Systems & Control Transactions

Research Article
Foundations of Computer Aided Process Design (FOCAPD 2024)
Breckenridge, Colorado, USA. July 14-18, 2024
Peer Reviewed Conference Proceeding

PSE PRESS

Fast, Accurate, and Robust Fault Detection and Diagnosis of Industrial Processes

Alireza Miraliakbar[a], Zheyu Jiang[a,*]

[a] School of Chemical Engineering, Oklahoma State University, Stillwater, Oklahoma, USA, 74078
* Corresponding Author: zheyu.jian@okstate.edu

ABSTRACT

Modern industrial processes are continuously monitored by a large number of sensors. Despite having access to large volumes of historical and online sensor data, industrial practitioners still face challenges in the era of Industry 4.0 in effectively utilizing them to perform online process monitoring and fast fault detection and diagnosis. To target these challenges, in this work, we present a novel framework named "FARM" for Fast, Accurate, and Robust online process Monitoring. FARM is a holistic monitoring framework that integrates (a) advanced multivariate statistical process control (SPC) for fast anomaly detection of nonparametric, heterogeneous data streams, and (b) modified support vector machine (SVM) for accurate and robust fault classification. Unlike existing general-purpose process monitoring frameworks, FARM's unique hierarchical architecture decomposes process monitoring into two fault detection and diagnosis, each of which is conducted by targeted algorithms. Here, we test and validate the performance of our FARM monitoring framework on Tennessee Eastman Process (TEP) benchmark dataset. We show that SPC achieves faster fault detection speed at a lower false alarm rate compared to state-of-the-art benchmark fault detection methods. In terms of fault classification diagnosis, we show that our modified SVM algorithm successfully classifies 17 out of 20 of the fault scenarios present in the TEP dataset. Compared with the results of standard SVM trained directly on the original dataset, our modified SVM improves the fault classification accuracy significantly.

Keywords: Fault Detection and Diagnosis, Process Monitoring, Statistical Process Control, Riemannian Manifold, Support Vector Machine

INTRODUCTION

Safe and efficient operation of an industrial plant depends on effective, continuous process monitoring (e.g., fault detection and diagnosis), which is enabled by advanced sensory systems that continuously generate streams of data to dictate the state of the plant. Despite having access to large volumes of historical and online sensor data, challenges remain in how these data could be used for effective online process monitoring. Existing techniques for process monitoring are inadequate because (a) fault scenarios in industrial systems and plants are complex, (b) sensors continuously produce massive arrays of big data streams that are often nonparametric (i.e., data streams may not follow any specific distribution) and heterogeneous (i.e., data streams may not follow the same distribution), and (c) there is an intrinsic trade-off between fault detection time and diagnostic accuracy.

To address this need, several process monitoring solutions have been developed over the past decades. Among them, dimensionality reduction techniques, such as principal component analysis (PCA), partial least squares (PLS) regression, as well as their different variations, are the most popular ones in the literature [1–3]. Dimensionality reduction techniques assume that the statistics characterizing the in-control profiles also span the subspace where out-of-control states (faults) lie in [4]. However, this assumption is generally invalid for industrial process monitoring as the process dynamics are quite complex and out-of-control states cannot be fully enumerated a priori. Also, plant operators often find it

difficult to interpret the results from PCA/PLS-based methods because the features are in the reduced space and do not have one-to-one mapping to the original sensor data sources. In addition, monitoring only the most significant subset of features often causes significant errors, as the fault may not be noticeable in the selected features. Lastly, dimensionality reduction techniques have no statistical guarantee on false alarm rate, making them unreliable for actual plant monitoring which requires false alarm to be low and controlled (e.g., ≤0.0027, the classic three-sigma limit) due to the significant money loss and safety issues of unplanned unit shutdown.

More recently, various machine learning (ML) tools such as support vector machine, decision tree, and deep neural network, have also been proposed and applied to process monitoring [5–8]. Nevertheless, existing ML methods still face problems such as overfitting and poor predictive accuracy. For example, while most published ML algorithms perform well during training and validation, their fault detection accuracies deteriorate and rarely exceed 90-95% in test sets. Considering the severe consequences in case of fault detection failure, such predictive accuracy is unacceptable. Furthermore, ML methods do not scale well with rare or new fault scenarios due to the lack of sufficient training data.

To target these challenges, in this work, we present a novel industrial process monitoring tool, which we named it as "FARM", for fast, accurate, and robust online fault detection and diagnosis. FARM is a holistic monitoring framework that integrates (a) advanced multivariate statistical process control (SPC) for fast anomaly detection of nonparametric, heterogeneous data streams, and (b) a modified support vector machine (SVM) for accurate and robust fault classification. Unlike existing general-purpose process monitoring frameworks, FARM's unique hierarchical architecture (see Figure 1) decomposes process monitoring into two fault detection and diagnosis, each of which is conducted by targeted algorithms. Only if a process anomaly is detected will the online data be sent to the fault classification/diagnosis module for accurate fault classification. Such hierarchical architecture successfully bypasses the intrinsic trade-off between fault detection speed and accuracy that is present in existing monitoring tools. Furthermore, using FARM, plant operators can choose a user-specified false alarm rate based on their expert knowledge of the process.

