IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 April 2024, accepted 15 May 2024, date of publication 21 May 2024, date of current version 31 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3403864

RESEARCH ARTICLE

Depth-Based Condition Monitoring and Contributing Factor Analysis for Anomalies in Combined Cycle Power Plant

MUNWON LIM[1], YUNHEE KIM[1], SHENGXIN JIN[1], SANGWOO HA[1], SUNG YONG CHANG[2], HAE SU KANG[2], GYU SANG PARK[2], MI LEE JOO[2], CHANG-SIK JUNG[3], YOUNGJIN CHO[3], AND SUK JOO BAE[1], (Member, IEEE)

[1]Department of Industrial Engineering, Hanyang University, Seoul 04763, South Korea
[2]Korea Electric Power Research Institute, Daejeon 34056, South Korea
[3]Rozetatech, Seongnam 13440, South Korea

Corresponding author: Suk Joo Bae (sjbae@hanyang.ac.kr)

This work was supported in part by Korea Institute of Energy Technology Evaluation and Planning (KETEP) Grant funded by Korea Government (MOTIE) (Development of Intelligent O&M Technologies for Standard Combined Cycle Power Plant and Development of Visualized Fire Protection System to Thermal Power Plant Through IIoT and Digital Twin Technology) under Grant 20217010100020 and Grant 1415186560.

ABSTRACT Unexpected fault or failure in the power plant have caused high maintenance costs, the loss of energy production, and even safety issues. Developments in sensor technologies and data analytics have aided proper preventive maintenance actions for the system to improve asset availability and reduce repair costs. Nevertheless, effective condition monitoring of a power plant experiences a considerable nuisance from challenging issues such as inherent data characteristics such as high correlations between process variables, irrelevant information from environmental noises, and system complexity. To resolve these problems, this paper proposes an integrated monitoring scheme for performing efficient corrective actions by identifying the variables related to anomalies in combined cycle power plants. The scheme includes a clustering-based linear discriminant analysis to extract key variables for reducing dimensionality to efficiently handle the data, followed by employing the Mahalanobis depth statistics for anomaly detection and causal analysis via contribution scores. The proposed monitoring scheme is applied to condition monitoring data of a combined cycle power plant in South Korea, which include two types of anomalous operations. The reliability and robustness of the proposed condition monitoring scheme are validated by comparing other state-of-the-art methods. The proposed method shows a potential in efficiently detecting anomalies during operation and even early detecting the precursors of anomalies. It is expected to prevent imminent faults or failures by taking proper actions to relevant key process parameters of combined cycle power plant in advance.

INDEX TERMS Anomaly detection, causal inference, condition-based maintenance, contribution scores, health monitoring.

I. INTRODUCTION

Power plants are complex and expensive assets consisting of various components with different functionalities. In such a large-scale system, anomalies of a subsystem or a component may lead to unexpected malfunctions or even breakdown of an equipment, which significantly affects the performance of a power plant. If the equipment is not operated due to anomalies and not recovered quickly through maintenance actions, undesirable results such as high maintenance costs, the loss of electricity production, and even safety issues may be induced [17]. Therefore, it is crucial to maintain the power

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

plant to a healthy status consistently to meet performance standards through condition monitoring and resulting maintenance actions. From the maintenance perspective, an early detection of potential anomalies during operation should be preceded to prevent sudden breakdown of a power plant [3].

Along with the deployment of advanced sensing technologies (e.g., smart sensors, internet of things (IoT)), data-driven maintenance has been recognized as an essential scheme for not only improving asset availability but also reducing the costs for repair or replacement [10]. Particularly for the complex system like power plant, big data analytics and artificial intelligence (AI) techniques for maintenance activities have facilitated a paradigm shift from time-based maintenance that cannot completely avoid unnecessary shutdowns to predictive maintenance or condition-based maintenance [15]. By integrating historical data into the diagnosis of system state, it becomes feasible to improve the performance of a power plant by sustaining it healthy status [8]. Condition-based maintenance permits the operator to understand current status of an equipment, effectively indicating when the maintenance is required so that the operation would not be suspended accidentally. For condition-based maintenance, condition monitoring has been widely implemented across a diverse array of industries including energy [33], manufacturing [27] and transportation [5]. In general, anomaly data have the patterns not conforming to a well-defined notion of normal behavior. Based on how much observations from equipment differ from in-control characteristics, the condition monitoring of a system has been meticulously devised.

This research mainly aims to develop a condition monitoring scheme for a combined cycle power plant (CCPP). CCPP is an electricity generating plant that utilizes cogeneration cycles, where the earlier cycle is conducted by a combustion turbine generator and the latter cycle by a steam turbine generator [14]. In a gas turbine, the exhaust with temperature of 400°C-600°C is released, containing enough thermal energy for subsequent use in the steam turbine to regenerate electricity. Incorporating both gas turbines and steam turbines, CCPP fulfills more efficient electricity generation than a conventional single-cycle power plant [24]. It is reported that CCPP can improve the efficiency of power generation by over 58%, compared to traditional power plant with only 40% of generating efficiency [23].

CCPP has a relatively complex structure consisting of various components to perform power generation. The performance of CCPP depends highly on consistent operation of its components that can affect the safety, availability, and reliability of the whole system. Thus, real-time health monitoring of the components is important to maintain the integrity of CCPP [9]. To assess the status of components online, various types of sensors are installed on critical parts of equipments in CCPP. Based on the measured data, plant-wide condition monitoring is established by individually monitoring each component, to help the operator timely detect system abnormality and accurately diagnose to take relevant maintenance actions.

However, there are several bottlenecks to construct an effective condition monitoring scheme for CCPP. Firstly, operating data in power plants have inherent characteristics; cross-variable association [33]. Some variables are highly associated with others based on their working mechanisms, where redundant information makes an interpretation of condition monitoring results more difficult. Secondly, an accurate anomaly detection or its positioning for a monitoring system is generally difficult because of unavoidable noises and measurement errors caused by the variation of operational environments or the performance of sensors [11]. Lastly, the complexity of CCPP system hinders the operator from easily identifying root causes for faults. In practice, faults in a power plant tend to occur randomly and it is not easy to secure sufficient representatives for all types of faults.

Addressing such issues, a number of studies based on mechanical, statistical, and machine learning models have been conducted to accurately detect anomalies by reducing the correlation, uncertainty, and computational complexity of power plant data, which is given in Table 1. To name a few mechanical approaches, Peng et al. [22] proposed a principal component analysis (PCA)-based model and a multi-flow model to diagnose fault types in a nuclear power plant, where mechanical simulations were implemented to quantitatively ensure the accuracy of results from the models. Bonilla-Alvarado et al. [3] introduced an empirical transfer function for updating a pre-defined physical model, wherein the dynamics of the model parameters were directly reflected into condition monitoring. In case of statistical models, Tobar et al. [29] employed a hybrid approach with similarity-based modeling, PCA, and Hotelling's T^2 test to statistically ensure an efficient and reliable plant operation. Sabouhi et al. [24] developed a performance model for CCPP using reliability block diagrams to illustrate inter-relationships between subsystems. For machine learning-based condition monitoring, Chen et al. [4] used correlation analysis and decision trees for feature selection, along with support vector machine (SVM) for fault prediction. Wang et al. [30] employed kernel PCA for fault detection and feature extraction, which was followed by SVM for fault type identification and subsequent similarity clustering to assess fault severity. Chen et al. [6] integrated generalized regression neural network and B-Spline transformation to reduce multicollinearity problems between sensors in CCPP.

In existing literature, various feature extraction or transformation techniques have been used to strengthen the precision of condition monitoring scheme. Kesgin and Heperkan [13] showed that complex structure of correlations and inter-dependencies between relevant input and output can be efficiently resolved by providing performance estimates with reduced dimensions. However, most of proposed condition monitoring techniques for CCPP to date are limited to either the systems with similar functionalities,

TABLE 1. Related works with respect to CM for power plants.

