On the Use of Decision Tree Induction for Discovery of Interactions in a Photolithographic Process

Dan Braha and Armin Shmilovici

***Abstract*—This paper delineates a comprehensive and successful application of decision tree induction to 1054 records of production lots taken from a lithographic process with 45 processing steps. Complex interaction effects among manufacturing equipment that lead to increased product variability have been detected. The extracted information has been confirmed by the process engineers, and used to improve the lithographic process.**

The paper suggests that decision tree induction may be particularly useful when data is multidimensional, and the various process parameters and machinery exhibit highly complex interactions. Another implication is that on-line monitoring of the manufacturing process (e.g., closed-loop critical dimensions control) using data mining may be highly effective.

***Index Terms*—Data mining, decision tree induction, photolithography, semiconductor process control, yield management.**

I. Introduction

MODERN semiconductor manufacturing, especially of integrated circuits, is very complex: contemporary processes contain up to 600 process steps [1]. Maintaining high yield levels at the end of manufacturing has been recognized as an important element to the competitiveness of semiconductor manufacturing companies [2]. A study into the practice of leading fabs [3] has revealed (among other findings) that

- They collect voluminous data into engineering databases, generated from various sensors tracking production data,[1] and information related to major events of yield losses.
- The engineering database of die yields is integrated with parametric measurements taken at the end of the manufacturing line for end-of-line yield analysis.
- Statistical process control (SPC) is adopted as a means of detecting manufacturing problems and improving the process performance. Large volumes of data are analyzed in order to quickly and comprehensively adjust the manufacturing process and equipment. Specific yield models are developed for their fabs. Automated statistical correlations are computed in order to ascertain what characteristics are common to low-yield wafers, and to automatically identifying the out-of-control conditions.

Traditionally, yield improvements have been achieved through the use of statistical and experimental design techniques [2]. Other methods, such as yield modeling and simulation techniques [2], [4] also depend on elaborate statistical techniques. Unfortunately, there are practical limitations in automating the statistical techniques when complex interactions and nonlinearities are involved in the underlying models. Moreover, the large amount of data in current semiconductor databases makes it almost impractical to manually analyze them for valuable decision-making information. This difficulty is mainly due to the large amount of records (each contains hundreds of attributes), which need to be simultaneously considered in order to accurately model the system's behavior.

The need for understanding complex interaction effects as well as the need for extracting useful knowledge from huge amounts of raw data using automated analysis and discovery tools have led to the development of knowledge discovery in databases (KDD) and data mining methodologies [5], [6]. The term KDD denotes the entire process of turning low-level data into high-level knowledge. Data mining is considered a particular step in an overall process that comprises the application of specific algorithms for extracting patterns. The data mining activity is often classified as follows: 1) *Discovery-based* methods, which look for hidden patterns in the data; 2) *Predictive-based* methods, which apply patterns to forecast future values; and 3) *Forensic-based* methods, which look for exceptional and unusual patterns. The most commonly used data mining techniques can be categorized in the following groups [7], [11]–[13]: statistical methods, artificial neural networks, decision trees, rule induction, case-based reasoning, Bayesian belief networks, and genetic algorithms.

There have been several attempts for applying KDD in the semiconductor industry. In [7], comprehensive applications of data mining within semiconductor manufacturing environments are described. In [8], an architecture is proposed for a generic integrated yield management system, without an actual application of it. A composite architecture that combine several data mining methods has been presented in [9], and has been applied to the refinement of a new dry cleaning technology that utilizes a laser beam for the removal of micro-contaminants.

Rietman *et al.* [21] isolated *manually* 22 (out of 300) important processing steps, and constructed a neural network system model from the records of 5646 unique lots. The neural network regression model was used for predicting four yield metrics. Kim and May [10], constructed four sequential neural net-

Manuscript received January 27, 2003; revised July 3, 2003. This work was supported by the consortium for control and management of manufacturing clusters in the semiconductor industry, as part of the MAGNET program of the Israeli Ministry of Trade and Industry.

