CONF-970246--2

Feature Analysis and Classification of Manufacturing Signatures Based on Semiconductor Wafermaps

Kenneth W. Tobin
Shaun S. Gleason
Thomas P. Karnowski

RECEIVED
FEB 06 1997
OSTI

MASTER

"The submitted manuscript has been authored by a contractor of the U.S. Government under contract No. DE-AC05-96OR22464. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes."

Work performed for SEMATECH, Austin, Texas, under CRADA No. SC92-1082 and prepared by OAK RIDGE NATIONAL LABORATORY, Oak Ridge, Tennessee 37831-6285, managed by LOCKHEED MARTIN ENERGY RESEARCH CORP. for the U.S. DEPARTMENT OF ENERGY under contract DE-AC05-96OR22464.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Feature Analysis and Classification of Manufacturing Signatures Based on Semiconductor Wafermaps

Kenneth W. Tobin^a, Shaun S. Gleason, Thomas P. Karnowski
Oak Ridge National Laboratory^b, Oak Ridge, Tennessee

Susan L. Cohen
SEMATECH, Austin, Texas

ABSTRACT

Automated tools for semiconductor wafer defect analysis are becoming more necessary as device densities and wafer sizes continue to increase. Trends towards larger wafer formats and smaller critical dimensions have caused an exponential increase in the volume of defect data which must be analyzed and stored. To accommodate these changing factors, automatic analysis tools are required that can efficiently and robustly process the increasing amounts of data, and thus quickly characterize manufacturing processes and accelerate yield learning. During the first year of this cooperative research project between SEMATECH and the Oak Ridge National Laboratory, a robust methodology for segmenting signature events prior to feature analysis and classification was developed. Based on the results of this segmentation procedure, a feature measurement strategy has been designed based on interviews with process engineers coupled with the analysis of approximately 1500 electronic wafermap files. In this paper, the authors represent an automated procedure to rank and select relevant features for use with a fuzzy pair-wise classifier and give examples of the efficacy of the approach taken. Results of the feature selection process are given for two uniquely different types of class data to demonstrate a general improvement in classifier performance.

Keywords: Feature analysis, feature ranking, feature selection, pattern recognition, classification, fuzzy pair-wise classifier, semiconductor, wafer inspection, electronic wafermap

1. INTRODUCTION

Automated analysis of semiconductor wafer defect data has become increasingly important over the past several years as a means of quickly understanding and controlling contamination sources and process faults which impact product yield. This paper discusses the automatic analysis of defect distributions on semiconductor wafers as sensed by in-line optical inspection tools.

1.1 Spatial Signature Analysis

Trends towards larger semiconductor wafer formats and smaller critical dimensions have caused an exponential increase in the volume of visual and parametric defect data that must be analyzed and maintained by the semiconductor device manufacturer. This explosion in the volume of data has necessitated the development of automation tools for wafer defect analysis. It has been estimated that up to 80% of the yield loss in the production of high-volume very-large-scale integrated (VLSI) circuits can be attributed to random visual pattern defects¹. Contamination particles that did not create problems with 1 μm design rules can now be categorized as "killer defects" as critical dimensions dip below 0.25 μm , i.e., defects which result in improper electrical device function. Spatial Signature Analysis (SSA) is an automated

^aK.W.T. (Correspondence): Email: tobinkwjr@ornl.gov; WWW: <http://www-ismv.ic.ornl.gov>; Telephone: 423-574-8521; Fax: 423-574-6663

^bWork Performed for SEMATECH, Austin, Texas, under CRADA No. SC92-1082 and prepared by OAK RIDGE NATIONAL LABORATORY, Oak Ridge, Tennessee, 37831-6285, managed by LOCKHEED MARTIN ENERGY RESEARCH CORP. for the U.S. DEPARTMENT OF ENERGY under contract DE-AC05-96OR22464.

procedure that has been developed by the authors to address the issue of intelligent data reduction while providing timely feedback on current manufacturing conditions. SSA performs a sophisticated defect clustering and signature classification of electronic wafermaps that represent visual pattern and particle defects².

