A Pattern Mining Framework for Inter-Wafer Abnormality Analysis

Nik Sumikawa, Li-C. Wang
Department of ECE, UC-Santa Barbara

Magdy S. Abadir
Freescale Semiconductor, Inc.

***Abstract*—This work presents three pattern mining methodologies for inter-wafer abnormality analysis. Given a large population of wafers, the first methodology identifies wafers with abnormal patterns based on a test or a group of tests. Given a wafer of interest, the second methodology searches for a test perspective that reveals the abnormality of the wafer. Given a particular pattern of interest, the third methodology implements a monitor to detect wafers containing similar patterns. This paper discusses key elements for implementing each of the methodologies and demonstrates their usefulness based on experiments applied to a high-quality SoC product line.**

1. Introduction

Analysis of wafer-level abnormalities has been a common practice in the industry for many years. In visual defect metrology, for example, abnormal wafer maps are detected, categorized and diagnosed to uncover systematic process issues. The work in [1] defines three types of abnormalities: (1) failing frequency based statistics indicating abnormal yield fluctuations, (2) failing location based statistics revealing abnormal concentration of failures (clustering), (3) spatial failing patterns (e.g. line, ring, arc, etc.) that can be correlated to some special causes such as scratches from material handling, non-uniformities in film thickness, edge-die effects, and so on.

A traditional wafer-level abnormality analysis system is yield driven and typically, performs the following two steps [1][2]: (1) identifying an abnormal wafer that potentially reflects a yield issue, and (2) recognizing a failing pattern on the wafer as belonging to a known issue.

The identification step is usually done by monitoring some failing frequency statistics based on counting the number of failing dies [1]. The pattern recognition step is more complicated, which can be implemented with various statistical techniques. For example, the work in [2] applied an imaging denoising technique [3] to remove statistically random failing dies. Then, a clustering method was applied to group remaining failing dies [4], followed by a classification method to put each group into a known category of issue. Other statistical learning methods for the pattern recognition step were also proposed [5]-[8]. In general, the pattern recognition step solves the following problem: Given a wafer w and a set C of known problematic pattern categories $\{c_1, \ldots, c_n\}$, decide if there is a failing pattern on w that closely resembles a $c_i \in C$.

This work considers a rather different problem. We do not assume that C is known in advance. Instead, we assume that a large population of wafers are given and the task includes finding, among these wafers, what patterns can be considered as *novel*. In this analysis, we equate *abnormality* to *novelty* and treat the discovery of abnormal patterns as part of the problem. With this assumption, a wafer pattern mining framework is proposed to support the discovery, analysis, and recognition of abnormal patterns. This framework consists of three methodologies:

1) **Abnormality Detection:** Identify wafers with patterns that are novel as compared to other wafers.
2) **Perspective Search:** Given a wafer of interest, identify a test perspective (i.e. a test or a group of tests) that exposes a pattern on the wafer, where the pattern is novel as compared to other wafers.
3) **Similarity Search:** Given a known abnormal pattern, detect wafers containing similar patterns.

We call our approach *inter-wafer* abnormality analysis because the abnormality of a wafer is measured relatively to others in a given population. This is in contrast to visual defect metrology where intra-wafer abnormalities were recognized according to known categories of problematic patterns.

It is important to note that in a test application context, the abnormality of a wafer can depend on the *test perspective*, i.e. the test or subset of tests used to define the wafer pattern. For example, given three tests $\{t_1, t_2, t_3\}$, a wafer may have an abnormal failing pattern based on t_1 individually, but not based on $\{t_1, t_2, t_3\}$ collectively. Hence, abnormalities are test perspective dependent. In Abnormality Detection, the test perspective is specified by the user.

In Perspective Search, a wafer is deemed abnormal by a user (without observing an abnormal failing pattern). The goal is to find a test perspective that results in a wafer pattern that is abnormal as compared to other wafers. In other words, we are interested in finding a test perspective that can visually expose the abnormality of a given wafer.

In Similarity Search, a pattern is given, and a monitor is built to detect wafers containing *similar* patterns. A key consideration for the search is how *similarity* should be defined. For example, similarity can be rotation invariant such that a pattern rotated by a certain degree is recognized as the same pattern.

This work discusses several key elements for implementing the three methodologies. These key elements include:

- **Encoding:** a 1-to-1 and onto mapping that converts a wafer map into a vector of values for further processing.
- **Transformation:** an onto mapping that converts an encoded vector into a *feature vector*.
- **Kernel:** a similarity measure function that takes two feature vectors as inputs and outputs a similarity value.
- **Learning algorithm:** a novelty detection algorithm that constructs a model to identify abnormal wafers.

This work is supported in part by SRC 2010-TJ-2093.

978-1-4799-0859-2/13/$31.00

This framework is evaluated using data from a high-quality SoC product line designed for the automotive market. Wafer test results from both parametric and non-parametric tests were analyzed. The following results were observed: Abnormality Detection was able to identify questionable wafers overlooked by the existing methods. Perspective Search was able to identify systematic wafer patterns correlated to the locations of dies that later were found to be customer returns. Similarity Search was able to effectively detect wafers containing patterns similar to a given pattern such as a line, arc, ring, cluster, etc.

The remainder of the paper is organized as the following. Section 2 discusses a motivation for the work. Section 3 explains the key elements to be considered in pattern mining. Section 4 describes the three methodologies. Sections 5, 6 and 7 discuss results based on the Abnormality Detection, Perspective Search, and Similarity Search, respectively. Section 8 concludes the paper.

2. A Motivation Example

Consider a scenario where we are given a wafer that is assumed to be abnormal, but we do not know what the abnormality is. For example, the wafer may contain a customer return. Our goal is to uncover the abnormality of the wafer so that the abnormality can be used as a way to detect future abnormal wafers such as those likely to contain a similar customer return.

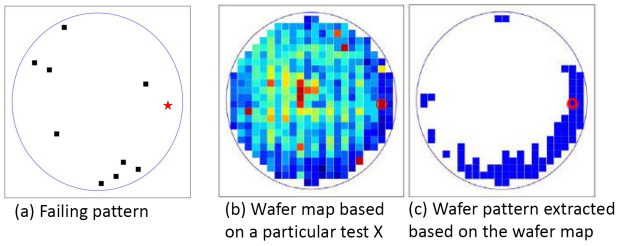


Fig. 1. **The abnormal pattern of a customer return wafer can only be observed based on a particular test X**

Figure 1-(a) shows the failing pattern of a wafer containing a known customer return. The parts failing wafer sort tests are shown as black squares and the customer return die is a red star. In this example, there is not much one can say about the abnormality of the wafer. There are too few failing dies to form a noticeable pattern and the failing parts seem to be randomly located. With an existing yield-driven abnormality recognition approach, the engineer would report that the customer return wafer was normal and there was nothing special about it.

Figure 1-(b) shows the wafer map based on the measured value of a parametric test X. Based on this map, we first smoothened the image by taking the location average [9], i.e. adjusting the value of each die by taking the average across its neighboring dies. Then, we set a threshold to classify the dies into those above and those below the threshold. This forms the wafer pattern shown in Figure 1-(c).

