Knowledge and Information Systems (2022) 64:75–93
https://doi.org/10.1007/s10115-021-01625-w

REGULAR PAPER

An integrated framework for diagnosing process faults with incomplete features

Roozbeh Razavi-Far[1] · Mehrdad Saif[2] · Vasile Palade[3] · Shiladitya Chakrabarti[2]

Received: 26 March 2020 / Revised: 31 October 2021 / Accepted: 7 November 2021 /
Published online: 26 November 2021

Abstract
Handling missing values and large-dimensional features are crucial requirements for data-driven fault diagnosis systems. However, most intelligent data-driven diagnostic systems are not able to handle missing data. The presence of high-dimensional feature sets can also further complicate the process of fault diagnosis. This paper aims to devise a missing data imputation unit along with a dimensionality reduction unit in the pre-processing module of the diagnostic system. This paper proposes a novel pooling strategy for missing data imputation (PSMI). This strategy can simplify complex patterns of missingness and incrementally update the pool. The pre-processing module receives incomplete observations, PSMI estimates missing values, and, then, the dimensionality reduction unit transforms completed observations onto a lower-dimensional feature space. These transformed observations are then fed as inputs to the fault classification module for decision making and diagnosis. This diagnostic scheme makes use of various state-of-the-art missing data imputation, dimensionality reduction and classification algorithms. This enables a comprehensive comparison and allows to find the best techniques for the sake of diagnosing faults in the Tennessee Eastman process. The obtained results show the effectiveness of the proposed pooling strategy and indicate that principal component analysis imputation and heteroscedastic discriminant analysis approaches outperform other imputation and dimensionality reduction techniques in this diagnostic application.

Keywords Data analysis · Missing data imputation · Dimensionality reduction · Fault diagnosis · Principal component analysis · Heteroscedastic discriminant analysis

✉ Roozbeh Razavi-Far
roozbeh@uwindsor.ca

[1] Department of Electrical and Computer Engineering and School of Computer Science, University of Windsor, Windsor, ON N9B 3P4, Canada

[2] Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada

[3] Center for Data Science, Coventry University, Coventry, CV1 5FB, UK

1 Introduction

Fault diagnosis plays an important role in improving the performance and increasing the safety and reliability of industrial processes [26]. Faults are unexpected events that avoid a system or its units to deliver the designated set of services. Faults are known to cause poor performances in equipment or cause them to operate adversely, which can lead to production stoppage and financial losses [9,18]. Hence, a fast and accurate diagnosis of faults is of paramount importance for industrial processes [6–8,23,30].

Fault diagnosis techniques can be divided into two categories [26]: data-driven and model-based methods. Model-based approaches use quantitative models and explicit equations to estimate the states or parameters of the system. However, given a complex system, it is hard to identify an accurate model, which follows the process behaviour in real operational conditions [26]. Data-driven diagnostic schemes usually use a set of observed features, in order to train a diagnostic model [14]. Although these data-driven diagnostic schemes tend to be accurate, fast and less complex, their performance is highly dependent on the quality and representativeness of the collected data [10,22].

Industrial processes are usually being monitored with various sensors and equipment. This often generates huge amount of data, which result in the curse of dimensionality [4, 16]. Collecting noisy and redundant features usually decreases the diagnostic performance. Mapping the data onto a lower-dimensional feature space is necessary, therefore, for a proper diagnosis of faults. It also reduces the required time and space, lowers the complexity of the decision-making process and increases the accuracy of the diagnostic system, while preserving the variability of the original data to a great extent [5,34]. Dimensionality reduction can be divided into two categories [42]: feature selection and feature extraction. Feature selection aims to remove redundant and irrelevant features from the original data, in order to reduce the dimension [16]. Feature extraction aims to transform the original data onto a lower-dimensional feature space. This paper makes use of feature extraction techniques in order to reduce the dimension of the feature space and the computational time, while increasing the accuracy of the diagnostic system [5].

Collecting incomplete observations is an unavoidable issue in the online industrial processes. This is due to data transmission errors, incorrect measurements and faulty sensors [15,25,31,32]. Most of the intelligent fault classifiers are designed to work with complete observations and are not suited to handle incomplete ones. Consequently, the missing observations turn out to be a serious challenge for the fault classifiers, causing negative impacts on the decision-making process [29,39]. This paper intends to deal with diagnosing faults under incomplete scenarios.

The two most commonly used methods of handling missing data are discarding and imputation. Discarding, which is the most common method, ignores missing observations. However, it leads to loss of critical information and dependencies, which may result in inaccurate results. Imputation, on the other hand, replaces the missing observations with the estimated values [27,33]. Missing data imputation techniques are classified under two major categories, as single and multiple imputation techniques [33]. Multiple imputation techniques are more complex and computationally expensive and, thus, not suitable for online monitoring applications. Single imputation techniques are usually simpler to implement and fast, making them an ideal choice for online monitoring applications [33].

Data-driven diagnostic systems are indeed predictive models that are constructed by means of data collected from the industrial systems. The absence of complete data hampers the diagnostic performance in both training and test phases. Incomplete observations during the

training (offline) phase are usually treated by means of complete observations. However, the main issue is during the test (online phase), where the observations become available one by one or in different batches for decision making. The data-driven diagnostic systems are usually cannot handle incomplete observations and thus the state of the systems is under question. The presence of missing data during the test/online phase can hamper diagnostic performance, no matter observations emerge in a batch-wise fashion or a one-by-one manner. In the former, the presence of a large number of missing values in the batch of data and their complex missingness pattern can complicate missing estimation. The latter suffers from the lack of donors for missing estimation since an incomplete observation emerges individually at once.

This paper proposes a novel pooling strategy for missing data imputation (PSMI), which aims to reduce the complexity of the missingness structure in the batch of data as well as the imputation error. PMSI forms a pool of donors to estimate missing values whenever an incomplete observation emerges during the online/test phase. During the online/test phase, PMSI enables the diagnostic system to treat missing data (one-by-one fashion) by resorting to the pool of donors and improves the imputation performance (batch-wise fashion) by simplifying complex patterns of missingness that may exist in the batch of data. Any arbitrary and random pattern of missingness can be transformed to a univariate missing pattern that can be simply treated by means of any missing data imputation technique. The proposed strategy also incrementally updates the pool and allows imputing the missing observations by resorting to both complete observations and recently imputed ones.

This paper also proposes a diagnostic scheme, which contains two main modules for the pre-processing and the decision-making processes. The former module includes several state-of-the-art missing data imputation and dimensionality reduction (DR) techniques. This module generates complete sets of observations with a lower number of features, in order to reduce the dimension of the data and increase the diagnostic accuracy. Thus, the proposed scheme enables a comprehensive comparison, which results in finding the best combination of techniques for diagnosing faults with missing observations in the Tennessee Eastman process.

