

Multivariate PCA-based Composite Criteria Evaluation Method for Anomaly Detection in Manufacturing Data

HyeokSoo Lee^{*,**}, Youngki Jo^{**}, Jongpil Jeong^{*}

^{*} Department of Smart Factory Convergence, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon, 16419, Korea

^{**} Research & Development Team, THIRA-UTECH Co., Ltd, CK Bldg 7, Hakdong-ro 5-gil Gangnam-gu, Seoul, 06044, Korea
 huxlee@g.skku.edu, ykjo0712@thirautech.com, jpjeong@skku.edu

Abstract— In recent years, manufacturing sites have become more intelligent and efficient by adopting various IT technologies. Among them, equipment and process abnormality detection is a topic of high interest for efficient factory operation. In this paper, we propose a method that can detect comprehensive abnormalities by utilizing the PCA algorithm, which is an unsupervised learning-based data analysis method that can easily analyze multivariate data and detect abnormalities in the data, the Hotelling T^2 method, which is suitable for multivariate data analysis, and the Box-Pierce statistical method to increase the detection criteria of abnormality detection data. To verify the effectiveness of the proposed method, experiments were conducted and validated using a chemical product production dataset. We expect that this method can be utilized for equipment and process anomaly detection in real-time at manufacturing sites.

Keywords— Multivariate Analysis, Anomaly Detection, Unsupervised Learning, Principal Component Analysis (PCA), Hotelling's T^2 Control Chart, Box-Pierce Test

I. INTRODUCTION

Along with the Fourth Industrial Revolution, major developed countries are increasingly emphasizing the importance of manufacturing, and the importance of smart factories that combine information and communication technologies in production and manufacturing is also increasing. The use of information and communication technology in the manufacturing industry began in earnest in the 80s. In particular, industries with high investment costs, such as the semiconductor and display industries, have automated their manufacturing sites by actively introducing information and communication technologies because production efficiency is significant. Currently, factories in these industries are mostly operated as unmanned automated factories, and all facilities and manufacturing operation systems are organically connected. Factory automation is a form of automating repetitive tasks according to certain rules by utilizing IT technology, and the smart factory, which has been emphasized recently, aims to transform into intelligence beyond such automation [1].

Recent advances in IT technologies such as artificial intelligence, sensors and the Internet of Things (IoT), big data, and robots have made it possible to build a flexible and intelligent manufacturing environment beyond the existing level of automation. By utilizing real-time data generated by various facilities, sensors, and the Internet of Things during the production process, it is possible to monitor the status of processes and facilities in real-time, detect abnormal conditions, and prepare for upcoming risks or problems before they occur. There are limitations to diagnosing and taking action on manufacturing equipment and process defects by relying on the experience of those in charge of the manufacturing site or related experts. Therefore, to increase manufacturing efficiency and utilization, it is necessary to proactively predict equipment failures and process defects using manufacturing data and AI technologies [2].

In this paper, we proposed a method for detecting and predicting abnormalities in facilities or processes by analyzing time series data generated by facilities at manufacturing sites in real time using unsupervised learning. There may be many variables in the data generated by facilities, and multivariate analysis methods are used to comprehensively analyze these many variables to predict abnormal situations. Therefore, we proposed a "Multivariate PCA-based Composite Criteria Evaluation Method" that can effectively predict abnormal situations by using the Principal Component Analysis (PCA) algorithm, multivariate analysis technique Hotelling T^2 control chart, and Box-Pierce statistical method, and verified it through experiments.

The paper is organized as follows:

In Section 2, we reviewed the techniques for multivariate analysis and their popularity in anomaly detection. We also reviewed the basic concepts of PCA algorithm, Hotelling T^2 control chart, and Box-Pierce statistics. In Section 3, the proposed method "Multivariate PCA-based Composite Criteria Evaluation Method" is defined and described. In Section 4, we conducted experiments on two experimental datasets and described the experimental results along with visualization results. The conclusions and future research directions are described in Section 5.

II. RELATED WORK

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. Multivariate Analysis

Multivariate analysis is a statistical method for simultaneously analyzing multiple variables on multiple measures and is an extension of univariate and bivariate analysis. It means performing multiple univariate analyses simultaneously, considering the dependent variables' statistical relationship [3].

There are the following types of multivariate analysis methods:

1) Multiple Regression: It is an analytical method that is appropriate when there is a relationship between one quantitative dependent variable and one or more quantitative independent variables. It allows you to predict changes in the dependent variable as a result of changes in multiple independent variables and to determine the extent to which the independent variables explain the dependent variable and the relative contribution of the independent variables to the dependent variable [4].

2) Multiple Discriminant Analysis (MDA): A method for classifying samples when each sample falls into two or more groups and the number of groups is known in advance. This method is appropriate when there are multiple categories and multiple groups based on the dependent variable [5].

3) Cluster Analysis: Cluster analysis is a method for finding a taxonomy for samples with no information about the population. Unlike discriminant analysis, cluster analysis does not have a predefined group. It checks the similarity between each sample and other samples and determines whether to classify them into a group [6].

4) Multivariate Analysis of Variance (MANOVA): It is a statistical method for determining the relationship between two or more categorical independent variables and multiple quantitative dependent variables and is an extension of univariate analysis of variance. Multivariate analysis of variance effectively tests hypotheses about each group's variance on two or more quantitative dependent variables [7].

