

Similarity Search on Wafer Bin Map through Nonparametric and Hierarchical Clustering

Jea Hoon Lee, Il-Chul Moon, and Rosy Oh

Abstract—Searching for and comparing similar wafer maps can provide crucial information for root cause analysis in the manufacturing process of integrated circuits. Owing to the high dimensionality and complexity of defect patterns, comparison of similar maps in their entirety is inefficient. This paper proposes an automated similarity ranking system with a novel feature set as a reduced representation of wafer maps. To detect systematic failure patterns across wafer maps, we use nonparametric Bayesian clustering based on the Dirichlet process Gaussian mixture model, and hierarchical clustering based on the symmetric Kullback–Leibler divergence. The proposed features are efficient because they require minimal computation and storage; furthermore, they allow for highly discriminative rankings of similar failure patterns. Thus they are suitable for large-scale analysis of wafer maps. The proposed method is experimentally verified using a real wafer map dataset from a semiconductor manufacturing company, and a subset of WM-811K.

Index Terms—Similarity ranking, nonparametric Bayesian clustering, wafer bin map, defect pattern

I. INTRODUCTION

INTEGRATED CIRCUIT (IC) manufacturing is a complex and sophisticated process consisting of hundreds of steps. Therefore, it is susceptible to defects. Early defect detection and root cause analysis are imperative, considering that defects are directly associated with product quality. Wafer bin maps (WBMs) are widely used as visual representations of defective dies on a wafer [1]. In several cases, WBMs exhibit the signature patterns of specific processes causing defects [2], [3]. Hence, comparing the process histories of similar failures in a database may provide sufficient information to perform defect diagnosis and root cause analysis without predetermined classes [4]. A similarity search on a WBM also enables the generation of new categories corresponding to new defect patterns [5].

With the development of machine learning, several machine learning algorithms have been applied to identify groups of similar WBMs without pattern labels [1], [6]. Although these methods successfully use deep neural networks such as variational autoencoder (VAE) [7], [8], they may fail to identify the top-N similar wafers in a cluster when a query wafer is to be compared with a set of already-known similar wafers. Using entire WBMs in a cluster to track the process history can lead to defect misdiagnosis if some wafers are not significantly similar.

The authors are with Department of Industrial & Systems Engineering, Korea Advanced Institute of Science and Technology, Daejeon, 34141, Republic of Korea (e-mail: jeahoon.lee@kaist.ac.kr; icmoon@kaist.ac.kr; rosy.oh5@gmail.com).

Corresponding author: Rosy Oh.

Using similarity as a criterion for including a WBM in a group to be examined, root cause analysis can be facilitated by comparing wafers with a higher degree of similarity. Therefore, similarity measurement and ranking are required to improve the efficiency of root cause analysis. As the notion of object similarity is inherently subjective, a means of mathematically measuring similarity is necessary to develop a similarity ranking system [9].

Considering the current wafer production rate and the sheer volume of accumulated WBM data, a reduced representation of WBMs should be employed to minimize computing cost and data storage, as well as maintain performance in WBM analysis. To this end, various methods have been proposed based on density [10], projection and geometrical features [11], [12], as well as features generated by convolutional neural networks (CNNs) [13], [14]. However, in most cases, these methods rely on complex feature processing and require high-dimensional features to achieve adequate discriminative power.

This paper proposes an automated similarity ranking method that includes systematic similarity measurement and a reduced WBM representation through nonparametric Bayesian clustering based on the Dirichlet process Gaussian mixture model (DPGMM) [15], and agglomerative hierarchical clustering method (HCM) based on the symmetric Kullback–Leibler divergence (SKLD) [16]. 1) The DPGMM detects local defect patterns on each WBM without specifying the number of clusters, and the patterns are used as resources of representative defect patterns. 2) Hierarchical clustering partitions all local defect patterns from all WBMs into a number of subgroups based on their similarity; thus, a subgroup indicates a collection of similarly located defective chips in the set of all WBMs. 3) The weight vector with respect to these subgroups represents the features of the reduced representations of defect patterns. The proposed weight vectors require only simple computation and minimal storage during the search for similar WBMs given a query. The main contributions of this study are summarized as follows:

- The proposed system enables automated similarity ranking without any predefined features and classes.
- The proposed system can be used for not only rare but also predefined defect patterns with improved performance and reduced computational time.

The remainder of this paper is organized as follows. In Section II, we review the related work. In Section III, we describe the proposed similarity ranking method. Section IV presents experiments conducted using real datasets, and we

TABLE I
RELATED WORK

Objectives	Description	Supervision	Reference	Method		Dataset
				Clustering	Metric	
Local clustering	Clustering defective chips on a wafer map to identify local defect patterns	Unsupervised	Wang <i>et al.</i> (2006) [17]	FCM/HCM		S(1), P(8)
			Yuan <i>et al.</i> (2011) [18]	SCM		S(20), P(6)
			Kim <i>et al.</i> (2018) [19]	iWMM		S(1), P(8)
			Jin <i>et al.</i> (2019) [20]	DBSCAN		WM-811K
WBM clustering	Clustering wafer maps to search for similar ones	Unsupervised	Hsu <i>et al.</i> (2007) [1]	ART1		WM-811K
			Zhang <i>et al.</i> (2013) [6]	SR/HCM		P(69, 82)
			Tulala <i>et al.</i> (2018)[7]	VAE/HCM		WM-811K
			Santos <i>et al.</i> (2019) [8]	CVAE/HCM		S(5k), P(6k)
			Hwang and Kim (2020) [4]	DPGMM-VAE		P(500)
Similarity ranking	Ranking wafer maps in order of increasing similarity to a queried one	Supervised	Nakazawa and Kulkarni (2018) [13]		Hamming distance	WM-811K
			Yu <i>et al.</i> (2019) [14]		Hamming distance	WM-811K
		Unsupervised	Wu <i>et al.</i> (2015)[12]		ED/ correlation	WM-811K
			Hsu <i>et al.</i> (2020) [9]		WMHD	P(1k)
			Proposed	DPGMM/HCM	JSD	P(160), WM-811K

* S(m) : synthetic dataset with m samples; P(m): private dataset with m observations

provides some suggestions for a limitation of the proposed system in Section V. The paper is concluded in Section VI.

II. RELATED WORK

Herein, we propose a similarity ranking system. To this end, we search for local clusters using the DPGMM and measure similarity using the Jensen–Shannon divergence (JSD). A comparison with other studies in terms of methodology, tasks, and datasets is shown in Table I. The studies are divided into three categories according to their main objectives: local clustering, WBM clustering, and similarity ranking.

A. Unsupervised Defect Pattern Analysis on WBMs

Unsupervised learning does not consider labels. In particular, clustering algorithms are widely applied not only for detecting defect patterns on a wafer (local clustering) but also for searching for wafer maps with similar defect patterns (WBM clustering).

Local clustering. Various machine learning algorithms have been employed to recognize defect patterns, such as fuzzy C means with the HCM [17], similarity-based clustering [18], the infinite warped mixture model [19], and density-based spatial clustering for applications involving noise [20].

WBM clustering. Various methods have been used, ranging from traditional clustering with statistical inference to recent deep generative methods using neural networks. Hsu *et al.* used a two-step adaptive-resonance-theory (ART1) neural network after a spatial randomness test to construct clusters of wafer maps [1]. Even though ART1 obviates the need for the number of clusters to be predetermined, it is not suitable for high-dimensional data. Zhang *et al.* applied sparse regression to extract robust features for missing data before performing wafer map clustering using the HCM [6]. Given the aforementioned inefficiency, most recent studies have focused on VAE. Tulala *et al.* used a VAE to extract features from the latent representation of wafer maps, and they clustered wafer maps using k-means and HCM [7].

Santos *et al.* also proposed a two-step clustering method based on features extracted from a convolutional VAE [8]. Hwang and Kim proposed a one-step clustering algorithm based on a VAE mixture framework using the DPGMM [4].