STRUCTURE AND WORKFLOW OF FARM

As mentioned earlier, FARM consists of two distinct yet interconnected modules. The first module performs fault detection by adopting the state-of-the-art quantile-based non-parametric SPC proposed by Ye and Liu [9]. Quantile-based nonparametric SPC can detect any process mean shift or anomaly from heterogeneous high-dimensional sensor data streams as early as possible while maintaining a pre-specified incontrol average run length. Inspired by the work of Smith et al. [10], 2the second module conducts fault classification through a modified SVM model. Both modules are connected as shown in Figure 1. FARM's workflow contains two steps: (1) offline training with historical data, followed by (2) online monitoring of real-time sensor data streams. During offline training, the parameters of the SPC module to be used for online monitoring are obtained using the historical in-control data. Also, the modified SVM module is trained by treating the faulty data's covariance matrices as features and the corresponding faulty scenario as labels.

Once offline training of FARM is complete, online sensor measurements will continuously be sent to FARM for simultaneous fault detection and diagnosis. First, they are monitored by the SPC module to detect any process anomaly in real time. Only if a process anomaly is detected will the online data be sent to the fault diagnosis module for accurate fault classification. Unlike general-purpose process monitoring frameworks, FARM's hierarchical architecture decomposes process monitoring tasks into two subtasks (fault detection and diagnosis), each of which is accomplished by specialized techniques. This allows fast, accurate, and robust fault detection and diagnosis to be simultaneously accomplished by FARM.

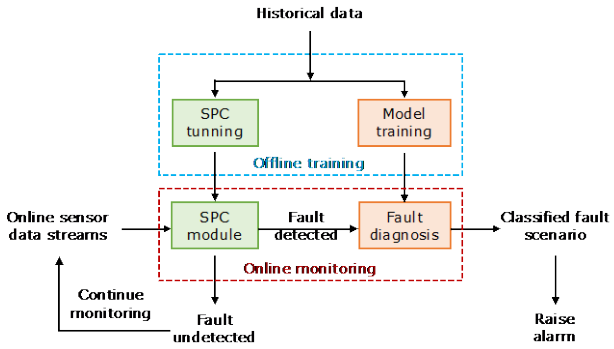

Figure 1. FARM's hiearchical structure consisting of fault detection and diagnosis modules.

Fault Detection

The backbone of FARM's fault detection module is the quantile-based non-parametric SPC algorithm proposed by Ye and Liu [9]. Jiang modified the original quantile-based SPC formulation of Ye and Liu [9] to monitor fully observable data streams [11]. Here, a brief description of the modified SPC formulation is presented. In offline training, the sensor measurements in each of the M historical in-control data streams X_j $(j = 1,2,\dots,M)$ are sorted in ascending order and partitioned into d number of quantiles $I_{j,1},\dots,I_{j,d}$ defined as:

$$I_{j,1} = (-\infty, q_{j,1}], I_{j,2} = (q_{j,1}, q_{j,2}], \dots, I_{j,d} = (q_{j,d-1}, +\infty) \quad (1)$$

For each $q_{j,i}$, two intervals called positive and negative cumulative intervals are defined as:

$$CI_{j,i}^{+} = [q_{j,i}, +\infty) \quad \text{and} \quad CI_{j,i}^{-} = (-\infty, q_{j,i}], \quad (2)$$

for every $i = 1, \dots, d-1$ and $j = 1,2, \dots, M$. With these positive/negative cumulative intervals identified from historical in-control data, one can detect anomalies in real time by detecting any upward/downward mean shift of online sensor data streams. To do this, for an online sensor data stream $X_j(t)$ where t stands for time, we define a binary variable $A_{j,i\in[1,\dots,d-1]}^{+}$ and $A_{j,i\in[1,\dots,d-1]}^{-}$ to indicate which positive and negative cumulative interval $X_j(t)$ lies in at time t, respectively:

$$A_{j,i\in[1,2,\dots,d-1]}^{+} = \begin{cases} 1 & \text{if } X_j(t) \in CI_{j,i}^{+} \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

$$A_{j,i\in[1,2,\dots,d-1]}^{-} = \begin{cases} 1 & \text{if } X_j(t) \in CI_{j,i}^{-} \\ 0 & \text{otherwise} \end{cases}. \quad (4)$$

