



Discriminative feature learning and cluster-based defect label reconstruction for reducing uncertainty in wafer bin map labels

Seyoung Park¹ · Jaeyeon Jang¹ · Chang Ouk Kim¹

Received: 2 January 2020 / Accepted: 31 March 2020 / Published online: 16 April 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Many studies have been conducted to improve wafer bin map (WBM) defect classification performance because accurate WBM classification can provide information about abnormal processes causing a decrease in yield. However, in the actual manufacturing field, the manual labeling performed by engineers leads to a high level of uncertainty. Label uncertainty has been a major cause of the reduction in WBM classification system performance. In this paper, we propose a class label reconstruction method for subdividing a defect class with various patterns into several groups, creating a new class for defect samples that cannot be categorized into known classes and detecting unknown defects. The proposed method performs discriminative feature learning of the Siamese network and repeated cross-learning of the class label reconstruction based on Gaussian means clustering in a learned feature space. We verified the proposed method using a real-world WBM dataset. In a situation where there the class labels of the training dataset were corrupted, the proposed method could increase the classification accuracy of the test dataset by enabling the corrupted sample to find its original class label. As a result, the accuracy of the proposed method was up to 7.8% higher than that of the convolutional neural network (CNN). Furthermore, through the proposed class label reconstruction, we found a new mixed-type defect class that had not been found until now, and we detected new types of unknown defects that were not used for learning with an average accuracy of over 73%.

Keywords Wafer bin map · Label uncertainty · Class label reconstruction · Unknown defect detection · Siamese network · G-means clustering

Introduction

With the advent of the fourth industrial revolution, semiconductors have been applied to various devices, including mobile and wearable devices. The need for semiconductor miniaturization has consequently increased. However, semiconductor miniaturization through enhancement and elaboration of existing process technologies has reached its limit. Consequently, new process technologies and 3D devices, such as extreme ultraviolet processing, fin field-effect transistors and gate-all-around devices, have been introduced (Ferain et al. 2011). This change in manufacturing technology increases the occurrence rate of defects; thus, it is essential to maximize yield through rapid defect detection (Liu and Chien 2013).

Forming integrated circuits on silicon wafers is a key objective of the semiconductor process, which is a fabrication process composed of several hundred to several thousands of steps. When the fabrication process is completed, electrical die sorting tests are performed to sort out defective chips in the wafer. As a result, a two-dimensional image wafer bin map (WBM), which expresses the defect of each chip with binary values of 0 and 1, is created (Kim et al. 2018). Figure 1 shows WBMs with six defect patterns having the same yield. The chips that pass all tests normally are marked in gray, and the other chips are marked in black. Figure 1a shows the normal WBM with no specific distribution pattern of the defective chips. Figures 1b–f show defective WBMs with different labels according to the distribution pattern of chips. In many cases, the types of abnormal processes that cause defects determine the defect pattern of the WBM (Wu et al. 2015). Therefore, it is a core task for yield improvement to accurately classify WBM defect patterns and take prompt action to determine the root cause of the defect based on the classification result.

✉ Chang Ouk Kim
kimco@yonsei.ac.kr

¹ Department of Industrial Engineering, Yonsei University, Seoul 03722, Republic of Korea

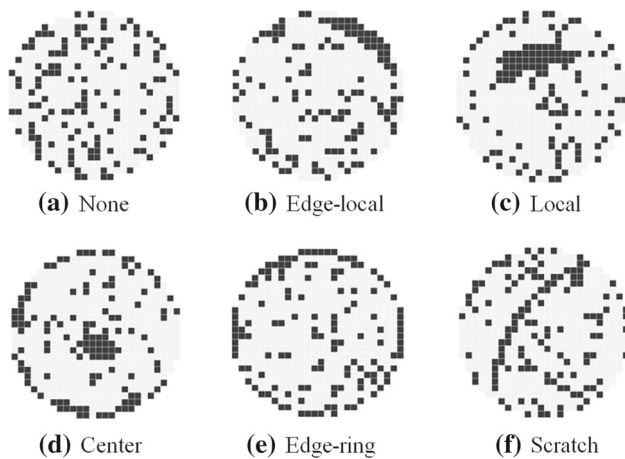
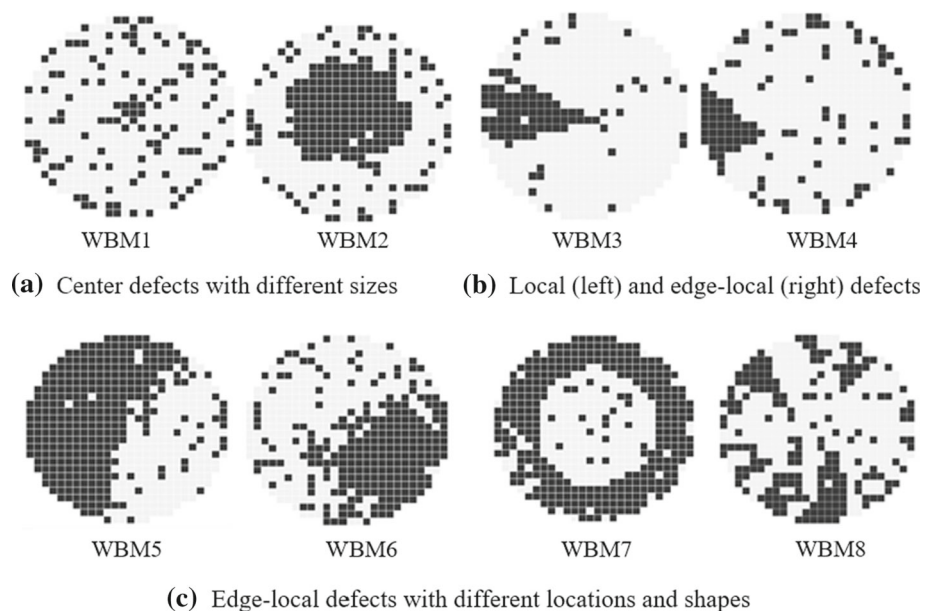


Fig. 1 WBM with different defect patterns

The recent increase in semiconductor output according to the continuously rising semiconductor demand has raised the need for the automatic classification of WBMs. Many studies on this topic have been performed (Chang et al. 2012; Kyeong and Kim 2018; Nakazawa and Kulkarni 2018; Wu et al. 2015). However, WBM class labels in the actual manufacturing field are often uncertain, and mislabeled data learning degrades the prediction performance of these classification models (Nettleton et al. 2010). The uncertainty of WBM labels is mainly caused by undetailed labeling and excessive subdivision of similar defect patterns in the manual labeling process performed by engineers (Liu and Chien 2013). Figure 2 shows the various types of mislabeling. In Fig. 2a, center defects of different sizes were considered a single label. In Fig. 2c, edge-local defects with different locations and shapes were considered a single label. In this case,

Fig. 2 Examples of WBMs with uncertain labels



defects of the same pattern can appear in different locations and sizes depending on the type of abnormal process. Hence, the labeling criteria need to be subdivided. Figure 2b shows excessive subdivision of similar patterns. The local defect and edge-local defect samples were classified into different classes, even if they had remarkably similar patterns. In addition, the error of considering a new defect pattern, resulting from a change in design and process conditions, as an existing known class label is a major factor that increases uncertainty (Adly et al. 2014). For example, WBM8 in Fig. 2c has a radial-type defect pattern that spreads from the wafer center to the edge; however, it is considered an edge-local defect.