D. Braha is with the School of Industrial Engineering and Management, Ben-Gurion University, Beer-Sheva 84105, Israel (e-mail: brahad@bgumail.bgu.ac.il).

A. Shmilovici is with the Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel (e-mail: armin@bgumail.bgu.ac.il).

Digital Object Identifier 10.1109/TSM.2003.818959

[1] For instance, data related to the machines used to process a lot; the operators attended the process; the batch of chemicals used, etc.

0894-6507/03$17.00 © 2003 IEEE

work subprocess models for a via formation process (one for each yield metric). A Genetic Algorithm was used to further optimize the parameters of the neural network models for process recipe update recommendations.

In this paper, we present an exploratory study within an actual lithographic process, which is composed of 45 subprocesses. Based on the records of 1054 unique lots, a decision tree induction algorithm has been employed in order to enhance the understanding of the intricate interactions between different processes, and to extract high-level knowledge that can be used to enhance the overall process quality. Unlike Rietman *et al.* [21], the important processing steps were identified *automatically* by the decision tree induction algorithm. In addition, the natural language like rules generated by the decision trees were found to better facilitate the corrective actions by the process engineers than neural networks "black box" models of (e.g., [10], [21].

Our study shows that data mining within semiconductor manufacturing environments can 1) build a series of *models*—without incorporation of any physical device models - that *consistently* represent the complex interactions among the consecutive processes as embedded in historical data; 2) deduce the set of parameter values that will obtain the highest success rate of the various target functions (e.g., enhance the photolitographic process); 3) identify conditions (e.g., specific uncalibrated machine) under which reduced yield is expected, and human intervention is needed; 4) effectively handle realistic industrial settings where data is dynamic and voluminous (i.e., thousands of records, dozens of attributes), and manual exploratory analysis is impractical; 5) represent the extracted high-level knowledge in a transparent manner (e.g., if–then rules) that helps the yield engineers to understand the derived models; and, thus, make decisions that improve the lithographic process; and 6) provide on-line monitoring and yield analysis; thus, enhancing the effectiveness of closed-loop critical dimensions (CD) control.

The paper is organized as follows: Section II outlines the lithographic process under examination. Section III provides a brief overview of decision tree induction. Section IV presents the experiments and the extracted high-level knowledge. Section V concludes the paper.

II. Overview of the Lithographic Process

A. The Lithographic Process Examined in this Paper[2]

To ensure a successful implementation of data mining within semiconductor manufacturing requires the understanding of the manufacturing process to which the data mining is applied. This will help to refine the goals and tasks of the data mining process; e.g., understanding the factors that might affect the reproduced critical dimensions or alignment offset results, etc.

Integrated circuits are built up from a number of patterned silicon, oxide, or metal layers, with specific characteristics. This research focuses only on data collected from the lithographic area. The goal of the lithographic process is to accurately transfer a set of opaque images from the masks, which represent the elements of the basic circuit design, to a substrate on the wafer, with virtually zero defects. A photoresist chemical is used to transfer the desired pattern by masking specific regions of the device from etching or ion implantation processes. During the process of wafer fabrication, a series of loops are performed, each one adding another layer to the device. Each loop is comprised of some or all of the major steps (elaborated in [1]) of photolithography, etching, stripping, diffusion, ion implantation, deposition, and chemical/mechanical planarization. Typical process parameters are detailed in [10]. At each stage, various inspections and measurements are performed to monitor the process and equipment. Supporting the entire process is a complex infrastructure of material supply, waste treatment, support, logistics, and automation.

The metrology information is used as a source of feedback to improve and confirm top quality operation. The main wafer features that are examined by the metrology cluster include: 1) *Develop Check*, which is concerned with examining the wafer for visual defects as well as measuring the reproduced critical dimensions; and 2) *Overlay Registration*, which is concerned with testing the relations between the different layers.

A product with poor results on the above tests is sent to re-work. That means that the resist must be stripped off, and the photolithography process is repeated. Higher rates of rework are frequently associated with die yield losses, higher cycle time, and out of control processes. The above tests are further detailed below:

- *CD metrology*: Critical dimensions are measured with scanning electrons microscopy (CD SEM). The CD SEM measures features on the wafer to insure that the steppers and etch tools are processing within the wanted targets. Development sites determine what layers and what sites on those layers have critical dimensions that need to be monitored closely in order to ensure speed performance and reliability at the end of line.

