Expert Systems with Applications 39 (2012) 4075–4083

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing

Dongil Kim [a], Pilsung Kang [d], Sungzoon Cho [a,*], Hyoung-joo Lee [b], Seungyong Doh [c]

[a] Seoul National University, 599 Gwanangno, Gwanak-gu, Seoul 151-744, Republic of Korea
[b] Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK
[c] Samsung Electronics Co. Ltd., 416 Maetan-dong, Yeongtong-gu, Suwon, Gyeonggi-do, Republic of Korea
[d] Information Technology Management Programme, International Fusion School, Seoul National University of Science & Technology (SeoulTech), 232 Gongreungno, Nowon-gu, Seoul, 139-743, Republic of Korea

ARTICLE INFO

Keywords:
Novelty detection
Faulty wafer detection
Semiconductor manufacturing
Virtual metrology
Dimensionality reduction

ABSTRACT

Since semiconductor manufacturing consists of hundreds of processes, a faulty wafer detection system, which allows for earlier detection of faulty wafers, is required. statistical process control (SPC) and virtual metrology (VM) have been used to detect faulty wafers. However, there are some limitations in that SPC requires linear, unimodal and single variable data and VM underestimates the deviations of predictors. In this paper, seven different machine learning-based novelty detection methods were employed to detect faulty wafers. The models were trained with Fault Detection and Classification (FDC) data to detect wafers having faulty metrology values. The real world semiconductor manufacturing data collected from a semiconductor fab were tested. Since the real world data have more than 150 input variables, we employed three different dimensionality reduction methods. The experimental results showed a high True Positive Rate (TPR). These results are promising enough to warrant further study.

1. Introduction

Semiconductor manufacturing consists of hundreds of processes, in each of which wafers that contain thousands of semiconductors are processed according to the recipes. A fault, such as the distortion of a wafer or the cut of a circuit, leads to a lower yields. The earlier a fault is detected, the better as this avoids increased manufacturing cost and increased lead time. To detect faulty wafers, a quality measurement process must be performed after each manufacturing process.

Statistical process control (SPC) methods have been typically used with Fault Detection and Classification (FDC) data (Qin, Cherry, Good, Wang, & Harrison, 2006; Su et al., 2007). In semiconductor manufacturing, each piece of process equipment contains hundreds of sensors to measure various conditions during the process, such as air temperature and exposure time. FDC data refer to those observations directly collected from the process sensors. SPC refers to the statistical methodology that inspects an important variable to be located in the specification range (Kourti & MacGregor, 1995). SPC based on FDC is commonly used in semiconductor manufacturing because of its immediacy; however, there are several limitations. First, SPC originally inspects and controls only a few of the hundreds of variables affecting wafer quality. Second, SPC monitors each variable independently while multiple variables interactively affect the quality of a wafer. Although SPC approaches such as the Principal Component Analysis (PCA) can deal with multiple variables, much of the original variable information can be lost. Third, these methods assume linearlity and unimodality of the data, which are often not true. Fig. 1 shows histograms of 20 different FDC variables from a real world semiconductor manufacturing plant. Clearly, the distributions are not unimodal. Finally, FDC does not measure wafer quality directly, but indirectly through the processing conditions. Thus, it is difficult to identify faulty wafers using only FDC data. Some FDC variables are irrelevant to wafer quality, and there is no guideline to distinguish faulty data from normal FDC data. Hence, defining a faulty wafer as a wafer having novel FDC data may cause many false alarms or misses.

Another way to detect a fault is by employing a metrology step after each process (Kang et al., 2009). However, metrology steps require extra cost, increased human resources and a longer cycle time (Cheng & Cheng, 2005; Chang, Kang, Hsu, Chang, & Chan, 2006), thus, only one wafer per process lot of 25 wafers is inspected. Metrology data have been rarely applied to fault detection due to their insufficiency. Thus, virtual metrology (VM) has been recently proposed (Chen et al., 2005). VM is defined as "the estimation of metrology values based on process data such as Fault Detection and Classification (FDC), context and previous metrology" (Besnard & Toprac, 2006). VM predicts metrology values without the need for a physical metrology function. Rather, in

* Corresponding author. Tel.: +82 2 880 6275; fax: +82 2 889 8564.
E-mail addresses: dikim01@snu.ac.kr (D. Kim), xfeel80@snu.ac.kr (P. Kang), zoon@snu.ac.kr (S. Cho), imhjlee@gmail.com (H.-j. Lee), sy.doh@samsung.com (S. Doh).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.09.088