This wafer pattern can be considered as novel because it occurrs infrequently on other wafers. Because of its novelty, this wafer pattern can be used as a special property to describe the customer return wafer.

To determine the novelty of the wafer pattern in Figure 1-(c), we need a methodology that can perform novelty detection on a collection of wafer patterns. In novelty detection, the wafer patterns are ranked based on their novelty. If a target wafer pattern is novel, the methodology will assign it a high novelty rank. This indicates that the pattern is quite different from the majority of the wafer patterns. This motivated the development of the Abnormality Detection methodology in this work. Hence, the wafer pattern in Figure 1-(c) was considered novel (highly ranked) after applying Abnormality Detection on a large number of wafers.

It is important to note that the wafer pattern in Figure 1-(c) is based on a particular test X and a given threshold. Given a large number of tests, finding the test perspective to expose an abnormal pattern demands another methodology. This motivated the development of the Perspective Search methodology. The result shown in Figure 1 was obtained with the Perspective Search to find the test X.

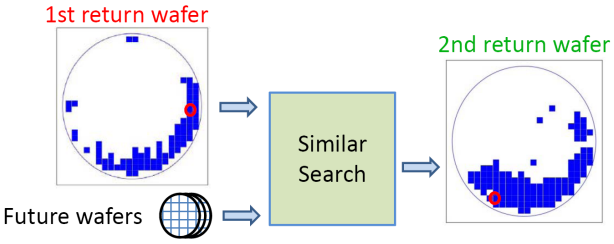
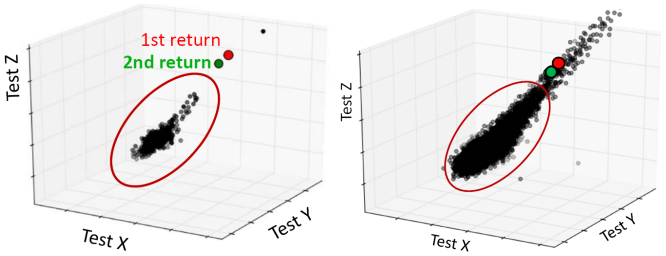


Fig. 2. **Abnormal pattern learned from one customer return wafer is used to detect another customer return wafer**

Suppose the wafer pattern in Figure 1-(c) was accepted as a property to describe the specialty of the customer return wafer. In application, we would like to monitor future wafers and recognize any wafer with a similar pattern. This motivated the development of the Similarity Search methodology.

Figure 2 shows the result of searching for similar wafer patterns based on the identified wafer pattern that describe the customer return wafer. The similarity search uncovered a second wafer (manufactured at a later time) with similar pattern. The second wafer also contained a customer return.



(a) Based on two return wafers (b) Based on two lots

Fig. 3. **Multivariate outlier model for the customer returns**

To illustrate how an abnormal wafer pattern can be used for screening potential customer returns, take the work in [10] as an example. This work proposed test model building methods for screening potential customer returns. Figure 3 shows a multivariate test model M built using the *reactive method* [10] for the first customer return. The model is based on the test X and two additional correlated tests Y and Z.

Figure 3-(a) shows the distribution of the dies from the two customer return wafers. The two customer return dies are clear outliers with respect to other dies. Hence, model M could screen out the two return dies with almost no overkills.

Figure 3-(b) shows the distribution of all the dies from the two lots containing the two customer return wafers. We see that the two customer return dies are marginal and are no longer clear outliers. Hence, an outlier model M would have to kill many dies in order to screen out the two returned dies. This shows that the multivariate outlier model M by itself is not a feasible solution for screening.

The abnormal wafer pattern in Figure 1-(c) provides a means to implement a hierarchical screen. With the Similarity Search, the hierarchical screen can be the following: If a wafer contains a wafer pattern similar to Figure 1-(c), then apply M on the wafer. As a result, M is only applied to few selected wafers and the yield impact is dramatically reduced.

With the results shown in Figure 2 and Figure 3-(a), we see that after learning from the first customer return, the wafer containing the second customer return could have been identified and the return could be screened as an outlier with minimal yield impact. In this application, the inter-wafer abnormality framework was indispensable for finding the test perspective X and the abnormal wafer pattern that led to the development of a feasible hierarchical screen for the returns.

3. Pattern Mining - Basic Concepts

As explained before, the abnormality of a wafer depends on the test perspective, i.e. the subset of tests and the corresponding test limits used to classify the "passing" and "failing" dies that define the wafer pattern. The abnormality also depends on three elements: the encoding, transformation, and kernel. Collectively, we call them the *algorithmic perspective*.

3.1. Finding abnormality with novelty detection

An abnormal wafer should have the property that there exists a pattern shown on the wafer that is "different" from most of the patterns on other wafers. This is a problem that can be solved by novelty detection [11], where the objective is to identify novel samples among a set of samples.

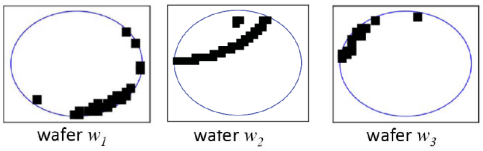


Fig. 4. **Is w_1 more similar to w_2 or to w_3?**

In order to measure "difference," one needs a kernel function $k()$ [12]. Given two wafers w_1 and w_2, $k(w_1, w_2)$ calculates a measure of *similarity* between the two wafers. The *kernel* function dictates how similarity should be measured, which in turn affects how novelty is defined. For example, Figure 4 shows three wafers. One may consider w_1 and w_2 more similar by reasoning that the arc pattern on w_2 is a shifted version of the arc pattern on w_1. Alternatively, one may argue that w_1 and w_3 are more similar because w_3 is similar to a rotated version of w_1. If w_1 and w_3 are more similar, then w_2 is the most novel wafer among the three. We see that the definition of the similarity measure is crucial in determining how novelty is perceived and consequently, how abnormality is defined.

3.2. The algorithmic perspective

We define *algorithmic perspective* P as 3-tuple (E, T, K):

1) E is the scheme to encode a wafer map
2) T is the transformation that converts a wafer pattern into a feature vector
3) K is the kernel function

3.2.1. **Encoding:** In this work, we only consider the wafer pattern represented as a black-and-white bit map. As explained in Section 2, a parametric wafer map can be converted into a binary wafer pattern depending on how we classify the dies into "passing" and "failing." Given a binary bit map, two possible encodings are: 0/1 encoding and $+1/-1$ encoding.