With this, we obtain two vectors $\mathbf{A}_j^{+}(t)$ and $\mathbf{A}_j^{-}(t)$ as:

$$\mathbf{A}_j^{+}(t) = [A_{j,1}^{+}, A_{j,2}^{+}, \dots, A_{j,d-1}^{+}], \quad (5)$$

$$\mathbf{A}_j^{-}(t) = [A_{j,1}^{-}, A_{j,2}^{-}, \dots, A_{j,d-1}^{-}]. \quad (6)$$

One can show that $\mathbb{E}[\mathbf{A}_j^{+}(t)] = \left[1 - \frac{1}{d}, 1 - \frac{2}{d}, \dots, 1 - \frac{d-1}{d}\right]$ and $\mathbb{E}[\mathbf{A}_j^{-}(t)] = \left[\frac{1}{d}, \frac{2}{d}, \dots, \frac{d-1}{d}\right]$ for $j = 1, \dots, M$ and $i = 1, \dots, d$. Therefore, by defining $\mathbf{A}_j^{+}(t)$ and $\mathbf{A}_j^{-}(t)$, the idea is to convert the task of detecting any mean shift in the distribution of $X_j(t)$ with respect to the distribution of historical in-control data into an equivalent task of detecting the upward (resp. downward) mean shift in the distribution of $A_{j,i}^{+}$ (resp. $A_{j,i}^{-}$) with respect to $\mathbb{E}[A_{j,i}^{+}]$ (resp. $\mathbb{E}[A_{j,i}^{-}]$). This transformation presents at least two major advantages. First, it has been shown that $A_{j,i}^{+}$ (resp. $A_{j,i}^{-}$) is more sensitive to upward (resp. downward) mean shifts than the original data streams themselves [9], thus allowing faster fault detection. And second, it allows nonparametric, heterogeneous data streams to be successfully monitored for the first time.

Quantile-based SPC implements the multivariate cumulative sum (CUSUM) procedure first proposed by Qiu and Hawkins [12, 13] to monitor multivariate big data streams of $\mathbf{A}_j^{+}(t)$ and $\mathbf{A}_j^{-}(t)$ for $j = 1, \dots, M$. This is achieved by defining $C_j^{+}(t)$ and $C_j^{-}(t)$ as:

$$C_j^{\pm}(t) = \left[\left(\boldsymbol{S}_j^{\pm,\text{obs}}(t-1) + \boldsymbol{A}_j^{\pm}(t)\right) - \left(\boldsymbol{S}_j^{\pm,\text{exp}}(t-1) + \mathbb{E}(\boldsymbol{A}_j^{\pm}(t)\right)\right]^{T} \cdot \left(\text{diag}\left(\boldsymbol{S}_j^{\pm,\text{exp}}(t-1) + \mathbb{E}(\boldsymbol{A}_j^{\pm}(t)\right)\right)^{-1} \cdot \left[\left(\boldsymbol{S}_j^{\pm,\text{obs}}(t-1) + \boldsymbol{A}_j^{\pm}(t)\right) - \left(\boldsymbol{S}_j^{\pm,\text{exp}}(t-1) + \mathbb{E}(\boldsymbol{A}_j^{\pm}(t)\right)\right] \quad (7)$$

In Equation (7), $\boldsymbol{S}_j^{\pm,\text{obs}}(t)$ and $\boldsymbol{S}_j^{\pm,\text{exp}}(t)$ are four vectors of size $d-1$ that are the CUSUM statistics initiated at $\boldsymbol{S}_j^{\pm,\text{obs}}(t=0) = \boldsymbol{S}_j^{\pm,\text{exp}}(t=0) = 0$:

$$\begin{cases} \boldsymbol{S}_j^{\pm,\text{obs}}(t) = 0, \boldsymbol{S}_j^{\pm,\text{exp}}(t) = 0, & \text{if } C_j^{\pm}(t) \le k \\ \boldsymbol{S}_j^{\pm,\text{obs}}(t) = \frac{\left(C_j^{\pm}(t)-k\right)}{C_j^{\pm}(t)}\left(\boldsymbol{S}_j^{\pm,\text{obs}}(t-1) + \boldsymbol{A}_j^{\pm}(t)\right) & \\ \boldsymbol{S}_j^{\pm,\text{exp}}(t) = \frac{\left(C_j^{\pm}(t)-k\right)}{C_j^{\pm}(t)}\left(\boldsymbol{S}_j^{\pm,\text{exp}}(t-1) + \mathbb{E}(\boldsymbol{A}_j^{\pm}(t)\right) & \text{if } C_j^{\pm}(t) > k \end{cases} \quad (8)$$