Optical inspection of semiconductor wafers has long been the primary means of detecting the sources of wafer defects. Semiconductor yield engineers use high resolution images of individual defects collected off-line to assess problems in the manufacturing process. Since high-resolution off-line defect review is time consuming and expensive, process engineers also use low resolution defect wafermaps from in-line optical inspection tools to determine the potential source of problems in the manufacturing process. They accomplish this by analyzing and sourcing unique spatial distributions or "signatures" of defects on the wafer surface by manually viewing images of the wafermap. Figure 1 shows an example of a wafermap containing various spatial signatures. Even when these spatial signatures do not contain significant portions of killer defects they provide a window into the manufacturing process which can be invaluable towards quickly identifying equipment problems. SSA attempts to emulate this human-level process for defect sourcing and to provide the fab engineer with faster time-to-results coupled with enhanced yield management strategies.

Figure 1 - Example of a high resolution defect (a), which makes up the various distributions shown in the wafermap (b).

1.2 Automatic Signature Classification

SSA begins the signature classification process by converting the electronic wafermap file into a grey-scale image where each pixel is assigned an intensity value according to the number of defects in the subtended area. Each pixel subsequently represents a defect density. Clusters or groupings of pixels, denoted as "objects", in this image are segmented into various high-level sets, or groupings, depending on their morphology^{3,4}. Each set is then characterized individually, i.e., objects belonging to each set have unique descriptive features which relate to the set. For example, elongated objects such as scratches or streaks (see Fig.1) are moved into a "curvilinear" set since they have curvilinear attributes such as elongation, compactness, location, orientation, etc. For this paper, examples will be given which are associated with the set defined as "global". Global objects are distributed over an entire wafer surface (e.g., the background pattern in the wafermap shown in Fig.1 excluding the streaks). They are generally sparsely distributed and have no highly clustered components and are treated as one wafermap object. The features used to distinguish globally distributed events are centralized geometric moments^{5,6}.

Once an object has been characterized, its features are sent to a classifier where a user-defined label is assigned to the result. For this work, a fuzzy k-Nearest Neighbor (kNN) approach has been adapted^{7,8} as described in the next section. For industrial pattern recognition problems, it is our experience that non-parametric classifiers such as Parzen or kNN⁹, apply well since information about the shape of the distribution of features in the multi-dimensional space of the classifier are not required. It is difficult to ascertain a statistical parameterization for the large variety of class types encountered. Also, in an industrial setting it is often required that the classifier system begin to classify new data with few training examples while providing reasonable accuracy. Bayesian classifiers¹⁰ and neural networks¹¹ generally require large sample populations to estimate the appropriate statistics for their method and are therefore difficult to implement for this industrial application. This is primarily due to the diverse nature of the patterns that arise for different manufacturing processes and facilities coupled with the length of time required to collect large sample populations. Also, over the period of time required to collect large sample sets, acceptable process variations can occur

which confuse the boundaries between classes. The fuzzy kNN classifier training set can readily be maintained over time (e.g., by including and excluding examples based on time and date), be modified often and can operate with relatively few examples for each class.

The focus of this paper is on an optimal selection of those features that will give high classification accuracy. There are many feature reduction methods described in the literature which can be applied to similar classification problems, e.g., linear transformations such as eigenvector projections or the Fisher discriminant⁷, or searching a directed graph, e.g., branch and bound techniques¹². The technique described herein is related to directed graph searching except that the feature search criteria are established *a-priori* based on a ranking of the features according to their class-discrimination potential. The method relies on using a pair-wise adaptation of the fuzzy kNN classifier so that the feature-space for any given pair of classes can be significantly reduced. For example, the classes defined in the global set for wafermap analysis are described by 25 geometric moments as mentioned above plus 3 non-moment features. It is rarely necessary that all 28 moments be used to discriminate between any given pair of classes in the set. In many cases, as will be shown later in the paper, as few as one feature is required to discriminate a test vector between a given pair of the defined classes. Typically, this method results in a requirement to use only a few features for any given pair. A reduction in feature dimensionality coupled with this pair-wise implementation results in improved classifier performance. It should be noted that the technique for feature ranking and selection described below is not limited in application to the fuzzy kNN classifier but should apply equally well to many other classifier types.

