XGB, shown in TABLE 5. Finally, "Log loss" is a metric used to assess the accuracy of probabilistic predictions made by a model. It measures how well the predicted probabilities align with the actual outcomes. A lower log loss indicates better performance. For example, the NDAMA_XGB model has a log loss of 22.79. Generally, higher accuracy, lower variance, and lower log loss are desirable characteristics for a model [61].

In Fig. 7, we present box and whisker plots illustrating the mean accuracy scores obtained through k=10 cross-validation. Our findings revealed that XGB exhibits consistently high accuracy with minimal variance, which is a promising result. Notably, the unexpectedly poor performance of NB stands out.

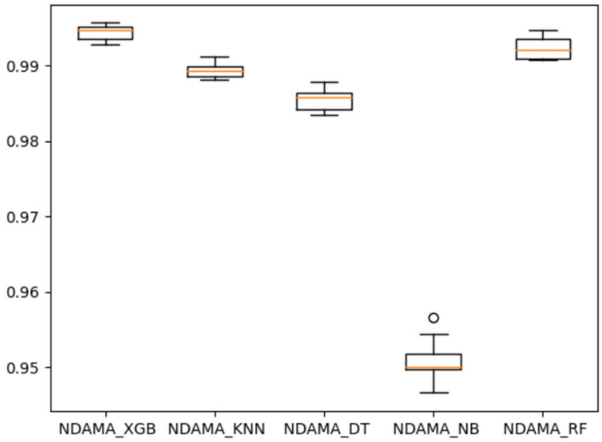


FIGURE 7. Comparison a box and whisker plot for mean accuracy of the proposed NDAMA methods.

The NDAMA-based method exhibited lower results than the other proposed predictive models for evaluation metrics. We provided multiclass ROC curves for each model compared with the average ROC curves in the target dataset in Fig. 8.

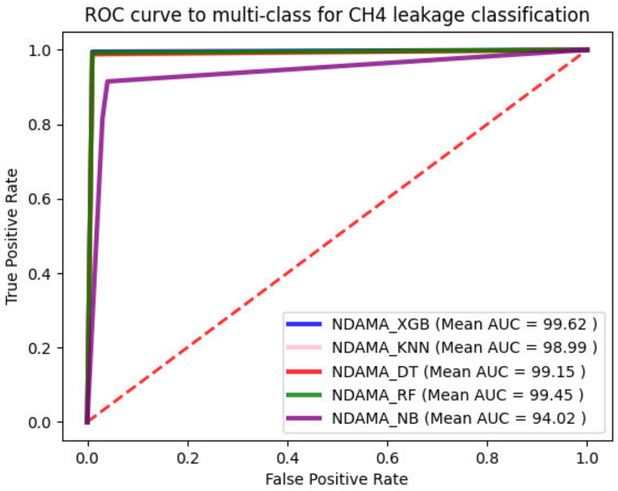


FIGURE 8. Receiver operating characteristic curves of the algorithms compared to the DAE-OE method.

The ROC curves for each comparative method on the target dataset with 10-fold cross-validation are clearly displayed in Fig 9. When the training set is divided into several subsets, it is feasible to compute the mean area under the curve and view the variance in the curve for 10-fold cross-validation. As a result, Fig. 9 shows the ROC curve for each cross-validation process with the mean. As noted above, we propose to find better model performance to predict XGBoost and RF for this dataset. Next, we compared our proposed methods to show the effects of different modules by the XGBoost algorithm, as shown in Fig 9.

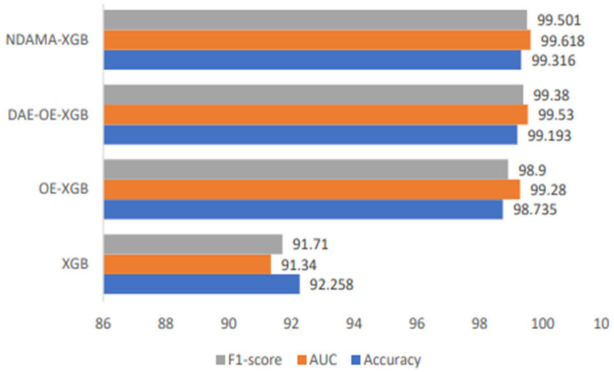


FIGURE 9. Comparison of the proposed modules and other guidelines for the XGBoost algorithm.

