Journal of Big Data

RESEARCH Open Access

Unsupervised outlier detection for time-series data of indoor air quality using LSTM autoencoder with ensemble method

Junhyeok Park[1], Youngsuk Seo[2] and Jaehyuk Cho[3*]

*Correspondence:
chojh@jbnu.ac.kr

[1] Department of Electronic and Information Engineering, Soongsil University, Seoul, Korea
[2] Department of Mathematics, Soongsil University, Seoul, Korea
[3] Department of Software Engineering, Jeonbuk National University, Jeonju Si, Jeollabuk Do, Korea

Abstract

The proposed framework consists of three modules as an outlier detection method for indoor air quality data. We first use a long short-term memory autoencoder (LSTM-AE) based reconstruction error detector, which designs the LSTM layer in the shape of an autoencoder, to build a reconstruction error-based outlier detection model and extract latent features. The latent feature class-assisted vector machine detector constructs an additional outlier detection model using previously extracted latent features. Finally, the ensemble detector combines the two independent classifiers to define a new ensemble-based decision rule. Furthermore, because real-time anomaly detection proceeds with unsupervised learning, more stable and consistent external detection rules are defined than when using a single ensemble model. Laboratory tests with five random cases were performed for objective evaluation. Thus, we propose a framework that can be applied to various industrial environments by detecting and defining stable outlier decision rules.

Keywords: Long short-term memory autoencoder, Environmental sensor, Anomaly detection, Multivariate time series, Condition monitoring

Introduction

Recently, as social interest in the environment has increased, the importance of detecting indoor air pollution (IAP), which can be harmful to the human body, has been emphasized. In particular, IAP is treated as a higher risk than outdoor air pollution because the time spent indoors is longer and because it is an enclosed space. Therefore, it is easy for pollutants to have an intensive and harmful effect on the body. Continuous exposure to the contaminants and their long-term accumulation in the body may lead to serious risks, such as respiratory diseases, cardiovascular disease, lung cancer, and bronchial asthma [1, 2]. To address these problems, researchers have devised various approaches to identify the causes of pollutants and control indoor air quality [3]. The Overall mission of this filed is to understand how IAP impacts human health and well-being, and to develop strategies to reduce exposure to IAP. As an important aspect of these solutions, controlling major factors such as total volatile organic compounds (TVOC) and CO_2, which have harmful effects on the human body, and detecting

outliers is being emphasized [4]. Although various methods are being studied to define the range of outliers for each factor, it was far from the actual real world application due to the problems such as multiple sources of solution, cost and availability of monitoring equipment. So new and differentiated approach is needed to identify universal outliers that can be applied to information and communications technology-based environmental monitoring sensors. In this way, we propose a new methodology that solves and approaches the IAP situation from an outlier perspective.

Outlier detection is a technique used to find abnormal values or patterns in large datasets. The "abnormal" that we defined is a concept that includes point, collective, and contextual outliers [5]. A point outlier is a specific data point with a large difference from other normal data. A collective outlier refers to data that show a change outside the predictable range among continuous patterns of change in consideration of the overall context of the data. Contextual outliers are extreme with respect to the whole set called global outliers [6]. To detect outliers in an indoor environment, real-time detection that can immediately identify air pollution is very important. Real-time outlier detection identifies anomalies in the data as indicators of the difference between the predicted values of the model and the measured values of the sensors [7]. Currently, algorithms that perform situation-specific outlier detection in real time from various types of sensor data are being actively researched [5]. When detecting situational outliers in the indoor environment in real time, the precondition for training the model becomes very difficult because the range of judgments that can be defined as outliers may be different for each environment. Therefore, we propose an artificial intelligence-based outlier detection model that can be robustly applied in several indoor environments.

In generally, to train the outlier detection model, only normal data used as training data. However, in the real world, because data are directly collected through a sensor, noise could be included in the data. Therefore, under the assumption that most data are normal samples with a certain level of noise, studies on unsupervised detection are conducted.

In outlier detection in the field of networks, an anomaly is defined as an exceptional pattern that does not follow the expected normal pattern of network traffic [8]. In autonomous driving, Resnet classification is used to diagnose component failures, and cases different from dynamic characteristics are defined as fault diagnoses [9]. Moreover, in natural images, when evaluating outlier detection methods through multi-class classification, class labels are adjusted to the existing classification datasets [10].

In time-series data research, many studies have been conducted on bearing defect diagnosis problems or multivariate outlier detection [9, 11]. In a univariate time series, a point or subsequence that exceeds a certain threshold can be considered an anomaly [1]. By contrast, in multivariate time-series data, the relationship between the variables, the abnormal values of each characteristic, and the abnormalities of the subsequences must be considered [12].

An outlier was defined as an exceptional observation value that deviated from the normally collected cluster. Based on the normal data category, data outside the corresponding decision boundary were determined as outliers. However, the range and value of outlier clusters vary depending on the environment. Indoor air quality is highly dependent on the structure of a building, materials used in its construction, ventilation cycle,

and behavior of its occupants [13]. Therefore, it is necessary to verify the linearity of the sensor data as the indoor environment changes. We developed an anomaly detection framework that can analyze key environmental factors based on self-collected data from sensor devices and detect abnormal patterns that can be identified as outliers. Moreover, we built an anomaly detection framework that can monitor the conditions of the environment and provide accurate alarms for dangerous situations.

We propose a long short-term memory autoencoder (LSTM-AE) as an algorithm to perform outlier detection on multivariate time-series data. The LSTM-AE is an algorithm that extracts low-dimensional compression characteristics that best represent data by reflecting the time-series characteristics in the data. Based on the latent characteristics extracted from the encoder, it reconstructs a restored value similar to the original value possible in the decoder [14]. In LSTM-AE, the threshold for discriminating outliers is determined by calculating the difference between the restored output value and the original input value, and the final decision rule is defined [15]. Such a detection method is called a reconstruction error-based outlier detection method (RE-DM). RE-DM has the advantage of comprehensively evaluating multivariate environmental data to derive reasonable decision-making rules even in an unsupervised environment.

However, when RE-DM is used independently, it is not only vulnerable to noise included in the data but also unstable near the threshold and shows somewhat inconsistent classification results. Therefore, the false positive rate (FPR), which indicates the positive prediction misclassification of the model, increases, causing several false alarms [16]. To solve this problem in the LSTM-AE, we constructed a sub-algorithm using the latent feature space extracted from the encoder. The sub-algorithm makes the existing one-dimensional decision rule more complex to output more careful and generalized classification results. In addition, by combining two independent models that operate with different mechanisms, outliers can be distinguished from diverse viewpoints. To verify that our model can be robustly applied in various environments, we perform laboratory (LAB) tests based on data collected from environmental monitoring sensors.

The LSTM-AE-based outlier detection framework presented in this study can detect various unpredictable abnormal situations by training deep neural networks (DNNs) in a multi-feature data pool that changes frequently depending on the indoor environment and timestamp for each domain. The main objective of this study is to perform low-cost and quick-time data anomaly detection with real-world data directly collected in a situation where the utilization of computing resources in various industries is limited.

Research methodology

As presented above, our proposed outlier detection method based on the ensemble of LSTM-AE and sub-algorithm builds a robust model considering the various uncertainty of the real-world. Through this methodology, we would like to propose a method for stably performing outlier detection in an unsupervised learning environment that has not yet been conquered.

Figure 1 shows the proposed anomaly detection framework, which can be divided into the following four major steps: (i) data cleansing stage, (ii) modeling stage, (iii) decision rule definition stage, and (iv) situational ab test stage.