Type of model	Literature	Monitoring object	Monitoring method
Mechanical model	Peng et al. [22]	Nuclear power plant	PCA-based multi-flow simulation
	Bonilla-Alvarado et al. [3]	Gas turbine-recuperated power plant	Empirical transfer function Dynamic modeling
Statistical model	Tobar et al. [29]	Natural gas power generation plant	Similarity-based PCA Hotelling's T^2 test
	Sabouhi et al. [24]	Separated plants for gas turbine and steam turbine in CCP	Reliability block diagrams between subsystems
Machine learning model	Chen et al. [4]	Thermal power plant	Correlation analysis-based decision trees and SVM
	Wang et al. [30]	Nuclear power plant	Kernel PCA and SVM
	Chen et al. [6]	Combined cycle generator	Generalized regression neural network B-Spline transformation

such as typical steam configurations and gas turbine control systems, or single-function plants. The interpretation of underlying anomaly effects is crucial for the prevention of failures in complex power systems. However, quantitative assessments regarding key contributing factors to anomalies and identification of their root causes have not been fully explored yet.

In this paper, we propose an integrated condition monitoring scheme of a CCPP for quantifying the influence of key variables on detected anomalies and uncovering their root causes to construct efficient maintenance strategy. To address high correlations between the variables with similar patterns, K-means clustering algorithm is employed. In a sequence, linear discriminant analysis (LDA) is applied to reduce the dimension of each cluster for emphasizing key features and mitigating data disturbance by removing unnecessary features. Based on selecting relevant variables from each cluster, Mahalanobis depth is introduced as a control statistic for condition monitoring, which enables anomaly detection in multidimensional data and facilitates causal inference. Through the integrated condition monitoring scheme, we aim to monitor the status of CCPP in real-time, and identify underlying factors of anomalies via contribution scores in case of out-of-control status. The proposed scheme is tailored to leverage critical information for anomaly detection, thereby filtering out extraneous variables from operational data. Consequently, it can efficiently take inherent correlation, uncertainty, and complexity within CCPP data into account.

The remainder of this paper is organized as follows. Section II illustrates K-means clustering-based LDA. Section III introduces our proposed scheme for condition monitoring and factor analysis contributing to anomalies based on the Mahalanobis depth. In Section IV, operational data of a CCPP in South Korea are analyzed to verify the performance of our proposed scheme by comparing with those of other existing methods. Finally, Section V concludes and discusses the directions for future research.

II. CLUSTERING-BASED LINEAR DISCRIMINANT ANALYSIS

To proactively perform anomaly detection in a large complex system, it is essential to extract key variables contributing to anomaly from a number of variables obtained from multiple sensors. In this section, we endeavor to extract key factors having critical impacts on system anomalies via a combined framework of K-means clustering algorithm and LDA. K-means clustering algorithm is executed to enhance the effectiveness of variable selection procedure by grouping all the variables into several clusters with similar characteristics. Then, LDA is applied to select key variables contributing to system anomalies from each cluster. Through the clustering-based LDA, pre-processed data with reduced dimension simplify computation and visualization, while removing unnecessary information for condition monitoring. This allows clustering the variables with similar characteristics and transforming complex data into interpretable forms.

Critical issues frequently encountered at condition monitoring of CCPP can be defined in terms of three key terms: high correlations between process variables, data disturbance by noises, and modeling complexity. In this work, the objective of introducing a K-means clustering algorithms is to reduce the dimensions of multivariate variables with similar characteristics to avoid multicollinearity problems between variables for making condition monitoring actions much simpler. After the clustering algorithm is employed to handle high correlations between variables to resolve the cross-variable problem, LDA is conducted to control unavoidable noises from operational environments. Through clustering-based LDA, high complexity caused by high correlations and environmental noises can be significantly resolved. The details of clustering-based LDA are given at the following.

A. K-MEANS ALGORITHM FOR VARIABLE CLUSTERING

Cluster analysis is a statistical method used to categorize a number of variables into several groups with high internal similarities by measuring the distance among individuals. The fundamental logic of cluster analysis is to classify objects in such a way that the variables within a same cluster exhibit homogeneous characteristics, while those in different clusters display heterogeneous characteristics. In this way, clustering aims to discern both the similarity within each cluster and the difference between objects in distinct clusters according to similarity measure [2], [7]. Usually, in the absence of prior information about a population, cluster analysis has served as a touchstone to better understand multivariate data.

As a method for minimizing the distance discrepancies, K-means clustering is commonly regarded as one of robust clustering techniques for handling highly correlated multivariate data, even if it is relatively simple and intuitive. K-means clustering groups all the data based on the centroids of each cluster, assigning each data point to the nearest centroid. The most commonly used metric for K-means clustering is the minimization of the sum of squared differences from a cluster center to each data point [25], [32]. Given the $N \times M$ multivariate dataset $\mathcal{X}$, the K-means clustering algorithm proceeds through the following steps:

- **Step 1**: Identify the optimal number of clusters that is represented by K and establish the maximum number of iterations.
- **Step 2**: Initiate the algorithm by positioning K centroids based on random initialization.
- **Step 3**: Associate each independent variable with the closest cluster by employing distance measurement.
- **Step 4**: Re-categorize data points into several clusters by comparing the distance between each data point and the centroids of all the clusters. This re-assignment is conducted based on the following membership function

$$z_{jk} = \begin{cases} 1, & \text{if } \mathbf{x}_{\cdot j} \text{ satisfies } \min\limits_{1 \le k \le K} \|\mathbf{x}_{\cdot j} - \mathbf{c}_k\|^2, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

  with respect to the jth set of variables, $\mathbf{x}_{\cdot j} = [x_{1j}, \ldots, x_{Nj}]^T$, and kth centroid, $\mathbf{c}_k = [c_{1k}, \ldots, c_{Nk}]^T$, for $j = 1, \ldots, M$. If the $\mathbf{x}_{\cdot j}$ belongs to kth cluster, then $z_{jk} = 1$, otherwise $z_{jk} = 0$.
- **Step 5**: Update the positions of cluster centroids based on new data point assignments. In this context, z_{jk} signifies the membership value of $\mathbf{x}_{\cdot j}$ to the centroid of cluster $\mathbf{c}_k$. The objective function for this method, which considers both the distance and the membership value of data points within each cluster, is defined as [19]

$$f = \sum_{j=1}^{M} \sum_{k=1}^{K} z_{jk} \|\mathbf{x}_{\cdot j} - \mathbf{c}_k\|^2. \tag{2}$$

- **Step 6**: If the positions of cluster centroids have shifted or the iteration has not reached pre-determined maximum, return to Step 3. Otherwise, output final cluster assignments.

The performance of K-means clustering algorithm is significantly affected by the choice of K value. To objectively determine the optimal number of K, elbow method can be employed [16], [28]. When clustering independent variables, the elbow method computes the sum of squares within clusters, which represents the total intra-cluster variation based on the number of clusters formed. This method aims to identify the elbow point as the optimal number of clusters, where the rate of decrease in the sum of squares within clusters starts to gradually diminish. The procedures for implementing the elbow method are given as follows:

- **Step 1**: Identify a range of cluster number K. If the range is set to $1, 2, \ldots, k$, the clustering results with k cases would be inspected and evaluated.
- **Step 2**: For each value of K, conduct K-means clustering algorithm to partition all the variables into the pre-determined number of clusters.
- **Step 3**: Calculate the sum of square errors (SSEs) for each K. The SSE is defined as the sum of squared distances between each variable and its cluster centroid as

$$SSE = \sum_{k=1}^{K} \sum_{\mathbf{x}_{\cdot j} \in C_k} \|\mathbf{x}_{\cdot j} - \mathbf{c}_k\|_2^2, \tag{3}$$

  for the kth cluster set, C_k.
- **Step 4**: Plot the calculated SSE on the y-axis over the number of clusters on the x-axis to visualize the elbow curve.
- **Step 5**: Find the elbow point where the slope in SSE gradually decreases. Finally, the number of cluster having elbow point is selected as the best choice of K.

