Machine Learning Approaches for Fault Detection in Semiconductor Manufacturing Process: A Critical Review of Recent Applications and Future Perspectives

V. Arpitha and A. K. Pani[*]
Department of Chemical Engineering,
Birla Institute of Technology and Science,
Pilani, Rajasthan, India – 333031

doi: https://doi.org/10.15255/CABEQ.2021.1973

Review
Received: May 14, 2021
Accepted: December 15, 2021

In modern industries, early fault detection is crucial for maintaining process safety and product quality. Process data contains information on the entire plant acting as a map for visualization of relationships between various plant units, making data-driven process monitoring a key technology for efficiency enhancement. This article focuses on review of process monitoring techniques reported for metal etching process, which is a batch operation carried out in semiconductor manufacturing industry. Various machine learning (and deep learning) techniques applied to date for fault detection and diagnosis of metal etching process are surveyed. Detailed survey of research work on different techniques and the reported results are presented in graphical (pie chart and bar chart) and tabular format. The review article further presents the pros and cons, gaps and future directions in the techniques applied in metal etching process.

Keywords:
metal etching process, semiconductor manufacturing, machine learning, process monitoring, fault detection

Introduction

In recent years, process safety and product quality are the major concerns of modern industries. Any abnormal behavior or fault occuring during plant operation, may lead to low process efficiency, unsafe conditions or process shutdown. In order to avoid such undesirable incidents, early detection of these faults become essential in ensuring process safety and downtime minimization. Therefore, automatic process monitoring techniques are implemented in industries for ensuring process safety and quality enhancement. Various process monitoring methods can be categorized as model-based, knowledge-based and data-based methods[1]. Model-based methods (i.e., process models developed from first principles) are based on the understanding of the relationships between different variables, giving very accurate results. However, as the process becomes complex, it becomes difficult to build such models and is expensive. The knowledge-based methods are entirely based on the prior knowledge of process behaviors and experiences of plant operators available. This whole process of creating the foundation of process knowledge is cumbersome and time consuming, and the results obtained are mostly intuitive. Compared to the two aforementioned methods, data-based methods require no prior process knowledge. They only require the data which is recorded and collected, and which is further used for modelling, monitoring, and control. Data-based methods have become hugely popular in the last decade. Recorded data contain the majority of process information, and therefore, stored process data can be effectively utilized for developing efficient data-based process monitoring models. An overview of a typical data-based process monitoring methodology is presented in Fig. 1.

Process monitoring, which is also known as fault detection and diagnosis, can further be categorized into a four-stage activity, namely, i) Fault detection (detecting any abnormal behavior), ii) Fault diagnosis (identifying variables related to the fault), iii) fault source identification (Root cause analysis), and iv) Process recovery (rectifying the fault and returning normal operating regime of process). All the aforementioned factors play a key role in enhancing the quality of production in process industries. Presently, various machine learning techniques are widely applied for fault detection and diagnosis. Machine learning techniques can be grouped into four different classes: unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning[1,2]. Out of these, supervised and unsupervised learning are widely adopted machine learning techniques and contribute to 80–90 percent

[*]Corresponding author: E-mail: akpani@pilani.bits-pilani.ac.in

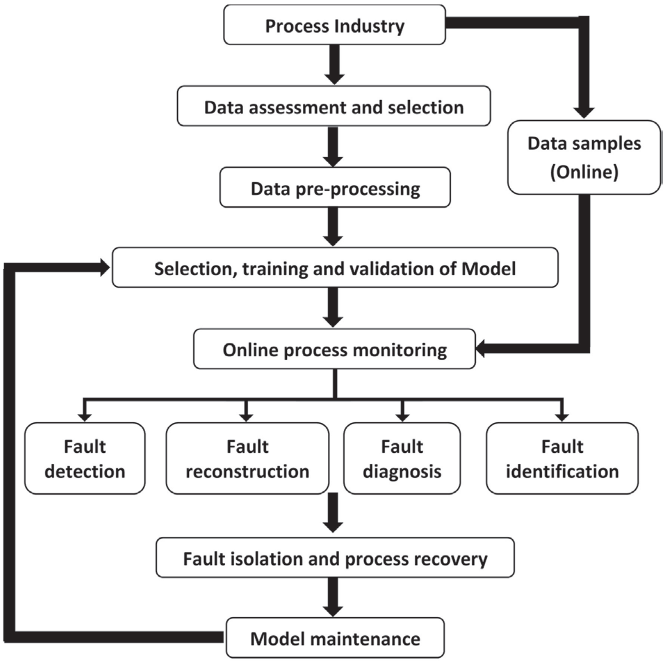

Fig. 1 – *Data-based process monitoring methodology*

of all industrial applications[2]. When data has no labels, and the main goal is to explore the data and to extract hidden structures within them, then the method used against this data are known as unsupervised learning methods. They are most commonly used for dimensionality reduction, information extraction, outlier detection, density estimation, process monitoring, etc. In process monitoring literature, principal component analysis (PCA), independent component analysis (ICA), support vector data description (SVDD), Gaussian mixture models (GMM), k-means clustering, kernel density estimation are some of the notable unsupervised learning methods. In contrast, the supervised learning methods deal with labeled data samples, which are either discrete or continuous. They add meaningful labels to the data so that when models are applied to these labeled data, they can classify or predict a likely label when new unlabeled data are encountered. When the label is of discrete value, supervised learning can be applied as classification techniques for fault detection. If the label value is continuous, then regression models can be constructed for prediction and estimation. Supervised learning methods find their application most commonly in fault classification and identification, soft sensor modeling, quality prediction and online estimation, process monitoring, key performance index prediction and diagnosis, etc. Most frequently used supervised methods include[2]: multivariate linear regression, artificial neural network (ANN), principal component regression (PCR), partial least square (PLS) regression, and support vector regression (SVR).

Any industrial process can be classified as a batch, semi-batch or continuous process. A continuous process is operated continuously as a single work unit at a time with no breaks in sequence, time or material flow, providing output at a constant rate, for example, distillation process, fertilizer industry, metal smelting process, etc. On the other hand, a batch process consists of steps that are performed in a defined order for a finite duration, with production following strictly the process specifications. Here, various grades of product can be produced as there is a scope of altering the operating conditions in a single batch process, for example, food processing, soap manufacturing, electrical machines, injection molding machine in plastic engineering, etc. The process of focus in this article is the semiconductor manufacturing industry, which is a batch process manufacturing sector. An important aspect of monitoring batch processes is to overcome the disadvantages of batch processing in terms of their operating and processing conditions that need to be controlled. Digital transformation is gradually taking over the operations in process industries, and there are continuous efforts being made in achieving smart factories in order to achieve the goal of a fourth industrial revolution (Industry 4.0). Intelligent manufacturing involves data mapping across end-to-end product life cycle, and helps manufacturers with current challenges in becoming more flexible and reacting to market changes with utmost ease. Sensors are made available in almost all processes for fault detection and diagnosis, because any human intervention during a crisis is not efficient and is time consuming. Efficient quality monitoring, process monitoring and process control reduces production cost, process downtime, improved product quality and effluent quality. Process data collected from the industry can be used to identify sensitive variations in the process and provide stable information over an extended time. It is not always necessary that all the sensors are sensitive to every process variable, and are stable enough to provide information for a long time. Thus, it is essential to choose appropriate sensors for efficient process utilization, and to apply various methods carefully to treat process data.

The fabrication of semiconductor devices through metal etching process removes selected layers of wafer for the purpose of pattern transfer, wafer planarization, isolation, and cleaning. Etching removes material only from the pattern traces after the circuit pattern is exposed by coating the wafer with a photoresist. This metal etching process is not a smooth process, and over time may face certain challenges like over-etching, process drifts, and shifts, etc. In order to rectify these problems in metal etching, fault detection and diagnosis becomes very important. Probably the first article reporting fault detection in metal etching was that of Wise *et al.,* who investigated the performances of multiway PCA, parallel factor analysis (PARAFAC), and trilinear decomposition for fault detection in an Aluminum (Al) stack etching process[3]. Their experimental dataset is since then publicly available and has been used by various researchers. The present work provides a thorough review of various data-based methods applied and analyzed on the semiconductor metal etching process data over the past 20 years and beyond. It has been observed that, among all the methods applied to the metal etching process, PCA-based techniques have been profoundly studied over the last two decades. The main reason for this could be their prevailing dimensionality reduction characteristics, which to a large extent, reduce the load on computation and storage. There are modified methods of GMM, k-nearest neighborhood (kNN), SVDD, Infinite GMM, Diffusion maps-based kNN (DM-kNN), Weighted Distance-based kNN (FD-wkNN), Sequential SVDD, Bagging SVDD (BagSVDD) that have also been applied in this field to focus more on the nonlinearity, multimodality, and non-Gaussian nature of the process data. Other than these commonly known machine learning methods, there are some infrequently used techniques like Parallel Factor Analysis 2 (PARAFAC2), Modified Independent Component Analysis (ICA), One-class Support Vector Machines (SVM), Random Forest Similarity Distance, Local Neighbor Normalized Matrix (LNNM), which have also been studied in order to explore different horizons and improve the scope of fault detection and its accuracy as time progresses.