In Equation (8), k is an allowance parameter that restarts the CUSUM procedure if no evidence of significant shift is detected after a while [14]. The value of k is obtained during offline training using historical in-control data. Then, one-sided local statistics W_j^{+} and W_j^{-} for respectively detecting upward and downward mean shifts of data stream j can be defined as:

$$W_j^{+}(t) = \max\left(0, C_j^{+}(t) - k\right), \quad (9)$$

$$W_j^{-}(t) = \max\left(0, C_j^{-}(t) - k\right). \quad (10)$$

If one wants to detect either upward or downward mean shifts, then a two-sided local statistic $W_j(t)$ can be defined as the maximum of the two one-sided local statistics:

$$\begin{cases} W_j(t=0) = 0, \\ W_j(t>0) = \max\left(W_j^{+}(t), W_j^{-}(t)\right). \end{cases} \quad (11)$$

Finally, to determine the stopping time T for raising the alarm by declaring the process is out-of-control, the top-r approach proposed by Mei [15] is adopted. First, at each time step t, the values of individual local statistics $W_j(t)$ for all data streams are ranked from largest to smallest: $W_{(1)}(t) > \cdots > W_{(k)}(t) > \cdots > W_{(M)}(t)$, in which $W_{(k)}(t)$ corresponds to the k^{th} largest local statistic. Next, the top r of the local statistics at time t is calculated, and the stopping time T, also known as the out-of-control run length, is defined as:

$$T = \inf\{t > 0 : \textstyle\sum_{(k)=1}^{r} W_{(k)}(t) \ge h\}, \quad (12)$$

where h is a threshold value that corresponds to the prespecified false alarm rate and can be obtained during offline training using historical in-control data. A commonly used h is obtained based on the false alarm rate of 0.27% (the classic 3σ limit).

Fault Classification and Diagnosis

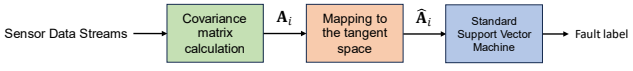

Figure 2. Flowchart of the modified SVM algorithm for improved facult classifcation.

In this section, we discuss how accurate fault

diagnosis can be achieved using a modified SVM module in FARM. Figure 2 illustrates how we modify standard SVM for fault classification by adding a data pre-processing step in the training step. To train the SVM model using the historical sensor data corresponding to different fault scenarios, we first compute the covariance matrix of the historical faulty data streams, followed by training the SVM model over the covariance matrix instead of the original faulty data streams. This modification is inspired by the fact that covariance matrices are symmetric and positive definite, and thus always lie on a Riemannian manifold. It has been recently shown that, by respecting this important geometric insight, one can greatly enhance the accuracy and interpretability of classification, regression, dimensionality reduction algorithms by conducting these computations on the tangent space of the manifold [10]. Inspired by this finding, we map the generated covariance matrices to their tangent space, which intersect the Riemannian manifold where these covariance matrices reside at the geometric mean of the covariance matrices (see Figure 3). This mapping is done through the logarithm operation as:

$$\widehat{\mathbf{A}}_i = \log_{\bar{A}}(\mathbf{A}_i), \tag{13}$$

where $\mathbf{A}_i$ is the covariance matrix of sensor data streams for dataset i calculated as:

$$\mathbf{A}_i = \frac{1}{N-1}\mathbf{X}_i\mathbf{X}_i^T, \tag{14}$$

where $\mathbf{X}_i$ is the original sensor data matrix containing M number of data streams values over N time steps. $\bar{\mathbf{A}}$ is the geometric mean of covariance matrices ($\mathbf{A}_i$), and $\widehat{\mathbf{A}}_i$ is the mapped matrix of matrix $\mathbf{A}_i$ to the tangent space as shown on Figure 3. The reader is encouraged to read the main reference explaining this mathematical calculation if interested [10].

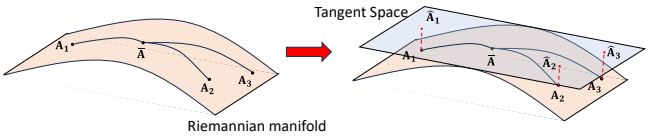

Figure 3. Illustration of (a) a Riemannian manifold and (b) the associated tangent space. The logarithmic map as well as the geodesic between the geometric mean $\bar{\mathbf{A}}$ and each covariance matrix $\mathbf{A}_i$ are also shown.

After this data preprocessing step, the mapped covariance matrices are used as input features, whereas the corresponding fault scenarios are used as labels to train a standard SVM model using a radial basis function (RBF) kernel.

During online monitoring stage, real-time sensor data streams are processed in the fault/anomaly detection module first. Only when a process anomaly is detected will the data streams be sent to the fault classification/diagnosis module. Such an arrangement will further enhance the accuracy and reliability of fault diagnosis module, as the data streams are certain to be faulty. Next, the covariance matrix for the sensor data streams is calculated, mapped to the tangent space of the Riemannian manifold, and used as the input to the trained SVM model to classify its fault label.