B. LINEAR DISCRIMINANT ANALYSIS FOR VARIABLE SELECTION

Through the K-means clustering algorithm and the elbow method, the variables with similar characteristics are grouped into the proper number of clusters. In a sequence, representative variables in each cluster can be identified by extracting significant variables contributing to anomalies. As a statistical method for classifying data and reducing its dimensionality, LDA has been widely employed across diverse domains such as pattern recognition, category classification, and feature dimension reduction. The objective of LDA is to discover a transformation matrix not only maximizing the distance between different classes, but also minimizing the internal variance of each class. By finding the linear decision boundaries that best separate given data, LDA can facilitate effective dimension reduction and removal of irrelevant information [12]. LDA generally secures highly reliable and robust results for labelled data by measuring the difference in means between classes, increasing the likelihood of clear separation, and establishing well-defined decision boundaries.

To efficiently reduce data dimensions, significant variables contributing to abnormal status should be defined *a priori*. For evaluating the likelihood that a specific data point belongs to a particular class, the conditional probability $P(\mathcal{W}_l|\mathbf{x}_{i\cdot})$ for a given observation $\mathbf{x}_{i\cdot}$ belonging to a specific class $\mathcal{W}_l$ is computed through the Bayes' rule

$$P(\mathcal{W}_l|\mathbf{x}_{i\cdot}) = \frac{P(\mathbf{x}_{i\cdot}|\mathcal{W}_l) \cdot P(\mathcal{W}_l)}{P(\mathbf{x}_{i\cdot})}, \tag{4}$$

where $\mathbf{x}_{i\cdot} = \left[x_{i1}, \ldots, x_{iM_k}\right]$ for the total number of variables in kth cluster, M_k. Here, $P(\mathcal{W}_l)$ represents the prior probability of $\mathcal{W}_l$ that indicates the probability of an observation included in the class $\mathcal{W}_l$. $P(\mathbf{x}_{i\cdot}|\mathcal{W}_l)$ denotes the likelihood of ith observation $\mathbf{x}_{i\cdot}$ given that it belongs to the class $\mathcal{W}_l$, reflecting the distribution of the class $\mathcal{W}_l$. Analogously, $P(\mathbf{x}_{i\cdot})$ is the probability of observed vector $\mathbf{x}_{i\cdot}$ regardless of its class.

In LDA, by evaluating $P(\mathcal{W}_l|\mathbf{x}_{i\cdot})$ across various classes, each data point is assigned to the class with the highest probability. For anomaly detection, we consider two classes $\mathcal{W}_0$ and $\mathcal{W}_1$ as normal and abnormal status of the system, respectively. Then, a decision boundary between two classes is established by maximizing the between-class scatter (S_B) and minimizing the within-class scatter (S_W). The expressions for S_B and S_W are given as

$$S_B = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T, \tag{5}$$

$$S_W = \sum_{\mathbf{x}_{i\cdot} \in \mathcal{W}_0} (\mathbf{x}_{i\cdot} - \boldsymbol{\mu}_0)(\mathbf{x}_{i\cdot} - \boldsymbol{\mu}_0)^T + \sum_{\mathbf{x}_{i\cdot} \in \mathcal{W}_1} (\mathbf{x}_{i\cdot} - \boldsymbol{\mu}_1)(\mathbf{x}_{i\cdot} - \boldsymbol{\mu}_1)^T, \tag{6}$$

respectively. In the above equations, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are the mean vectors of normal and abnormal status, respectively [26]. Through the linear transformation G into lower dimensional space, the optimal decision boundary can be obtained from the following problem

$$\max_G \frac{G^T S_B G}{G^T S_W G}, \tag{7}$$

which maximizes S_B and minimizes S_W simultaneously [31]. After simplification, the optimization problem of Eq. (7) is identical to the eigenvalue problem of $S_W G - \lambda S_B G = 0$. A single set of eigenvalue and eigenvector is calculated because the decision boundary of LDA is determined by two classes for anomaly detection. For more details on eigenvalue-based linear equations, see the reference [20].

In the context of LDA, the dimension reduction with significant variables for each cluster is conducted by examining the eigenvector corresponding to the linear decision boundaries. Firstly, the importance of variables is arranged in the order of elements with high values within the computed eigenvector. Subsequently, to determine the variable set exhibiting the highest diagnostic accuracy, variables are combined in descending order of importance. Then, the classification results from the combination of the variables are compared. By projecting the original data onto selected variable set with the highest performance, features can be generated in the reduced-dimensional space for each cluster. As one of the performance evaluation metrics for classification problems, $F1$-score is defined as

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

where the values of precision and recall are defined as $Precision = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and $Recall = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, respectively. Table 2 shows the confusion matrix for performance evaluation.

TABLE 2. Confusion matrix for performance evaluation.

Confusion matrix		Predicted: $\mathcal{W}_0$ (Normal)	Predicted: $\mathcal{W}_1$ (Abnormal)
Actual	$\mathcal{W}_0$	True Negative (TN)	False Positive (FP)
	$\mathcal{W}_1$	False Negative (FN)	True Positive (TP)

III. CONDITION MONITORING AND ANOMALY CONTRIBUTION ANALYSIS

In this section, we propose an integrated framework for real-time condition monitoring and contribution analysis for anomalies using the concept of multivariate control chart. Utilizing the primary information extracted from original operation data, control statistics over operational time are defined to monitor and diagnose the current status of a system. In case of any abnormal state, a quantitative analysis of contributing factors to anomalies is conducted for clear inference for following maintenance actions.

As one of control statistics, Mahalanobis depth offers distinct advantages in terms of robustness and reliability, particularly in the identification of outliers or anomalies within complex datasets. Mahalanobis depth considers the covariance structure that captures how variables are interrelated each other. This approach allows the Mahalanobis depth to precisely measure the deviation of observations from usual patterns in multivariate space. Furthermore, as a non-parametric approach, its capacity of adjusting to overall dispersion of data enables reliable anomaly detection without the need for stringent distributional assumption for observed data. Through this metric, it becomes possible to effectively inspect the current state based on complex data and investigate abnormal causes for system improvement.

For monitoring purpose, we suggest the following procedure consisting of three phases: (1) define control statistics and their control limits (CLs), (2) monitor and detect anomalies via the control chart, and (3) derive the contribution score based on the Mahalanobis depth. We will explain the procedure in details at the following.

A. PHASE I: DETERMINATION OF CONTROL STATISTICS AND CONTROL LIMITS

Mahalanobis depth is a sophisticated metric for assessing the similarity between a specific observation at a given time point and overall distribution. By quantifying the centrality of target observation $\mathbf{x}$ within a (multivariate) given distribution $\mathcal{F}$, Mahalanobis depth is defined as [21]

$$MD(\mathbf{x}_{i\cdot}) = \frac{1}{1 + (\mathbf{x}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot})^T S^{-1} (\mathbf{x}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot})}, \tag{8}$$

for the ith sample mean vector $\bar{\mathbf{x}}_{i\cdot}$, and sample covariance matrix S. Through this equation, Mahalanobis depth simplifies complex multivariate data into a single measure, indicating how closely an observation aligns with the central trend of the distribution.

Because the Mahalanobis depth is a nonparametric measure, CLs are established without any assumptions about a specific distribution of the data. Alternatively, reference data should be used to determine appropriate CLs for detecting out-of-control state. In general, reference data is the baseline data when the system is in-control state, implying typical behaviors of the system. Assuming that all observations are in a normal state, CLs are calculated based on a pre-defined significance level α. Also known as "type I error", α represents the probability of false alarm, that is, falsely declaring the probability that the system is out-of-control even though it is actually in-control. The consideration of type I error is particularly crucial for complex and large-scale systems like CCPP, because a substantial amount of costs is associated with false alarms. The lower control limit (LCL) based on the Mahalanobis depth statistic is set as

$$LCL_{MD} = P_{MD}(\alpha), \tag{9}$$

where $P_{MD}(\alpha)$ is the 100αth percentile of the Mahalanobis depth distribution. The distribution of Mahalanobis depth can be empirically obtained using the density of Mahalanobis depth statistics from reference data.