5) Factor Analysis: Factor analysis is a statistical approach that analyzes the interrelationships among many variables to describe them in terms of a common underlying dimension. It is a way to compress the information contained in a large number of individual variables into a smaller set of variables with minimal loss of information. A typical factor analysis method is PCA [8, 17].

6) Canonical Correlation Analysis: Canonical correlation analysis is a logical extension of multiple regression analysis that investigates the relationship between one quantitative dependent variable and multiple quantitative independent variables. The basic principle is to find the linear combination of each variable group that maximizes the

correlation between the dependent variable and the independent variable group and to find the set of weights of each variable group that maximizes the correlation between the dependent variable and the independent variable group [9].

B. Anomaly Detection

Anomaly detection is the search for objects or materials that exhibit unexpected patterns in data, and it is a method of creating models based on training data to find data with different characteristics from existing data. It can be used in a variety of fields, including security, medicine, finance, and manufacturing, to predict malicious behavior such as credit card fraud, cyber intrusions, and terrorist acts, as well as abnormal situations such as system failures, equipment breakdowns, and defective products. Anomaly detection is often confused with the problem of classification. Classification is about finding the boundary that separates two categories, while anomaly detection is about finding outlier data given multiple categories [10, 11].

For anomaly detection, there are the following methods:

1) Statistical Methods: Statistical methods are based on the assumption that outliers appear in the lower regions of the probability distribution. They fit a normal data model to the given data and use statistical inference to determine if new objects follow that model. The distribution you use depends on the complexity of your data. If your data is simple, you can use simple distributions like Normal, Poisson, or Multinomial. However, more complex models, such as mixed models or Hidden Markov Models, require iterative computations to estimate, which can take a long time depending on the convergence rate and criteria [10].

2) Classification-based Methods: Classification-based methods train data into one class or multiple classes depending on the number of labels and treat objects that do not correspond to a class as outliers. Representative classification-based method algorithms are Autoencoder, Bayesian network, Support Vector Machine (SVM) [12], and Decision Trees [13]. The Autoencoder algorithm has an input layer and an output layer with the same number of nodes. It can also have one or more hidden layers and uses an encoder to compress the input data and a decoder to restore the input data. If the error is large, it is judged as an abnormal situation [10].

3) Clustering-based Methods: Clustering-based methods can detect outliers with the following principles. The first is that outliers are those with a long distance from the center of the cluster. There are methods such as the Self-Organizing Map (SOM) [14] and k-Means [15] algorithm that cluster training data and compare test data with clusters to determine outliers. The second method is to determine whether an object is an anomaly by the size or density of the cluster to which it belongs. It is based on the assumption that normal belongs to large or dense clusters, while anomalies belong to small or sparse clusters [10].

4) Nearest Neighbor-based Methods: Nearest neighbor analysis leverages the basic idea that normal data occurs in dense groups of data, while anomalies occur at a distance. Nearest neighbor anomaly detection methods need to measure the distance or similarity between two pieces of data. The distance or similarity between two pieces of data is calculated differently for continuous and categorical data. For continuous data, we usually use Euclidean distance, and for categorical data, we use a simple matching coefficient. Also, for multivariate data, we need to calculate the value of each attribute and then combine them. A typical method for nearest neighbor analysis is the k-nearest Neighbor algorithm [10, 16].

5) Spectral-based Methods: Spectral-based methods are based on the assumption that data can be sent to a lower-dimensional subspace, where the normal and abnormal are sharply separated. A typical method is Principal Component Analysis [17]. When the data is divided into a low-dimensional space, normal objects that satisfy the correlation structure of the data have low values, and anomalous objects that deviate from the correlation structure have high values. It is suitable for processing high-dimensional data and can be used as a preprocessing for applying other techniques. However, it is only useful if the normal and outliers are properly separated in the low-dimensional space to which you send the data [10].

C. Principal Component Analysis (PCA)

Principal Component Analysis was first invented by Carl Pearson in 1901, but in the 1930s, Harold Hotelling, who was unaware of its existence, developed it separately and named it Principal Component Analysis [17]. PCA is a technique for transforming high-dimensional data into low-dimensional data, and it uses orthogonal transformations to transform data in high-dimensional space into low-dimensional space. For example, in the case of data with more than two dimensions, when the data is scaled on a single axis, the axis with the largest variance is called the first principal component, and the axis with the second largest variance is called the second principal component. Furthermore, the principal components after the first principal component are defined as having the largest variance under the constraint that they are orthogonal to the previous principal components. The important components are orthogonal because they are the eigenvectors of the covariance matrix. Principal component analysis is the simplest of the eigenvector-based multivariate analyses, and if a multivariate dataset appears to be simply a set of coordinates, it can be reduced to a lower dimension so that it can be analyzed effectively [17, 18].

If the data contains many variables, the dimensionality of the data may be large, resulting in slow learning and analysis speeds and inefficient analysis results. In this case, manifold learning reduces the dimensionality of the data for efficient learning and analysis. A manifold is a space where data exists, and data can exist in various dimensions. By reducing the dimensionality given by the data, you can view the data in a smaller dimension, requiring fewer parameters and making

learning and analysis more efficient. However, on the learning side, while reducing dimensionality speeds up learning, it does not necessarily mean that reducing dimensionality improves the performance of the model. Therefore, it is necessary to determine the appropriate dimensionality by checking the results of the analysis. The most representative dimensionality reduction algorithm for manifold learning is PCA [17, 18].