As we move to an era demanding a higher level of data processing, deep learning finds its place of existence in this world. Deep learning techniques benefit complex systems with multiple variables. They can tackle a large number of highly correlated variables for diagnosis and abnormal operating situations through various composition of nonlinearities. Over the past few years, deep learning techniques, namely, Stacked Sparse Auto-Encoder (SSAE) and Denoised Auto-Encoder (DAE) have also been explored for this particular process. All these techniques, along with their variations and how they have benefited in detecting the faults developed in the metal etching process, as well as their shortcomings, if any, are also elaborated further herein in chronological order for each section.

There are review or survey works reported on fault detection and diagnosis methods for different industries. This article is probably the first ever attempt to review various fault detection methods worked upon in the metal etching process for semiconductor manufacturing industry. The present work aims to serve as a ready reference for researchers working in the field of fault detection in etching process, and guide them in the right direction for different scopes of novelty that they can possibly explore. It also encourages researchers to try to overcome the drawbacks of certain techniques that have already been investigated in this process.

The article is organized as follows: A detailed description of the metal etching process carried for semiconductor manufacturing is presented in the next section describing the metal etching process and scope of the present work. In this section, along with process description, different types of associated problems, the relevant variables, and the different possible faults are also mentioned. The section following process description, is the most important part of this article. In this section, we present a review of all the process monitoring techniques reported so far for the semiconductor industry metal etching process. The section is categorized into different subsections based on the type of machine learning techniques applied. Following the detailed review, an analysis of the survey work is presented along with future perspectives and areas of research are presented followed by concluding remarks.

Description of the metal etching process and scope of the present work

Etching is a process of material removal from a wafer surface. This helps to create patterns on the wafer permanently. Etching removes material only from the pattern traces after the circuit pattern is exposed by coating the wafer with a photoresist. Etching chemicals (etchants) are used to remove the material of interest. The main purpose of the etching process in this article is the fabrication of semiconductor devices. The goal of this process is to etch a certain metal wafer or the metal oxide used as a mask of a wafer surface exposing only the portion of metal required. This is achieved mostly through dry etching or more specifically through plasma etching. Plasma etching is one of the most commonly used techniques in the semiconductor processing industry to develop micro- and nano-scale patterns on a silicon wafer. Fig. 2 presents a typical plasma etching flow diagram[4]. In plasma etching, the etchant is introduced in a gas phase. Radio frequency (RF) electrodes are used to produce plasma, which in turn ionizes the gas. The oxide layer undergoes a reaction with the ionized gas and the material is removed. Since the etching process is selective to the material to be removed, the material under the resist or oxide layer or the resist itself is protected.

When it comes to large-scale production, there are many challenges being faced in this process, namely, incomplete etching, overetching, and undercutting, process drift and shift. It becomes diffi-

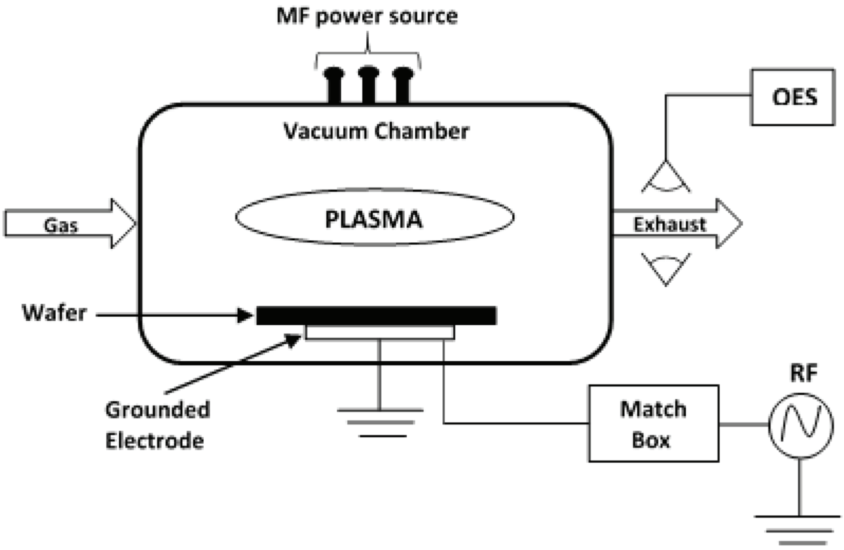

Fig. 2 – *Flow diagram of a plasma etching process*

cult to monitor such faults in a large industry through human intervention. Instead, process sensors are used to monitor the difficulties, and which in turn help to identify the origin of the fault. They play a primary role in fault detection. For the process studied in this work, three sensor systems are most commonly used[3]. Firstly, the Machine State Sensors (MSS) are built into the processing tool collecting machine data during wafer processing, like gas flow rate, chamber pressure. Radio Frequency Monitors (RFMs) are used to measure voltage and current, and phase relationships at frequency 13–56 MHz in the RF control system. Optical Emission Spectroscopy (OES) is used to monitor plasma in the range of 245–800 nm at three locations above the wafer using fiber optics.

The major problems, and the reasons for their occurrence, that are often faced during metal etching a semiconductor are as follows:

Issues faced during the process

- Overetching and undercutting.
- Over time the process could undergo process drift and shifts, as the process is non-stationary, i.e., varying mean and covariance over time.

Reasons for problem occurrence

- Aging of etcher over clean cycle-residue accumulation inside the chamber.
- Fluctuations in the incoming materials because of variations in upstream processes, causing a lag.
- Drift in the sensors used for process monitoring.

This article is a review of different fault detection techniques applied on an Al stack etching process performed on the commercially available Lam9600 plasma metal etching tool[3]. In this process, the TiN/Al-0.5%Cu/TiN/oxide stack is etched with inductively coupled BCl_3/Cl_2 plasma. The different machine state variables presented in Table 1 are taken into account for monitoring. This is a public dataset available in the Eigenvector Research Data Archive (http://www.eigenvector.com/data/Etch/index.html). The dataset consists of values for 129 wafers, among which 108 normal wafers were taken during three experiments (experiments 29, 31 and 33), and 21 wafers (seven in each of the 3 experiments) were intentionally induced with faults by altering the transformer-coupled plasma (TCP) power, RF power, pressure, BCl_3 and Cl_2 flow rate, and He chuck pressure, which is presented in Table 2. Forty-three wafers were processed in each experiment, and they were conducted several weeks apart.

A number of fault detection and process monitoring methods are reported in the literature which are able to identify maximum number of faults in this system. In this review work, different machine learning based fault detection techniques (linear and non-linear) that have been applied to this process so far have been compiled and studied.

Review of process monitoring techniques applied in metal etching industry

This section reviews all the process monitoring techniques applied in the last two decades to the aforementioned metal etching process. The review is presented in different categories. Each category presents a review of all the works concerning a particular technique.

Table 1 – *Machine state variables for process monitoring*[3]

1	BCl_3 flow	8	RF tuner	15	TCP impedance
2	Cl_2 flow	9	RF load	16	TCP top power
3	RF bottom power	10	Phase error	17	TCP reflectance
4	RFB reflected power	11	RF power	18	TCP load
5	Endpoint A detector	12	RF impedance	19	Vat valve
6	Helium pressure	13	TCP tuner		
7	Chamber pressure	14	TCP phase error		

Table 2 – *Fault types detected in the metal etching process*[3]

Fault No.	Fault description
1	TCP+50
2	RF-12
3	RF+10
4	Pr+3
5	TCP+10
6	BCl_3+ 5
7	Pr-2
8	Cl_2-5
9	He Chuck
10	TCP+30
11	Cl_2+5
12	RF+8
13	BCl_3- 5
14	Pr+2
15	TCP-20
16	TCP-15
17	Cl_2-10
18	RF-12
19	BCl_3+10
20	Pr+1
21	TCP+20

Principal Component Analysis (PCA) based techniques

Among all the fault detection techniques, multivariate statistical fault detection methods, like the principal component analysis (PCA) and its modified versions have drawn maximum interest in application to the metal etching process of semiconductor manufacturing[5]. Wise *et al.* were among the first authors to have implemented PCA in the field of semiconductor metal-etching industry[3]. They conducted a comparative research on PCA and multiway PCA (organizing of data into time-ordered blocks of the original data) along with other techniques for the metal etching process. PCA can identify a combination of variables that depict a major trend in a dataset, while the multiway PCA considers that data are collected in a sequential manner, providing identical results even after reordering of samples. This technique was termed multi-way because it handles time-series data ordered into different blocks according to time where each block may represent a process run or a single sample. They are of particular importance for the analysis of batch process data. Analysis shows PCA worked better on raw data than multiway PCA, multiway PCA did not perform better than PCA of the raw data, but with different arrangements of data, MPCA showed better results on the machine state data and OES, but not the RFM. Faults that change the shape of the process trajectory (stretching or shortening of the etching process) without altering the overall mean and covariance are detected by multiway PCA, but not PCA. Even though multiway PCA performed better than PCA, PCA is preferred due to its simple model nature. It was observed that none of the methods applied could account for process drifts.