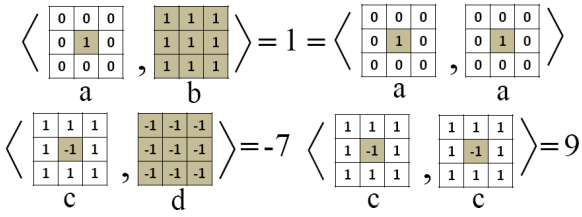


Fig. 5. **Illustrating the impact of encoding**

Figure 5 illustrates the importance of encoding. On the top, two bit maps a and b are shown with a 0/1 encoding ("0"/"1" represents a passing/failing die). Assume no transformation is used. Further assume the kernel is the *dot product* ($\langle a, b\rangle = \sum_i (a_i b_i)$). We see that $\langle a, b\rangle = 0{+}0{+}0{+}0{+}1{+}0{+}0{+}0{+}0 = 1$ (9 bits). We see that $\langle a, a\rangle = 1$ as well. In other words, b is considered equivalent to a. On the bottom the same two bit maps are encoded as c and d based on $+1/-1$ encoding. In this case, we see that $\langle c, d\rangle = -7$ which does not equal $\langle c, c\rangle$=9. The value "9" indicates identical patterns, i.e. all bits are the same pairwise. We see that with $+1/-1$ encoding, bit map d is considered very different from bit map c.

This example illustrates that if one intends to consider *containment* to be the same as equivalence, then 0/1 encoding should be used. Otherwise, $+1/-1$ should be used.

3.2.2. **Transformation:** Transformation takes a bit map and converts it into a feature vector $v = (v_0, \ldots, v_{n-1})$ based on n features $f_0, \ldots, f_{n-1}$. Transformation may not be one-to-one, i.e. different bit maps can be converted into the same feature vector v. Transformation, which can also be seen as the step of *feature generation*, is at the core of many image processing techniques [13]. For example, to make the representation *rotation invariant*, one can apply a histogram transform that converts an image into a histogram with a fixed number of features [14].

Many popular transforms, such as wavelet transform, are developed for handling gray-scale and color images (or video frames) [13]. A wafer map in our case is a black-and-white bit map. Hence, simpler transforms are considered in this work.

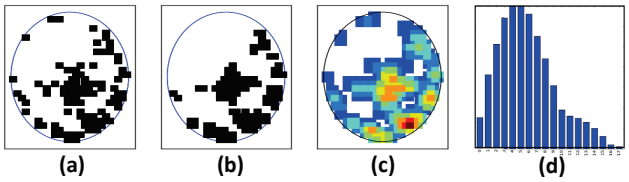


Fig. 6. **(a) Original wafer map (b) 2×2-raster smoothening (c) 3×3 local binary pattern transform (d) distance transform**

Figure 6 illustrates three types of transforms implemented in this work. Plot (a) shows the original wafer map. Plots (b)-(d) shows the results of applying the respective transform.

- **Raster Smoothening:** A raster scan is performed on the wafer image based on a window of size $i \times i$. For example, assume $i = 2$. In each step there are four dies to be considered. If the majority of dies are with label -1, a -1 is associated with the step. Otherwise, a $+1$ is associated with the step. There are N steps, where each step associates a value with the die in the top left corner of the window. As shown in plot (b) of Figure 6, this transform removes isolated failing dies, i.e. "smoothening" the image by removing noisy "pixels."
- **Local Binary Pattern (LBP):** In LBP [15], each die is no longer labeled with pass/fail. Instead, each die is labeled with the number of fails in a local region. For example, each die is assigned the number of failing dies in the neighboring eight dies and itself. As a result, the wafer is no longer represented as a bit map. Instead it becomes a colored map where a color is assigned to an integer value. This is shown in plot (c).
- **Distance Transform:** Distance transform [16] converts a bit map into a gray-scale image by replacing each white pixel with the distance to the nearest black pixel. Suppose the maximum distance from a white pixel to a black pixel on an image is n. Essentially, distance transform maps an image into n features $f_0, \ldots, f_{n-1}$, where feature f_i represents the number of pixels with distance $= i$. This results in a histogram. For example, if a black pixel represents a passing die (or vice versa) and there are k passing (failing) die, we have $f_0 = k$. Plot (d) shows the result of distance histogram of the wafer map in plot (a).

To gain the intuition behind the three transforms, consider matching a pattern appearing on two given wafers. Raster smoothening filters out the noise and makes the pattern on a wafer more apparent for comparison. This improves the precision of the pattern matching. On the other hand, LBP blurs the shape of a major pattern with a color coding. This provides more degrees of freedom for pattern matching.

Distance transform is rotation invariant. This means that if one wafer is an exact rotation of the other, the two resulting histograms are the same. However the location information is lost in the distance transform. As a result, wafers with different patterns may result in the same histogram. With distance transform, the probability of "aliasing" is much higher than raster smoothening (note that LBP transform is 1-to-1 and onto). Hence, the information lost in distance transform is the highest among the three. Section 5.6 compares the performance of the three transforms in more detail.

3.2.3. **Kernel:** Given two feature vectors v and u, a kernel $k(v, u)$ measures the similarity between the two vectors. Common kernels include Gaussian ($k = e^{-g\sum_{i=0}^{n-1}(v_i-u_i)^2}$), dot-product ($k = \sum_{i=0}^{n-1}(v_i u_i)$), polynomial ($k = (\sum_{i=0}^{n-1}(v_i u_i) + R)^d$ for some R and d), and intersection kernel ($k = \sum_{i=0}^{n-1}\min(v_i, u_i)$) [12]. Different kernels capture the notion of similarity in different ways. For example, the Gaussian kernel is based on the "distance" between two vectors $\sum(v_i - u_i)^2$ while the dot-product measures the angle when $||v|| = ||u|| = 1$, i.e. $\cos(u, v) = (\sum_{i=0}^{n-1}(v_i u_i))/(||u||||v||)$. They behave differently in an outlier analysis context [17]. Construction and/or selection of kernels is an important aspect when applying the kernel-based learning algorithms [18].

In this work, we discovered that using the Gaussian kernel gives the most reliable results. The intersection kernel is only useful with the distance transform. Dot-product and polynomial kernels can deliver similar results as Gaussian, but they are more sensitive to the encoding and the transform.

4. Pattern Mining Methodologies

4.1. Abnormality Detection Methodology

The Abnormality Detection methodology solves a novelty detection problem based on N wafers $w_1, \ldots, w_N$. Figure 7 illustrates the methodology. The user supplies one or more test perspectives, and gives a number ν representing the upper bound on the percentage of abnormal wafers to be found. The Abnormality Detection ranks the wafers based on their novelty and classifies up to $\nu\%$ of the wafers as novel. As discussed earlier, novelty detection depends on the test perspective as well as the algorithmic perspective.

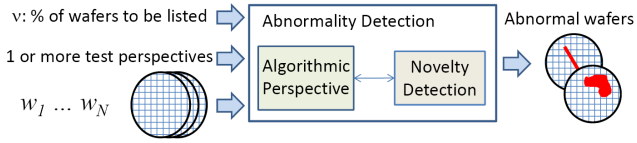


Fig. 7. **Abnormality detection methodology**

4.1.1. **Novelty detection algorithm:** As shown in Figure 7, the novelty detection algorithm is independent of the algorithmic perspective. Once the algorithmic perspective is fixed, one can employ a variety of algorithms to explore novelty. In this work, we adopt the support vector (SV) method for novelty detection [11]. The SV method is convenient from a user perspective because it allows an input specifying an *upper bound* ν on the % of wafers to be classified as novel.