The PCA algorithm proceeds in the following steps: [17, 18]

- Normalization is performed to center the axis without changing the data. Normalization scales all variables so that their mean is zero.
- Find the covariance matrix for the normalized data.
- Find the eigenvalues and eigenvectors for the calculated covariance matrix.
- The obtained eigenvalues are sorted in descending order to find the largest values.
- The eigenvector values corresponding to these large eigenvalues become the coefficient values that make up the principal components.
- Check the ratio of the Eigenvalues, determine which principal components are significant, and reduce the dimensionality by excluding the remaining principal components.

D. Hotelling's T^2 Control Chart

Hotelling's T^2 test, proposed by Harold Hotelling in 1947, is an extension of the t-test and a typical multivariate test method and assumes that the data is normally distributed and follows the F-distribution [19, 20, 21]. If data has p quality attributes and the number of observations is n, then the data has a $p \times n$ dimensional matrix structure. In this case, the statistics for individual observations are described by equation (1). [19, 20, 21]

$$T^2 = (X_i - \bar{X})^T S^{-1} (X_i - \bar{X}) \quad (1)$$

The T^2 statistic takes into account the correlation between the independent variables by using the inverse of the covariance. Under the assumption that the data is normally distributed, the statistic of the Hotelling T^2 control chart follows the F distribution. This is equivalent to equation (2). [19]

$$T^2 \sim \frac{(n-1)p}{n-p} F_{(p, n-p)} \quad (2)$$

For a significance level α , the Upper Control Limit (UCL) value can be defined in equation (3).

$$UCL = \frac{(n-1)p}{n-p} F_{\alpha}(p, n-p) \quad (3)$$

Hotelling's T^2 control chart assumes that normal data follows a normal distribution and uses a covariance matrix, and if Hotelling's T^2 statistic is higher than the control limit, it

can also detect abnormal conditions. In particular, multivariate statistical methods are effective when multiple variables are highly correlated with each other. Therefore, multivariate statistical methods are a useful way to use multiple variables together to monitor processes more accurately and detect process abnormalities [19, 20, 21].

E. Box-Pierce (BP) Test

The Box-Pierce test was published in 1970 by G. E. P. Box and David A. Pierce [22]. It can be used in time series analysis to determine the existence of autocorrelation, and is a method for testing the null hypothesis "the data are independently distributed" and the alternative hypothesis "the data are not independently distributed and are time series correlated"[22, 23]. The Box-Pearce test is defined as the square of the covariance coefficient ρ for a given data set multiplied by time h and multiplied by the number of data n . The relevant equation is shown in equation (4) and (5). The Box-Pierce test follows a chi-square distribution [22, 23].

$$\hat{\rho} = X_t - X_{t-1} \quad (4)$$

$$Q = n \sum_{k=1}^h \hat{\rho}_k^2 \quad (5)$$

The Box-Pierce Q statistic is approximated from a chi-square distribution to an F distribution using the actual and predicted values. The value of the Box-Pierce Q statistic and the UCL value are defined by equations (6) and (7) below.

$$BP \sim \frac{(n^2 - 1)p}{n(n - p)} F(p, n - p) \quad (6)$$

$$UCL = \frac{(n^2 - 1)p}{n(n - p)} F_{\alpha}(p, n - p) \quad (7)$$

III. PROPOSED METHOD

This paper proposes a method to analyze unlabeled data to detect and predict anomalies. To do this, we need a method based on unsupervised learning and a method of multivariate data analysis. In the case of univariate data analysis, there is a problem that as the number of variables increases, the number of control charts also increases, and when detecting anomalies for each variable, the frequency is often too frequent to be significant. Therefore, a multivariate data analysis method is needed to improve efficiency by managing multiple variables with one control chart regardless of the number of variables and to detect abnormal situations that cannot be detected by univariate data analysis [3, 10, 11].

We reviewed methods for anomaly detection and multivariate data analysis in the previous section. In this paper, we first utilize PCA, which is commonly used for anomaly detection and multivariate data analysis, as the basic algorithm

[18, 26, 27]. Based on the eigenvalues and eigenvector values of the covariance matrix obtained from the PCA algorithm calculation, multivariate analysis is performed. For this purpose, the Hotelling T^2 control method, which detects anomalies by considering the correlation between quality attribute values [24, 25], and the Box-Pierce Q statistical method, which checks for autocorrelation using the residuals between predicted and actual values in time series data, are used together. This method is defined as a "Multivariate PCA based Composite Criteria Evaluation Method". Figure 1 shows the detailed steps of the "Multivariate PCA-based Composite Criteria Evaluation Method".

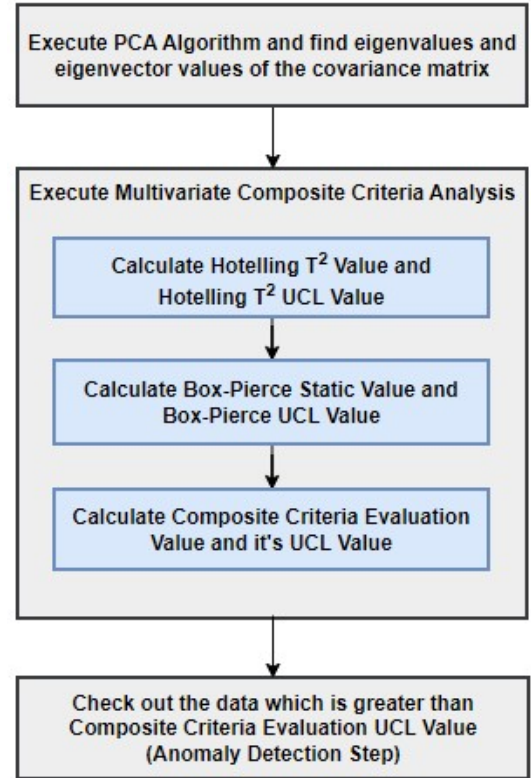


Figure 1. Multivariate PCA-based Composite Criteria Evaluation Method

We want to calculate the Composite Criteria Evaluation Value UCL values using a chi-square distribution, and since Hotelling T^2 UCL and Box-Pierce UCL are F-distributed, we approximate them with a chi-square distribution. The relevant equations can be found in (8) and (9).

