Dynamic Bhattacharyya Bound-Based Approach for Fault Classification in Industrial Processes

Benben Jiang, *Member, IEEE*, and Bofan Zhu

***Abstract*—Data-driven fault diagnosis has attracted increasing research interest with a recent trend of aiming at large-scale and complex systems. In this article, we propose a method under a probabilistic framework, named dynamic Bhattacharyya bound (DBB), to extract features for fault diagnosis. An information criterion is adopted to determine the order of dimensionality reduction and time lags when applying the proposed approach. Compared with conventional diagnostic approaches, the proposed DBB approach has several advantageous features. First, the DBB approach minimizes an upper bound of the Bayes error which is a direct manifestation of the misclassification rate. Second, pairwise Bhattacharyya bounds between different faults are summed up in the objective function, enabling it to address the fault diagnosis of multiple faults that may have large overlaps. The proposed method is validated through the Tennessee Eastman process and it shows advantageous performance than other methods such as Fisher discriminant analysis (FDA), dynamic FDA, and LP-DFDA.**

***Index Terms*—Bayes error, dimensionality reduction (DR), dynamic Bhattacharyya bound (DBB), fault classification, Tennessee Eastman process (TEP).**

I. Introduction

FAULT diagnosis is an essential procedure to guarantee safe and efficient operations of industrial processes. With an increasing demand for production efficiency, energy conservation, and environmental protection, modern industry presents a new trend featured by a large amount of operation units and complex interactions among interconnected units. As a result, the occurrence of process faults may lead to huge economic losses and severe safety issues due to the propagation of their effects to downstream and neighboring units. Therefore, it is critical to detect and diagnose the faults timely with process monitoring techniques, including fault diagnosis, to ultimately improve the reliability and safety of the entire process plant [1]–[4].

Specifically, for fault diagnosis (or equivalently, fault classification), the objective is to classify the types of faults that have already occurred, so that corresponding maintenance endeavors can be deployed to recover the system from downtime. Among the many fault diagnosis techniques, the data-driven approach plays an important role for large-scale and complex industry processes, where traditional methods based on qualitative and empirical knowledge become impractical given the unavailability of physical models for such large systems. On the other hand, the accessibility of large amount of historical data reflecting the operation mechanism and state makes it possible to extract fault signatures without the necessity of physical models. In data-driven fault diagnosis, the data produced under different fault types are divided into different categories, and the types of faults can be identified by classifying their data to the corresponding categories. Principal component analysis (PCA) [5], [6], Fisher discriminant analysis (FDA) [8], [9], partial least squares (PLS) [10] are basic data-driven approaches to improve the proficiency of fault diagnosis via dimensionality reduction (DR). The PCA method reduces the dimensionality of the raw data by seeking several principal coordinates along which the covariance of principal components is maximized. Thus, in PCA, the major variations in the raw data are reserved and small variations such as noise are discarded. The FDA method defines a between-class scatter matrix characterizing the variations among different fault classes, and a within-class scatter matrix representing the variations within fault classes. Optimal projection vectors are obtained by maximizing the scatter between classes while minimizing the scatters within classes. The PLS method stores fault information in the dependent matrix and process variables in the independent matrix. This method finds optimal loading vectors that maximize the covariance between fault matrix and process variable matrix. Thus, the correlation between predicted variables and predictors is maximized and the large variations within the predictors are reserved simultaneously. For all the methods mentioned above, after DR, discriminant functions are constructed in different ways for fault classification [8], [11], [12]. In addition, data-driven methods have also been widely used in industrial cyber-physical systems [7], [23], together with extensive development of MATLAB toolbox [26].

Manuscript received January 1, 2021; revised January 15, 2021; accepted January 24, 2021. Date of publication February 3, 2021; date of current version September 29, 2021. Paper no. TII-21-0005. *(Corresponding author: Benben Jiang.)*

Benben Jiang is with Tsinghua University, Beijing 100084, China (e-mail: bbjiang@tsinghua.edu.cn).

Bofan Zhu is with the Beijing University of Chemical Technology, Beijing 100029, China (e-mail: zhubofan21@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2021.3056533.

Digital Object Identifier 10.1109/TII.2021.3056533

In statistical process monitoring, the discriminant analysis methods are carried out assuming the independent and identical distribution (i.i.d.) of samples. Such an assumption only holds when the sample interval is sufficiently large, which happens only to processes with slow dynamics [13]. However, most industrial data have strong serial correlations and neglecting such correlation information may lead to increased false detection and diagnosis results. To handle these scenarios, FDA-, PCA- and PLS-based methods are enhanced by stacking past variable values with current variable values [14]. In this way, serial correlation information can be captured when performing DR [15]. In addition, the overlaps between different types of fault data can be reduced, which improves the performance of fault diagnosis [15].

In general, any decision rule has a probability of classifying a sample to a wrong class due to the randomness of samples. Bayes error, the probability that a sample is assigned to the wrong class [16], can evaluate the performance of a decision rule. It measures the lowest classification error associated with a classifier. The Bayes error always exists due to the irreducible error in classification; in practice, different classes always have certain overlap in the true population. Thus, it is necessary to account for the Bayes error during DR to optimize the classifier performance. Based on the framework of Bayes error, a group of classification techniques, known as the Bayes error feature selection methods, have been proposed and implemented in a variety of fields, such as pattern recognition [16] and speech recognition [17]. Some of these methods are designed to find upper bounds for Bayes error, such as Chernoff bound [18] and Bhattacharyya bound (BB) [19]. Some of these methods are designed to find the diversity of data information, such as interclass divergence [20]. In other DR methods, such as FDA, selecting the projection vectors optimizes a criterion that maximizes the ratio between interclass and intraclass scatter matrices. However, such a criterion is not a direct quantification of the classification accuracy. In contrast, our proposed approach, during DR, selects the loadings that directly minimize the Bayes error of the subsequent discriminant function. This interpretation implies that our approach can yield a classification performance closer to the theoretical optimum than other DR methods. The criteria used by traditional DR methods may not be aligned with optimizing the Bayes error of the adopted discriminant function. Therefore, using the Bayes error bound to guide the design of classifiers for fault diagnosis is of great importance to directly improve the classification performance. Moreover, the BB is shown to have a closed-form expression for Gaussian variables. This motivates us to use the BB-based DR for achieving the optimal projection vectors for fault diagnosis of industrial processes.