  Since the process is normally quite repeatable, CD's are normally measured on only a small sample of wafers from a process lot. A lot typically contains 24 wafers that are processed together. In the current process, one wafer from each lot is measured on 5 different sites having predetermined geometric patterns (called CD cells) that represent the technology (e.g., 0.7 micron). Since the wafer is not ideally planarized, different features are located at different focal lengths. Since the "walls" are not exactly vertical, different measurements are performed for the topographically high features and for the topographically low features. The measurements are taken twice: after photo (DI) and after etch (FI). To illustrate, for a particular product, which has a mask CD of 0.69 microns (as specified by the mask shop), the average DI measurements for one wafer from the lot were 0.650 (respectively 0.735) for the low (respectively high) feature. The standard deviations were 0.014 and 0.007, respectively. The FI measurements for the same wafer were 0.710 (respectively 0.777) for the low (respectively high) feature, with standard deviations of 0.036 and 0.058, respectively.

[2] This study has been conducted in a major wafer foundry that strategically focuses on advanced flash memory and CMOS image sensor technologies. Reflecting various customers' specifications, the fab runs multiple products in different batch sizes. Hence, the learning rate for any new production run has to be very fast.

The value of the CD measured after the development inspection (DI) is used in a feed-forward manner to allow customization of the etch process recipe on a lot-by-lot basis. Variability in the DI–CD value for a lot can be partly compensated for by manipulating the associated lot etch recipe. Using the underlying process knowledge, tool engineers monitor the FI–CD, resulting from the post-etch inspection, and make process adjustments accordingly. These adjustments ensure that the etcher is not processing out of the desired target.

- *Overlay metrology*: The measurement of overlay entails the determination of the centerline (position) of one target, from an underlying lithographic level relative to a second target in the upper lithographic level. Special targets (e.g., concentric squares) are measured to yield "x" and "y" centerline values for the target. Overall, 14 measurements are made, and the difference in centerline values between the upper and lower levels yields the measured "overlay offset" vector with both an "x" and "y" component. The vector component values are then used for process control, stepper feedback, and rework determination. To illustrate, the average alignment offsets associated with the top layer of a certain wafer from a particular lot were -0.06 for the x axis, and -0.03 for the y axis. The maximum misalignments were 0.20 for the x axis, and 0.11 for the y axis. In this case, since the misalignments are much smaller than half the CD, the chips on the wafer are expected to pass the functionality tests.

B. Setting the Data Mining Goals

As mentioned above, the metrology information is used as a source of feedback to improve and confirm top quality operation. Currently, the process engineers acquire some knowledge pertaining to interactions among various process parameters (e.g., exposure duration versus etching duration) that enables them to manually set the tools, and accordingly reduce the process variability. For instance, the exposure and etching durations are inversely related (within a bounded range). Thus, if the exposure duration is too large, the process engineers could reduce the etching duration; thus reducing variability in the shape of the grooves formed during the etching process. Thus, it has been recognized by the process engineers that discovering *additional* complex interactions among the *input* lithographic process parameters (including parameters such as equipment types, processing dates, product geometrical features, etc.) and the *output* CD and overlay measurements will help to further improve the lithographic process, and to semi-automate the feedback process as shown in Fig. 1. To this end, a data mining methodology has been explored, with the ultimate goal of developing an integrated closed-loop CD control that will provide a wide range of high-level "process knowledge" as illustrated in the lower box of Fig. 3 (e.g., if the resist bottle was opened more than two days ago, than the exposure time should be reduced by 10 ms). The data mining methodology is outlined in the following section.