$$T^2 \quad UCL = \frac{(n - 1)p}{n - p} F_{\alpha}(p, n - p) \sim \chi_{\alpha}^2(p) \quad (8)$$

$$BP \quad UCL = \frac{(n^2 - 1)p}{n(n - p)} F_{\alpha}(p, n - p) \sim \chi_{\alpha}^2(p) \quad (9)$$

Equation (10) explains how the Composite Criteria Evaluation Value and corresponding UCL values are obtained. To simplify the formulas, we will denote the Hotelling T^2 statistic as α_1 , the Box-Pierce statistic as α_2 , the Hotelling T^2

UCL value as β_1 , and the Box-Pierce UCL value as β_2 . For the Composite Criteria Evaluation UCL, the Hotelling T2 UCL value and Box-Pierce UCL value obtained earlier are combined.

$$\text{Composite Criteria Evaluation Value} = \frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2} \quad (10)$$

IV. EXPERIMENT AND RESULTS

The experiment was conducted using data from a chemical manufacturing process. A total of ten datasets were used for the experiment, and the labeled dataset was used to check the results of the analysis. In the case of unsupervised learning, the evaluation and adequacy of experimental results are limited and difficult, so we decided to use a labeled dataset to verify the analysis results. When learning of experiment, we removed the labels from the dataset conducted the experiment, and compared the experimental results with the label values in the dataset to check the accuracy.

The experiment was verified by first visualizing the raw data with the labels removed. The deviation of data values between data variables was too large to analyze the data immediately. Therefore, the data sets were scaled and normalized, and the normalized data was visualized to check the distribution. Then, we conducted experiments with the proposed method and checked the Hotelling T² statistic, Box-Pierce statistic, and related UCL values. The final Composite Criteria Evaluation Value and UCL values were then checked, and the accuracy of the analysis results was confirmed by comparing the obtained values with the label values. Figure 2 describes the analysis steps to run the experiment.

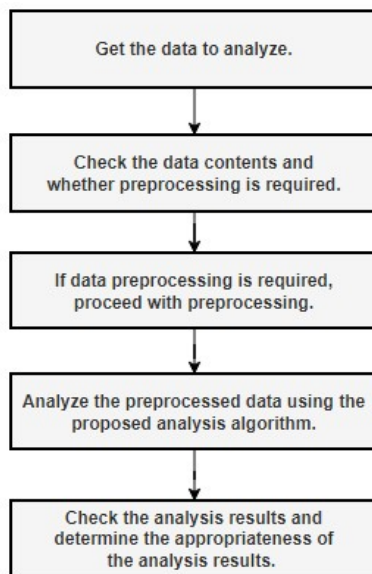


Figure 2. Processes for analyzing data

The experiment was performed with a total of 10 datasets. Each dataset consisted of 165 observations and included two abnormalities. Even though the datasets have the same number of data and the same number of abnormalities, they are all

independent and different. In the experiment, the measurement results (labeled values) that can identify abnormalities were removed from the dataset and the experiment was conducted with unsupervised learning after the experiment, the experimental results were compared with the measurement results (labeled values) to check the accuracy of the analysis prediction. The following Table 1 summarizes the experimental results.

TABLE 1. SUMMARY OF EXPERIMENTAL RESULTS

Data Set #	Number of data	Number of anomalies in the measurement	Number of anomaly predictions
1	165	2	2
2	165	2	1
3	165	2	2
4	165	2	1
5	165	2	2
6	165	2	2
7	165	2	2
8	165	2	2
9	165	2	2
10	165	2	3
Total	1,650	20	19

In the dataset perspective, 7 out of 10 predictions were correct, which is about 70% accuracy, but in the case of the data, 3 out of 1,650 predictions were incorrect, which is 99.82% accuracy. The following describes the detailed analysis procedure and results for two datasets that were correctly predicted

1) Experiment #1: The first data is a time series dataset with 7 variables and 165 measurements. The label values include 2 bad results. The raw data can be visualized in Figure 3 and the distribution of data is shown in Figure 4.