Various state-of-the-art missing data imputation techniques are used in the pre-processing module of the diagnostic system to treat incomplete observations. These missing data imputation techniques are compared together in terms of normalized root mean square (NRMS) error. The attained results by each missing data imputation technique show the efficiency of the proposed strategy. PMSI is also compared with an un-pooled strategy, in which incomplete observations are accumulated, and, then, treated by missing data imputation techniques. The attained results show the efficiency of PMSI in terms of NRMS and the performance of the decision-making module. Besides, PMSI enables online decision making, while the un-pooled strategy cannot immediately make a decision and requires to collect a certain number of observations, which results in delay in decision making that is a very important issue in online diagnostic applications.

The remainder of the paper is structured as follows: Sect. 2 of the paper briefly presents the design of the diagnostic system. Section 3 presents the selected techniques for missing data imputation and dimensionality reduction. Section 4 discusses and compares the experimental results. Conclusions are presented in Sect. 5.

2 Design of the diagnostic system

Most intelligent diagnostic classifiers rely on the complete datasets to learn and diagnose process faults [32]. However, fault classification becomes a challenging task in the presence of missing data. On the other hand, incomplete features are often being collected due to failures in sensors, transmission devices and system process. Missing data mechanisms can be broadly classified under three categories, namely Missing at Random (MAR), Missing Not at Random (MNAR) and Missing Completely at Random (MCAR) [35]. This paper focuses on the MCAR structure, that is usually generated due to failures in the sensors and data collection process.

Diagnostic accuracy is highly dependent on the availability of useful data, and, thus, presence of missing data can negatively impact on the diagnostic performance [32]. Therefore, to improve the diagnostic performance, there is a need to estimate missing values prior to fault classification.

Figure 1 illustrates the workflow of the diagnostic system. The proposed diagnostic model contains two modules for pre-processing and fault classification. The pre-processing module itself contains two units for missing data imputation and dimensionality reduction, which aim to estimate missing values and reduce the feature dimension, respectively. In the first unit, this work proposes a pooling strategy for missing data imputation (PSMI).

The collected data is split into training and test subsets using a tenfold cross-validation strategy as shown in Fig. 1. In order to evaluate the performance of the diagnostic system, an incomplete scenario is generated by inducing missing values in various observations of the test subsets, in a completely random order. The pattern of missingness, which is created in the test subset, is usually an arbitrary pattern, which further complicates the imputation process. PSMI is proposed in order to reduce the impact of the pattern of missingness. In this strategy, as illustrated in Fig. 1, the complete set of observations of the training subset are forming a pool, that is used during the imputation process. The test subset contains missing features with an arbitrary pattern of missingness. However, test observations are fed into the pool one by one. A test observation without any missing feature is automatically added to the pool. If a test observation contains any missing feature, the imputation technique makes use of the pool of observations and the current missing observation (i.e. the target observation) to estimate the missing features of the target observation and, thus, creates a complete observation. This imputed observation is then added to the pool for the future use, while also being fed to the subsequent unit for dimensionality reduction. The dimensionality reduction unit transforms observations onto reduced feature space and, then, sends reduced observations to the fault classification module for decision making.

The imputation is performed whenever a new incomplete observation is added to the pool. PSMI uses three state-of-the-art algorithms for estimation of missing data including extreme learning machine imputation (ELMI), k-nearest neighbours imputation (kNNI) and principal component analysis imputation (PCAI). PSMI helps in transforming any arbitrary pattern of missingness that may appear in the test subset into a simpler pattern of missingness (i.e. univariate or monotone). Algorithm 1 describes the pooling strategy.

The pool P is initialized using the training subset $X_{m\times n}$, where m stands for the number of observations and n stands for the number of features. The test subset $Y_{o\times n}$ contains o observations. Each of the test observation y_t can be split into complete y_t^{obs} and incomplete y_t^{mis} vectors, where $t = 1, \ldots, o$.

Each complete observation y_t is fed into the pool. For an incomplete observation, the imputation module splits the pool subset P into observed P^{obs} and target P^{mis} features, with

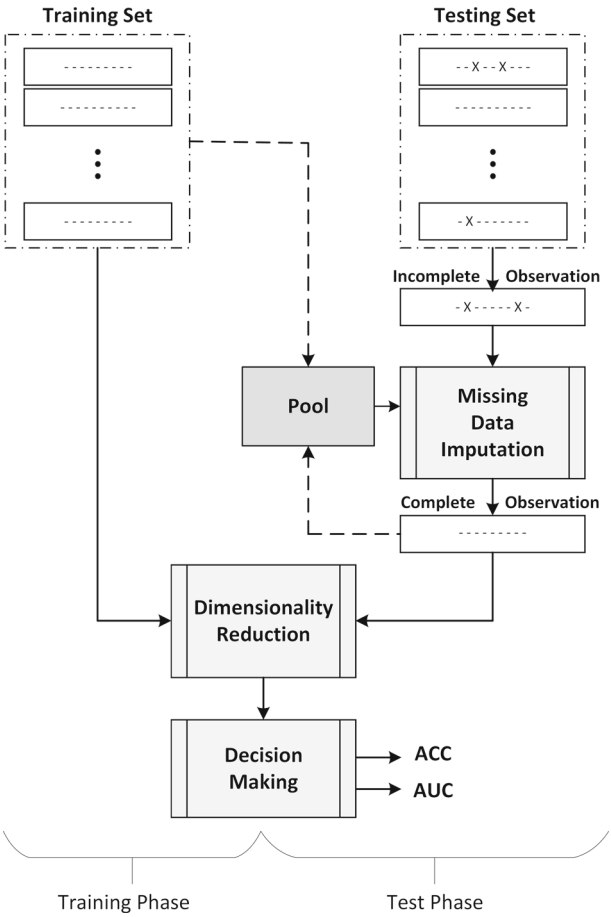

Fig. 1 Block diagram of the diagnostic system. The incomplete observations are imputed through PSMI and, consequently, transformed onto a lower-dimensional feature space by means of the dimensionality reduction unit and, then, fed into the fault classification module for decision making

the number of target features identical to the missing features. The observed features are fed as the training inputs to one of the imputation techniques Ω_w, where $w = 1, 2, 3$, along with the target features P^{mis} as the targets to train a prediction model $\hat{P}$. This step is defined inside a function, called $CreateModel$. The next step, which is defined inside another function, called $PredictData$, uses that predictive model $\hat{P}$ and complete features y_t^{obs} to estimate the missing features $\hat{y_t}^{\text{mis}}$. The complete and estimated features are combined together to create a complete subset.

The completed subset is then fed into the dimensionality reduction (DR) unit. High-dimensional features can often lead to unreliable predictions and a large computational time. This requires the subset to be transformed onto a more informative feature subset with a lower dimension. Here, the DR module uses two state-of-the-art supervised methods, namely neighbourhood components analysis (NCA) and heteroscedastic discriminant analysis (HDA), and two unsupervised methods, namely extreme learning machines (ELM) and principal components analysis (PCA). These state-of-the-art DR methods are selected and used in the DR

Algorithm 1: Pooling strategy for missing data imputation.