Recently, studies in the visual recognition field have attempted to reduce the uncertainty of image labeling through noise relabeling. For example, Veit et al. (2017) and Li et al. (2017) proposed a method of modifying a noisy sample label through a cleaning model trained using a clean dataset. They could maximize the classification performance by reflecting modified class label information in the image classification model. However, this method requires manual sorting of clean data; thus, it is not appropriate for WBM data, which have many samples that have been mislabeled due to human error. To overcome this limitation, studies have suggested methods to find and modify noisy samples automatically in the classification model training process (Köhler et al. 2019; Wang et al. 2018). However, these methods can consider a noisy sample only as an outlier or as belonging to an existing class.

To overcome this limitation and apply class relabeling to a WBM dataset, we introduce an extended class label reconstruction method. The proposed method trains an embedding model that makes samples with similar patterns closer in a feature space and performs class label reconstruction based

on the clustering result in a learned feature space. The class label reconstruction process not only relabels the class of a sample but also provides the cluster label of a sample and information about whether a sample is noisy or clean. These pieces of information are reflected again in the training embedding model. In addition to enabling class relabeling, this repeated training allows the creation of new classes for samples that cannot be considered to belong to an existing class and enables the detection of new defect patterns. We introduce the Siamese network for embedding (Chopra et al. 2005) and propose a new loss function to generate discriminative features by reflecting class label reconstruction results. The class label reconstruction is based on the result of the Gaussian means (G-means) clustering algorithm (Hamerly and Elkan 2004), which assumes that samples in a cluster are distributed in accordance with a Gaussian distribution while automatically learning the optimal number of clusters.

We verified the proposed method using WM-811K (Wu et al. 2015), which is a real-world WBM dataset. For this experiment, the class labels of some training data were changed to different random class labels. The experimental results showed that with training progress, the noisy samples found their original class labels, and accordingly, the classification accuracy for the test dataset steadily increased. This provided a basis for the proposed method to perform better than existing classification benchmarks in the presence of labeling uncertainty. In addition, we observed that samples with different patterns even in the same class were divided into multiple groups and that mixed-type defect samples, which were hardly seen with existing classes, had new class labels. In addition, we performed a classification experiment after assuming some defect classes of the WBM as unknown classes, and the proposed method showed a detection rate higher than 73% on average for unknown class samples that had not been trained. This shows that the proposed method can also be used for detecting new defect patterns.

The remaining parts of this paper are organized as follows. Section “[Literature review](#)” presents a brief review of existing studies on WBM defect pattern classification and class relabeling. Section “[Background](#)” introduces the Siamese network and G-means clustering algorithm, which are base models for the proposed method. Section “[Methodology](#)” provides a detailed description of the proposed class label reconstruction method. Section “[Experiment](#)” presents the experimental design and results using WM-811K. Finally, Section “[Conclusion](#)” discusses the results of this paper and future research directions.

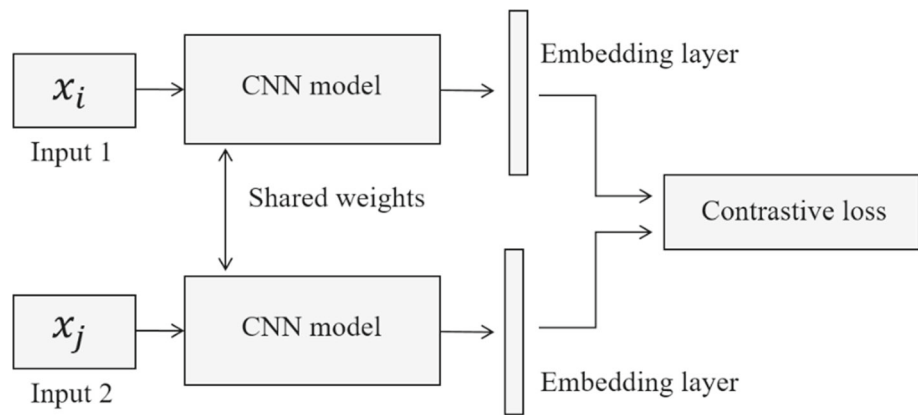
Literature review

With the rising importance of accurate abnormal process detection and yield maximization issues, many studies have

been conducted to maximize the defect pattern classification performance of the WBM through machine learning methods. Past studies on the WBM have focused on extracting manual features that workers consider important and using them as inputs of shallow machine learning models for the classification of defect patterns (Chang et al. 2012; Wang et al. 2006; Wu et al. 2015). Recently, studies have been conducted on the convolutional neural network (CNN), which is an image classification model based on deep learning that learns important features for classification by itself (Kyeong and Kim 2018; Nakazawa and Kulkarni 2018). However, existing studies have assumed that the data labels used during training are correct.

To solve the problem of degrading classification performance caused by uncertain labels of the training data, studies have often used noisy sample removal (Guan et al. 2011) or introduced a loss function that reduces the influence of noisy samples (Patrini et al. 2017; Vahdat 2017). Recently, however, noise label modification methods have been researched under the assumption that classification model performance can be improved by modifying the label of a noisy sample and using the modified label information in classification model training. For example, Veit et al. (2017) proposed a label cleaning model that only trains images with clean labels given by people manually. They modified noise labels through the model and used the modified label information to train the CNN classifiers. Li et al. (2017) obtained soft labels by calibrating the classification results of a model that learned a clean dataset using knowledge graphs established based on Wikipedia information. In that study, a soft label was a class membership probability set that single data were expected to have, and soft labels were considered in the loss function of the classification model together with the original data labels. However, both of those methods require data sorting by engineers. Kohler et al. (2019) observed that the SoftMax score distribution of the CNN is different between noisy samples and clean samples. Thus, they proposed a method for detecting noisy samples and modifying their class labels by adopting the SoftMax score vector threshold. Wang et al. (2018) determined the class label for a sample and identified whether it was noisy or not using a Siamese network that had learned contrastive loss and classification loss simultaneously for discriminative feature learning. They calculated a score that determines whether a sample is noisy or not in a feature space using the local outlier factor algorithm, which is a density-based outlier detection algorithm, and adjusted the weight of the classification loss in total loss when training the model. The greatest limitation of these existing studies is that they decreased the extent of reflection when training the detected noisy samples or relabeled them only with existing classes. Thus, they failed to consider the subdivision of groups depending on the differences in the patterns of the same class, the possibility of existing new classes, and

Fig. 3 The structure of the Siamese network



the appearance of unknown samples. Therefore, this study focuses on proposing a new class label reconstruction method designed to overcome these limitations.

Background

Siamese network

The Siamese network was proposed to learn an embedding space based on similarities between images (Chopra et al. 2005). In general, the Siamese network learns embedding, which minimizes the distance between samples if the sample pair (x_i, x_j) belongs to the same class and maximizes it otherwise. According to this learning purpose, the Siamese network is composed of two base networks that share network weights, as shown in Fig. 3, and is learned to minimize the contrastive loss function $CL(x_i, x_j, Y_{ij})$ in the following equation:

$$CL(x_i, x_j, Y_{ij}) = Y_{ij} \frac{1}{2} d_{ij}^2 + (1 - Y_{ij}) \frac{1}{2} \max\{0, \alpha - d_{ij}\} \quad (1)$$

where Y_{ij} is a similarity indicator of sample pairs and has a value of 1 if the sample pairs have the same class and 0 otherwise; d_{ij} is the Euclidean distance of the sample pair (x_i, x_j) in the embedding space; and α is the margin of the minimum distance that the two dissimilar samples must maintain.

