applied sciences

MDPI

Article

Factorial Analysis for Gas Leakage Risk Predictions from a Vehicle-Based Methane Survey

Khongorzul Dashdondov [1] and Mi-Hwa Song [2,*]

[1] Department of Computer Engineering, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Korea; khongorzul63@gmail.com
[2] School of Information and Communication Science, Semyung University, Jecheon 27136, Korea
* Correspondence: mhsong@semyung.ac.kr

Abstract: Natural gas (NG), typically methane, is released into the air, causing significant air pollution and environmental and health problems. Nowadays, there is a need to use machine-based methods to predict gas losses widely. In this article, we proposed to predict NG leakage levels through feature selection based on a factorial analysis (FA) of the USA's urban natural gas open data. The paper has been divided into three sections. First, we select essential features using FA. Then, the dataset is labeled by k-means clustering with OrdinalEncoder (OE)-based normalization. The final module uses five algorithms (extreme gradient boost (XGBoost), K-nearest neighbors (KNN), decision tree (DT), random forest (RF), Naive Bayes (NB), and multilayer perceptron (MLP)) to predict gas leakage levels. The proposed method is evaluated by the accuracy, F1-score, mean standard error (MSE), and area under the ROC curve (AUC). The test results indicate that the F-OE-based classification method has improved successfully. Moreover, F-OE-based XGBoost (F-OE-XGBoost) showed the best performance by giving 95.14% accuracy, an F1-score of 95.75%, an MSE of 0.028, and an AUC of 96.29%. Following these, the second-best outcomes of an accuracy rate of 95.09%, F1-score of 95.60%, MSE of 0.029, and AUC of 96.11% were achieved by the F-OE-RF model.

Keywords: gas leak prediction; factor analysis; PCA; K-means; XGBoost

Citation: Dashdondov, K.; Song, M.-H. Factorial Analysis for Gas Leakage Risk Predictions from a Vehicle-Based Methane Survey. *Appl. Sci.* **2022**, *12*, 115. https://doi.org/10.3390/app12010115

Academic Editors: Keon Myung Lee and Mi-Hye Kim

Received: 12 October 2021
Accepted: 21 December 2021
Published: 23 December 2021

1. Introduction

In general, methane is a potent greenhouse gas, and as a source of energy, it is in decline. It makes worrying contributions to air pollution and environmental change. Furthermore, the loss of methane gas into the outside air at high levels is a significant health concern that exacerbates pneumonia, asthma, and other respiratory diseases. In addition, gas-related accidents occur frequently, but preventing and responding to such accidents is complicated. According to statistics from 2019, there have been 84 gas accidents and 624 gas-related fires in South Korea [1,2], 65 percent of which were related to LPG and 35 percent related to natural gas leaks. In addition, looking at the causes of natural gas accidents in recent years, a significant percentage has shifted from negligence to those caused by inadequate or aged facilities, corrosion of pipe connections, carelessness in the handling of containers, and self-injury [3]. In the past, research on gas leaks and accidents focused on accident responses through measurements and monitoring using physical sensors. The safety of gas pipelines is predicted using pipeline reliability and time-dependent reliability analysis methods and can also predict maintenance performance.

The primary causes of gas pipeline accidents are external impacts, corrosion of piping, structural and material defects, or natural disasters such as earthquakes. For the prevention of and response to gas leaks and related accidents, information such as the location and degree of the leakage is required and an understanding of whether a leak occurred. However, recent research trends around gaseous fuels seek to understand accident-prevention flows by predicting rather than measuring and monitoring responses by, for instance, pipe risk prediction models using the pressure, output voltages, and protection capabilities [4,5].

Regarding this, various studies have been conducted over the past several decades. Most gas leak detection methods can detect leaks in real-time or minutes after a leak has occurred. Leak size estimation is possible in most detection systems. Cable sensors, soil monitoring devices, ultrasonic flow meters, pressure point analysis methods, and digital signal processing do not have this ability. Technologies related to gas leak detection are mainly divided into non-technical, hardware-based, and software-based methods. Hardware-based methods generally perform better regarding sensitivity [6]. For example, a vapor-sampling system such as LEOS [7] can detect 0.05% leaks, and most non-technical optical methods can detect gas concentrations in the ppm range. A detailed overview of techniques based on signal processing, state estimation, and knowledge-based methods used for chronic leaks of onshore and offshore platforms has been discussed [8].

Data for this paper were collected via a hardware-based method with IoT-based gas detection sensors installed in a vehicle. Therefore, this study measures the data values of the elements in an environment affected by gas and predicts the level of the gas leakage risk without directly measuring the gas loss data based on the relationship between the gas data and the environmental information. These types of predictions can be made quickly using modern ML algorithms.

We propose the feature selection method over natural gas leakage data based on the FA method. The novelty and distinct advantages of our research are as follows:

- The main contribution of this paper is the novelty FA method combined with OE normalization to improve the accuracy of the results of standard ML algorithms by 6~10%.
- Another advantage is that the data are normalized by the OE method, and logarithmic transformation is performed, theoretically resulting in a data distribution identical to that in earlier work [9]. Subsequently, imbalanced data were included in this gas data used to generate the gas leakage level given the k-mean clustering method.
- In addition, the study was also implemented using actual open data that had not previously been used with the ML algorithm, which future researchers will widely use for comparative research.
- It is also possible to use this method to create gas leakage data levels for air assessments in Korea.
- In practice, this level is considered a step forward in forecasting to determine leakages depending on the elements of environmental data information. Advances in the detection of gas leaks in everyday life can also help develop studies that accurately identify the health effects of gas leakages, suggesting the possibility of preventing such leaks and developing medical applications.
- Finally, the limitation of this study is that the multi-variable outlier detection method was not used in the data pre-processing. Our following study hoped that the feature selection method would improve the accuracy results after the multivariate outlier detection method was using Mahalanobis distance [10] and a deep autoencoder.

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

2. Related Work

This section introduces a factor-analysis-based feature-selection method related to the proposed method. Researchers have studied a pilot project of this mapping approach to explore the first step in understanding the effects of NG leaks [9,11]. NG masses were measured using a Picarro CH4 sensor and a Google Street view machine [11]. This refers to gas sensors that are resistant to fire and wired and wireless transmitters that can be used in high-sensitivity facilities. In addition, the vehicle used is an IoT-based

remote monitoring system, with a dual-antenna diagnostic solution used for real-time data aggregation analysis.