In this article, an approach based on BB of Bayes error is proposed for diagnosing faults. The BB model is further extended by constructing the data matrix with lagged values of variables, regarded as dynamic BB (DBB), for better capturing the dynamic information in the data. Additionally, the optimal DR order and time lags of DBB are determined by utilizing an information criterion (IC).

The rest of this article is organized as follows. The BB method is briefly described in Section II. Section III describes the DBB-based fault diagnosis method. The validity of the DBB-based diagnostic method was confirmed by the Tennessee Eastman process (TEP) in Section IV. Finally, Section V concludes this article.

II. Revisit of BB Based on Bayes Error

Consider a general classification problem, in which we classify the d-dimensional input $\boldsymbol{x} \in R^d$ into C categories. The principle of classification for class i consists of the prior probability λ_i and the probability density function $p_i(\mathbf{x}), p = 1, \ldots, C$. Observation $\boldsymbol{x}$ is classified into class j if $j = \arg\max_{1\le i\le C}\lambda_i p_i(\mathbf{x})$. The error rate of this classifier is called Bayes error and is defined as follows:

$$\varepsilon = 1 - \int_{R^d} \max_{1\le i\le C} \lambda_i p_i(\boldsymbol{x}) d\boldsymbol{x}. \tag{1}$$

Assume that there is a projection vector $\boldsymbol{w}$ that projects the d-dimensional vector x of class i into a lower k-dimensional space: $\boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x}$, with $\boldsymbol{w}$ a $k \times d$ matrix of rank $k \le d$. After DR, the prior probability of class i remains the same but the density function in reduced space is changed to $p_i^{\mathbf{w}}(\mathbf{y})$. The Bayes error then becomes

$$\varepsilon_{\boldsymbol{w}} = 1 - \int_{R^p} \max_{1\le i\le C} \lambda_i p_i{}^{\boldsymbol{w}}(\boldsymbol{y}) d\boldsymbol{y}. \tag{2}$$

Theoretically, DR, in general, induces some information loss since the mapped variable $\boldsymbol{y}$ in reduced space cannot preserve all information of the original variable $\boldsymbol{x}$. Correspondingly, one usually has $\varepsilon \le \varepsilon_{\boldsymbol{w}}$. For a given proper dimension p, the feature extraction problem can be transformed into an optimization problem of finding the optimal projection vector by

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in R^{k\times d},\ \mathrm{rank}(\boldsymbol{w})=k}{\arg\min} \varepsilon_{\boldsymbol{w}}. \tag{3}$$

However, the Bayer error ε is hard to optimize and instead, an upper bound of ε BB is utilized to solve the feature extraction problem above.

Now let us prove the following upper bound for the Bayes error:

$$\varepsilon \le \sum_{1\le i\le j\le C} \sqrt{\lambda_i \lambda_j} \int_{R^d} \sqrt{p_i(\boldsymbol{x}) p_j(\boldsymbol{x})} d\boldsymbol{x}. \tag{4}$$

Proof: The left equation can be written as [16], [19]

$$\begin{aligned}\varepsilon &= \int_{R^d} \sum_{i=1}^{C} \lambda_i p_i(\boldsymbol{x}) d\boldsymbol{x} - \int_{R^d} \max_{1\le i\le C} \lambda_i p_i(\boldsymbol{x}) \\ &= \int_{R^d} \min_{1\le i\le C} \sum_{j\ne i}^{C} \lambda_j p_j(\boldsymbol{x}) d\boldsymbol{x}.\end{aligned} \tag{5}$$

There exists a permutation of the indices $\sigma_{\boldsymbol{x}}$: $\{1, \ldots, C\} \to \{1, \ldots. C\}$ so that the terms $\lambda_1 p_1(\mathbf{x}), \ldots, \lambda_C p_C(\mathbf{x})$ are listed in increasing order as $\lambda_{\sigma_{\mathbf{x}}(1)} p_{\sigma_{\mathbf{x}}(1)}(\mathbf{x}) \le \cdots \le \lambda_{\sigma_{\mathbf{x}}(C)} p_{\sigma_{\mathbf{x}}(C)}(\mathbf{x})$[16], [19]. Thus, for $1 \le k \le C - 1$

$$\lambda_{\sigma_{\boldsymbol{x}}(k)} p_{\sigma_{\boldsymbol{x}}(k)}(\boldsymbol{x}) \le \sqrt{\lambda_{\sigma_{\boldsymbol{x}}(k)} p_{\sigma_{\boldsymbol{x}}(k)}(\boldsymbol{x})\, \lambda_{\sigma_{\boldsymbol{x}}(k+1)} p_{\sigma_{\boldsymbol{x}}(k+1)}(\boldsymbol{x})}. \tag{6}$$

Thus, (5) can be transformed into [16], [19]

$$\begin{aligned}\min_{1\le i\le C}\sum_{j\neq i}\lambda_i p_j(\boldsymbol{x}) &= \sum_{k=1}^{C-1}\lambda_{\sigma_{\boldsymbol{x}}(k)}p_{\sigma_{\boldsymbol{x}}(k)}(\boldsymbol{x})\\ &\le \sum_{k=1}^{C-1}\sqrt{\lambda_{\sigma_{\boldsymbol{x}}(k)}p_{\sigma_{\boldsymbol{x}}(k)}(\boldsymbol{x})\,\lambda_{\sigma_{\boldsymbol{x}}(k+1)}p_{\sigma_{\boldsymbol{x}}(k+1)}(\boldsymbol{x})}\\ &\le \sum_{1\le i\le j\le C}\sqrt{\lambda_i p_j(\boldsymbol{x})\,\lambda_j p_j(\boldsymbol{x})} \qquad (7)\end{aligned}$$

which verifies the inequality (4). ■

If we assume that the distribution is normal with means $\boldsymbol{\mu}_i$ and covariance $\sum_i$, (4) reduces to a simpler expression

$$\varepsilon \le \sum_{1\le i\le j\le C}\sqrt{\lambda_i\lambda_j}e^{-g(i,j)} \qquad (8)$$

where

$$\begin{aligned}g(i,j) &= \frac{1}{8}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)^T\left[\frac{\sum_i+\sum_j}{2}\right]^{-1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)\\ &\quad+\frac{1}{2}\log\left(\frac{\left|\frac{\sum_i+\sum_j}{2}\right|}{\sqrt{\left|\sum_i\right|\left|\sum_j\right|}}\right). \qquad (9)\end{aligned}$$

Equation (9) is called the Bhattacharyya distance between distributions p_i and p_j[3].