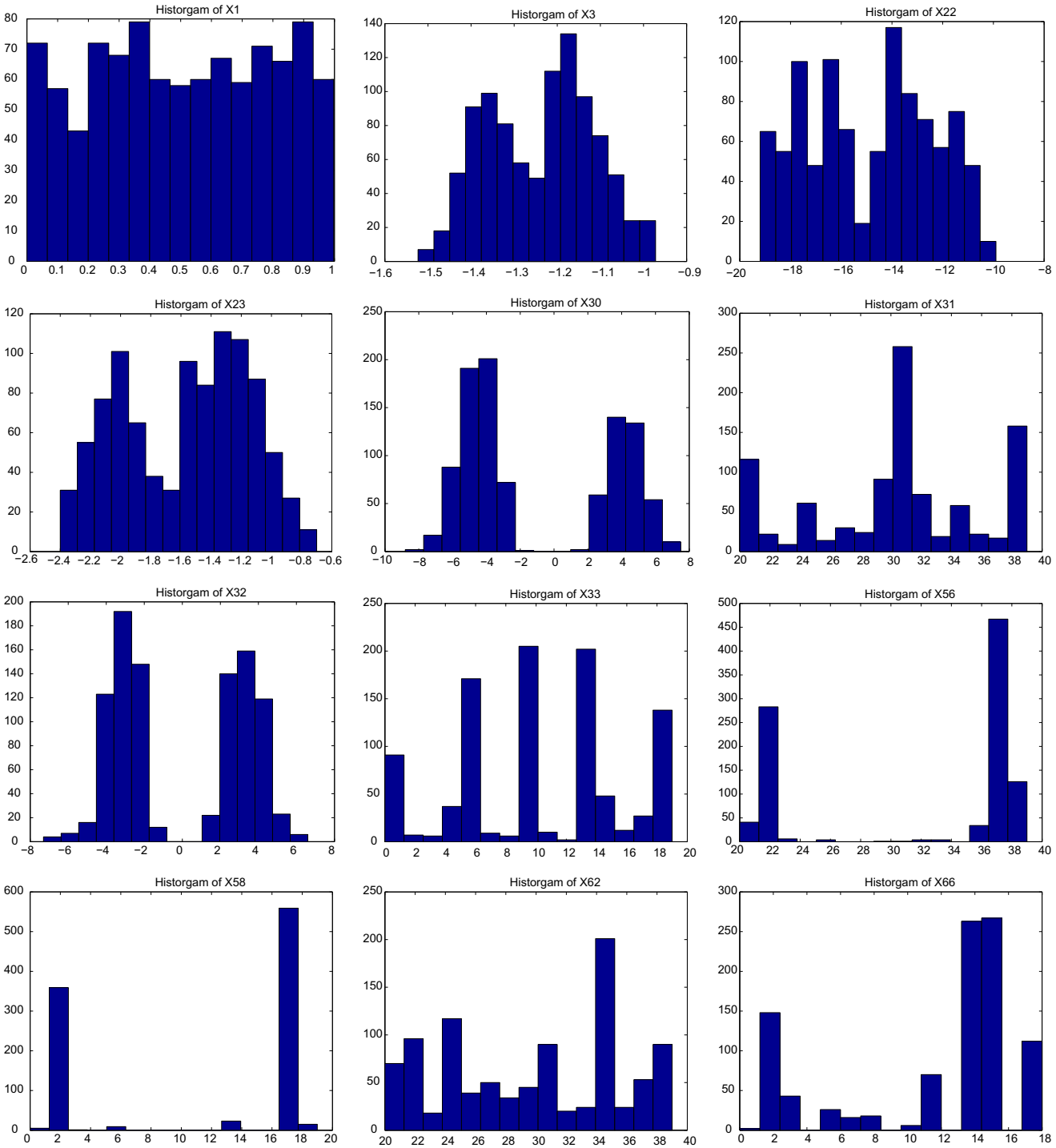

Fig. 1. Histograms of non-uni-modal distributions.

VM, regression models are employed to estimate a real-valued target to predict the metrology value. Since VM is not a physical operation but a numerical computation, it requires much less time and cost than the does real metrology step. Moreover, VM provides estimated metrology values for all wafers while real metrology provides only one out of every 25 wafers. However, for fault detection, VM has some limitations, as regression models tend to underestimate the deviations of predictors. Since faulty wafers are defined as wafers which have large deviations in the predictors, i.e. metrology values, even an accurate regression model has a large probability for missing many faulty wafers.

In this paper, we propose a machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. The proposed method detects faulty wafers directly from FDC data and metrology data, as shown in Fig. 2. We trained machine learning-based novelty detection models with FDC data to detect faulty wafers whose corresponding metrology values were outside the fabrication specification limit. Novelty detection methods are employed for two reasons. First, fault detection datasets are highly class imbalanced; faulty wafers are a very small portion of the total wafers. Second, faulty wafers are not defined as a set or a hypersphere, merely as any wafer different from the majority of normal

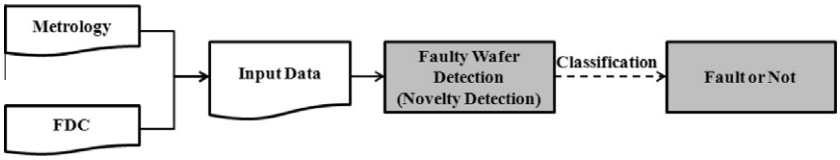

Fig. 2. The faulty wafer detection system.

wafers. Moreover, machine learning-based novelty detection methods have been successfully employed to detect multi-modal distributions as well as nonlinear distributions (Markou & Singh, 2003). We employed seven different machine learning-based novelty detection methods including 1-SVM, Gaussian mixture model, etc., and applied our proposed approach to real world semiconductor data. The real data had more than 150 variables, the treatment of which involved three different dimensionality reduction methods: stepwise linear regression, stepwise one class support vector machines (1-SVM) and Principal Component Analysis (PCA). By operating a faulty wafer detection model independent of the VM, we expected to detect faulty wafers after each process step, assuring that faulty wafers did not progress through the remaining process steps.

The remainder of this paper is organized as follows. In Section 2, we summarize the literature of commonly used SPC methods as well as novelty detection methods currently in practice. In Section 3, we introduce the semiconductor manufacturing data that were used and the process that we conducted to produce a faulty wafer detection system, including preprocessing, dimensionality reduction and novelty detection. In Section 4, we provide our experimental settings as well as the experimental results. In Section 5, we summarize our results and conclude the paper.