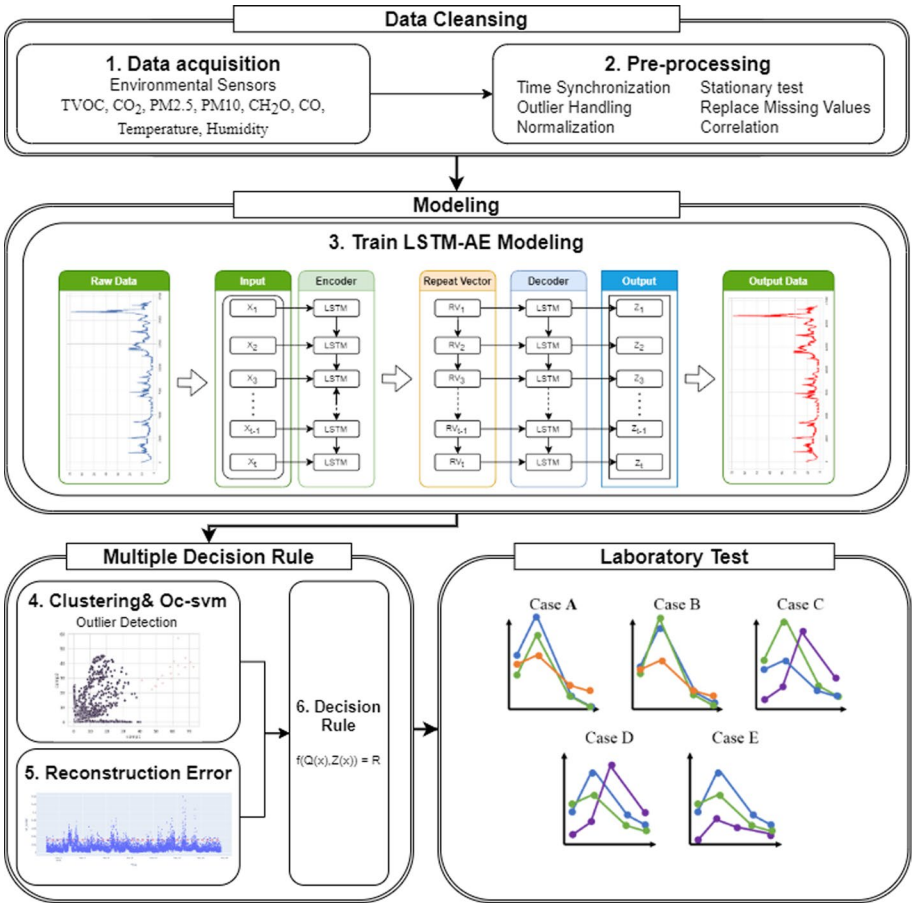


Fig. 1 Framework of the proposed method: flowchart for outlier detection

First, (i) data cleansing is the process of extracting and refining data collected by the sensors. The main components to be used for model training are extracted, and missing values and outliers generated in the sensing process are corrected. (ii) The modeling step trains the LSTM-AE model and extracts the latent characteristics contained in the data based on the trained encoder part. In the case of the existing outlier detection method using LSTM-AE, an outlier detection rule based on the reconstruction error using the difference between the original value and the restored value is used; however, we wanted to define an additional device for more consistent and accurate decision-making. (iii) The decision rule definition step is the process of constructing a sub-algorithm based on the lthrough the steadiness test and that the data that could be used for future atent characteristics passed through the encoder part of the LSTM-AE and combining it with the RE-DM. By combining two independent rules, we propose a new outlier detection rule with a reduced FPR. (iv) Finally, a LAB test was performed to evaluate the performance of the framework defined above. In summary, steps (i) to (iii) represent the process of establishing the proposed anomaly detection framework, and step (iv) is the process of building a LAB test to evaluate the objective performance of the proposed anomaly detection framework.

(a) (b) (c)

Fig. 2 Internet-of-Things environmental sensor device: **a** actual circuit diagram, **b** sensor device structure, **c** sensor device configuration

Table 1 Measurements of performance characteristics of the sensor devices

Sensor type	Detection range	Method
TVOC	0–60,000 ppb	Semiconductor
CO_2	400–60,000 ppm	
PM 2.5	1–1000 $\mu g/m^3$	Optical
PM 10	1–1000 $\mu g/m^3$	
CH_2O	0–5 ppm	Electrochemical
CO	0—1000 ppm	
Temperature	−20–80℃	Semiconductor
Humidity	0–100% (RH)	

Data and devices

To comprehensively measure indoor air quality in the same environment, we used various hazardous substances, such as particulate matter 2.5 (PM2.5), particulate matter 10 (PM10), TVOC, CO_2, CO, and CH_2O, and a temperature and humidity sensors that provide basic information about the environment. Figure 2 shows the circuit design, internal structure diagram, and configuration of the designed environmental devices.

In addition, Table 1 summarizes the specifications and operation methods of the sensors used in the self-manufactured environmental devices.

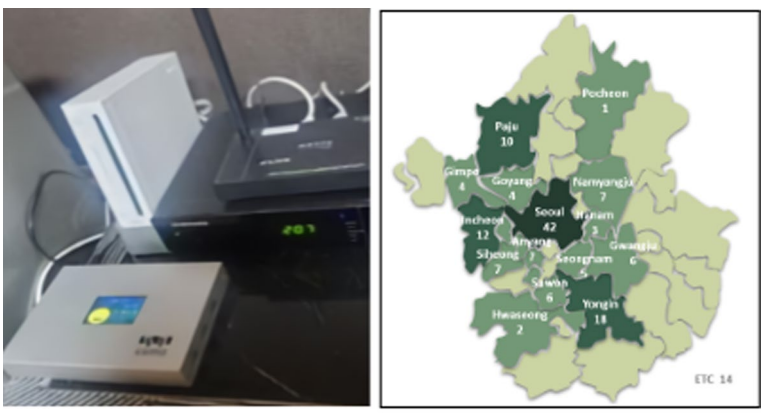
The operating principles of the sensors are semiconductor, optical, and electrochemical, depending on the gas and particles to be detected. Various communication methods such as wireless local area networks, long-term evolution, and fifth-generation environmental sensors can be used. In addition, information is collected through various complex devices, such as a universal asynchronous receiver transmitter, an inter-integrated circuit, and an analog-digital converter, installed inside the sensor. Functions, such as device function reset, and sensor cycle change can be performed through uplink and downlink transmissions and reception.

Outliers may occur because of mechanical defects depending on the operating principle or communication method of each sensor used in the device. The classification of outlier detection technologies can be divided into statistical-based, nearest-neighbor-based, clustering, and classification-based technologies [17]. In addition, a problem can be overlooked as a malfunction of the sensor or communication problem when labeling and processing the types of outliers [18]; thus, it was decided not to classify outliers that occur according to the characteristics of its sensor. In conclusion, the aim was to predict

abnormal elements of harmful factors regardless of the operation principle or communication method of the sensor by measuring with various sensors.

In addition, the experimental target group and target area were selected to collect data from environmental sensors. Samples of at least five different clusters (e.g., region, age, gender, facility, residential environment) were extracted, and participants were recruited from the metropolitan area and central regions of Korea (Seoul, Gyeonggi). In addition, the device was installed in 134 locations: general households, industrial complex areas, and smart city areas. In the case of an industrial complex area, it was installed in a location where ventilation was smooth and local abnormalities did not occur. Because the indoor environment was different for each installation site, the installation sites were selected considering a certain level of common characteristics, such as ventilation and height, to control for some of the environmental conditions. Finally, the 134 sensors were periodically monitored, and it was confirmed that no significant defects or errors occurred in the sensors during measurement. The sensors used in this study have been certified for stability through several tests with standard test methods and procedures performed by domestic specialized laboratories, such as the Korea Electronics Technology Institute and Korea Research Institute of Standards and Science. Figure 3 shows a diagram of the installation of environmental sensors at home and locations of the 134 test beds.

When measuring multiple gases with a sensor, a disadvantage is that the measurement range becomes too wide if it is measured from too far away [19]. Therefore, environmental sensors were placed at an appropriate distance within a controllable range, and cross-validation was performed between the same devices by placing several sensors in the same space to check the allowable error between the sensors were considered and evaluate the reliability factors of non-ideal sensors. When dividing the training and test sets, the study was conducted by not overlapping the divided range of each sensor. As a result, we secured a total of 17,000 real-world data, of which 11,000 learning data, 3000 verification data, and 3000 test data. Through this verification process, the uncertainty inherent in the real-world data was resolved [20], and it was possible to substitute other sensor values for the defect values from the sensor [21]. The device was installed in the



(a) (b)

Fig. 3 Verification of sensor stability: **a** sensor installation at home and **b** test bed location configuration

computer center of Soongsil University in Dongjak-gu, Seoul, and was operated in compliance with the regulations and operation manual of "indoor air quality for multi-use facilities," as prescribed by the Ministry of Environment of the Republic of Korea. In addition, in the configuration of the indoor environment, ventilation occurs at a certain level, and the sensors were placed at a certain distance from the door and window, as shown in Fig. 4.

Preprocessing

Time-series data tuning Univariate time-series data form a sequence of single observations at successive time points. The data index is generally considered to be the observed row, but time is an implicit variable [22]. The time-series data measured by the environmental monitoring sensor showed continuous characteristics and were collected sequentially. Because it is important to set the time at equal intervals, the values of each sensor were measured and simultaneously collected every 2 min and transmitted to the server as an unequally spaced series, owing to the characteristics of the sensor's mechanical defects. Therefore, it is necessary to set the time intervals equally using approximation and interpolation [23, 24].

Stationarity

In time-series analysis, it is important to ensure that the data are stable and free from autocorrelation. Before using the clustering algorithm to detect outliers in the environmental monitoring sensor, the data were checked for normality. Here, normality refers to the behavior in which the mean and standard deviation of the data change over time, and data with such behavior are considered abnormal. We performed an augmented Dickey–Fuller (ADF) test to quantitatively verify normality.