A typical process dataset usually contains values of variables that are highly correlated. This information is hidden in the form of combinations of variables: the latent structure. Hence, it is better to model the non-correlated groups of variables separately using a non-linear model or local linear models. Camacho and Picó, came up with a new approach of PCA called multiphase PCA (MPPCA) to deal with the non-linear nature of semiconductor manufacturing batch process, which the traditional linear models like PCA/PLS fail to consider[6]. Here, multistage models are constructed by designing different models for different stages. The stage depicts the time segment corresponding to one unit operation. The models are generated independently and they are local to the stage. MPPCA model is applied for the detection of phases (the segments of the batch that are well approximated by a linear model) in a batch process. The results obtained after comparative studies show that MPPCA gives the highest prediction power by adjusting the model to the nature of the process. This model does not require post processing, unless one wants to reduce the number of submodels. This way, the computational load and application of a non-linear model to the entire process are avoided. No prior knowledge of the process is required, as the parameters are intuitive and defined to be independent of the specific process. This article opens a new scope in the research of various design of other algorithms within the MPPCA framework, and uses fuzzy transitions between phases or Gaussian mixture models to improve the non-linear dynamic modelling.

Since multiway PCA appeared to show satisfactory results, Han *et al.* applied this technique for end point detection (end point is the instance where plasma etching is stopped in order to prevent overetching) in plasma etching using multiple wafers for real-time prediction of EPD (end point detection) with normal wafer data, and used the entire optical emission spectra for EPD[7]. Two models are included in this algorithm, namely, single wafer model, and multiple wafer model. The former model uses the PCA loading vector of the first wafer as a predictor for the EPD time of the second wafer, and the latter uses the multiway PCA loading vector of previous wafers for the EPD time of the target wafer. When this method was applied in the metal etching process, multiway PCA was more robust and reliable than PCA for EPD, as PCA with the single wavelength signal could not overcome the noise issues or include the drift of the process.

As the process becomes complex, the number of variables increases, and with faster sampling rates, the storage grows drastically. In order to break down the size of the existing PCA model, Good *et al.* came up with a unified PCA (UPCA) model, where the problem was broken down into smaller units, and PCA model was applied to those units[8]. This way, the computational time for model generation and adaptation was reduced dramatically and could automatically group variables. This model does not take into account the correlations that practically exist among variables; thus, UPCA is only an approximation of the PCA model. It has been applied on a wafer electrical test (WET) data, where 484 variables were grouped into 15 blocks with little or no interactions between blocks. Results demonstrated that, when no interblock correlations were considered, PCA and UPCA were identical, but the model size had reduced by 89.4 % with UPCA. The model has the ability to create and modify models *in situ,* and to use the process knowledge to disallow insignificant correlation between unrelated variables. The drawback of this model was that it caused additional false alarms and missing data than that with PCA modelling. Hence, this model would be appropriate and efficient to group variables automatically.

Most of the monitoring methods assume that the collected batch dataset has a three-way array, implying that all the batches should have the same batch length and each variable an identical sampling rate. In reality, most of the batch processes do not comply with these assumptions because as the chemical composition of the raw material and environmental conditions vary, batch lengths become unequal. Hence, Wang and Yao[9] proposed a new model called multivariate functional kernel principal component analysis (MFKPCA), where the variable trajectories are not considered as just vectors, but as smooth functions because they often show functional behaviors. This characteristic of the variables makes it easier to transform the collected three-way data into a two-way function matrix. This proposed model combines the functional data analysis (FDA) and nonlinear PCA. The FDA handles the subtle differences between the variable trajectories of normal and faulty samples, as well as the unequal length of batch by eliminating random noises. To the developed multivariate functional data, the kernel trick is applied to handle the nonlinear relationships among variables and sampling times. In addition to SPE and T^2 statistics for fault detection, mean squared error (MSE) statistics is also utilized in MFKPCA to take into account the fitting errors that leave behind variation information about variable trajectories between normal batches and faulty ones. Results from the metal etching process showed that the average testing accuracy of MFKPCA was higher than other methods, and showed similar missing alarm rates as that of multiway PCA and multiway KPCA, but lower false alarm rate. Guo *et al.* came up with a modified approach of PCA known as weighted difference principal component analysis (WDPCA) to deal with nonlinearity and multimodal characteristics of complex industrial processes[10]. This approach in the preprocessing stage uses weighted difference method on the original data to get rid of the multimodal and nonlinear characteristics. To this preprocessed data, PCA model is applied effectively, since the data approximately follows Gaussian distribution. The authors applied WDPCA along with 5 other models (PCA, KPCA, ICA, kNN, LOF) on a semiconductor manufacturing industry and compared their results. The analysis showed that WDPCA was able to detect all 20 faults with SPE index and almost all with the T^2 index, whereas the SPE index of PCA could detect only 17 faults. The performance of KPCA, ICA, kNN, and LOF were poor in detecting all the faults. Compared with the other 5 methods, there were no false alarm rates of the T^2 index and missing alarm rates of the SPE index of the WDPCA model.

In real scenarios, the faulty samples are usually in an extreme imbalance ratio in comparison with the normal samples. This creates the issue of class imbalance during fault detection. Wang *et al.* worked on a feature extraction method based on kernel principal component analysis (KPCA), a self-supervised learning method to achieve rapid detection of faults or abnormal samples[11]. This is a data-driven method of fault detection, which only uses normal data samples for training, where principal components are extracted using KPCA from the raw data, and based on how high the reconstruction error of the testing sample is than the training sample in the feature space, the fault is identified. The reconstruction error of the sample is compared with the threshold, which is the maximum reconstruction error of the training normal samples. During testing, all the samples, including normal and faulty ones, are projected on the eigenvector. Since eigenvector are the principal components of the samples, faulty samples are thereby detected as those with high reconstruction error. Results showed that this approach of self-learning was able to detect all the 21 faults in the metal etching process.

It is known to us that the wafer processing industry deals with three-dimensional array and nonlinear data characteristics, and for this process a simple multiway PCA may not serve the purpose. Hence, Zhang *et al.* proposed a novel technique called multiway principal polynomials (MPPA)[12]. This nonlinear modelling technique captures the nonlinear nature of the data by replacing the straight PCs in PCA with curved principal polynomials. This method learns a low-dimensional representation from a process data on the basis of sequential principal polynomials. When tested on the wafer processing data, SPE of MPPA performed better than MPCA, FD-kNN and PC-kNN in effectively capturing the nonlinear nature of process data.

Hybrid PCA techniques

In this section, combination of two or more fault detection techniques keeping PCA as a base technique are reviewed.

To eliminate the drawbacks of certain multivariate statistical fault detection methods, namely, nonlinearity and multimode handling, He and Wang proposed a modified fast pattern recognition-based method that combines the benefits of both PCA and k nearest neighbors (FD-kNN), principal component-based kNN (PC-kNN)[13]. Even though FD-kNN alone can overcome the mentioned characteristics of a semiconductor batch process, it becomes highly complex in computation and needs intensive storage/memory requirement. The new approach, PC-kNN, is able to sort this issue. Two major steps are involved in this method; firstly, PCA is applied to the original process data for dimensionality reduction and extracting the key process features. Secondly, fault detection is done by applying FD-kNN on the obtained principal subspace by PCA. The conducted experiments suggested that PC-kNN was capable of detecting abnormalities based on local neighborhoods, and handled nonlinearity and multimodal distributions naturally. PC-kNN tended to out-perform both PCA T^2 and FD-kNN. The faults detected by PC-kNN and SPE were combined to obtain a full coverage of fault detection since PC-kNN does not take into account the residual subspace. As a future aspect, one can try to figure out how a single PC-kNN model can be used for the entire process plant, as they can handle multimodal data, rather than using hundreds of context-specific PCA models.