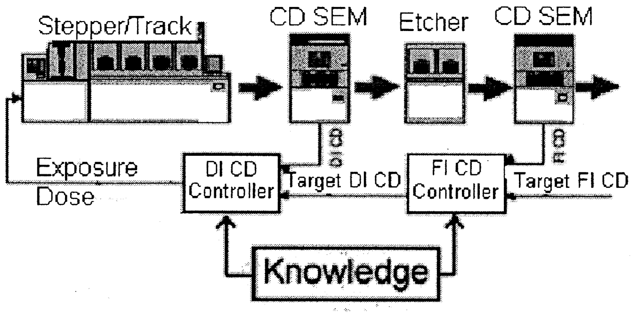

Fig. 1. Closed-loop CD Control.

III. Introduction to Decision Tree Induction

A. Introduction

In this paper, we employ a data mining method (i.e., decision tree classification) that falls under the category of *classification*. To illustrate, consider that each data item, which corresponds to a list of stepper input variables (i.e., a parameter setting of the stepper), is classified to several categories according to the output variable representing the CD's metrological data. Given a historical dataset of wafer input variables and their corresponding CD classes, the classification method can identify the CD class to which a *new* set of stepper input variables is most likely to fit.

In practice, one class is more desirable than the others (e.g., the class with the smallest variability in the CD measurements). We would like to identify the operating conditions of the stepper that *consistently* derive that class. Since the classification is performed based on experimental data, it means that the model generated for that class (e.g., a decision tree) already takes into account any physical constraints on the stepper's input parameters (e.g., maximum exposure time for a given resist thickness). An on-line or off-line control procedure will use the classification model to determine the input parameters of the stepper that generate the lowest CD variability, and will tune the stepper accordingly.

B. Decision Trees

In this section, we provide a brief description of the decision tree classification method employed in this paper. A decision tree algorithm uses training examples (based on historical data) to construct a classification model, which describes the connection between classes and attributes. Once it has learned, the classification model can classify new, unknown instances.

In this paper, a highly effective and widely used classification-based algorithm called C4.5 [7], [13], [14] is employed. C4.5 employs a top-down, greedy construction of a decision tree. It begins with checking which attribute (*input variable*) should be tested at the root of the tree. Each instance attribute is evaluated using a statistical test (called *information gain*) to determine how well it classifies the training examples. The best attribute is selected and used as the test at the root node of the tree. A descendant of the root node is then created for each possible value of this attribute (*discrete* or *continuous* values are allowed), and the training examples are sorted to the appropriate descendant nodes. The entire process is then repeated using the training examples associated with each descendant node to select the best attribute to test at that point in the tree. The above algorithm for constructing a decision tree is detailed and illustrated in Appendix A.

The generated decision tree can be used to classify a particular data item by starting at the root of the tree, and moving through it until a terminal node (leaf) indicating a class is encountered (only discrete output values are allowed). Each nonterminal node represents a test or decision to be carried out on a single attribute value (i.e., input variable value) of the considered data item, with one branch and subtree for each possible outcome of the test. When a terminal node is reached, a decision is made. At each nonleaf decision node, the attribute (i.e., input variable) specified by the node is tested, which leads to the root of the subtree corresponding to the test's outcome.

By placing the attributes with higher information gain closest to the root, the algorithm favors selecting shorter trees over longer ones. The algorithm can also backtrack to reconsider earlier choices by using a pruning method called *rule post-pruning* [13], [14]. The "rule post-pruning" method is utilized in order to overcome the *overfitting* problem, which often arises in learning tasks.

C. Understanding the Model

The decision trees are read from top-down. When there is a decision to be made, an indentation in the structure will be noticed. The tree employs a case's attribute values to map it to a leaf designating one of the classes. Every leaf of the tree is followed by "(n)" or "(n/m)." For instance, a data item that is assigned (according to its attribute values) to the leftmost leaf of the decision tree presented in Fig. 2 is classified as "500-ms exposure times (71/0)" for which n is 71 and m is 0 (for brevity, 71/0 is written as 71). A simple majority classifier is applied if a data item can be classified to more than one class. The value of n is the number of cases in the historical dataset that are mapped to this leaf, and m (if it appears) is the number of them that are classified incorrectly by the leaf. Consequently, the leaf's prediction accuracy is estimated by the Laplace ratio $((n-m+1)/(n+2))$.