2. Existing faulty wafer detection methods

2.1. Statistical process control

In this section, we discuss SPC methods that are commonly used in manufacturing. Univariate SPC is used to control an important single target variable so that it is within the predefined specification range (Kourti & MacGregor, 1995). A univariate SPC chart records a single quality variable over its target value throughout the manufacturing process. If the quality variable is outside of its specification range, univariate SPC reports a fault. Since the univariate SPC has limitations, multivariate approaches have also been widely used in practice. Multivariate SPC is one of the first methods to use the Mahalanobis distance (Kourti & MacGregor, 1995). In this approach, there is a target vector, τ, which is predefined or is usually defined as a mean vector of some multivariate input data. The control variable, χ^2, is defined as the Mahalanobis distance between the process data and the target vector, as in Eq. (1)

$$\chi^2 = (y-\tau)^T\Sigma^{-1}(y-\tau), \tag{1}$$

where Σ denotes the covariance matrix of the input data. Based on this concept, Hotelling's T^2 became one of the most popular control variables in the manufacturing field (Kourti & MacGregor, 1995). Hotelling's T^2 is almost the same as the Mahalanobis distance except that Hotelling's T^2 uses the sample covariance matrix S rather than Σ. Hotelling's T^2 is defined as in Eq. (2).

$$T^2 = (y-\tau)^T S^{-1}(y-\tau), \quad \text{where} \quad S = \frac{1}{n-1}\Sigma_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^T. \tag{2}$$

PCA-based control approaches were proposed and have been widely used in real world manufacturing (Kourti & MacGregor, 1995; Qin, 2003; Qin et al., 2006; Yue, Qin, Markle, Nauert, & Gatto, 2000). One such approaches accounts for the reconstruction error in the process data. PCA can be defined as in Eq. (3) with a vector $\mathbf{x}$ and the projection vector $\mathbf{P}$, and the reconstruction error of test data $\mathbf{x}$ is defined as the squared error in Eq. (4). If SPE is larger than a threshold δ, $\mathbf{x}$ is classified as faulty.

$$\mathbf{x} = \hat{\mathbf{x}} + \tilde{\mathbf{x}}, \quad \text{where} \quad \hat{\mathbf{x}} = \mathbf{P}\mathbf{P}^T\mathbf{x} \equiv \mathbf{C}\mathbf{x} \quad \text{and} \quad \tilde{\mathbf{x}} = (\mathbf{I}-\mathbf{C})\mathbf{x}. \tag{3}$$

$$\text{SPE} \equiv \|\tilde{\mathbf{x}}\|^2 = \|(\mathbf{I}-\mathbf{C})\mathbf{x}\|^2. \tag{4}$$

Another PCA-based method calculates the Hotelling's T^2 of projected data from PCA, as in Eq. (5).

$$T^2 = \mathbf{x}^T\mathbf{P}^T\Lambda^{-1}\mathbf{P}\mathbf{x}, \quad \text{where} \quad \Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_l\}. \tag{5}$$

Partial Least Squares (PLS)-based methods have also been proposed (Geladi & Kowalski, 1986; Kourti & MacGregor, 1995; Wilson, Irwin, & Lightbody, 1997). PLS projects both input data X and corresponding target variable Y using an independent PCA (outer relation). Then PLS determines the regressive relationship between X and Y (inner relation). Wilson et al. (1997) proposed a neural network-based regression method to identify a non-linear regression function of the inner relation of PLS and trained the RBF networks to fit the non-linear inner relation of PLS.

However, those multivariate control methods have limitations. Mahalanobis distance-based methods (Hotelling's T^2) and PCA-based methods include an assumption that the data are uni-modularly distributed. If the manufacturing data follows a multi-modal distribution, a more realistic assumption, the multivariate methods tend to produce an erroneous result.

2.2. Virtual metrology

In VM, regression models were employed to estimate actual metrology values of wafers. As shown in Fig. 3, FDC data and the previous metrology values were used to predict the next metrology values with various linear or non-linear regression models such as linear regression, k-Nearest Neighbors (k-NN)-based regression, decision tree regression, neural networks regression and support vector regression. Also, to reduce the number of input dimensions, various dimensionality reduction methods such as stepwise linear regression, genetic algorithm and PCA were employed (Kang et al., 2009; Kang, Kim, Lee, Doh, & Cho, 2011). Their performances were good enough to estimate actual metrology values. However regression models had a limitation to detect faulty wafers.

In order to determine the VM's capability to detect faulty wafers, we conducted a preliminary experiment with real world semiconductor manufacturing data. As shown in Table 1, there were eight targets to be predicted from two different equipment chucks. The VM models were trained to estimate metrology values using FDC data. Various regression models as well as preprocessing and dimensionality reduction methods were employed, following the methodology of Kang et al. (2011). Mean Absolute Specification Error (MASE) can be defined as $\frac{1}{n}\sum_i^n \frac{|y_i-\bar{y}_i|}{SpecRange}$ where y_i and $\bar{y}_i$ indicate

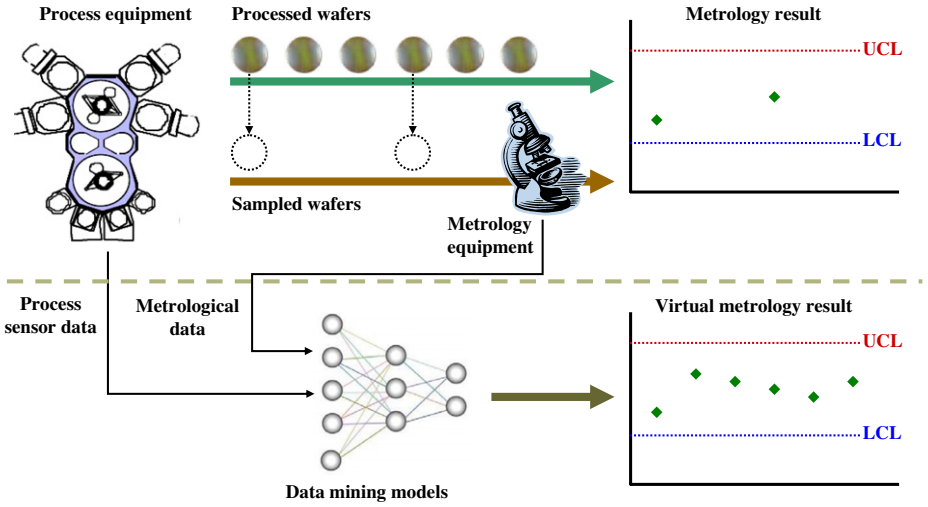

Fig. 3. The concepts of actual metrology (upper) and virtual metrology (lower) (Kang et al., 2009).