CASE STUDY: TENNESSEE EASTMAN PROCESS

Abstracted from a real chemical process, the Tennessee Eastman Process (TEP) is a nonlinear open-loop unstable process that has been widely used in various computational studies as benchmark case for plant-wide control, process monitoring, and data-driven optimization [16]. As shown in the schematic of Figure 4, the TEP consists of 4 major unit operations: a reactor, a stripping column, a separator, and a product condenser. The process involves the production of two liquid product components G and H from four gaseous reactants A, C, D and E with an additional inert B and a by-product F. The process is continuously monitored by a total of 52 process variables, including 11 manipulated and 41 measured variables.

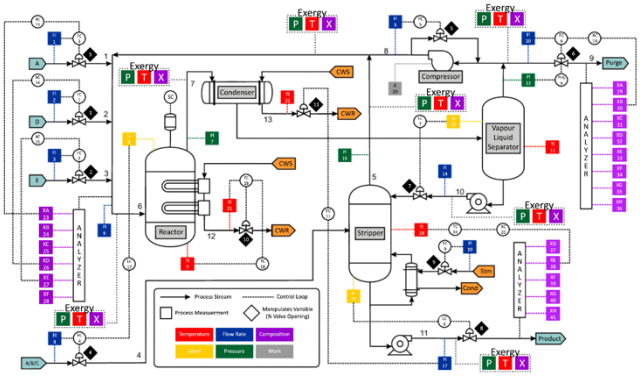

Figure 4. Schematic of TEP (figure extracted from [17]).

Fault Detection Module Performance

Table 1 lists the comparison results of our SPC module with respect to two benchmark fault detection algorithms, which are PCA-T^2 and SVM [11]. The data used for this study is obtained by the MATLAB graphical user interface (GUI) originally developed by Andersen et al. [18]. Overall, a total of 50 hours (simulation) of normal operation data were generated using this GUI to determine the threshold value h in Equation (12) and to construct the quantiles $I_{j,1}, \ldots, I_{j,d}$ as well as the cumulative intervals $CI_{j,i}^{\pm}$. In addition to normal operation (in-control) data, the

GUI can generate process data for 28 different fault scenarios. Here, we select three faults, namely IDV 2, 3, and 13 (see Table 1 for description), to compare the performance of the SPC algorithm with other benchmarks.

Table 1. Description of faults for comparison study of multiple fault detection benchmarks.

Fault #	Description	Fault Type
IDV 2	B composition in stream 4 with A/C ratio constant	Step
IDV 3	D feed temperature in stream 2	Step
IDV 13	Reaction kinetics	Slow drift

Table 2 summarizes the comparison results of fault detection speed and the corresponding false alarm rate of all three monitoring frameworks, quantified by out-of-control run length (i.e., how many additional observations are needed to declare out-of-control status and raise alarm after the actual fault is introduced) for each algorithm. As we can see, among the three monitoring frameworks, quantile-based SPC framework yields the fastest fault detection speed in all three fault scenarios, while maintaining the lowest false alarm rate. Given that a lower false alarm rate generally sacrifices fault detection speed due to more conservative monitoring behavior, the quantile-based SPC framework achieves a win-win situation compared to other benchmark algorithms.

Table 2. Fault detection results in terms of out-of-control run length (false alarm rate) for SPC , PCA-T^2 and SVM for TEP dataset [11].

Fault #	SPC	PCA-T^2	SVM
IDV 2	125 (0.27%)	216 (0.5%)	180 (0.8%)
IDV 3	95 (0.27%)	366 (0.5%)	16815 (83%)
IDV 13	128 (0.27%)	1131 (0.5%)	675 (12.7%)

Fault Diagnosis Module Performance

For fault diagnosis, we experimented various classification algorithms using the TEP dataset developed by Rieth et al. [19], which consists of 500 simulation cases of normal (in-control) operation as well as 20 fault scenarios. To illustrate, we present three representative models here.

First, we highlight the "best model" obtained by following training procedure illustrated in Figure 2. Figure 5 shows the confusion matrix obtained through 10-fold cross-validation of these 20 faults. Clearly, the modified SVM model demonstrated outstanding classification performance for all faults except for faults IDV 3, 9, and 15. This result outperforms a number of fault diagnosis algorithms in the literature. It is worth noting that faults IDV 3, 9, and 15 correspond to "step change in temperature of reactor feed D", "random variation in temperature of reactor feed D", and "sticking value failure for condenser cooling water valve", respectively. And these three faults are well-known to be particularly challenging to differentiate due to the close similarity of their dynamic behaviors to the overall process. To tackle this longstanding challenge of successful differentiation of these faults, new, creative methodologies need to be developed.