Ge and Song proposed an adaptive method that can overcome the difficulties seen in multiway PCA method called the substatistical PCA method[14]. Hence, a substatistical PCA method is introduced that avoids future value estimation, and can be used for non- Gaussian process data by employing support vector data description (SVDD). Using the kernel learning technology, SVDD is able to deal with small data samples and form a flexible boundary around the process monitoring while adapting to the shape of the samples. This method is also able to obtain cross-information between data blocks with lower time complexities. Even though most of the conventional issues of multivariate statistical process control (MSPC) are resolved by this method, it is still not able to handle nonlinearity, and is applicable for stationary cases only.

Yu came up with a new model that eliminates the major drawback of PCA model, i.e., the assumption that the data are Gaussian distributed, and the confidence bounds are set based on this assumption, which is not the case in reality[15]. This approach combines the benefits of PCA, i.e., dimensionality reduction and the Gaussian mixture model (GMM), i.e., handling nonlinearity and multimodal batch trajectories, along with two process state quantifying indexes, negative log likelihood probability (NLLP), and Mahalanobis distance (MD) with failure probability (BIP) being proposed to assess the process states for fault detection. In order to avoid inputs with high dimensionality and sparsity, PCs generated by the PCA model are fed as inputs to the GMM model. Hence, the model takes the name Principal Component Gaussian Mixture Model (PCGMM). When this model was applied on a semiconductor dataset, results showed that the charts of PCGMM for both NLLP and MD indices detected all 20 faults and their failure probabilities (BIP) were all 100 %. In conclusion, PCGMM-NLLP and PCG-

MM-ND outperformed PCA-T^2 and PCA-SPE control charts in detecting various faults for multimodal data characteristics.

Even though KPCA manages to work brilliantly on nonlinear data, it fails to consider the multimodal nature of the data, which happens to be the case for semiconductor manufacturing process. Thereby, Zhang *et al.* proposed a model that combined NND rule with KPCA to reduce the impact of multimodality, known as NND based KPCA[16]. Multimodal trajectories are eliminated by this technique by applying the NND rule on the raw dataset first, guaranteeing approximately Gaussian distribution, thereby making the application of kernel PCA on the data in the NND subspace effective, and which just handles nonlinear nature of the data. In comparison to FD-kNN, KPCA and other multivariate statistical process monitoring (MSPM) methods, NND-KPCA demonstrated 100 % fault detection rate, and had a lower false alarm rate than KPCA.

Gaussian Mixture Model (GMM) based techniques

A novel model, Infinite Gaussian Mixture Model (GMM) was proposed by Chen *et al.*[17], to provide confidence bounds to detect any process deviation from normal operation without the assumption of Gaussian data. This model is a special case of Dirichlet process mixture, and is a limit of the finite GMM. It provides a more accurate calculation of confidence bounds as it uses a Bayesian approach to estimate the probability density function (PDF) of the process data. Chen and Zhang proposed a model that deals with online monitoring for batch processes that are not Gaussian-distributed using the Gaussian Mixture Model method (GMM)[18]. The historical set of data collected based on normal operating conditions, is used to set confidence bounds for monitoring statistics without the assumption of Gaussian data, to detect deviation of the process. Here the entire data of old batch is needed, and only up to the current time as that of the new batch for monitoring statistics. Firstly, MPCA is applied on the batch data to extract and obtain low-dimensional representation of the process. Next, GMM is applied to obtain the joint PDF (Probability Density Function) of the predicted monitoring statistics from MPCA at each time step. The diagnosis of the fault detected is done by contribution analysis method in which the variable that defies the confidence bounds is identified. From the case study of semiconductor manufacturing industry, it is clear that the GMM is a promising method in terms of lower false alarm, and is accurate at calculating the confidence bounds in the online monitoring of batch processes.

Yu combined the benefits of the local and nonlocal preserving projection (LNPP) and GMM, and came up with a new model called GMM with Local and Nonlocal preserving projection[19]. This model aims at extracting local and nonlocal information from the process data to improve the performance of GMM with dimensionality reduction. LNPP has the ability to distinguish directions keeping the local and nonlocal structural information intact for given data. Unlike the PCA, LNPP is able to identify low-dimensional information hidden within high-dimensional observations. Based on the extracted information from LNPP, GMM is applied for the estimation of PDF of semiconductor manufacturing process data. GMM, along with a quantification index, Mahalanobis distance, was proposed to detect the process states with faults. Bayesian inference-based method was proposed to provide the process failure probability. The proposed model outperformed PCA-based monitoring models.

k-Nearest Neighbor Rule (kNN) based techniques

He and Wang developed the fault detection method using the k-nearest neighbor rule (FD-kNN) to handle characteristics like nonlinearity, multimodal batch trajectory due to product mix, and variable durations during process steps in a semiconductor process industry, that are difficult to identify with multivariate statistical fault detection methods such as principal component analysis (PCA)[20]. The kNN rule classifies unlabeled samples based on their similarities with samples in the training set, making it suitable for application in pattern classification. The normal operation data are the only data available as a training set. This barrier is overcome by adapting the traditional kNN rule and making use of only normal operation data. The FD-kNN method reported in the article is built on the idea that the trajectories of the incoming normal sample and that of the training samples (i.e., of normal operation data only) are similar, whereas the trajectory of an incoming faulty sample shows deviation from that of the normal training samples. The FD-kNN method handles nonlinearity and multimodal trajectories naturally, as the method makes no assumption of linearity and detects faults based on local neighborhoods. This method takes a crucial place in online process monitoring, as the preprocessing is done automatically. In addition, in industries with limited preprocessing data, FD-kNN method seems to outperform the PCA method. This does not necessarily imply that FD-kNN outperforms PCA in all cases. Li *et al.* proposed a new scheme, Just-in-time (JIT) and kNN-based integrated model. JIT detection method can store the current measured data in the database, enabling it to be flexible and adaptive inherently[21]. It can also detect where the query is not normal through online and adaptive approach. Based on the Mahalanobis distance between the normal samples, the raw data sets are simplified and updated. The time-varying control limit (CL) of the updated database is regulated through kNN rule combined with SPC method. This is done every time a fault detection has to be conducted. From the case study of semiconductor process industry, it is very clear that the proposed model is well suited for nonlinear, dynamic, and multimodal processes. A novel method for fault detection, called Diffusion maps-based kNN (DM-kNN), has been presented by Li and Zhan[22]. This model reduces training sample storage and deals with nonlinear dataset. Diffusion maps is defined on a graph of data points by constructing a Markov chain. It is a robust nonlinear manifold dimensionality reduction technique while preserving the intrinsic geometrical structure of the dataset. High-dimensional data is transformed to low-dimensional featured space with intrinsic dimensionality. To this low-dimensional dataset, kNN rule is applied to detect potential faults. When this method was applied on a metal etching process, the results demonstrated that the model detected all the 20 faults, and developed method demonstrated effective monitoring with superior fault detection performance in comparison with other techniques like MPCA, FD-kNN, and PC-kNN.

Zhou *et al.* introduced a new fault detection method, which merges the benefits of random projection and kNN (RPkNN)[23]. Random projection not only reduces the computational complexity and storage space, but also preserves the distance of pairwise samples in the random subspace, which is not possible in the case of PC-kNN. This reduces the false alarm rate and the missing detection. This is further combined with the kNN rule to deal with multimodal nature of batch and nonlinearity. On analysis of a semiconductor process data, results showed that the fault detection capability of RP-kNN was identical to that of FD-kNN, but resulted in dramatic reduction in computational complexity and storage space. When the distance distortion of RP is compared with that of PCA, it is observed that RP causes very limited changes and most of the information is retained, which makes dimensionality reduction much more effective than PCA.

When the dispersion degree of different modes are similar, the detection performance of kNN is acceptable. However, when the dispersion degrees of different modes are not similar, minor faults go undetected. Guo *et al.* proposed a kNN-based probability density (PD-kNN) model that ignores the different degrees of dispersion between modes, and classifies the modes based on probability density[24]. The proposed model overcomes the shortcoming of kNN by applying PD to determine the mode of the new test data. The test data is detected by the existing kNN model using the training data of its respective mode. Hence, PD-kNN model detects weak faults having a lower dispersion degree that are submerged by the normal data in the mode with higher dispersion degree in a multimodal data. Zhang *et al.* came up with an alternative model to the PD-kNN called Weighted Distance-based kNN (FD-wkNN) to deal with the same issue of detecting weak faults[25]. To overcome this problem, adjustments of the squared distance of a sample to *k*-nearest neighbors (D^2) of different modes are done to the same scale through weighted parameters.