Similarly, we define the Bhattacharyya distance between $p_i^{\mathbf{w}}$ and $p_j^{\mathbf{w}}$in our projection space to be $g_{\boldsymbol{w}}(i,j)$. Combining (4) and (8), the Bayesian error rate (2) in the projection space is obtained as follows:

$$\varepsilon_{\boldsymbol{w}} \le \sum_{1\le i\le j\le C}\sqrt{\lambda_i\lambda_j}e^{-g_{\boldsymbol{w}}(i,j)}. \qquad (10)$$

Take the right-hand side of inequality (10) as an upper bound of the Bayes error $\varepsilon_{\boldsymbol{w}}$ in the projection space. Then, minimizing its upper bound will give rise to an approximate solution of the original problem.

III. Proposed BB Approach for Diagnosing Multiple Faults

Necessary assumption: We assume that each class of data is Gaussian distributed with means $\boldsymbol{\mu}_i$ and covariance $\sum_i$, where i represents the ith class (or fault).

In order to apply the strategy mentioned in the previous section to diagnose multiple faults, we need to rewrite the objective function into the following form for diagnosing C faults:

$$J_{\boldsymbol{w}} = \frac{2}{C(C-1)}\sum\nolimits_{1\le i\le j\le C}\sqrt{\lambda_i\lambda_j}e^{-g_w(i,j)}. \qquad (11)$$

When dealing with multiclass problems, the objective function (11) is a sum of all pairwise BBs [21] between faults. For the problem with many classes, as the number of terms in the objective function increases, the complexity increases accordingly. To find the local minimum for such a problem, the conjugate gradient method is employed in this article. Thus, the gradient of the objective function (11) with respect to $\boldsymbol{w}$ has to be derived. According to (9), we have

$$\begin{aligned}g_{\boldsymbol{w}}(i,j) &= \frac{1}{2}\mathrm{trace}\left\{\left(\boldsymbol{w}C_{ij}\boldsymbol{w}^T\right)^{-1}\boldsymbol{w}B_{ij}\boldsymbol{w}^T\right\}\\ &\quad+\frac{1}{2}\log\frac{\left|\boldsymbol{w}C_{ij}\boldsymbol{w}^T\right|}{\sqrt{\boldsymbol{w}\sum_i\boldsymbol{w}^T}\sqrt{\boldsymbol{w}\sum_j\boldsymbol{w}^T}} \qquad (12)\end{aligned}$$

where

$$B_{ij} = \frac{1}{4}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)^T \qquad (13)$$

$$C_{ij} = \frac{1}{2}\left(\sum\nolimits_i+\sum\nolimits_j\right), 1\le i\le j\le C. \qquad (14)$$

Thus, the gradient of $J_{\boldsymbol{w}}$ with respect to $\boldsymbol{w}$ is shown to be

$$\frac{\partial J_{\boldsymbol{w}}}{\partial \boldsymbol{w}} = -\sum_{1\le i\le j\le C}\sqrt{\lambda_i\lambda_j}e^{-g_{\boldsymbol{w}}(i,j)}\frac{\partial g_{\boldsymbol{w}}(i,j)}{\partial \boldsymbol{w}} \qquad (15)$$

where

$$\begin{aligned}\frac{\partial g_{\boldsymbol{w}}(i,j)}{\partial \boldsymbol{w}} &= \frac{1}{2}\left(\boldsymbol{w}C_{ij}\boldsymbol{w}^T\right)^{-1}\\ &\quad\times\left[\boldsymbol{w}B_{ij}\boldsymbol{w}^T\left(\boldsymbol{w}C_{ij}\boldsymbol{w}^T\right)^{-1}\boldsymbol{w}C_{ij}-\boldsymbol{w}B_{ij}\right]\\ &\quad+\left(\boldsymbol{w}C_{ij}\boldsymbol{w}^T\right)^{-1}\boldsymbol{w}C_{ij}-\frac{1}{2}\\ &\quad\times\left[\left(\boldsymbol{w}\sum\nolimits_i\boldsymbol{w}^T\right)^{-1}\boldsymbol{w}\sum\nolimits_i+\left(\boldsymbol{w}\sum\nolimits_j\boldsymbol{w}^T\right)\boldsymbol{w}\sum\nolimits_j\right]. \qquad (16)\end{aligned}$$

A. DBB Model

In Section II, the BB method assumes that the observation data at a certain moment are independent of the historical observation sequence. This assumption is valid only when the sampling interval is long. However, for most modern industrial processes, sampling intervals are short and thus, the assumption of serial independence is not satisfied. In addition to considering the cross correlation between different variables, autocorrelation of each variable in time needs to be considered. The DBB method constructs an augmented matrix by stacking past values of each variable together with the current observation. In this way, the dynamic relationship between system variables can be effectively extracted, thereby accurately describing the dynamic behavior of the system.

Assume that the observation dataset $\boldsymbol{X}$ without stacking lagged variables contains d observation variables, and each variable has n observation values. $\boldsymbol{X}$ is expressed as

$$\boldsymbol{X} = \begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1d}\\ x_{21} & x_{22} & \cdots & x_{2d}\\ \vdots & \vdots & \ddots & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{nd}\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_1^T\\ \boldsymbol{x}_2^T\\ \vdots\\ \boldsymbol{x}_n^T\end{bmatrix} \qquad (17)$$

where, $\boldsymbol{x}_i \in R^d$, $i = 12,\ldots,n$, is a d-dimensional observation vector. Extending $\boldsymbol{X}$ by appending previous l observations of

each variable sequentially yields the augmented matrix as

$$\boldsymbol{X}(l)=\begin{bmatrix}\boldsymbol{x}_t^T & \boldsymbol{x}_{t-1}^T & \cdots & \boldsymbol{x}_{t-l}^T\\ \boldsymbol{x}_{t-1}^T & \boldsymbol{x}_{t-2}^T & \cdots & \boldsymbol{x}_{t-l-1}^T\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{x}_{t+l-n}^T & \boldsymbol{x}_{t+l-n-1}^T & \cdots & \boldsymbol{x}_{t-n}^T\end{bmatrix} \tag{18}$$

where $\boldsymbol{x}_t^T$ is the d-dimensional observation vector at time t.

The BB model established in (11) and (18) is known as the DBB. Compared with the method based on a single observation vector [26], [27], the augmented vector method can capture dynamic information from the data and improve the performance of fault diagnosis. However, how to find the optimal lag l is an essential problem. It risks losing information if l is too small and overfitting if l is too large. In this article, a variant of the IC [24], [29] in system identification is used to determine the time lag l, which will be detailed in Section III-C.