To find novel samples, the SV method solves a quadratic optimization problem by finding a maximum margin hyperplane to separate most of the sample points from the origin in the kernel-induced feature space [11]. Figure 8 illustrates the intuition underlying this idea.

In the input space, suppose there are 11 samples $w_1, \ldots, w_{11}$ located in a two-dimensional space according to a probability distribution $\mathcal{D}$. Suppose we draw two random samples s_1, s_2 from $\mathcal{D}$. The samples s_1, s_2 are likely to be

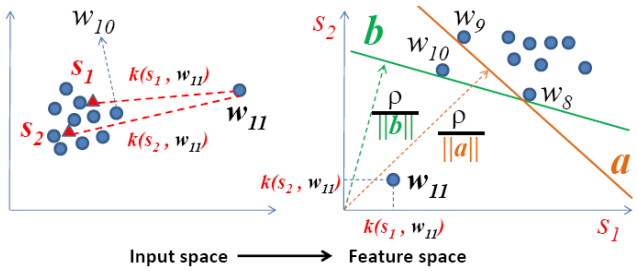


Fig. 8. **Support vector method for novelty detection**

located in the high density region as shown in the figure. For each sample w_i, we then map it into a kernel-induced feature space with the coordinate $(k(s_1, w_i), k(s_2, w_i))$. The figure shows how w_{11} is mapped into this feature space.

Notice that w_{11} is the farthest from s_1, s_2 in the input space. Hence, $k(s_1, w_{11})$ and $k(s_2, w_{11})$ are the smallest, indicating that w_{11} is most dissimilar to s_1, s_2 among all samples. Hence, in the feature space, w_{11} is the closest to the origin $(0, 0)$.

To identify novel samples in the feature space, one can now try to find a maximum-margin hyperplane (a line in the two-dimensional space) that separates most of the points from the origin. In the figure, two hyperplanes are shown, characterized as $\mathbf{a} = (a_1, a_2)$ and $\mathbf{b} = (b_1, b_2)$. The margins, measured as the distance from the hyperplane to the origin, are $\frac{\rho}{||\mathbf{a}||}$ and $\frac{\rho}{||\mathbf{b}||}$ for $\mathbf{a}$ and $\mathbf{b}$, respectively. We see that to maximize the margin of a hyperplane $\mathbf{h} = (h_1, h_2)$, we need to minimize $||\mathbf{h}||$. However, this cannot be done without constraints.

In the method, the constraint is given by the upper bound on the number of novel samples classified by the hyperplane. For example, suppose the upper bound is 2. Then, both $\mathbf{a}$ and $\mathbf{b}$ are possible solutions. There is a tradeoff between the two hyperplanes. The hyperplane $\mathbf{a}$ has a larger margin but classifies w_{11} and w_{10} as novel samples, which may not be desirable. The hyperplane $\mathbf{b}$ has a smaller margin and classifies only w_{11} as the novel sample which may be more desirable. This shows that strictly enforcing maximum margin may not always lead to the desired outcome. To relax the maximum margin requirement, a slack variable η_i can be introduced for each novel sample classified by a hyperplane. The η_i represents the distance from the novel sample to the hyperplane. For a non-novel sample, we have $\eta_i = 0$. The idea is to minimize $||\mathbf{h}|| + \sum \eta_i$ instead of minimizing $||\mathbf{h}||$ as stated above. This is called a soft-margin SV method [11].

Figure 8 illustrates the SV method with two dimensions. Suppose the feature space is n-dimensional based on n samples $s_1, \ldots, s_n$ where $n \to \infty$. We denote the resulting feature vector of w_i as $\phi(w_i)$. Let $\mathbf{h} = (h_1, \ldots, h_n)$ denote a hyperplane in the n-dimensional kernel-induced feature space. Note that the dot product $\langle \mathbf{h}, \phi(w_i) \rangle$ tells which side w_i is located with respect to $\mathbf{h}$. If $\langle \mathbf{h}, \phi(w_i) \rangle \geq \rho$ (same ρ shown in the figure), then w_i is on the side of the hyperplane not containing the origin (non-novel side). Otherwise, it is on the same side with the origin (novel side).

Formally, given m samples $w_1, \ldots, w_m$ and the upper bound % ν the SV method solves the following quadratic optimization problem:

$$\min \frac{1}{2}||\mathbf{h}||^2 + \frac{1}{\nu m}\sum_i \eta_i - \rho \quad (1)$$

$$\text{subject to } \langle \mathbf{h}, \phi(w_i) \rangle \geq \rho - \eta_i, \eta_i \geq 0 \quad (2)$$

The standard way is by using the Lagrangian method and finding the dual optimization problem. The dual is:

$$\min \frac{1}{2}\sum_{ij} \alpha_i \alpha_j k(w_i, w_j) \quad (3)$$

$$\text{subject to } 0 \leq \alpha_i \leq \frac{1}{\nu m}, \sum_i \alpha_i = 1 \quad (4)$$

The kernel $k()$ corresponds to the feature mapping $\phi()$ noted above. Each α_i is associated with a sample w_i. If $\alpha_i > 0$, w_i is called a support vector. Conceptually, support vectors define the hyperplane. Pictorially in Figure 8, w_9 and w_8 are support vectors for $\mathbf{a}$ and w_{10} and w_8 are support vectors for $\mathbf{b}$.

Once the optimal α's are found, the decision function for novelty detection is $f(w) = \text{sign}((\sum_i \alpha_i k(w_i, w)) - \rho)$. Hence, a sample w is novel if and only if $f(w) < 0$. Because $f(w)$ measures the distance to the hyperplane, it can also be used as a measure for the degree of novelty: the more negative the number is, the more novel the sample is with respect to the others. Hence, $f(w)$ can also be used to rank novel samples according to their degrees of novelty.

4.1.2. **Finding novelty across multiple perspectives:** In Figure 7, if multiple test perspectives are entered, it will raise the question how to compare wafer failing patterns from two different test perspectives. Note that for a stop-on-fail data, there is no overlap between the wafer failing patterns on the same wafer based on two non-overlapping test perspectives.

Suppose two non-overlapping patterns have the same shape and size, we need a transform that will treat them as the same pattern. In other words, the transform needs to be insensitive to the location change and rotation of a pattern. Distance transform therefore becomes useful in this case. Later in the experimental section, we will demonstrate that distance transform is an effective tool for comparing test perspectives and evaluating their relative importance. This is useful when one has no idea which test perspective(s) to begin with.

4.2. Perspective Search Methodology

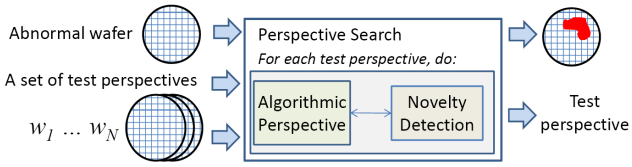


Fig. 9. **Perspective Search methodology**

Figure 9 illustrates the Perspective Search methodology. A wafer is given that is already classified as abnormal. This abnormality is decided not based on observation of an abnormal pattern. Instead, the wafer contains a different type of anomaly. For example, the wafer may be classified as abnormal because it contains a customer return. A set of test perspectives are given for the search. The objective is to determine if there

exists a test perspective that exposes a failing pattern on the given wafer that is seen as novel in comparison to other wafers in the same perspective. If such a perspective can be found, the perspective is reported and the corresponding failing pattern is displayed for visual inspection.