Gaussian means clustering

Clustering algorithms have been used as a useful tool for data mining, data compression, and outlier detection. However, most clustering algorithms have a limitation: the number of clusters must be set in advance, which is difficult to know without prior information about data distribution. G-means clustering is a hierarchical clustering method that automatically finds the optimal k by repeatedly perform-

ing k -means clustering (MacQueen 1967) while sequentially increasing the number of clusters k (Hamerly and Elkan 2004). When training the G-means clustering algorithm, the data that belong to each cluster are converted to scores based on the characteristics of the data distribution in the cluster. The scores of good clusters are assumed to follow a Gaussian distribution. The algorithm repeatedly executes k -means clustering while increasing k until all clusters are found to be adequate. Let us assume that there are child centroids c_1 and c_2 that divide the dataset that belongs to cluster c into two groups (c_1 and c_2 can be easily found by the k -means algorithm with $k = 2$). Then, $v = c_1 - c_2$ is assumed to be a vector that determines the main directionality of cluster c , and a score $s_i = g(\langle z_i, v \rangle / \|v\|^2)$ is defined for a feature vector $z_i \in c$, where g is the standard Gaussian normalization function. Finally, the Anderson–Darling statistical test verifies whether the scores of each cluster follow the Gaussian distribution (Anderson and Darling 1952). Let us assume that when the elements of score set S_c of cluster c are sorted sequentially, the i th element is $s_{(i)}$, $i = 1, 2, \dots, n_c$, and F is the cumulative distribution function of the standard Gaussian distribution. Then, the Anderson–Darling statistic $A^2(S_c)$ is calculated by Eq. (2). If $A^2(S_c)$ is smaller than the threshold calculated according to significance level ϵ , cluster c is determined as a good cluster.

$$\begin{aligned} A^2(S_c) &= -\frac{\rho(n_c)}{n_c} \sum_{i=1}^{n_c} (2i-1) [\log(F(s_{(i)})) + \log(1-F(s_{(n_c+1-i)}))] \\ &\quad - n_c \end{aligned} \quad (2)$$

Here, $\rho(n_c) = 1 + \frac{4}{n_c} - \frac{25}{n_c^2}$ is a weight applied when the mean and variance in the score set are estimated from the data (Stephens 1974).

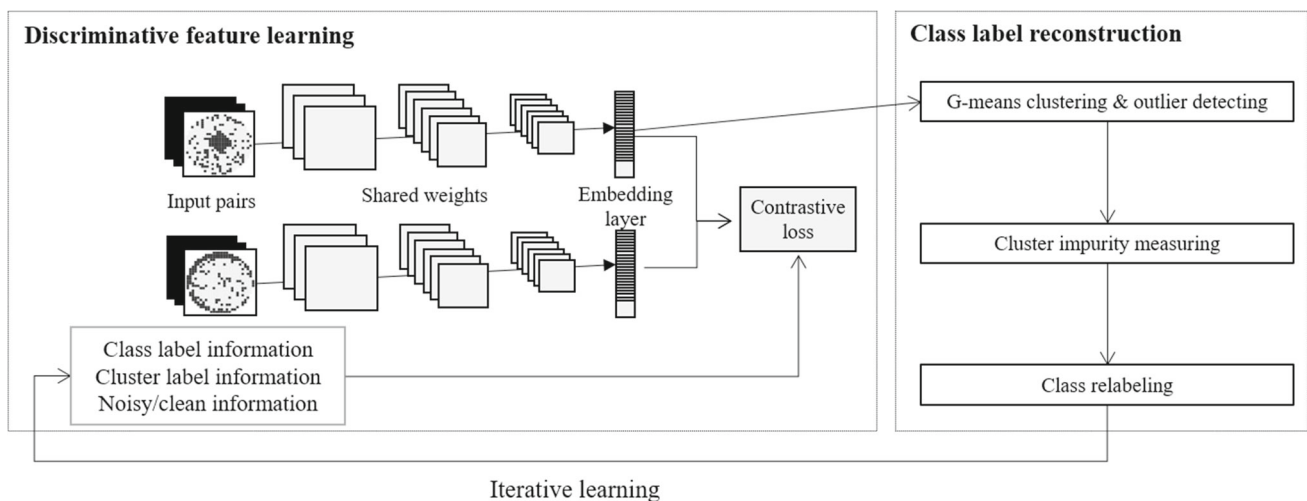


Fig. 4 The iterative learning framework for the proposed method

Methodology

In this section, we introduce the proposed uncertain class label reconstruction method. As shown in Fig. 4, the proposed method is composed of discriminative feature learning using the Siamese network and the class label reconstruction that uses the G-means clustering result in the learned feature space. We extend the contrastive loss of the existing Siamese network to consider not only the class label information but also the cluster label information. The extended contrastive loss also considers whether the sample is noisy. Furthermore, we propose methods of discovering noisy samples through outlier detection, measuring the impurity level of an individual cluster and modifying the class labels of samples that belong to the cluster based on the G-means clustering result.

Initially, the cluster label of each sample is set the same as the class label, and discriminative feature learning and class label reconstruction are performed alternately. The result of feature learning provides a better feature space for G-means clustering, thus enabling more sophisticated class label reconstruction. The class label, cluster label, and noise/clean information modified through class label reconstruction enable the learning of strong discriminative features. The iterative learning process enables the subdivision of patterns of the same class and the discovery of new classes. Furthermore, the clustering results and the modified class label information are used to determine whether incoming data belong to one of the known classes or there are unknown defects.

Discriminative feature learning

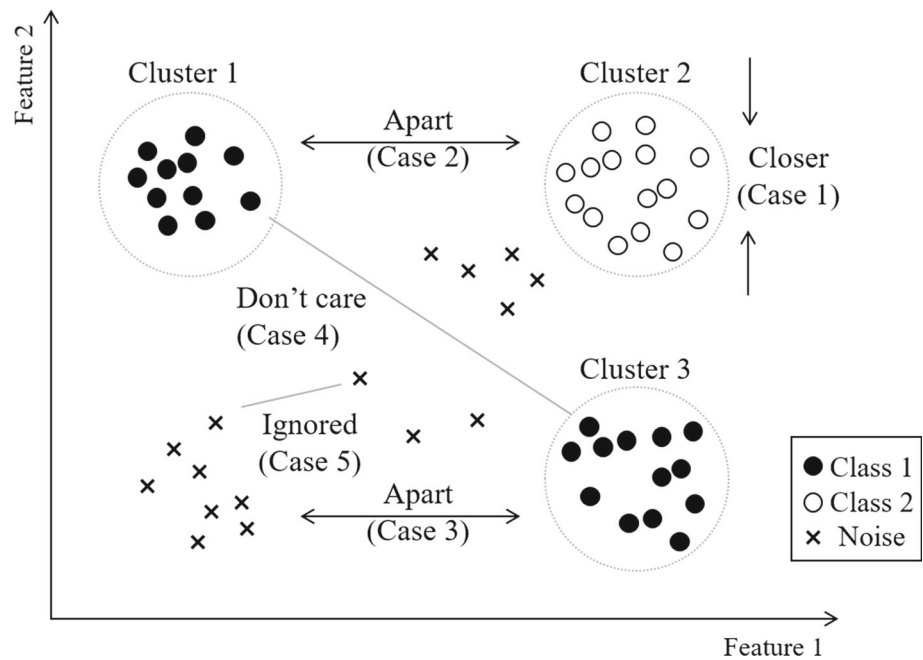
The Siamese network can learn embedding, which makes the samples of the same class closer and the samples of different

classes farther using class label information. Through this process, the Siamese network automatically learns the discriminative features that distinguish similar and dissimilar samples in the feature space, and the discriminative features can be used to find dissimilar noisy samples (Wang et al. 2018). Furthermore, the network can learn more powerful discriminative features by removing the noisy samples found in the middle of learning or reducing the degree of reflecting them during learning. In conclusion, through continuous repetition of feature learning and noise removal, the Siamese network can have high-performance noise detection capacity.