B. Multiple Projections Learning for the DBB Approach

This section introduces a learning method based on the Gram–Schmidt process [22], which involves seeking a DBB projection matrix $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_a]\in R^{d\times a}$, where a represents the order of dimension reduction. We denote $\boldsymbol{S}_t$, $\boldsymbol{S}_b$, and $\boldsymbol{S}_w$ as the total scatter matrix, the between-class scatter matrix, and the within-class scatter matrix, respectively. To ensure the orthogonality of the columns in the projection matrix $\boldsymbol{W}$, we add the constraint $\boldsymbol{W}^T\boldsymbol{S}_t\boldsymbol{W}=\boldsymbol{I}$ with $\boldsymbol{S}_t=\boldsymbol{S}_b+\boldsymbol{S}_w$, i.e.,

$$\boldsymbol{w}_i^T\boldsymbol{S}_t\boldsymbol{w}_j=\begin{cases}1, i=j\\ 0, i\neq j\end{cases}. \tag{19}$$

Assuming that the first r, $1\le r<a$, projection vectors, i.e., $\boldsymbol{W}_r=[\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_r]\in R^{d\times r}$, are known, the representation of the next projection vector $\boldsymbol{w}_{r+1}$ can be obtained by

$$\begin{aligned}&\min_{\boldsymbol{w}_{r+1}} J_{\boldsymbol{w}_{r+1}}=\frac{2}{C(C-1)}\sum_{1\le i\le j\le C}\sqrt{\lambda_i\lambda_j}e^{-g_{w+1}(i,j)}\\ &\text{s.t. }\boldsymbol{w}_{r+1}^T\sum\boldsymbol{w}_{r+1}=1,\\ &\quad\boldsymbol{w}_{r+1}^T\sum\boldsymbol{w}_i=0,\quad i\le r.\end{aligned} \tag{20}$$

In order to solve the above problem, we introduce the following equation to create conditions for converting it into an unconstrained problem:

$$\boldsymbol{A}_{r+1}\triangleq\boldsymbol{I}_d-\boldsymbol{W}_r\boldsymbol{W}_r^T\boldsymbol{S}_t. \tag{21}$$

Obviously, when the rth projection vector is determined, $\boldsymbol{W}_{r+1}$ is in the subspace spanned by the column vectors in formula (21). Thus, there is a vector $\boldsymbol{v}_{r+1}\in R^d$ satisfying

$$\begin{aligned}&\boldsymbol{w}_{r+1}=\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}\\ &(\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1})^T S_t\boldsymbol{w}_i=0\quad\forall 1\le i\le r.\end{aligned} \tag{22}$$

Substituting (22) into (20) to eliminate the second constraint, we obtain an unconstrained optimization problem

$$\min_{\boldsymbol{w}_{r+1}}\frac{2}{C(C-1)}\sum\nolimits_{1\le i\le j\le C}\sqrt{\lambda_i\lambda_j}e^{-g_{w+1}(i,j)} \tag{23}$$

where

$$\begin{aligned}g_{\boldsymbol{w+1}}(i,j)=&\frac{1}{2}\text{trace}\\ &\left\{\left(\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}C_{ij}\boldsymbol{v}_{r+1}^T\boldsymbol{A}_{r+1}^T\right)^{-1}\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}B_{ij}\boldsymbol{v}_{r+1}^T\boldsymbol{A}_{r+1}^T\right\}\\ &+\frac{1}{2}\log\frac{\left|\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}C_{ij}\boldsymbol{v}_{r+1}^T\boldsymbol{A}_{r+1}^T\right|}{\sqrt{\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}\sum_i\boldsymbol{v}_{r+1}^T\boldsymbol{A}_{r+1}^T}\sqrt{\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}\sum_j\boldsymbol{v}_{r+1}^T\boldsymbol{A}_{r+1}^T}}.\end{aligned} \tag{24}$$

The conjugate gradient method mentioned in Section II is used to solve (23). The obtained solution is denoted as $\boldsymbol{v}_{r+1}^*$.Then, the projection vector $\boldsymbol{W}_{r+1}$ is normalized to satisfy the first constraint in (22)

$$\boldsymbol{w}_{r+1}=\frac{\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}^*}{\sqrt{\left(\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}^*\right)^T\boldsymbol{S}_t\boldsymbol{A}_{r+1}\boldsymbol{v}_{r+1}^*}}. \tag{25}$$

Since this optimization problem is nonlinear, we use the first FDA projection vector as the initial conditions to facilitate the algorithm for better solutions.

C. Information Criterion

There are two hyperparameters to be determined in Sections III-A and III-B—the time lag l and the DR order a. Conventionally, cross validation is adopted to determine the optimal hyperparameters, where a subset of training data is used as training set and the rest as validation set. Under each combination of the selected hyperparameter values, n-fold cross validation can be performed to obtain the averaged validation error. The optimal hyperparameter values are chosen as those giving the smallest validation error. However, cross validation has the disadvantage of heavy computation burden and thus, is not applicable to small datasets. To this end, we propose a method based on the following IC:

$$\text{IC}=f(a,l)+\frac{a}{n} \tag{26}$$

where $f(a,l)$ is the misclassification rate from the training set, obtained by projecting data into the first a DBB loading vectors. l is the time lag and n is the average number of training samples per class.

D. DBB-Based Fault Diagnosis

In summary, the detailed flow chart of obtaining optimal DBB projection vectors is demonstrated in Fig. 1. Note that this is an offline training step and thus, the main algorithmic complexity in the optimization step shall not pose any issue. Once the projection matrix $\boldsymbol{W}$ is obtained, it can be used to diagnose the fault in any given test data. Specifically, for a test dataset, we first stack observations in a similar fashion as (18). Then,

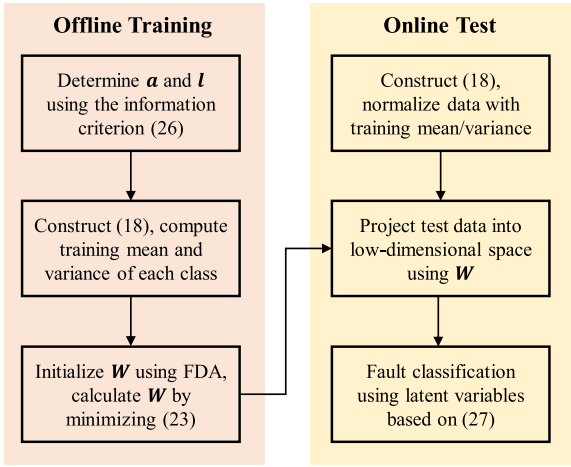

Fig. 1. Training procedure flow chart.