The performance of the extracted decision trees on the training cases is further analyzed by deriving the *confusion matrix*. The confusion matrix is of size $n \times n$ (n is the number of different classes), where each of its elements n_{ij} defines the number of training cases that have actual class i, and are classified wrongly as class j. This enables to evaluate the frequency of false alarms (i.e., assigning a case to a wrong class) from the classification system. The total number of misclassifications (obtained from summing the nondiagonal elements in the confusion matrix) is used to evaluate the *error rate* of the classification system. For instance, a decision tree, which misclassifies 6 of the 1054 given cases, has an error rate of 0.57%.

A decision tree can also be represented as a set of if–then rules by creating one rule for each path from the root node to a leaf node. The engineers involved in the lithographic process have indicated that representing the extracted knowledge by "if–then" rules is generally easier to understand than the tree-based representation. The set of if–then rules, which is *equivalent* to the tree-based representation in Fig. 2, is shown in Table I:

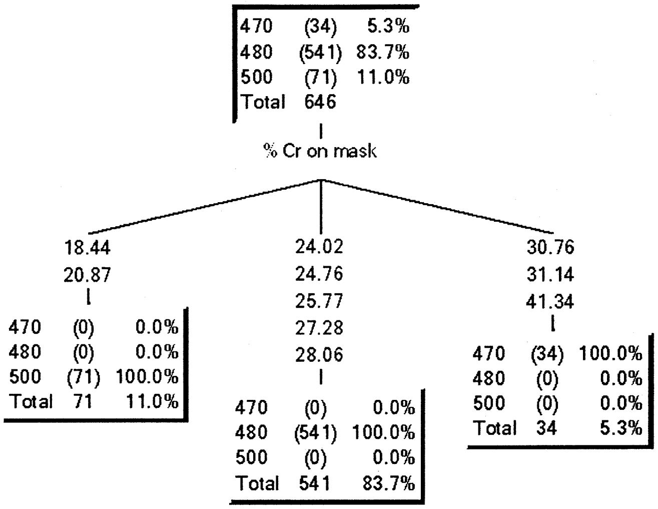

Fig. 2. Decision tree for the relation between exposure times and %chrome.

TABLE I
EQUIVALENT RULE-BASE REPRESENTATION TO DECISION TREE OF FIG. 2

Rule 1: (71, lift 9.1) **IF** %Cr on mask = [18.44, 20.87]
THEN exp_time = 500 (0.9863)

Rule 2: (541, lift 1.19) **IF** %Cr on mask = [24.02, …, 28.06]
THEN exp_time = 480 (0.9982)

Rule 3: (34, lift 18.9) **IF** %Cr on mask = [30.76, …, 41.34]
THEN exp_time = 470 (0.9722)

In the above representation, each rule consists of [22]: 1) An arbitrary rule number (from the leftmost leaf first), which serves to identify the rule; 2) One or more conditions that must all be satisfied if the rule is to be applicable; 3) A class predicted by the rule; 4) Statistics "(n, lift x)" or "(n/m, lift x)," which summarize the performance of the rule; and 5) A value between 0 and 1, which indicates the rule's prediction *accuracy*.

Similarly to a leaf, n is the number of training cases covered by the rule (that is, each training case satisfies all of the rule's conditions), and m (if it appears) shows how many out of them do not belong to the class predicted by the rule. The rule's prediction *accuracy* (or confidence) is estimated by the Laplace ratio $((n-m+1)/(n+2))$. The "*lift x*" label is the estimated accuracy of the rule divided by the *prior probability* of the predicted class.[3] This is because predicting a class that is less prevalent in the population of data items is considered more difficult. The prior probability for class j is computed as $(n_j+1)/(\sum_{j=1}^{c} n_j + c)$ where c is the number of classes and n_j is the number of training cases that belong to class j.

When a rule set is used to classify a new data item, it may happen that several *conflicting* rules can be applied (i.e., different classes are predicted). In this research, each applicable rule votes for its predicted class with a voting weight that is

[3] The *prior probability* of the predicted class is estimated by the frequency of historical data items that are classified as the predicted class.

equal to its prediction *accuracy* value (confidence), the votes are aggregated, and the class with the highest total vote is chosen as the final prediction.