However, discriminative feature learning using the contrastive loss of existing Siamese networks in Eq. (1) is not appropriate for a situation where there is uncertainty in the WBM defect label because the following two cases are not considered. First, the existing Siamese network can only use defect label information, and as a result, samples that belong to the same class but can be considered different patterns cannot be subdivided into multiple groups. In this case, the learning of the Siamese network progresses in the direction of binding groups having different patterns in one group, which is likely to distort the feature space. Second, failing to consider or giving a low weight to noisy samples in the learning process can prevent the noisy samples from forming a new significant cluster or from being incorporated into a more appropriate class.

We assume five pair conditions that can occur in an environment where class label uncertainty exists, and we propose a new Siamese network learning loss function that considers these conditions. We determine the pair conditions by considering the class label and cluster label and considering whether the sample is noisy or clean. The five pair conditions and the learning strategy for each pair condition are as follows (see Fig. 5):

Fig. 5 The five cases of sample pairs in a feature space



- Case 1 When two samples have the same class label and cluster label, this pair can be considered to have the same pattern, and the two samples are learned to minimize the distance between them.
- Case 2 When two samples that belong to different classes have different cluster labels, this pair is considered to have different patterns, and the two samples are learned to maximize the distance between them.
- Case 3 When only one of the two samples is noisy, this pair is considered to have different patterns, and the two samples are learned to maximize the distance between them.
- Case 4 When two samples that belong to the same class have different cluster labels, they are not reflected during learning because it is unclear whether they have the same pattern or not.
- Case 5 When two samples are both noisy, the pair is not reflected during learning.

We propose a new contrastive loss function, as shown in Eq. (3), to consider the above five pair cases during Siamese network learning. Noisy samples are assumed to have different class labels from those of all the other samples. To additionally consider whether a noisy sample is included in a pair, we introduce modified class similarity indicator U_{ij} . Here, U_{ij} has a value of 1 if two clean samples have the same class label and 0 otherwise. Thus, when a noisy sample is included in a pair, class similarity indicator U_{ij} is always zero.

$$CL(x_i, x_j, U_{ij}, C_{ij}, I_{ij}) = I_{ij} \left(U_{ij} C_{ij} \frac{1}{2} d_{ij}^2 + (1 - U_{ij}) \frac{1}{2} \max \{0, \alpha - d_{ij}\} \right) \quad (3)$$

where C_{ij} is the cluster similarity indicator of the sample pair; 1 is assigned if the two samples have the same cluster, and 0 is assigned otherwise. I_{ij} is an indicator that is assigned 1 if it includes a clean sample and 0 otherwise.

Class label reconstruction and prediction

Before class label reconstruction, we clustered WBM in feature space using the G-means clustering algorithm. In the embedding space, closely located data points share similar latent features. We assume that samples that belong to the same cluster share the same defect pattern as iterative learning converges, except for some noise. Based on this assumption, the clustering result and the class label information, we perform class label reconstruction, which consists of outlier detection, cluster impurity measuring, and class relabeling.

- (1) Outlier detection: The G-means clustering algorithm might allow the inclusion of some outliers because it assumes that the distance scores in the cluster follow the Gaussian distribution, which allows some outliers with low probability. Therefore, samples whose distance from the cluster center is larger than a certain value are assumed to be outliers and are determined as noise. In other words, if the Euclidean distance between $z_i = f(x_i)$, where f is the embedding function of the Siamese

network, and sample x_i belongs to cluster c and centroid μ_c is larger than threshold δ , sample x_i is determined to be an outlier. The outlier samples are treated as noise and are not considered in the cluster impurity measurement or class relabeling stage.

- (2) Cluster impurity measuring: The impurity level of a cluster is an important measure when deciding the relabeling strategy for the samples that belong to the cluster. For example, if the impurity is low because a cluster is mostly composed of samples of one class, the cluster can be considered one group representing the class. However, if a cluster is composed of samples of different classes in similar ratios, the cluster cannot represent the class. For the noisy samples (except outliers) that belong to one cluster, a new “noise” class label is assigned temporarily. Let us assume that there is a class label set Υ_c of the samples that belong to cluster c and a composition ratio (prior probability) of class y samples, P_c^y , for $y \in \Upsilon_c$. As shown in Eq. (4), we propose entropy based on prior probability as cluster impurity measurement Imp_c . Furthermore, the representative class of cluster c is set as $R(c) = \operatorname{argmax}_{y \in \Upsilon_c} P_c^y$.

$$Imp_c = \sum_{y \in \Upsilon_c} (-P_c^y \times \log_{C+1}(P_c^y)) \quad (4)$$

- (3) Class relabeling: To divide the clusters into three categories according to the impurity level, we set the lower and upper limits of impurity, θ^{low} and θ^{high} , respectively. For each cluster category, a different relabeling strategy is adopted, as shown in Fig. 6. For the first category, if $Imp_c \leq \theta^{low}$, cluster c is assumed to be the clean group of representative class $R(c)$. In this case, the samples with a different class from $R(c)$ are considered initially mislabeled samples, and they are assigned class label $R(c)$. Second, in the case of $\theta^{low} < Imp_c \leq \theta^{high}$, we assume cluster c as the noisy group of $R(c)$. In this case, to minimize the risk of incorrect relabeling, the samples with a different class from $R(c)$ are treated as noise (see Fig. 6b). Finally, in the case of cluster c , where $Imp_c > \theta^{high}$, since the reliability of the representative class is low, we assign a new class label to the samples (see Fig. 6c). In this process, a class label can be assigned to some noisy samples, but subsequent iterative discriminative feature learning and class label reconstruction will determine the samples as noisy. In this case, relabeling can inform the existence of a new class, such as the multidefect class.

The representative class information of the clusters for which class label reconstruction has been completed is used for class prediction of the incoming data. To explain this in detail, let us assume that centroid μ_c and representative class

$R(c)$ of cluster c are both decided and reliable. In this case, if the Euclidean distance between $z_{new} = f(x_{new})$ and μ_c is smaller than δ , x_{new} , is considered to belong to c . If z_{new} does not belong to any cluster, it can be determined to be unknown. However, because there is a possibility that z_{new} belongs to multiple clusters, class label y^* is assigned to incoming data x_{new} using the following rule:

$$y^* = \begin{cases} R\left(\operatorname{argmin}_{c \in C} \|z_{new} - \mu_c\|^2\right) & \text{if } \min_{c \in C} \|z_{new} - \mu_c\|^2 \leq \delta \\ \text{“Unknown”} & \text{Otherwise} \end{cases} \quad (5)$$

where C is the cluster label set.