2. FUZZY PAIR-WISE CLASSIFIER

For this work a fuzzy kNN method has been adapted to perform a pair-wise classification of a test vector to a user-defined training set of class examples. The fuzzy kNN classifier assigns a membership to a test vector according to the following relationship⁸,

$$\mu(x)_i = \frac{\sum_{j=0}^{k-1} \mu_{ij} \left(1/\|x-x_j\|^{2/(m-1)} \right)}{\sum_{j=0}^{k-1} \left(1/\|x-x_j\|^{2/(m-1)} \right)}, \quad (1)$$

where $\mu(x)_i$ is the membership of the test point, x , to class i , $\|x-x_j\|$ is the L -norm distance between the test vector, x , and class vector x_j , k is the number of nearest neighbors, and m sets the strength of the fuzzy distance function. μ_{ij} is the membership value of the j -th neighbor to the i -th class and is determined *a-priori* from the training data⁸. The implementation of Eq. (1) for this application is outside the scope of this discussion. The point to note is that fuzzy membership results are integral to measuring and handling classification ambiguity. Class ambiguity measurements derived from fuzzy membership values are used to help select relevant features and will be further discussed below.

The motivation for achieving a significant dimensionality reduction in feature space is to improve classifier performance by discarding "noisy" features, i.e., those features which attribute little or no discriminatory power to the algorithm. By using only features intrinsic to the classification problem, a significant performance improvement can be realized. From the point of view of a Bayesian classifier, performance cannot be degraded through the addition of superfluous features. For real-world applications, though, the assumptions associated with a Bayes classifier are rarely valid¹². This effect can be explained for distance-based classifiers like kNN by considering the L -norm distance terms in Eq. (1), which, for the point of demonstration, can be expressed as,

$$d(x, x_i) = \|x - x_i\| = \left[\sum_{\text{discriminating}} (x - x_i)^p + \sum_{\text{noise}} (y - y_i)^p \right]^{1/p}, \quad (2)$$

where p defines the L -norm distance measure used, e.g., $p = 1$, Manhattan distance, $p = 2$, euclidean distance, etc. It has been assumed in this equation that the class features can be segmented into two distinct groups: those that contribute to the discrimination of classes, and those that do not. The *discriminating* term in Eq. (2) will provide a statistically significant contribution to a distance measure such as that in Eq. (1), while the *noise* term will add a positive bias to the summation in a random fashion, i.e., it will tend to wash-out the discriminatory power of those features which best represent the class. By reducing the dimension of the feature space, it is intended that the *noise* term in Eq. (2) be

mitigated. The technique described in this paper attempts to provide this segmentation such that there is minimal overlap between the resulting groupings of features.

A pair-wise implementation of the fuzzy kNN classifier has been designed to maximize the reduction in the number of features required to correctly classify a test vector within a given pair of classes. A pair-wise comparison is performed for all unique class pairs, i.e., there are $N(N-1)/2$ unique pair-wise combinations for N classes. The pair-wise fuzzy membership result, μ_{ij} , for a test vector, x , can be expressed by the following,