INPUTS:
$X_{m\times n}$ is training subset containing m observations.
$Y_{o\times n}$ is the test subset containing o observations.
n stands for the number of features.
DEFINITIONS:
n^{mis} stands for the number of missing features in an observation.
n^{obs} stands for the number of complete features in an observation.
Ω_w is the imputation technique where $w = 1, 2, 3$
PROCESS:
INITIALIZE the pool P with X.
for $\forall y_t \in Y (t = 1, \ldots, o)$ **do**
  if $y_t^{\text{mis}} = \emptyset$ **then**
    APPEND y_t to the pool P
  else
    TRAIN a predictive model using P
    $\left[\hat{\mathcal{P}}\right] = CreateModel(\Omega_w, P^{\text{obs}}, P^{\text{mis}})$
    FEED y_t^{obs} into the predictive model
    $\left[\hat{y}_t^{\text{mis}}\right] = PredictData(\hat{\mathcal{P}}, y_t^{\text{obs}})$.
    The complete observation is then $\hat{y}_t = (y_t^{\text{obs}}, \hat{y}_t^{\text{mis}})$
    APPEND $\hat{y}_t$ to the pool P
  end
end

unit for the sake of comparison. Several low-dimensional models are generated through different DR methods by means of the training subset. Each test observation is applied to these models, in order to be transformed onto a lower-dimensional feature space. These completed observations of the lower dimension are then fed to the fault classification module for the decision making. This module also contains four classifiers, namely extreme learning machine (ELM), k-nearest neighbours (kNN), linear discriminant analysis (LDA) and heteroscedastic discriminant analysis (HDA). These state-of-the-art classifiers are selected and used in the fault classification module for the sake of comparison. The fault classification performance is analysed through the accuracy (ACC) and the area under the curve (AUC) measures [28].

3 Diagnostic scheme

The proposed scheme is designed for fault classification from high-dimensional features under incomplete scenarios. It contains pre-processing and decision-making modules. The pre-processing module uses the pooling strategy for missing data imputation (PSMI) along with the dimensionality reduction (DR) unit. They are designed into a serial configuration to impute missing values and return complete observations and, then, reduce the data into a lower dimension feature space. This provides the decision-making module with complete and informative observations of the lower dimension.

3.1 Missing data imputation

The pooling strategy for missing data imputation (PSMI) consists of three imputation techniques that uses the pool of observations to train a model that can be used to estimate missing scores in the target observation. Once the imputation is performed, the completed observation is appended into the pool.

3.1.1 Principal component analysis imputation (PCAI)

PCAI makes use of the regularized iterative PCA algorithm to impute missing scores [12,19]. The algorithm initially estimates the missing values with the feature means and, then, applies PCA on the completed dataset. Thereafter, it imputes the missing values with the regularized reconstruction with a predefined order [19]. These steps are repeated until convergence criterion has been met.

3.1.2 k-Nearest neighbour imputation (kNNI)

kNNI is based on the k-nearest neighbour algorithm, which replaces the missing feature values with a weighted mean of the k-nearest neighbour features. The weights for the neighbours are inversely proportional to the similarly measure, like Euclidean distances or Pearson correlation, from the neighbouring features [3,38].

3.1.3 Extreme learning machine imputation (ELMI)

ELMI estimates missing values by means of the extreme machine learning regression algorithm [24]. The algorithm trains a regression model using the complete features as inputs and the missing features, target features, as targets [17]. Using the trained model, the algorithm regresses the complete features of the incomplete subset to impute the missing values of the incomplete, target, features.

3.2 Dimensionality reduction

The DR unit consists of four state-of-the-art feature extraction methods. These techniques calculate the transformation matrix W [41]. This transformation matrix W, applies linear transformation in order to project the complete dataset X onto a lower-dimensional feature space, while retaining maximum knowledge about the original dataset.

3.2.1 Heteroscedastic discriminant analysis (HDA)

HDA is a linear dimensionality reduction algorithm, which is proposed for datasets with the normal distribution between each class, allowing for the heteroscedasticity of the given data [21].

Considering an input data of $X = X_1 \cup X_2 \cup \cdots \cup X_l$ belonging to l classes of n observations, such that $X_i = \{\mathbf{x}_{1i}, \mathbf{x}_{2i}, \ldots, \mathbf{x}_{ni}\}$, the multi-class HDA criterion aims to attain the matrix W, that maximizes the following function:

$$\zeta_{\mathrm{HDA}}(W) = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \rho_i \rho_j \mathrm{tr}\Big\{\big(W\sigma_\omega W^T\big)^{-1} W S \sigma_\omega^{1/2} \Big[\big(\sigma_\omega^{-1/2}\sigma_{ij}S_\omega^{-1/2}\big)^{-1/2} \times \sigma_\omega^{-1/2}\sigma_{\eta_{ij}}\sigma_\omega^{-1/2} \big(\sigma_\omega^{-1/2}\sigma_{ij}\sigma_\omega^{-1/2}\big)^{-1/2} + \frac{1}{\pi_i\pi_j}\big(\log\big(\sigma_\omega^{-1/2}\sigma_{ij}\sigma_\omega^{-1/2}\big) - \pi_i \log\big(\sigma_\omega^{-1/2}\sigma_i\sigma_\omega^{-1/2}\big) - \pi_j \log\big(\sigma_\omega^{-1/2}\sigma_j\sigma_\omega^{-1/2}\big)\big)\Big]\sigma_\omega^{1/2} W^T\Big\} \quad (1)$$

where ρ_i stands for the a priori probability of class i, σ_i stands for the within-class covariance matrix of class i, μ_i stands for the mean vector of class i, μ stands for the estimated overall mean, σ_ω stands for the average within-class scatter matrix, σ_ϵ stands for the between-class scatter matrix, σ_{ij} stands for the average pairwise within-class scatter matrix, $\sigma_{\epsilon_{ij}}$ stands for the pairwise between-class scatter matrix, π_i and π_j stand for the relative priors [21], which can be calculated using the following set of equations:

$$\begin{aligned}
\rho_i &= |X_i|/|X| \\
\sigma_\omega &= \sum_{i=1}^{k} \rho_i \sigma_i \\
\sigma_\epsilon &= \sum_{i=1}^{k} (\mu_i - \mu)(\mu_i - \mu)^T \\
\mu &= \sum_{i=1}^{k} \rho_i \mu_i \\
\sigma_{\epsilon_{ij}} &= (\mu_i - \mu_j)(\mu_i - \mu_j)^T \\
\sigma_{ij} &= \pi_i\sigma_i + \pi_j\sigma_j \\
\pi_i &= \rho_i/(\rho_i + \rho_j) \\
\pi_j &= \rho_j/(\rho_i + \rho_j).
\end{aligned}$$