Table 1
The results of fault detection with VM.

Chuck	Target	MASE (%)	MASE Limit (%)	TPR (%)	FPR (%)	TPR-FPR (%)
Chuck 1	Y_1	8.14	<10	0	0	0
Chuck 1	Y_2	6.20	<10	0	0	0
Chuck 1	Y_3	7.38	<10	0	0	0
Chuck 1	Y_4	6.03	<10	0	0	0
Chuck 2	Y_1	7.82	<10	0	0	0
Chuck 2	Y_2	6.27	<10	0	0	0
Chuck 2	Y_3	7.36	<10	0	0	0
Chuck 2	Y_4	5.82	<10	0	0	0

the real target value and the estimated target value, respectively. 'SpecRange' can be defined as the difference between the upper specification limit and the lower specification limit of the target variable. In these data, the upper specification limit was '0.015,' while the lower specification limit was '−0.015.' If a MASE value was lower that the predefined 'MASE Limit,' the VM model was acceptable for use in the real semiconductor manufacturing process line. Both 'SpecRange' and 'MASE Limit' used in this paper were fit to the actual values used in the semiconductor manufacturing process. TPR stands for the True Positive Rate while FPR stands for the False Positive Rate. As shown in Table 1, those eight VM models were acceptable in terms of 'MASE Limit,' indicating their acceptabilities as VM models. However, none of them identified any faulty wafer. In order to explain this, real target values and estimated target values from the VMs were plotted in Fig. 4. Squares indicate the real metrology values while circles indicates the VM values. As shown, the VM model tended to underestimate the deviation in the metrology values. With the upper specification limit of 0.015, the actual specification limit of those data, there were 13 faulty wafers; however, the VM models just considered all wafers to be normal.

2.3. Machine learning-based novelty detection

One of the early approaches was a *k*-Nearest Neighbors (*k*-NN)-based method (He & Wang, 2007). This method did not employ the statistical process control concept; rather it employed a machine learning algorithm, *k*-NN, to detect faulty wafers. The researchers

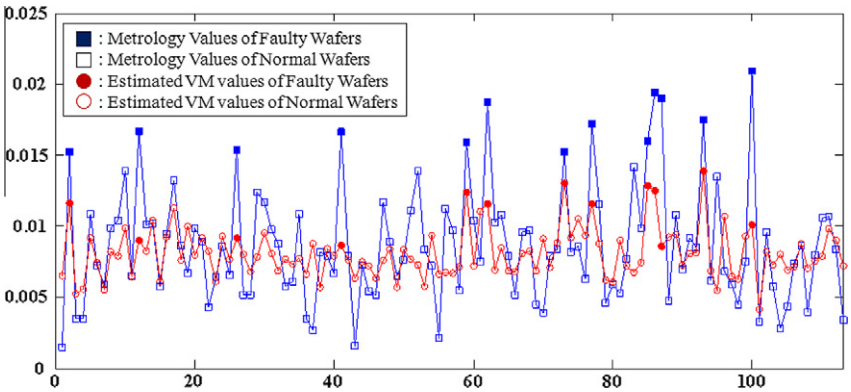

Fig. 4. Real metrology values and metrology values estimated by VM for 113 wafers.

conducted the k-NN method with FDC process data and calculated their sum of the squared distance, on which they based their threshold. When a new test data point was collected, they identified its nearest neighbors from the training data and calculated its squared distance. If the squared distance was larger than the threshold, this test data point was considered a faulty wafer. Even though this approach used only FDC data, it overcame the major problems of SPC, linearity and unimodality of data. The result of their paper showed potential for applying machine learning algorithms to semiconductor manufacturing data.

3. Faulty wafer detection system

3.1. Overview

In this research, we propose a faulty wafer detection system based on machine learning-based novelty detection methods. Input data were FDC data and metrology data from the previous metrology step, while faulty wafers were defined as wafers having corresponding metrology values outside of the specification limits. Fig. 2 shows the concept of the faulty wafer detection system, while Fig. 5 shows the process steps for constructing the faulty wafer detection system using actual manufacturing data. We conducted the proposed method with real semiconductor manufacturing data and evaluated the performance of the proposed method.

3.2. Data description

The FDC data and metrology data were collected from a photo process in a Korean semiconductor manufacturing company over a seven and a half month period in 2007. The FDC data were collected from sensors of equipment in the photo process. The equipment contains two different chucks to operate the same process. Each sensor monitors the process status, such as heating duration, mask overlay, etc. Metrology data were collected from two metrology machines, one of which was located before the photo process, and one was located after the photo process. The metrology data from the first machines were used as input variables, and the metrology data from the second machine were used to derive the target variables. Metrology machine only inspects one wafer out of every 25 wafers. This limited data collection was one of the main motivations of this research.

3.3. Data preprocessing

The same input variables that were used for VM were also employed in the faulty wafer detection system. The FDC data extracted from the current process step and the metrology (or VM) values from the previous metrology step are used for input variables, as in Fig. 3. First, we discarded the variables which indicate meta-information of the process such as LOT ID, WAFER ID, STATUS and EXPOSED. We then discarded other input variables which contained no useful information. For categorical variables, those variables which had only one value were discarded. For numerical variables, those variables which had zero standard deviation values were discarded. The number of variables is summarized in Table 2.