$$\mu(x)_{ij} = \begin{bmatrix} \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} & \begin{pmatrix} \mu_0 \\ \mu_2 \end{pmatrix} & \begin{pmatrix} \mu_0 \\ \mu_3 \end{pmatrix} & \begin{pmatrix} \mu_0 \\ \mu_4 \end{pmatrix} & \dots \\ & \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} & \begin{pmatrix} \mu_1 \\ \mu_3 \end{pmatrix} & \begin{pmatrix} \mu_1 \\ \mu_4 \end{pmatrix} & \dots \\ & & \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} & \begin{pmatrix} \mu_2 \\ \mu_4 \end{pmatrix} & \dots \\ & & & \begin{pmatrix} \mu_3 \\ \mu_4 \end{pmatrix} & \dots \\ & & & & \dots \end{bmatrix} \quad \forall i=0, \dots, N-1, j=i, \dots, N-1, \quad (3)$$

where each element of μ_{ij} is determined by applying Eq. (1) to the class pair ij . The fuzzy membership of the test vector for each membership pair is then combined to give a membership to each class in the set which can be expressed by,

$$M_i(x) = \frac{1}{N-1} \begin{bmatrix} \sum_{\forall j \neq 0} \mu_{0j}^0 \\ \sum_{\forall j \neq 1} \mu_{1j}^0 \\ \cdot \\ \cdot \\ \sum_{\forall j \neq 2} \mu_{(N-1)j}^0 \end{bmatrix}, \quad (4)$$

where the superscript on μ_{ij}^0 selects the first membership element of μ_{ij} . Note that the range for each membership value, M_i , is $[0,1]$, but that $\sum M_i$ does not necessarily equal 1. It should also be noted that a crisp classification is assigned to the test vector, x , after defuzzification of the membership values given by Eq. (4). The defuzzification process is outside the scope of this paper but is designed to allow a test vector to be classified as *unknown*. Data given the *unknown* label are considered to have ambiguous classification results and therefore low confidence that a correct classification can be made based on the available information.

Although significant feature dimension reduction can be achieved by use of the pair-wise method described above, an increase in computational expense is incurred. Where the fuzzy kNN algorithm of Eq. (1) must be applied N times for an N -class problem, the pair-wise fuzzy kNN implementation requires $N(N-1)$ applications to achieve the resulting memberships given by Eq. (4). This increased computational burden has proven to be worth the benefit of improved classifier performance.

3. FEATURE MEASUREMENTS AND RANKING

Figure 2 shows several examples of global signatures which have been collected during semiconductor manufacturing. Central moments have been chosen as descriptive features for these patterns for the following reasons: (1) the entire distribution plays a role in the signature, i.e., there are no background or noise levels to contend with; (2) the patterns consist of distinct and reoccurring structures which make them amenable to moment descriptions, and; (3) moment descriptions provide good empirical discrimination between classes. The application of moments to image analysis is described thoroughly in the literature^{5,6}.

The data shown in Fig.2 are examples of the many distinct patterns which can arise in semiconductor manufacturing due to various particle contamination and faulty mechanical processes. The location, skew, density, extent, etc. of the wafer defects all describe various potential manufacturing problems and can guide the fabrication engineer in the determination of the defect source. These examples are typical but by no means do they cover the entire gamut of potential patterns which may arise. For this reason it is important that the feature description and the classifier be able to accommodate large variability from one semiconductor application to the next.

Figure 2 - Example data from six different classes of globally distributed signatures. This data is representative of true manufacturing signatures found in industry.

To facilitate the process of selecting the best performing features which describe the data shown above (or other class data) two methods of feature ranking were investigated: (1) a fuzzy set theoretic measure which describes the distribution of a class' feature data in terms of an index of fuzziness¹³, and; (2) the following simple statistical feature evaluation index (FEI) which uses the class mean and standard deviation for a given feature,

$$FEI_{ij} = \frac{(m_i - m_j)^2}{\sigma_i^2 + \sigma_j^2}, \quad (5)$$

where i and j are indices for the class pair being evaluated. Although tested fairly extensively on semiconductor data, it was determined that the fuzzy set theoretic approach was not a strong discriminator of the class type under consideration. The statistical FEI strongly considers the width of the distribution of a feature across the class pair, i.e., the denominator of Eq. (1), while the fuzzy method focused more on the distribution of the class features within a class. Therefore the statistical FEI method proved more robust for this application and was selected for use.