B. PHASE II: ONLINE MONITORING

After setting the Mahalanobis depth statistics and the LCL based on reference data, condition monitoring can be conducted to continuously assess the status of the system. For newly acquired data, the Mahalanobis depth statistics are re-calculated to evaluate sample mean vector and covariance matrix obtained from the reference data. Using the real-time Mahalanobis depth, we can quantitatively diagnose how much new observations deviate from normal behaviors of the system. The Mahalanobis depth values for new observations are sequentially compared against pre-established LCL, and if it falls below the LCL, the system is considered to be out-of-control (or anomaly). By establishing such an online monitoring scheme, system anomalies can be detected quickly prior to the occurrence of impending failures.

Once an abnormal operation is detected, the system operator needs to decide to either allow these anomalies within the system's tolerance or implement maintenance actions immediately. In case that the control statistics are consecutively under the LCL over an inspection period, the online monitoring is interrupted and corrective measures should be initiated to preserve the system's integrity. By recovering the system to its standard operating condition, the efficiency and reliability of CCPP can be sustained.

C. PHASE III: ANOMALY CONTRIBUTION ANALYSIS

To conduct proper maintenance actions for anomalies, it is necessary to investigate which factors have significantly influenced the system's abnormality. For this purpose, contribution analysis is employed to pinpoint the variables that are primarily responsible for the anomalies. The contribution analysis decomposes the Mahalanobis depth statistic into several components, where each element quantifies the influence of extracted variables (or features) [1]. This dissection helps the operator identify key variables that may be root causes of system disturbances.

In contribution analysis for anomalies, Mahalanobis depth statistic can be mathematically represented as the sum of weighted squared deviations for each variable, using the sample covariance matrix for its computation. Mahalanobis depth is divided into individual contribution scores for each variable and Eq. (8) is reformulated as

$$MD(\mathbf{x}_{i\cdot}) = \frac{1}{1 + \sum_{j_1 \in j} \sum_{j_2 \in j} w_{j_1 j_2}(x_{ij_1} - [\bar{x}_{i\cdot}]_{j_1})(x_{ij_2} - [\bar{x}_{i\cdot}]_{j_2})}, \tag{10}$$

where x_{ij_n} and $[\bar{x}_{i\cdot}]_{j_n}$ are the j_nth individual data of $\mathbf{x}_{i\cdot}$ and the j_nth element of $\bar{\mathbf{x}}_{i\cdot}$ for $n = 1, 2$ and $j = 1, \ldots, K$, respectively. Through the decomposition, the weighted squared deviation for jth variable, $w_{jj}(x_{ij} - \bar{x}_{ij})^2$ is then calculated, and the importance for each variable is ranked based on their contribution scores w_{jj} for $j = 1, \ldots, K$. This allows for identifying primary sources behind observed anomalies. Once major contributed factors are identified, appropriate corrective actions can be suggested.

Because a periodical re-calculation of mean vector and covariance matrix is essential to conduct real-time condition monitoring, the sample mean and covariance matrix at phase III are continuously updated with respect to new operating data. Accordingly, the CLs should be re-calculated based on updated Mahalanobis depth statistics. By following these steps, real-time condition monitoring can be effectively performed, supporting the earlier detection of anomalies and their root causes, then prompt actions for the system maintenance.

IV. APPLICATION: COMBINED CYCLE POWER PLANT

In this section, monitoring data of a CCPP in South Korea was analyzed to illustrate the proposed condition monitoring scheme to early detect the anomalies caused by combustion oscillations within the CCPP. The data was collected at one-minute sampling intervals from a total of 1,106 sensors, spanning a period of 3 days. During this data collection period, the CCPP system experienced two types of anomalous operations for approximately 1 hour and 44 minutes, and 2 hours and 27 minutes, respectively. For the sake of convenience, each anomaly will be designated as Case A and Case B. Each fault scenario involves the issues with both the gas turbine and the steam turbine. Case A concerns a failure linked to combustion dynamics pressure, which is caused by the problems with the fuel supply, whereas Case B involves a malfunction in the heat recovery steam generator of the steam turbine. The details on the variables and the interval of anomaly are confidential, thus we did not openly list them in the analysis. Given that two distinct anomalies were identified within the collected data, pre-processing and K-means clustering were uniformly performed to the data set. Subsequent procedure of dimensionality reduction using LDA and contribution analysis based on the Mahalanobis depth requires divergent approaches tailored to each type of anomaly. Consequently, the analyses for the two scenarios were independently conducted to accommodate unique characteristics of each anomaly.

A. DATA PRE-PROCESSING AND EXPLORATION

To facilitate anomaly detection process, data pre-processing was performed by defining proper ranges of individual variables. In practice, out of the 1,106 variables in the dataset, 916 variables were used for this analysis, excluding irrelevant variables to condition monitoring. In advance of data reduction, a preliminary correlation analysis was performed to assess the presence of high correlations between variables. This step is essential because high correlation between variables can affect the performance of anomaly detection. As a standard technique for calculating the relationship between variables $\mathbf{x}_{\cdot j_1}$ and $\mathbf{x}_{\cdot j_2}$, the Pearson's correlation coefficient was employed as

$$\rho_{j_1,j_2} = \frac{\sum_{i=1}^{N}(x_{ij_1} - [\bar{x}_{\cdot j_1}]_i)(x_{ij_2} - [\bar{x}_{\cdot j_2}]_i)}{\sqrt{\sum_{i=1}^{N}(x_{ij_1} - \bar{x}_{\cdot j_1})^2}\sqrt{\sum_{i=1}^{N}(x_{ij_2} - \bar{x}_{\cdot j_2})^2}}.$$

In the correlation analysis for the CCPP, a number of variables were highly correlated with each other, as shown in Fig. 1.

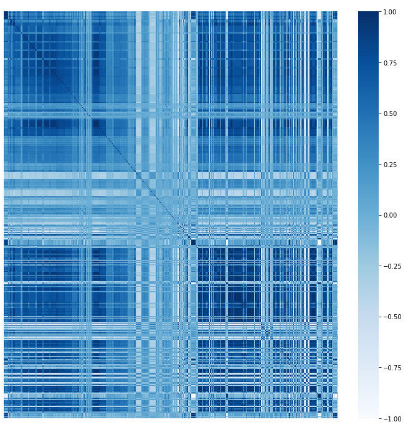
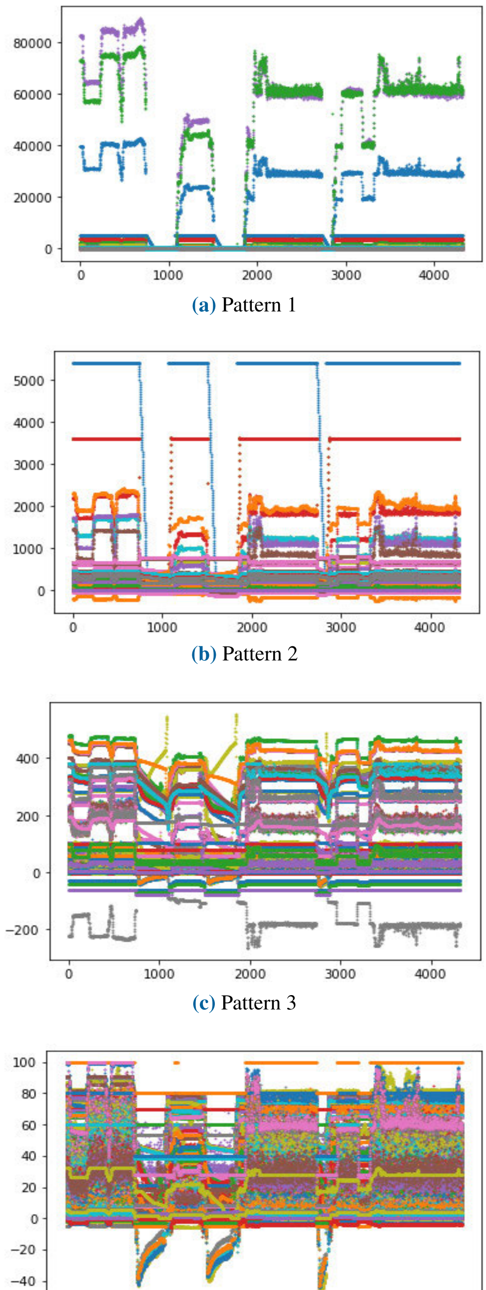


FIGURE 1. Visualization of the correlation matrix of all independent variables.