Hierarchical clustering. The HCM builds a hierarchy of clusters based on a measure of dissimilarity between the clusters. It starts with each data point in its own cluster, and iteratively merges the two most similar clusters [21]. Because the HCM detects abnormal patterns rather than merging them into the main clusters [6], it has been used to extract features for the WBM analysis with simply taking linkage based on the Euclidean distance. For local clustering, Wang *et al.* employed the HCM with a single linkage to the partitions from fuzzy C means [17]. For WBM clustering, Zhang *et al.* employed complete-link HCM and developed the modified the L-method to determine the appropriate number of clusters [6]. Tulala *et al.* used Ward’s criterion [7] and Santos *et al.* used average linkage [8].

B. Similarity Ranking

Few studies have focused on WBM similarity ranking methods, even though they provide information that may be useful in the identification of potential manufacturing problems. Nakazawa and Kulkarni suggested a CNN for defect pattern recognition and applied the Hamming distance to the features extracted by a fully connected layer to retrieve wafers [13]. Yu *et al.* also used a CNN and the Hamming distance but with dimension-reduced features via principal component analysis (PCA) to measure similarity [14]. It should be noted that the aforementioned distance measurements for similarity ranking can be implemented without labels. However, it is difficult to ensure the same performance in the absence of labels because similarity ranking is performed using discriminative features for labeled data.

Wu *et al.* proposed a set of Radon-based and geometry-based features without pattern labels. Furthermore, they used a two-step similarity ranking method to reduce the time cost:

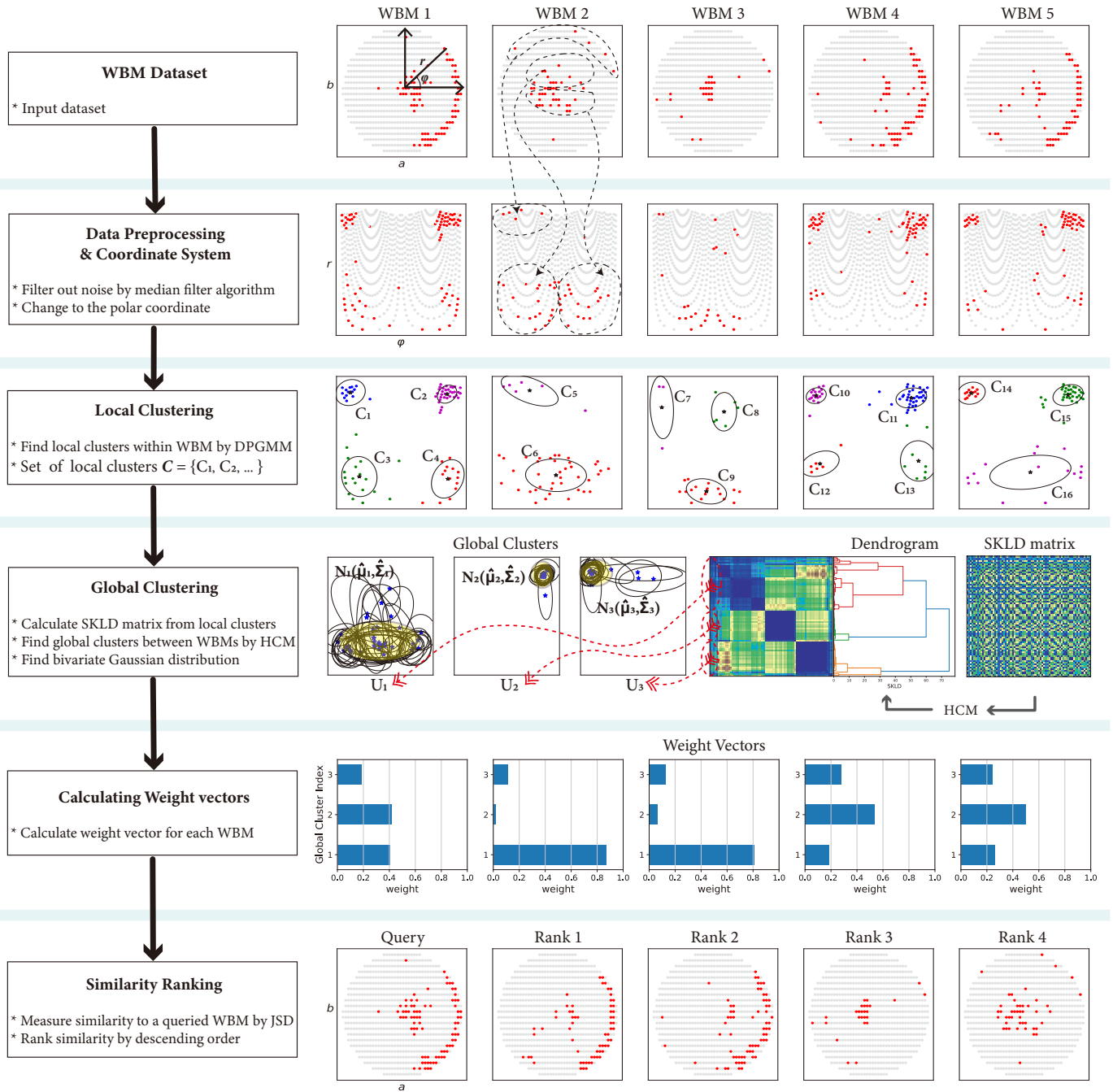


Fig. 1. Flowchart of the proposed similarity search system.

rank similarity using two-dimensional normalized correlation coefficients among the top-N similar WBMs based on the Euclidean distance of the extracted features [12]. Hsu *et al.* proposed an integrated similarity ranking framework that measures similarity without pattern labels [9]. They employed weighted and modified Hausdorff distances using enhanced defect features from a mountain clustering algorithm, determined by the pointwise comparison of all defective chips from two wafers. Even though this approach is an integrated WBM similarity measurement system and yields exceptional results, it uses the entire WBM for the similarity measurement, and this is disadvantageous for big data migration.

III. METHODOLOGY

A. Overall Framework

The general procedure of the proposed similarity ranking system is shown in Fig. 1. Initially, the wafer maps are converted to polar coordinates to ensure that *local clusters* can be obtained using the DPGMM. Aggregation of all local clusters in the dataset using the HCM based on the SKLD is employed to generate subgroups, termed *global clusters*. The WBMs are then represented as a mixture model of bivariate Gaussian distributions for the global clusters with a weight vector that is subsequently used to measure and rank similarity

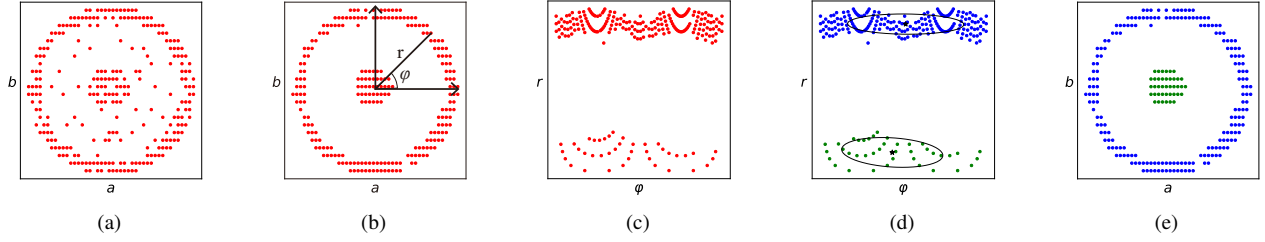


Fig. 2. (a) Original WBM, (b) median filtered wafer maps using a filter size of 3×3 , (c) polar coordinate system, (d) DPGMM result in polar coordinates, and (e) DPGMM result in Cartesian coordinates.

TABLE II
LIST OF NOTATIONS

Symbol	Description	Value	Note
\mathbf{y}_i	Status of chips on WBM i	Normal = 0; Defective = 1	$\mathbf{y}_i \in \{0, 1\}^J$
\mathbf{x}_{ij}	Polar coordinates of a defective chip on WBM i	$(\varphi_{ij}, r_{ij})^T$	$j \in A_d(i)$
\mathbf{z}_{ij}	Polar coordinates of a chip on WBM i	$(\varphi_{ij}, r_{ij})^T$	$j \in \{1, \dots, J\}$
φ_{ij}	Angular coordinate	$\varphi \in (0, 360]$	
r_{ij}	Radial coordinate	$r > 0$	
\mathcal{C}	Set of local clusters		
C_b	Local cluster		$b \in \{1, \dots\}$
U_g	Global cluster		$g \in \{1, \dots, G\}$
\mathcal{N}_g	Bivariate Gaussian distribution		
f_g	Probability density function (pdf) of \mathcal{N}_g		
$\hat{\mu}_g$	Estimated mean vector for \mathcal{N}_g	$\mu_g \in \mathbb{R}^2$	
$\hat{\Sigma}_g$	Estimated covariance matrix for \mathcal{N}_g	$\Sigma_g \in \mathbb{R}^{2 \times 2}$	
$A_d(i)$	Index set for defective chips in WBM i		

based on the JSD.