Further, we present a list of environmental and gas features in raw data properties of NG found in mobile-device-based methane gas research [12]. When NG is released into the air, it creates major climate and environment problems. The fluctuations in the sensitivity of the devices used to measure the CH4 level are relatively high. One study is related to the present study [13], where ordinal-encoder (OE) normalization was used as the target feature of CH4 (g min). Then, k-means clustering for the labeling of CH4 for the data pre-processing part was used. The earlier work [9,11] sourced the flow rate of the CH4 leaks. A machine-based approach to environmental engineering has been widely used to predict natural gas leaks [14,15]. Natural gas pipelines were subjected to real-world acoustic emission signal coefficient and cluster analyses [16]. A comprehensive review and a model for predicting mercury emissions based on the coal supply characteristics and operating conditions have been discussed [17]. Researchers used factor analysis to reduce the high-dimensional variables' size and obtain the optical properties during the cluster analysis [18]. As mentioned above, we also compared the results of the proposed method with several statistical studies based on other similar gas data. However, no other researchers have utilized machine-learning algorithms on these data types before. Using a comprehensive database, factors such as various pressures, temperatures, and composition ranges successfully implement natural gas pressure and emissions as an intelligent model for predicting evolution [19]. Some authors have proposed systems for detecting leaks in water and natural gas networks in residential and construction environments [20]. The Gaussian mixed model (GMM), the hidden Markov model (HMM), and the one-class support vector machine (OC-SVM) were validated according to comparative perspective statistical modeling, and the features were selected by modifying the sequence character selection (SFS). The best result was 95.32% for the hourly and daily features for both the GMM and HMM on the Almanac of Minutely Power dataset (AMPds) [21] and the Department of International Development (DFID) [2] datasets.

Standard machine-learning techniques are divided into supervised, semi-supervised, and unsupervised. The data we use are unsupervised data without labels. XGBoost is widely regarded as flexible as it contains evaluation functions and a variety of optimization options. It can also utilize tree pruning using the Greedy algorithm, which makes it powerful for overfitting. With a special connection to other algorithms and excellent utility, it is capable of ensemble learning, which is another strength [22]. When creating a tree, XGBoost uses CART (classification and regression). All leaves are related to the model's final score, unlike the decision-making tree that only considers the result values of leaf nodes. The purpose of this study was to compare the XGBoost, KNN, DT, RF, NB, and MLP [4,10,20,23–26] classification techniques for feature selection using factor analysis (FA) [5] and OE normalization for risk predictions of NG leaks. The system identifies pipeline gas leaks with an input vector based on the Manhattan distance of certain features by exploiting the k-nearest neighbor classifier. The result is a maximum accuracy of 90%, depending on the relevant threshold value [27]. Furthermore, in another work [28], average accuracy rates were 99.28% and 99.97% using features extracted directly from acoustic emission (AE) signals to train the KNN classifier in a pipeline system. Another study compared different prediction methods regarding the valve internal leakage rate [29]. In that paper, factor analysis was also introduced to realize the dimensionality reduction in the AE signals of a valve. Their accuracy rates were between 74.5% and 98.8%. In other research [30], k-means clustering was used to contribute to constructing the cascade support vector data description (Cas-SVDD) and improving the accuracy of pipeline leak detection. In addition, a combination of CNN and the depiction of underwater acoustic signals in the time-frequency range were studied [31] to provide a new understanding of underwater pipeline error recognition methods, with the recognition accuracy validated via a pipeline leakage rate of 95.5%. In another work [32], applied convolutional neural networks (CNN) were used to classify whether images (CCTV frames) are of different

classes of natural gas leaks in onshore wellheads. Those experimental results showed an accuracy rate of 99.78%. Two studies [33,34] implemented a feature-selection method based on SVM classifiers to identify leak severity levels of gas pipelines for a real-time monitoring system. The corresponding accuracy rates were 95.6% and 99%.

In this paper, the F-OE method is suggested to increase the accuracy of natural gas data level detection. The main contributions of this research are that we propose a novel factor analysis method and OE-based ML method to predict NG leakage risk levels in open data in an unsupervised mode. Here, after we performed a ten-scale logarithmic transformation, the overall content distribution pattern of the original data did not change significantly until the k-means clustering. Therefore, our method is suitable for the early prediction of NG leaks in the air. It presents higher accuracy and higher AUC scores than the techniques. The model is evaluated for its accuracy, mean squared error (MSE), F1-score, and receiver operating characteristic curve (ROC).

3. Methodology

In this paper, we used Picaro's vehicle-based methane gas open data. Figure 1 shows the general system architecture of the proposed method. In this method, initially, we cleaned data by removing null and constant variables from rows and columns. In this case, we selected 17 features from 33 features [12]. Next, we set the features for the factorial analysis based on a principal component analysis (PCA). Here, we selected seven component features from the original 17 attributes. After feature selection, we normalized the data using the OrdinalEncoder technique. In addition, we standardized these features in terms of the Z score. We also divided data into two parts, referring to them as the gas and environment datasets. In the gas attribute, we used k-means clustering for labeling. After labeling, we combined gas and environment data by conferring the labeled dataset.

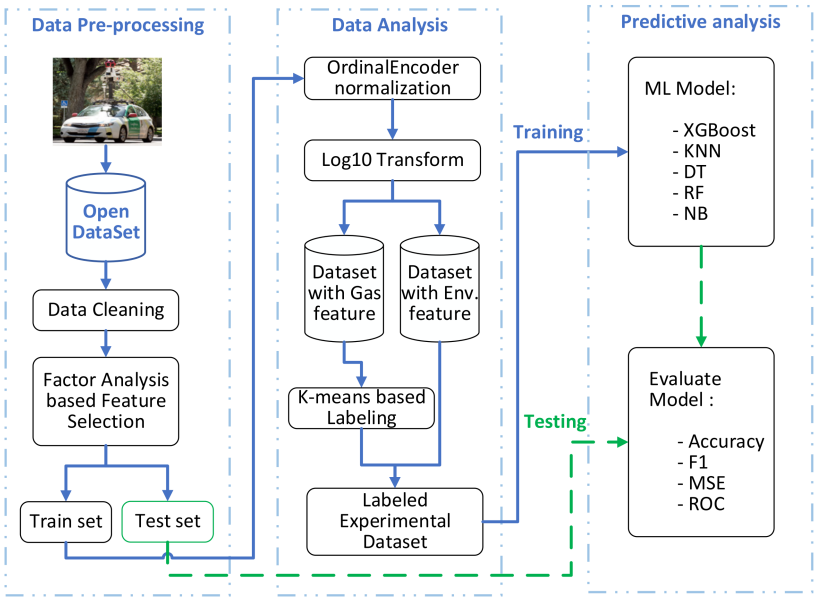


Figure 1. The general system architecture of the proposed method.