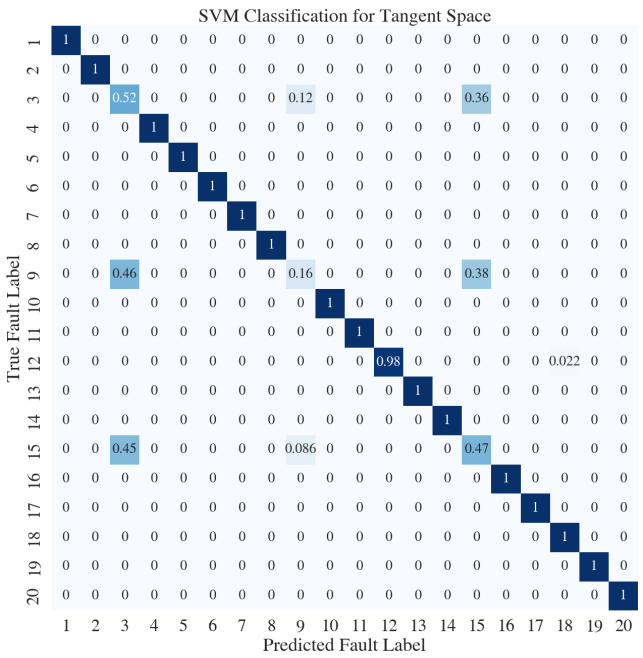

Figure 5. Confusion matrix (after 10-fold cross validation) of fault diagnosis results for our proposed modified SVM model.

As a direct comparison, Figure 6 shows the confusion matrix for the case where standard SVM without any data pre-processing is used for training and validation. It is clear that fault classification accuracy deteriorates significantly in this case.

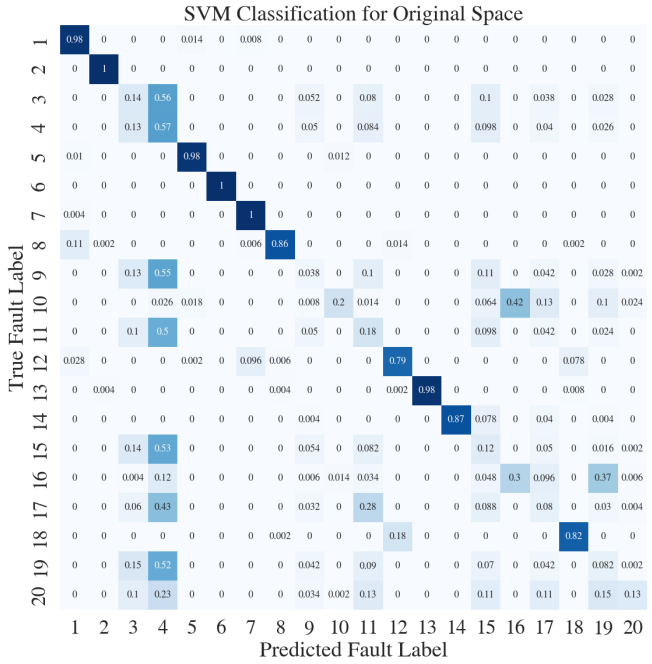

Figure 6. Confusion matrix (after 10-fold cross validation) of fault diagnosis results for standard SVM model without the introduced data pre-processing step.

Finally, we present the results for another fault diagnosis algorithm based on principal geodesic analysis (PGA) discussed by Smith et al. [10]. PGA is a counterpart of principal component analysis (PCA) applied on the tangent space of the Riemannian manifold, as it identifies the geodesics that capture the most variance in the data. In other words, in PGA, we simply apply PCA technique to the mapped covariance matrices of faulty data streams for dimensionality reduction. To determine the number of principal geodesics (which are the "principal components" in PGA) needed, we perform sensitivity analysis and identify that 29 principal geodesics are required to capture 99% of the variance in the original dataset containing covariance matrices on the Riemannian manifold. Furthermore, four distance measures, namely Euclidean, Mahalanobis, Manhattan, and Cosine are tested and compared. Clustering is done by assigning a point to its closest cluster based on the distance mesure used. We identify that, among these four measures, the cosine distance offers the best fault classification performance. Figure 7 shows the confusion matrix of PGA-cosine approach with the 10-fold cross validation. As we can see, in general, 12 out of the 20 faults can be fully classified, whereas faults IDV 3, 5, 9, 10, 12, 13, 15, and 18 cannot. Although its accuracy is yet to match with the best model, the PGA-Cosine algorithm performs much better than standard SVM without data pre-processing.