The target variables were derived from the metrology values. In these real world data, there were four different target variables, Y_1, Y_2, Y_3 and Y_4, hence, four different models were constructed. For each model, if the metrology value was outside of the its specification range, the corresponding target value was set to be fault, as in Fig. 6. The number of faulty wafers among all wafers is summarized in Table 3. Because it was a highly unbalanced problem, we proposed to use novelty detection algorithms to train the model using only normal wafers, and all of the data were normalized.

3.4. Dimensionality reduction

The input variables were collected from equipment sensors and the previous metrology step. Even though the sensors monitored the information for other purposes, not all of the information were relevant to the faulty wafer detection model. In addition, the number of input variables was relatively large, resulting in the dimensionality problem. Those irrelevant variables and the curse of dimensionality may degrade the model performance. Hence, a smaller subset of input variables that contribute to model performance must be selected. We used dimensionality reduction approaches.

There are two types of dimensionality reduction methods. The first type, variable selection, chooses input variables which contribute to the model performance and discards irrelevant input variables. We used two variable selection methods: stepwise linear regression (Johnson & Wichern, 1998) and stepwise 1-SVM. Stepwise methods start with a single input variable that best fits the input-target relationship, which is evaluated using the model performance, linear regression and 1-SVM. Other input variables are kept in the candidate set. Through forward selection, a new variable from the candidate set is added to the model, if the variable increases the model performance. In backward selection, a variable can be eliminated if it does not affect the model performance or if it decreases the model performance. By repeating those forward selection and backward selection steps, we can obtain a smaller subset of important input variables.

The second type of dimensionality reduction method, variable construction, assembles a new set of input variables by combining the original variables. We used one variable construction method, PCA (Jolliffe, 1986), which calculates new variables by projecting eigenvectors of the input data. Each eigenvalue indicates the degree of variance in the input data that was covered by the corresponding eigenvector. Hence, projected data onto the k eigenvectors corresponding to the k largest eigenvalues are used for training. We used an input variable set that alleviated at least 70% in the variance of the original input data.

3.5. Machine learning-based novelty detection

We used machine learning-based novelty detection methods to detect faulty wafers for the following reasons. First of all, faulty wafers cannot be defined with a specific characteristic. Rather, faulty wafers are defined as all wafers different from the normal wafers. Because of this ambiguity, it is better to use novelty detection models than it is to use conventional binary classification models. Second, the data are highly imbalanced. The fraction of normal wafers is approximately 98–99%, while the fraction of faulty wafers is approximately 0.5–2%. Conventional binary classification models tend to place too much emphasis on the majority class.

We have used several machine learning-based novelty detection methods. For density estimation methods, Gaussian density

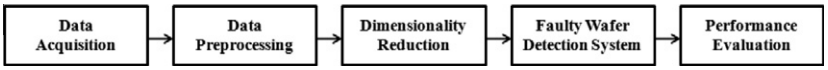

Fig. 5. The process steps of the faulty wafer detection system.

Table 2
The number of data and the number of variables.

Chuck	# of wafers	# of FDC variables	# of metrology variables	# of input variables after preprocessing
Chuck 1	2583	148	15	117
Chuck 2	2509	148	15	117

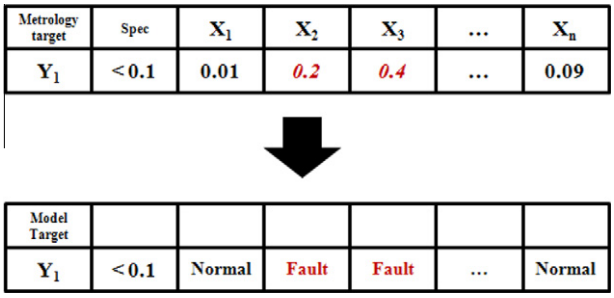

Fig. 6. Target variable determination for the faulty wafer detection system.

Table 3
The number of faulty wafers.

Chuck	# of wafers	Y_1	Y_2	Y_3	Y_4
Chuck 1	2583	60 (2.3%)	6 (0.2%)	18 (0.7%)	15 (0.6%)
Chuck 2	2509	22 (0.9%)	9 (0.4%)	24 (1.0%)	9 (0.4%)

estimation (Barnett & Lewis, 1994; Bishop, 1995), Gaussian mixture model (GMM) (Bishop, 1995) and Parzen window (Bishop, 1995) methods were employed. The Gaussian density estimation method estimates a single Gaussian distribution for the input data. Then when new test data arrive, it determines how far each test data point is from the center of the Gaussian distribution. If the distance between a data point and the center of the Gaussian distribution is greater than the threshold distance, the data point is classified as novel. However, the Gaussian density estimation assumes that the input data are generated from a single Gaussian distribution, a weak point of Gaussian density estimation; GMM and Parzen window density estimations were designed to overcome that issue. The GMM estimates the distribution of the input data using a linear combination of several individual Gaussian distributions, while the Parzen window density estimation applies a Gaussian distribution to each data point, and then the linear combination of each Gaussian distribution becomes the distribution of all input data.