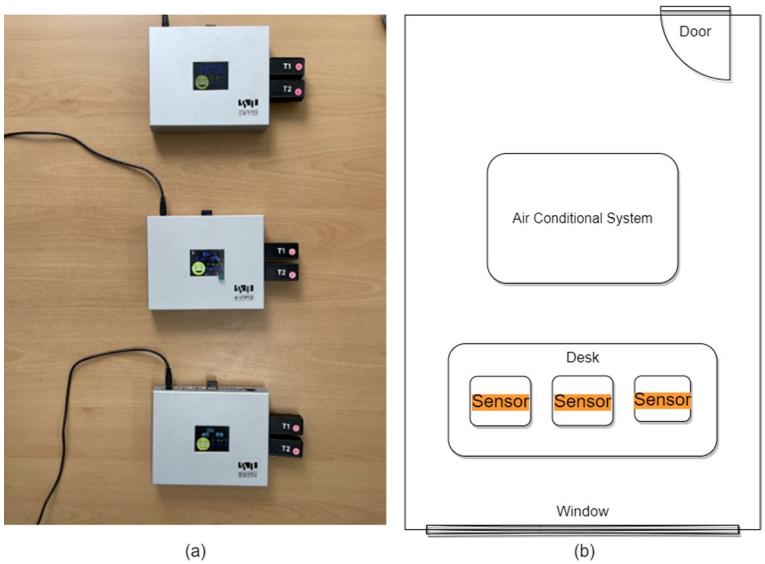


Fig. 4 Verification of sensor reliability: **a** actual sensor device installation and **b** experimental environment configuration

Equation (1) is the amount of change at time t for each sensor gas data point; y_t is the gas data point at time t, α is a constant, β is a coefficient for the time trend, and p is the drift order of the autoregression process. Furthermore, γ is the unit root and represents the influence of the previous y_{t-1}. Therefore, to determine whether γ is negative, the unit root test is performed under the null hypothesis that γ is 0 (Eq. 2), and γ is negative is the alternative hypothesis for the ADF test. Equation (3) can be computed and compared with the relevant threshold for the Dickey–Fuller test. In the ADF test, the null hypothesis is that there is a unit root, and the null hypothesis is rejected. A parameter was considered normal if the null hypothesis was less than 0.05, and the test statistic of the parameter was less than the critical value [25]. Thus, we confirmed that the normality was satisfied. For this test, only negative values were important, so if the calculated test statistic was less than the critical value, the null hypothesis was rejected, and the unit root did not exist. It was verified that each sensor data point was stable by proving that there was no unit root through the steadiness test and that the data that could be used for future modeling.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \tag{1}$$

$$H_0 : \gamma = 0, H_1 : \gamma < 0 \tag{2}$$

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \tag{3}$$

Missing values

A missing value is a case in which a specific part of the data is absent or contains meaningless values [26]. The environmental sensors have two types of missing values. First one occurs when the measured value of the sensor changes rapidly, and the sensor cannot express the corresponding value [27]. Second, missing values may occur depending on the measurement range of the sensor [28]. The sensors used in this study had approximately 1% of missing values, which were processed through linear interpolation. Linear interpolation is a method of replacing missing values according to a linear distance using the values of both adjacent endpoints. In general, linear interpolation is used to deal with missing values in genotyping and machine translation [29]. Equation (4) shows the formula for linear interpolation. When the data values at two points p_1 and p_2 were $f(p_1)$ and $f(p_2)$, any random between p_1 and p_2 is the expression for the imputation $f(p)$ for the point p. Linear interpolation was used for each gas in the sensor, and values were missing at two time points at most; thus, they could be interpolated. Here, $d_1$is the distance from p_1 to p, and d_2 is the distance from p to p_2.

$$f(p) = \frac{d_2}{d_1 + d_2} f(p_1) + \frac{d_1}{d_1 + d_2} f(p_2) \tag{4}$$

Correlation analysis

Multivariate data refer to data that consist of several independent variables. The correlation coefficient used to determine the correlation between variables is one approach that

directly reflects the relationship between the two sequences [30]. In this study, the correlation between variables was investigated using the Pearson correlation coefficient. The correlation coefficient is an indicator of the relationship between two variables regardless of the unit of measurement, and since the correlation of each pair of variables is an important factor in the proposed model, a correlation analysis was performed. In Eq. (5), r_{xy} is the correlation coefficient. The correlation coefficient is obtained by dividing the covariance by each standard deviation, and x and y are gas data points.

$$r_{xy} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2} \sqrt{\sum_i^n (Y_i - \bar{Y})^2}} \tag{5}$$

Process method

LSTM-AE The multi-feature indoor air quality data used in this study are time-series data in which characteristic changes are observed along the time domain. When performing outlier detection on such time-series data, it is very important to consider the characteristics and patterns of the data [31]. In general, a recurrent neural network-based model is used in this case, but it is difficult to expect good performance for long-sequence data because of well-known problems, such as the vanishing or exploding gradient [32]. Therefore, fundamental architectural transformation is required to solve this long-dependency problem, and the LSTM has been proposed to overcome the problem. The LSTM layer adjusts the input/output information from the nonlinear gate unit to and from the memory cell. This allows the model to adaptively learn the long-term dynamic information of the sequence, demonstrating superior performance in time-series data modeling [33]. However, most recurrent neural network frameworks, including LSTM, operate under supervised learning and are unsuitable for unsupervised learning environments, including outlier detection. Data collected from multivariate environmental sensors are unlabeled and cannot be used in a supervised learning framework, as it is very difficult to select objective evaluation criteria [14]. As such, the essence of the problem we want to point out is that most of the data obtained in the real world is unlabeled data, so we cannot fully utilize the existing high-performance artificial intelligence model raised. Thus, as additional procedures are required, we propose an outlier detection model that automatically learns time-series characteristics in an unsupervised learning environment based on the LSTM layer.

As shown in Fig. 5, LSTM-AE is a model in which LSTM layers are stacked in an autoencoder (AE) structure. AE is an unsupervised learning algorithm that is primarily used to learn latent features from unlabeled data that best represent the data. An AE is composed of an encoder and decoder. First, the encoder extracts the latent features by compressing the input value into a low-dimensional bottleneck layer. Then, the decoder restores the low-dimensional vector back to the original input value. This restored output is used to calculate the difference between the original input, and the parameter is modified to minimize the difference. Outlier detection using the AE model determines the outliers for new inputs by setting a threshold based on this reconstruction error [15].

Most existing studies that perform outlier detection using LSTM-AE have one-dimensional decision rules that compare reconstruction errors with pre-calculated thresholds.

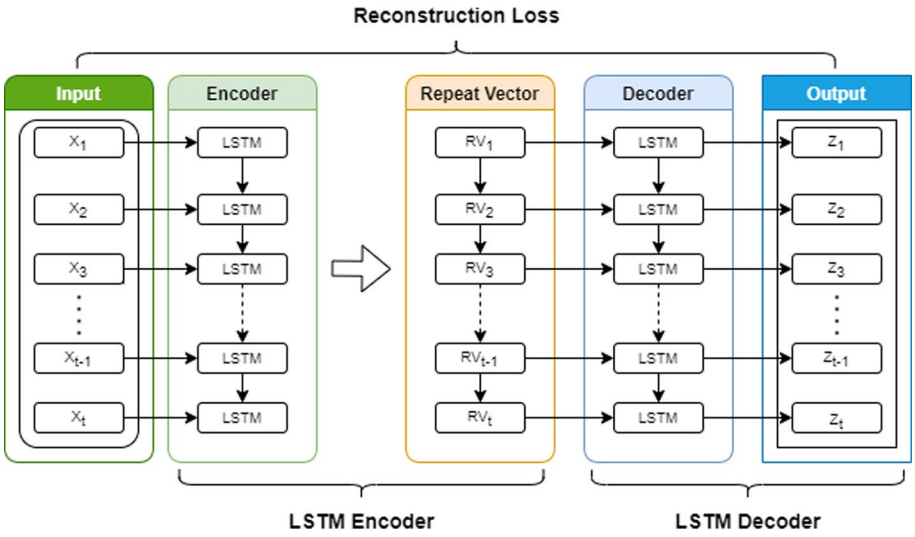


Fig. 5 LSTM-AE architecture

However, such simple decision rules tend to cause performance instability, depending on the quality of the training data. If the model responds sensitively to small noise, the FPR can rise dramatically, causing numerous false alarms in the commercialization stage [16]. Sensing data, in particular, such as actual data, are bound to contain more noise, so detection models with simple decision rules inevitably result in poor performance. Therefore, in addition to RE-DM, we defined a sub-algorithm using latent features generated by encoders. Finally, we propose an accurate and consistent decision rule by combining the discrimination results of RE-DM and the sub-algorithm.