Support Vector Data Description (SVDD) based techniques

A process monitoring task requires to distinguish normal samples from the faulty samples, grouping the normal data as one class. Hence, this monitoring can be considered as a one-class classification problem. One-class SVDD was applied for the first time for batch process monitoring by Ge *et al.*[26] The goal of this approach is to group all the normal process data samples into one class, so as to differentiate from the faulty samples with no Gaussian limitation, and make it efficient for nonlinear cases. A sub-SVDD model has also been developed for multiphase batch processes. This method proves to be computationally efficient, as the model only incorporates a quadratic optimization step. Khediri *et al.* presented a procedure called Kernel k-means (KK-means) clustering-based local SVDD to monitor processes with multimodal and nonlinear characteristics[27]. This model is based on separate models for different process modes. Using the clustering method, specifically Kernel k-means, if the separation boundaries between clusters are non-linear, the process modes are separated based on the similarities using sum-of-square criterion. Different SVDD models are used for each cluster. This way, not only the faults are identified, but also the process modes in which they occur. For accurate fault detection based on a spatiotemporal pattern classification approach, Chang *et al.* developed a new classifier known as Sequential SVDD. The proposed technique models the characteristic areas of flow and sequential flow of process data[28]. SVDD handles the non-Gaussian and nonlinear data characteristics, whereas the sequential modelling handles the sequential characteristics. As soon as a process abnormality occurs, the fault can be detected quickly by checking the start and end-points of the process using the ordered SVDDs. This information is used for the diagnosis of fault sources. In this model, PCA is first applied on the original dataset to reduce the dimensionality and for feature extraction. From these featured areas, critical ones that are abnormal are selected. A sequential classifier describing the selected data areas are configured with SVDD and their sequential relationships with a Gaussian distribution. To improve the functioning and performance of SVDD model, Ge and Song developed an ensemble form of the same: Bagging SVDD (BagSVDD)[29]. Instead of using a single SVDD model, an ensemble model technique (bagging) is used to develop multiple models based on various SVDD techniques using subdatasets. Monitoring performances of different sub-SVDD models are combined using voting-based (uses cut off value) method and Bayesian-based (uses probability) method to form a final monitoring result.

Based on functional data description of the batch dataset, Yao *et al.* extended the conventional SVDD to the functional aspect, called functional SVDD[30]. In this approach, a three-way array is transformed into a two-way array by considering each variable's time-varying trajectory as sampled functional data. It is very well known to us that the SVDD method of monitoring captures the spherical boundary around the normal data and sets off control limits based on support vectors (SVs). When applied for a complex batch process, the accuracy of monitoring decreases with this set control limit. Hence, Wang *et al.* proposed a Dynamic hypersphere-based SVDD (DH-SVDD) model for batch process monitoring. DH-SVDD improves the monitoring accuracy of the existing SVDD models[31]. Here, a static hypersphere is built on the historical training data first, and then combining a test sample and the training data, a dynamic hypersphere is built. If the current sample is faulty, then a significant change between the static and dynamic hyperspheres will be detected. To validate this, the model was tested on the semiconductor etching process, and based on the results, the accuracy of fault detection was the best among the eight tested methods.

Deep learning techniques

When it comes to studying and analyzing complex process data, there is a large number of highly correlated variables, which can be easily tackled these days with the help of deep learning tools. Traditional machine-learning techniques require preprocessing of raw data based on user's prior knowledge and expertise before application of techniques on the data. On the other hand, deep-learning techniques possess multiple processing layers, which carry out feature extraction as well as modeling. In the context of industrial applications, they are capable of diagnosing faulty process conditions by composing a number of nonlinear approximations. Lv *et al.* used a stacked sparse auto-encoder (SSAE) network, which identified minute details and changes in a fault signal[32]. By increasing the number of layer to be stacked, more nonlinearities can be characterized through higher order correlations. The detection performance when illustrated on metal etching process data showed superiority over many other process monitoring methods, and highlighted that the performance can be further improved by increasing the number of normal samples in the training phase.

While using standard classification algorithms, there is a high chance that information is lost from the trace data (sequences of sensor readings) while extracting statistical features and therefore, there is a possibility of failure to consider class imbalance situations. Since most of the data used for fault detection and classification (FDC) are class-imbalanced, many times faulty wafers are also classified as normal wafers, and the sensor noise and wafer-to-wafer (W2W) variations are not considered in usually implemented techniques, one-class classifiers are implemented. In order to incorporate W2W variations and sensor noises, Jang *et al.* developed a one-class FDC model based on Denoised auto-encoder (DAE) and analyzed the residual trace[33]. DAE is a feed-forward neural network with only one hidden layer. It can decompose trace data into ideal trace, sensor noise, W2W variations and abnormal patterns, and provide an ideal trace as an output. So if the wafer is faulty, the abnormal patterns and sensor noises will be available in the difference between the input and output traces (residual trace). To segregate the noises from the abnormal patterns, any residuals in the denoised residual trace (DRT) below the threshold will be considered as sensor noises and be removed, and the remaining residuals will be carriers of information regarding abnormal patterns. Etch results showed that a limited number of normal samples were sufficient to train the model, remove W2W variations and sensor noises, as well as detect abnormal patterns. In addition, using the DRT information, the proposed method could identify the process parameters responsible for the wafer faults, and help in shaping and give time details of abnormal pattern occurrence.

Miscellaneous techniques

As discussed earlier in this article, Wise *et al.* were among the first researchers to study the models developed from the obtained data of the three sensor systems of a metal etching process, namely, principal component analysis (PCA), multiway PCA, trilinear decomposition (TLD), and parallel factor analysis (PARAFAC)[3]. TLD and PARAFAC are multiway methods. However, unlike MPCA (which depends on PCA and rearrangement of the original data), TLD and PARAFAC handle and interpret the convolute time and variable information better. After analysis and comparison of the models, it was observed that PARAFAC performed marginally better than TLD. This could be due to the imaginary solutions obtained in TLD, which tend to be a problem. It was also observed that even though the overall performance of the sensors was similar, OES sensor appeared to degrade the most as models changed from local to global. This is due to the large amount of drift in the OES signals due to residue build-up. Whereas, there were minimal variations in the sensitivity of RFM models (local or global), hence suggesting that RFM sensors are the most stable and/or least sensitive to many changes that do not affect processing. In summary, PARAFAC worked the best, closely followed by PCA on the means and TLD. However, due to the simplicity of PCA algorithm, it is preferred for practice.

It is not always the case that a batch process takes the same amount of time in every run. Even if so, they will not necessarily follow exactly the same time trajectory. They may take longer time in some processing steps and less time in others. Hence, the data obtained for each batch may be of different length. Focusing on this very aspect, Wise *et al.* developed and worked on stretching of time axis, using the PARAFAC2 model[34]. Unlike PCA, TLD, and PARAFAC techniques, this approach does not approximate the range of process trajectories as the sum over fixed time profiles. A major advantage of PARAFAC2 model is that the original data of unequal batch lengths can be used directly without going through preprocessing methods. From the results obtained after fault detection, one can infer that the sensitivity of PARAFAC2 model was somewhat higher in comparison with the MPCA, but not significantly more sensitive than PARAFAC. Besides, if fitted time profiles were not considered, PARAFAC2 model would rearrange the data records without stipulating a fault, making it a major disadvantageous factor of the model. The main benefit of the PARAFAC2 model is that the data records of varying length do not require preprocessing.

Lee *et al.* introduced a novel MSPM method which is a modified version of the existing independent component analysis (ICA)[35]. ICA is itself an improved method of the principal component analysis (PCA) overcoming the shortcomings of PCA, like its inability to provide lower-order representation for non-Gaussian data. ICA extracts the important components that influence the process, and the ICs extracted are monitored rather than the original data. Nonetheless, it is nearly impossible to know how many ICs need to be extracted to obtain a stable ICA model, which is why it increases the computational load. The order of the ICs cannot be determined in ICA unlike PCA (where PCs are arranged in descending order). All these disadvantages are overcome in the modified ICA. This model uses PCA to estimate the initial ICs, and the dominant ICs are calculated using the conventional ICA without changing the variance. When this novel method was applied on the semiconductor etching process for fault detection and diagnosis, it gave more satisfactory results than the PCA. Pattern changes were reflected better by the extracted dominant ICs. It also revealed the group of process variables responsible for the out-of-control action of the processes from the contribution plots. This analysis shows that the proposed method is promising for process monitoring.

The focus of Yu and Wang was on Neural Networks (NNs) as a technique of multivariate statistical process control (MSPC)[36]. The reason being exceptional noise tolerance in real time, which requires no hypothesis on statistical distribution of monitored measurements. Most of the NNs are based on supervised learning, which means that the abnormalities of the process needs to be known before hand as a feed to the training dataset. This is very difficult to acquire in real industries, but the normal operating datasets are much easier to extract. Hence, a new approach of NNs was introduced, where only the normal operating datasets are required for the neural system. This approach is based on the Self-organizing Map (SOM) in which the training is adaptive to the data that has been input with no human intervention during learning. Using the Minimum Quantization Error (MQE) calculation, this approach is able to provide a much more accessible and quantitative estimate for the current process state. Based on these estimated MQE values, a MQE chart is prepared to study the process behavior. After conducting experimental studies on a bivariate process and the semiconductor batch process, it was concluded that the MQE chart was much more effective and robust than other MSPC tools in detecting minute process shifts.