Experiments

Experimental design

The WM-811K dataset used in this study consisted of 811,456 WBMs collected from real-world fabrications (Wu et al. 2015). The original dataset contained image data of various sizes, and some data labels were not recorded. Consequently, this experiment verified the proposed method using the labeled data of 26×26 size that contained the largest labeled data. The dataset used in this experiment consisted of 13,203 normal (nondefect class) data and 837 data in eight defect classes (center, edge-ring, donut, edge-local, local, near-full, random, and scratch defects). Hence, the sample size imbalance between normal and abnormal data was large, and the amount of data for each defect class was small. Therefore, we applied random undersampling to the normal data and data augmentation using a random rotation of $[-90^\circ, 90^\circ]$ to the defect image data to make the sample size of each class in [1000, 2000].

The optimal base CNN structure of the Siamese network was selected through validation. We set 80% of the total dataset as the validation data for each class and set whether the Siamese network classified the classes of input pairs correctly as the performance indicator. It was assumed that all of the labels of the validation data were correctly assigned. Consequently, class label reconstruction was not considered in the validation. In detail, we trained the existing Siamese network based on 80% of the validation dataset pairs and then selected a CNN structure that showed a high classification accuracy for the remaining 20% of the data. For validation, the contrastive loss of Eq. (1) at $\alpha = 1$ was applied. If α was larger than 0.5, it was determined that the classes of the pair were not the same. We applied a grid search to find the CNN structure. To minimize the search range, we set the CNN layer structure as a convolutional layer-convolutional layer-max pooling layer-fully connected layer and the optimizer as an Adam optimizer (Kingma and Ba 2014) in advance. For the

Fig. 6 The three types of class label reconstruction based on impurity level

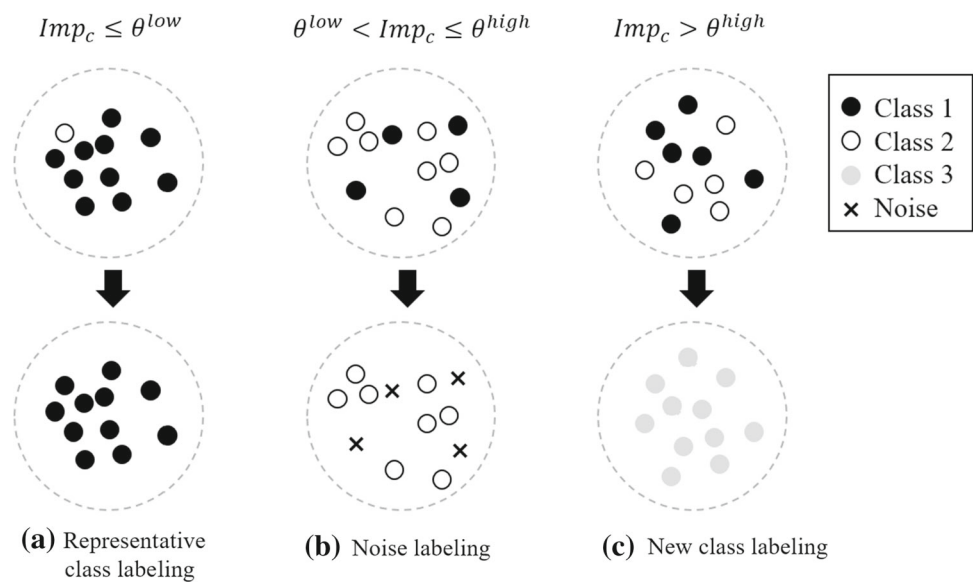


Table 1 Optimal network parameters from the grid search

Network parameters	Grid search result
Learning rate	0.001
Filter size in the 1st convolutional layer	3 × 3
Filter size in the 2nd convolutional layer	4 × 4
Number of feature maps in the 1st convolutional layer	32
Number of feature maps in the 2nd convolutional layer	64
Number of nodes in the fully connected layer	16

activation function of the convolutional layers, the rectified linear unit (Nair and Hinton 2010) was used, and the max-pooling filter size was set to 2×2 . In addition, the learning rate, filter size in the convolutional layers, number of feature maps in the convolutional layers, and number of nodes in the fully connected layer were found through a grid search. Table 1 lists the optimal network parameters as a result of the search. To obtain optimal hyperparameters δ , θ^{high} , and θ^{low} , we executed a grid search on the validation dataset. A randomly selected 10% of the dataset was corrupted to have different class labels different from the original labels. The search ranges for δ , θ^{high} , and θ^{low} were set as [0.1, 1.0], [0.7, 1.0], and [0.1, 0.4], respectively. We adjusted θ^{high} and θ^{low} at intervals of 0.05 and δ at intervals of 0.1, and we selected the hyperparameter set that yielded the best validation accuracy. The selected δ was 0.7, θ^{high} was 0.9, and θ^{low} was 0.3.

To verify the proposed method in various ways, we designed three experiments. The first experiment examined whether the proposed method can improve the accuracy of the test dataset by correcting the class labels with the training

progress in a situation where the class labels of some training data are incorrect. In addition, we examined whether the class samples with similar patterns are clustered well in the embedding space by visualizing them in a 2D space. The second experiment observed the changes in the classification performances of the proposed method and the benchmark classification models according to the corruption ratio, which is the ratio of the incorrect class labels of the training data. From the experiment, we could demonstrate that the proposed method is robust to the uncertainty of the class labels. The third experiment examined the performance of the proposed model detecting unknown samples after setting some random classes as unknown classes. In every experiment, 80% of the data were used as the training dataset, and 20% of the data were used for the test dataset for each class. In the third experiment, all data were used as the test dataset for unknown classes.

Class label reconstruction results

To verify whether the mislabeled data could find their original class label through the proposed method, we corrupted 10% of the training data by randomly assigning different class labels. Then, we performed discriminative feature learning and class label reconstruction for 10 epochs in total. We assumed that iterative learning converged if the class label change rate was less than 0.1% after 10 epochs. Here, discriminative feature learning means 20 epochs of Siamese network learning. Table 2 shows the test accuracies, noise rates, class label change rates, number of clusters, and number of classes in each epoch. All measurements are training results except for the test accuracies. First, the accuracy increased as the initial high noise rate sharply decreased, which demonstrates that the samples initially classified as

Table 2 Iterative class label reconstruction results using 10% of the corrupted training dataset

Epoch	Test accuracy	Noise rate (%)	Class label change rate (%)	Number of clusters	Number of classes
1	58.2	18.8	1.4	31	9
2	72.6	5.9	2.4	44	9
3	77.7	1.9	3.7	69	9
4	76.4	0.7	3.4	69	9
5	81.2	0.4	2.3	78	9
6	87.3	0.3	1.3	69	9
7	84.6	0.5	1.6	79	9
8	88.4	0.4	0.8	79	9
9	89.6	0.1	0.5	75	9
10	89.6	0.2	0.1	71	10

noisy found their original class labels well, resulting in increasing accuracy. Furthermore, from epoch 4, even though the noise rate decreased negligibly, discriminative feature learning and class label change drove an increase in accuracy. In addition, an increase in the cluster number did not induce an increase in the class number. This indicates that even though the groups in the class were subdivided, there was a low risk of indiscriminate new class generation. Finally, a new class was generated in epoch 10; however, there was no decrease in accuracy. This suggests that samples that were not classified correctly into existing classes formed a new class. It seems that too many clusters were generated compared to the number of classes. A large number of clusters might not be efficient for engineers to find root causes. However, considering that class labels are based on human judgement and thus are noisy (Liu and Chien 2013), cluster information can help engineers find root causes by supplementing revised class information.