Latent feature clustering and one-class sub-algorithm

The clustering technique, which is widely used in data mining, is a representative unsupervised algorithm that binds data with the same characteristics and distributions into clusters. Density-based spatial clustering of applications with noise (DBSCAN) was first introduced in 1996 as an algorithm to generate clusters of unspecified shapes by calculating the density of neighboring data [34]. An advantage of DBSCAN is that it can be used even when the number of clusters is not known in advance, especially in unsupervised learning environments [35]. The data used in this study are anomalous and have an unpredictable distribution; therefore, DBSCAN is suitable in this situation [16].

In the entire framework, DBSCAN was performed as a preliminary task to obtain a normal latent feature space. In general, clustering is performed on pure input data without any processing procedures [14]. However, this results in poor performance in terms of explanatory power and compression because it does not reflect the high-level characteristics derived by the model [36]. Therefore, we performed clustering on the latent features extracted from LSTM-AE. Thus, a high-level cluster that reflects the time-series and nonlinear properties can be created.

Then, we built a sub-algorithm based on the previously obtained normal clusters. In an actual industrial environment, only normal classes can be used as initial knowledge; therefore, a one-class classification (OCC)-based approach is required [37]. A one-class

support vector machine (OC-SVM) is a special case of the support vector machine (SVM) algorithm used to detect outliers, producing the smallest hyperplane containing training samples based on the premise that most data are normal data [38]. It is possible to determine whether to discriminate outliers based on the inside and outside of the decision boundary created by this hyperplane [31]. Additional quantitative evaluation is possible by calculating the distance difference between the decision boundary and the point.

As such, the latent features extracted through the bottleneck layer improved the performance of the OCC classifier. Next, we define consistent and accurate outlier detection decision rules by combining the discrimination results of the OC-SVM with those based on the reconstruction error of the LSTM-AE.

Decision rule combining OC-SVM and reconstruction threshold

In general, certain outlier detection methods perform better than others because the characteristics of the time-series data generated by each device are different [39]. In addition, the decision rules applied in a single model show somewhat distorted and biased results because the views defining the outliers are different, as are the characteristics and advantages of each model. Therefore, we propose a new outlier decision rule that is reasonable and consistent by combining two different outlier detection models. Recently published papers on the air quality of the environment [40, 41] reported excellent performance in terms of accuracy and efficiency when using several single models in an ensemble method. In this study, we defined a decision rule that combined the previous two models using a voting classifier, an ensemble method. The voting classifier combines the output results of multiple classifiers into one decision rule and is categorized as hard or soft voting. The hard-voting classifier combines multiple classification results using the majority vote method. By contrast, the soft-voting classifier derives the final result by utilizing the decision probabilities of several independent models. In this study, a soft-voting classifier was constructed using the reconstruction error and SVM score [42].

The method of detecting outliers based on the reconstruction error of the LSTM-AE can be seen as a model-based approach [15]. If the model-based approach learns the features of normal data and determines the outliers, then the OC-SVM algorithm is a content-based approach that detects outliers based on whether a particular point is included in the hyperplane formed by most normal data [43]. We intend to develop a generalized model that maintains the advantages of the two models but eliminates the disadvantages by combining two single models approached from different perspectives.

Laboratory test

Objective evaluation of unsupervised learning models, such as the LSTM-AE, has a limit because there are no separate labels. Therefore, to overcome this limitation, an objective evaluation of the model for the self-collected labeled data was performed by conducting a LAB test. The LAB test evaluated the outlier detection performance for various cases by periodically generating different events.

To objectively verify the proposed framework, we conducted LAB tests on five abnormal situations and continuously monitored environmental sensors [44] to determine the

level of outliers and detection performance of the model. The evaluation was conducted for approximately two weeks.

The LAB test process was conducted as follows.

1. First, to increase data reliability, the same three sensor devices were arranged at equal intervals for data collection.
2. The spray method was applied four times a day at regular intervals. The reason for the regular intervals was that it was necessary to secure time for ventilation and device initialization for each event.
3. To maintain the test environment constant, environmental factors that could be variables in the test were controlled. Ambient environmental factors refer to various factors that can cause sensor defects or unintended outliers.
4. For the subsequent calculation of the classification index and the objective evaluation of the model, annotation was performed separately for each abnormality section.

Results

Preprocessing

It was necessary to verify the normality of the indoor air quality data used in the study before full-scale modeling because it is time-series data. Table 2 shows the *p*-value output of the ADF test. In Table 2, it can be seen that the *p*-value of each harmful factor is lower than 0.05 and that the *t*-value is smaller than the critical value within the significance level of 1%. Therefore, the null hypothesis that this is not a stationary time series can be rejected at a significance level of 5% [45]. A threshold is set according to the number of observations, and thus, the null hypothesis that there is a unit root can be rejected [46]. Since the null hypothesis was rejected, we are confident that all seven parameters are normal [47]. Therefore, temperature, humidity, TVOC, CO, CO_2, CH_2O, and PM do not require a difference and are in a steady state. In other words, because there is no unit root, normality is guaranteed.

In addition, approximately 0.92% of the values were not measured by the sensor owing to external factors. After confirming that the sections where the missing values occurred did not significantly affect the surrounding values, this study replaced the missing values using the linear interpolation method.

Additionally, as shown in Fig. 6, environmental substances with strong correlations were identified. TVOC-CO_2, TVOC-CH_2O, and TVOC-Humidity showed a strong

Table 2 Augmented Dicky–Fuller test at various levels

Type	Temperature	Humidity	TVOC	CO	CO_2	CH_2O	PM10
ADF test	−5.336495	−4.812659	−5.780521	−12.57404	−6.217924	−8.065061	−5.708918
Critical Value (1%)	−3.959300	−3.959301	−3.959301	−3.959301	−3.959301	−3.959301	−3.959301
Critical Value (5%)	−3.410747	−3.410748	−3.410748	−3.410748	−3.410748	−3.410748	−3.410748
Critical Value (10%)	−3.127201	−3.127202	−3.127202	−3.127202	−3.127202	−3.127202	−3.127202
p-value	0.000048	0.000445	0.000006	5.1291e-20	7.5653e-07	5.0535e-11	0.000009

correlation, and Temperature-TVOC showed a weak correlation. Before performing multivariate imputation for each variable, modeling was performed considering the characteristics of the variables that showed correlation. The reason for considering variables with clear correlation in the multivariate layer is that it has an important effect in understanding the correlation information restored from the bottleneck of the LSTM-AE [48].

Modeling

LSTM-AE

To train the LSTM layer, the data must be converted into a 3D format:[samples, timesteps, features]. Here, timestep is a hyperparameter representing the size of the window that the LSTM will accept as short-term memory, and the larger the timestep value, the longer the range of series accepted in the short-term memory. To find the optimal timestep, we performed a grid-based search task for four cases [1, 3, 5, 7]. As a result, when the timestep was 3, the model showed the best performance in terms of reconstruction error and cluster cohesion. Therefore, considering that the data used in this study were measured in 2-min cycles, a window was formed in units of 6 min to train the short-term memory in the LSTM layer.

The reconstruction error scatter and distribution plots are shown in Fig. 7. More than 75% of the learning data had a reconstruction error of less than 3% within the interquartile range. This means that the model can restore an output that is similar to the input within an error range of approximately 3%, and it can be seen that feature learning has performed very well. In addition, the reconstruction error scatterplot in Fig. 7(a) shows

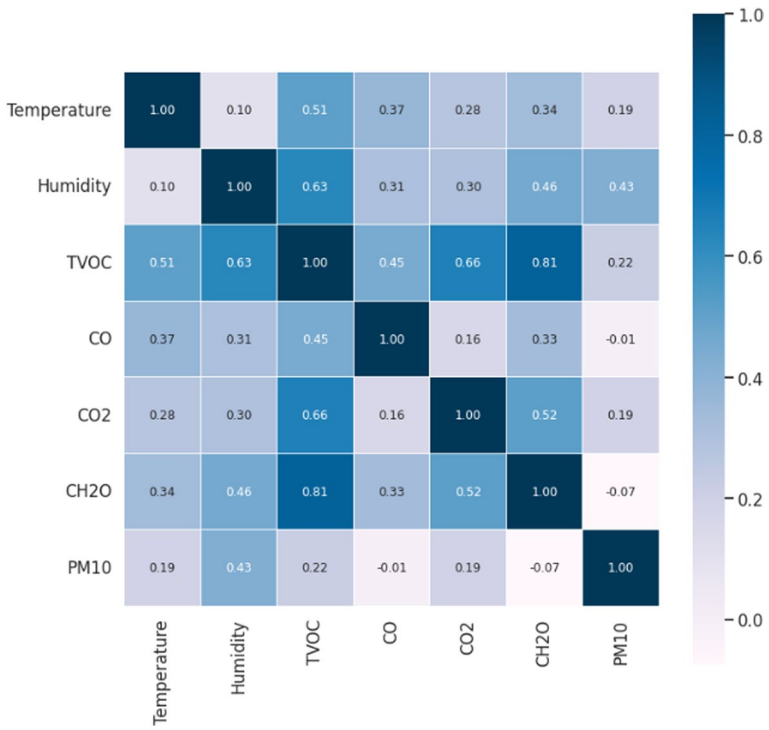


Fig. 6 Pearson correlation between environmental substances

strong cohesion at most points, except for periodically observed anomaly points, which can be seen as the model learning the latent features that best represent normal data, and it can be expected to show collective compressed features in the future.