In this labeled dataset, we launched the training of the machine learning models for predictive analysis. Behind the training step, we tested predictive and evaluative models for their accurate measurements.

3.1. Factorial-Analysis-Based Feature Selection

Feature selection proposes a simple and effective technique to eliminate redundant and irrelevant data. Removing the unrelated attributes improves accuracy, reduces the computation time, and facilitates an enhanced assuming of the learning model. Factor analysis is a technique used to reduce many variables to a few factors. This method

calculates all variables' most common variance levels and scores them. Factorial analysis has several possible approaches. In this paper, we used the most common method for the principal component analysis to extract factors from the dataset. PCA extracts the maximum variance and sets these values as the first factor [35]. The variance explained by the first factor is then calculated, after which the maximum variance of the second factor is calculated. This process ultimately moves to the last element. This model led to a determinant of 0.022, no multicollinearity, a KMO factor of 0.541 compatible sample sizes, and significance on the 0.001 level (alpha < 0.005), allowing rejection of the null hypothesis and therefore showing a correlation between the variables. Here, we utilized a factor analysis. Table 1 shows the descriptive statistics for all features, including the means, standard deviations, and the communality value extraction method by PCA of the components used in the factor analysis. The number of cases was N = 69,831.

Table 1. Descriptive statistics and communalities of the factor model (N = 69,831, Initial = 1).

Features	Mean	Std. Deviation	Communalities Extraction
CavityPressure	139.99	0.0076	99.4
CavityTemp	45.02	0.1948	69.3
DasTemp	48.27	4.5963	79.9
EtalonTemp	44.71	0.0429	89.9
WarmBoxTemp	45.01	0.0359	88.0
OutletValve	28,652.2	402.65	86.6
GPS_ABS_LAT	33.5	0.0155	31.9
WS_WIND_LON	1.01	24.393	51.7
WS_WIND_LAT	−5.65	11.604	49.2
WS_COS_HEADING	−0.003	0.7081	54.7
WS_SIN_HEADING	−0.033	0.6717	36.9
WIND_N	1.46	18.535	62.9
WIND_E	−4.005	17.589	58.1
WIND_DIR_SDEV	25.22	24.215	51.9
CAR_SPEED	4.17	5.5035	57.8
CH4	2.005	0.2032	83.2

Given that PCA is an iterative evaluation process, it starts with one as the initial estimate of the commonality (this is the total variability of all seven components for our dataset) and then continues the analysis until the initial collaboration is found. This means that the initial value on the diagonal of the correlation matrix is determined by the axial factoring of the main factor in terms of the square of the variable with the other attributes.

These results are shown as the percentages of features of the values explained by the corresponding coefficient for the given variable. In other words, we find that the factor model defines about 83.2% of the variation in CH4. Likewise, 99.4% of the CavityPressure, 89.9% of the EtalonTemp, 88.0% of the WarmBoxTemp, 86.6% of the OutletValve, and 79.9% of the DasTemp are explained by our factor model. These results present the best outcome of explaining the variation in the CavityPressure, CavityTemp, DasTemp, EtalonTemp, WarmBoxTemp, OutletValve, WIND_N, and CH4 factors in the factor analysis. Hence, they appear as percentages of the features of the value explained by the coefficient for the variable. We demonstrate that the factor model explains approximately 83.2% of the variation in CH4. Likewise, 99.4% of the CavityPressure, 89.9% of the EtalonTemp, 88.0% of the WarmBoxTemp, 86.6% of the OutletValve, and 79.9% of the DasTemp are explained by our factor model. The results present the best outcome of explaining the variation in the CavityPressure, CavityTemp, DasTemp, EtalonTemp, WarmBoxTemp, OutletValve, WIND_N, and CH4 in the factor analysis.

We sought to determine values that are close to the initial values. This model shows that most of the features of these variables are explained. In this case, the model is better for some variables than others. The model best describes CavityPressure, DasTemp, EtalonTemp, WarmBoxTemp, OutletValve, and CH4, with the outcomes not bad for other vari-

ables, such as CavityTemp, WS_WIND_LON, WS_COS_HEADING, WIND_N, WIND_E, WIND_DIR_SDEV, and CAR_SPEED. However, for other variables such as GPS_ABS_LAT, WS_WIND_LAT, and WS_SIN_HEADING, the model does not work very well and only explains about half of the changes. PCA extracted this model. Table 2 shows a rotated component matrix of the suggested factor model. We also extracted the PCA model and rotated Varimax with the Kaiser normalization method, which converged in 15 iterations. In this case, we extracted seven components.

Table 2. Rotated component matrix.

Features	Component						
	1	2	3	4	5	6	7
DasTemp	0.868						
OutletValve	0.848						
GPS_ABS_LAT							
WS_WIND_LAT							
EtalonTemp		0.943					
WarmBoxTemp		0.931					
WIND_N			−0.783				
WS_WIND_LON			0.688				
WS_SIN_HEADING			0.507				
WIND_E				0.682			
WIND_DIR_SDEV				0.661			
WS_COS_HEADING							
CavityTemp					0.684		
CAR_SPEED					0.590		
CH4						0.908	
CavityPressure							0.997