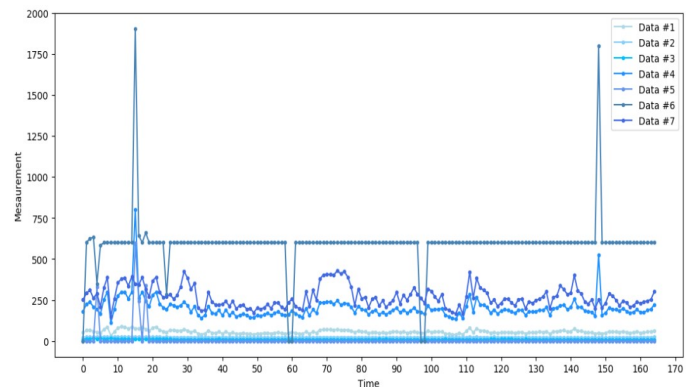


Figure 3. Raw data visualization results

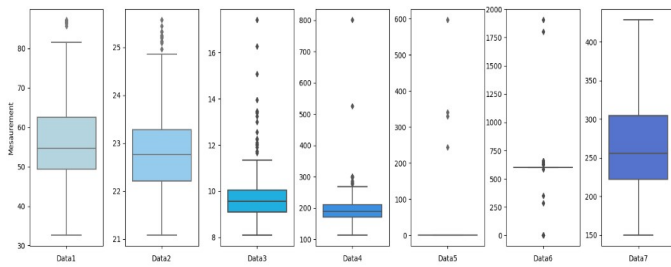


Figure 4. Distribution of raw data

When we visualized the raw data and its distribution, we found that the deviation between the data was considerable, so we normalized the data for efficient data analysis. Figures 5 and 6 show the distribution of the data after data normalization.

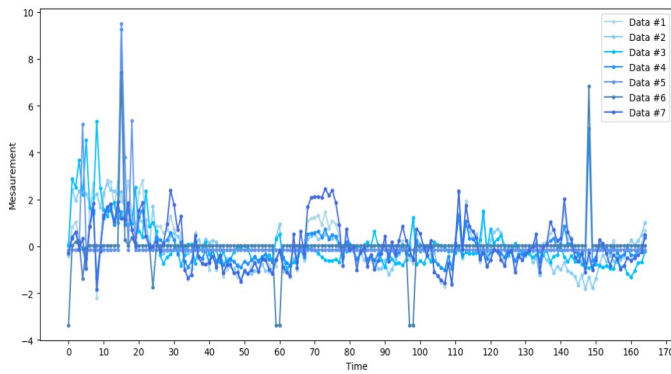


Figure 5. Normalized data visualization results

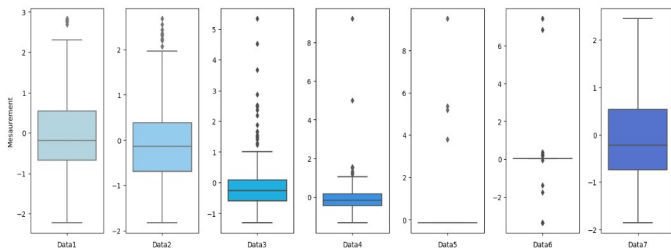


Figure 6. Distribution of normalized data

The proposed method, "Multivariate PCA-based Composite Criteria Evaluation Method", is executed to get Hotelling T^2 statistical value, Hotelling T^2 UCL value, Box-Pierce statistical value, and Box-Pierce UCL value. Figure 4 shows the results for these values.

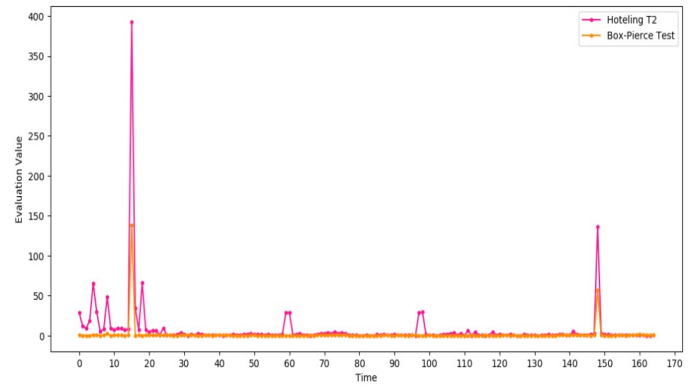


Figure 7. Hotelling T^2 and Box-Pierce statistics visualization results

Utilizing the Hotelling T^2 statistical value, Hotelling T^2 UCL value, Box-Pierce statistical value, and Box-Pierce UCL value values obtained earlier, we finally get the Composite Criteria Evaluation result and UCL value. Figure 5 shows the Composite Criteria Evaluation result and UCL value. The UCL value 15.2496315 is indicated by the red dotted line, and a total of two anomalies are predicted that exceed the UCL value.

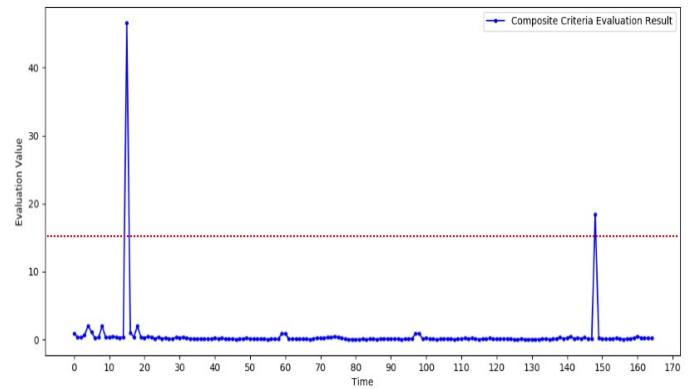


Figure 8. Composite Criteria Evaluation visualization results

The labeled values in the dataset are the results of post-production inspections, with most results between 10.00 and 11.00, and the average of the normal results in the dataset is 10.3933. The 16th and 149th inspection results in the dataset exceed this average criterion with values of 15.8655 and 15.8805, respectively, and can be categorized as abnormal. By checking the Composite Criteria Evaluation results analyzed by unsupervised learning with the proposed method, we can see that the two cases of exceeding the threshold UCL coincide with the points of abnormal conditions mentioned above.