Figure 8 shows the results of applying LSTM-AE to the training data and test data. In the figure, whenever the input value changes, it can be seen that the overall trend is well followed according to the corresponding amount of change. However, there is a somewhat conservative change when the values change rapidly. Thus, in abnormal sections that deviate from the existing trend and show different patterns, the recovery rate of the data decreases, resulting in a larger reconstruction error. Existing studies determine outliers based on the presence or absence of such a reconstruction error above the threshold, which not only depends excessively on the reconstruction performance of the model but also creates a number of false alarms owing to simple decision rules [14]. Therefore, we define an additional sub-algorithm by utilizing the latent feature space of the LSTM-AE encoder part. In addition, a more complex and consistent outlier determination rule was proposed by combining this sub-algorithm and the detection model based on the reconstruction error.

Latent feature DBSCAN clustering and OC-SVM

We extracted latent features by constructing the bottleneck layer of the LSTM-AE into four nodes. DBSCAN was performed on the corresponding latent features for pre-operation and noise removal for the OC-SVM definition.

Figure 9(a) shows two latent feature components with the best degree of cohesion and separation among the four nodes to visualize the process of forming a DBSCAN cluster. The area where a large number of data points are gathered in the lower left is a normal cluster, and the data on the right side are noise included in the training data. Earlier,

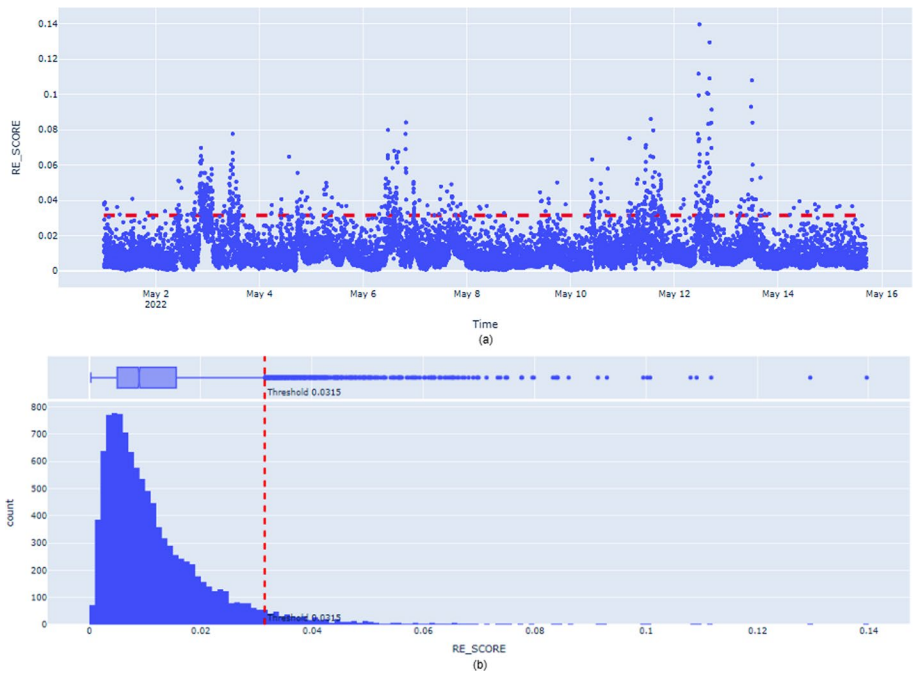


Fig. 7 Reconstruction error: **a** scatter plot and **b** distribution plot

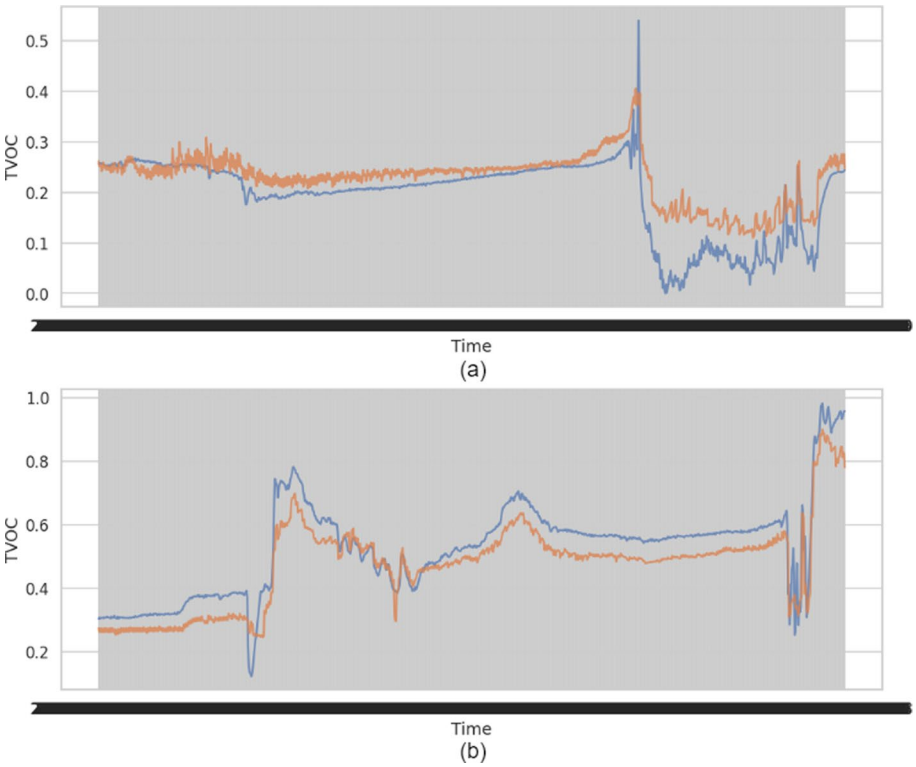


Fig. 8 Truth and prediction error comparison: **a** training set and **b** test set

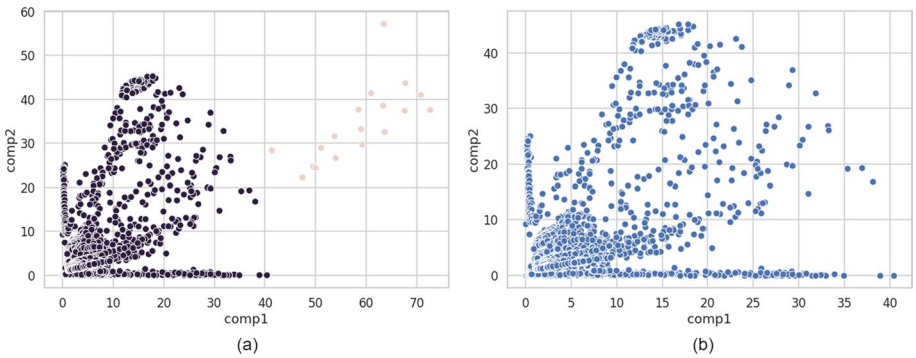


Fig. 9 DBSCAN clustering results **a** total data points and **b** normal data cluster points

it was mentioned that noise in the data adversely affects the normal data characteristic learning of the LSTM-AE. In addition, OC-SVM is performed under the assumption that most of the data are normal; therefore, we tried to remove this fine noise through clustering.

DBSCAN created clusters based on two important hyperparameters. The first is the maximum radius of the neighborhood, and the second is the minimum number of points that can create an independent cluster. We judged the dense population of pointers in the lower left-hand corner of Fig. 9(a) as a normal cluster and adjusted the hyperparameters to include all those points. As a result, a cluster representing the normal data distribution could be obtained (Fig. 9b).

Normal clusters of latent feature spaces were obtained using the DBSCAN. However, DBSCAN only creates a normal cluster within the learning data and does not provide a scoring method to determine outliers in new data samples. Therefore, an additional sub-algorithm is defined to determine the outliers in a latent feature space. The OC-SVM is a representative OCC outlier detection algorithm that can be trained with one class. As a result of applying the OC-SVM algorithm on latent features, a hyperplane to convert the outlier score for the new sample was obtained. In Fig. 10(a), it can be observed that a hyperplane containing most of the data was created, except for some data located at the boundary of the cluster.

Figure 10(b) shows the results of performing OC-SVM on the training and test data. It can be seen that the majority of the test data are well projected within the normal cluster, and a small number of outliers colored yellow were detected outside the cluster. This is the result of the OC-SVM determination of data outside the decision boundary as an outlier, and it is randomly spread.

Decision rule combining OC-SVM and reconstruction threshold

This section discusses the significant performance differences in consistency and accuracy when using reconstruction error-based and OC-SVM-based decision rules independently and presents the results of the new decision rule combining the two algorithms.