Given k perspectives, the Perspective Search essentially runs the novelty detection k times and identifies in which perspective the given wafer is most novel. This can be slow for a large number of N wafers. To speed up the search, for each perspective, we randomly sample i wafers (say $i = 20$) from the wafer population. The perspective becomes "valid" if the abnormal wafer is ranked as the most novel wafer when compared to the i wafers using the novelty detection algorithm. The process can be repeated a number of times so that if a perspective becomes "valid" in any of the runs, it is valid. At the end, all valid perspectives are verified again with the large population to determine which perspective results in the highest novelty rank for the given abnormal wafer.

4.3. Similarity Search Methodology

Figure 10 illustrates the Similarity Search methodology. The goal of similarity search is to find all wafers that contain a pattern similar to a target wafer pattern.

Let p denote the target wafer pattern to be searched for. Let $k(p, w_i)$ denote the similarity measure between p and wafer w_i. The key in similarity search is to ensure *invariance*. This invariance can be with respect to rotation, horizontal flipping, vertical flipping, etc.

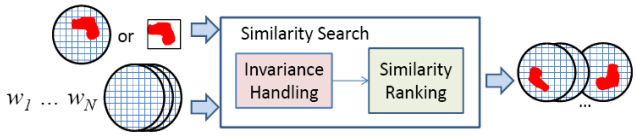


Fig. 10. **Similarity Search methodology**

The simple solution to achieving invariance is by taking the distance transform. However, as mentioned above, distance transform causes substantial information loss and may not be effective for accurate pattern matching. The alternative is to explicitly create a set of samples $p_1, \ldots, p_t$ such that each p_i is a rotated (or horizontally flipped or vertically flipped) sample from p. Then, the similarity between p and w_i is redefined as $k'(p, w_i) = \max_i\{k(p_i, w_i)\}$.

5. Experiments - Abnormality Detection

The proposed pattern mining framework was evaluated using data from the production line of an automotive SoC. Due to the extremely high quality requirement, each part is tested with a comprehensive test suite including more than 1000 parametric tests, as well as scan and BIST tests. The evaluation included thousands of wafers sampled over more than one year of production that includes the initial period, where the production line was still undergoing fine-tuning. Therefore, it was desirable to mine the wafer data to reveal abnormal patterns for improving process and/or test quality.

5.1. Abnormality Based on Specific Test Perspectives

Yield is a common indicator used to detect wafer abnormality. Wafers with substantial yield loss are usually scrapped to avoid potential quality issues. While this type of analysis can capture obvious abnormalities, it is not effective for detecting subtle abnormalities specific to a test or a subset of tests.

First, there are too many test perspectives to inspect manually. It is impractical to ask a person to exhaustively examine all test perspectives. Second, the existence of an abnormal wafer pattern does not imply that the number of failing dies based on the particular test perspective is larger than others. Hence, yield-based methods tends to overlook those abnormalities that do not cause substantial yield loss.

We applied the Abnormality Detection methodology to search for abnormalities that might have been overlooked. The novelty detection was run based on each individual test, groups of correlated tests, as well as a set of pre-defined test bins each comprising a subset of tests. In this experimental section, we report selected results based on a few test perspectives, where interesting abnormalities were found. The goal is to explain the feasibility and effectiveness of the methodology, rather than presenting a comprehensive list of found abnormalities.

In the experiments, unless otherwise stated, the default algorithmic perspective is the following. The encoding is $+1/-1$ encoding. The transform is 2×2-raster smoothening and the kernel is the Gaussian kernel. They were described in Section 3.

5.2. An example of wafer abnormality found

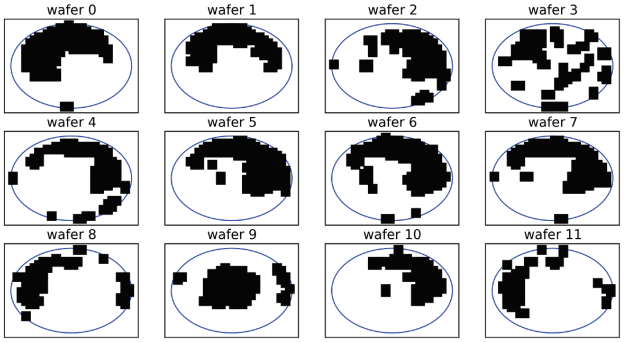


Fig. 11. **Top 12 novel wafers found based on a test perspective T_c comprising current tests**

Figure 11 shows the top 12 abnormal wafers found based on a test perspective T_c consisting of current tests. We note that the black "boxes" on all wafer maps shown in this paper are enlarged for better visualization. As a result, each wafer map looks like it has a much larger number of "failing" dies than there really were. Also note that the classification of a die as failing is test limit dependent.

These wafers were not considered abnormal because their yield loss was not high enough. However, when one focused on the particular current test perspective T_c, abnormal failing patterns were revealed.

More interestingly, wafer 0 and wafer 1 are from the same lot (lot A). Wafers 2, 4-7, and 10 are from the same lot (lot

B). This shows possible systematic behavior across the two lots. Result in Figure 11 led us to inspect all wafers in lot A and lot B respectively based on the particular perspective T_c.

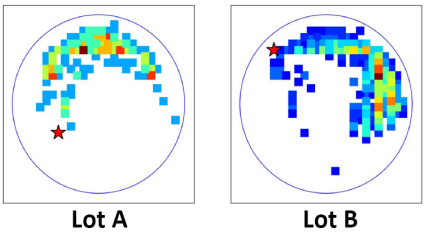


Fig. 12. **Colored wafer (heat) maps showing the number of failing dies across all wafers in the lot based on test perspective T_c**

Figure 12 shows two wafer (heat) maps based on the total number of failing dies across all wafers in each lot. The systematic behavior can be seen clearly in both lots. Notably, each lot contained a part that later was found to be a customer return. The location of the customer return on the wafer is marked as the red star. We see that both customer returns are located on the edge of the abnormal patterns.

This example demonstrates the importance of searching for abnormalities based on a particular test perspective and how the result can lead to discovery of additional systematic trends (issues) that are otherwise overlooked. The abnormalities seen in Figure 12 were later diagnosed to be caused by a test-related issue and fixed by modifying the test program.

5.3. Additional examples of wafer abnormalities

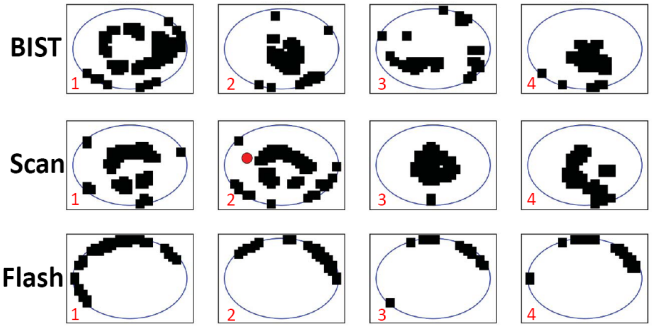


Fig. 13. **Top 4 novel wafers based on different test perspectives: (a) BIST tests, (b) scan tests, and (c) a parametric flash tests**

Figure 13 shows three more results based on three additional test perspectives: a set of BIST tests, a set of scan tests, and a parametric flash test. We see that abnormal patterns detected include: ring, semi-ring, cluster and arc. Also, the 2nd most abnormal wafer based on the scan perspective contains a part later found to be a customer return. Again, these abnormalities were overlooked by existing yield-based detection methods.