To simplify the following description of the feature selection process, and to demonstrate the generality of the feature evaluation approach, the Iris class data will also be used as an example⁷. The Iris data set contains 150 measurements

of sepal length, sepal width, petal length, and petal width for three types of Iris flower: Iris-virginica, Iris-versicolor, and Iris-setosa. This is a three class problem based on four feature measurements and has been referenced often in the literature since R. A. Fishers 1936 paper on discriminant analysis. Examples of the feature evaluation approach as applied to both the Iris data and semiconductor manufacturing data will be given in the results section.

4. FEATURE SELECTION

One could consider the feature selection problem as a search through all possible permutations of measured features using a performance metric to select the best combination. Obviously, this brute force approach is unwieldy since for a J-dimensional problem, there are $2^J - 1$ permutations which must be considered. There has been much work done in the field to determine an efficient means of minimizing the search through various orderings of features e.g., step-wise search techniques like backward and forward selection⁹, and branch and bound methods¹². The method applied in this research uses the feature ranking criteria to order the class features in terms of their discriminatory power. Features are selected based on their rank and added to an ordered list, that consists of a set of feature selection masks, $\{W_n | n = 0, 1, \dots, J-1\}$ which begins with the highest ranking feature only, followed by the highest ranking two features, etc. Based on this ordering scheme, there will never be any more than J permutations which must be tested, where J is the number of measured features. For this work, the performance criteria is based on a hold-one-out⁷ test (HOO) performed using the class training data and the feature selection list.

HOO performance is determined by holding one test vector out from a test set prior to training, followed by showing the test vector to the trained classifier. This procedure is continued until all the test data has been held out once. The resulting expected performance is determined as the fraction of correct classifications made per class or for the entire test set. For the pair-wise fuzzy kNN classifier this entails calculating the classifier performance for each pair of classes in the training data once for each of the ordered sets of features, i.e., $JN(N-1)$ times.

Although still in the testing stage, one other piece of information is being investigated for determining the best combination of features to use for optimal class performance. The test vector x , is returned from the classifier with a fuzzy membership to each pair of classes, μ_{ij} (see Eq. (3)), this membership is used to calculate an index of fuzziness (ambiguity)¹³ for each tested pair according to the relationship,

$$\gamma(A) = \frac{2}{\sqrt{J}} \left[\sum_q [\mu_A(x_q) - \mu_{\bar{A}}(x_q)]^2 \right]^{1/2} \quad (6)$$

where μ_A represents the fuzzy membership of the vector x_q to class A, and $\mu_{\bar{A}}$ is the crisp membership of x_q to class A, where $\mu_{\bar{A}} = 0$ if $\mu_A \leq 0.5$, and $\mu_{\bar{A}} = 1$ if $\mu_A > 0.5$, and q is a summation index over the number of training samples. Note that $\|\mu_A - \mu_{\bar{A}}\|$ is defined as the distance between fuzzy set A and its ordinary set. The index of fuzziness expresses the amount of fuzziness inherent in the set A. For the limiting case where all members of set A have fuzzy membership 0.5, then $\gamma_{\max} = 1$, i.e., the set is maximally fuzzy. For the case where all membership values in the set A are 0 or 1, then $\gamma_{\min} = 0$, i.e., the set is crisp. The index of fuzziness adds additional information to the feature selection process which quantitatively describes the set ambiguity which can be interpreted as a degree confidence in the data.

An example of the application of this approach for a comparison of the class Iris-virginica to Iris-setosa is given for the Iris data and is shown in Table I. The FEI value is shown along the first row for each of the measured features. Following the FEI data are feature selection mask vector W_n . The mask is represented as a vector whose element is 1 if the corresponding feature is to be used and whose element is 0 if the feature is not used. There are four masks developed for the four-dimensional space of the Iris data. The column labeled "HOO performance" gives the resulting classification performance obtained by showing the selected features to the classifier. The right-most column labeled "Index of fuzziness" contains the fuzzy index result corresponding to the classification performance. Note that the performance is relatively high for all feature selection masks but that the maximum performance is obtained for mask W_3 . Also note that the ambiguity is a minimum for W_1 .