Mahadevan and Shah proposed a supervised model one-class SVM, which is a variant of the traditional SVM algorithm[37]. This model is trained only from normal data, unlike the traditional SVM, which is developed from both normal and faulty data. The objective of this technique is to detect any outliers (faulty samples). This approach handles nonlinearity with the help of kernel functions. Just like the PCA and DPCA uses T^2 and SPE statistics as distance metrics and threshold for fault detection, this model is based on a single nonlinear distance metric measure. Due to its ability in handling multimodality, the one-class SVM model had better results than PCA technique in detecting faulty wafers in the semiconductor etching process.

Yu constructed Hidden Markov Model (HMM) combining local and global information of Gaussian component hidden in the HMM to identify process faults for nonlinear and multimodal process[38]. For this purpose, two novel quantification statistical models are introduced, namely, MDNLLP (Mahalanobis distance combined with negative log likelihood probability), and BIP (Bayesian inference-based probability). Based on the information obtained by these quantification models, process faults are detected around nonlinear and multimodal operating areas. The implementation of this model of HMM-based MDNLLP and BIP on the semiconductor batch process proved to be effective in fault detection with very few false alarm.

Puggini *et al.* applied an unsupervised method called Random Forest Similarity Distance, which uses random forests and decision trees to identify the faulty wafers merely by observing the chemical signatures during plasma etching[39]. Random forest is an ensemble learning method where a single decision is obtained based on multiple outputs. To detect the anomaly, the distance of the new wafer from normally behaving data is measured so the dissimilarity in the distance can be compared with the previously observed normal behaving data. Results show that the method can capture the effects of process drifts, along with identifying faulty wafers effectively. A major drawback of this method is that, when the training dataset varies with time, it is not possible to compare their distance measurements, because the similarity distance evaluated is a relative measurement of proximity.

A data-mining based algorithm, incremental clustering-based fault detection, proposed by Kwak *et al.* detects faulty wafers even in the case of severely imbalanced class distribution and with process drifts[40]. Here, the normal data is clustered as one, and once a new wafer is added, its class label is calculated using Mahalanobis distance. The statistical summary (prototype) of the closest cluster is updated to the new wafer, and if it happens to be a faulty one, then a new single-member cluster is created, and merging operation is initiated for cluster overlaps ahead.

An unsupervised technique, Nonlocal structure constrained neighborhood preserving embedding (NSC-NPE) was proposed by Miao *et al.* for dimensionality reduction[41]. NSC-NPE is an algorithm based on global information. Both local and nonlocal information are considered at the same time with metric preserving properties. The local scatter is minimized and nonlocal scatter is maximized for the best possible mapping of data structure. A detailed data structure can be mapped out from this information either hidden in the neighborhoods or of distinguished remote data points belonging to different neighborhoods.

To deal with multimodality and unequal length of data, Guo *et al.* developed a method called Local neighbor normalized matrix (LNNM)[42]. Nonlinear relations are identified between modes and within modes. Initially, local weighted algorithm (LWA) is used for preprocessing of the unequal length of batch data, and thereafter LNNM is constructed for the equal length of data generated. Using the K-means algorithm mode, clustering is done, and in each mode, LOF method is used for determining the first control limits to remove outliers. MPCA is applied to each mode to determine control limits of multiple modes. Based on the results, the approach of LNNM was more effective in fault detection than most of the common techniques like MPCA, KNN, and MKPCA, with a fault detection rate of 100 % and false alarm rate of 9.1 %.

Locality preserving projections (LPP), a MSPM method like the PCA when combined with a new difference preprocessing data algorithm for nonlinear and multimodal data gives a new model, Difference locality preserving projections (DIF-LPP) developed by Guo *et al.*[43] This method requires no prior knowledge of the process, and processes the data into a Gaussian fit and a single mode before applying LPP to the dataset. DIF can also be combined with other MSPM techniques to improve fault detection.

The survey work aforepresented is summarized in chronological order in Table 3. It may be noted that all the works reported in Table 3 are based on the public dataset available at http://www.eigenvector.com/data/Etch/index.html. All works focus on detecting all or some of the 21 artificially introduced faults, which are described in Table 2, by utilizing the same input variables presented in Table 1.

Critical analysis of the review

Fig. 3 presents a pie chart explaining the percentage of techniques that have been applied to this process.

It can be noticed in Fig. 3 that PCA-based techniques (traditional/modified PCA or PCA in combination with some other techniques) still account for almost fifty percent of all techniques. Among machine learning techniques, in the last decade, independent component analysis (ICA) has become an attractive alternative to PCA for non-Gaussian data. However, there are only a few investigations of ICA-based techniques applied to the metal etching process. Furthermore, there is also scope for further exploration of various deep learning techniques for fault detection and diagnosis in this process.

It is known that there are 21 faults present in the metal etching process public dataset, and many articles have reported their fault detection rates

Table 3 – *Summary of literature review on fault detection in metal etching process*

Researchers	Year	Techniques applied	
Wise *et al.*[3]	1999	PCA-based	Multiway PCA (MPCA)
Camacho and Picó[6]	2006		Multiphase PCA (MPPCA)
Han *et al.*[7]	2008		Multiway PCA for EPD
Good *et al.*[8]	2010		Unified PCA (UPCA)
Wang and Yao[9]	2015		Multivariate functional kernel PCA (MFKPCA)
Guo *et al.*[10]	2017		Weighted Difference PCA (WDPCA)
Wang *et al.*[11]	2018		Kernel PCA (KPCA)
Zhang *et al.*[12]	2018		Multiway Principal Polynomials Analysis (MPPA)
He and Wang[13]	2010	Hybrid PCA-based	Principal Component-based kNN (PC-kNN)
Ge and Song[14]	2010		Substatistical PCA (SVDD combined with PCA)
Yu[15]	2011		Principal Component Gaussian Mixture Model (PCGMM)
Zhang *et al.*[16]	2017		Nearest neighbor difference rule–based KPCA (NND-KPCA)
Chen *et al.*[17]	2006	GMM-based	Infinite GMM
Chen and Zhang[18]	2012		MPCA and GMM
Yu[19]	2012		Local and nonlocal preserving projection (LNPP) and GMM
He and Wang[20]	2007	kNN-based	Fault detection method using the k-nearest neighbor rule (FD-kNN)
Li *et al.*[21]	2012		Just-in-time (JIT) and kNN-based integrated model
Li and Zhang[22]	2014		Diffusion maps-based kNN (DM-kNN)
Zhou *et al.*[23]	2015		Random Projection and kNN (RPkNN)
Guo *et al.*[24]	2018		kNN-based probability density (PD-kNN)
Zhang *et al.*[25]	2019		Weighted Distance-based kNN (FD-wkNN)
Ge *et al.*[26]	2011	SVDD-based	One-class SVDD
Khediri *et al.*[27]	2012		Kernel k-means (KK-means) clustering-based local SVDD
Chang *et al.*[28]	2012		Sequential SVDD
Ge and Song[29]	2013		Bagging SVDD (BagSVDD)
Yao *et al.*[30]	2014		Functional SVDD (FSVDD)
Wang *et al.*[31]	2018		Dynamic hypersphere-based SVDD (DH-SVDD)
Lv *et al.*[32]	2018	Deep learning-based	Stacked sparse auto-encoder (SSAE)
Jang *et al.*[33]	2019		Denoised auto-encoder (DAE)
Wise *et al.*[3]	1999	Miscellaneous	Trilinear decomposition (TLD) and parallel factor analysis (PARAFAC)
Wise *et al.*[34]	2001		Parallel factor analysis 2 (PARAFAC2)
Lee *et al.*[35]	2006		Modified Independent Component Analysis (ICA)
Yu and Wang[36]	2009		Self-organizing Map (SOM)
Mahadevan and Shah[37]	2009		One-class Support Vector Machines (SVM)
Yu[38]	2010		Hidden Markov Model (HMM)
Puggini *et al.*[39]	2015		Random Forest Similarity Distance
Kwak *et al.*[40]	2015		Incremental Clustering-based Fault Detection
Miao *et al.*[41]	2015		Nonlocal structure constrained neighborhood preserving embedding (NSC-NPE)
Guo *et al.*[42]	2016		Local neighbor normalized matrix (LNNM)
Guo *et al.*[43]	2018		Difference locality preserving projections (DIF-LPP)