Figure 11(a) shows a diagram that expresses the actual outliers included in the LAB test data from the CO_2 perspective. As shown in the figure, outliers appear mainly at extreme and inflection points, and they occur at predetermined time intervals for each LAB test case.

Figure 11(b) shows the results of the detection of outliers based on the reconstruction error, and more outliers were detected than the actual outliers. Therefore, more outliers occurred than in the general case, which considerably affected the normal data discrimination of the LSTM-AE. In fact, it can be seen that the reconstruction error was high overall. Therefore, when using independent decision rules based on reconstruction errors, as previously mentioned in Sect. 2, it is expected that a high FPR will appear.

Figure 11(c) shows the results of the detection of outliers based on the OC-SVM. In the figure, adjacent data points show similar detection results. This can be interpreted

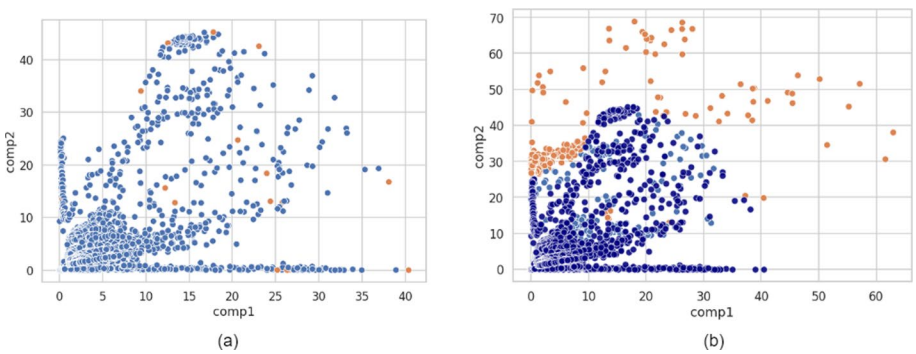


Fig. 10 Latent feature space OC-SVM prediction results **a** training data and **b** training and test data

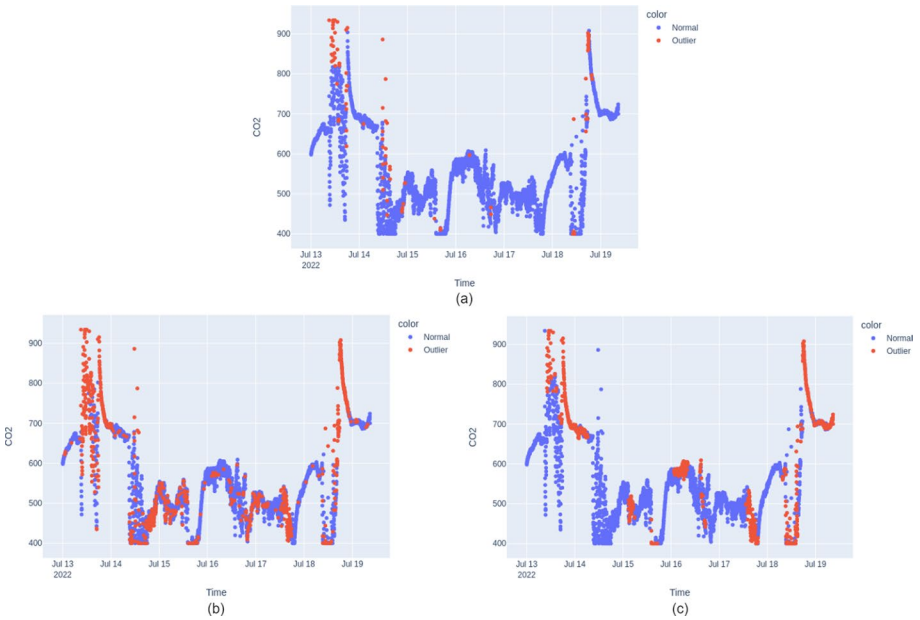


Fig. 11 CO_2 plot over time: **a** actual outliers, **b** reconstruction error-based model, and (**c**) OC-SVM-based model

as a result of distinguishing between pointers included in the hyperplane and pointers not based on their location. Similar to the reconstruction error-based determination method, it is predictable that a high FPR will appear.

Figure 12 shows the results of the LAB test using the hard-voting-based decision rule. In the figure, the number of outliers similar to the actual outlier distribution has been detected, and the outliers are determined for a specific section periodically. In comparison with the previously created independent models, we can also graphically confirm that the hard voting-based decision model performs more consistent and stable detection, while the normal data are easily confused with outlier data, even with small noise.

Figure 13 shows the results of defining the outlier detection rules using soft-voting-based decision rules. In the figure, the number of outliers is significantly reduced in comparison with the previous rules; the outliers are identified only when the model is nearly certain. This is the result of the deliberate adjustment of the rules to reduce false alarms, which can also be interpreted as a more stable outlier detection.

Table 3 summarizes the classification metrics of theconducted LAB test. In the table, it can be observed that the ensemble based onthe voting classifier shows better performance than when the models were usedindependently. In particular, the soft-voting classifier exhibited excellentaccuracy. Therefore, we propose this decision rule as the final model. Insummary, rather than using independent models separately in this outlierdetection process, we demonstrated that ensemble models combining the twoindependent models perform more complex and reasonable decision-making.

Fig. 12 Hard-voting-based CO_2 plot

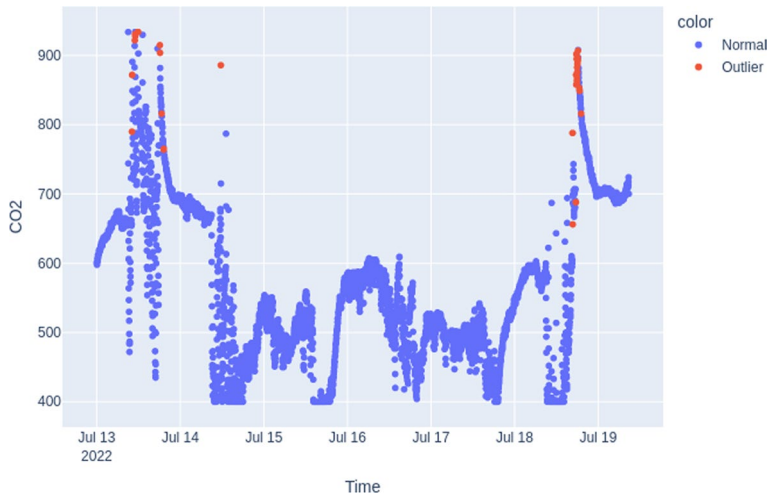


Fig. 13 Soft-voting-based CO_2 plot

Table 3 Results of laboratory test for model validation

Model	Accuracy
Reconstruction error model	0.7992
OC-SVM model	0.7479
Hard-voting model	0.9072
Soft-voting model	0.9766

We table as above for an objective comparison of outlier detection performance and general performance guidelines on other datasets. The compared algorithms are promising algorithms that are evaluated as SOTA-level in the industry for performing outlier detection algorithms on the ECG5000 dataset. From the table above, we can

see that in terms of accuracy indicators, our proposed model performs very well, even considering that the datasets are different.

Laboratory test verification

In order to verify the generalization performance of our model in a diverse environment, we installed about 300 test bed sensors across the country to establish an experimental environment. Figure 14 is a picture of the current status and specific location of the test bad, and it can be seen that it is installed in a wide range of locations, mainly in major administrative areas.

Table 4 summarizes the LAB test results of the soft-voting model performed for each case. These data include major indicators used for prediction, such as latent components 1 and 2, SVM scores, and reconstruction error scores. The table shows the accuracy of each case. Latent components 1 and 2 are key elements that reduce time-series data in window units to a lower dimension, and SVM and reconstruction error scores are major predictive result indicators of each independent model. In addition, among the five cases, the average accuracy score was 0.9568, and no significant differences were found for any of the cases. Because each case was independently conducted assuming a different abnormal situation, the abnormal values detected by the sensor were different in each case. Therefore, when combined with the previous