Compared with other state-of-the-art AE-based methods for gas leak detection, our proposed NDAMA method shows the best performance in all cases. Table 6 shows the results compared with other state-of-the-art AE-based methods for gas leak detection.

TABLE 6. The comparisons of classification applications of ML models using other methods for gas leakage.

CLASSIFIER ALGORITHMS	ACCURACY (%)	F1-SCORE (%)
LW-LSTM-VAE-S (STD) [19]	-	80.25
LW-LSTM-VAE-M (STD) [19]	-	84.46
BAE-OCSVM [20]	91.0	93.0
CAE-OCSVM [20]	93.0	94.0
LSTM-AE-OCSVM [20]	98.0	97.0
AE_DCNN-VAPSO-SVM [21]	96.61	-
IACGAN model [22]	95.0	-
PCA_KD [23]	-	81.80
ICA_KD [23]	-	98.50
RP-1dConvLSTM-AE_OCSVM [24]	96.40	96.58
NDAMA-RF	**99.29**	**99.36**
NDAMA-XGB	**99.51**	**99.53**

V. CONCLUSION

This study proposed a method consisting of three modules to predict gas leakage. Preparing efficient training data through data preprocessing and data labeling modules has dramatically improved the productive performance of machine learning algorithms. Using this method to create gas leakage data levels for air assessments in Korea is also possible. In other words, we used a DAE 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 easily distinguished by the AE model. The data were normalized using OE transformations and 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 accuracy scores achieved after $k = 10$ cross-validation, the NDAMA_XGB model is accurate, which means it predicts the target variable accurately around 99.51% (95% CI, 99.39-99.63) of the time. The proposed NDAMA methods significantly improved the accuracies of the DAE-OE and other baseline methods. Finally, the log loss of the NDAMA_XGB model is 22.79, so it can be concluded that the model has high accuracy, low variability, and low log loss, so it is an effective model. As a consequence, the proposed framework emerged as the best predictive model, capable of significantly outperforming existing state-of-the-art baseline models.

TABLE 7. The environmental feature description of the target dataset.

Class	Total
CavityTemp	Temperature in the cavity of the instrument
DasTemp	Room temperature
EtalonTemp	Temperature of Etalon plate
WarmBoxTemp	Temperature of the instrument warm box
OutletValve	outlet valve of the cavity
Gps_Abs_Lat	Latitude measured by GPS
Gps_Abs_Long	Longitude measured by GPS
Ws_Wind_Lon	wind direction longitude
Ws_Wind_Lat	wind direction latitude
Ws_Cos_Heading	wind speed conversion of cosine
Ws_Sin_Heading	wind speed conversion of sin
Wind_N	wind direction by north (0.00 C)
Wind_E	wind direction by east (90.00 C)
Wind_Dir_Sdev	standard deviation of horizontal wind direction
Ws_Rotation	wind rotation
Car_Speed	car speed installed sensor
CH4	CH4 methane ratio (ppm)

APPENDIX

FEATURE DESCRIPTIONS

We conducted a general description of the environmental features used in the experimental study for the target dataset, as shown in Table 7.

REFERENCES

[1] Z. D. Weller, D. K. Yang, and J. C. von Fischer, "An open source algorithm to detect natural gas leaks from mobile methane survey data," *PLoS One*, vol. 14, no. 2, Feb. 2019, Art. no. e0212287.

[2] J. C. von Fischer, D. Cooley, S. Chamberlain, A. Gaylord, C. J. Griebenow, S. P. Hamburg, J. Salo, R. Schumacher, D. Theobald, and J. Ham, "Rapid, vehicle-based identification of location and magnitude of urban natural gas pipeline leaks," *Environ. Sci. Technol.*, vol. 51, no. 7, pp. 4091–4099, Apr. 2017.

[3] Y. Liu, Y. Zhou, K. Yang, and X. Wang, ''Unsupervised deep learning for IoT time series,'' *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14285–14306, Feb. 2023.