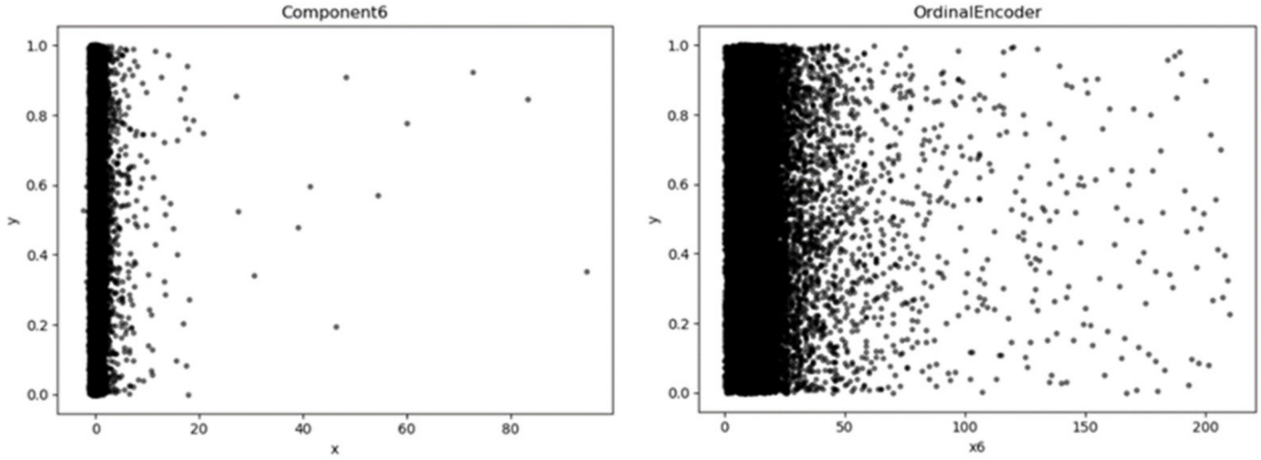
The rotated component matrix is called the load, which is the main output for the principal component analysis, which then includes calculating the relationships between each variable and the calculated components. In Table 2, moderate to strong correlations between DasTemp, OutletValve, and component 1; EtalonTemp, WarmBoxTemp and component 2; WS_WIND_LON, WS_SIN_HEADING, and component 3 (here WIND_N shows a positive correlation with component 3) are shown.

Furthermore, WIND_E, WIND_DIR_SDEV, and component 4; CavityTemp, CAR_SPEED and component 5; CH4 and component 6; and CavityPressure component 7 were found, with other features and each component showing very low correlations. Accordingly, the first component appears to measure DasTemp, and the second component measures EtalonTemp; the third component measures WS_WIND. The fourth component appears to measure WIND_E, the fifth component measures CavityTemp, the sixth component measures CH4, and the seventh component measures CavityPressure.

3.2. Data analysis

3.2.1. Data Normalization Using Ordinal Encoder and Log10 Transform

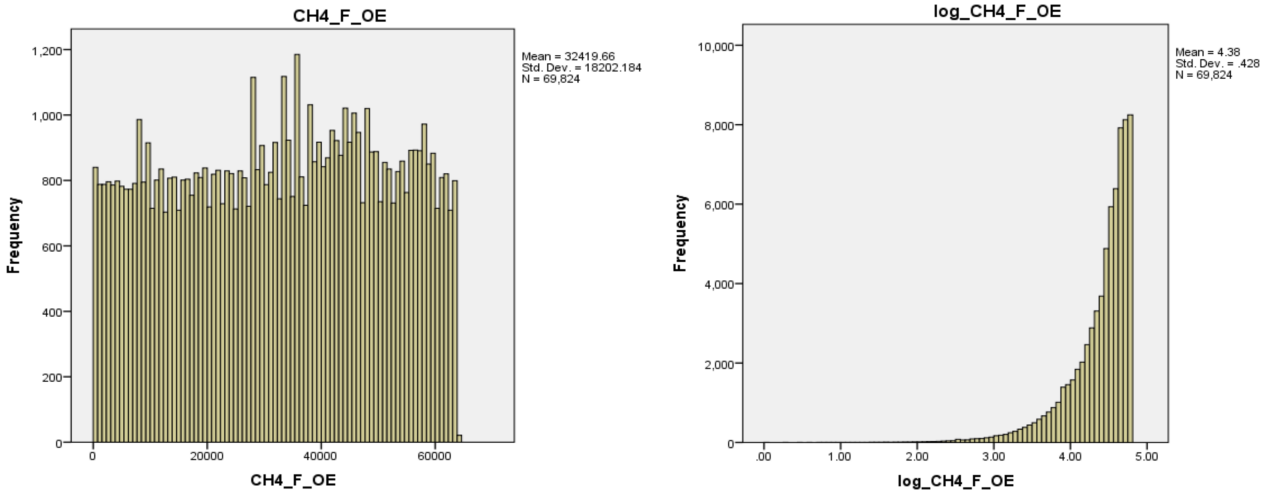
The normalization technique organizes a database to minimize duplicate and redundancy data. We encode categorical variables as an integer array. The input of this transformer is identical to the integer or a string array and represents a value obtained according to the category (discrete) characteristics. This section converts features into ordinal integers. As a result, one integer column (0 to n − 1) appears in one element, and n is the number of categories. We implemented OE normalization for all selected components [13]. Figure 2 shows plots of component 6 with and without OE.



(**a**) (**b**)

Figure 2. Plots of CH4 components data with OE normalization data for NG: (**a**) Below FA-selected 6th component named by CH4 and (**b**) normalization OE of component 6-CH4.

Additionally, we transformed the log10 scale transform with all seven components after OE. The results are shown in Figure 3. The distribution of the initial values of the data mentioned in earlier work [9] is similar.



(**a**) (**b**)

Figure 3. Histogram of F-OE-CH4 component data vs. log transform with F-OE-CH4 components for NG: (**a**) histogram of component 6 named by CH4-F-OE, and (**b**) Log10 scale transformed histogram of CH4-F-OE.

3.2.2. Data Labeling Using K-Means Clustering

We select the 69,824-row dataset from six environment datasets and one gas dataset in this session. The first open dataset has no label. Therefore, we used the simple and commonly used k-means algorithm to make the label for our experimental dataset. It does not influence the classification process, and the target value data needs a class label. Hence, the K-means algorithm with a ten-scale log transform was applied to assign the class information and subsequently define the risk of CH4 as low, medium, or high. K-means is a multi-variable classification method developed by MacQueen in 1967 [35]. The main concept is distributing the variables of the nearest class n values into k subgroups. The basic idea is to divide the variables into k subgroups of n values in the most comparable

class. First, all samples belong to group k. The method then calculates the Euclidean norm between the model to be measured and the core point of each class and recalculates the new core coordinates. It then computes the Euclidean norm until all samples to be measured can no longer be allocated. In Figure 4, we illustrate the k-means clustering results using a histogram. Here, the final cluster center includes the low, medium, and high marks, which reached 1.839, 2.116, and 2.638, respectively, nearly identical to the median threshold of 2.16 (ppm) for defining elevated CH4 levels in earlier work [11].