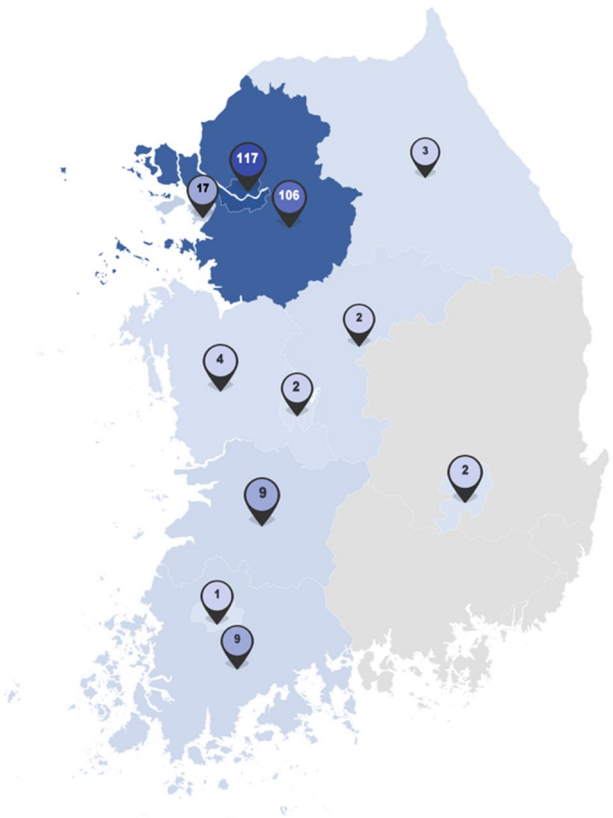


Fig. 14 Test bed installation status and location

Table 4 Comparison with other outlier detection algorithm

Study	Model	Dataset	Performance
Pankaj Malhotra at al. 2016 [49]	SAE-C	ECG5000	Acc 0.934
Gilles Vandewiele at al. 2019 [50]	GENDIS	ECG5000	Acc 0.94
Joao Pereira at al. 2019 [51]	F-t ALSTM-FCN	ECG5000	Acc 0.9496
Proposed Model	LSTM-AE with Ensemble Method	Indoor Environment data	Acc 0.9766

Table 5 Results of the laboratory test for model validation

Label	comp1 mean	comp2 mean		SVM_score	RE_score	Accuracy
Case A	3.850266	14.650074	242.539284		0.075596	0.9211
Case B	1.367582	3.634454	362.680808		0.035693	0.9480
Case C	4.224086	3.098875	371.604030		0.031248	0.9609
Case D	5.048500	0.328430	352.984160		0.034976	0.9778
Case E	10.387705	19.196732	148.181216		0.118963	0.9766
Total	4.975628	8.181713	295.597900		0.059295	0.9568

results, it can be interpreted that the model we built exhibited relatively consistent performance in various environments and cases.

Table 5 is the performance of our model for five abnormal situation sections conducted through the laboratory test. All five abnormal situations were experimented with a slightly different variation, and Case A showed a slightly lower accuracy, but overall, it can be seen that all five abnormal situations showed consistent and excellent accuracy. From this table, we can estimate that our model shows more general and robust performance for various abnormal situations.

Discussion

Previously, it was mentioned that LSTM-AE should be trained with normal data because the model can extract latent features of normal data better than those of abnormal data. The results affect both the reconstruction error-based model and the OC-SVM. However, the initial original data we collected were unlabeled and contained noise due to several physical and environmental factors. Therefore, the most important aspect of this study was the quality evaluation and refinement of the data. We not only conducted various data validation procedures, such as normality verification and correlation measurement, to prove objective rationality for data quality but also processed the data several times to remove noise. Through these processes, we were able to obtain refined analytical data. As a result, the LSTM-AE could be trained stably.

Conclusion

We propose a new method of LSTM-AE based Outlier Detection with Ensemble Method to perform robust & general outlier detection even in the Unsupervised environment. This paper is significant in that it proposed a standard method to implement

Ensemble Method through sub-algebra as well as an artificial neural network-based anomaly detection approach, and it has contribution in that it proved its performance through real-world data collected by hands. Through this technology, it can contribute to protecting the health of many citizens living in daily life, including the city center, which is a densely populated area. Through this framework, we overcame the limitations of the previously proposed outlier detection methods and defined more stable outlier decision rules.

Additionally, we conducted additional LAB tests to demonstrate the effectiveness of the entire framework in real-world abnormal situations. By applying this framework to five different cases, key values used as major indicators of outlier detection, such as the SVM score and reconstruction error, were derived, and an ensemble method was applied to the values. Through this LAB test, we demonstrated that our proposed framework is consistent and outperforms other methods. In addition, a method for performing objective verification of the model in an unsupervised learning environment was proposed.

One of the main advantages of this framework is that it is relatively lightweight, in terms of time and memory. The LSTM-AE requires fewer parameters than other existing DNN models. Therefore, it requires relatively less hardware resources, such as central processing units and random-access memory. Ensemble methods are generally used in building machine learning models. This is because in the case of machine learning, the training of a single model is simple and the number of parameters is small; therefore, multiple models can be trained simultaneously. However, deep learning requires a long time and effort to build an independent neural network. Therefore, in this study, the sub-algorithm derived from the LSTM-AE was used instead of constructing several deep learning models. Through this method, we achieved performance improvements using a single DNN model without the need to build multiple heavyweight DNN models.

This study highlighted the limitations and problems of previous outlier detection studies and suggested solutions, and this framework has a generalized structure that can be applied to various industrial environments. Through this research, it is expected that a more consistent and superior outlier detection model can be constructed using various multi-feature data collected from sensor devices widely adopted in the real world.

Abbreviations

AE	Autoencoder
LAB	Laboratory
ADF	Augmented Dickey–Fuller
DBSCAN	Density-based spatial clustering of applications with noise
DNN	Deep neural network
FPR	False positive rate
IAP	Indoor air pollution
RE-DM	Reconstruction error-based outlier detection method
LSTM	Long short-term memory
OCC	One-class classification
SVM	Support vector machine
TVOC	Total volatile organic compounds

Acknowledgements
Not applicable.

Author contributions
JP suggested the overall framework and technical idea. He led the development of artificial neural network modeling and outlier detection rule algorithms, and participated in the entire research process, including real world data preprocessing and visualization. YS managed the environmental sensor used in the experiment, refined the data, and performed statistical tests. He led the control of the surrounding environment and overall experimentation during the simulation

process, and contributed to the summary and diagramming of the experiment results. Based on his rich theoretical knowledge and numerous research experiences, JC provided technical advice and research support for the technical proposals used in this study. He presented technical improvements and development directions for an outlier detection problem solution using artificial neural networks, and supported creation of a simulation environment and review of manuscripts. All authors read and approved the final manuscript.

Funding
This work was supported by a Korea Environmental Industry & Technology Institute (KEITI) grant funded by the Korean government (Ministry of the Environment), Project No. RE202101551, the development of IoT-based technology for collecting and managing big data on environmental hazards and health effects. This research was also supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MIST) (No. 2019-0-00135, Implementation of 5G-based Smart Sensor Verification Platform). This paper was supported by research funds for newly appointed professors of Jeonbuk National University in 2022.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no Competing interests.

Received: 9 January 2023 Accepted: 2 May 2023
Published online: 17 May 2023

References

1. Zheng D, Li F, Zhao T. Self-adaptive statistical process control for anomaly detection in time series. Expert Syst Appl. 2016;57:324–36. https://doi.org/10.1016/j.eswa.2016.03.029.
2. Akl EA, Gaddam S, Gunukula SK, Honeine R, Jaoude PA, Irani J. The effects of waterpipe tobacco smoking on health outcomes: a systematic review. Int J Epidemiol. 2010;39:834–57. https://doi.org/10.1093/ije/dyq002.
3. Tran VV, Park D, Lee Y-C. Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality. Int J Environ Res Public Health. 2020. https://doi.org/10.3390/ijerph17082927.
4. Bai M, Liu J, Chai J, Zhao X, Yu D. Anomaly detection of gas turbines based on normal pattern extraction. Appl Therm Eng. 2020;166:114664doi. https://doi.org/10.1016/j.applthermaleng.2019.114664.
5. Haque MA, Mineno H. Contextual outlier detection in sensor data using minimum spanning tree based clustering, 2018 International conference on computer, communication, chemical, material and electronic engineering (IC4ME2). IEEE. 2018. https://doi.org/10.1109/IC4ME2.2018.8465643.
6. Lillstrang M, Harju M, del Campo G, Calderon G, Röning J, Tamminen S. Implications of properties and quality of indoor sensor data for building machine learning applications: two case studies in smart campuses. Build Environ. 2022;207:108529doi. https://doi.org/10.1016/j.buildenv.2021.108529.
7. Hill DJ, Minsker BS. Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ Model Softw. 2010;25:1014–22. https://doi.org/10.1016/j.envsoft.2009.08.010.
8. Bhuyan MH, Bhattacharyya DK, Kalita JK. Network anomaly detection: methods, systems and tools. Ieee Commun Surv tutorials. 2013;16:303–36. https://doi.org/10.1109/SURV.2013.052213.00046.
9. Qian L, Pan Q, Lv Y, Zhao X. Fault detection of bearing by resnet classifier with model-based data augmentation. Machines. 2022. https://doi.org/10.3390/machines10070521.
10. Bergmann P, Batzner K, Fauser M, Sattlegger D, Steger C. The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. Int J Comput Vis. 2021;129(4):1038–59.
11. Filzmoser P, Garrett RG, Reimann C. Multivariate outlier detection in exploration geochemistry. Comput Geosci. 2005;31(5):579–87. https://doi.org/10.1016/j.cageo.2004.11.013.
12. Li J, Pedrycz W, Jamal I. Multivariate time series anomaly detection: a framework of hidden markov models. Appl Soft Comput. 2017;60:229–40. https://doi.org/10.1016/j.asoc.2017.06.035.
13. Feichtinger D. Architecture and the challenges of indoor air quality. Field actions science reports. J F Actions. 2020;21:40–3.
14. Li T, Wang Z, Liu S, Lin W-Y. Deep unsupervised anomaly detection, Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021; pp. 3636–3645.
15. Amarbayasgalan T, Pham VH, Theera-Umpon N, Ryu KH. Unsupervised anomaly detection approach for time-series in multi-domains using deep reconstruction error. Symmetry. 2020;12:1251doi. https://doi.org/10.3390/sym12081251.
16. Apostol E-S, Truică C-O, Pop F, Esposito C. Change point enhanced anomaly detection for IoT time series data. Water. 2021. https://doi.org/10.3390/w13121633.
17. Grira N, Crucianu M, Boujemaa N. Unsupervised and semi-supervised clustering: a brief survey. Rev Machine Learning Tech Processing Multimedia Content. 2004;1:9–16.