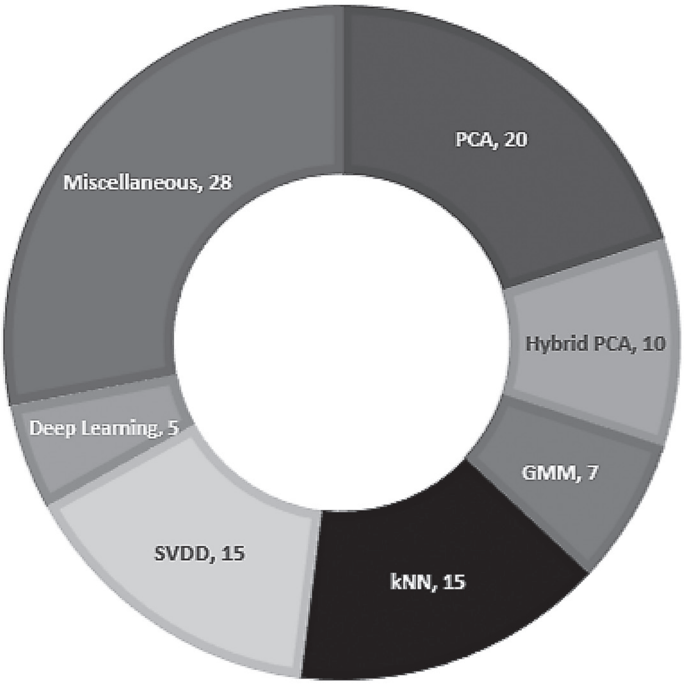

Fig. 3 – *Percentage distribution of techniques analyzed*

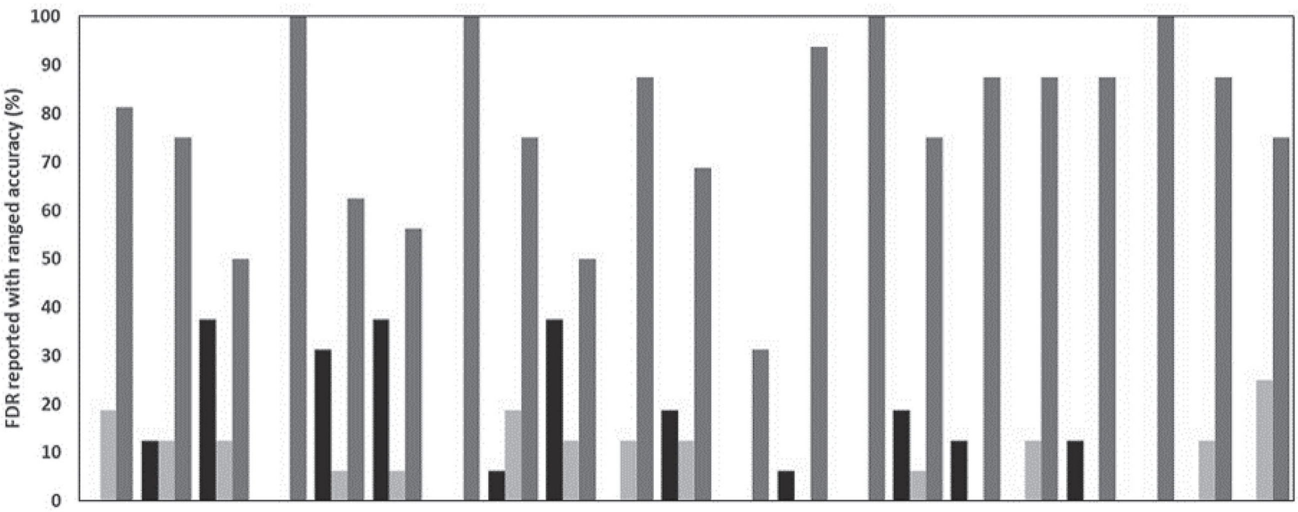

Fig. 4 – *Graph of FDR reported for all 21 faults*

(FDR) with certain accuracy. The FDR results reported in each research work are compiled for all the 21 faults and presented as a bar chart in Fig. 4 for readers' quick insight.

Fig. 4 represents the percentage of articles that have detected the faults, categorized based on their reported ranges of detection rate. In Fig. 4, low depicts the FDR range of a particular fault falling between 0–20 %. Similarly, medium depicts the FDR range of 20–70 %, and high depicts the FDR range of 70–100 %. There are 16 articles that have reported the FDR of each fault. From the graph, it is clear that fault numbers 1, 4, 7, 10, 13, 14, 16, 17, 18, 19, and 20 have been reported with high FDR, out of which fault numbers 4, 7, 14, and 19 have a very high FDR comparatively. Most of the articles have studied only 20 faults, excluding fault number 12, thus, not much information can be gathered for this fault regarding its FDR. Other faults have not been reported with satisfactory FDRs, and a lot of work and analysis needs to be done to improve their FDRs.

Figs. 3 and 4 are expected to serve as ready references for researchers working in this area. These two figures in combination indicate the gaps in existing research in fault detection of metal etching process. They present the techniques frequently applied as well as the fault(s) that pose a challenge in successful detection.

Conclusion and future directions

The different problems associated with the metal etching process can be minimized by development of an efficient process monitoring system. Effective monitoring in semiconductor manufacturing will lead to reduced usage of test wafers as well as reduced scrap. Traditional monitoring techniques include univariate statistical process control charts, such as Shewhart, CUSUM or EWMA charts. However, multivariate nature of the metal etching process renders univariate techniques less effective. Batch processing, multimodal batch trajectory, and highly nonlinear characteristics associated with the metal etching process makes it more difficult and challenging for design of efficient monitoring system. This article presents a review of the different techniques reported so far for fault detection in the metal etching process. From this review, it is very clear that many researchers have explored this field of fault detection techniques in metal etching process over the past two decades. Even though numerous machine learning techniques have been applied and analyzed, deep learning techniques have found very limited application in this process. In the coming years, more research based on deep learning can be applied to this field. Furthermore, independent component analysis (ICA) and its variants, which is an improved version of PCA handling non-Gaussian data, is yet to be explored thoroughly for the metal etching process. More modified versions of ICA, either existing or newly developed methods, can broaden the scope of fault detection in semiconductor industry. In addition, not many researchers have included in their research the fault number 12, making it difficult to gather data for improving its FDR. Thus, more focus needs to be put on detecting the less investigated faults as well as faults that have low reported FDRs so far.

CONFLICT OF INTEREST:

The authors declare that they have no conflict of interest.

References

1. *Ge, Z., Song, Z.,* Bagging support vector data description model for batch process monitoring, J. Process Control. **23** (2013) 1090.
doi: https://doi.org/10.1016/j.jprocont.2013.06.010
2. *Ge, Z., Song, Z., Ding, S. X., Huang, B.,* Data mining and analytics in the process industry: The role of machine learning, IEEE Access. **5** (2017) 20590.
doi: https://doi.org/10.1109/ACCESS.2017.2756872
3. *Wise, B. M., Gallagher, N. B., Butler, S. W., White Jr, D. D., Barna, G. G.,* A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process, J. Chemom. **13** (1999) 379.
doi: https://doi.org/10.1002/(SICI)1099-128X(199905/08)-13:3/4%3C379::AID-CEM556%3E3.0.CO;2-N
4. *Prakash, P. K. S., McLoone, S. F.,* Plasma etch process virtual metrology using aggregative linear regression, International Conference of Soft Computing and Pattern Recognition (SoCPaR) IEEE. 2011, 538.
doi: https://doi.org/10.1109/SoCPaR.2011.6089153
5. *He, Q. P., Wang, J.,* Principal component based k-nearest-neighbor rule for semiconductor process fault detection, American Control Conference IEEE. 2008, 1606.
doi: https://doi.org/10.1109/ACC.2008.4586721
6. *Camacho, J., Picó, J.,* Multi-phase principal component analysis for batch processes modelling, Chemom. Intell. Lab. Syst. **81** (2006) 127.
doi: https://doi.org/10.1016/j.chemolab.2005.11.003
7. *Han, K., Park, K. J., Chae, H., Yoon, E. S.,* Multi-way principal component analysis for the endpoint detection of the metal etch process using the whole optical emission spectra, Korean J. Chem. Eng. **25** (2008) 13.
doi: https://doi.org/10.1007/s11814-008-0003-8
8. *Good, R. P., Kost, D., Cherry, G. A.,* Introducing a unified PCA algorithm for model size reduction, IEEE Trans. Semicond. Manuf. **23** (2010) 201.
doi: https://doi.org/10.1109/TSM.2010.2041263
9. *Wang, H., Yao, M.,* Fault detection of batch processes based on multivariate functional kernel principal component analysis, Chemom. Intell. Lab. Syst. **149** (2015) 78.
doi: https://doi.org/10.1016/j.chemolab.2015.09.018
10. *Guo, J., Wang, X., Li, Y., Wang, G.,* Fault detection based on weighted difference principal component analysis, J. Chemom. **31** (2017) e2926.
doi: https://doi.org/10.1002/cem.2926
11. *Wang, T., Qiao, M., Zhang, M., Yang, Y., Snoussi, H.,* Data-driven prognostic method based on self-supervised learning approaches for fault detection, J. Intell. Manuf. **31** (2020) 1611.
doi: https://doi.org/10.1007/s10845-018-1431-x
12. *Zhang, X., Li, Y.,* Multiway principal polynomial analysis for semiconductor manufacturing process fault detection, Chemom. Intell. Lab. Syst. **181** (2018) 29.
doi: https://doi.org/10.1016/j.chemolab.2018.08.005
13. *He, Q. P., Wang, J.,* Large-scale semiconductor process fault detection using a fast pattern recognition-based method, IEEE Trans. Semicond. Manuf. **23** (2010) 194.
doi: https://doi.org/10.1109/TSM.2010.2041289
14. *Ge, Z., Song, Z.,* Semiconductor manufacturing process monitoring based on adaptive substatistical PCA, IEEE Trans. Semicond. Manuf. **23** (2010) 99.
doi: https://doi.org/10.1109/TSM.2009.2039188
15. *Yu, J.,* Fault detection using principal components-based Gaussian mixture model for semiconductor manufacturing processe, IEEE Trans. Semicond. Manuf. **24** (2011) 432.
doi: https://doi.org/10.1109/TSM.2011.2154850
16. *Zhang, C., Gao, X., Xu, T., Li, Y.,* Nearest neighbor difference rule–based kernel principal component analysis for fault detection in semiconductor manufacturing processes, J. Chemom. **31** (2017) e2888.
doi: https://doi.org/10.1002/cem.2888