the obtained loading matrix $\boldsymbol{W}$ is used to project the data to the latent space, in which the following discriminant function [24] can be used for fault classification:

$$g_j(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_{j,\text{mean}})^T\boldsymbol{W} \times \left(\frac{1}{n_j-1}\boldsymbol{W}^T\boldsymbol{S}_j\boldsymbol{W}\right)^{-1}\boldsymbol{W}^T(\boldsymbol{x}-\boldsymbol{x}_{j,\text{mean}}) - \frac{1}{2}\ln\left[\det\left(\frac{1}{n_j-1}\boldsymbol{W}^T\boldsymbol{S}_j\boldsymbol{W}\right)\right]. \quad (27)$$

For each observation, it is grouped into class j if

$$g_j(\boldsymbol{x}) > g_i(\boldsymbol{x}) \qquad \forall i \neq j. \quad (28)$$

IV. Application to TEP

The TEP is a well-known benchmark in process control and monitoring, and it simulates the closed-loop operation of an actual industrial process with high fidelity. The TEP has 52 process variables with 41 controlled variables and 12 manipulated variables except for the agitation speed of the reactor's stirrer. The simulation data have one healthy scenario and 21 fault scenarios of different types. The sampling interval is 3 min. Each simulation scenario includes 480 samples of training data and 800 samples of testing data. Note that all the simulations are conducted under Gaussian-distributed disturbance and noise [24], and thus, all fault classes can meet the necessary assumption on the distribution of the data. More details about the TEP are given in [13].

A. Experiment 1: IDV3, IDV4, IDV9, IDV11, IDV14

In this experiment, we use IDVs 3, 4, 9, 11, and 14 as the fault set to verify our method [25]. Details about IDVs 3, 4, 9, 11, and 14 can be found in Table I

It is observed that both IDV 3 and IDV 9 are related to the D feed temperature and thus, although of different types, they are difficult to separate. On the other hand, IDV 4, 11, and 14 are all related to the properties of reactor cooling water. Specifically, IDV 4 and 11 are faults in the temperature of reactor cooling water, while IDV 14 is regarding the flow rate. In general, it is expected that the data from these faults may possess significant overlap due to the reasons above. In this example, we examine our algorithm by diagnosing these five faults.

TABLE I
Description of Faults Selected for the Case Studies

	Fault	Description	Type
Experiment 1	IDV 3	D feed temp	Step
	IDV 4	Reactor cooling water inlet temp	Step
	IDV 9	D feed temp	Random variation
	IDV 11	Reactor cooling water inlet temp	Random variation
	IDV 14	Reactor cooling water valve	Sticking
Experiment 2	IDV 16	Unknown	Unknown
	IDV 17	Unknown	Unknown
	IDV 18	Unknown	Unknown
	IDV 19	Unknown	Unknown

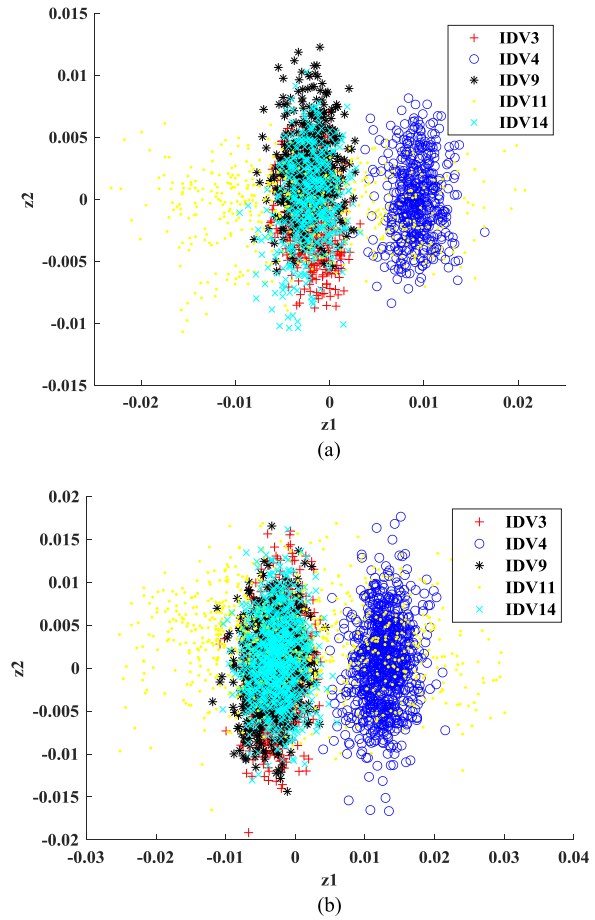

Fig. 2. Projections of the training and testing data for Experiment 1 onto the first two FDA loading vectors. (a) Training data. (b) Testing data.

To verify our heuristic, we project both training data and test data from these five faults into the first two FDA directions, as shown in Fig. 2. Clearly, these fault classes demonstrate a large overlap. IDV3, IDV9, IDV14 are strongly overlapped, whereas IDV4 is well-separated from this cluster although it has a slight overlap with the sparsely scattered IVD11.

To determine the optimal hyperparameters for the DBB-based fault diagnosis, we grid the time lag l and the reserved order of

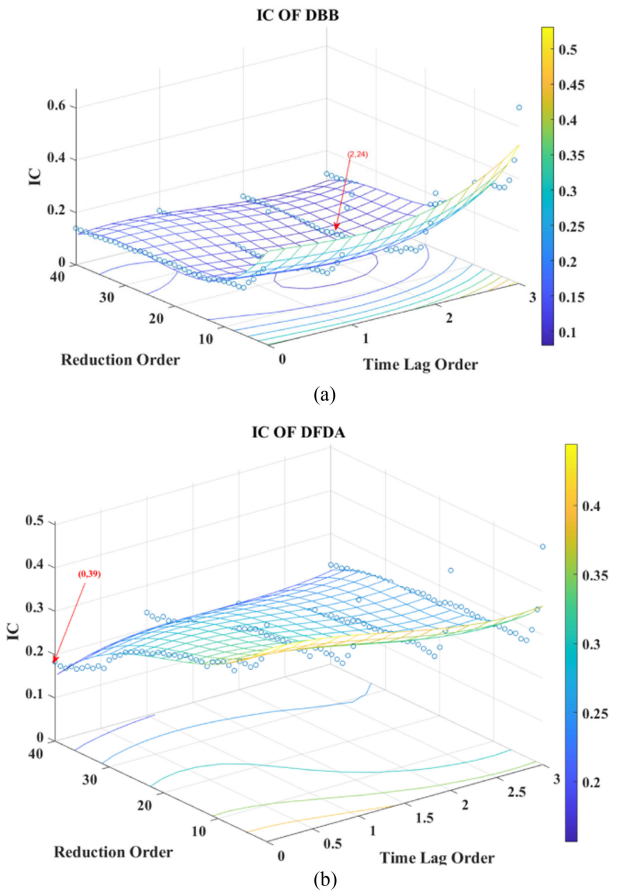

Fig. 3. IC plots for two comparison schemes: (a) DBB-based method and (b) DFDA-based method.