2) Experiment #2: The second data is a time series dataset with 12 variables and 165 measurements. This dataset also has a total of 2 bad results in its labels. The raw data can be visualized in Figure 9 and the distribution of data is shown in Figure 10.

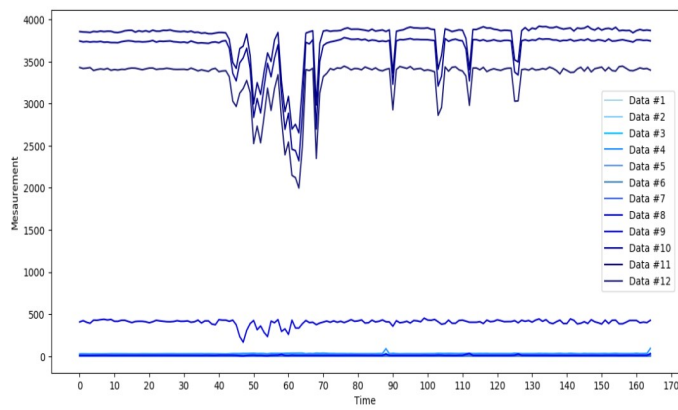


Figure 9. Raw data visualization results

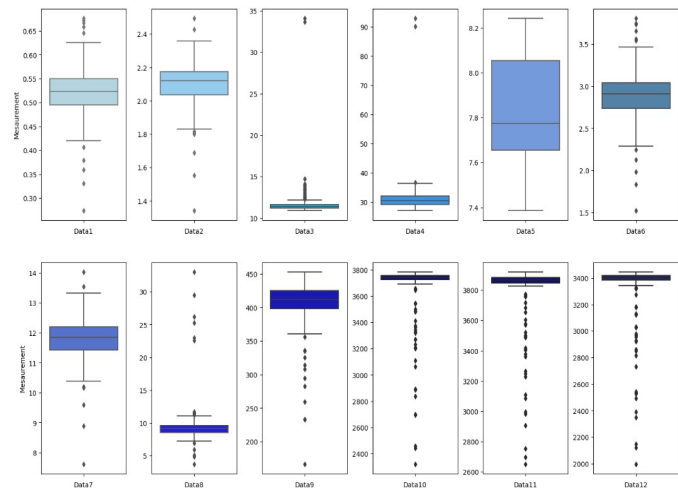


Figure 10. Distribution of raw data

When we checked the raw data, we found that this dataset also had a very large variance between the data, so we needed to normalize the data for efficient data analysis. Figures 11 and 12 show the distribution of the data after data normalization.

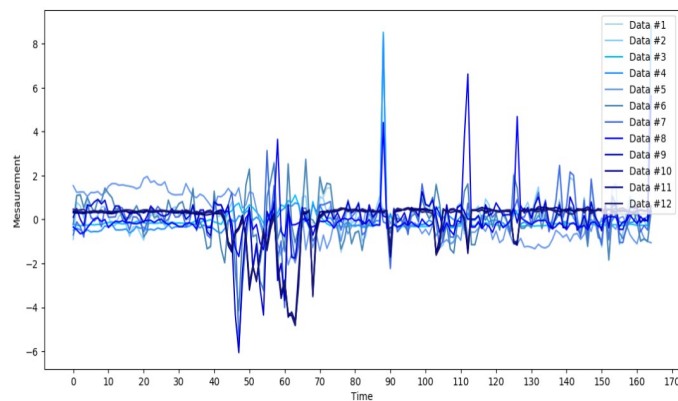


Figure 11. Normalized data visualization results

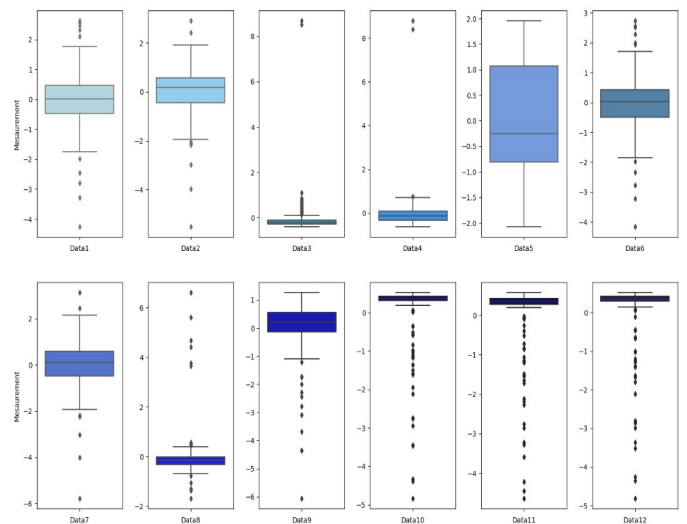
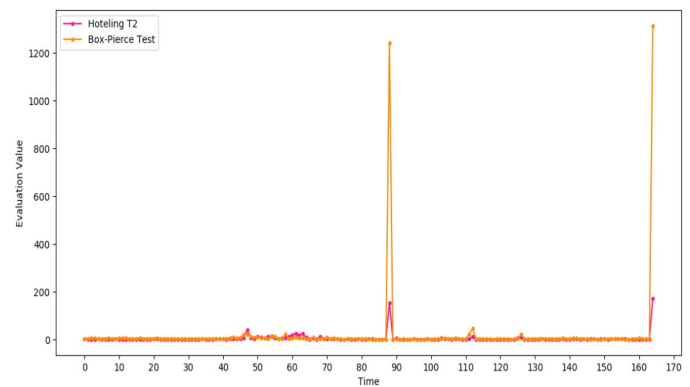


Figure 12. Distribution of normalized data

The proposed method is run and got Hotelling T^2 statistical value, Hotelling T^2 UCL value, Box-Pierce statistical value, and Box-Pierce UCL value. Figure 13 shows what they look like.