A list of notations appears in Table II. We assume that there are J chips on wafer map $i \in \{1, \dots, I\}$. Then, the status of chips, that is, binary values indicating “normal” or “defective,” is collected in the vector $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})$, and the location information in polar coordinates is summarized as $\mathbf{z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{iJ})^T$, $\mathbf{z}_{ij} = (\varphi_{ij}, r_{ij})^T$. The location information regarding defective chips corresponds to a subset of \mathbf{z}_i , denoted by $\mathbf{x}_i = \{\mathbf{z}_{ij}; j \in A_d(i)\}$.

B. Data Preprocessing and Coordinate System

Median filtering. We applied a median filter to remove noise from the wafer map. The median filter is a nonlinear filter that replaces the current value with the median value in the neighboring region. It is quite effective in impulsive noise removal and minimizes image blur [22]. Fig.2(b) shows the result of applying a median filter with a size of 3×3 to the original wafer map in Fig.2(a). It can be seen that most of the noise was removed.

Polar coordinates Most defect patterns tend to form ring-type spatial patterns, such as donuts and edges, owing to the structural characteristics of semiconductor manufacturing equipment. However, these patterns are not well represented by the DPGMM in Cartesian coordinates (a, b) . Thus, we express the WBMs in polar coordinates, that is, a two-dimensional coordinate system in which each point is determined by its distance r from the origin (radial coordinate), and the directional angle φ of the position vector (angular coordinate):

$$\varphi = \text{atan2}(b/a) \quad \text{and} \quad r = \sqrt{a^2 + b^2}.$$

Then, we obtain an elliptic shape that can be better modeled, as shown in Fig. 2. The WBM in Fig. 2(b) exhibits two patterns: center and edge. In polar coordinates, these patterns take the form of two distinguishable clusters (Fig. 2(c)), which are then identified as two local clusters by the DPGMM (Fig. 2(d)). Clearly, the two local clusters indeed represent the initial center and edge patterns when expressed in Cartesian coordinates, shown in green and blue in Fig. 2(e), respectively.

C. Local Clustering by DPGMM

In local clustering, groups of defective chips are generated on each WBM. We use the DPGMM to identify local clusters. A nonparametric Bayesian method is used because the number of clusters should not be specified so that subjectivity may be eliminated. In the DPGMM, for each defective chip $j \in A_d$, on a WBM, a random variable X_j is normally distributed with probability π_k as

$$X_j | \theta, s_j = k \sim N(\mu_k, \Sigma_k),$$

where $\theta_k = (\mu_k, \Sigma_k)$ is the set of parameters with mean vector $\mu_k \in \mathbb{R}^2$ and covariance matrix $\Sigma_k \in \mathbb{R}^{2 \times 2}$ for component k . Here, the component label s_j , $j \in A_d$ is generated by a categorical distribution with an infinite set of mixture proportions as a parameter π , that is, $s_j \sim \text{Cat}(\pi)$.

The Dirichlet process (DP) is used as a nonparametric prior over the parameters of the mixture components:

$$\theta_k | H \sim H,$$

where $H \sim DP(H_0, \alpha)$ with two parameters: a base distribution H_0 and a concentration parameter $\alpha > 0$. Another representation of the DP, termed stick-breaking construction [23], is as follows: We consider two infinite collections of independent random variables $\beta_k \sim \text{Beta}(1, \alpha)$ and $\theta_k \sim H_0$. Then, H can be expressed as

$$H = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k},$$

which is a discrete distribution, where $\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j)$, and δ_{θ_k} is the probability measure degenerate at θ_k . A set of mixture proportions π generated from the stick-breaking process follows the Griffiths-Engen-McCloskey distribution denoted by $\pi \sim \text{GEM}(\alpha)$ [24]. The joint density function is given by

$$p(\mathbf{x}, \mathbf{s}, \pi, \theta) = p(\pi) p(\theta) \prod_{j \in A_d} p(x_j | s_j, \theta_{s_j}) p(s_j | \pi),$$

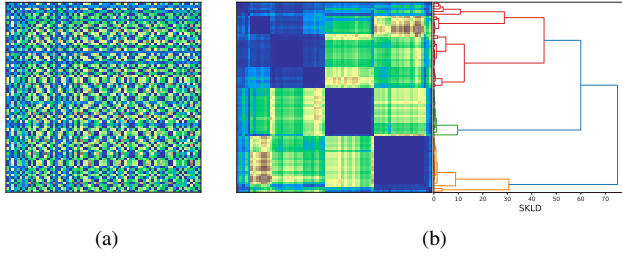


Fig. 3. (a) SKLD matrix before HCM (b) SKLD matrix and dendrogram after HCM.

where $p(\theta)$ is the prior distribution of the parameters of the Gaussian distribution.

We first obtain local clusters from each WBM using the DPGMM. Let $\mathcal{C} = \{C_b\}_{b=1}^B$ be the set of all local clusters with at least three defective chips. That is, a cluster with fewer than three defective chips is not included in \mathcal{C} . We note that these clusters are not from a specific WBM but from the entire set of WBMs. For example, if we obtain three local clusters from WBM 1, two local clusters from WBM 2, and one local cluster from WBM 3, then the set of all local clusters is $\mathcal{C} = \{C_1, C_2, \dots, C_6\}$.

D. Global Clustering by HCM

In the proposed system, the weights of the global clusters used to represent a WBM are defined as discriminative features. A global cluster is a subgroup of the set of local clusters \mathcal{C} obtained by the DPGMM. Herein, we explain the generation of the global clusters, denoted by U_g , $g \in \{1, \dots, G\}$, from the local clusters and the distribution of defective chips in U_g .

Distance measurement. We first introduce the distance measurement used in hierarchical clustering. The Kullback–Leibler divergence (KLD) is widely used to quantify the differences between two probability distributions [25], [26]. For two probability distributions P and Q with probability density functions p and q , respectively, the KLD from P to Q is defined as

$$D_{\text{KLD}}(P||Q) := \int p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) d\mathbf{x}.$$

We note that $D(P||Q) \geq 0$, where equality holds if and only if $P = Q$. However, as $D(P||Q) \neq D(Q||P)$, the KLD fails to be symmetric, and thus it cannot be used as a distance measure.

Accordingly, we use the SKLD, defined by

$$D_{\text{SKLD}}(P||Q) = D_{\text{KLD}}(P||Q) + D_{\text{KLD}}(Q||P). \quad (1)$$

For two bivariate Gaussian distributions, D_{SKLD} can be represented analytically as

$$D_{\text{KLD}}(P||Q) = \frac{1}{2} \left[\log \frac{|\Sigma_q|}{|\Sigma_p|} - 2 + \text{tr}\{\Sigma_q^{-1}\Sigma_p\} + (\mu_q - \mu_p)^T \Sigma_q^{-1}(\mu_q - \mu_p) \right]. \quad (2)$$

Fig. 3(a) shows a colormap for the $B \times B$ symmetric matrix of the SKLD for \mathcal{C} . A smaller distance between two

distributions of local clusters corresponds to a color closer to navy blue.

Hierarchical clustering. We obtain global clusters by dividing the local clusters in the set \mathcal{C} into groups. We adopt the HCM based on the SKLD with complete linkage. Further details can be found in [16], [27]. The HCM sequentially merges groups at each iteration according to decreasing similarity; this is presented in a dendrogram. Cutting the dendrogram at a given threshold distance yields the global clusters, which are distinct, whereas the objects in each cluster are broadly similar. Even though there are a few criteria (such as the L-method) that can be used to determine the cutting point [6], they do not provide appropriate subgroups, in that the subgroups function as not only reduced representation of entire WBMs but also discriminative features. Therefore, in Subsection III-G, we provide a criterion for determining the threshold to ensure better similarity searching performance.