DR a in the interval [0, 3] and [1, 40], respectively. For each grid, the conjugate gradient is used to obtain the optimal loading vectors for DBB approach, upon which the IC can be computed. A 2-D IC surface is fitted to the computed values of the IC on each mesh grid, as shown in Fig. 3(a). The optimal hyperparameters are thus determined as $l = 2$ and $a = 24$ from the training data.

We make two comparisons between DBB-based methods (i.e., the methods based on BB and DBB), DFDA-based methods (i.e., FDA and dynamic FDA), and locality-preserving FDA (LP-DFDA) [26] in this case study. The motivation to select (D)FDA and LP-DFDA for comparison is as follows. (D)FDA is the most classical DR-based method for fault diagnosis and it adopts a different criterion during DR. The LP-DFDA is a nonlinear version (considering local nonlinear manifolds in the data) of FDA. It can almost provide the best performance reported in the literature for DR-based methods. By comparing with these methods, we can clearly demonstrate the necessity of considering DBB criterion during DR. For the first comparison, all methods use the same hyperparameter values determined above, i.e., $l = 2$ and $a = 24$. For BB and FDA, where the serial correlations among process variables are neglected, we simply set $a = 24$. Apparently in this case, the selected hyperparameters are not optimal for the DFDA-based and LP-DFDA methods. Thus, in the second comparison, for DFDA

TABLE II
CLASSIFICATION RESULTS FOR FDA, DFDA, BB, AND DBB FOR COMPARISON 1

Misclassification rate (%)	FDA	DFDA	LP-DFDA	BB	DBB
IDV 3	74.38	63.53	56.34	53.00	59.40
IDV 4	12.63	14.79	38.64	13.75	15.04
IDV 9	69.13	78.57	63.36	57.38	51.63
IDV 11	39.38	47.49	14.68	24.38	9.02
IDV 14	66.13	71.55	0	0.25	0.00
Overall	52.33	55.19	34.60	29.75	27.02

TABLE III
CLASSIFICATION RESULTS FOR DFDA, LP-DFDA, AND DBB FOR COMPARISON 2

Misclassification rate (%)	(D)FDA-based	LP-DCVA	DBB-based
IDV 3	73.25	55.07	59.40
IDV 4	17.75	12.39	15.04
IDV 9	65.38	56.45	51.63
IDV 11	34.88	26.16	9.02
IDV 14	12.38	0.130	0.00
Overall	40.12	30.04	27.02

and LP-DFDA methods, we seek their respective optimal hyperparameters by plotting the IC surface of DFDA, as shown in Fig. 3(b) for DFDA. It is found that the optimal values are $l = 0, a = 39$ for DFDA and $l = 1, a = 7$ for LP-DFDA (the IC surface is omitted). The detailed fault diagnosis results are listed in Tables II and III, respectively, for these two comparison schemes.

From Fig. 3(b) one can see a clear downward trend as the order increases, which indicates that the DBB-based method prefers less complex models, compared with the DFDA-based method. The optimal reduction order a and time lag l for the DFDA-based method in this case can be determined as $l = 0$ and $a = 39$ from training sets. $l = 0$ indicates that the original data are not dynamically enhanced for DFDA for this specific set of faults.

Table II describes the performance of four fault diagnosis methods on the test data: FDA, DFDA, LP-DFDA, BB, and DBB. With the FDA method, the misclassification rates of IDV3, IDV4, and IDV9 are as high as 74.38%, 66.13%, and 69.13%, respectively, i.e., these faults are incorrectly diagnosed most of the time. In contrast, the BB-based method, although also neglecting serial correlations, achieves better diagnosis performance for all five faults. The overall misclassification rate for the BB-based method is 29.75%, which is a factor of 1.76 enhanced discriminability than FDA's 52.33%. As remarked in the above sections, the main reasons for the superior performance of the BB-based method are: 1) the BB-based method minimizes the Bayes error through minimizing an upper bound, which is a direct manifestation of the misclassification rate and 2) the objective function sums up the BB between all pairs of faults and thus, the separation between these faults is not dominated by classifying only a few faults.

In addition, the fault diagnosis results of DFDA, LP-DFDA, and DBB are also included in Table II. By stacking lagged variables into the data matrix, the dynamic information between consecutive samples can be fully captured. From Table II, the

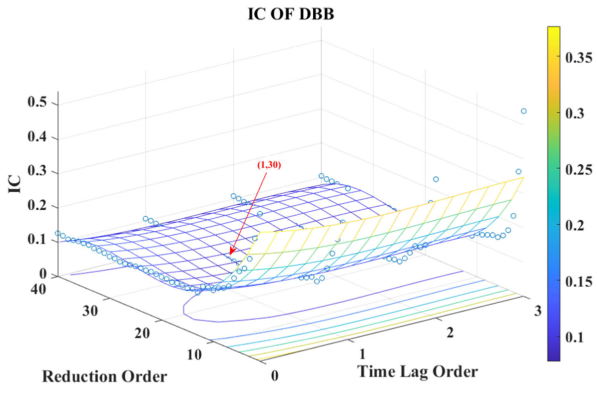

Fig. 4. Surface fitting based on DBB-based method IC data.

LP-DFDA shows 34.60% misclassification rate (better than FDA and DFDA), whereas the DBB method demonstrates the best classification performance with only 27.02% misclassification rate. This is a factor of 1.94, 2.04, 1.28, and 1.10 less than that of the FDA, DFDA, LP-DFDA, and BB, respectively. Since the selected hyperparameters may not be optimal for DFDA (and LP-DFDA), its performance is surprisingly worse than that of the FDA.