These results suggest that possibly all four features are required to separate the two classes based on performance. But what of the ambiguity? Is it possible that having minimal ambiguity in the data could lead to more robust performance? To investigate this possibility, a selection criteria was developed which can select from the list of choices presented by the data based on both the performance and the ambiguity. The data used to generate Table I consists entirely of training data. It is proposed that the system can provide more robust decision making capabilities for previously unseen data if a combination of expected performance and ambiguity are used. For this reason, a feature selection metric has been defined which combines the class performance, P_{ij} , with the index of fuzziness, γ_{ij} , where i and j represent the class pair under test. This is given by the relationship,

$$\Gamma_{ij} = P_{ij}(1 - \gamma_{ij}) \quad (7)$$

TABLE I Feature selection masks, W_n , for Iris-virginica vs. Iris-setosa						
Feature	sepal length	sepal width	petal width	petal length	HOO performance	Index of fuzziness
FEI	0.0201	0.0138	0.112	0.1369		
W_0	0	0	0	1	0.95	0.209
W_1	0	0	1	1	0.94	0.0839
W_2	1	0	1	1	0.92	0.1239
W_3	1	1	1	1	0.97	0.1781

Note that the performance and the index of fuzziness are both bounded over the range [0,1]. Therefore the feature selection metric is also bounded over this range. Some of the difficulties associated with the appropriate selection of the optimal feature mask, W_n , can be observed from the semiconductor wafer data shown in Fig. 3. This data represents the performance P_{ij} , index of fuzziness (ambiguity), γ_{ij} , and the feature selection metric, Γ_{ij} , for the two-class comparison "dense" versus "crescent" (see Fig. 2). This data is plotted against the number of ranked features tested.

Note that the performance alone does not provide enough information to make a decision about the number of features which should be used to label this pair of classes. There are four constant-performance plateaus in the top plot from which to select the appropriate number of features, i.e., for any of these plateaus, similar performance can be expected. The plot of ambiguity shows that the decision boundary between the class pairs becomes more confused as more features are added. The selection criteria function shown in the final plot maximizes the performance while minimizing the ambiguity and the choice reduces to selecting one, two, or three features. Since the selection criteria function is constant for these three choices, the smallest number of features is chosen from the set to computationally simplify the comparison, i.e., as indicated, one feature is selected for this class pair.

Figure 3 - Graphs showing performance, index of fuzziness (ambiguity), and the selection criteria for the comparison of the class pair "dense" and "crescent" shown in Fig. 2.

5. RESULTS

The Iris flower data and the semiconductor wafer data will be used to demonstrate the results of applying the feature ranking and selection procedure described above. In general, the results for this method have proved to be very effective at dramatically reducing the feature-space of the problem while improving the performance of the fuzzy kNN classifier. Figure 4 shows several examples from the semiconductor wafer data that demonstrate how the pair-wise performance varies as more features are used with the classifier. Note in some instances (e.g., the dense / ring comparison) the performance actually improves prior to degrading as more features are considered, while in others (e.g., the dense / medium comparison) the performance simply degrades as any new features are added. In all cases there is some optimal performance point achieved by using less than the total number of features measured.