Bivariate Gaussian distribution of global clusters. After the global clusters U_g , $g \in \{1, \dots, G\}$, have been obtained, a mathematical representation is required. To this end, we propose a bivariate Gaussian distribution \mathcal{N}_g , which can be identified by the mean and covariance matrix. Using the maximum likelihood method, we obtain the sample mean and sample covariance matrix of the defective chips in the global cluster as the two estimated parameters:

$$\hat{\mu}_g = \frac{1}{|U_g|} \sum_{s=1}^{|U_g|} \mathbf{x}_{g,s}$$

$$\hat{\Sigma}_g = \frac{1}{|U_g|} \sum_{s=1}^{|U_g|} (\mathbf{x}_{g,s} - \hat{\mu}_g)^T (\mathbf{x}_{g,s} - \hat{\mu}_g),$$

where $\mathbf{x}_{g,s}$ are the polar coordinates of defective chip s in U_g , and $|U_g|$ is the number of defective chips in U_g .

E. Calculating Weight Vectors

Weight vectors indicating the relative presence of the global clusters are used as a reduced representation for WBMs; these are the proposed features. The weight vector of wafer map i is denoted by $\mathbf{w}_i = (w_{i,1}, w_{i,2}, \dots, w_{i,G}) \in [0, 1]^G$, $i \in \{1, \dots, I\}$, and is obtained as

$$w_{i,g} = \frac{\sum_{j \in A_d} \ell_{ij,g}}{\sum_{s=1}^G \sum_{j \in A_d} \ell_{ij,s}},$$

where $\sum_{g=1}^G w_{i,g} = 1$ with $w_{i,g} \in [0, 1]$,

$$\ell_{ij,g} = \frac{f_g(\mathbf{x}_{ij}|\hat{\mu}_g, \hat{\Sigma}_g)}{\sum_{s=1}^G f_s(\mathbf{x}_{ij}|\hat{\mu}_s, \hat{\Sigma}_s)}$$

is the normalized density function of \mathbf{x}_{ij} , which represents the polar coordinates of defective chip $j \in A_d$ on WBM i , and f_g is the pdf of \mathcal{N}_g with mean vector $\hat{\mu}_g$ and covariance matrix $\hat{\Sigma}_g$. Higher weight values indicate a pronounced presence of the respective global cluster on a WBM.

As the condition $\sum_{g=1}^G w_{i,g} = 1$ ensures that the integral of the joint pdf $p(\mathbf{x}_i)$ in (3) is equal to 1, the finite GMMs

can be regarded as a clustering method for defect patterns on WBM i , with the pdf in the form

$$p(\mathbf{x}_i) = \sum_{g=1}^G w_{i,g} f_g(\mathbf{x}_i | \hat{\mu}_g, \hat{\Sigma}_g), \quad (3)$$

where $w_{i,g}$ is the g th mixture proportion. Hence, it is assumed that the cluster indicator c_i follows an independent categorical distribution with parameter \mathbf{w}_i :

$$c_i \sim \text{Cat}(\mathbf{w}_i). \quad (4)$$

That is, the features of defect patterns are summarized by the weight vector, and similar WBMs tend to have similarly distributed weights.

F. Similarity Ranking

Similarity ranking is used to retrieve and rank wafers according to their degree of similarity to a queried wafer exhibiting a specific pattern to be examined. As defect patterns tend to result from specific processes, similarity ranking can provide crucial information for root cause analysis.

The framework of the proposed system involves two steps: The JSD between the weight vector of a query and those of the other WBMs is first calculated. Subsequently, the similarity of all candidates to the query is ranked in ascending order of the JSD. A higher similarity rank indicates that the distribution of the corresponding WBM is more similar to that of the queried WBM.

The JSD is also symmetric and is widely used as a distance measure. For two categorical distributions P and Q for the cluster indicators of two WBMs in (4), with parameters \mathbf{w}_p and \mathbf{w}_q , respectively, their JSD is given by

$$D_{\text{JSD}}(P||Q) = \frac{1}{2} D_{\text{KLD}}(P||M) + \frac{1}{2} D_{\text{KLD}}(Q||M), \quad (5)$$

where

$$D_{\text{KLD}}(P||M) = \sum_{g=1}^G w_{p,g} \log \left(\frac{w_{p,g}}{m_g} \right),$$

and $m_g = 0.5(w_{p,g} + w_{q,g})$. The proposed features allow for an efficient computation of this quantity.

G. Assumptions and Hyperparameters

Number of global clusters. Determining the number of clusters, G , in hierarchical clustering is difficult. Therefore, we propose the following criterion: We compare the reconstructed with the actual WBMs, changing the number of cluster groups in the result of hierarchical clustering. The WBMs are reconstructed using a normalized likelihood function and their weight vector. The details of the process are as follows: The normalized likelihood matrix for the i th WBM is given by $J \times G$ dimensional matrix $\mathbf{L}_i = [\ell_{ij,g}^*]$, where

$$\ell_{ij,g}^* = \frac{f_g(\mathbf{z}_{ij} | \hat{\mu}_g, \hat{\Sigma}_g)}{\sum_{s=1}^J f_g(\mathbf{z}_{ij} | \hat{\mu}_s, \hat{\Sigma}_s)},$$

is the normalized likelihood function of \mathcal{N}_g for \mathbf{z}_{ij} , which is the vector of the polar coordinates of chip j on WBM i . Then,

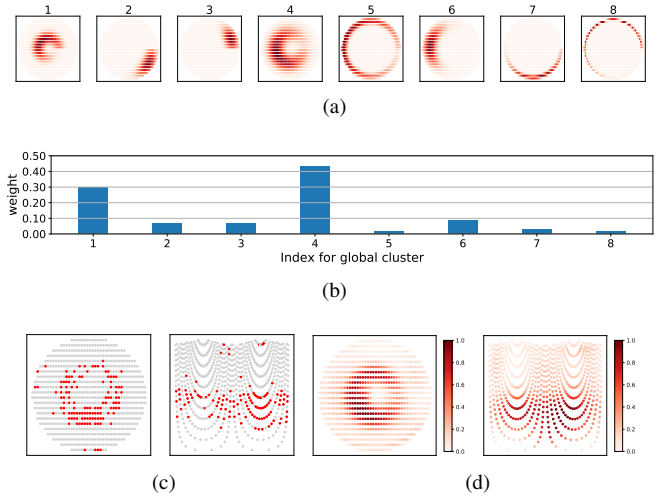


Fig. 4. (a) Normalized likelihood, (b) weights for eight global clusters, (c) example of original WBM, and (d) reconstructed map.

the status $\mathbf{y}_i^* \in \mathbb{R}^J$ of the reconstructed WBM is defined by the weighted mean of the normalized likelihood:

$$\mathbf{y}_i^* = \mathbf{w}_i \mathbf{L}_i^T.$$

We recall that the status of a selected WBM \mathbf{y}_i is a binary vector consisting of zeros (normal chip) or ones (defective chip). Hence, for a proper comparison, min-max normalization is applied to \mathbf{y}_i^* to ensure that its range becomes $[0, 1]$, and the result is denoted by \mathbf{y}_i^{**} :

$$\mathbf{y}_i^{**} = \frac{\mathbf{y}_i^* - \min(\mathbf{y}_i^*)}{\max(\mathbf{y}_i^*) - \min(\mathbf{y}_i^*)}.$$

Finally, the reconstruction quality is measured in terms of the mean squared error (MSE). The MSE for the normalized version of the reconstructed WBMs is given by

$$\text{MSE} = \frac{1}{I} \sum_{i=1}^I \text{MSE}_i, \quad \text{MSE}_i = \frac{1}{J} (\mathbf{y}_i - \mathbf{y}_i^{**})^T (\mathbf{y}_i - \mathbf{y}_i^{**}).$$

We prefer a similarity searching system with a certain number of global clusters with lower MSE values.

An example of a reconstructed map with eight global clusters is shown in Fig. 4. Fig. 4 (a) shows the normalized likelihood matrix \mathbf{L}_i , and the small figures indicate the column vector in Cartesian coordinates for each global cluster. Fig. 4(b) shows the bar graph for the weight vector of the WBM in Fig. 4(c), which represents normal and defective chips in gray and red color, respectively. Fig. 4(d) shows the reconstructed map \mathbf{y}_i^{**} .