5.4. Indirect inference

It is important to note that results from pattern mining are often used as guides to pursue further analysis. They rarely are used to conclude an analysis. For example, the result shown in Figure 11 guided us to pursue the analysis on lots A and B based on the test perspective T_c. The abnormal wafers shown in Figure 11 do not contain a customer return. However, as we included the analysis of all wafers in each lot, Figure 12 shows the systematic trend that is correlated to the location of a customer return contained in another wafer of the same lot. Figure 12 is more conclusive than Figure 11.

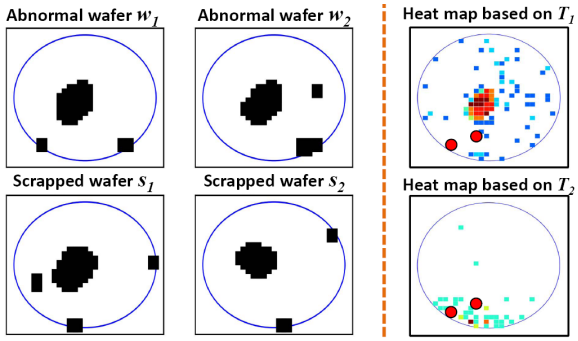


Fig. 14. **An example of indirect inference from perspective 1 to perspective 2 for identifying the abnormality of interest**

Figure 14 shows another example of indirect inference. Two wafers w_1 and w_2 from the same lot were found to be abnormal based on the test perspective T_1. This case was particularly interesting because there were two wafers s_1 and s_2 from the same lot that were scrapped. As shown in the figure, w_1 and w_2 behave similarly to s_1 and s_2 based on T_1.

It turned out that the lot was later found to contain two customer returns. Figure 14 shows the heat map based on all wafers in the lot for perspective T_1. As we can see, the two customer returns do not show strong correlation to the cluster failing pattern based on the T_1 perspective.

While w_1 and w_2 were reported among the top 10 novel wafers based on T_1, the same lot contains other wafers reported among the top 30 novel wafers based on another test perspective T_2. Initially, T_1 was selected for investigation. After the lot was found to contain more than one abnormal wafers, the next most important perspective T_2 was selected for investigation. Figure 14 shows the heat map based on T_2. This time, the T_2 heat map reveals an abnormal pattern highly correlated to the locations of the two returns.

5.5. Comparing multiple test perspectives

In a production test setting, there can be many tests and test bins. This presents a challenge for a user to apply Abnormality Detection as it may be unclear which test perspective to start with. Therefore, it is desirable to have a method that can compare abnormalities across different test perspectives.

In Section 4.1.2, we discuss this need and point to the use of distance transform as a solution. Figure 15 compare three test perspectives of three different types (current, BIST and scan) based on a distance transform. In this transform, each die is encoded with the distance to the nearest passing die. Feature f_i therefore represents the number of dies with distance i. For example, f_1 is the number of failing dies with at least one adjacent passing die. Feature f_2 is the number of failing dies where the shortest distance to a passing die is via another failing die. f_3 is the number of failing dies where the shortest distance to a passing die is via two failing dies. Figure 15 presents a 3-dimensional plot based on f_1, f_2 and f_3.

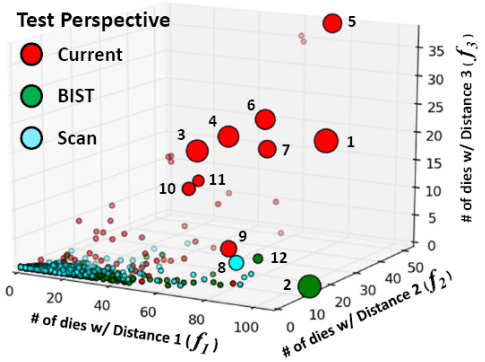


Fig. 15. **Comparing test perspectives based on visualization in a 3-D space defined with features f_1, f_2 and f_3 in distance transform**

Each dot in Figure 15 is a wafer projected into this 3-dimensional space. Outliers can be observed in this plot. For example, wafer with label "5" (based on the current perspective) is far away from the majority of other wafers. We see that this wafer has much larger values in features f_1, f_2 and f_3. This indicates that there are more failing dies surrounded by other failing dies on this wafer. In other words, this wafer is likely to contain a cluster of failing dies.

The figure shows that the current test perspective has more outliers than the other two perspectives. This indicates that with the current perspective, more wafers are likely to contain "big clusters." Hence, the current perspective can be a good starting perspective to look for abnormal wafers.

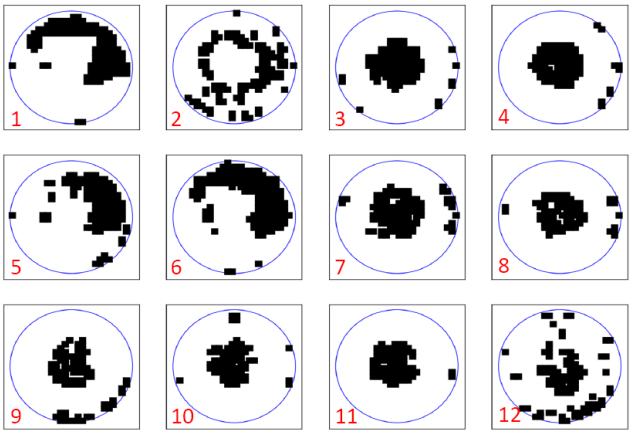


Fig. 16. **Abnormal wafers found based on multiple test perspectives collectively using distance transform**

Figure 16 shows the result of novelty detection performed on all wafer maps shown in Figure 15. The numbers in red are the novelty rank. These correspond to the labels in Figure 15. From Figure 15, we see that the wafers with labels/ranks 2 and 12 are based on the BIST perspective. The wafer 8 is based on the scan perspective. The rest are based on the current perspective. More interestingly, the outliers shown in Figure 15 are found to be the most abnormal wafer maps in Figure 16 (though the outlier ranking is not the same as the novelty ranking). This shows that while Figure 15 is not entirely accurate, this plot presents an effective visualization method for comparing different test perspectives.

The result in Figure 16 shows that distance transform can enable novelty detection across multiple perspectives. For example, in Figure 16 abnormal wafer maps are ranked and found from all three perspectives, not restricted with just one.