For example, the variables with similar patterns over operating time are shown in Fig. 2. Through these observations, it was noted that clustering procedure should be conducted to reduce the dimensionality of the data in advance.

B. DIMENSION REDUCTION VIA K-MEANS CLUSTERING-BASED LDA

To facilitate condition monitoring scheme, clustering the variables with similar patterns is required. For this purpose, K-means clustering was conducted to group all the variables into several clusters with similar characteristics. Because K-means clustering is greatly affected by the number of clusters K, the elbow method was introduced to determine the best number of clusters. Based on SSEs for each cluster, the elbow plot is given in Fig. 3. The elbow method shows that the optimal number of clusters is 3, and the results of K-means clustering with $K = 3$ are given in Fig. 4. The number of variables in each cluster, C_1, C_2 and C_3, are 537, 275 and 104, respectively.



(a) Pattern 1

(b) Pattern 2

(c) Pattern 3

(d) Pattern 4

FIGURE 2. Multiple similar patterns with high correlation coefficients.

After clustering independent variables into three groups, LDA-based dimension reduction was executed to each group separately to extract significant variables contributing to

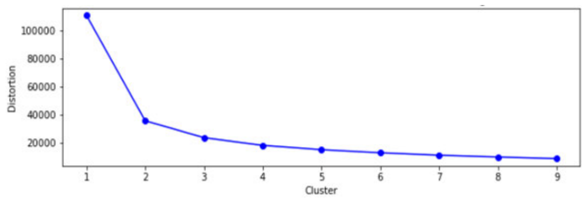
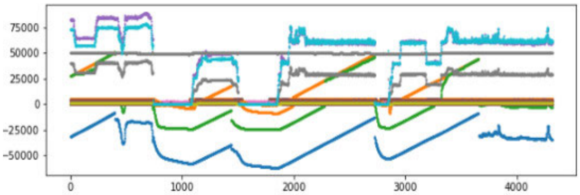
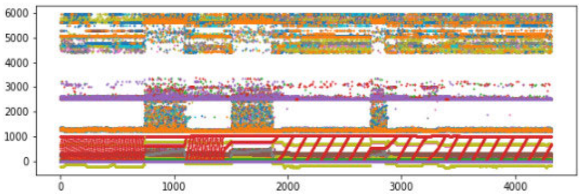


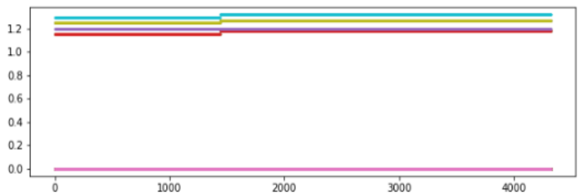
FIGURE 3. Elbow plot for CM data of CCP in south korea.



(a) Class 1 (C_1)

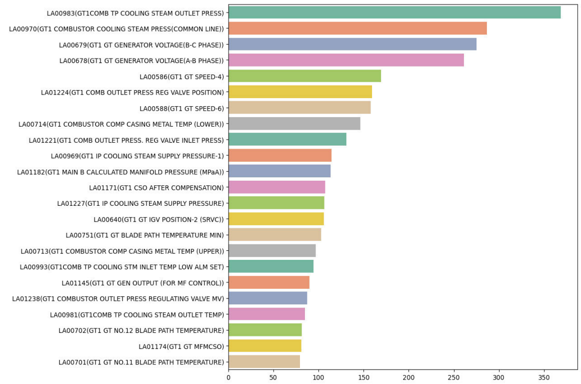


(b) Class 2 (C_2)

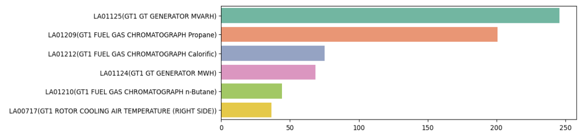


(c) Class 3 (C_3)

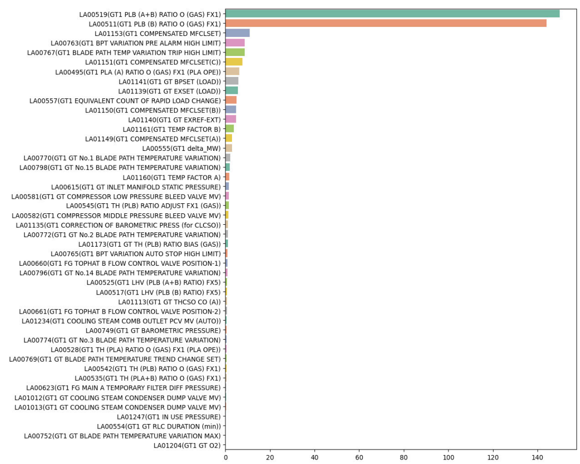
FIGURE 4. Visualization of K-means clustering results with $K = 3$.



(a) C_1



(b) C_2



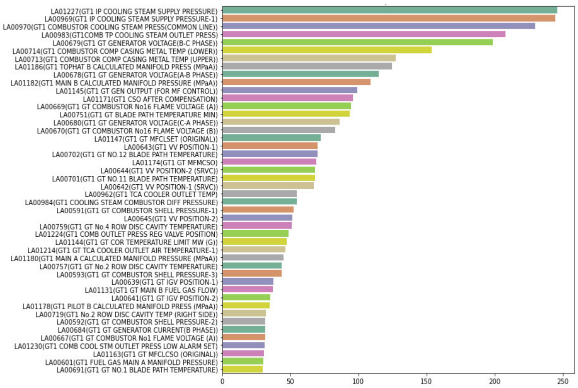
(c) C_3

FIGURE 5. The results of feature extraction for each cluster for Case A.

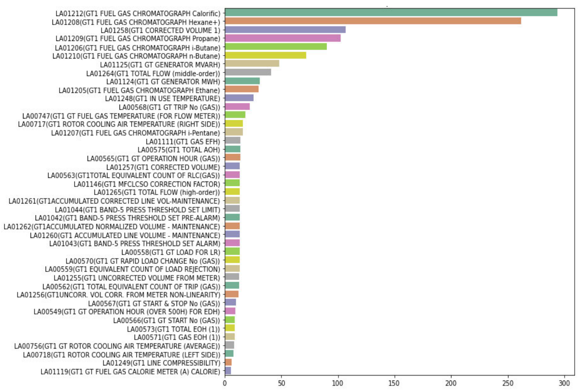
detected anomalies. To enhance the accuracy of anomaly detection, we identified key variables with the higher eigenvector elements. The composition of each cluster was determined by the $F1$-score such that improved classification performance could be provided. For Case A and Case B, corresponding features were individually given in Fig. 5 and Fig. 6, respectively. In each cluster, important features with the number of (23, 6, 46) for Case A, and (46, 43, 20) for Case B were selected as key variables for the clusters C_1, C_2, and C_3, respectively. The significant variables from each cluster were linearly combined by LDA, aiming to distinguish normal and abnormal decision boundaries. As a result, the variables were clustered into three dimensions, one dimension per cluster for each anomalous operation case.

To evaluate the effectiveness of dimension reduction from LDA, we compared classification results of cluster-based LDA with cluster-based PCA and original data without dimension reductionfor two types of anomalous operations. PCA is one of representative methods for dimension reduction to generate reduced variables with a linear projection of interrelated variables [29]. Using PCA, the dimensions were reduced to a small number of principal components (PCs) explaining more than 80% of the total variance for each cluster. As a result, 2, 35, and 11 PCs for both cases were selected.