[4] K. Cao, H. Ding, B. Wang, L. Lv, J. Tian, Q. Wei, and F. Gong, ''Enhancing physical-layer security for IoT with nonorthogonal multiple access assisted semi-grant-free transmission,'' *IEEE Internet Things J.*, vol. 9, no. 24, pp. 24669–24681, Dec. 2022, doi: 10.1109/JIOT.2022.3193189.

[5] B. Cheng, M. Wang, S. Zhao, Z. Zhai, D. Zhu, and J. Chen, ''Situation-aware dynamic service coordination in an IoT environment,'' *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2082–2095, Aug. 2017, doi: 10.1109/TNET.2017.2705239.

[6] T. Arshad, J. Zhang, I. Ullah, Y. Y. Ghadi, O. Alfarraj, and A. Gafar, ''Multiscale feature-learning with a unified model for hyperspectral image classification,'' *Sensors*, vol. 23, no. 17, p. 7628, Sep. 2023.

[7] M. Shang and J. Luo, ''The Tapio decoupling principle and key strategies for changing factors of Chinese urban carbon footprint based on cloud computing,'' *Int. J. Environ. Res. Public Health*, vol. 18, no. 4, p. 2101, Feb. 2021, doi: 10.3390/ijerph18042101.

[8] L. Liu, S. Fu, and C. Han, ''Investigation on diesel spray flame evolution and its conceptual model for large nozzle and high-density of ambient gas,'' *Fuel*, vol. 339, May 2023, Art. no. 127357, doi: 10.1016/j.fuel.2022.127357.

[9] K. S. Kumar, Y. E. B. Vidhya, R. Selvaraj, S. Seshadri, S. M. S. Nagendra, and N. J. Vasa, ''Dual absorption broadband photoacoustic technique to eliminate interference in gas mixtures,'' *IEEE Sensors J.*, vol. 23, no. 6, pp. 5703–5712, Mar. 2023.

[10] M. F. Hendrick, R. Ackley, B. Sanaie-Movahed, X. Tang, and N. G. Phillips, ''Fugitive methane emissions from leak-prone natural gas distribution infrastructure in urban environments,'' *Environ. Pollut.*, vol. 213, pp. 710–716, Jun. 2016.

[11] C. B. Alden, S. C. Coburn, R. J. Wright, E. Baumann, K. Cossel, E. Perez, E. Hoenig, K. Prasad, I. Coddington, and G. B. Rieker, ''Single-blind quantification of natural gas leaks from 1 km distance using frequency combs,'' *Environ. Sci. Technol.*, vol. 53, no. 5, pp. 2908–2917, Mar. 2019.

[12] Y. Wang, Q. Yuan, T. Li, Y. Yang, S. Zhou, and L. Zhang, ''Seamless mapping of long-term (2010–2020) daily global XCO_2 and XCH_4 from the greenhouse gases observing satellite (GOSAT), Orbiting Carbon Observatory 2 (OCO-2), and CAMS global greenhouse gas reanalysis (CAMS-EGG4) with a spatiotemporally self-supervised fusion method,'' *Earth Syst. Sci. Data*, vol. 15, no. 8, pp. 3597–3622, Aug. 2023.

[13] J. Wilkinson, C. Bors, F. Burgis, A. Lorke, and P. Bodmer, ''Correction: Measuring CO_2 and CH_4 with a portable gas analyzer: Closed-loop operation, optimization and assessment,'' *PLoS One*, vol. 14, no. 3, Mar. 2019, Art. no. e0206080.

[14] J. E. U. Cabus, Y. Cui, and L. B. Tjernberg, ''An anomaly detection approach based on autoencoders for condition monitoring of wind turbines,'' in *Proc. 17th Int. Conf. Probabilistic Methods Appl. Power Syst.*, Manchester, U.K., 2022, pp. 1–6.

[15] H. Jiang, Z. Xiao, Z. Li, J. Xu, F. Zeng, and D. Wang, ''An energy-efficient framework for Internet of Things underlaying heterogeneous small cell networks,'' *IEEE Trans. Mobile Comput.*, vol. 21, no. 1, pp. 31–43, Jan. 2022, doi: 10.1109/TMC.2020.3005908.