W can be calculated by solving the eigenvalue decomposition of the following equation and, then, selecting the largest m eigenvectors.

$$\sigma_{\mathrm{HDA}} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \rho_i \rho_j \mathrm{tr}\Big\{\sigma_\omega^{-1}\sigma_\omega^{1/2}\Big[\big(\sigma_\omega^{-1/2}\sigma_{ij}\sigma_\omega^{-1/2}\big)^{-1/2} \times \sigma_\omega^{-1/2}\sigma_{\epsilon_{ij}}\sigma_\omega^{-1/2}\big(\sigma_\omega^{-1/2}\sigma_{ij}\sigma_\omega^{-1/2}\big)^{-1/2} + \frac{1}{\pi_i\pi_j}\big(\log\big(\sigma_\omega^{-1/2}\sigma_{ij}\sigma_\omega^{-1/2}\big) - \pi_i \log(\sigma_\omega^{-1/2}\sigma_i\sigma_\omega^{-1/2}\big) - \pi_j \log\big(\sigma_\omega^{-1/2}\sigma_j\sigma_\omega^{-1/2}\big)\big)\Big]\sigma_\omega^{1/2}\Big\}. \quad (2)$$

3.2.2 Neighbourhood components analysis (NCA)

NCA is a dimensionality reduction algorithm, which is proposed for learning a Mahalanobis distance measure that is used in the kNN classification algorithm [13]. This algorithm is restricted to learning the Mahalanobis distance metrics that can be represented by the symmetric positive semi-definite metrics and estimated using the inverse square roots [13].

This algorithm makes use of a differentiable cost function, that is based on a stochastic neighbour assignment in the transformed space, which can be considered an effective measure compared to estimating the actual leave-one-out classification error [13].

The probability ρ_{ij} that the jth point is selected by the ith point as a neighbour can be computed as follows:

$$\rho_{ij} = \frac{\exp(-\|W\mathbf{x}_i - W\mathbf{x}_j\|^2}{\sum_{k\neq i}\exp(-\|W\mathbf{x}_i - W\mathbf{x}_k\|^2} \tag{3}$$

where W stands for the transformation matrix. The probability ρ_i stands for the correct classification of the point i is given by the following equation:

$$\rho_i = \sum_{j\in l_i}\rho_{ij} \tag{4}$$

where l_i stands for all points of the same class i. This algorithm requires to obtain the matrix W, which can maximize the number of correctly classified observations, as follows:

$$\zeta_{\text{NCA}}(W) = \sum_i\sum_{j\in l_i}\rho_{ij} = \sum_i \rho_i. \tag{5}$$

The gradient-based method along with the gradient operator of Eq. (6) is used to maximize $\zeta_{NCA}(W)$:

$$\frac{\nabla\zeta_{\text{NCA}}}{\nabla W} = 2W\sum_i\left\{\rho_i\sum_k\rho_{ik}x_{ik}x_{ik}^T - \sum_{j\in l_i}\rho_{ij}x_{ij}x_{ij}^T\right\} \tag{6}$$

where $x_{ij} = \mathbf{x}_i - \mathbf{x}_j$.

3.2.3 Principal component analysis (PCA)

PCA is a popular unsupervised linear dimensionality reduction technique. It seeks for a projection that best represents the data [2,40].

Considering an input data of $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the projected data is given by $\mathbf{y}_i = W\mathbf{x}_i$ for $i \in (1, \ldots, n)$. The scatter matrix is calculated by the following equation:

$$\sigma = \sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^T \tag{7}$$

where

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j. \tag{8}$$

σ is decomposed into ψ and Λ, where $\psi = (\psi_1, \psi_2, \ldots, \psi_d)$ stands for the eigenvectors and $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$ stands for the eigenvalues. The eigenvectors and eigenvalues are arranged so that $\lambda_1 > \lambda_2 > \cdots > \lambda_d$. In order to reduce the data to a lower dimension m, the matrix W can be reformulated as $W = [(\psi_1^T, \psi_2^T, \ldots, \psi_m^T)]^T$.

3.2.4 Extreme learning machine (ELM)

ELM is a generalized form of the single hidden layer feedforward network (SLFN) [17]. Considering an input data of $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, b hidden nodes are generated to approximate the input–output relation among these n pairs of input observations with a zero error. This can be done by resorting to random weights and bias, (υ_i, ν_i), and the output weight vector Γ_i, which satisfy the following:

$$\sum_{i=1}^{b} \Gamma_i f(\upsilon_i \cdot \mathbf{x}_j + \nu_i) = \mathbf{y}_j + \varepsilon_j, \quad j = 1, \ldots, n \tag{9}$$

where ε stands for the error or noise, b stands for the number of hidden nodes, f stands for the hidden layer activation function and y_j stands for the output vector at the jth row. This can be reformulated as follows:

$$G\Gamma = \Upsilon \tag{10}$$

where

$$G = \begin{bmatrix} g(\mathbf{x}_1) \\ \vdots \\ g(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} f(\upsilon_1 \cdot \mathbf{x}_1 + \nu_1) & \cdots & f(\upsilon_b \cdot \mathbf{x}_1 + \nu_b) \\ \vdots & \ddots & \vdots \\ f(\upsilon_1 \cdot \mathbf{x}_n + \nu_1) & \cdots & f(\upsilon_b \cdot \mathbf{x}_n + \nu_b) \end{bmatrix}_{n \times b} \tag{11}$$

$$\Gamma = \begin{bmatrix} \Gamma_1^T \\ \vdots \\ \Gamma_b^T \end{bmatrix}_{b \times o} \tag{12}$$

and

$$\Upsilon = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix} = \begin{bmatrix} \mathbf{y}_{11} & \cdots & \mathbf{y}_{1o} \\ \vdots & \ddots & \vdots \\ \mathbf{y}_{n1} & \cdots & \mathbf{y}_{no} \end{bmatrix}_{n \times o}. \tag{13}$$

G stands for the randomized matrix of the hidden layer output, o stands for the number of target features and Υ stands for the training data target matrix. The ith column of the matrix G stands for the ith hidden node output w.r.t. the set of n inputs, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and the jth row of the matrix G stands for hidden layer feature mapping w.r.t. the jth input $\mathbf{x}_j$, i.e. $g(\mathbf{x}_j) = \{f(\upsilon_1 * \mathbf{x}_j + \nu_1)), \ldots, f(\upsilon_L * x_j + \nu_L)\}$.

For the purpose of dimensionality reduction [20], the input bias is set to 0 (i.e. $\nu_i = 0$) and the orthogonal random weight matrix υ, is calculated using a modified version of Eq. 11 where $\upsilon = \{\upsilon_1, \upsilon_2, \ldots, \upsilon_b\}$ and $\upsilon^T \upsilon = \mathbb{I}$. W can be calculated using the following equation:

$$W = \Gamma = \upsilon^T \psi \psi^T \tag{14}$$

where ψ stands for the eigenvectors of the covariance matrix $X^T X$ [20].