Settings of DPGMM We employed a DPGMM with a variational inference algorithm using BayesianGaussian mixture objects [28]. The algorithm is approximated with a truncated distribution with a fixed maximum number of components (stick-breaking representation), which is specified according to the data. The DP parameter is set as $\alpha = 0.1$, and the prior for Gaussian parameters is used as the empirical distribution.

Data cleansing The following data cleansing process is required to implement the proposed system. For local clustering,

TABLE III
CONFUSION MATRIX OF SIMILARITY RANKING

True label	Model prediction	
	Similar (S)	Dissimilar (D)
Similar (S)	TP	FN
Dissimilar (D)	FP	TN

TABLE IV
EXAMPLE OF SIMILARITY RANKING

Similarity ranking	Query	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7
True label	1	1	1	2	1	3	2	3
True label (S/D)		S	S	D	S	D	D	D
Model prediction (S/D)		S	S	S	D	D	D	D
Matching result		TP	TP	FP	FN	TN	TN	TN

we remove WBMs with fewer than two defective chips in the dataset after median filtering because the map does not have any specific pattern and can thus be considered a non-pattern. After the DPGMM, we use the majority of local clusters to reflect the general local defect pattern. Specifically, a local cluster with fewer than three defective chips is not included in the set \mathcal{C} . For model evaluation, a category of defect patterns is removed if it affects fewer than ten WBMs.

The details of the public dataset are as follows. Because of the same issue as in clustering, the pattern categories “near-full” and “none” were removed from WM-811K. For model evaluation in a setting similar to that in a private dataset, after data cleansing as described above, we selected a public dataset with more than four categories and more than 100 observations. This is because less than 100 data points are not critical mass to compare the performance in various defect patterns.

IV. EXPERIMENTS

A. Model evaluation

To determine whether two WBMs are similar, we regard similarity ranking as a binary classification problem with two classes: similar (S) and dissimilar (D). The prediction is determined by the ranking threshold γ . Specifically, a ranking higher than the threshold is classified as S. The performance of this system is measured by comparing the predicted and true labels, resulting in four possible combinations: true positive (TP), false positive (FP), true negative (TN), and false negative (FN), as shown in Table III. Table IV shows the model prediction and corresponding matching results if we set $\gamma = 3$.

Receiver operator characteristic (ROC) curves are often used in binary classification problems, but they are overoptimistic when true examples are significantly rarer than negative examples [29], [30]. The main purpose is to evaluate whether retrievals have been performed well in a similar sequence. To evaluate detection performance on imbalanced data, we used the precision–recall (PR) curve at every threshold in the ranked sequence, so that the y- and x-axis correspond to precision (Pc) and recall (Rc), respectively:

$$Pc = \frac{TP}{TP + FP} \quad \text{and} \quad Rc = \frac{TP}{TP + FN}.$$

TABLE V
DISTRIBUTION OF DEFECT PATTERN TYPES (PRIVATE DATASET)















Total	A	B	C	D	E	F	G
							
160	20	25	11	17	23	41	23

TABLE VI
DESCRIPTION OF PUBLIC DATASETS

ID	Size	Total	Center	Donut	Edge-local	Edge-ring	Local	Random	Scratch
									
Set 1	35 * 40	404	60	33	82	60	101	68	0
Set 2	26 * 26	843	90	0	289	23	297	74	70
Set 3	33 * 29	370	20	0	130	55	107	12	46
Set 4	39 * 37	710	148	17	359	0	126	31	29
Set 5	44 * 41	592	18	12	390	17	101	13	41
Set 6	25 * 27	525	59	0	254	0	157	36	19
Set 7	25 * 26	254	28	0	58	70	63	23	12
Set 8	34 * 34	438	246	0	39	29	91	23	10
Set 9	32 * 29	308	26	0	181	14	61	15	11
Set 10	30 * 34	688	58	0	306	0	241	18	65
Set 11	29 * 26	448	20	0	212	55	101	0	60
Set 12	41 * 33	290	40	0	130	13	70	0	37
Set 13	39 * 37	267	25	0	112	11	96	0	23
Set 14	25 * 27	2335	2192	0	100	15	11	17	0
Set 15	35 * 31	288	21	44	99	0	109	0	15
Set 16	41 * 42	427	45	234	42	0	80	26	0

The curve indicates the trade-off between precision and recall at thresholds. As it is insensitive to the number of TNs, the PR curve, unlike the ROC curve [31], is more often used to evaluate detection performance [32], [33].

The area under the PR curve, termed average precision (AP), provides summarized information regarding the PR curve as a single numeric metric. Here, we use the approximated AP obtained by the weighted mean of precision:

$$AP = \sum_{\gamma=1}^I Pc(\gamma) [Rc(\gamma) - Rc(\gamma - 1)],$$

where $Pc(\gamma)$ and $Rc(\gamma)$ are the precision and recall given γ , respectively.

B. Dataset Description

Private dataset. The performance of the proposed method was evaluated using a real-world WBM dataset provided by a semiconductor manufacturing company. The dataset contains 160 WBMs with seven manually labeled defect patterns: center and edge (A), top local (B), center (C), left and right (D), edge ring (E), donut (F), and random (G). Table V provides a summary of the defect pattern types. The figures in the table depict samples of defect patterns, where normal and defective chips are indicated in gray and red colors, respectively. We note that pattern labels were not used in the proposed similarity ranking system but in model evaluation.

Public dataset. We also verified the performance of the proposed method using a subset of the WM-811K dataset [12]. The public dataset consists of nine categories of wafer map

TABLE VII
COMPARISON OF AP (PRIVATE DATASET)

Type	Proposed	Hsu et al.	Wu et al. ₁	Wu et al. ₂	ECD
A	0.974 ± 0.065	0.912 ± 0.081	0.948 ± 0.069	0.733 ± 0.241	0.794 ± 0.261
B	0.955 ± 0.049	0.702 ± 0.109	0.945 ± 0.077	0.845 ± 0.132	0.785 ± 0.184
C	0.555 ± 0.225	0.366 ± 0.229	0.630 ± 0.244	0.650 ± 0.188	0.755 ± 0.096
D	0.908 ± 0.123	0.968 ± 0.042	0.689 ± 0.119	0.574 ± 0.233	0.675 ± 0.300
E	0.946 ± 0.105	0.681 ± 0.158	0.618 ± 0.237	0.665 ± 0.189	0.684 ± 0.171
F	0.941 ± 0.100	0.934 ± 0.065	0.905 ± 0.125	0.696 ± 0.128	0.486 ± 0.269
G	0.689 ± 0.160	0.212 ± 0.068	0.149 ± 0.027	0.244 ± 0.045	0.370 ± 0.025
Overall	0.882 ± 0.173	0.719 ± 0.284	0.725 ± 0.302	0.638 ± 0.244	0.621 ± 0.266

TABLE VIII
COMPARISON OF COMPUTATIONAL TIME (PRIVATE DATASET, SECOND)

Process	Proposed	Hsu et al.	Wu et al. ₁	Wu et al. ₂	ECD
Preprocess	101.57	0.25	-	5.7	-
Similarity ranking	0.39	8301.92	10.50	5.2	0.40
Total	101.96	8302.17	10.50	11.0	0.40

* A hyphen indicates that there is no preprocessing in the method.

patterns (center, donut, edge-local, edge-ring, loc, random, scratch, near-full, and none) with various wafer map shapes. We divided the public dataset into subsets according to wafer map size. This is because the experiments were conducted on wafer maps of the same dimensions to avoid information distortion owing to the normalization process.

Based on the data cleansing criteria for public datasets, we selected 16 datasets with seven categories. The wafer map figures in the first row of Table VI are samples of defect patterns in dataset 1 (set 1); however, each dataset has various different wafer maps. Even though the seven defect pattern categories are shared by all 16 public datasets, the dataset characteristics, such as the size of the wafer map, are different. Table VI shows the descriptions of the 16 public datasets, including the dimensions of the wafer map (size), total number of maps (total), and the number of maps for each type of defect pattern (center-scratch).