Figure 7 shows the visualization of the clustering results of the training dataset in the feature space in a 2D space through t-stochastic neighbor embedding (Maaten and Hinton 2008). Figure 7a shows that with training progress, the initially dispersed samples were clustered by class and multiple pattern clusters were formed, even in the same class. Furthermore, Fig. 7b shows that the number of noisy samples decreased as the class labels were corrected, as the noisy samples that had been dispersed at first were selected as members of specific clusters according to pattern similarity. In particular, in the last epoch, a new class was formed between edge-local class samples and local class samples. As shown in Fig. 8, this new class clearly had a different pattern compared with the patterns of the edge-local and local class samples. The newly generated class is considered a multitype defect that contains all edge-local and local patterns. This result reveals that the proposed method generates a new class for samples that are difficult to explain by the existing known class label only.

Figure 9 shows the clusters that have representative defect patterns by class among the 71 clusters that were generated after 10 epochs of iterative learning. The class label reconstruction results show that the proposed method can subdivide the same class samples into multiple groups with different patterns. For example, as shown in Fig. 9a, the center defect class was subdivided into multiple groups according to the distribution size of the defect cells at the center. In addition, Figs. 9b–d reveal that clusters that were considered to have different patterns were subdivided well for each class. Using the subdivided cluster label information obtained through the proposed method enables more detailed abnormal process detection compared to the existing method of using only class label information, and this can improve the yield.

Prediction performance comparison results

To verify the performance of the proposed method when the class label has uncertainty, we compared the results with the results of the following representative classification models based on accuracy: CNN, K-nearest neighborhood (KNN) (Altman 1992), decision tree (Quinlan 1999), random forest (Ho 1998), and naive Bayes (Domingos and Pazzani 1997). For the CNN structure, we applied the SoftMax output layer, which has the same structure as that of the base model of the Siamese network. For the other classification models, we applied the default model parameters provided by the scikit-learn library of the Python programming language (Pedregosa et al. 2011).

Table 3 shows the accuracies of the proposed method and the comparison classification models for each corruption level when the ratio of data with corrupted class labels was set from 0 to 40%. First, the shallow machine learning models, except the CNN, were found to be inappropriate for WBM classification because their accuracies were clearly

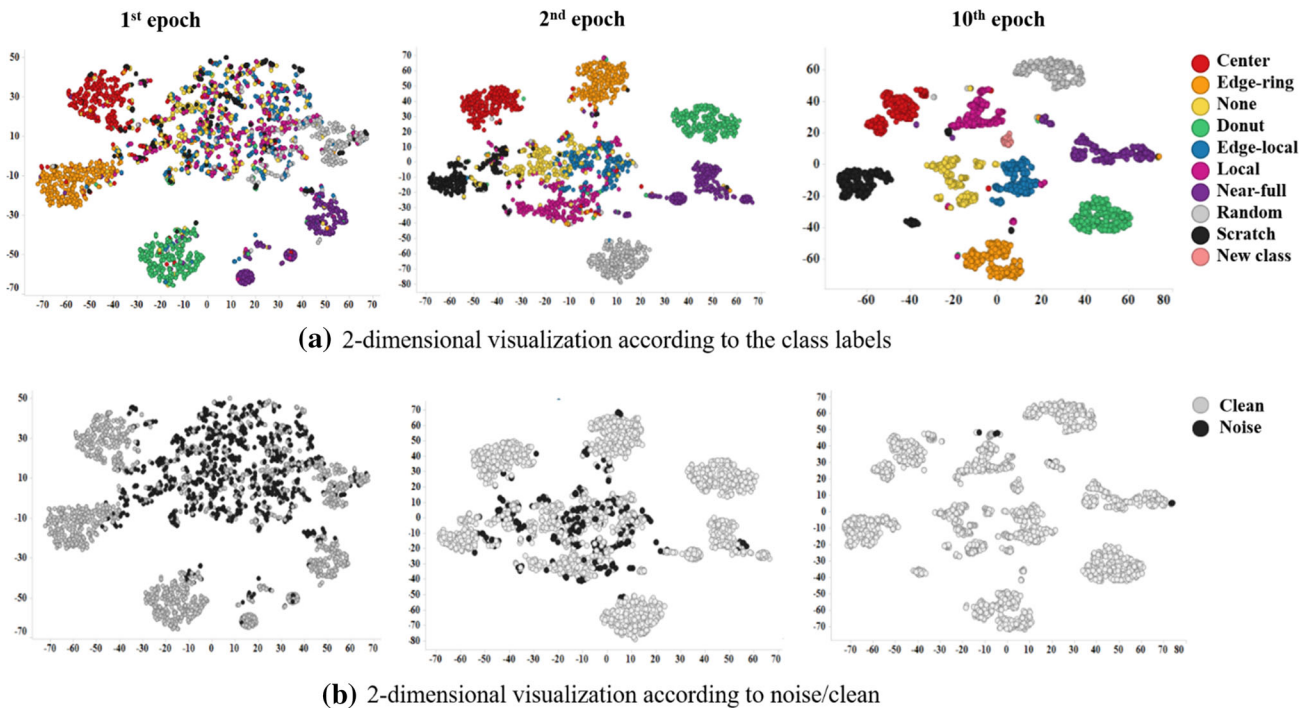


Fig. 7 2-Dimensional visualization of the clustering results for the corrupted training dataset

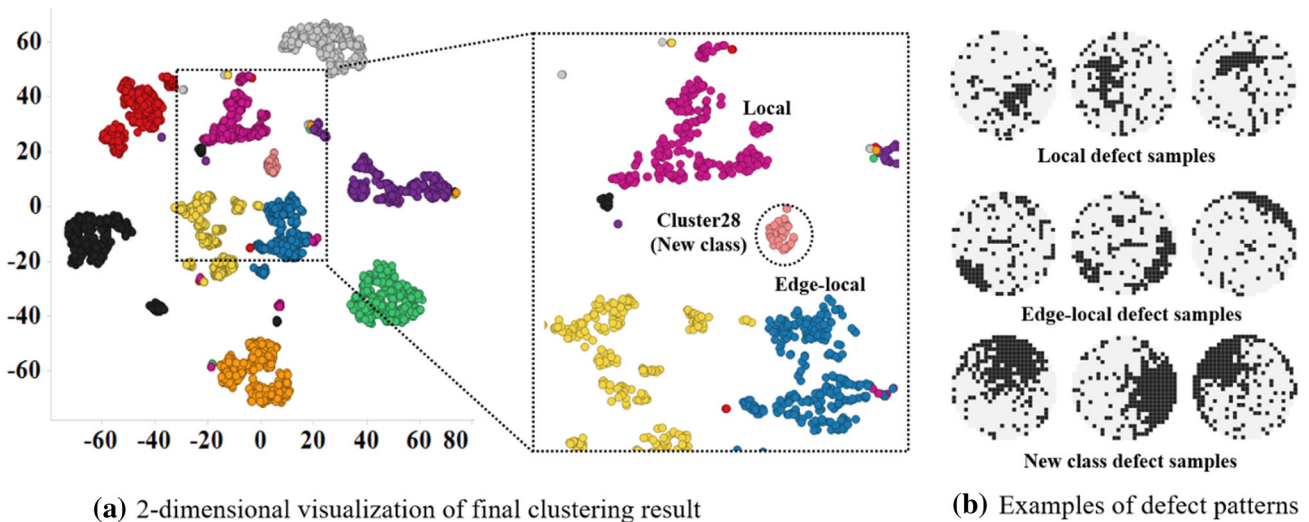


Fig. 8 New class generation after the tenth epoch of the class label reconstruction

low for every corruption level. The proposed method showed an accuracy that was only 1% higher than that of the CNN in corruption-free situations. However, as the corruption level increased, the difference between the proposed method and CNN increased, and the difference increased to 7.8% at a 30% corruption level. The proposed method can classify unknown classes, unlike other models; thus, there is a significant difference in accuracy. This difference in classification performance indicates that when the class label has uncertainty, the class label reconstruction of the proposed method

is effective. However, when the uncertainty is large enough that the corruption level is 40%, the performance difference between the compared models is not large. This suggests that the class label reconstruction works well only to a certain level of uncertainty.