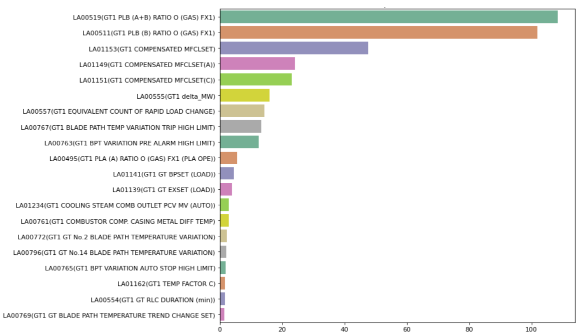
To objectively evaluate the performance of anomaly detection, three types of classifiers, SVM, random forest (RF), and artificial neural network (ANN), were applied for classification between two classes: normal and abnormal status. For the details on the three classification methods, refer to [18]. Because monitoring data in CCPP was imbalanced with a small amount of abnormal data, classification performance was evaluated based on the $F1$-score, including Recall and Precision. For each classifier, hyper-parameter tuning was conducted to secure better classification performance and 10-fold cross-validation was carried out.



(a) C_1



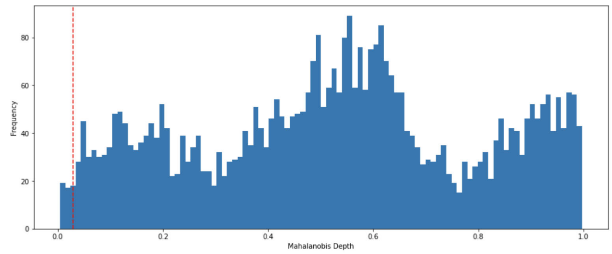
(b) C_2



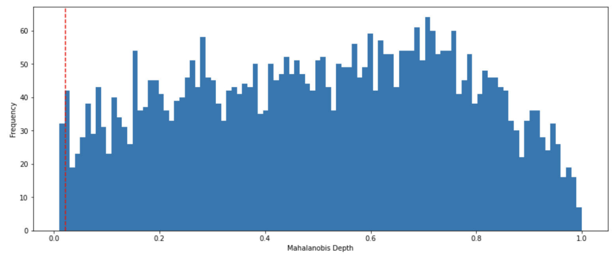
(c) C_3

FIGURE 6. The results of feature extraction for each cluster for Case B.

The comparison results for each cluster using original data, PCA- and LDA-based reduced data are given in Table 3. Table 3 shows that cluster-based LDA method explains the total variability of original data using only three features. Note that cluster-based LDA method is robust to the types of classifiers and provides stable $F1$-scores compared to the other methods. Moreover, the Recall metric serves as a crucial indicator within the monitoring framework to assess the promptness and accuracy in identifying actual anomalous operations. From this standpoint, cluster-based LDA provides an effective detection results compared to the other methods. Although some of classifiers derived from original data and PCA-based reduced data demonstrate enhanced precision, a significant proportion of observations are categorized as False Negative (FN), consequently resulting in reduced Recall accuracy.



(a) Case A



(b) Case B

FIGURE 7. Distribution of the MD statistics in normal status with LCL (red line).

C. MAHALANOBIS DEPTH-BASED CONDITION MONITORING AND CONTRIBUTION ANALYSIS

After reducing dimensionality via LDA, monitoring standards were built to assess the status of the CCPP and identify influential factors contributing to abnormal operation. In the initial phase, Mahalanobis depth was used as the control statistic for dimension-reduced data in the CCPP data set. Subsequently, the LCL was determined at a pre-specified significance level from reference data. In this application, the significance level is conservatively set to $\alpha = 0.01$ to minimize false alarm rate. Based on observed data at normal operating condition, the LCL for Case A and Case B was set to be 0.0289 and 0.02270, respectively. Along with Mahalanobis depth for the reduced data from LDA, the LCL line was given in Fig. 7. As shown in Fig. 7, Mahalanobis depth-based non-parametric condition monitoring scheme is appropriate for CCPP operating data, where the Mahalanobis depth statistics for both cases seem not to follow a specific parametric distribution, e.g., normal distribution.

Upon configuring the control statistics and CLs, online condition monitoring was conducted in Phase II. Our monitoring scheme detected the system anomalies spanning from the 2,732-2,836 time points (104 minutes)for Case A and 745-892 time points (minutes) for Case B, where the Mahalanobis depths lie under the LCL consecutively. The period exactly coincides with the abnormal period about that the operator gave information *a priori*. Notably, the online monitoring scheme can also detect the precursors for anomalies at least 5 minutes before an engineer actually

TABLE 3. The results of dimension reduction and classification.

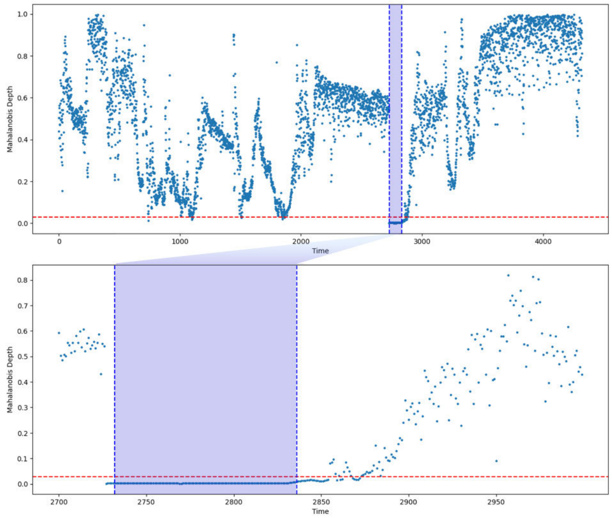
			Original data	PCA	LDA
The number of variables		C_1	537	2	1
		C_2	275	35	1
		C_3	104	11	1
		Total	916	48	3
Case A	SVM	$F1$-score	95.66%	98.36 %	99.37%
		Recall	91.79%	96.77%	100.00%
		Precision	100.00%	100.00%	98.76%
	RF	$F1$-score	98.82%	98.21%	99.52%
		Recall	97.96%	96.55%	100.00%
		Precision	99.71%	100.00%	99.05%
	ANN	$F1$-score	98.55%	98.51%	99.30%
		Recall	97.14%	97.08%	100.00%
		Precision	100.00%	100.00%	98.63%
	Mean	$F1$-score	98.68%	98.36%	99.40%
		Recall	97.55%	96.80%	100.00%
		Precision	99.85%	100.00%	98.81%
Case B	SVM	$F1$-score	93.43%	94.41%	98.59%
		Recall	89.70%	94.04%	97.25%
		Precision	97.65%	94.94%	100.00%
	RF	$F1$-score	96.30%	97.25%	98.82%
		Recall	92.86%	94.64%	97.92%
		Precision	100.00%	100.00%	99.78%
	ANN	$F1$-score	82.64%	93.90%	98.91%
		Recall	82.69%	93.98%	98.15%
		Precision	96.22%	93.93%	99.70%
	Mean	$F1$-score	89.47%	95.18%	98.77%
		Recall	87.77%	94.22%	97.77%
		Precision	98.11%	96.29%	99.83%

TABLE 4. The results of dimension reduction and classification.

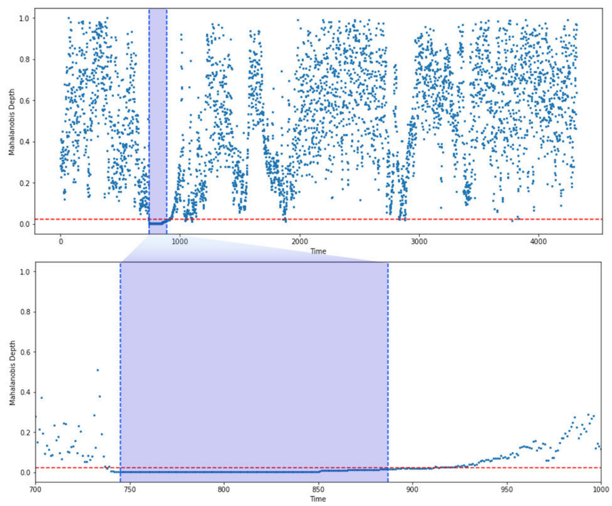
			Normal	Abnormal
Case A	Variance-covariance matrix		$\begin{bmatrix} 0.9356 & 0.1731 & -0.1050 \\ 0.1731 & 1.0217 & -0.0971 \\ -0.1050 & -0.0971 & 0.9990 \end{bmatrix}$	$\begin{bmatrix} 3.6108 & -0.4057 & -0.5696 \\ -0.4057 & 0.1190 & 0.0624 \\ -0.5696 & 0.0624 & 1.0389 \end{bmatrix}$
	Contribution Score	C_1	35.35%	3.14%
		C_2	32.29%	89.91%
		C_3	32.36%	6.95%
Case B	Variance-covariance matrix		$\begin{bmatrix} 0.8560 & 0.2575 & -0.4281 \\ 0.2575 & 0.4961 & -0.2693 \\ -0.4281 & -0.2693 & 0.7273 \end{bmatrix}$	$\begin{bmatrix} 5.0853 & 7.5678 & -6.4446 \\ 7.5678 & 15.2987 & -8.9864 \\ -6.4446 & -8.9864 & 8.7387 \end{bmatrix}$
	Contribution Score	C_1	26.54%	69.31%
		C_2	40.45%	3.81%
		C_3	33.00%	26.88%

detect them. Using our online monitoring scheme, it is expected to prevent imminent fault or failure by giving warning signs and taking proper actions in advance.