[16] J. Wang, S. Shao, Y. Bai, J. Deng, and Y. Lin, ''Multiscale wavelet graph autoencoder for multivariate time-series anomaly detection,'' *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[17] H. Kim, J. U. Ko, K. Na, H. Lee, H.-S. Kim, J.-D. Son, H. Yoon, and B. D. Youn, ''Opt-TCAE: Optimal temporal convolutional auto-encoder for boiler tube leakage detection in a thermal power plant using multisensor data,'' *Expert Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119377.

[18] H. Gao, B. Qiu, R. J. D. Barroso, W. Hussain, Y. Xu, and X. Wang, ''TSMAE: A novel anomaly detection approach for Internet of Things time series data using memory-augmented autoencoder,'' *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 5, pp. 2978–2990, Mar. 2022.

[19] D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, ''Lightweight long short-term memory variational auto-encoder for multivariate time series anomaly detection in industrial control systems,'' *Sensors*, vol. 22, no. 8, p. 2886, Apr. 2022.

[20] Z. Zuo, L. Ma, S. Liang, J. Liang, H. Zhang, and T. Liu, ''A semi-supervised leakage detection method driven by multivariate time series for natural gas gathering pipeline,'' *Process Saf. Environ. Protection*, vol. 164, pp. 468–478, Aug. 2022.

[21] D. Yang, N. Hou, J. Lu, and D. Ji, ''Novel leakage detection by ensemble 1DCNN-VAPSO-SVM in oil and gas pipeline systems,'' *Appl. Soft Comput.*, vol. 115, Jan. 2022, Art. no. 108212.

[22] X. Miao, H. Zhao, B. Gao, and F. Song, ''Corrosion leakage risk diagnosis of oil and gas pipelines based on semi-supervised domain generalization model,'' *Rel. Eng. Syst. Saf.*, vol. 238, Oct. 2023, Art. no. 109486.

[23] F. Harrou, K. R. Kini, M. Madakyaru, and Y. Sun, ''Uncovering sensor faults in wind turbines: An improved multivariate statistical approach for condition monitoring using SCADA data,'' *Sustain. Energy, Grids Netw.*, vol. 35, Sep. 2023, Art. no. 101126.

[24] Z. Zuo, H. Zhang, L. Ma, T. Liu, and S. Liang, ''Leak detection for natural gas gathering pipelines under multiple operating conditions using RP-1dConvLSTM-AE and multimodel decision,'' *IEEE Trans. Ind. Electron.*, early access, Jul. 2023, doi: 10.1109/TIE.2023.3294645.

[25] S. Lu, J. Guo, S. Liu, B. Yang, M. Liu, L. Yin, and W. Zheng, ''An improved algorithm of drift compensation for olfactory sensors,'' *Appl. Sci.*, vol. 12, no. 19, p. 9529, Sep. 2022, doi: 10.3390/app12199529.

[26] N. V. S. Korlapati, F. Khan, Q. Noor, S. Mirza, and S. Vaddiraju, ''Review and analysis of pipeline leak detection methods,'' *J. Pipeline Sci. Eng.*, vol. 2, no. 4, Dec. 2022, Art. no. 100074.

[27] S.-B. Zhu, Z.-L. Li, S.-M. Zhang, Ying-Yu, and H.-F. Zhang, ''Deep belief network-based internal valve leakage rate prediction approach,'' *Measurement*, vol. 133, pp. 182–192, Feb. 2019.

[28] V. V. Krivetskiy, M. D. Andreev, A. O. Efitorov, and A. M. Gaskov, ''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,'' *Sens. Actuators B, Chem.*, vol. 329, Feb. 2021, Art. no. 129187.

[29] S. Lu, Y. Ban, X. Zhang, B. Yang, S. Liu, L. Yin, and W. Zheng, ''Adaptive control of time delay teleoperation system with uncertain dynamics,'' *Frontiers Neurorobotics*, vol. 16, Jul. 2022, Art. no. e928863, doi: 10.3389/fnbot.2022.928863.

[30] Y. Song and S. Li, ''Gas leak detection in galvanised steel pipe with internal flow noise using convolutional neural network,'' *Process Saf. Environ. Protection*, vol. 146, pp. 736–744, Feb. 2021.

[31] D. Khongorzul, M-H. Kim, and S. M. Lee, ''Ordinalencoder based DNN for natural gas leak prediction,'' *J. Korea Converg. Soc.*, vol. 10, no. 10, pp. 7–13, 2019.