Unknown sample detection results

In addition, we verified whether the proposed method can determine unknown class samples well. We designed three

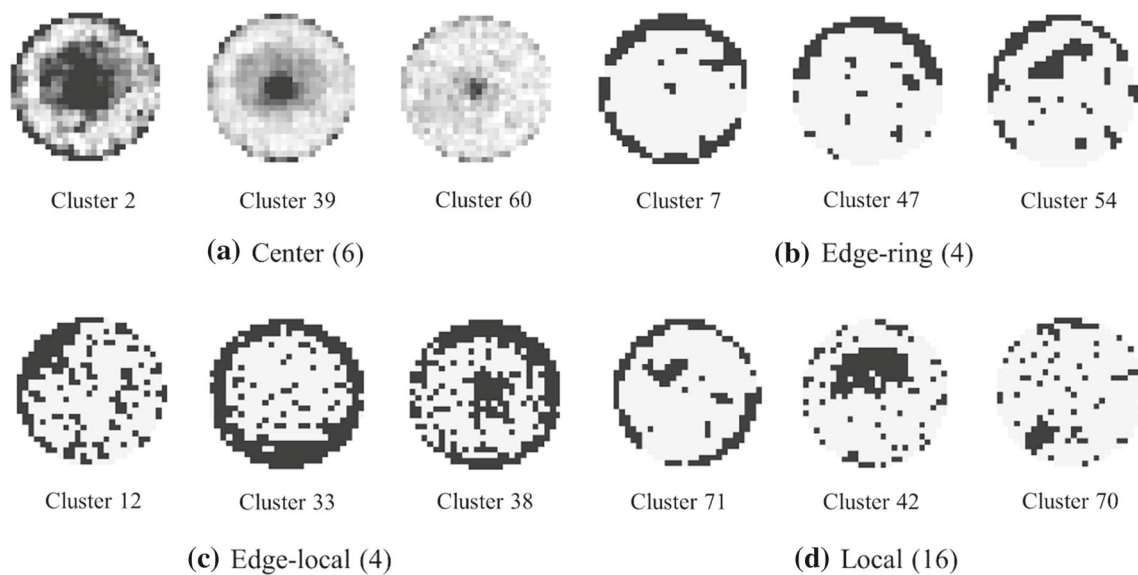


Fig. 9 Segmentation results derived with the proposed method

Table 3 Performance comparison results of the proposed method and benchmarks of WBM classification with different corruption ratios

Method	Corruption ratio				
	0%	10%	20%	30%	40%
Proposed	91.2	89.6	83.7	75.6	64.2
CNN	90.2	84.7	80.2	67.8	63.3
KNN	74.4	72.6	68.7	65.0	58.8
Decision tree	62.7	62.3	61.6	61.3	59.3
Random forest	68.8	65.6	65.3	62.3	61.1
Naive Bayes	62.5	59.5	55.3	59.6	44.6

The best results are shown in bold for each corruption ratio

scenarios, where the unknown class sets are {donut}, {donut, scratch}, and {donut, scratch, random}. Corruption of the training dataset class label was not considered in this study.

Table 4 shows the accuracy for each defect pattern according to different numbers of unknown classes. Figure 10 shows a visualization of the results of class label reconstruction in feature space in a situation where there are three unknown classes in 2D using t-stochastic neighbor embedding. Figure 10a shows the visualization results of the training samples, and Figs. 10b and c show the visualization results of the nine class test samples, including an unknown class. As shown in Table 4, the proposed method derived a high accuracy above 85% for known classes used in learning except for some defect patterns. In the case of the donut defect, which is one of the unknown classes, the accuracy was 96–97%, until the number of unknown classes was two. Thus, the performance did not degrade for known classes as well, and the accuracy was 87.5% when the number of unknown classes became three. This high performance is presumably

because the donut defect samples generate clearly different features from those of other known classes, as shown in Fig. 10b. However, the proposed method shows low detection ratios of 70.3% and 44.8% for random and scratch defects, respectively, when three unknown classes exist. This seems to suggest not only that the two class samples generate features similar to the nondefect and edge-local classes but also that the similarities of features are small between the same class samples. This result contradicts the formation of features, which were clearly different from those of other classes when random and scratch defects were used in training, as shown in Fig. 7. In other words, for more elaborate and unknown defect detection, an improvement to distinguish unknown and known class samples in the embedding space is required. However, the results showing the possibility of simultaneous performance of class label reconstruction and unknown sample detection are encouraging.

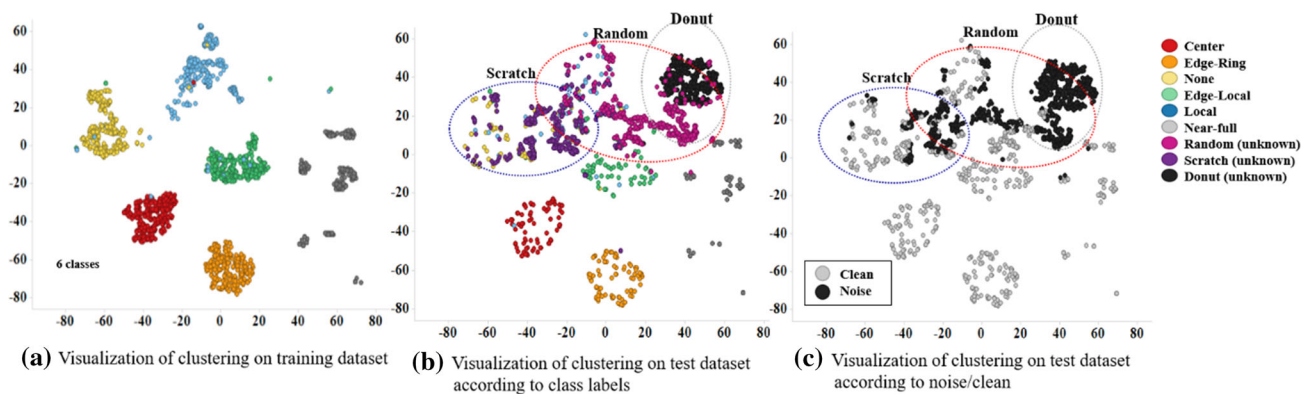
Conclusion

This study proposed a class label reconstruction method that is applicable to situations where the labeling of a WBM is uncertain. The proposed method is designed to subdivide one class with multiple patterns into multiple groups to create a new class for samples that cannot be classified into existing classes and to detect unknown class samples. Experiments using WM-811K, which is a real-world WBM dataset, revealed that these functions of the proposed method worked well. In particular, the findings showed that a new defect class of a multidefect type could be created. Furthermore, we showed that the proposed method can maximize the clas-

Table 4 The class-specific accuracies (%) according to the number of unknown classes

Number of unknown classes	Center	Edge-ring	None	Edge-local	Local	Near-full	Random	Scratch	Donut
1	96.3	98.9	84.1	91.5	86.1	100.0	92.2	62.8	96.3
2	97.5	100.0	76.1	84.5	89.4	100.0	96.7	42.1	97.5
3	97.5	100.0	85.0	88.2	80.9	100.0	70.3	44.8	87.5

The results for unknown classes are shown in bold

**Fig. 10** Visualization results on the training dataset and test dataset for the situation where three unknown cases exist

sification accuracy for the test dataset by removing the class label corruption. This enables the proposed method to perform better than the compared classification models when labeling has uncertainty.