Table II shows the number of features selected for each pair-wise comparison of the three classes in the Iris data. Note again that the feature space has been reduced from four dimensions to one or two dimensions. Table III shows the confusion matrix for the training set of Iris data which is composed of 100 samples. The left-most column represents the known labels for the data whereas the top row represents the computer classification results. The overall expected performance for the classifier using the HOO technique is 96% for the reduced feature space. Table IV was generated by testing the classifier with Iris data which was held back during training (50 samples). The test performance of 93% is only slightly less than the expected performance of 96% shown in Table III. It has been anticipated that the feature selection technique which incorporates the class ambiguity described by Eq. (7) will tend to mitigate discrepancies between the expected classifier performance (e.g., based on training with the HOO method) and the operation of the trained classifier in the field (represented in this case by the Iris test data in Table IV). At this point in the research, there has not been enough semiconductor wafer data collected and tested to verify or refute this claim for this application but funded research is ongoing and the results have been encouraging thus far.

Figure 4 - Performance of various pair-wise class comparisons of the semiconductor wafer data as a function of the number of features used.

TABLE II Number of features out of 4 selected for each of the pair-wise comparisons for the Iris data. Results obtained using the feature selection metric given by Eq. (7)			
	Iris setosa	Iris viricolor	Iris virginica
Iris-setosa		1	2
Iris-viricolor			1

Figure 5 is an image representation of the fuzzy membership results obtained for the semiconductor wafer data prior to crisp classification (see Eq. (4)). Each column of the data contains the $M_i(x_q)$, where $i = 0, 1, \dots, N-1$ and N is the number of classes. The feature vector x_q represents one of the training samples and $q = 0, 1, \dots, Q-1$ with Q the number of samples considered. The semiconductor wafer training set contains 6 classes of data with 28 measured features per sample, distributed among 45 samples. Each row of the grey-scale image represents the membership of the different class examples to the different classes in the set. A light intensity represents a high class membership whereas a dark intensity represents a low class membership. Table V shows the number of features selected to distinguish the unique class pairs in the training set. Note the dramatic reduction in the feature-space for this problem which results in at most nine features being used for the medium / sparse comparison. More often than not only one feature is required to successfully distinguish class pairs. Table V shows the resulting confusion matrix and the expected classifier performance of 93%.

Note from the confusion table for the semiconductor wafer data in Table VI that there is some mis-classification, or confusion, between the classes "dense" and "ring". Referring to Fig. 2, it is understandable based on the data that there is some confusion between these two classes. In particular, compare the first and third examples under "dense" to the first and second examples under "ring". It is apparent from the fuzzy matrix shown in Fig. 5 that one example in the "dense" class has a higher membership to the "ring" class. The mis-classified example is the third example in the "dense" class (see both Fig. 2 and Fig. 5). The first two examples in the "ring" class have higher memberships to the "dense" class. Needless to say, the error in these classifications are not related to the features selected, but rather to the similarity of the three examples in question.

TABLE III Confusion matrix showing classification results using the HOO technique for the Iris training data. Results obtained using the feature selection metric given by Eq. (7)						
known classification	computer classification			unknown	per-class performance	total performance
Iris-setosa	31	0	2	0	94%	96%
Iris-virsicolor	0	35	0	0	100%	
Iris-virginica	2	0	30	0	94%

Figure 5 - Grey-scale display of fuzzy membership values for the semiconductor wafer data shown in Fig. 2 for the 6 class, 28 feature problem.

TABLE IV Confusion matrix showing classification results using the HOO technique for the Iris test data. The classifier was trained with the data shown in Table II which includes using the feature selection metric given by Eq. (7).						
known classification	computer classification			unknown	per-class performance	total performance
Iris-setosa	14	0	2	1	84%	93%
Iris-virsicolor	0	15	0	0	100%	
Iris-virginica	0	0	17	1	96%	

TABLE V Number of features out of 28 selected for each of the pair-wise comparisons for the semiconductor wafer data shown in Fig. 2. Results obtained using the feature selection metric given by Eq. (7)						
	dense	ring	medium	sparse	crescent	edge
dense		3	1	1	4	1
ring			2	1	3	1
medium				9	2	1
sparse					1	1
crescent						1

In all cases tested to date, the feature ranking and selection procedure has resulted in a reasonable and conservative selection of relevant features for the classification problems tested. Most mis-classification issues have been associated with similarities in the underlying data samples, whether considering the semiconductor wafer data or the Iris data. In all cases tested, classifier performance has been improved by reducing the number of features to those intrinsic to making the distinction between the unique class pairs.