Since the modified SVM model showed superior performance over PGA-Cosine method, the confusion matrix of this method is compared with 10-fold confusion matrix of the ridge classifier model presented by Smith et al. [10], which trained on the mapped covariances. Figure 8 depicts the difference of confusion matrices between the modified SVM (A) and Ridge (B) classifiers. Positive numbers show that the prediction probability values of modified SVM model were higher than Ridge classifier. Conversely, the negative values indicate that the Ridge model's prediction probabilies were more than the modified SVM. Lastly, zero means that both models had the same prediction probability. As can be seen, both models have the same accuracy for all faults excepts faults 9 and 15, which the ridge classifier had a better accuracy than the modified SVM model by looking at the diagnoal values of the difference matrix.

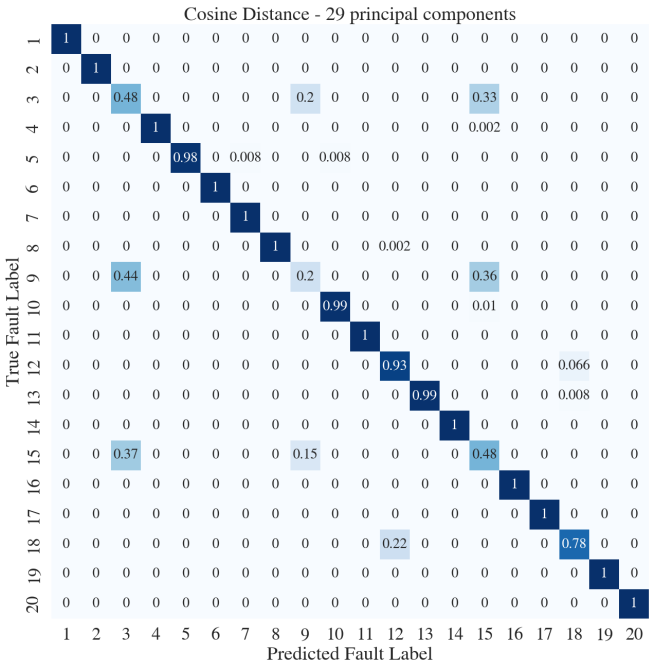

Figure 7. Confusion matrix (after 10-fold cross validation) of fault diagnosis results for PGA-Cosine classification algorithm with 29 principal components being selected.

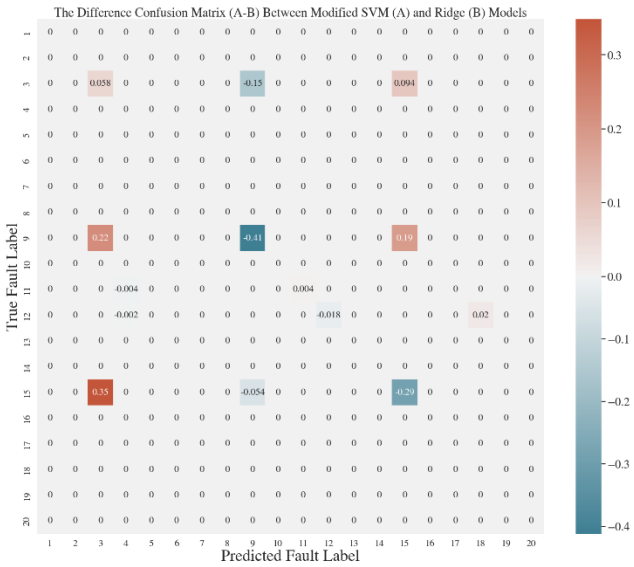

Figure 8. The difference confusion matrix between 10-fold cross validation matrices of the modified SVM classifier presented in this study (A) and the ridge classifier presented by Smith et al. (B) [10].

CONCLUSION

In this work, we present a fast, accurate, and robust algorithmic framework named FARM for industrial process monitoring. FARM is a holistic framework that synergistically performs fault detection and diagnosis tasks to improve monitoring performance. The fault detection module inside FARM adopts an advanced quantile-based SPC approach that can detect any mean shift of non-parametric and heterogenous multivariate data streams as soon as possible while maintaining a pre-specified false alarm rate. Meanwhile, the fault diagnosis module inside FARM implements a modified SVM algorithm for fault classification. Compared to standard SVM approach, our modified SVM algorithm includes an important data preprocessing step that makes use of the manifold insight of covariance matrix to greatly enhance classification accuracy. By validating and evaluating the performance of our FARM framework using the TEP dataset, we observe that 1) our fault detection module can achieve fast anomaly detection speed at a low false alarm rate, and 2) our fault diagnosis module successfully classifies 17 out of 20 fault scenarios at 100% accuracy. Unfortunately, faults IDV 3, IDV 9, and IDV 15 of the TEP dataset, which are known to be hard to classify, still face challenges in differentiating among one another with high accuracy. Our future work involves revamping the FARM framework to improve the classification accuracy of these hard-to-differentiate faults.