In a sequence, we endeavored to detect influential factors that directly affect the anomalies. Contribution analysis for anomalies was performed to identify key variables to the anomalies in Phase III. For this purpose, contribution scores were individually calculated for each variable, based on the sum of weighted squared deviations. Along with variance-covariance matrices, Table 4 presents contribution scores for both normal and abnormal periods separately for comparison. While all the classes evenly contribute to normal status, it is clearly noted that the cluster 2 (C_2) for Case A and the cluster 1 for Case B from LDA mainly contribute to



(a) Case A



(b) Case B

FIGURE 8. Visualization of the CM result based on MD with LCL (red line) and abnormal period (blue dashed lines).

anomalies with almost 90% and 70% of contribution scores, respectively. Based on the anomaly contribution analysis, it is necessary to delve into the clusters with high contribution scores more closely to identify main key variables.

To evaluate the accuracies of the contribution analyses, we compared expert interpretation of fault causes for each case against data-driven analysis results. Case A is classified into a failure type related to combustion dynamics pressure in gas turbine, attributed to fuel supply issues. The data-driven analysis indicated that cluster 2, containing variables closely associated with fuels, presented a high contribution score, aligning with identified cause of the failure. Similarly, Case B pertains to a fault in the heat recovery steam generator of steam turbine. Here, the variables in Cluster 1, which is related to temperature control, combustion flame, and pressure, also scored high contribution. From these facts, we could confirm that the contribution analysis via the Mahalanobis depth has a potential in identifying key variables that critically affect operational anomalies in CCPP.

V. CONCLUSION

In this study, we propose an efficient online condition monitoring scheme tailored for CCPP. To address intricate challenges inherent in sensing data within a complex system, which are characterized by high correlation, data disturbance, and limited interpretability, we introduce a comprehensive condition monitoring scheme for a large-scale complex system of CCPP. This scheme incorporates clustering-based dimension reduction, data-depth analysis, and anomaly contribution analysis sequentially. Through K-means clustering and LDA, we extract a concise set of groups significantly influencing system anomalies. Then, Mahalanobis depth is employed as a measure of control statistics, enabling precise diagnosis and evaluation of plant operation status. Additionally, by calculating contribution scores, we endeavor to identify key influential factors contributing to anomalies, which can support proactive fault prevention by initiating efficient maintenance actions. Empirical analysis of operational data from a CCPP demonstrates the potential of our condition monitoring scheme in timely and preemptive detection of operational anomalies. Notably, the integration of nonparametric control charts based on Mahalanobis depth statistics further enhances the capability of earlier anomaly detection in a complex system with a large number of monitoring variables that is not easy to apply parametric control charts. By means of the proposed method, it is expected to prevent imminent faults or failures by taking proper actions to relevant factors in advance. The contribution of this work can be briefly summarized as follows:

- The proposed method provides maintenance-efficient condition monitoring by mainly focusing on key variables influencing the anomalies in a combined cycle power plant.
- It is possible to quantify the degrees of variables having a major impact on the anomalies when an abnormal status is diagnosed.
- The proposed method help the operator find causal inference for abnormal status, contributing to reliability enhancement and maintenance improvement in a combined cycle power plant.

Admittedly, the diagnosis of anomalies and inference of their causes in a complex system still pose challenging problems for future research. The consideration of external factors such as environmental changes, equipment degradation, and variations in operating conditions is expected to enhance the robustness of online condition monitoring scheme. Based on the proposed condition monitoring scheme, optimal time-based or condition-based maintenance policy can be devised to improve the reliability of power generating equipment and reduce operational costs. By determining appropriate time points for repair or replacement of components or subsystems, the integrity of power plants is expected to be greatly improved. Finally, this approach can be easily extended to other similar complex systems such as nuclear power plants, chemical plants, and semiconductor manufacturing processes for condition monitoring.

REFERENCES

[1] S. J. Bae, G. Do, and P. Kvam, "On data depth and the application of nonparametric multivariate statistical process control charts," *Appl. Stochastic Models Bus. Ind.*, vol. 32, no. 5, pp. 660–676, Sep. 2016.

[2] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Behav. Sci.*, vol. 12, no. 2, pp. 153–155, Mar. 1967.

[3] H. Bonilla-Alvarado, K. M. Bryden, L. Shadle, D. Tucker, and P. Pezzini, "Development of real-time system identification to detect abnormal operations in a gas turbine cycle," *J. Energy Resour. Technol.*, vol. 142, no. 7, Jul. 2020, Art. no. 070903.

[4] K.-Y. Chen, L.-S. Chen, M.-C. Chen, and C.-L. Lee, "Using SVM based method for equipment fault detection in a thermal power plant," *Comput. Ind.*, vol. 62, no. 1, pp. 42–50, Jan. 2011.

[5] C. Chen, D. Zhang, P. S. Castro, N. Li, L. Sun, S. Li, and Z. Wang, "IBOAT: Isolation-based online anomalous trajectory detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 806–818, Jun. 2013.

[6] J. Chen, H. Li, D. Sheng, and W. Li, "A hybrid data-driven modeling method on sensor condition monitoring and fault diagnosis for power plants," *Int. J. Electr. Power Energy Syst.*, vol. 71, pp. 274–284, Oct. 2015.

[7] R. C. Dubes and A. K. Jain, *Algorithms for Clustering Data*. New York, NY, USA: Prentice-Hall, 1988.

[8] M. Fast and T. Palmé, "Application of artificial neural networks to the condition monitoring and diagnosis of a combined heat and power plant," *Energy*, vol. 35, no. 2, pp. 1114–1120, Feb. 2010.

[9] A.-T. W. K. Fahmi, K. R. Kashyzadeh, and S. Ghorbani, "A comprehensive review on mechanical failures cause vibration in the gas turbine of combined cycle power plants," *Eng. Failure Anal.*, vol. 134, Apr. 2022, Art. no. 106094.

[10] Y. Han and Y. H. Song, "Condition monitoring techniques for electrical equipment—A literature survey," *IEEE Trans. Power Del.*, vol. 18, no. 1, pp. 4–13, Jan. 2003.

[11] P. Hundi and R. Shahsavari, "Comparative studies among machine learning models for performance estimation and health monitoring of thermal power plants," *Appl. Energy*, vol. 265, May 2020, Art. no. 114775.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer, 2013.

[13] U. Kesgin and H. Heperkan, "Simulation of thermodynamic systems using soft computing techniques," *Int. J. Energy Res.*, vol. 29, no. 7, pp. 581–611, 2005.

[14] R. Kehlhofer, B. Rukes, F. Hannemann, and F. Stirnimann, *Combined-Cycle Gas and Steam Turbine Power Plants*. PennWell Books, LLC, 2009.

[15] C. Kirchsteiger, L. Balling, and J. Teichmann, "Probabilistic outage analysis of a combined-cycle power plant," *Power Eng. J.*, vol. 9, no. 3, pp. 137–141, Jun. 1995.

[16] F. Liu and Y. Deng, "Determine the number of unknown targets in open world based on elbow method," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 5, pp. 986–995, May 2021.