3.3 Fault classifiers

The fault classification task is performed using four state-of-the-art data-driven classifiers, namely heteroscedastic discriminant analysis (HDA) [21], linear discriminant analysis (LDA) [11,37], k-nearest neighbour (kNN) [1] and extreme learning machines (ELM) [17]. All of these intelligent classifiers are tuned to handle classification of multi-class datasets.

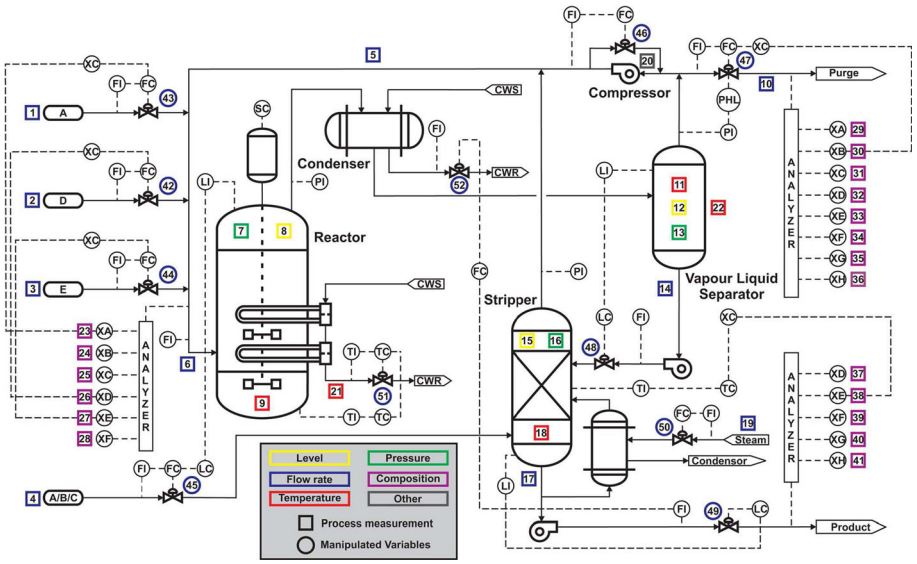

Fig. 2 Tennessee Eastman process workflow [36]

Table 1 Description of the process faults and normal state

Index	Description
0	Normal
1	A/C Feed ratio, B Composition constant (stream 4)
2	B Composition, A/C ratio constant (stream 4)
6	A Feed loss (stream 1)
7	C Header pressure loss—reduced availability (stream 4)

4 Experimental results

In this paper, a pooled missing data imputation strategy and multi-class dimensionality reduction methods are used to form the pre-processing module of the diagnostic system. This proposed diagnostic technique is validated using an industrial process, called Tennessee Eastman (TE) process [6]. The Tennessee Eastman process is a popular industrial process, that has been extensively used to evaluate designed diagnostic and control systems [6]. The TE process consists of an exothermic two-phase reactor, a flash separator and a reboiled striper, as shown in Fig. 2. The simulated Tennessee Eastman process contains 1 normal condition and 21 faulty conditions.

In this paper, various sets of observations from the normal state and four faulty states, i.e. faults 1, 2, 6 and 7, are selected to evaluate the proposed diagnostic system. The description of the selected faulty classes is reported in Table 1. The training subset contains 500 observations from the normal state and 480 observations from each faulty states. Each observation contains 52 features. The testing subset contains 960 observations with 52 features.

The incomplete test subset is imputed using both pooled and un-pooled strategies. Table 2 shows the comparison between normalized root mean square (NRMS) values based on the

Table 2 Comparison between pooled (PMSI) and un-pooled missing data imputation strategies in terms of NRMS

MDI	Pooled	Un-pooled
PCAI	5.68e−5	1.23e−4
kNNI	5.92e−4	5.95e−4
ELMI	1.8e−3	2.7e−3

Table 3 Comparison between pooled (PMSI) and un-pooled missing data imputation strategies in terms of ACC and AUC of the decision-making module

MDI	ACC		AUC	
	Pooled	Un-pooled	Pooled	Un-pooled
PCAI	0.950	0.923	0.928	0.912
kNNI	0.946	0.922	0.915	0.905
ELMI	0.925	0.919	0.923	0.907

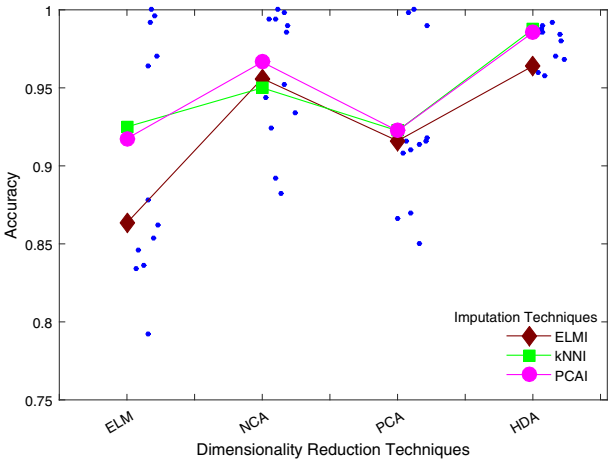

Fig. 3 The distribution of the accuracy measures attained through each dimensionality reduction method combined with missing data imputation methods

imputed dataset generated by the different missing data imputation techniques. It can be seen that the pooled strategy generates a slightly lower NRMS value, which indicates that the imputed scores are closer to the real scores. The major benefit of the pooled strategy is, however, the reduction in complexity of the missing data imputation technique achieved through univariate and monotone patterns.

The incomplete observations are imputed using both pooled and un-pooled strategies. Table 3 shows the comparison between the averaged ACC and AUC values attained by classifiers through each imputation technique and strategy. It can be seen that the pooled (PMSI) strategy outperforms the un-pooled strategy in terms of both ACC and AUC, which indicates that the imputed scores by means of PMSI can improve the diagnostic performance. The obtained results confirm the efficiency of PMSI in handling random missing features in online diagnostic applications.