[32] O. Lyudchik, *Outlier Detection Using Autoencoders*, document CERN-STUDENTS-Note-2016-079, 2016.

[33] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga, ''Outlier detection with autoencoder ensembles,'' in *Proc. SIAM*, Houston, TX, USA, 2017, pp. 90–98.

[34] T. Kieu, B. Yang, and C. S. Jensen, ''Outlier detection for multidimensional time series using deep neural networks,'' in *Proc. 19th IEEE Int. Conf. Mobile Data Manage. (MDM)*, Aalborg, Denmark, Jun. 2018, pp. 125–134.

[35] T. K. Kathlyn, H. Zabiri, C. Aldrich, X. Liu, and A. A. A. M. Amiruddin, ''Fault detection and identification in an acid gas removal unit using deep autoencoders,'' *ACS Omega*, vol. 8, no. 22, pp. 19273–19286, May 2023.

[36] T. Amarbayasgalan, K. H. Park, J. Y. Lee, and K. H. Ryu, ''Reconstruction error based deep neural networks for coronary heart disease risk prediction,'' *PLoS One*, vol. 14, no. 12, Dec. 2019, Art. no. e0225991.

[37] K. Dashdondov, M. H. Kim, and K. Jo, ''Deep autoencoder-based framework for the classification of natural gas leaks grade using multivariate outlier detection,'' in *Proc. ACM KDD Conf. URBCOMP*, vol. 22, 2022, pp. 1–6.

[38] E. von Luetschwager, J. C. Fischer, and Z. D. Weller, ''Characterizing detection probabilities of advanced mobile leak surveys: Implications for sampling effort and leak size estimation in natural gas distribution systems,'' *Elementa, Sci. Anthropocene*, vol. 9, no. 1, p. 00143, 2021.

[39] X. Wang, H. Feng, T. Chen, S. Zhao, J. Zhang, and X. Zhang, ''Gas sensor technologies and mathematical modelling for quality sensing in fruit and vegetable cold chains: A review,'' *Trends Food Sci. Technol.*, vol. 110, pp. 483–492, Apr. 2021, doi: 10.1016/j.tifs.2021.01.073.

[40] T. Takano and M. Ueyama, ''Spatial variations in daytime methane and carbon dioxide emissions in two urban landscapes, Sakai, Japan,'' *Urban Climate*, vol. 36, Mar. 2021, Art. no. 100798.

[41] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou, ''Autoencoder for words,'' *Neurocomputing*, vol. 139, pp. 84–96, Sep. 2014.

[42] H. Min, Y. Fang, X. Wu, X. Lei, S. Chen, R. Teixeira, B. Zhu, X. Zhao, and Z. Xu, ''A fault diagnosis framework for autonomous vehicles with sensor self-diagnosis,'' *Expert Syst. Appl.*, vol. 224, Aug. 2023, Art. no. 120002, doi: 10.1016/j.eswa.2023.120002.

[43] Z. Zhang, X. Lai, S. Du, W. Yu, and M. Wu, "Early warning of loss and kick for drilling process based on sparse autoencoder with multivariate time series," *IEEE Trans. Ind. Informat.*, vol. 19, no. 11, pp. 11019–11029, Feb. 2023.

[44] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 665–674.

[45] X. Zhou and L. Zhang, "SA-FPN: An effective feature pyramid network for crowded human detection," *Int. J. Speech Technol.*, vol. 52, no. 11, pp. 12556–12568, Sep. 2022, doi: 10.1007/s10489-021-03121-8.

[46] Y. Lin and J. Wang, "Probabilistic deep autoencoder for power system measurement outlier detection and reconstruction," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1796–1798, Mar. 2020.

[47] C. Spandonidis, P. Theodoropoulos, F. Giannopoulos, N. Galiatsatos, and A. Petsa, "Evaluation of deep learning approaches for oil & gas pipeline leak detection using wireless sensor networks," *Eng. Appl. Artif. Intell.*, vol. 113, Aug. 2022, Art. no. 104890.

[48] G. K. Gupta, *Introduction to Data Mining With Case Studies*. New Delhi, India: PHI Learning Private Ltd., 2014.