For non-probabilistic methods, *k*-means Clustering-based novelty detection (KMC) (Hastie, Tibshirani, & Friedman, 2002) and 1-SVM (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001; Lee & Cho, 2007) have been employed. KMC conducts a *k*-means clustering algorithm on the input data. Then, when a new data point arrives, the distance between that data point and the closet cluster center is calculated. If the distance is greater than the threshold, that data point is classified as novel. 1-SVM is a novelty detection version of support vector machines that identifies a hyperplane $\mathbf{w}$ that separates a fraction of the patterns of the origin in a feature space using a maximal margin, as shown in Fig. 7, in the optimization problem using only positive data $\mathbf{X}^+$:

$$\begin{aligned} &\text{Minimize } \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n^+}\sum_{\mathbf{x}_i \in \mathbf{X}^+} \xi_i^+, \\ &\text{s.t. } \mathbf{w}^{\mathbf{T}}\Phi(\mathbf{x}_i) \geqslant \rho - \xi_i^+, \quad \xi_i^+ \geqslant 0. \end{aligned} \tag{6}$$

For reconstruction methods, PCA-based and KPCA-based novelty detection methods have been employed. Applying PCA, input data

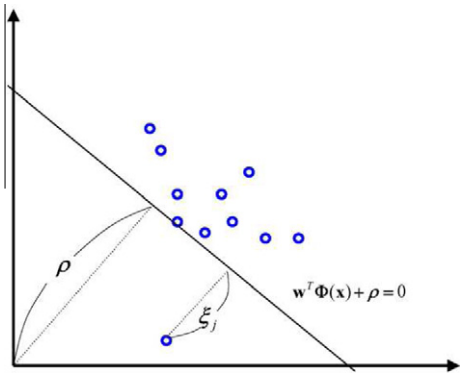

Fig. 7. The concepts of one class support vector machines (Lee et al., 2008).

are mapped into a linear subspace defined by a set of eigenvectors, $\mathbf{V}$, which explains the largest variance in the input data. The reconstruction of a data point $\mathbf{x}$ can be calculated as $\hat{\mathbf{x}} = \mathbf{V}\mathbf{V}^T\mathbf{x}$. Then the reconstruction error can be calculated as $e(\mathbf{x}) = \|\mathbf{x} - \mathbf{V}\mathbf{V}^T\mathbf{x}\|^2$. If the reconstruction error is greater than the threshold value, that data point is classified as novel. KPCA is a non-linear version of PCA that includes the kernel mapping function, Φ. The reconstruction error of KPCA can be calculated as $e(\Phi(\mathbf{x})) = \|\Phi(\mathbf{x}) - \mathbf{V}_\Phi\mathbf{V}_\Phi^T\Phi(\mathbf{x})\|^2$. All of the variable selection methods and novelty detection methods are summarized in Table 4.

3.6. Performance evaluation

We employed the True Positive Rate (TPR) and False Positive Rate (FPR) as measures of performance. True positive and false positive are defined based on the confusion matrix, as in Fig. 8. TPR can be defined as $\frac{TP}{TP+FN}$, while FPR can be defined as $\frac{FP}{TN+FP}$. TPR, also known as sensitivity or hit ratio, is concerned with how accurately a model detects faulty wafers. FPR, which also can be calculated as $1 - Specificity$, is concerned with how incorrectly a model classifies normal wafers. A larger TPR is preferred, along with a smaller FPR. We used TPR-FPR as the unified measure to evaluate model performance.

4. Experimental results

4.1. Cross validation result

In order to evaluate the performance of the machine learning-based novelty detection method, the experiments can be classified

Table 4
The number of data and the number of variables.

Dimensionality reduction methods	Novelty detection methods
Stepwise LR	Gaussian density estimation
Stepwise 1-SVM	Gaussian mixture model
PCA	Parzen window
	k-Means clustering
	1-SVM
	PCA
	KPCA

		Predict	
		Normal	Fault
Actual	Normal	TN	FP
	Fault	FN	TP

Fig. 8. Confusion matrix.

Table 5
The number of input variables obtained from each dimensionality reduction method.

Chuck	Target	Stepwise LR	Stepwise 1-SVM	PCA
Chuck 1	Y_1	23	7	4
Chuck 1	Y_2	27	8	4
Chuck 1	Y_3	27	18	4
Chuck 1	Y_4	23	6	4
Chuck 2	Y_1	21	7	4
Chuck 2	Y_2	22	11	4
Chuck 2	Y_3	28	14	4
Chuck 2	Y_4	28	6	4

into two different evaluation types, cross validation (CV) and moving windows (MW). CV is an experimental platform widely used to evaluate model performance. This method divide the dataset into k-folds, and evaluates data in onefold with a model trained with data in the other $k-1$-folds. After evaluating all of the data, the total performance of the model can be evaluated. On the other hand, MW mimics the real world environment by considering the influence of time. MW trains a model with data from a certain time period and uses it to evaluate the next time period. We can easily evaluate the performance of a model using CV, while performance can be evaluated more practically using MW.

To evaluate the performance with CV, the dataset was divided into five exclusive folds. The model was trained model with only fourfolds and was used to test the other onefold until all the data were tested at once. We searched the model parameters using fivefold cross validation of each training dataset. Table 5 shows the number of input variables selected from each dimensionality reduction method, illustrating that the number of selected variables was much smaller than the number of original input variables. We concluded that there are many irrelevant and redundant variables in semiconductor manufacturing data.

Table 6 shows the experimental results. Faulty wafer detection models which showed the best results for the corresponding target variables were summarized. As shown in Table 6, stepwise 1-SVM showed the best results among the other dimensionality reduction methods for all target variables. For the novelty detection model, 1-SVM showed the best TPR-FPR results for all target variables, with TPRs greater than 80% for seven of eight target variables, with three values greater than 90%. This model detected faulty wafers with high True Positive Rates.

4.2. Moving window result

To evaluate the model performance in a real world environment, we used the MW learning scheme as in Fig. 9. The semiconductor data were collected over a period of seven and a half months. In the moving window learning, the test period was always one month while the training period was varied from 2 to 5 months. We searched the model parameters using fivefold cross validation for each training dataset.

Experimental results evaluated using moving window learning are summarized in Table 7. For moving window learning, stepwise linear regression was selected as the best dimensionality reduction model. For the novelty detection model, KMC and PCA both had the maximum TPR-FPR. The results of TPR were greater than 50% for all target variables, and were grater than 80% for three target variables. There was no training time period which showed a significantly superior performance.