Figure 13. Hotelling T^2 and Box-Pierce statistics visualization results

Based on the Hotelling T^2 statistical value, Hotelling T^2 UCL value, Box-Pierce statistical value, and Box-Pierce UCL value values obtained earlier, we finally calculate the Composite Criteria Evaluation result and UCL value. Figure 14 shows their values. The UCL value 30.8467355 is marked with the red dotted line, and there are two anomalies are predicted.

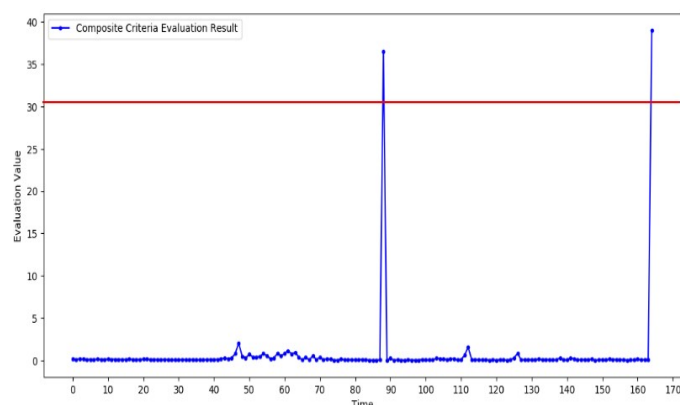


Figure 14. Composite Criteria Evaluation visualization results

The labeled values in the dataset are the results of post-production inspections, with most results between 10.00 and 11.00, and the average of the normal results in the dataset is 10.3932. The 89th and 165h inspection results in the dataset exceed this average criterion with values of 15.834 and 15.819, respectively, and can be categorized as abnormal. By checking the Composite Criteria Evaluation results analyzed by unsupervised learning with the proposed method, we can see that the two cases of exceeding the threshold UCL coincide with the points of abnormal conditions mentioned above.

V. CONCLUSIONS

The topic of this paper is research on how to detect abnormal situations by analyzing data generated in the manufacturing production process based on unsupervised learning. Most of these data are in the form of multivariate data with many variables. During machine learning, PCA, which is easy to analyze multivariate data and can detect abnormal situations in the data, is used for initial data analysis, and based on the analyzed results, the Hotelling T2 method, which is suitable for multivariate data analysis, and Box-Pierce statistical method, which is used to prevent abnormal situations from being detected too frequently in various variables, are used together to derive comprehensive abnormality detection results, and "Multivariate PCA based Composite Criteria Evaluation Method" is proposed. An experiment was conducted to verify the effectiveness of the proposed method. Experiments were conducted and verified using 10 datasets from a manufacturing plant that produces chemical products. In the case of the proposed algorithm, it is difficult to verify the results against the experimental results because it is in the form of unsupervised learning. So, we used datasets with measurement results that classified abnormal and normal states, removed the results, tested with datasets with only variables, and compared the experimental results with the measurement results of the dataset. The experimental results confirmed that 7 out of 10 datasets match exactly. In the case of mismatched datasets, one dataset detected only one of the two anomalies, and the other dataset detected anomalies correctly, but there was misrecognized a normal situation as an anomaly once. In terms of datasets, the accuracy was 70%, and in terms of data, the accuracy was about 99.82%. This

method can be utilized for real-time abnormality detection of facilities and processes in the field. The detected anomalies can be shared with the manufacturing operation system, and the person in charge can check the equipment status or process variables and take early action if there is a problem. In the future, we would like to study how to increase the efficiency of manufacturing operations by more closely linking to the functions of the manufacturing operation system, how to separate the exact type of anomaly detection into equipment anomaly detection or process anomaly detection, and how to expand parameter management by linking anomaly detection with process or equipment parameters.