For comparison 2, we update the optimal hyperparameters for DFDA as $l = 0$ and $a = 39$ from Fig. 3(a), and $l = 1$ and $a = 7$ for LP-DFDA. The detailed results are in Table III, where the overall misclassification rate of DFDA is 40.12%, a much better performance than that in comparison 1. However, this rate is still a factor of 1.48 worse than that of DBB. The DBB method has a factor of 1.44 better than LP-DCVA (38.81%). Again, the reason comes from the advantage of DBB in directly reducing Bayes error relative to the DFDA method.

B. Experiment 2: IDV16, IDV17, IDV19, IDV20

In this experiment, we investigate the fault diagnosis of IDV 16, IDV 17, IDV 19, and IDV 20 (see Table I), which are unknown faults, to further evaluate the performance of the DBB-based fault diagnosis method. Similar to the previous example, the first step is to determine the optimal hyperparameter values. To this end, we plot the IC value at different combinations of the reduction order a and time lag l for the DBB-based method, as shown in Fig. 4, together with a fitted 2-D surface. The optimal hyperparameters in this case are determined as $l = 1$ and $a = 30$ from the training data. As a comparison, for the methods based on BB and FDA, the optimal orders are chosen as $a = 24$. The DFDA-based method selects the best order determined by IC as $l = 1$ and $a = 24$, and the best hyperparameter for the LP-DFDA is shown to be $l = 1$ and $a = 40$.

As shown in Fig. 5, the DR generated by DBB has better visual performance than FDA, BB, DFDA, and LP-DFDA in terms of less overlap between the data from IDVs 16, 17, 19, and 20 on the 2-D reduced space. For each method, the right plot demonstrates the histogram of normalized pairwise L_2 distances between all pairs of IDVs in the 2-D space. It is observed that the distributions of distances between the classes generated by FDA, DFA, and LP-DFDA are fairly unbalanced, with some distances being very small while others being large. In contrast,

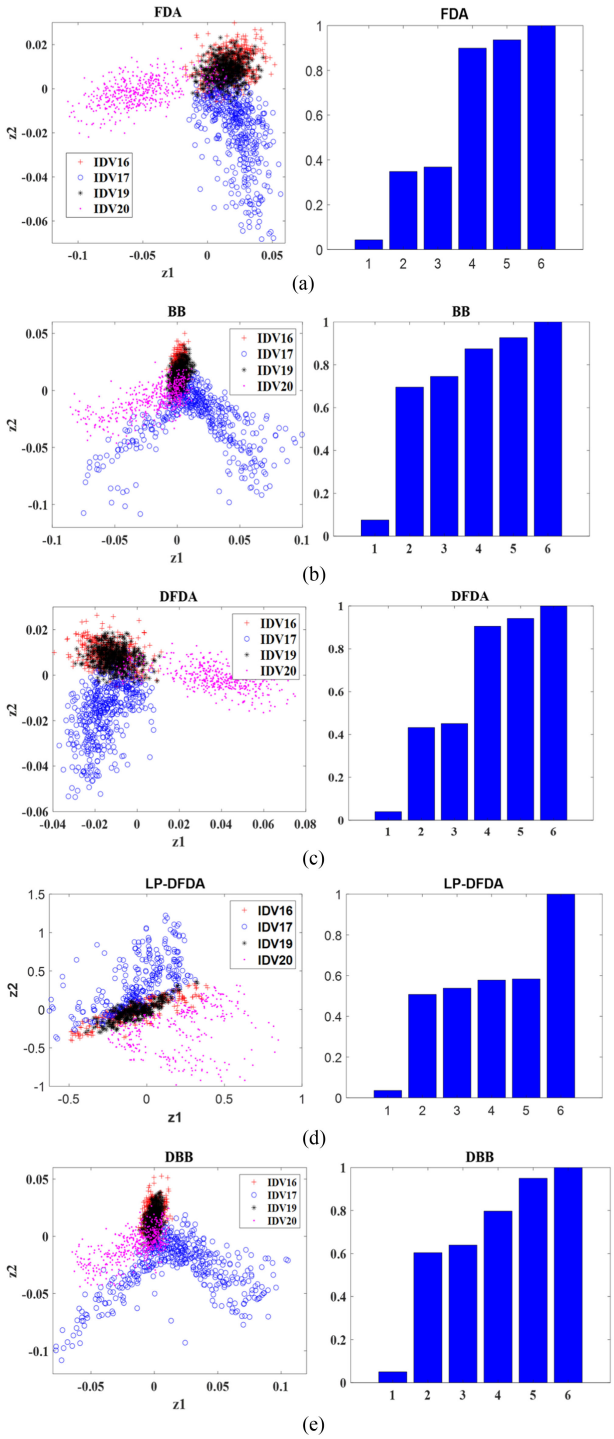

Fig. 5. Projections of training data using different methods for IDV 16, 17, 19, and 20 onto the first two loading vectors (left) and the histogram of distances between fault classes (right): (a) FDA; (b) BB; (c) DFDA; (d) LP-DFDA; and (e) DBB.

the distances between the classes generated by BB and DBB are more uniform, which implies better separability, and this is consistent with the scatter plots on the left.

Table IV illustrates the classification results of each method on the testing data. As shown in Table IV, the overall misclassification rate for DBB is 8.01% for the test data, compared to 12.67% of DFDA and 10.14% of LP-DFDA, which is a factor of

TABLE IV
CLASSIFICATION RESULTS OF FDA, DFDA, LP-DFDA, BB, AND DBB FOR THE TEST DATA

Misclassification rate (%)	FDA	DFDA	LP-DFDA	BB	DBB
Fault 16	28.75	23.65	7.76	16.28	8.51
Fault 17	16.13	12.14	3.63	18.25	7.51
Fault 19	10.38	3.50	20.03	9.38	1.00
Fault 20	12.25	11.39	9.14	19.13	15.02
Overall	16.88	12.67	10.14	15.78	8.01

1.58 and a factor of 1.27 enhanced diagnostic performance for DFDA and LP-DFDA, respectively. Specifically, DBB shows much superior performance than the other methods in almost all fault classes except for Fault 20 than (D)FDA and BB. Also, the BB-based method gives lower misclassification rate than FDA. This observation shows that the (D)BB-based approaches are consistently better than the (D)FDA-based approaches. In addition, the DBB method shows improved test performance than LP-DFDA. In addition, both DFDA and DBB approaches present improved performance than their respective counterparts without considering lagged observations in the data. Thus, taking account of serial correlation is of significance in seizing the dynamic patterns in the TEP data and practical data where fast sampling is ubiquitous.