18. Horsburgh JS, Jones AS, Stevens DK, Tarboton DG, Mesner NO. A sensor network for high frequency estimation of water quality constituent fluxes using surrogates. Environ Model Softw. 2010;25:1031–44. https://doi.org/10.1016/j.envsoft.2009.10.012.
19. Chen L-J, Ho Y-H, Hsieh H-H, Huang S-T, Lee H-C, Mahajan S. An anomaly detection framework for large-scale PM2. 5 sensing systems. IEEE Internet Things J. 2017;5:559–70. https://doi.org/10.1109/JIOT.2017.2766085.
20. Elouedi Z, Mellouli K, Smets P. Assessing sensor reliability for multisensor data fusion within the transferable belief model. IEEE Trans Syst Man Cybernetics Part B (Cybernetics). 2004. https://doi.org/10.1109/TSMCB.2003.817056.
21. Qi H, Iyengar SS, Chakrabarty K. Distributed sensor networks—a review of recent research. J Franklin Inst. 2001;338:655–68. https://doi.org/10.1016/S0016-0032(01)00026-6.
22. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, Stork J. Comparison of different methods for univariate time series imputation in R. ArXiv. 2015. https://doi.org/10.48550/arXiv.1510.03924.
23. Belmonte JC, Manzano J, Arbiol J, Cirera A, Puigcorbe J, Vila A, Sabate N, Gracia I, Cane C, Morante J. Micromachined twin gas sensor for CO and O2 quantification based on catalytically modified nano-SnO2. Sens Actuators B. 2006;114:881–92. https://doi.org/10.1016/j.snb.2005.08.007.
24. Robinson P. Estimation of a time series model from unequally spaced data. Stoch Process Appl. 1977;6:9–24. https://doi.org/10.1016/0304-4149(77)90013-8.
25. Bhandari S, Bergmann N, Jurdak R, Kusy B. Time series analysis for spatial node selection in environment monitoring sensor networks. Sensors. 2017. https://doi.org/10.3390/s18010011.
26. Mealli F, Rubin DB. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. Biometrika. 2015;102:995–1000. https://doi.org/10.1093/biomet/asv035.
27. Panapakidis IP, Bouhouras AS, Christoforidis GC.2018. A missing data treatment method for photovoltaic installations IEEE International Energy Conference (ENERGYCON). IEEE. 2018.https://doi.org/10.1109/ENERGYCON.2018.8398780.
28. Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JM, Finkelstein SN. Missing data in medical databases: Impute, delete or classify? Artificial Intelligence Med. 2013. https://doi.org/10.1016/j.artmed.2013.01.003.
29. Browning BL, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. 2016. https://doi.org/10.1016/j.ajhg.2015.11.020.
30. Son YS, Baek J. A modified correlation coefficient based similarity measure for clustering time-course gene expression data. Pattern Recognit Lett. 2008;29:232–42. https://doi.org/10.1016/j.patrec.2007.09.015.
31. Ergen T, Kozat SS. Unsupervised anomaly detection with LSTM neural networks. IEEE Trans Neural Networks Learn Syst. 2019;31:3127–41. https://doi.org/10.1109/TNNLS.2019.2935975.
32. Zhao H, Sun S, Jin B. Sequential fault diagnosis based on LSTM neural network. IEEE Access. 2018;6:12929–39. https://doi.org/10.1109/ACCESS.2018.2794765.
33. Nguyen H, Tran KP, Thomassey S, Hamad M. Forecasting and anomaly detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. Int J Inf Manag. 2021;57:102282. https://doi.org/10.1016/j.ijinfomgt.2020.102282.
34. Wang W, Hu X, Du Y. Algorithm optimization and anomaly detection simulation based on extended Jarvis-Patrick clustering and outlier detection. Alexandria Eng J. 2022;61:2106–15.
35. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996;96(34):226–31.
36. Zhang C, Liu J, Chen W, Shi J, Yao M, Yan X, Xu N, Chen D. Unsupervised Anomaly Detection Based on Deep Autoencoding and Clustering. Secur Communication Netw. 2021;2021:1–8.
37. Arellano-Espitia F, Delgado-Prieto M, Gonzalez-Abreu A-D, Saucedo-Dorantes JJ, Osornio-Rios RA. Deep-compact-clustering based anomaly detection applied to electromechanical industrial systems. Sensors. 2021;21:5830. https://doi.org/10.3390/s21175830.
38. Mourão-Miranda J, Hardoon DR, Hahn T, Marquand AF, Williams SC, Shawe-Taylor J, Brammer M. Patient classification as an outlier detection problem: an application of the one-class support vector machine. Neuro Image. 2011. https://doi.org/10.1016/j.neuroimage.2011.06.042.
39. Munir M, Siddiqui SA, Chattha MA, Dengel A, Ahmed S. Fusead: unsupervised anomaly detection in streaming sensors data by fusing statistical and deep learning models. Sensors. 2019;19:2451. https://doi.org/10.3390/s19112451.
40. Araujo LN, Belotti JT, Alves TA, de Souza Tadano Y, Siqueira H. Ensemble method based on artificial neural networks to estimate air pollution health risks. Environ Model Softw. 2020;123:104567doi. https://doi.org/10.1016/j.envsoft.2019.104567.
41. Simmons JA, Splinter KD. A multi-model ensemble approach to coastal storm erosion prediction. Environ Model Softw. 2022;150:105356. https://doi.org/10.1016/j.envsoft.2022.105356.
42. Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng. 2021;2:40–6. https://doi.org/10.1016/j.ijcce.2021.01.001.
43. Bicego M, Figueiredo MA. Soft clustering using weighted one-class support vector machines. Pattern Recogn. 2009;42:27–32. https://doi.org/10.1016/j.patcog.2008.07.004.
44. Yousif M, Burdett H, Wellen C, Mandal S, Arabian G, Smith D, Sorichetti RJ. An innovative approach to correct data from in-situ turbidity sensors for surface water monitoring. Environ Model Softw. 2022;155:105461doi. https://doi.org/10.1016/j.envsoft.2022.105461.
45. Witt A, Kurths J, Pikovsky A. Testing stationarity in time series. Phys Rev E. 1988. https://doi.org/10.1016/j.aej.2021.08.009.
46. Dickey DA, Fuller WA. Likelihood ratio statistics for autoregressive time series with a unit root Econometrica, J Econometric Soc. 1981. https://doi.org/10.2307/1912517.
47. Rakib M, Haq S, Hossain MI, Rahman T. IoT Based Air Pollution Monitoring & Prediction System, 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET), IEEE, 2022; p. 184–189.doi: https://doi.org/10.1109/ICISET54810.2022.9775871.
48. Narayanan S, Marks R, Vian JL, Choi J, El-Sharkawi M, Thompson BB. Set constraint discovery: missing sensor data restoration using autoassociative regression machines, Proceedings of the 2002 International Joint Conference on

Neural Networks. IJCNN'02 (Cat. No. 02CH37290), IEEE, 2002; pp. 2872–2877.doi: https://doi.org/10.1109/IJCNN.2003.1224050.
49. Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P, Shroff G. LSTM-based encoder-decoder for multi-sensor anomaly detection. ArXiv. 2016. https://doi.org/10.48550/arXiv.1607.00148.
50. Vandewiele G, Ongenae F, De Turck F, GENDIS. Genetic discovery of shapelets. Sensors. 2021;21(4):1059. https://doi.org/10.3390/s21041059.
51. Karim F, Majumdar S, Darabi H, Chen S. LSTM fully convolutional networks for time series classification. IEEE Access. 2017;6:1662–9. https://doi.org/10.1109/ACCESS.2017.2779939.

Publisher's Note