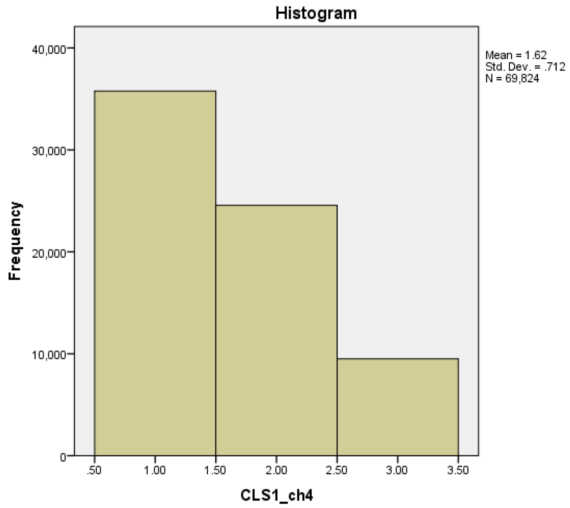


Figure 4. Histogram of clusters for the gas dataset with CH4.

4. Experimental Study

4.1. Experimental Dataset

The datasets were used to propose a model for gas leak predictions in experimental data from earlier work [9,11]. Therefore, it is referred to here as "03/15/2017–03/25/2017" for Sample_Raw open data [12]. Initially, we removed a row of missing values and features unrelated to gas leaks, after which there were a total of 69,824 records from 78,771 records originally and 17 features from 33. Afterward, we selected features according to the F-OE model of natural gas level detection, which was, in this case, seven features. Figure 5 shows the procedure used to create the target dataset. Regarding the default settings of the training (70%) and testing (30%) sets, we defined three class labels, low, medium, and high, for the target features of an experimental dataset. The descriptive statistics for the target variables are described in Table 3.

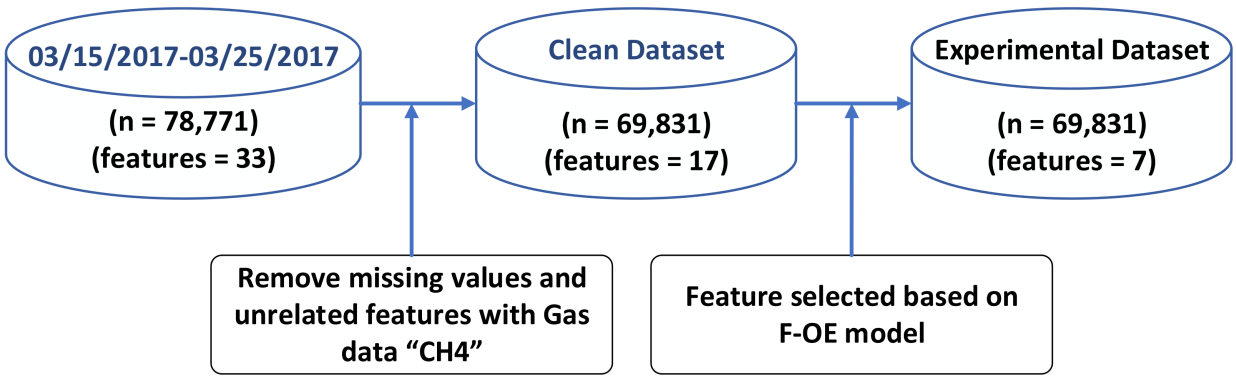


Figure 5. The experimental dataset preparation procedure of the NG dataset.

Table 3. Description of classes for experimental datasets.

Class	Total	Train 70%	Test 30%
Low	35,765	25,111	10,654
Medium	24,561	17,121	7440
High	9505	6649	2856
Total	69,831	48,881	20,950

4.2. Classifiers

We compared the proposed F-OF model with those based on XGBoost, KNN, DT, RF, and NB. According to the compared algorithms, we selected the best values of the input parameters by changing these values until the performance increased. Table 4 shows the parameter configuration for the classification models compared here.

Table 4. Parameter setup for the compared classifiers.

Algorithm	Parameters	Optimal Values
XGBoost	n_estimators: The number of boosting stages to perform. Gradient boosting is robust to over-fitting; accordingly, a large number usually results in better performance. This was configured to be between 50 and 400 and increased in steps of 50. learning_rate: The learning rate shrinks the contribution of each tree according to the learning_rate value. max_depth: Maximum depth of the individual regression estimators. The maximum depth limits the number of nodes in the tree. We tune this parameter for the best performance; the best value depends on the interaction of the input variables.	n_estimators = 300, learning_rate = 0.1, max_depth = 12
KNN	n_neighbors: Number of neighbors. It is configured between 2 and 30.	n_neighbors = 3
DT	criterion: 'gini' for the Gini impurity and 'entropy' for the information gain measurements of the criterion of a split was used to identify the best decision tree splitting candidate.	criterion = 'gini'
RF	n_estimators: The number of trees in the forest. It was configured to be between 10 and 100 and increased in steps of 10. criterion: "gini" and "entropy" were used as splitting criteria.	n_estimators = 100, criterion = 'gini'

The experiment was conducted on a computer with a 3.6 GHz Intel (R) Core (TM) i7-4790 CPU, an NVIDIA GeForce GTX Graphics 960 graphics card, and 16 GB of RAM. The open-source libraries Scikit-learn and Keras, written in Python, were used for the machine learning algorithms. SPSS 20.0 conducted the all-logarithm transformation, factorial, and k-means clustering analyses. We compared the proposed approach to other machine-learning-based algorithms, specifically XGBoost, RF, KNN, DT, NB, and MLP, on an open dataset.

4.3. Evaluation Metrics

The performance evaluation in this study was performed using the metrics of accuracy, AUC, F1 score, and MSE. We found the precision and recall values using the equations below [24].

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{1}$$

and

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{2}$$

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

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

We studied three-class cases, where the average of the F1 score for each category is weighed against the average parameter from Equation (3). Accuracy is a measure of the degree to which the calculated value is close to the actual value. Accuracy is the sum of the true positive fractions and the actual negative fractions of all the test data, expressed as Equation (4).

$$\text{Accuracy} = \frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}} \tag{4}$$

In addition, one of the measures we evaluated was the mean squared error (MSE) relative to the actual values for risk predictions of NG leaks [36].

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

with m and n referring to the number of observations, where m is the number of data instances and n serves to predict NG leakage. Here, X and Y are the actual and predicted values for the i and jth data points, respectively.