V. CONCLUSION

In this article, we proposed a novel DBB method for fault diagnosis. The proposed method seeked to minimize the Bayes error during DR, in contrast to traditional DR such as FDA. A Bhattacharyya upper bound was formulated as a quantification of Bayes errors in distinguishing the faults. To further grip the serial correlation information due to dynamics of the process, lagged variables were also incorporated into the data matrix. For multiple faults, the objective function of DBB was formulated by summing up the pairwise DBB for all pairs of fault classes. A novel IC was proposed to account for both the training performance and model complexity for selecting the hyperparameters. Finally, the proposed method was examined by the benchmark TEP. The DBB-based method yielded superior fault diagnosis performance than FDA, DFDA, LP-DFDA, and the BB-based approach. Future work includes validating the proposed algorithm using real applications and also applications to fault diagnosis of vibration signals [27], [28].

REFERENCES

[1] S. Yin, S. X. Ding, A. Haghani, H. Hao, and P. Zhang, "A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process," *J. Process Control*, vol. 22, no. 9, pp. 1567–1581, 2012.

[2] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.,* vol. 63, no. 5, pp. 3137–3147, May 2016.

[3] F. Cheng, Q. He, and J. Zhao, "A novel process monitoring approach based on variational recurrent autoencoder," *Comput. Chem. Eng.*, vol. 129, 2019, Art. no. 106515.

[4] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.

[5] J. Yu, "Local and global principal component analysis for process monitoring," *J. Process Control*, vol. 22, no. 7, pp. 1358–1373, 2012.

[6] A. Papaioannou and S. Zafeiriou, "Principal component analysis with complex kernel: The widely linear model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1719–1726, Sep. 2014.

[7] S. Yin, J. Rodriguez-Andina, and Y. Jiang, "Real-time monitoring and control of industrial cyber-physical systems: With integrated plant-wide monitoring and control framework," *IEEE Ind. Electron. Mag.*, vol. 13, no. 4, pp. 38–47, Dec. 2019.

[8] K. Zhong, M. Han, T. Qiu, and B. Han, "Fault diagnosis of complex processes using sparse kernel local fisher discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1581–1591, May 2020.

[9] P. Hu, D. Sang, Y. Shang, and Y. Xiang, "Multi-view linear discriminant analysis network," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5352–5365, Nov. 2019.

[10] C. Tong, T. Lan, H. Yu, and X. Peng, "Distributed partial least squares based residual generation for statistical process monitoring," *J. Process Control*, vol. 75, pp. 77–85, 2019.

[11] Y. Dong and S. J. Qin, "New dynamic predictive monitoring schemes based on dynamic latent variable models," *Ind. Eng. Chem. Res.*, vol. 59, no. 6, pp. 2353–2365, 2020.

[12] M. Barker and W. Rayens, "Partial least squares for discrimination," *J. Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[13] L. H. Chiang, M. E. Kotanchek, and A. K. Kordon, "Fault diagnosis based on fisher discriminant analysis and support vector machines," *Comput. Chem. Eng.*, vol. 28, no. 8, pp. 1389–1401, 2004.

[14] B. Jiang, D. Huang, X. Zhu, F. Yang, and R. D. Braatz, "Canonical variate analysis-based contributions for fault identification," *J. Process Control*, vol. 26, pp. 17–25, 2015.

[15] B. Jiang, X. Zhu, D. Huang, and R. D. Braatz, "A combined canonical variate analysis and fisher discriminant analysis (CVA–FDA) approach for fault diagnosis," *Comput. Chem. Eng.*, vol. 77, pp. 1–9, 2015.

[16] F. Keinosuke, *Introduction to Statistical Pattern Recognition*. Amsterdam, The Netherlands: Elsevier, 2013.

[17] S. George and M. Padmanabhan, "Minimum Bayes error feature selection for continuous speech recognition," in *Proc. Neural Inf. Process. Syst. Conf.*, 2001, pp. 800–806.

[18] H. Herman, "Chernoff bound," in *International Encyclopedia of Statistical Science*. Berlin, Germany: Springer, 2001, pp. 242–243.

[19] G. Xuan, P. Chai, and M. Wu, "Bhattacharyya distance feature selection," in *Proc. 13th Int. Conf. Pattern Recognit.,* Vienna, Austria, 2002, pp. 195–199.

[20] J. E. Cavanaugh, "Criteria for linear model selection based on Kullback's symmetric divergence," *Aust. New Zealand J. Statist.*, vol. 46, no. 2, pp. 257–274, 2015.

[21] K. Brinker, "Multilabel classification via calibrated label ranking," *Mach. Learn.*, vol. 73, no. 2, pp. 133–153, 2008.

[22] K. Swirydowicz, J. Langou, S. Anathan, U. Yang, and S. Thomas, "Low synchronization Gram–Schmidt and generalized minimal residual algorithms," *Numer. Linear Algebr. Appl.*, vol. e2343, pp. 1–20, Oct. 2020.

[23] Y. Jiang and S. Yin, "Recursive total principal component regression based fault detection and its application to vehicular cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1415–1423, Apr. 2018.

[24] L. H. Chiang, E. L. Russell, and R. D. Braatz, *Fault Detection and Diagnosis in Industrial Systems*. London, U.K.: Springer-Verlag, 2001.

[25] Y. Jiang and S. Yin, "Recent advances in key-performance-indicator oriented prognosis and diagnosis with a MATLAB toolbox: DB-KIT," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 2849–2858, May 2019.

[26] Q. Lu, B. Jiang, R. B. Gopaluni, P. D. Loewen, and R. D. Braatz, "Locality preserving discriminative canonical variate analysis for fault diagnosis," *Comput. Chem. Eng.*, vol. 117, pp. 309–319, 2018.

[27] A. Glowacz *et al.*, "Detection of deterioration of three-phase induction motor using vibration signals," *Meas. Sci. Rev.*, vol. 10, no. 6, pp. 241–249, 2019.

[28] W. Caesarendra, M. Pratama, B. Kosasih, T. Tjahjowidodo, and A. Glowacz, "Parsimonious network based on a fuzzy inference system (PANFIS) for time series feature prediction of low-speed slew bearing prognosis," *Appl. Sci.*, vol. 8, no. 12, 2018, Art. no. 2656.

[29] B. Jiang, Z. Guo, Q. Zhu, and G. Huang, "Dynamic minimax probability machine-based approach for fault diagnosis using pairwise discriminate analysis," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 2, pp. 806–813, Mar. 2019.