The completed and transformed observations are used as inputs for the purpose of fault classification. The classification performance is measured and compared in terms of accuracy and receiver operating characteristic curve for the different techniques. Accuracy (ACC) can

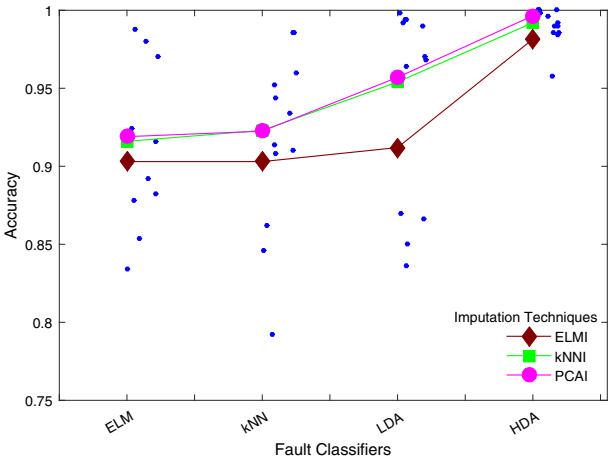

Fig. 4 The distribution of the accuracy measures attained by each fault classifier through missing data imputation methods

be termed as the degree of closeness of prediction of the normal and faulty classes to the original classes. The receiver operating characteristic (ROC) curve is a plot that illustrates the classification performance by varying the discrimination threshold. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) of each fault class. The area under the ROC curve (AUC) is selected as a performance measure. For the multi-class dataset, ROC curve is generated for every pair of the normal and faulty classes and the average AUC is then calculated.

Figure 3 illustrates the distribution of the performance measures, in terms of accuracy, attained by each dimensionality reduction method. The attained accuracy measures through each dimensionality reduction method are shown by blue circles, where the average of those measures attained through each imputation method are connected together by solid lines with distinct markers. This enables to compare the performance of missing data imputation methods combined with each dimensionality reduction method. Figure 3 shows that PCAI outperforms other imputation methods, where combined with most of dimensionality reduction techniques. This figure also shows a clear advantage of using HDA as a dimensionality reduction (DR) technique along with all three missing data imputation techniques. Both kNNI and PCAI perform closely to each other, however, PCAI results in a better performance measure on average.

Figure 4 illustrates the distribution of the performance measures, in terms of accuracy, attained by each fault classifier (FC). The attained accuracy measures through each fault classifier are shown by blue circles, where the average of those accuracy measures attained by each missing data imputation method are connected together by solid lines with distinct markers. This enables to compare the performance of each missing data imputation method combined with each fault classifier. Figure 4 shows that PCAI outperforms other imputation methods, where combined with most of the fault classifiers. The figure shows that HDA outperforms other classifiers, regardless of the missing data imputation technique that is used to estimate the missing values a priori. PCAI shows a slight advantage over kNNI for missing data imputation technique when combined with all four classifiers. HDA outperforms other selected dimensionality reduction techniques since it makes use of the linear transformation and directed distance matrix. HDA uses the weighted sum of the corresponding directed

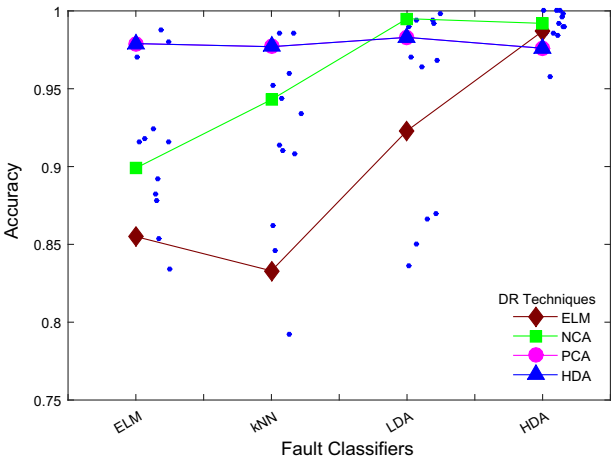

Fig. 5 The distribution of the accuracy measures attained by each fault classifier through dimensionality reduction methods

distance matrices instead of the well-known between-class scatter matrix. On the other hand, PCAI outperforms other missing data imputation techniques since it considers the global similarity among all features.

Figure 5 illustrates the distribution of the performance measures, in terms of accuracy, attained by each fault classifier. The attained accuracy measures through each fault classifier are shown by blue circles, where the average of those measures attained by each dimensionality reduction method are connected together by solid lines with distinct markers. This enables to compare the performance of each dimensionality reduction method combined with each fault classifier. Figure 5 shows that HAD and NCA outperform other imputation methods, where combined with most of classifiers. The figure shows HDA generates the best overall performance as a fault classifier combined with all four dimensionality reduction techniques. Both PCA and HDA are shown to generate a steady performance when combined with any classifier. NCA and ELM show huge improvement in performance when combined with HDA as a fault classifier.

Figure 6 illustrates the performance obtained through each missing data imputation technique in terms of diagnostic accuracy and AUC. The boxes in the figure depict the distribution range of accuracy or AUC values between the first and third quartile, the lines inside each box depict the median value and the dash lines depict the outlier range. Panel (a) shows PCAI slightly outperforms other imputation techniques in terms of accuracy. Imputation by PCAI for certain sets of observations results in 100% diagnostic performance. On the other hand, ELMI produces the largest range of variation in diagnostic performance among all imputation techniques. Panel (b) shows PCAI slightly outperforms compared to other imputation techniques in terms of AUC.

Figure 7 illustrates the diagnostic performance obtained through each dimensionality reduction technique in terms of accuracy and AUC. Panels (a) and (b) show that the supervised techniques (NCA and HDA) outperform the two unsupervised (ELM and PCA) dimensionality reduction techniques. HDA results in the best performance as a dimensionality reduction tool and also causes the least variation in diagnostic performance.

Figure 8 illustrates the distribution of the diagnostic performance obtained through each fault classifier in terms of accuracy and AUC. The figure reiterates that HDA obtains the best

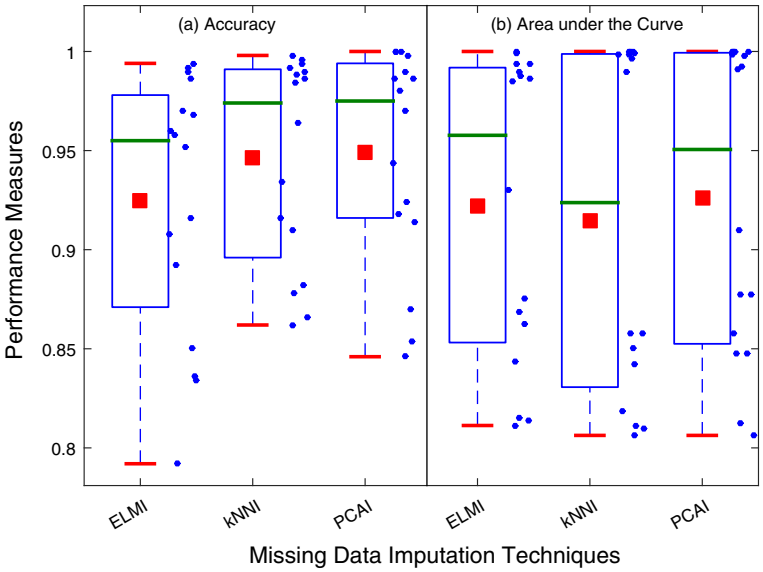

Fig. 6 Distribution of the diagnostic performance attained through each imputation technique