C. Baseline Description

To evaluate the performance of the proposed system, we compared it with the following four similarity-searching methods. 1) Hsu et al. [9]: we set the parameters as $m = 0.1$ for the mountain function, and $S_{\text{out}} = 0.1 \times \max_d$ for maximum matching tolerance, which are determined by the AP among the candidate combinations. Here, \max_d is the maximum distance of defective chips from the origin in a dataset so that S_{out} varies depending on the dataset. For comparison with method proposed by Wu et al. [12], we consider two versions with different proportion of the wafer maps retrieved in the first step, and set the weight between similarity scores as $\beta = 0.5$. Note that because the number of wafers in each dataset is different, in the first step we decided to find similar wafer maps that is proportional to the size of the dataset. We denote the proportion as p_1 . 2) Wu et al.₁: we set $p_1 = 1$, which is determined by AP among candidate proportions. The similarity ranking is obtained based only on the normalized correlation coefficient without searching similar wafer maps first. 3) Wu et al.₂: we set $p_1 = 0.5$ for the two-step similarity

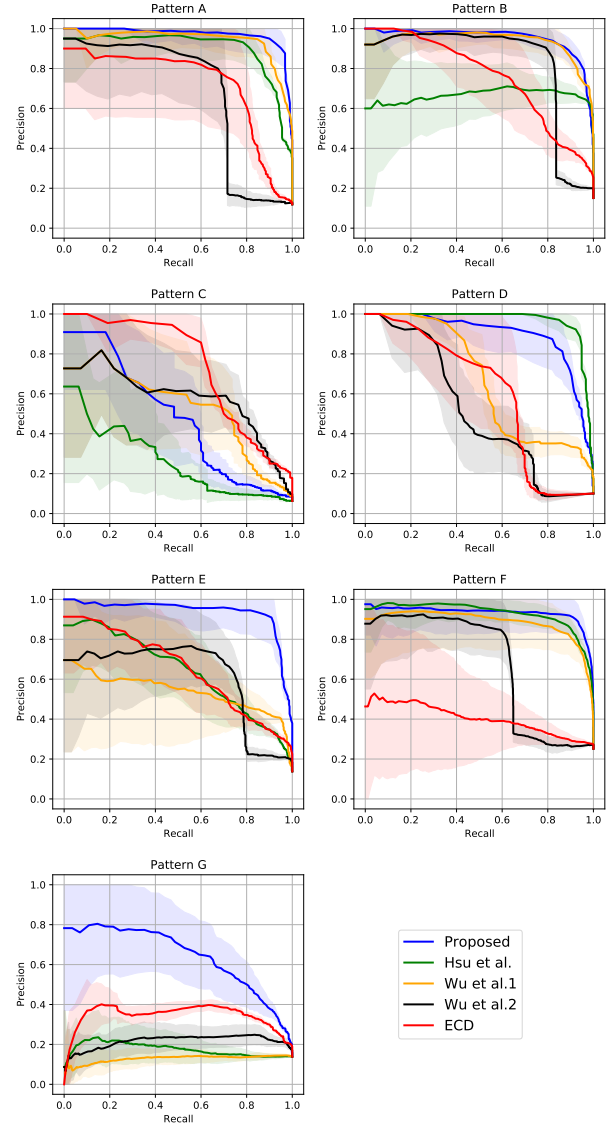


Fig. 5. Precision-recall curves for private dataset.

ranking method. 4) ECD : it provides a rank in ascending order of the Euclidean distance between a query $\mathbf{y}_i^\#$ and the other WBM \mathbf{y}_{-i} , that is, $\{(\mathbf{y}_i^\# - \mathbf{y}_{-i})^T(\mathbf{y}_i^\# - \mathbf{y}_{-i})\}^{0.5}$.

D. Results from Private Dataset

Regarding the private dataset, we used a set of 28 global clusters specified as per the criterion in Subsection III-G. The PR curve and AP were used to evaluate the retrieval performance of the aforementioned similarity ranking methods. As all WBMs were to be queried, each method was iterated 160 times. Table VII shows the mean and standard deviation (sd) of the AP for each defect pattern and for all queries for the five methods (mean ± sd). In general, the proposed method performs better than the other four except for pattern C. This is because the number of WBMs for which pattern C was detected is small; moreover, there is a pattern appearing as a donut, indicating a label issue. The PR curves shown in Fig. 5 for the five methods also graphically demonstrate the benefits

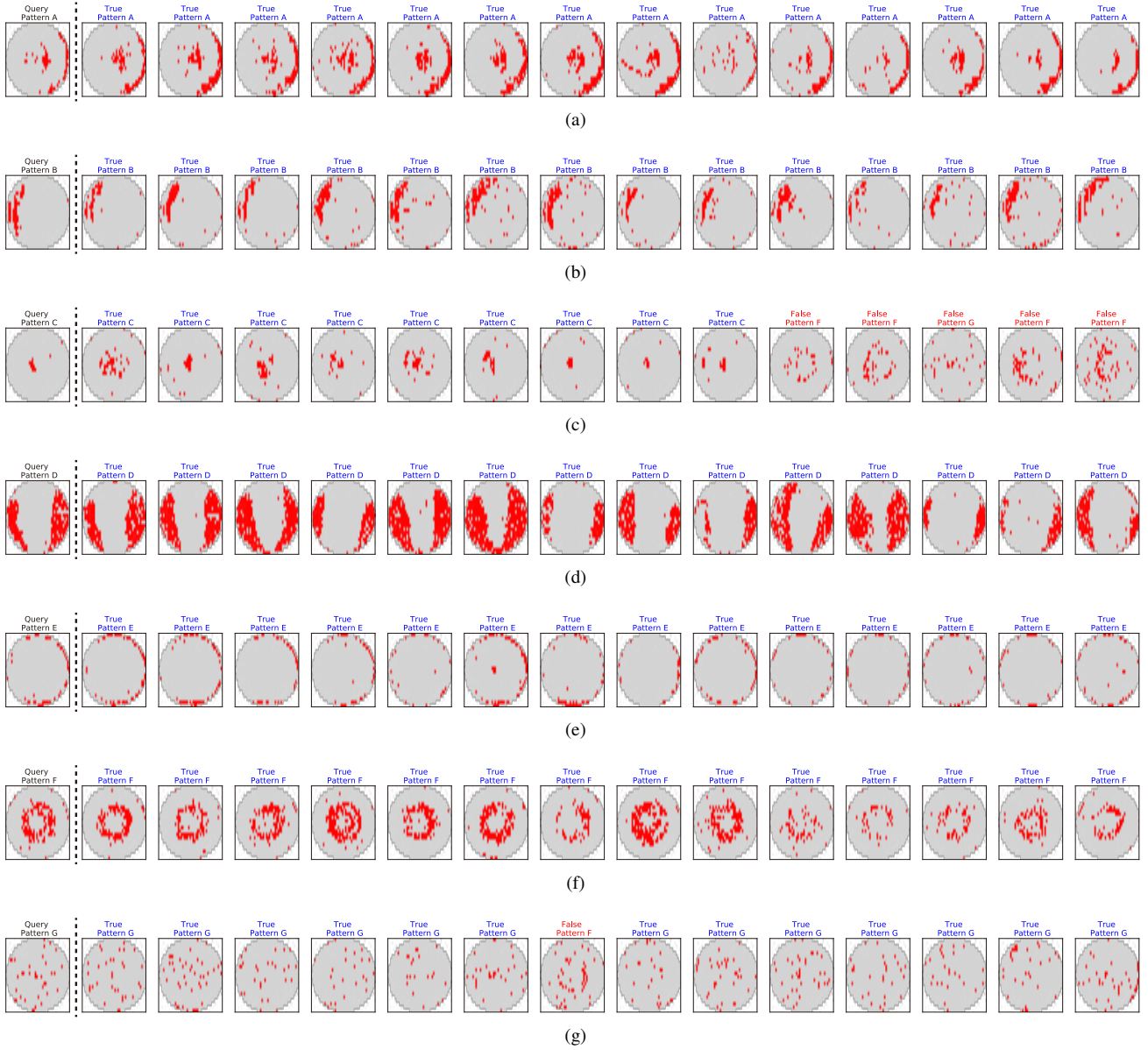


Fig. 6. Similarity ranking result of the proposed method for the private dataset: Patterns (a) A, (b) B, (c) C, (d) D, (e) E, (f) F, and (g) G.

of the proposed method, particularly in similarity search for edge patterns. The figure shows the mean PR curve (solid line) and sd (shaded area) of the iterations for each method.