5.6. Comparing the three transforms

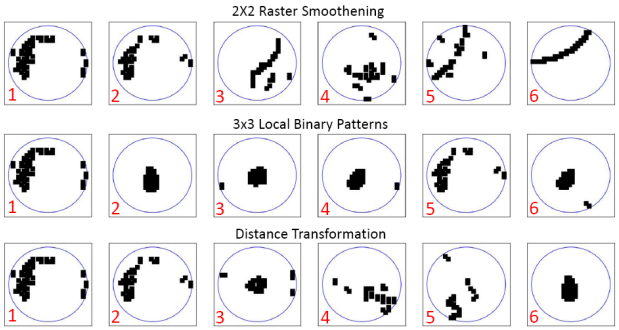


Fig. 17. **Effects of using different transforms - results based on a particular current test perspective and Gaussian kernel**

Section 3 discusses three different types of transforms implemented in this work. Figure 17 illustrates their differences in the context of novelty detection with a fixed current test perspective and the same Gaussian kernel.

Figure 17 shows that raster smoothening tends to find abnormal patterns that spread a longer distance across the wafer (i.e. a long line or arc). LBP tends to find clusters. Distance transform tends to find additional patterns other than lines/arcs and clusters. This result demonstrates the importance of transform and its affect on novelty detection.

5.7. Robustness of the novelty detection algorithm

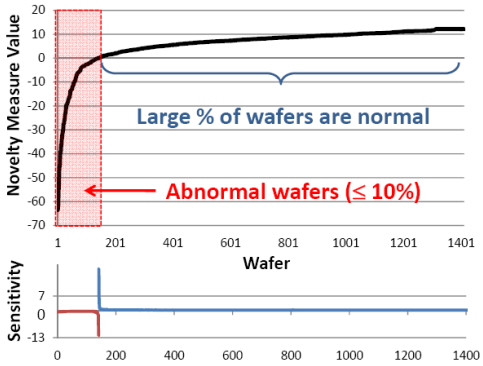


Fig. 18. **Novelty measures and their sensitivity to kernel parameter g based on the Gaussian kernel with $\nu = 0.1$ (10%)**

As discussed in Section 3.2.3, the Gaussian kernel was found to be the most effective kernel in our experiments. Given two feature vectors u and v, the Gaussian kernel measures their similarity as $k(u,v) = e^{-g\sum_{i=0}^{n-1}(u_i - v_i)^2}$. The parameter g is called a Gaussian width. Figure 18 shows how sensitive the novelty detection ranking is to this parameter g.

We sampled g from 0.001 to 0.1 with an increment of 0.001. In Figure 18, the top plot shows the average novelty measure for a collection of 1.4K wafers based on the repeated runs of novelty detection each with a different g. The bottom plot

shows the standard deviation of the novelty measure divided by the average novelty measure for each wafer. We call this quantity the sensitivity. Note that the support vector method was run with a $\nu = 0.1$ (up to 10% wafers to be classified as novel). Hence, $\sim$10% of the wafers have their novelty measures less than zero.

Figure 18 shows that the novelty measure is not sensitive to the parameter change. Only the wafers close to the novelty decision boundary are susceptible. In other words, the ranking of top novel wafers (with large negative novelty measure values) is not sensitive to the parameter change. The ranking of most of the non-novel wafers (with positive novelty measure values) is also not sensitive. The ranking is only sensitive for wafers with novelty measure values close to zero. From a user perspective, such wafers are less interesting because the focus is usually on the highly-ranked novel wafers.

6. Experiments - Perspective Search

This section presents selected results based on Perspective Search to illustrate its potential usage. In each run, a wafer was given for the search. Each wafer was of interest because there was a die on the wafer that was found to be a customer return. Our task was to find a test perspective that could reveal an abnormal pattern on this wafer.

Figure 19 shows results based on five wafers from five lots labeled as A-E. The test perspective identified for the wafer in lot A was a BIST perspective. The perspective for the wafer in lot B was a flash related test. These two perspectives are similar because they both intended to test a particular functional aspect of a flash. The same scan-test perspective was found for wafers in lot C and lot D. Another scan-test perspective was found for the wafer in lot E.

In Figure 19, a heat map is shown below each corresponding wafer pattern. Each heat map plots the accumulated failing parts, based on the test perspective, across all wafers in the lot. The correlation between the location of each customer return and the abnormal pattern is revealed in the corresponding heat map. More interestingly, the heat maps from lots A and B are similar and failure analysis later found that customer returns in lots A and B were due to the same metal-related process abnormality. Similar findings were discovered for the customer returns in lots C and D as well. These two returns shared the same reason due to a Vth-related process abnormality. The customer return in lot E was due to a latent defect, caused by a wear-out of the silicide. The heat map shows that the part could have been identified as a weak device.

7. Experiments - Similarity Search

An abnormal pattern may be known in advance or identified through Abnormality Detection and/or Perspective Search. In Similarity Search, the goal is to detect wafers with similar patterns. Section 4.3 explains that a crucial consideration for the search is to achieve pattern matching with *invariance*. Note that the search can be independent of the test perspective, i.e. if concern is not on the test that caused the pattern but the

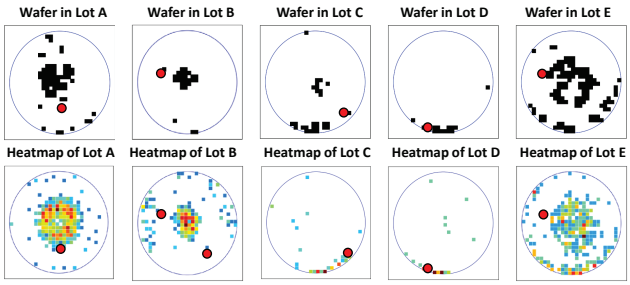


Fig. 19. **Wafer patterns identified on wafers containing customer returns. The respective heatmaps reveal that the pattern is systematic across many wafers in the lot.**

existence of the pattern itself. Hence, the search can be applied with multiple test perspectives.

7.1. Invariance search with distance transform

As discussed before, the distance transform is invariant to rotation and also insensitive to location change. In the first experiment, we show the result of the search based on distance transform and the Gaussian kernel. In Figure 20, a wafer map is given (highlighted by the red box). The 11 most similar wafers are shown. We see that Similarity Search identified wafers with similar patterns rotated. We also see that after finding the most similar patterns on wafer 1-9, the search continues to find the next group of similar patterns, which are cluster-like patterns contained in the original pattern.

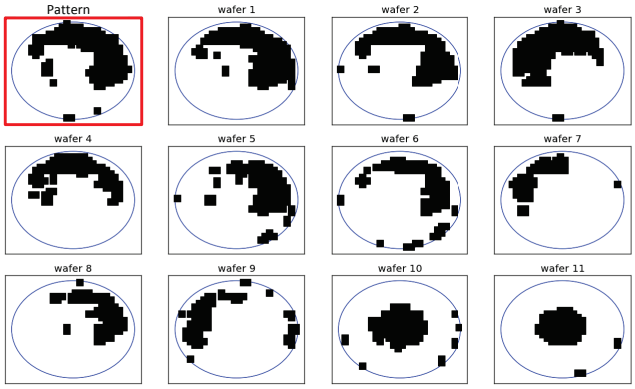


Fig. 20. **Similarity Search using the wafer "Pattern". The 11 most similar wafer patterns are shown in Wafer 1-11**

Figure 21 shows two additional results based on two different wafers. In case (a), we see that the search can successfully find similar cluster-like patterns. In case (b), we see that the search is not effective for finding similar line patterns. As discussed in Section 3, the distance transform results in the loss of the location information of a failing die. This is why the transform is not effective for identifying a line pattern where the relative locations of failing dies are important to determine the line shape.