5. Results and Discussion

The accuracy, F1-score, MSE, and ROC curve measurements of the performance results are shown in Table 5, where the highest values of evaluation scores are marked in bold. If we did not use the F-OE model to predict the gas leakage level, the XGBoost algorithm showed the highest performance, with an accuracy rate of 94.99%, F1-score 95.53, and ROC 96.17% compared to the other machine learning algorithms of KNN, DT, RF, and NB. However, a factor analysis with OE (F-OE) model-based XGBoost and RF algorithms outperformed when using only the XGBoost algorithm. The proposed F-OE-XGBoost model achieved the highest accuracy of 95.14%, F1-score of 95.75%, MSE of 0.028, and AUC of 96.29%. Following this, the F-OE-RF model achieved the second-best accuracy rate of 95.09%, F1-score of 95.60%, MSE of 0.029, and AUC of 96.11%. The F-OE-NB model showed lower results compared to the other predictive models of the evaluation metrics. Table 5 shows that they were reduced from 17 dimensions to 7 dimensions by factor analysis, the accuracy of gas leak detection increased in all algorithms. Then indicates that the proposed factor analysis-based feature reduction method is suitable for predicting gas leak detection.

Table 5. Evaluation results of the compared algorithms on the experimental dataset (%).

	Algorithms	Accuracy	F1	MSE	AUC
7 features with OE	F-OE-XGBoost	**95.141**	0.957579	0.028226	96.29
	F-OE -KNN	**95.018**	0.952601	0.031551	95.88
	F-OE -DT	90.859	0.937716	0.041496	94.78
	F-OE -RF	**95.093**	0.956036	0.029196	96.11
	F-OE -NB	65.709	0.731945	0.173381	76.16
7 features	F-XGBoost	94.993	0.955336	0.029722	96.17
	F-KNN	93.079	0.947657	0.034163	95.07
	F-DT	56.678	0.678796	0.204025	72.82
	F-RF	69.709	0.703196	0.179618	74.02
	F-NB	70.916	0.762133	0.140000	76.74
17 features	XGBoost	88.382	0.896736	0.065044	83.71
	KNN	84.902	0.850603	0.096150	79.39
	DT	80.267	0.857235	0.095768	82.14
	RF	88.702	0.896528	0.064185	82.86
	NB	64.821	0.658825	0.207160	56.49

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

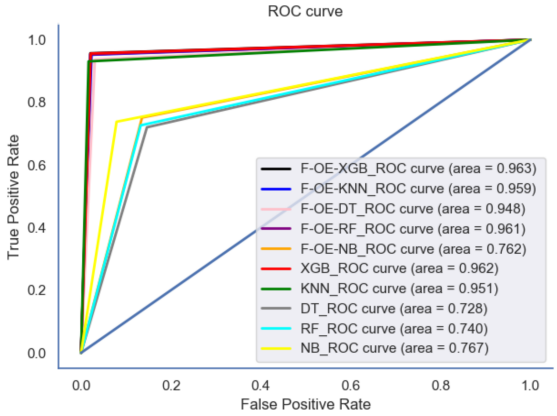


Figure 6. Receiver operating characteristic curves of the algorithms compared to the F-OE method.

In addition, the ROC curve for each level of the experimental dataset in Figure 7 was determined. As mentioned above, we proposed to find a better model performance to predict medium and high-level classes for the experimental datasets. The F-OE model-based XGBoost shows the following higher ROC scores of 96.7% (low level), 96.1% (medium level), and 96.1% (high level) compared to the others for all class levels.

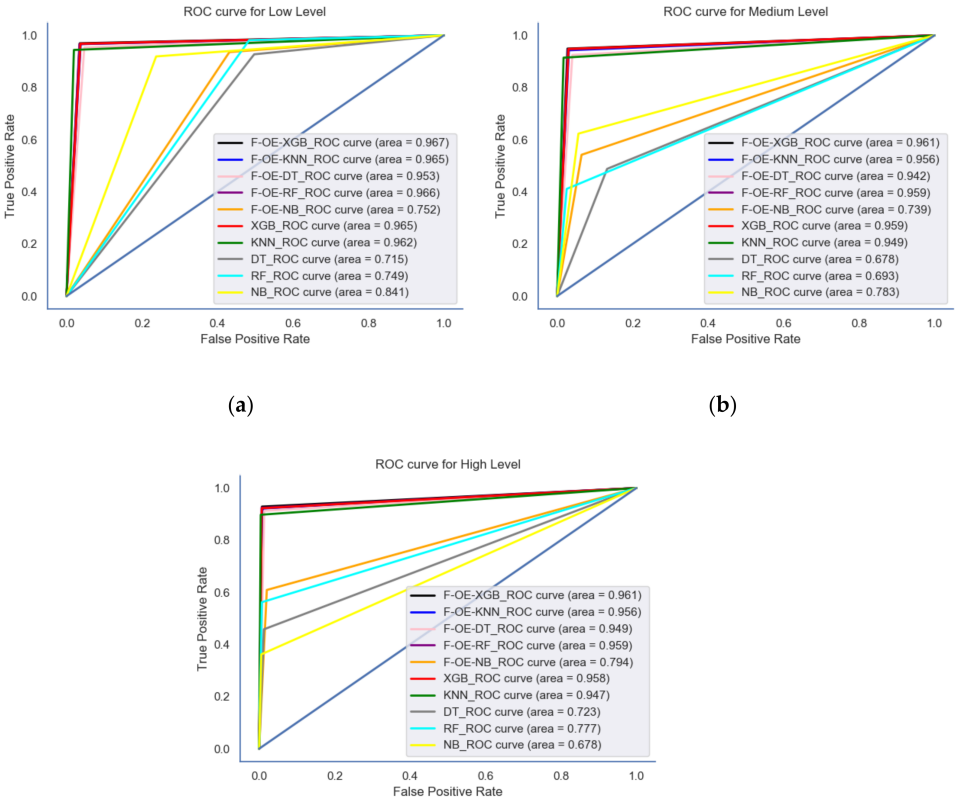


Figure 7. Multi-class ROC curves of the compared F-OE algorithms for low, medium, and high levels: (**a**) ROC curve for the low-level class, (**b**) ROC curve for the medium-level class, and (**c**) ROC curve for the high-level class.

Finally, we tuned the hyperparameter of XGBoost using the grid search infrastructure in Scikit-learn on the experimental dataset. Evaluating the effectiveness of the Bayesian hy-

perparameter optimization method on the XGB model increases the model's efficiency [17]. They then assessed values for the max depth between 1 and 12 (2, 4, 6, 8, 10, and 12).