[49] O. A. M. López, A. M. López, and J. Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham, Switzerland: Springer, 2022.

[50] M. Dix, A. Chouhan, S. Ganguly, S. Pradhan, D. Saraswat, S. Agrawal, and A. Prabhune, "Anomaly detection in the time-series data of industrial plants using neural network architectures," in *Proc. IEEE 7th Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Aug. 2021, pp. 222–228.

[51] J.-W. Yang and K. Dashdondov, "In-depth examination of machine learning models for the prediction of ground temperature at various depths," *Atmosphere*, vol. 14, no. 1, p. 68, Dec. 2022.

[52] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

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

[54] A. Nasri, A. Boujnah, A. Boubaker, and A. Kalboussi, "A smart gas sensor using machine learning algorithms: Sensor types based on IED configurations, fabrication techniques, algorithmic approaches, challenges, progress and limitations: A review," *IEEE Sensors J.*, vol. 23, no. 11, pp. 11336–11355, Apr. 2023.

[55] X. Liang, Z. Huang, S. Yang, and L. Qiu, "Device-free motion & trajectory detection via RFID," *ACM Trans. Embedded Comput. Syst.*, vol. 17, no. 4, pp. 1–27, Jul. 2018, doi: 10.1145/3230644.

[56] Z. Ding, W. Song, and S. Zhan, "A measurement system for the tightness of sealed vessels based on machine vision using deep learning algorithm," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[57] D. M. Vlachogiannis, S. Moura, and J. Macfarlane, "Intersense: An XGBoost model for traffic regulator identification at intersections through crowdsourced GPS data," *Transp. Res. C, Emerg. Technol.*, vol. 151, Jun. 2023, Art. no. 104112.

[58] K. Kalinaki, O. A. Malik, D. T. C. Lai, R. S. Sukri, and R. B. H. A. Wahab, "Spatial–temporal mapping of forest vegetation cover changes along highways in Brunei using deep learning techniques and Sentinel-2 images," *Ecological Informat.*, vol. 77, Nov. 2023, Art. no. 102193.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[60] W. Wu, C. Song, J. Zhao, and G. Wang, "Knowledge-enhanced distributed graph autoencoder for multiunit industrial plant-wide process monitoring," *IEEE Trans. Ind. Informat.*, early access, Jun. 2023.

[61] M. Krishnan, Y. Lim, S. Perumal, and G. Palanisamy, "Detection and defending the XSS attack using novel hybrid stacking ensemble learning-based DNN approach," *Digit. Commun. Netw.*, early access, Oct. 2022.

KHONGORZUL DASHDONDOV received the B.S. and M.S. degrees in mathematics from the School of Mathematics and Computer Science, National University of Mongolia, Mongolia, in 1998 and 2000, respectively, and the Ph.D. degree from the Mobile/Multimedia Communication Research Laboratory, Department of Radio and Communication Engineering, Chungbuk National University, Cheongju, South Korea, in 2013. Since 2017, she has been a Postdoctoral Research Fellow with the Ubiquitous Game Laboratory, Chungbuk National University. Since 2023, she has also been an Assistant Professor with the Department of Computer Engineering, Gachon University, Cheongju. Her research interests include queueing theory, artificial intelligence, deep learning, big data analysis, and healthcare analytics.

MI-HYE KIM received the B.S., M.S., and Ph.D. degrees in mathematics from Chungbuk National University, Cheongju, South Korea, in 1992, 1994, and 2001, respectively. She is currently a Professor with the Department of Computer Engineering, Chungbuk National University. Since 2004, she has been the Leader of the Ubiquitous Game Laboratory, Chungbuk National University. She is currently a Professor with Chungbuk National University. Her research interests include fuzzy measures and fuzzy integrals, digital therapeutics, gesture recognition, and medical serious games.

KYURI JO received the B.S. and Ph.D. degrees in computer science and engineering from Seoul National University, in 2013 and 2018, respectively. She was a Postdoctoral Research Fellow with the Bio and Health Informatics Laboratory, Seoul National University. Since September 2019, she has been an Assistant Professor with the Department of Computer Engineering, Chungbuk National University, Cheongju, South Korea. Her current research interests include bioinformatics, machine learning, and multi-omics analysis.