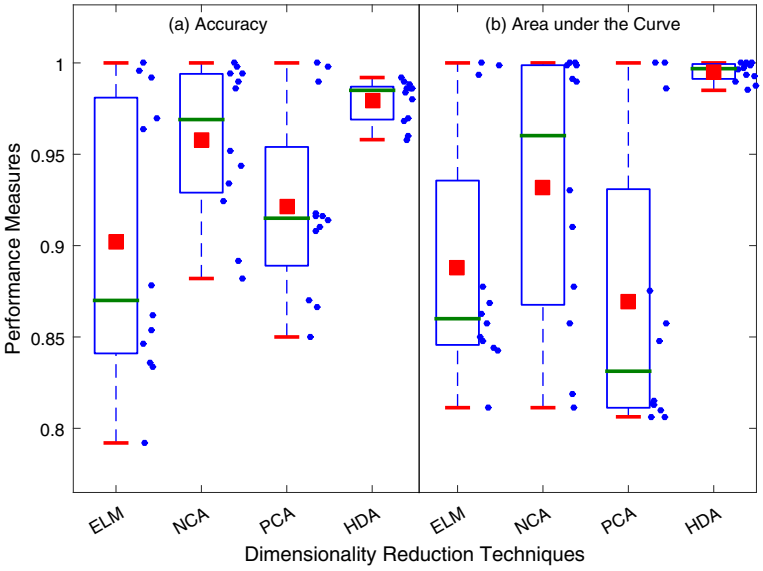

Fig. 7 Distribution of the diagnostic performance attained through each dimensionality reduction technique

performance and the most stable results compared to other classifiers in terms of accuracy and AUC, as seen in both panels.

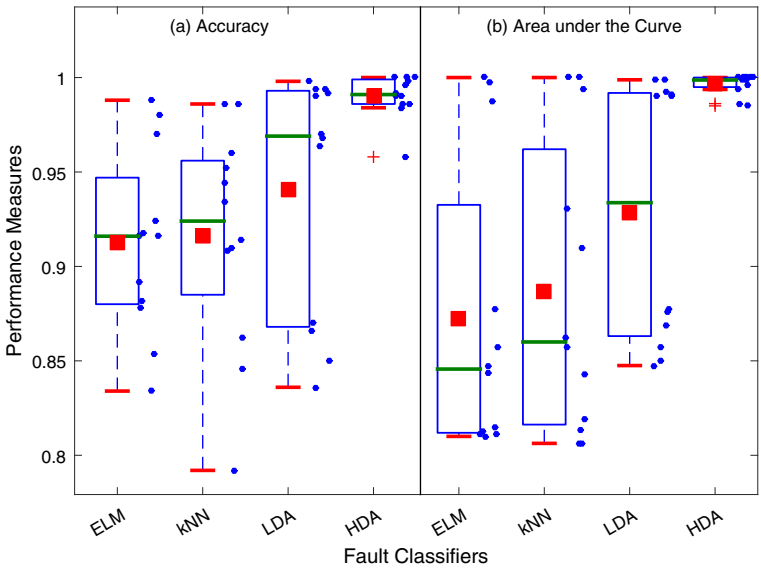

Fig. 8 Distribution of the diagnostic performance attained by each fault classifiers in terms of accuracy

5 Conclusion

This paper proposes an efficient scheme for fault diagnosis with incomplete and high-dimensional features. The proposed scheme contains three modules, for missing data imputation, dimensionality reduction, and fault classification. This work proposes a missing data imputation strategy, which uses a pooling mechanism to create a complete set of features and alleviates complex missingness structures into simple structures, such as univariate and monotone structures. The fault classification performances are evaluated both in terms of accuracy and area under curve performance measures. The obtained results confirm the efficiency of the scheme in handling random missing features in an online industrial application. A comparative study of the results illustrates that, in terms of both accuracy and area under curve, PCAI generates the best performance in the missing data imputation, while HDA outperforms other competitors as both dimensionality reduction and fault classification tools. Thus, it can be concluded that the combination of PCAI-HDA will be an ideal tool for the proposed diagnostic system.

References

1. Altman N (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46(3):175–185
2. Atouni M, Verron S, Kobi A (2015) Fault detection with conditional Gaussian network. Eng Appl Artif Intell 45:473–481
3. Batista G, Monard M (2002) A study of k-nearest neighbour as an imputation method. HIS 87:251–260
4. Bellman RE (1961) Adaptive control processes. Princeton University Press, Princeton
5. Cao W, Haralick R (2009) Affine feature extraction: a generalization of the fukunaga–koontz transformation. Eng Appl Artif Intell 22(1):40–47
6. Downs J, Vogel E (1993) A plant-wide industrial process control problem. Comput Chem Eng 17(2):245–255