We plotted the relationship between each series of max depth values for the given n estimators in Figure 8. We have achieved the best result with the settings of n-estimators = 350 and max depth = 12.

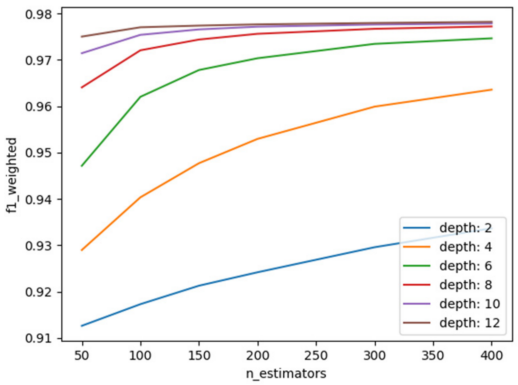


Figure 8. The plot of results of the optimal 'max_depth' and 'n_estimators' parameters according to the F1-weighted score for the experimental dataset in XGBoost.

As mentioned above, in practice, it is possible to make forecasts for determining leakages depending on the elements in the given environmental data information. Furthermore, one limitation of the F-OE method is that it did not use the multivariate outlier detection method for data pre-processing. Therefore, the detection accuracy results may be better if an outlier detection method is used. Accordingly, in our following study, we will focus on several multivariate outlier detection methods.

Figure 9 shows the time performance of algorithms used in our experimental data by seconds. However, for this dataset, the complexity time of the F-OE-based KNN, DT, RF, and NB algorithms work faster than the F-OE-based XGBoost algorithm. Our proposed F-OE-based algorithms achieve higher accuracy compared with the other algorithms. In conclusion, our proposed algorithm's accuracy complexity is more elevated than different algorithms, and the computation time of our algorithm is constant.

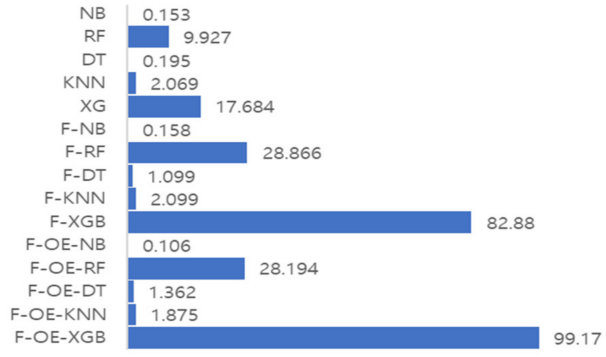


Figure 9. Comparison of computation time of algorithms (sec).

6. Conclusions

This paper presents predicted natural gas leakage levels determined using feature selection based on a factor analysis of actual open data. We conducted training with OrdinalEncoder normalization and ten-scale log transformation on a cleaned dataset and then classified gas data using the k-means clustering method. The analysis revealed various component factors associated with gas leakages. A PCA-based factorial analysis model extracted these features. The F-OE method analysis found that gas leaks were closely related to temperature, vehicle speed, and other wind factors. In addition, gas leakages

were not related to the location factors for this dataset. Finding a better model predicts medium and high levels for an unbalanced experimental dataset. According to the test results, the proposed F-OE-XGBoost algorithm has accuracy, F1-score, MSE, and AUC outcomes of 95.14%, 95.75%, 0.028, and 96.29%, respectively. Following this, the F-OE-RF model achieved the second-best accuracy rate of 95.09%, F1-score of 95.60%, MSE of 0.029, and AUC of 96.11%. The results here demonstrate that the proposed method is suitable for the early prediction of NG losses in the air. The system was implemented using SPSS and Python and tested its performance on actual open data.

Author Contributions: K.D. conceived and designed the experiments; K.D. performed the experiments; K.D. analyzed the data and discussed the results; K.D. wrote the paper; M.-H.S. supervised, checked, gave comments, and approved this work. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Ministry of Public Safety and Security. *2019th Yearbook of Disaster, Ministry of Public Safety and Security*; Ministry of Public Safety and Security: Sejong, Korea, 2019.
2. Department for International Development. Live Data Page for Energy and Water Consumption. Available online: http://data.gov.uk/dataset/dfid-energy-and-water-consumption (accessed on 8 March 2021).
3. Kim, Y.K.; Sohn, H.G. Disasters from 1948 to 2015 in Korea and power-law distribution. In *Disaster Risk Management in the Republic of Korea*; Springer: Singapore, 2017; pp. 77–97. [CrossRef]
4. Deichmann, J.L.; Hernández-Serna, A.; Campos-Cerqueira, M.; Aide, T.M. Soundscape analysis and acoustic monitoring document impacts of natural gas exploration on biodiversity in a tropical forest. *Ecol. Indic.* **2017**, *74*, 39–48. [CrossRef]
5. Zadkarami, M.; Safavi, A.A.; Taheri, M.; Salimi, F.F. Data driven leakage diagnosis for oil pipelines: An integrated approach of factor analysis and deep neural network classifier. *Trans. Inst. Meas. Control* **2020**, *42*, 2708–2718. [CrossRef]
6. USDT. *Leak Detection Technology Study for PIPES Act*; Tech. Rep.; U.S. Department of Transportation: Washington, DC, USA, 2007.
7. Bryce, P.; Jax, P.; Fang, J. Leak-detection system designed to catch slow leaks in offshore Alaska line. *Oil Gas J.* **2002**, *100*, 53–59.
8. Behari, N.; Sheriff, M.Z.; Rahman, M.A.; Nounou, M.; Hassan, I.; Nounou, H. Chronic leak detection for single and multiphase flow: A critical review on onshore and offshore subsea and arctic conditions. *J. Nat. Gas Sci. Eng.* **2020**, *11*, 103460. [CrossRef]
9. Weller, Z.D.; Yang, D.K.; Fischer, J.C. An open-source algorithm to detect natural gas leaks from mobile methane survey data. *PLoS ONE* **2019**, *14*, e0212287.
10. Dashdondov, K.; Kim, M.H. Mahalanobis Distance Based Multivariate Outlier Detection to Improve Performance of Hypertension Prediction. *Neural Process. Lett.* **2021**, *2*, 1–3. [CrossRef]
11. von Fischer, J.C.; Cooley, D.; Chamberlain, S.; Gaylord, A.; Griebenow, C.J.; Hamburg, S.P.; Salo, J.; Schumacher, R.; Theobald, D.; Ham, J. Rapid vehicle-based identification of location and magnitude of urban natural gas pipeline leaks. *Environ. Sci. Technol.* **2017**, *51*, 4091–4099. [CrossRef] [PubMed]
12. Zachary, D.W.; Duck, K.Y.; von Joseph, C.F. *Instruction for Processing Mobile Methane Survey Data to Detect Natural Gas Leaks*; Colorado State University: Fort Collins, CO, USA, 2018. Available online: https://github.com/JVF-CSU/MobileMethaneSurveys/tree/master/Scripts/SampleRawData (accessed on 10 October 2018).
13. Khongorzul, D.; Kim, M.H.; Lee, S.M. OrdinalEncoder based DNN for natural gas leak prediction. *J. Korea Converg. Soc.* **2019**, *10*, 7–13.
14. Xue, P.; Jiang, Y.; Zhou, Z.; Chen, X.; Fang, X.; Liu, J. Machine learning-based leakage fault detection for district heating networks. *Energy Build.* **2020**, *223*, 110161. [CrossRef]
15. Xu, Y.; Zhao, X.; Chen, Y.; Yang, Z. Research on a mixed gas classification algorithm based on extreme random tree. *Appl. Sci.* **2019**, *9*, 1728. [CrossRef]
16. Lei, Y.; Jiang, W.; Jiang, A.; Zhu, Y.; Niu, H.; Zhang, S. Fault diagnosis method for hydraulic directional valves integrating PCA and XGBoost. *Processes* **2019**, *7*, 589. [CrossRef]
17. Janizadeh, S.; Vafakhah, M.; Kapelan, Z.; Mobarghaee Dinan, N. Hybrid XGboost model with various Bayesian hyperparameter optimization algorithms for flood hazard susceptibility modeling. *Geocarto Int.* **2021**, *21*, 1–20. [CrossRef]
18. Zhu, S.B.; Li, Z.L.; Zhang, S.M.; Liang, L.L.; Zhang, H.F. Natural gas pipeline valve leakage rate estimation via factor and cluster analysis of acoustic emissions. *Measurement* **2018**, *125*, 48–55. [CrossRef]