We consider our future research going in two directions. First, the clustering results of the proposed method divide the samples of one class into detailed groups, thus enabling the detection of multiple patterns in the same class. However, the experimental results showed that the number of clusters was too large compared to the number of classes. Additionally, there is a possibility that the sample distribution, which includes the isolated noisy sample and initial clustering result in the initial embedding space, affects the final class label reconstruction result. Therefore, methods to more strictly control the subdivision of patterns and initially isolated noise should be researched to enable better interpretation of the clustering results. Second, we showed the possibility of detecting unknown samples through the proposed method. However, in the feature space of the Siamese network, some unknown samples were remarkably close to the existing class area in the feature space. This not only increases the incorrect classification ratio for unknown classes but also distorts the interpretation of the distribution information of unknown class samples. Therefore, future studies should propose a method that can detect unknown samples more accurately by researching an embedding learning method that allows the unknown samples to be more easily separated from the existing class area.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) (Grant No. NRF-2019R1A2B5B01070358).

References

- Adly, F., Yoo, P. D., Muhaidat, S., & Al-Hammadi, Y. (2014). Machine-learning-based identification of defect patterns in semiconductor wafer maps: An overview and proposal. In *2014 IEEE international parallel & distributed processing symposium workshops* (pp. 420–429). IEEE. <https://doi.org/10.1109/IPDPSW.2014.54>.
- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185. <https://doi.org/10.2307/2685209>.
- Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *The Annals of Mathematical Statistics*, *23*(2), 193–212. <https://doi.org/10.1214/aoms/1177729437>.
- Chang, C.-W., Chao, T.-M., Horng, J.-T., Lu, C.-F., & Yeh, R.-H. (2012). Development pattern recognition model for the classification of circuit probe wafer maps on semiconductors. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, *2*(12), 2089–2097. <https://doi.org/10.1109/TCPMT.2012.2215327>.
- Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR)* (vol. 1, pp. 539–546). IEEE. <https://doi.org/10.1109/CVPR.2005.202>.
- Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*(2–3), 103–130.
- Ferain, I., Colinge, C. A., & Colinge, J.-P. (2011). Multigate transistors as the future of classical metal–oxide–semiconductor field-effect

- transistors. *Nature*, 479(7373), 310–316. <https://doi.org/10.1038/nature10676>.
- Guan, D., Yuan, W., Lee, Y.-K., & Lee, S. (2011). Identifying mislabeled training data with the aid of unlabeled data. *Applied Intelligence*, 35(3), 345–358. <https://doi.org/10.1007/s10489-010-0225-4>.
- Hamerly, G., & Elkan, C. (2004). Learning the k in k-means. In *Advances in neural information processing systems* (vol. 17, pp. 1–8). <https://papers.nips.cc/paper/2526-learning-the-k-in-k-means>
- Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832–844. <https://doi.org/10.1109/34.709601>.
- Kim, J., Lee, Y., & Kim, H. (2018). Detection and clustering of mixed-type defect patterns in wafer bin maps. *IIEE Transactions*, 50(2), 99–111. <https://doi.org/10.1080/24725854.2017.1386337>.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *The third International Conference on Learning Representations, ICLR 2015*, 1–15. <https://arxiv.org/abs/1412.6980>
- Köhler, J. M., Autenrieth, M., & Beluch, W. H. (2019). Uncertainty based detection and relabeling of noisy image labels. In *2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)* (pp. 33–37). <https://arxiv.org/abs/1906.11876>
- Kyeong, K., & Kim, H. (2018). Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks. *IEEE Transactions on Semiconductor Manufacturing*, 31(3), 395–402. <https://doi.org/10.1109/TSM.2018.2841416>.
- Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., & Li, L.-J. (2017). Learning from noisy labels with distillation. In *2017 IEEE international conference on computer vision (ICCV)* (Vol. 2017-October, pp. 1928–1936). IEEE. <https://doi.org/10.1109/ICCV.2017.211>.
- Liu, C.-W., & Chien, C.-F. (2013). An intelligent system for wafer bin map defect diagnosis: An empirical study for semiconductor manufacturing. *Engineering Applications of Artificial Intelligence*, 26(5–6), 1479–1486. <https://doi.org/10.1016/j.engappai.2012.11.009>.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (vol. 1, pp. 281–297).
- Nair, V., & Hinton, G. E. (2010). Rectified linear units improve Restricted Boltzmann machines. In *ICML 2010 - Proceedings, 27th international conference on machine learning* (pp. 807–814). Retrieved from <https://www.scopus.com/inward/record.uri?eid=2-s2.0-77956509090&partnerID=40&md5=70e2e88c9faa609cc4bd7221fc47e5ca>.
- Nakazawa, T., & Kulkarni, D. V. (2018). Wafer map defect pattern classification and image retrieval using convolutional neural network. *IEEE Transactions on Semiconductor Manufacturing*, 31(2), 309–314. <https://doi.org/10.1109/TSM.2018.2795466>.
- Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4), 275–306. <https://doi.org/10.1007/s10462-010-9156-z>.
- Patrini, G., Rozza, A., Menon, A. K., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2233–2241). IEEE. <https://doi.org/10.1109/CVPR.2017.240>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Quinlan, J. (1999). Simplifying decision trees. *International Journal of Human-Computer Studies*, 51(2), 497–510. <https://doi.org/10.1006/ijhc.1987.0321>.
- Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347), 730–737. <https://doi.org/10.1080/01621459.1974.10480196>.
- Vahdat, A. (2017). Toward robustness against label noise in training deep discriminative neural networks. In *Advances in neural information processing systems* (pp. 5597–5606). <https://arxiv.org/abs/1706.00038>.
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., & Belongie, S. (2017). Learning from noisy large-scale datasets with minimal supervision. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 6575–6583). IEEE. <https://doi.org/10.1109/CVPR.2017.696>.
- Wang, C.-H., Kuo, W., & Bensmail, H. (2006). Detection and classification of defect patterns on semiconductor wafers. *IIE Transactions*, 38(12), 1059–1068. <https://doi.org/10.1080/07408170600733236>.
- Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., & Xia, S.-T. (2018). Iterative learning with open-set noisy labels. In *2018 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 8688–8696). IEEE. <https://doi.org/10.1109/CVPR.2018.00906>.
- Wu, M.-J., Jang, J.-S. R., & Chen, J.-L. (2015). Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Transactions on Semiconductor Manufacturing*, 28(1), 1–12. <https://doi.org/10.1109/TSM.2014.2364237>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.