REFERENCES

- [1] R. Agnieszka, B. Arne, B. Marcel and S.M. Erik, "The Smart Factory: Exploring Adaptive and Flexible Manufacturing Solutions", *Procedia Engineering*, vol. 69, 2014.
- [2] Y. Shen, R.J. Juan and J. Yuchen, "Real-Time Monitoring and Control of Industrial Cyberphysical Systems: With Integrated Plant-Wide Monitoring and Control Framework", *IEEE Industrial Electronics Magazine*, vol. 13, Dec, 2019.
- [3] P. Luke and G. Hessameddin, *Multivariate Data Analysis*, 8th ed, Cengage Learning EMEA, 2018.
- [4] M. Tucker and M. Brian, "Multiple Regression in L2 Research: A Methodological Synthesis and Guide to Interpreting R2 Values", *The Modern Language Journal*, vol. 102, p 713-731, Dec, 2018.
- [5] C. E. Brown, *Applied Multivariate Statistics in Geohydrology and Related Sciences*, 1st ed, Heidelberg: Springer-Verlag, Dec, 2011.
- [6] J. R. Kettenring, "The Practice of Cluster Analysis", *Journal of Classification*, vol. 23, Dec, 2006.
- [7] F. L. Huang, "MANOVA: A Procedure Whose Time Has Passed?", *SAGE Journals*, vol. 64, Dec, 2019.
- [8] T. Mohsen and W. Angela, "Factor Analysis: a means for theory and instrument development in support of construct validity", *International Journal of Medical Education*, vol. 11, Jun, 2020.
- [9] Guangdong Liu, Shanshan Yang, Wei Liu, Shengshu Wang, Penggang Tai, Fuyin Kou, Wangping Jia, Ke Han, Miao Liu and Yao He, "Canonical Correlation Analysis on the Association Between Sleep Quality and Nutritional Status Among Centenarians in Hainan", *Frontiers in Public Health*, vol. 8, Nov, 2020.
- [10] Chandola. V, Banerjee and A. Kumar, "Anomaly Detection: A Survey", *ACM Computing Surveys*, vol. 41, pp. 1-58, Nov. 2009.
- [11] L. M. Ghinea, M. Miron and M. Barbu, "Semi-Supervised Anomaly Detection of Dissolved Oxygen Sensor in Wastewater Treatment Plants", *MDPI Sensors*, vol. 23, Sep. 2023.
- [12] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines", *IEEE Intelligent Systems and their Applications*, vol. 13, Jul-Aug, 1998.
- [13] M. Oded and R. Lior, *Data Mining and Knowledge Discovery Handbook*, 1st ed, New York, US: Springer-Verlag, Dec, 2005.
- [14] T. Kohonen, "The self-organizing map", *Proceedings of the IEEE*, vol. 78, Sep, 1990.
- [15] Claude Sammut, Geoffrey I. Webb, *Encyclopedia of Machine Learning*, 1st ed, New York, US: Springer, Nov, 2010.
- [16] Oliver Kramer, *Dimensionality Reduction with Unsupervised Nearest Neighbors*, 1st ed, Heidelberg: Springer-Verlag, Nov, 2013.
- [17] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed, New York, US: Springer-Verlag, 2002.
- [18] L.E. Mujica, J. Rodellar, A. Ferná'ndez and A. Gu'emes, "Q-statistic and T2-statistic PCA-based measures for damage assessment in structures", *SAGE Journals*, vol. 10, Nov. 2010.
- [19] National Institute of Standards and Technology (NIST) Engineer Statistics Handbook. [Online]. Available: <https://www.itl.nist.gov/div898/handbook/pmc/section5/pmc543.htm/>
- [20] R. E. Schumacker, "Using R with Multivariate Statistics", SAGE Publishing, 2015.

- [21] M. R. Piña-Monarez, "Practical decomposition method for T² hotelling chart", *The International Journal of Industrial Engineering*, vol. 20, pp. 401–411, Nov. 2013.
- [22] M. Tucker and M. Brian, "The multiple testing problem for Box-Pierce statistics", *Electronic Journal of Statistics*, vol. 8, pp. 497–522, Nov. 2013.
- [23] WIKIPEDIA Ljung–Box test. [Online]. Available: https://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test/
- [24] T.Y. Heo, H.B. Jeon, S.M. Park, Y.J. Lee, "Development of Real-Time Water Quality Abnormality Warning System for Using Multivariate Statistical Method", *Journal of Korean Soc. Environ. Eng.*, vol. 37, pp. 137–144, Feb. 2015.
- [25] J.H. Kim, S.B. Kim, "Local T² Control Charts for Process Control in Local Structure and Abnormal Distribution Data", *Journal of Korean Society for Quality Management*, vol. 40, pp. 337–346, Aug. 2012.
- [26] A. Liang, Y. Hu, G. Li, "The impact of improved PCA method based on anomaly detection on chiller sensor fault detection", *International Journal of Refrigeration*, Sep. 2023.
- [27] T.H.K. Mohammed, K. Abdelmalek, F.H. Mohamed, B. Abderazak, M. Majdi, "Improving Kernel PCA-based algorithm for fault detection in nonlinear industrial process through fractal dimension", *Process Safety and Environmental Protection*, vol. 179, pp. 525–536, Sep. 2023.

HyeokSoo Lee received a M.S. degree in Computer Science from Fairleigh Dickinson University, New Jersey US, in 1995. He is currently working toward a Ph.D. degree in smart factory convergence

at Sungkyunkwan University, Suwon Korea. He is currently an employee as a head of the Research & Development team of THiRA-UTECH Co, Ltd. His research interests are mainly in Smart Factory, Reinforcement Learning, Deep Learning, Robotics, and Data Analysis.

Youngki Jo received the B.S. and M.S. degree in Industrial Engineering from Kumoh National Institute of Technology, Korea in 2018, and 2020. He is currently an employee as a research engineer of the Research & Development team of THiRA-UTECH Co, Ltd. His research interests include generative adversarial networks, deep learning, and time-series modeling and forecasting.

Jongpil Jeong received the bachelor's degree in engineering from Sungkyunkwan University, Suwon, South Korea, and the master's and Ph.D. degrees in computer engineering from Sungkyunkwan University, in 2003 and 2008, respectively. He has been with Sungkyunkwan University, since 2008, and has been an Associate Professor with the Department of Smart Factory Convergence, Sungkyunkwan University, since 2016. He is the Principal Investigator of MAKE UNIC, a key support area for smart manufacturing with Sungkyunkwan University. His research interests include smart factory, industrial AI, anomaly detection, manufacturing data analysis, AI-based fault diagnosis and prediction, 5G-based smart manufacturing, industrial IoT applications, AI platforms, cloud platforms, and industrial security.