[17] B. Lindemann, B. Maschler, N. Sahlab, and M. Weyrich, "A survey on anomaly detection for technical systems using LSTM networks," *Comput. Ind.*, vol. 131, Oct. 2021, Art. no. 103498.

[18] M. Lim, B. Vidakovic, and S. J. Bae, "Fault classification via energy based features of two-dimensional image data," *Commun. Statist.-Theory Methods*, vol. 52, no. 11, pp. 3939–3959, Jun. 2023.

[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probabilities*, vol. 1, 1967, pp. 281–296.

[20] A. M. Martinez and M. Zhu, "Where are linear feature extraction methods applicable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1934–1944, Dec. 2005.

[21] B. M. Mun, M. Lim, and S. J. Bae, "Condition monitoring scheme via one-class support vector machine and multivariate control charts," *J. Mech. Sci. Technol.*, vol. 34, no. 10, pp. 3937–3944, Oct. 2020.

[22] M.-J. Peng, H. Wang, S.-S. Chen, G.-L. Xia, Y.-K. Liu, X. Yang, and A. Ayodeji, "An intelligent hybrid methodology of on-line system-level fault diagnosis for nuclear power plant," *Nucl. Eng. Technol.*, vol. 50, no. 3, pp. 396–410, Apr. 2018.

[23] A. Poullikkas, "An overview of current and future sustainable gas turbine technologies," *Renew. Sustain. Energy Rev.*, vol. 9, no. 5, pp. 409–443, Oct. 2005.

[24] H. Sabouhi, A. Abbaspour, M. Fotuhi-Firuzabad, and P. Dehghanian, "Reliability modeling and availability analysis of combined cycle power plants," *Int. J. Electr. Power Energy Syst.*, vol. 79, pp. 108–119, Jul. 2016.

[25] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020.

[26] F. Song, D. Mei, and H. Li, "Feature selection based on linear discriminant analysis," in *Proc. Int. Conf. Intell. Syst. Design Eng. Appl.*, vol. 1, Oct. 2010, pp. 746–749.

[27] G. A. Susto, M. Terzi, and A. Beghi, "Anomaly detection approaches for semiconductor manufacturing," *Proc. Manuf.*, vol. 11, pp. 2018–2024, Jan. 2017.

[28] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 336, Apr. 2018, Art. no. 012017.

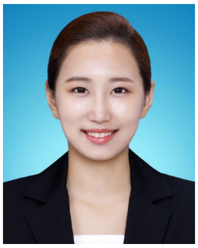
[29] F. A. Tobar, L. Yacher, R. Paredes, and M. E. Orchard, "Anomaly detection in power generation plants using similarity-based modeling and multivariate analysis," in *Proc. Amer. Control Conf.*, Jun. 2011, pp. 1940–1945.

[30] H. Wang, M.-J. Peng, Y. Yu, H. Saeed, C.-M. Hao, and Y.-K. Liu, "Fault identification and diagnosis based on KPCA and similarity clustering for nuclear power plants," *Ann. Nucl. Energy*, vol. 150, Jan. 2021, Art. no. 107786.

[31] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1569–1576, 2004.

[32] K. R. Zalik, "An efficient K-means clustering algorithm," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1385–1391, 2008.

[33] Y. Zhang, Z. Y. Dong, W. Kong, and K. Meng, "A composite anomaly detection system for data-driven power plant condition monitoring," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4390–4402, Jul. 2020.



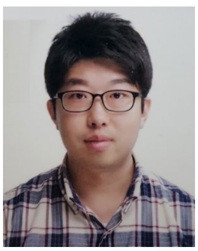
MUNWON LIM received the Ph.D. degree in industrial engineering from Hanyang University, Seoul, South Korea, in 2024. She is currently a Postdoctoral Researcher with Hanyang University. Her current research interests include anomaly detection based on signal and image processing, prognostics, and health management of big-sized sensor data.

YUNHEE KIM received the bachelor's degree in industrial management engineering from the Hankuk University of Foreign Studies, Yongin, South Korea. She is currently pursuing the Ph.D. degree with Hanyang University, Seoul, South Korea. Her research interests include reliability engineering and abnormal detection of large complex systems.

SHENGXIN JIN received the bachelor's degree in applied physics from South China University of Technology, Guangzhou, China, in 2022. She is currently pursuing the master's degree with Hanyang University, Seoul, South Korea. She has a keen interest in research areas that encompass statistics, reliability engineering, and big data analysis.

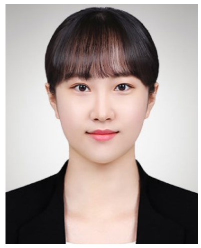
SANGWOO HA received the bachelor's degree in industrial management engineering from Kyonggi University, Suwon, South Korea. He is currently pursuing the M.S. degree with Hanyang University, Seoul, South Korea. His research interests include reliability engineering, statistics, and warranty analysis.

SUNG YONG CHANG received the M.S. degree in material science and engineering from Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2004. He is a Principal Researcher with Korea Electric Power Cooperation Research Institute (KEPCO RI), where he joined in 2004. His research interests include the process/control analysis and simulation, and furthermore, development/implementation of operational optimization solutions for gas turbine based combined cycle power plant, life assessment of boiler tubes of fossil power plant, and rejuvenation of hot gas path components of gas turbine. He received the National Academy of Engineering of Korea (NAEK) Next Generation Award, in 2017, for work on gas turbine super alloys.



HAE SU KANG received the M.S. degree in mechanical system engineering from Chungnam National University, Daejeon, South Korea, in 2016. From 2012 to 2018, he was a Researcher with Korea Electric Power Cooperation Research Institute (KEPCO RI), where he has been a Senior Researcher, since 2019. His research interests include the process/control analysis and simulation, and furthermore, development/implementation of operational optimization solutions for gas turbine based combined cycle.

GYU SANG PARK received the M.S. degree in aerospace engineering from Jeonbuk National University, Jeonju, South Korea, in 2017. From 2018 to 2020, he was a Researcher with Korea Electric Power Corporation Research Institute (KEPCO RI), where he has been a Senior Researcher, since 2021. He has engaged in designing thermodynamic models for combined cycle gas turbine engines and predicting their performance. Additionally, he is conducting research on using artificial intelligence to predict and diagnose anomalies in power generation equipment.



MI LEE JOO received the M.S. degree in aerospace engineering from Jeonbuk National University, Jeonju, South Korea, in 2019. Since 2021, she has been a Researcher with Korea Electric Power Cooperation Research Institute (KEPCO RI). Her research interests include the process analysis and simulation, and furthermore, the gas turbine combined cycle power plant power prediction.

CHANG-SIK JUNG received the master's degree in industrial engineering from Hanyang University, in 1997. After the graduation, he was with Hyundai Information Technology Company, where he was a Department Head, from 1997 to 2013. Then, he joined Hyosung ITX, as a Division Chief and the Executive Director, from 2014 to 2019. Since 2020, he has been the Executive Director with Rozetatech.

YOUNGJIN CHO graduated from the Department of Industrial Engineering, Hanyang University, in 1991. He began his career in 1985, developing computerization programs for general hospitals. In 1991, he founded Daeshin Computer and Hard & Soft. In 1993 (two years later), he established Jungsoft, followed by the founding of Jungsoft USA Inc., Silicon Valley, USA, in 1998. In 2015, he founded Rozetatech. In 2020, he took on the role of the IoT Wireless Subcommittee Chairman with Korea Fire Industry Association. As of 2021, he is the Vice President of the Korean Society of Industrial and Systems Engineering.

SUK JOO BAE (Member, IEEE) received the Ph.D. degree from the School of Industrial and Systems Engineering, Georgia Institute of Technology, in 2003. From 1996 to 1999, he was a Reliability Engineer with Samsung SDI, South Korea. He is currently a Professor with the Department of Industrial Engineering, Hanyang University, Seoul, South Korea. His research interests include reliability evaluation of light displays and nano devices via accelerated life and degradation testing, statistical robust parameter design, and process control for large-volume on-line processing data. He is a member of INFORMS, ASA, and IMS. He is an Associate Editor of IEEE TRANSACTIONS ON RELIABILITY.