Table VIII shows the computational time, that is, the time required to obtain features (preprocess) and calculate the similarity ranking (similarity ranking), as well as the total time (total). Even though the proposed method requires more time for preprocessing, the time to calculate the similarity ranking is similar to that of the ECD; as a result, it requires minimal time compared with the method in [9] and [12].

Fig. 6 shows the similarity search results for seven queries with different defect patterns. The first wafer map from the left is the queried map, and the other maps are retrieved from the dataset. It can be seen that the retrieved wafer maps are similar to the queried map for all defect patterns. It should be noted that there are only 11 wafer maps with pattern C, including the query, and Fig. 6(c) shows that nine maps out

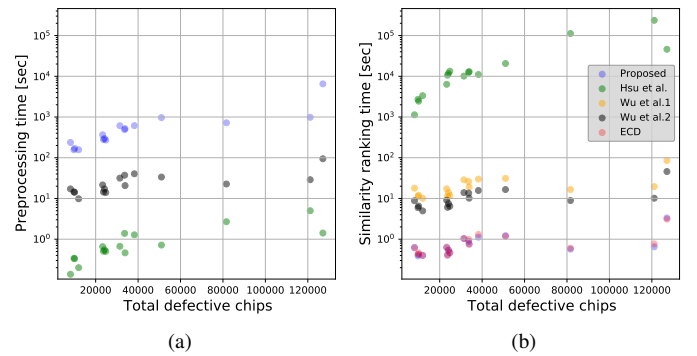


Fig. 7. (a) Scatter plot of preprocessing time vs total defects. (b) Scatter plot of similarity ranking time vs total defects.

of ten were detected and ranked higher.

TABLE IX
COMPARISON OF AP AND COMPUTATIONAL TIME (PUBLIC DATASET)

ID	G	AP					Computational time (s)									
		Proposed	Hsu et al.	Wu et al. ₁	Wu et al. ₂	ECD	Proposed		Hsu et al.		Wu et al. ₁		Wu et al. ₂		ECD	
							Prep	SR	Prep	SR	Prep	SR	Prep	SR	Prep	SR
Set 1	60	0.496 ± 0.241	0.344 ± 0.178	0.463 ± 0.221	0.441 ± 0.210	0.441 ± 0.223	720.0	0.6	2.7	112061.5	-	16.6	22.6	5.1	-	0.6
Set 2	57	0.449 ± 0.196	0.378 ± 0.174	0.413 ± 0.174	0.413 ± 0.119	0.394 ± 0.122	966.3	1.2	0.7	20615.1	-	31.1	33.6	5.2	-	1.2
Set 3	60	0.417 ± 0.142	0.367 ± 0.135	0.402 ± 0.140	0.374 ± 0.117	0.374 ± 0.122	296.5	0.5	0.5	11683.5	-	13.8	17.1	4.6	-	0.5
Set 4	59	0.704 ± 0.281	0.593 ± 0.255	0.646 ± 0.295	0.624 ± 0.280	0.558 ± 0.279	614.2	1.1	1.3	11034.5	-	29.9	40.4	5.5	-	1.3
Set 5	59	0.610 ± 0.262	0.554 ± 0.259	0.584 ± 0.277	0.598 ± 0.292	0.546 ± 0.251	489.3	0.9	1.4	12497.5	-	26.1	37.3	5.5	-	1.0
Set 6	60	0.516 ± 0.204	0.424 ± 0.177	0.498 ± 0.215	0.474 ± 0.165	0.448 ± 0.145	519.6	0.8	0.5	13000.6	-	19.1	20.7	4.5	-	0.7
Set 7	59	0.434 ± 0.252	0.348 ± 0.195	0.407 ± 0.221	0.445 ± 0.187	0.415 ± 0.264	156.9	0.4	0.2	3344.3	-	10.1	9.8	4.1	-	0.4
Set 8	57	0.681 ± 0.319	0.514 ± 0.235	0.662 ± 0.348	0.628 ± 0.333	0.618 ± 0.381	366.0	0.6	0.7	6350.8	-	17.3	21.5	4.7	-	0.6
Set 9	60	0.604 ± 0.273	0.496 ± 0.255	0.593 ± 0.264	0.575 ± 0.250	0.541 ± 0.209	271.5	0.5	0.5	13369.8	-	11.8	14.0	4.5	-	0.5
Set 10	60	0.493 ± 0.202	0.437 ± 0.167	0.464 ± 0.196	0.454 ± 0.151	0.410 ± 0.126	610.7	1.0	0.7	10092.5	-	28.7	31.6	5.3	-	1.0
Set 11	60	0.429 ± 0.206	0.410 ± 0.194	0.394 ± 0.170	0.427 ± 0.185	0.355 ± 0.117	237.5	0.6	0.1	1131.5	-	17.9	17.2	4.4	-	0.6
Set 12	60	0.512 ± 0.245	0.445 ± 0.194	0.444 ± 0.230	0.469 ± 0.221	0.385 ± 0.197	170.3	0.4	0.3	2445.3	-	11.9	14.1	4.6	-	0.5
Set 13	59	0.490 ± 0.206	0.451 ± 0.191	0.444 ± 0.203	0.461 ± 0.185	0.401 ± 0.166	156.7	0.4	0.3	2716.9	-	11.3	14.5	4.6	-	0.4
Set 14	50	0.958 ± 0.158	0.929 ± 0.181	0.950 ± 0.183	0.924 ± 0.200	0.934 ± 0.219	6500.9	3.3	1.4	45947.2	-	85.5	94.2	7.4	-	3.1
Set 15	60	0.546 ± 0.192	0.480 ± 0.173	0.519 ± 0.177	0.465 ± 0.134	0.459 ± 0.130	279.8	0.4	0.5	10646.4	-	11.3	14.4	4.6	-	0.4
Set 16	46	0.609 ± 0.262	0.482 ± 0.232	0.613 ± 0.261	0.582 ± 0.242	0.602 ± 0.233	983.4	0.6	5.0	236187.5	-	19.5	28.9	5.6	-	0.8

* A hyphen indicates that there is no preprocessing in the method.

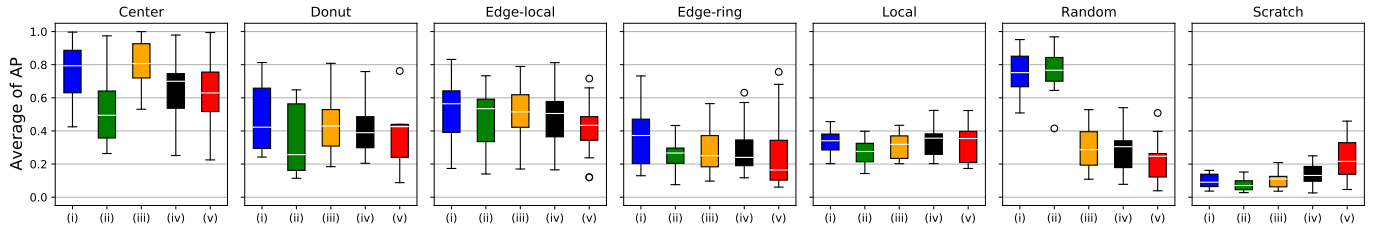


Fig. 8. Boxplots of average AP on public datasets for each defect pattern by five methods: (i) Proposed, (ii) Hsu et al., (iii) Wu et al.₁, (iv) Wu et al.₂, and (v) ECD.

E. Results from Public Dataset

Table IX shows the mean and sd of the AP for 100 selected queries by stratified random sampling with proportional allocation on each public dataset, and presents the computational time for preprocessing (prep) and similarity ranking (SR) in seconds. Moreover, the number of global clusters G for each public dataset is presented. As shown in the experiments with the private dataset, in general, it is also confirmed that the proposed system shows improved performance than the other methods even if the size, the number, and the defect patterns are different. For set 14, the AP is higher than other sets. This is not surprising, though, as 94% of the defect patterns in set 14 belong to the same type (center) as described in Table VI. Note that the results of public datasets except set 14 is much lower than that of private dataset. This is presumably because the differences between the patterns in the private dataset are clear and the corresponding labels are relatively accurate as shown in Table V.