19. Shamshirband, S.; Hadipoor, M.; Baghban, A.; Mosavi, A.; Bukor, J.; Várkonyi-Kóczy, A.R. Developing an ANFIS-PSO model to predict mercury emissions in combustion flue gases. *Mathematics* **2019**, *7*, 965. [CrossRef]
20. Fagiani, M.; Squartini, S.; Gabrielli, L.; Severini, M.; Piazza, F. A statistical framework for automatic leakage detection in smart water and gas grids. *Energies* **2016**, *9*, 665. [CrossRef]
21. Makonin, S.; Popowich, F.; Bartram, L.; Gill, B.; Bajic, I.V. AMPds: A public dataset for load disaggregation and eco-feedback research. In Proceedings of the 2013 IEEE Electrical Power Energy Conference, Halifax, NS, Canada, 21–23 August 2013; pp. 1–6.
22. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
23. Zhang, M.L.; Pena, J.M.; Robles, V. Feature selection for multi-label naive Bayes classification. *Inf. Sci.* **2009**, *179*, 3218–3229. [CrossRef]
24. Khongorzul, D.; Lee, S.M.; Kim, Y.K.; Kim, M.H. Image denoising methods based on DAECNN for medication prescriptions. *J. Korea Converg. Soc.* **2019**, *10*, 17–26. [CrossRef]
25. Hemmati-Sarapardeh, A.; Hajirezaie, S.; Soltanian, M.R.; Mosavi, A.; Nabipour, N.; Shamshirband, S.; Chau, K.W. Modeling natural gas compressibility factor using a hybrid group method of data handling. *Eng. Appl. Comput. Fluid Mech.* **2020**, *14*, 27–37. [CrossRef]
26. Mohammadi, M.R.; Hadavimoghaddam, F.; Pourmahdi, M.; Atashrouz, S.; Munir, M.T.; Hemmati-Sarapardeh, A.; Mosavi, A.H.; Mohaddespour, A. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. *Sci. Rep.* **2021**, *11*, 17911. [CrossRef]
27. Quy, T.B.; Kim, J.M. Real-time leak detection for a gas pipeline using a k-NN classifier and hybrid ae features. *Sensors* **2021**, *21*, 367. [CrossRef]
28. Quy, T.B.; Muhammad, S.; Kim, J.M. A reliable acoustic emission-based technique for the detection of a small leak in a pipeline system. *Energies* **2019**, *12*, 1472. [CrossRef]
29. Zhou, M.; Zhang, Q.; Liu, Y.; Sun, X.; Cai, Y.; Pan, H. An integration method using kernel principal component analysis and cascade support vector data description for pipeline leak detection with multiple operating modes. *Processes* **2019**, *7*, 648. [CrossRef]
30. Zhao, H.; Li, Z.; Zhu, S.; Yu, Y. Valve internal leakage rate quantification based on factor analysis and wavelet-BP neural network using acoustic emission. *Appl. Sci.* **2020**, *10*, 5544. [CrossRef]
31. Xie, J.; Xu, X.; Dubljevic, S. Long range pipeline leak detection and localization using discrete observer and support vector machine. *AIChE J.* **2019**, *65*, e16532. [CrossRef]
32. Melo, R.O.; Costa, M.G.; Costa, F.C.F. Applying convolutional neural networks to detect natural gas leaks in wellhead images. *IEEE Access* **2020**, *8*, 191775–191784. [CrossRef]
33. Rui, X.; Qunfang, H.; Jie, L. Leak detection of gas pipelines using acoustic signals based on wavelet transform and Support Vector Machine. *Measurement* **2019**, *146*, 479–489. [CrossRef]
34. Xie, Y.; Xiao, Y.; Liu, X.; Liu, G.; Jiang, W.; Qin, J. Time-frequency distribution map-based Convolutional Neural Network (CNN) model for underwater pipeline leakage detection using acoustic signals. *Sensors* **2020**, *20*, 5040. [CrossRef] [PubMed]
35. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
36. Nabipour, N.; Mosavi, A.; Baghban, A.; Shamshirband, S.; Felde, I. Extreme learning machine-based model for Solubility estimation of hydrocarbon gases in electrolyte solutions. *Processes* **2020**, *8*, 92. [CrossRef]