Fig. 10 compares the best models resulted from cross validation and the results from moving windows. As shown in Fig. 10, the model performances were degraded when moving windows learning was employed. Table 8 summarizes and compares the best models of both CV and MW. In the CV experiment, the novelty detection model identified 82.99% of faulty wafers, with an FPR of 34.61%. In the moving window experiment, the novelty detection model identified 73.49% of faulty wafers and exhibited an

Table 6
Experimental results of the best models in terms of TPR-FPR evaluated from fivefolds cross validation (CV).

Chuck	Target variable	Dimensionality reduction	Novelty detection	TPR (%)	FPR (%)	TPR-FPR (%)
Chuck 1	Y_1	Stepwise 1-SVM	1-SVM	81.67	49.35	32.32
Chuck 1	Y_2	Stepwise 1-SVM	1-SVM	83.33	21.81	61.52
Chuck 1	Y_3	Stepwise 1-SVM	1-SVM	82.22	34.50	47.72
Chuck 1	Y_4	Stepwise 1-SVM	1-SVM	97.33	45.44	51.89
Chuck 2	Y_1	Stepwise 1-SVM	1-SVM	99.09	49.90	49.19
Chuck 2	Y_2	Stepwise 1-SVM	1-SVM	93.33	23.36	69.97
Chuck 2	Y_3	Stepwise 1-SVM	1-SVM	42.50	17.75	24.75
Chuck 2	Y_4	Stepwise 1-SVM	1-SVM	84.44	34.80	49.64

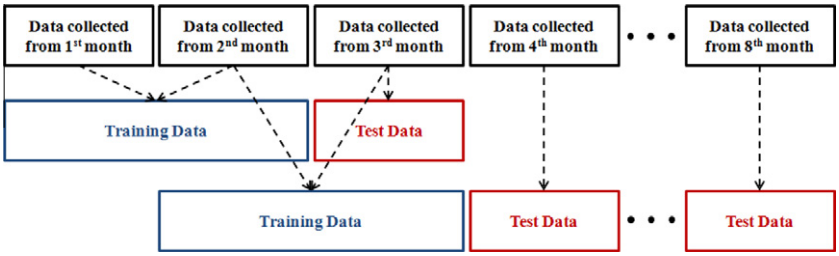

Fig. 9. Moving window learning scheme.

Table 7
Experimental results of the best models in terms of TPR-FPR, as evaluated by moving window (MW).

Chuck	Target variable	Dimensionality reduction	Novelty detection (month)	Training period	TPR (%)	FPR (%)	TPR-FPR (%)
Chuck 1	Y_1	Stepwise LR	KMC	3	55.87	31.10	24.77
Chuck 1	Y_2	Stepwise LR	PCA	5	75.00	39.44	35.56
Chuck 1	Y_3	Stepwise LR	PCA	3	80.00	56.27	23.73
Chuck 1	Y_4	Stepwise LR	KMC	5	67.86	37.37	30.49
Chuck 2	Y_1	Stepwise LR	KMC	2	91.67	58.98	32.69
Chuck 2	Y_2	Stepwise LR	KMC	4	72.22	53.51	18.71
Chuck 2	Y_3	Stepwise LR	PCA	2	61.94	42.98	18.96
Chuck 2	Y_4	Stepwise LR	KMC	4	83.33	47.16	36.17

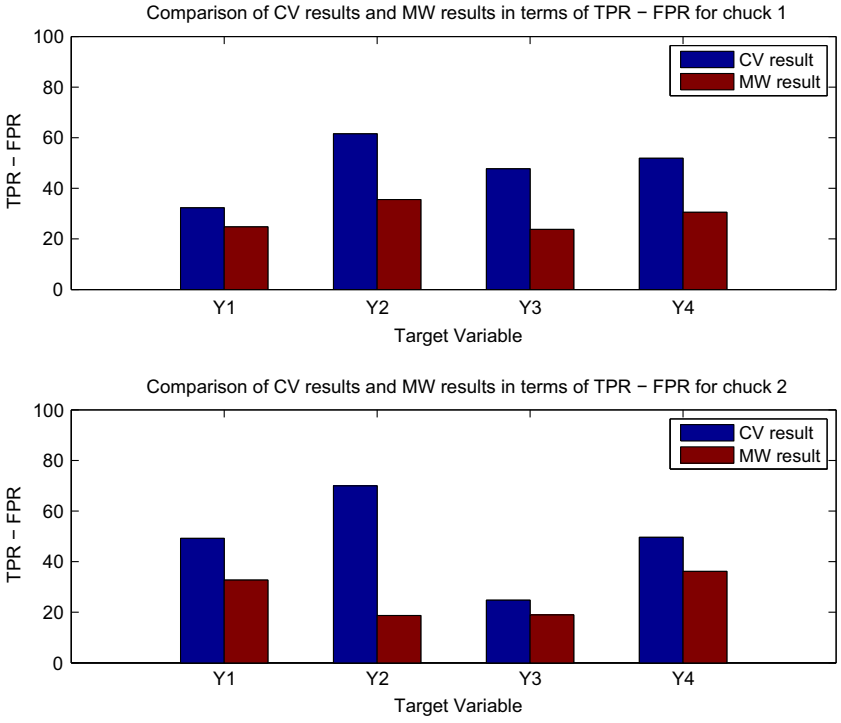

Fig. 10. Comparison of the best models of CV and MW in terms of TPR-FPR.

Table 8
Summary of the average experimental results for the best models of CV and MW.

Training type	TPR (%)	FPR (%)	TPR-FPR (%)
Cross validation	82.99	34.61	48.38
Moving window	73.49%	45.85	27.64

FPR of 45.85%. In both cases, the TPR was sufficiently high to continue with this research. The performance was degraded when moving window learning scheme was employed.