7.2. Invariance search with explicit invariant samples

Instead of using the distance transform, Section 4.3 discussed an alternative of using artificially-created samples from the original pattern to achieve an invariance search. For the

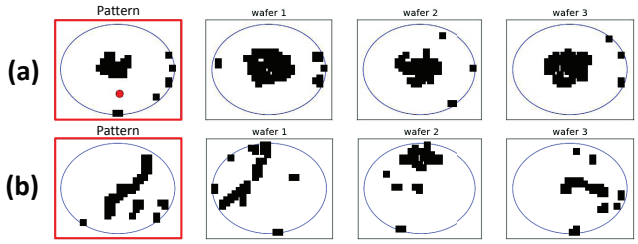


Fig. 21. **The 3 most similar wafers after performing Similarity Search using a (a) cluster and (b) line pattern**

experiments in this section, we consider three target patterns shown in Figure 22.

For a line or an edge, the pattern is rotated 10° each time to produce a collection of 36 samples. Then, the 36 samples are used simultaneously in the similarity search. In the search, a raster scan is performed on a given wafer map based on the $n \times n$ window size. The matching score for a wafer is determined by the best matching result during this raster scan.

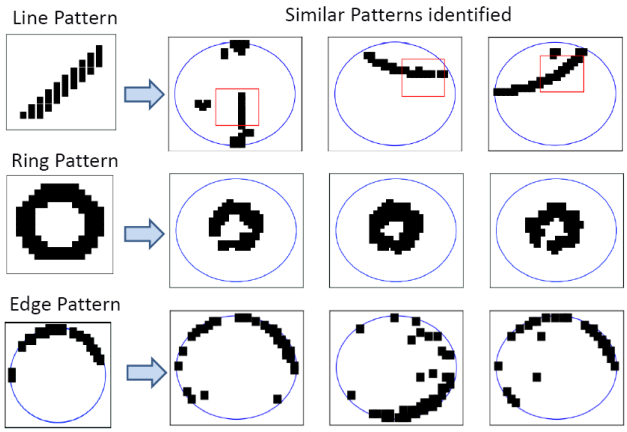


Fig. 22. **Line, ring and edge patterns and the wafers containing the respective similar patterns identified by the Similarity Search**

Figure 22 shows the top 3 wafers matching each of the given patterns, respectively. The best matched $n \times n$ window for the line pattern on each wafer is highlighted with a red box. We see that the search was able to effectively identify similar line or edge patterns in different locations with different orientations.

In the three examples, the algorithmic perspective was based on the $+1/-1$ encoding, the 2×2 raster smoothening transform and the dot-product kernel. We used the dot-product kernel for its simplicity. Using a Gaussian kernel would deliver comparable results. However, the computation of Gaussian function is more complex than dot product. Because the search is often implemented as an on-line monitor, a simpler implementation is preferred for its efficiency.

8. Conclusion

This work presents a pattern mining framework consisting of three methodologies to support the discovery, analysis and recognition of test-dependent inter-wafer abnormalities. The framework is evaluated based on an SoC product line for the automotive market. We show that the Abnormality Detection methodology could uncover abnormalities usually overlooked with a yield-based method. These abnormalities provided guides to uncover subtle systematic issues that were test related. The Perspective Search methodology was used to facilitate the diagnosis of customer returns. We show that in several instances the search was able to identify systematic trends correlated to the known returns. Finally, we show that the Similarity Search was effective for recognizing wafers containing patterns similar to a target pattern.

This work uses customer return analysis as an example application. The pattern mining framework can also be applied in other scenarios. For examples, it can be used to support the analysis of abnormal low-yield wafers and of burn-in fails. These applications will be discussed in future work.

References

[1] S. Cunningham and S. MacKinnon. Statistical methods for visual defect metrology. *IEEE Trans. Semi. Manuf.*, vol. 11, no. 1, pp. 48-53, 1998.

[2] Tao Yuan, Way Kuo, and Suk Joo Bae. Detection of Spatial Defect Patterns Generated in Semiconductor Fabrication Processes *IEEE Tran. on Semi Manuf.*, Vol 24, No 3, 2011.

[3] S. Byers and A. E. Raftery. Nearest-neighbor clutter removal for estimating features in spatial point processes. *J. Am. Stat. Assoc.*, vol. 93, no. 442, pp. 577-584, 1998.

[4] M.-S. Yang and K.-L. Wu. A similarity-based robust clustering method. *IEEE Trans Pattern Anal. Mach. Intell.*, vol. 26, no. 4, pp. 434-448, 2004.

[5] Fei-Long Chen and Shu-Fan Liu A Neural-Network Approach To Recognize Defect Spatial Pattern In Semiconductor Fabrication. *IEEE Tran. on Sem. Manufacturing*, Vol 13, No 3, 2000

[6] B. Kundu, K. P. White Jr. and C. Mastrangel Defect clustering and classification for semiconductor devices. *The 2002 45th Midwest Symposium on Circuits and Systems*, 2002

[7] J. Y. Hwang and W. Kuo. Model-based clustering for integrated circuits yield enhancement. *Eur. J. Oper. Res.*, Vol 178, No 1, pp. 143-153, 2007.

[8] K. P. White Jr., B. Kundu and C. Mastrangel Classification of Defect Clusters on Semiconductor Wafers Via the Hough Transformation. *IEEE Tran. on Semi. Manufacturing*, 2008

[9] W. Robert Daasch, et al. Neighbor selection for variance reduction in IDDQ and other parametric data. IEEE International Test Conference (ITC). 2002, pp. 1240-1248.

[10] Nik Sumikawa, et al. Screening Customer Returns with Multivariate Test Analysis. In *IEEE ITC*, 2012.

[11] Schölkopf, B et al. Support vector method for novelty detection. Neural Information Processing Systems, 2000.

[12] B. Scholkopf and A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization,and Beyond. *The MIT Press*, 2001

[13] Theodoridis, Sergios; Koutroumbas, Konstantinos. Pattern Recognition, 2nd edition, Elsevier academic press, 2003.

[14] O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. IEEE Tran. on Neural Networks, 1999.

[15] T. Ojala, M. Pietikinen, and D. Harwood. A Comparative Study of Texture Measures with Classification Based on Feature Distributions, *Pattern Recognition*, vol 29, pp. 51-59, 1996.

[16] P. F. Felzenswalb and D. P. Huttenlocher. Distance transforms of sampled functions. *Cornell Computing and Information Science Technical Report*, 2004.

[17] Wu, Sean H. et al. A Study of Outlier Analysis Techniques for Delay Testing. In *IEEE ITC*, 2008.

[18] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis. *Cambridge University Press* 2004.