ACKNOWLEDGEMENTS

We gratefully acknowledge financial support from Oklahoma State University College of Engineering, Architecture, and Technology's Startup Fund No. 1-155160 and from the National Science Foundation award TI-2331080.

REFERENCE

1. Jackson, J.E., Mudholkar, G.S.: Control Procedures for Residuals Associated with Principal Component Analysis. Technometrics. 21, 341–349 (1979). https://doi.org/10.2307/1267757
2. Geladi, P., Kowalski, B.R.: Partial least-squares regression: a tutorial. Anal. Chim. Acta. 185, 1–17 (1986). https://doi.org/10.1016/0003-2670(86)80028-9
3. Fezai, R., Mansouri, M., Taouali, O., Harkat, M.F., Bouguila, N.: Online reduced kernel principal component analysis for process monitoring. J. Process Control. 61, 1–11 (2018). https://doi.org/10.1016/j.jprocont.2017.10.010
4. Woodall, W.H., Spitzner, D.J., Montgomery, D.C., Gupta, S.: Using Control Charts to Monitor Process and Product Quality Profiles. J. Qual. Technol. 36, 309–320 (2004). https://doi.org/10.1080/00224065.2004.11980276
5. Zhao, H., Hu, Y., Ai, X., Hu, Y., Meng, Z.: Fault detection of Tennessee Eastman process based on topological features and SVM. IOP Conf. Ser. Mater. Sci. Eng. 339, 012039 (2018). https://doi.org/10.1088/1757-899X/339/1/012039
6. Onel, M., Kieslich, C.A., Pistikopoulos, E.N.: A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: Application to the Tennessee Eastman process. AIChE J. 65, 992–1005 (2019). https://doi.org/10.1002/aic.16497
7. Chebel-Morello, B., Malinowski, S., Senoussi, H.: Feature selection for fault detection systems: application to the Tennessee Eastman process. Appl. Intell. 44, 111–122 (2016). https://doi.org/10.1007/s10489-015-0694-6
8. Heo, S., Lee, J.H.: Fault detection and classification using artificial neural networks. 10th IFAC Symp. Adv. Control Chem. Process. ADCHEM 2018. 51, 470–475 (2018). https://doi.org/10.1016/j.ifacol.2018.09.380
9. H. Ye, K. Liu: A Generic Online Nonparametric Monitoring and Sampling Strategy for High-Dimensional Heterogeneous Processes. IEEE Trans. Autom. Sci. Eng. 19, 1503–1516 (2022). https://doi.org/10.1109/TASE.2022.3146391
10. Smith, A., Laubach, B., Castillo, I., Zavala, V.M.: Data analysis using Riemannian geometry and

applications to chemical engineering. Comput. Chem. Eng. 168, 108023 (2022). https://doi.org/10.1016/j.compchemeng.2022.108023
11. Jiang, Z.: Online Monitoring and Robust, Reliable Fault Detection of Chemical Process Systems. In: Kokossis, A.C., Georgiadis, M.C., and Pistikopoulos, E. (eds.) Computer Aided Chemical Engineering. pp. 1623–1628. Elsevier (2023)
12. Qiu, P., Hawkins, D.: A Rank-Based Multivariate CUSUM Procedure. Technometrics. 43, 120–132 (2001)
13. Qiu, P., Hawkins, D.: A Nonparametric Multivariate Cumulative Sum Procedure for Detecting Shifts in All Directions. J. R. Stat. Soc. Ser. Stat. 52, 151–164 (2003)
14. Xian, X., Zhang, C., Bonk, S., Liu, K.: Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation. J. Qual. Technol. 53, 135–153 (2021). https://doi.org/10.1080/00224065.2019.1681924
15. Y. Mei: Quickest detection in censoring sensor networks. In: 2011 IEEE International Symposium on Information Theory Proceedings. pp. 2148–2152 (2011)
16. Downs, J.J., Vogel, E.F.: A plant-wide industrial process control problem. Ind. Chall. Probl. Process Control. 17, 245–255 (1993). https://doi.org/10.1016/0098-1354(93)80018-I
17. Hu, M., Hu, X., Deng, Z., Tu, B.: Fault Diagnosis of Tennessee Eastman Process with XGB-AVSSA-KELM Algorithm. Energies. 15, (2022). https://doi.org/10.3390/en15093198
18. Andersen, E.B., Udugama, I.A., Gernaey, K.V., Khan, A.R., Bayer, C., Kulahci, M.: An easy to use GUI for simulating big data using Tennessee Eastman process. Qual. Reliab. Eng. Int. 38, 264–282 (2022). https://doi.org/10.1002/qre.2975
19. Rieth, C.A., Amsel, B.D., Tran, R., Cook, M.B.: Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation, https://doi.org/10.7910/DVN/6C3JR1, (2017)