5. Conclusions

Since faulty wafers result in increased costs and longer lead times, accurate and timely detection of faulty wafers is very important in semiconductor manufacturing. The conventional methods such as SPC and VM have some limitations in faulty wafer detection. In this paper, we proposed a machine learning-based novelty detection method for faulty wafer detection in semiconductor manufacturing. Machine learning based-novelty detection can detect faulty wafers without the need for assumptions of distribution, and nonlinear problems can be addressed. Also, this type of method can successfully classify novel and minority wafers.

The faulty wafer detection steps that we conducted are as following. The semiconductor manufacturing data were preprocessed to act as the training data for the novelty detection method. We used three different dimensionality reduction methods and seven different novelty detection methods. We conducted experiments with real world data collected from a semiconductor manufacturer. The experiments can be classified into two settings, cross validation and moving windows. The overall experimental result indicated a possibility of detecting faulty wafers with these approaches. As shown in Table 8, the TPRs were adequate in both experiments. The experimental results of CV are better than those of MW, possibly because the training data set in the training window was too small. If the training period were increased, the model performance may be increased.

There are some limitations and future works of this research. First of all, the TPR/FPR ratio can be controlled by setting misclassification cost in the semiconductor manufacturing. In this paper, we set the misclassification cost to be the same. However, a further research for controlling the trade-off of TPR and FPR by determining the misclassification cost will be needed. Second, since the moving windows scheme is more realistic than the cross validation experiments, further research on the moving windows scheme will be required. Also, we plan to reduce the outliers in the data as they can degrade model performance. In addition, we plan to carry

out an economic analysis on the actual use of this method. We also plan to apply this approach to other semiconductor manufacturing processes.

Acknowledgements

This work was supported by the Brain Korea 21 program in 2006–2011, Seoul R&D Program (TR080589M0209722), and Mid-career Researcher Program funded by the NRF (National Research Foundation) and MEST (No. 400-20110010). This work was also supported by the Engineering Research Institute of SNU and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0021893).

References

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. New York, USA: Wiley and Sons.

Besnard, J., & Toprac, A. (2006). Wafer-to-wafer virtual metrology applied to run-to-run control. In *Proc. of the 3rd ISMI symposium on manufacturing effectiveness, USA*.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York, USA: Oxford University Press.

Chang, Y. J., Kang, Y., Hsu, C. L., Chang, C. T., & Chan, T. Y. (2006). Virtual metrology technique for semiconductor manufacturing. *International Joint Conference on Neural Networks, Vancouver*, 5289–5293.

Cheng, J., & Cheng, F. T. (2005). Application development to virtual metrology in semiconductor industry. In *The 32nd annual conference of IEEE industrial electronics society, USA* (pp. 124–129).

Chen, P., Wu, S., Lin, J., Ko, F., Lo, H., Wang, J., et al. (2005). Virtual metrology: A solution for wafer to wafer advanced process control. *IEEE International Symposium on Semiconductor Manufacturing, USA*, 155–157.

Geladi, P., & Kowalski, B. R. (1986). Partial least squares regression: A tutorial. *Analytica Chimica Acta, 185*, 1–17.

Hastie, T., Tibshirani, R., & Friedman, J. (2002). *The element of statistical learning: Data mining, inference, and prediction*. Springer.

He, Q. P., & Wang, J. W. (2007). Fault detection using the *k*-nearest neighbor rule for semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing, 20*(4), 345–354.

Johnson, A. R., & Wichern, D. W. (1998). *Applied multivariate statistical analysis*. Prentice Hall.

Jolliffe, I. T. (1986). *Principal component analysis*. Springer-Verlag.

Kang, P., Kim, D., Lee, H., Doh, S., & Cho, S. (2011). Virtual metrology for run-to-run control in semiconductor manufacturing. *Expert Systems with Applications, 38*(3), 2508–2522.

Kang, P., Lee, H., Cho, S., Kim, D., Park, J., Park, C., et al. (2009). A virtual metrology system for semiconductor manufacturing. *Expert Systems with Applications, 36*, 12554–12561.

Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivatiate projection methods. *Chemometrics and Intelligent Laboratory Systems, 28*, 3–21.

Lee, H., & Cho, S. (2007). Focusing on non-respondents: Response modeling with novelty detectors. *Expert Systems with Applications, 33*, 522–530.

Lee, H., Cho, S., & Shin, M. (2008). Supporting diagnosis of attention-deficit hyperactive disorder with novelty detection. *Artificial Intelligence in Medicine, 42*(3), 199–212.

Markou, M., & Singh, S. (2003). Novelty detection: A review – Part 2: Neural networks based approaches. *Signal Processing, 83*, 2499–2521.

Qin, S. J. (2003). Statistical process monitoring: Basics and beyond. *Journal of Chemometrics, 17*, 480–502.

Qin, S. J., Cherry, G., Good, R., Wang, J., & Harrison, C. A. (2006). Semiconductor manufacturing process control and monitoring: A fab-wide framework. *Journal of Process Control, 16*, 179–191.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*, 1443–1471.

Su, A. J., Jeng, J. C., Huang, H. P., Yu, C. C., Hung, S. Y., & Chao, C. K. (2007). Control relevant issues in semiconductor manufacturing: Overview with some new results. *Control Engineering Practice, 15*, 1268–1279.

Wilson, D. J. H., Irwin, G. W., & Lightbody, G. (1997). Neural networks and multivariate SPC. *IEE-Colloquium-(Digest), 174*, 1–5.

Yue, H. H., Qin, S. J., Markle, R. J., Nauert, C., & Gatto, M. (2000). Fault detection of plasma etchers using optical emission spectra. *IEEE Transactions on Semiconductor Manufacturing, 13*(3), 374–385.