TABLE VI									
Confusion matrix showing classification results using the HOO technique for the semiconductor wafermap training data shown in Fig.2 using the feature selection metric given by Eq. (7).									
known classification	computer classification						unknown	per-class performance	total performance
dense	8	1	0	0	0	0	0	89%	93%
ring	2	8	0	0	0	0	0	80%	
medium	0	0	5	0	0	0	0	100%	
sparse	0	0	0	8	0	0	0	100%	
crescent	0	0	0	0	8	0	0	100%	
edge	0	0	0	0	0	5	0	100%	

6. CONCLUSIONS

The proposed technique for ranking and selecting relevant features to improve performance for the automatic classification of defect patterns on semiconductor wafers has been very successful to date. The feature selection technique has resulted in improved performance for the unique implementation of the pair-wise fuzzy kNN classifier. The approach, which is based on using both the pair-wise classifier expected performance and the class pair ambiguity, is unique for this application and holds promise for helping to improve robustness of the trained classifier system in the field. Robustness in this instance refers to matching the classifiers performance in the field with the expected performance based on the classifier training data and the HOO technique. This supposition will be tested in the coming months as field trials of the classification system begin.

Finally, it should be noted that the proposed feature ranking and selection method has applicability to a wide range of feature ranking and classification methods and is not limited to the techniques described herein. For example, a method based on using a Fisher discriminant for feature ranking coupled with a pair-wise implementation of a neural network classifier should provide improved performance over a non-pair-wise implementation which uses all measured features. Although the kNN problem has been posed in terms of separating the class features into *discriminatory* and *noisy* sets (i.e., see Eq.(2)), other non-distance-based analogies can be made which are applicable to classifiers that are not-distance-based.

REFERENCES

1. T. Hattori, "Detection and Identification of Particles on Silicon Surfaces", *Particles on Surfaces, Detection, Adhesion, and Removal*, Edited by K. L. Mittal, Marcel Dekker, Inc., New York, p.201.
2. S. S. Gleason, K. W. Tobin, and T. P. Karnowski, "Spatial Signature Analysis of Semiconductor Defects for Manufacturing Problem Diagnosis", *Solid State Technology*, July, 1996.
3. K. W. Tobin, S. S. Gleason, T. P. Karnowski, M. H. Bennett, "An Image Paradigm for Semiconductor Defect Data Reduction", *SPIE's 1996 International Symposium on Microlithography*, Santa Clara Convention Center, Santa Clara, CA, March 10-15, 1996.
4. Gleason, S.S., and Tobin, K.W., "Directional-based Dilation for Connection of Piece-Wise Objects: A Semiconductor Manufacturing Case Study", *International Conference on Image Processing*, Lausanne, Switzerland, Sept 16-19, 1996.
5. S. X. Liao, and M. Pawlak, "On Image Analysis by Moments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 3, March 1996, p.252.
6. C. The, and R. Chin, "On Image Analysis by the Method of Moments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 4, July 1988, p.496.
7. R. O. Duda, and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973, P.95.
8. J. M. Keller, M. R. Gray, and J. A. Givens, Jr., "A Fuzzy K-Nearest Neighbor Algorithm", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-15, No. 4, p. 580, July/August 1985.
9. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd Ed., Academic Press, Inc., San Diego, 1990.
10. J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons Limited, West Sussex, England, 1994.
11. S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan College Publishing Company, New York, 1994.
12. W. Siedlecki, and J. Sklansky, "On Automatic Feature Selection", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 2, No. 2, 1988, p.197.
13. S. K. Pal, and B. Chakraborty, "Fuzzy Set Theoretic Measure for Automatic Feature Evaluation", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-16, No. 5, September/October 1986.