17. *Chen, T., Morris, J., Martin, E.,* Probability density estimation via an infinite Gaussian mixture model: Application to statistical process monitoring, J. R. Stat. Soc. Ser. C Appl. Stat. **55** (2006) 699.
doi: https://doi.org/10.1111/j.1467-9876.2006.00560.x
18. *Chen, T., Zhang, J.,* On-line multivariate statistical monitoring of batch processes using Gaussian mixture model, Comput. Chem. Eng. **34** (2010) 500.
doi: https://doi.org/10.1016/j.compchemeng.2009.08.007
19. *Yu, J.,* Semiconductor manufacturing process monitoring using Gaussian mixture model and Bayesian method with local and nonlocal information, IEEE Trans. Semicond. Manuf. **25** (2012) 480.
doi: https://doi.org/10.1109/TSM.2012.2192945
20. *He, Q. P., Wang, J.,* Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes, IEEE Trans. Semicond. Manuf. **20** (2007) 345.
doi: https://doi.org/10.1109/TSM.2007.907607
21. *Li, J., Li, Y., Yu, H., Xie, Y., Zhang, C.,* Adaptive fault detection for complex dynamic processes based on JIT updated data set, J. Appl. Math. 2012.
doi: https://doi.org/10.1155/2012/809243
22. *Li, Y., Zhang, X.,* Diffusion maps based k-nearest-neighbor rule technique for semiconductor manufacturing process fault detection, Chemom. Intell. Lab. Syst. **136** (2014) 47.
doi: https://doi.org/10.1016/j.chemolab.2014.05.003
23. *Zhou, Z., Wen, C., Yang, C.,* Fault detection using random projections and k-nearest neighbor rule for semiconductor manufacturing processes, IEEE Trans. Semicond. Manuf. **28** (2014) 70.
doi: https://doi.org/10.1109/TSM.2014.2374339
24. *Guo, J., Wang, X., Li, Y.,* kNN based on probability density for fault detection in multimodal processes, J. Chemom. **32** (2018) e3021.
doi: https://doi.org/10.1002/cem.3021
25. *Zhang, C., Gao, X., Li, Y., Feng, L.,* Fault detection strategy based on weighted distance of k nearest neighbors for semiconductor manufacturing processes, IEEE Trans. Semicond. Manuf. **32** (2018) 75.
doi: https://doi.org/10.1109/TSM.2018.2857818
26. *Ge, Z., Gao, F., Song, Z.,* Batch process monitoring based on support vector data description method, J. Process Control. **21** (2011) 949.
doi: https://doi.org/10.1016/j.jprocont.2011.02.004
27. *Khediri, I. B., Weihs, C., Limam, M.,* Kernel k-means clustering based local support vector domain description fault detection of multimodal processes, Expert Syst. Appl. **39** (2012) 2166.
doi: https://doi.org/10.1016/j.eswa.2011.07.045
28. *Chang, H. J., Song, D. S., Kim, P. J., Choi, J. Y.,* Spatiotemporal pattern modeling for fault detection and classification in semiconductor manufacturing, IEEE Trans. Semicond. Manuf. **25** (2011) 72.
doi: https://doi.org/10.1109/TSM.2011.2172469
29. *Ge, Z., Song, Z., Gao, F.,* Review of recent research on data-based process monitoring, Ind. Eng. Chem. Res. **52** (2013) 3543.
doi: https://doi.org/10.1021/ie302069q
30. *Yao, M., Wang, H., Xu, W.,* Batch process monitoring based on functional data analysis and support vector data description, J. Process Control. **24** (2014) 1085.
doi: https://doi.org/10.1016/j.jprocont.2014.05.015
31. *Wang, J., Liu, W., Qiu, K., Yu, T., Zhao, L.,* Dynamic hypersphere based support vector data description for batch process monitoring, Chemom. Intell. Lab. Syst. **172** (2018) 17.
doi: https://doi.org/10.1016/j.chemolab.2017.11.002
32. *Lv, F., Wen, C., Liu, M., Bao, Z.,* Higher-order correlation–based multivariate statistical process monitoring, J. Chemom. **32** (2018) e3033.
doi: https://doi.org/10.1002/cem.3033
33. *Jang, J., Min, B. W., Kim, C. O.,* Denoised residual trace analysis for monitoring semiconductor process faults, IEEE Trans. Semicond. Manuf. **32** (2019) 293.
doi: https://doi.org/10.1109/TSM.2019.2916374
34. *Wise, B. M., Gallagher, N. B., Martin, E. B.,* Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch, J. Chemom. **15** (2001) 285.
doi: https://doi.org/10.1002/cem.689
35. *Lee, J. M., Qin, S. J., Lee, I. B.,* Fault detection and diagnosis based on modified independent component analysis, AIChE J. **52** (2006) 3501.
doi: https://doi.org/10.1002/aic.10978
36. *Yu, J. B., Wang, S.,* Using minimum quantization error chart for the monitoring of process states in multivariate manufacturing processes, Comput. Ind. Eng, **57** (2009) 1300.
doi: https://doi.org/10.1016/j.cie.2009.06.009
37. *Mahadevan, S., Shah, S. L.,* Fault detection and diagnosis in process data using one-class support vector machines, J. Process Control. **19** (2009) 1627.
doi: https://doi.org/10.1016/j.jprocont.2009.07.011
38. *Yu, J.,* Hidden Markov models combining local and global information for nonlinear and multimodal process monitoring, J. Process Control. **20** (2010) 344.
doi: https://doi.org/10.1016/j.jprocont.2009.12.002
39. *Puggini, L., Doyle, J., McLoone, S.,* Fault detection using random forest similarity distance, IFAC-PapersOnLine **48** (2015) 583.
doi: https://doi.org/10.1016/j.ifacol.2015.09.589
40. *Kwak, J., Lee, T., Kim, C. O.,* An incremental clustering-based fault detection algorithm for class-imbalanced process data, IEEE Trans. Semicond. Manuf. **28** (2015) 318.
doi: https://doi.org/10.1109/TSM.2015.2445380
41. *Miao, A., Ge, Z., Song, Z., Shen, F.,* Nonlocal structure constrained neighborhood preserving embedding model and its application for fault detection, Chemom. Intell. Lab. Syst. **142** (2015) 184.
doi: https://doi.org/10.1016/j.chemolab.2015.01.010
42. *Guo, J., Yuan, T., Li, Y.,* Fault detection of multimode process based on local neighbor normalized matrix, Chemom. Intell. Lab. Syst. **154** (2016) 162.
doi: https://doi.org/10.1016/j.chemolab.2016.02.010
43. *Guo, J., Zhong, L., Li, Y.,* Fault detection based on difference locality preserving projections for the semiconductor process, J. Chemom. **32** (2018) e3035.
doi: https://doi.org/10.1002/cem.3035
44. *Alauddin, M., Khan, F., Imtiaz, S., Ahmed, S.,* A bibliometric review and analysis of data-driven fault detection and diagnosis methods for process systems, Ind. Eng. Chem. Res. **57** (2018) 10719.
doi: https://doi.org/10.1021/acs.iecr.8b00936