In addition, the proposed method is efficient even on a high-dimensional dataset, as seen in Fig.7 and Table IX. The figure shows the association between the total number of defective chips in a dataset and the time to compute the preprocessing (a), and the similarity ranking time (b) with base-10 log scale for the y-axis. In the proposed system, the preprocess is time-consuming for a large dataset because of using the HCM; however, once the global clusters are obtained from a dataset, the computational time for similarity ranking is similar to that of the ECD.

Fig.8 shows the box plots of the average AP on 16 public datasets for each defect pattern. In the center and local patterns, the Wu et al.₁ method shows comparable performance, but similarity searching is not efficient as shown in Table IX. The time can be shortened by two-step method, Wu et al.₂, but it is confirmed that the performance is worse and the method is still time-consuming.

V. DISCUSSION

For scratch type pattern, the proposed method is slightly inferior performance than the ECD. To further improve the performance, we could incorporate the similarity ranking method based on the ECD in the proposed system. For example, a two-step similarity ranking method as in Wu *et al.* can be adopted: first, search the top- N similar wafer maps based on the ECD of the WBMs; then, rank the top- N candidates based on the JSD of the weight vectors. This method is also one way to reduce the computational time for massive dataset. Alternatively, one may consider a two-way similarity ranking system where ECD method is employed for scratch type while using our proposed system for other types in practice. However, it is the intent of this paper to propose a one-step comprehensive approach for ranking similarity on a wafer map. We believe that incorporating ECD in the current similarity system is beyond the scope of this paper and we leave this as a future study for practitioners.

VI. CONCLUSION

We proposed an automated similarity ranking system for WBMs. It searches and ranks wafer maps according to the similarity and location of their defect patterns. This is because the system is not rotation-invariant and can therefore identify the exact location of these patterns; thus it can provide crucial information for root cause analysis.

We identified global clusters as systematic failure patterns across wafer maps through nonparametric Bayesian clustering based on the DPGMM, and the HCM based on the SKLD. The proposed features are provided by the weight vectors with respect to the global clusters. Using the estimated bivariate Gaussian distribution of the global clusters, each WBM can be represented by a finite Gaussian mixture model with its weight vector. The proposed features are effective because they require minimal computation and storage; moreover they have high discriminatory power in ranking similar failure patterns, and thus they are suitable for large-scale analysis of wafer maps. The efficiency and performance of the proposed system were experimentally confirmed using a private dataset and a subset of WM-811K.

ACKNOWLEDGEMENTS

This research was supported by SK Hynix AICC.

REFERENCES

- [1] S.-C. Hsu and C.-F. Chien, "Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing," *International Journal of Production Economics*, vol. 107, no. 1, pp. 88–103, 2007.
- [2] F.-L. Chen and S.-F. Liu, "A neural-network approach to recognize defect spatial pattern in semiconductor fabrication," *IEEE Trans. Semicond. Manuf.*, vol. 13, no. 3, pp. 366–373, 2000.
- [3] K. S.-M. Li, P. Y.-Y. Liao, K. C.-C. Cheng, L. L.-Y. Chen, S.-J. Wang, A. Y.-A. Huang, L. Chou, G. C.-H. Han, J. E. Chen, H.-C. Liang *et al.*, "Hidden wafer scratch defects projection for diagnosis and quality enhancement," *IEEE Transactions on Semiconductor Manufacturing*, vol. 34, no. 1, pp. 9–16, 2021.
- [4] J. Hwang and H. Kim, "Variational deep clustering of wafer map patterns," *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 3, pp. 466–475, 2020.
- [5] S. Park, J. Jang, and C. O. Kim, "Discriminative feature learning and cluster-based defect label reconstruction for reducing uncertainty in wafer bin map labels," *Journal of Intelligent Manufacturing*, vol. 32, no. 1, pp. 251–263, 2021.
- [6] W. Zhang, X. Li, S. Saxena, A. Strojwas, and R. Rutenbar, "Automatic clustering of wafer spatial signatures," in *Proceedings of the 50th Annual Design Automation Conference*, 2013, pp. 1–6.
- [7] P. Tulala, H. Mahyar, E. Ghalebi, and R. Grosu, "Unsupervised wafermap patterns clustering via variational autoencoders," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.
- [8] T. Santos, S. Schrunner, B. C. Geiger, O. Pfeiler, A. Zernig, A. Kaestner, and R. Kern, "Feature extraction from analog wafermaps: A comparison of classical image processing and a deep generative model," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 2, pp. 190–198, 2019.
- [9] C.-Y. Hsu, W.-J. Chen, and J.-C. Chien, "Similarity matching of wafer bin maps for manufacturing intelligence to empower industry 3.5 for semiconductor manufacturing," *Computers & Industrial Engineering*, vol. 142, p. 106358, 2020.
- [10] M. Fan, Q. Wang, and B. van der Waal, "Wafer defect patterns recognition based on optics and multi-label classification," in *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. IEEE, 2016, pp. 912–915.
- [11] J. Yu and X. Lu, "Wafer map defect detection and recognition using joint local and nonlocal linear discriminant analysis," *IEEE Trans. Semicond. Manuf.*, vol. 29, no. 1, pp. 33–43, 2016.
- [12] M.-J. Wu, J.-S. R. Jang, and J.-L. Chen, "Wafer map failure pattern recognition and similarity ranking for large-scale data sets," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 1, pp. 1–12, 2015.
- [13] T. Nakazawa and D. V. Kulkarni, "Wafer map defect pattern classification and image retrieval using convolutional neural network," *IEEE Trans. Semicond. Manuf.*, vol. 31, no. 2, pp. 309–314, 2018.
- [14] N. Yu, Q. Xu, and H. Wang, "Wafer defect pattern recognition and analysis based on convolutional neural network," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 4, pp. 566–573, 2019.
- [15] R. M. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.
- [16] J. Yang, E. Grunsky, and Q. Cheng, "A novel hierarchical clustering analysis method based on kullback–leibler divergence and application on dalaimiao geochemical exploration data," *Computers & Geosciences*, vol. 123, pp. 10–19, 2019.
- [17] C.-H. Wang, S.-J. Wang, and W.-D. Lee, "Automatic identification of spatial defect patterns for semiconductor manufacturing," *International journal of production research*, vol. 44, no. 23, pp. 5169–5185, 2006.
- [18] T. Yuan, W. Kuo, and S. J. Bae, "Detection of spatial defect patterns generated in semiconductor fabrication processes," *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 3, pp. 392–403, 2011.
- [19] J. Kim, Y. Lee, and H. Kim, "Detection and clustering of mixed-type defect patterns in wafer bin maps," *Iise Transactions*, vol. 50, no. 2, pp. 99–111, 2018.
- [20] C. H. Jin, H. J. Na, M. Piao, G. Pok, and K. H. Ryu, "A novel dbscan-based defect pattern detection and classification framework for wafer bin map," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 3, pp. 286–292, 2019.
- [21] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th Ed. Wiley, 2011.
- [22] C.-J. Huang, C.-F. Wu, and C.-C. Wang, "Image processing techniques for wafer defect cluster identification," *IEEE Design & Test of Computers*, vol. 19, no. 2, pp. 44–48, 2002.
- [23] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica sinica*, pp. 639–650, 1994.
- [24] W. J. Ewens, "Population genetics theory-the past and the future," in *Mathematical and statistical developments of evolutionary theory*. Springer, 1990, pp. 177–227.
- [25] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [26] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997.
- [27] D. Müllner *et al.*, "fastcluster: Fast hierarchical, agglomerative clustering routines for r and python," *Journal of Statistical Software*, vol. 53, no. 9, pp. 1–18, 2013.
- [28] D. M. Blei, M. I. Jordan *et al.*, "Variational inference for dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.
- [29] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [30] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.
- [31] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.
- [32] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 821–830.
- [33] F. Ahmed and A. Courville, "Detecting semantic anomalies," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3154–3162.