7. Farajzadeh-Zanjani M, Hallaji E, Razavi-Far R, Saif M (2021) Generative-adversarial class-imbalance learning for classifying cyber-attacks and faults—a cyber-physical power system. IEEE Trans Dependable Secure Comput. https://doi.org/10.1109/TDSC.2021.3118636
8. Farajzadeh-Zanjani M, Hallaji E, Razavi-Far R, Saif M (2021) Generative adversarial dimensionality reduction for diagnosing faults and attacks in cyber-physical systems. Neurocomputing 440:101–110
9. Farajzadeh-Zanjani M, Hallaji E, Razavi-Far R, Saif M, Parvania M (2021) Adversarial semi-supervised learning for diagnosing faults and attacks in power grids. IEEE Trans Smart Grid 12(4):3468–3478
10. Farajzadeh-Zanjani M, Razavi-Far R, Saif M (2016) Efficient sampling techniques for ensemble learning and diagnosing bearing defects under class imbalanced condition. In: 2016 IEEE symposium series on computational intelligence (SSCI). pp 1–7
11. Fisher R (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188
12. Folch-Fortuny A, Arteaga F, Ferrer A (2016) Missing data imputation toolbox for MATLAB. Chemom Intell Lab Syst 154:93–100
13. Goldberger J, Roweis S, Hinton G, Salakhutdinov R (2004) Neighbourhood components analysis. In: Advances in neural information processing systems, vol 17. MIT Press, pp 513–520
14. Grimble M, Johnson M (2005) Advanced textbooks in control and signal processing. Springer, Berlin
15. Hallaji E, Razavi-Far R, Saif M (2021) DLIN: Deep ladder imputation network. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2021.3054878
16. Hancer E, Xue B, Zhang M, Karaboga D, Akay B (2018) Pareto front feature selection based on artificial bee colony optimization. Inf Sci 422:462–479
17. Huang G (2014) An insight into extreme learning machines: random neurons, random features and kernels. Cogn Comput 6:376–390
18. Jing C, Gao X, Zhu X, Lang S (2014) Fault classification on Tennessee Eastman process: PCA and SVM. In: 2014 International conference on mechatronics and control (ICMC)
19. Josse J, Husson F (2013) Handling missing values in exploratory multivariate data analysis methods. J SFdS 153(2):79–99
20. Kasun LLC, Yang Y, Huang GB, Zhang Z (2016) Dimension reduction with extreme learning machine. IEEE Trans Image Process 25(8):3906–3918
21. Loog M, Duin R (2004) Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. IEEE Trans Pattern Anal Mach Intell 26(6):732–739
22. Monsef H, Ranjbar A, Jadid S (1997) Fuzzy rule-based expert system for power system fault diagnosis. IEE Proc Gener Transm Distrib 144(2):186–192
23. Oliveira J, Pontes VK, Sartori I, Embirucu M (2017) Fault detection and diagnosis in dynamic systems using weightless neural networks. Expert Syst Appl 84:200–219
24. Razavi-Far R, Chakrabarti S, Saif M, Zio E (2019) An integrated imputation–prediction scheme for prognostics of battery data with missing observations. Expert Syst Appl 115:709–723
25. Razavi-Far R, Cheng B, Saif M, Ahmadi M (2020) Similarity-learning information-fusion schemes for missing data imputation. Knowl Based Syst 187:104805
26. Razavi-Far R, Davilu H, Palade V, Lucas C (2009) Model-based fault detection and isolation of a steam generator using neuro-fuzzy networks. Neurocomputing 72(13):2939–2951
27. Razavi-Far R, Farajzadeh-Zanajni M, Wang B, Saif M, Chakrabarti S (2021) Imputation-based ensemble techniques for class imbalance learning. IEEE Trans Knowl Data Eng 33(5):1988–2001
28. Razavi-Far R, Farajzadeh-Zanjani M, Saif M (2017) An integrated class-imbalanced learning scheme for diagnosing bearing defects in induction motors. IEEE Trans Ind Inform 13(6):2758–2769
29. Razavi-Far R, Farajzadeh-Zanjani M, Saif M, Chakrabarti S (2020) Correlation clustering imputation for diagnosing attacks and faults with missing power grid data. IEEE Trans Smart Grid 11(2):1453–1464
30. Razavi-Far R, Kinnaert M (2012) Incremental design of a decision system for residual evaluation: a wind turbine application*. In: IFAC proceedings. 8th IFAC symposium on fault detection, supervision and safety of technical processes, vol 45(20). pp 343–348
31. Razavi-Far R, Palade V, Zio E (2014) Optimal detection of new classes of faults by an invasive weed optimization method. In: 2014 International joint conference on neural networks (IJCNN). pp 91–98
32. Razavi-Far R, Zio E, Palade V (2014) Efficient residuals preprocessing for diagnosing multi-class faults in a doubly fed induction generator, under missing data scenarios. Expert Syst Appl 41(14):6386–6399
33. Scheffer J (2002) Dealing with missing data. Res Lett Inf Math Sci 3:153–160
34. Sharma N, Saroha K (2015) Study of dimension reduction methodologies in data mining. In: International conference on computing, communication and automation (ICCCA2015)
35. Sim J, Kwon O, Lee K (2016) Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets. Expert Syst Appl 46:486–493
36. Wang G, Li J, Sun C, Jiao J (2018) Least squares and contribution plot based approach for quality-related process monitoring. IEEE Access 6:54158–54166

37. Yang X, Rui S, Zhang X, Xu S, Yang C, Liu PX (2019) Fault diagnosis in chemical processes based on class-incremental FDA and PCA. IEEE Access 7:18164–18171
38. Zhang S (2012) Nearest neighbor selection for iteratively KNN imputation. J Syst Softw 85(11):2541–2552
39. Zhang Z, Dong F (2014) Fault detection and diagnosis for missing data systems with a three time-slice dynamic Bayesian network approach. Chemom Intell Lab Syst 138:30–40
40. Zhu J, Ge Z, Song Z (2017) Distributed parallel PCA for modeling and monitoring of large-scale plant-wide processes with big data. IEEE Trans Ind Inform 13(4):1877–1885
41. Zhu Y, Wang Z, Gao D, Li D (2017) GMFLLM: a general manifold framework unifying three classic models for dimensionality reduction. Eng Appl Artif Intell 65:421–432
42. Zhu Z, Song ZH (2011) A novel fault diagnosis system using pattern classification on kernel FDA subspace. Expert Syst Appl 38:6895–6905

Roozbeh Razavi-Far received the B.Sc. degree in Electrical and Computer Engineering, the M.Sc. and Ph.D. degrees from Amirkabir University of Technology and achieved a second Ph.D. degree from Politecnico di Milano. He was a postdoctoral fellow at the Université libre de Bruxelles, Belgium; an NSERC postdoctoral fellow and a lecturer at University of Windsor, Canada. He is currently with the Faculty of Engineering at University of Windsor. He is also a Cross-appointed Professor with the School of Computer Science at University of Windsor. His research focuses on machine learning, data mining, big data analytics, computational intelligence, cybernetics and security of cyber-physical systems. He has served as an Associate Editor and the Guest Editor for several prestigious journals. He is the Chapter Chair of IEEE Computational Intelligence, and Systems, Man and Cybernetics Societies, Windsor section.

Mehrdad Saif received the B.S., M.S. and DEng degrees in Electrical Engineering from Cleveland State University, OH, USA, in 1982, 1984 and 1987, respectively. During his graduate studies, he worked on research projects sponsored by NASA Lewis (now Glenn) Research Center as well as Cleveland Advanced Manufacturing Program (CAMP). In 1987, he joined the School of Engineering Science at Simon Fraser University (SFU), BC, Canada. From 2002 to 2011, he was the Director of the School of Engineering Science. Since July 2011, he has been the Dean of the Faculty of Engineering at the University of Windsor, Windsor, ON, Canada. Dr. Saif served two terms (1995, 1997) as the Chairman of the Vancouver Section of the IEEE Control Systems Society and is currently a member of the editorial board of the IEEE Systems Journal, International Journal of Control and Computers, IEEE CDC and ACC. He has published about 300 journal and conference papers as well as an edited book.

Vasile Palade received the Ph.D. degree from the University of Galati, Galati, Romania, in 1999. He joined Coventry University, Coventry, UK, in 2013, after working for several years as a Lecturer with the Department of Computer Science, University of Oxford, Oxford, UK. He has authored 200 articles in journals and conference proceedings and several books. He is currently a Professor of Artificial Intelligence and Data Science with the Centre for Data Science, Coventry University. His research interests include deep learning and neural networks, various nature-inspired optimization algorithms, computer vision and natural language processing. He has delivered keynote talks and chaired international conferences on machine learning and applications. He is an associate editor for several reputed journals.

Shiladitya Chakrabarti received his BSc and MEng degrees (honours) in computer engineering from the University of Windsor, Canada. His research interests include data mining, machine learning and pattern recognition.