Limitations of decision tree induction: 1) Since they seek for an ordered combination of attributes, from the most important—on the top of the tree—to the least influencing, there are some patterns of interactions and combinations of attributes which can not be captured with decision trees; 2) The algorithm is computationally demanding, which means that in large datasets, tradeoffs have to be made between accuracy and runtime, such as subsampling the records, or inspecting only part of the possible relations; 3) Numeric fields have to be broken into ranges, which means that some patterns will not be inspected, and that there are missing data in some of the ranges.

IV. Experiments and Extracted Knowledge

A. Feature Selection and the Experimental Setting

The geometries of metal layers are complex, and small manufacturing variations can easily interact with manufacturing variations of the lower layers, thus affecting the device's functionality. Moreover, the exposure conditions of metal layers may be set almost individually for each product in order to control the CDs after develop and after etch. Consequently, it has been decided to collect data associated with the top metal layer in order to understand the relationship between process parameters, noise, and final CD measurements.

A dataset that records the successful production of 1054 lots during the period May 2000 to March 2001 has been analyzed. The dataset records the history of 13 different 0.7-micron products.

The *input* variables (attributes or features) include the following information types: 1) product type; 2) lot label; 3) processing steps involved in the top metal layer production along with their processing dates (the wafer fab involves at most 45 such processes). For instance, the fabrication of a certain product involves only 43 of the processing steps; 4) The percentage of chrome on the mask (e.g., 24.76%); 5) The stepper field size in the x and y dimensions (e.g., $x = 16.92$ mm, $y = 18.92$ mm; 6) product CD as provided by the mask shop (e.g., 0.69 micron); 7) stepper exposure time as applied to the specific product (e.g., 470 ms); 8) stepper focus offset as applied to the specific product (e.g., -0.4 micron); and 7) daily reflectance measurements from the metal sputter related to the uniformity of the metallization process (e.g., 58.5%).

The *output* (or dependent) variables include the following: 1) average and standard deviation values of both the high and low topography features *before* and *after* the etch process; and 2) The average and maximum values of the alignment offset results in the x and y directions (measured *after* the etch process).

In the dataset, 52 081 attribute values (about 50%) were missing as a result of the coding structure of the processing steps (i.e., less processes for some of the products), and the fact that some measurements were skipped when there were consecutive same product lots. In general, there are three common strategies to handle missing attribute values: 1) eliminate the record from the experiment—which results in loss of information; 2) fill each missing value with the most prevalent value of that attribute (see [7], [13]–[15]); 3) define another category for the missing value. The third option was used with the software.[4]

In applying the decision tree induction algorithm (see Section III), the records in the database have been randomly selected such that 66% of them have been used for constructing the decision tree. The remaining records have been used estimating the *prediction accuracy* of the constructed model. In addition, aggressive tree pruning (see Section III) has been selected in order to obtain concise classification models that can be effectively used and interpreted by the production engineers. Each experiment has been repeated five times to check for the consistency of the predictions.[5] The maximum and minimum prediction accuracy is reported in the next subsections.

B. Illustration of Experimental Results[6]

1) Diagnostic Analysis: The purpose of this set of experiments is to detect anomalies in the production process that can be further analyzed and corrected by the production engineers. For each of the output variables, a decision tree model has been constructed. In total, seven processes have been detected to have strong influence over the output measurements. In all cases, the first level of the decision trees has been found to be the processing periods where lots begun the corresponding process. The second level of the decision trees has been found to be either the lot labels or the identification of specific machine. The prediction accuracy of the generated decision trees has been relatively good (i.e., 86% to 98%).

To illustrate, consider Fig. 3 that presents the generated decision tree for the output attribute "*maximum* alignment offset *in x axis*." The decision tree is read from top-down. Each tree-node reports the average, standard deviation, and number of cases mapped to this node. Starting from the root of the tree, it can be seen that the average and standard deviation associated with the "maximum alignment offset in x axis" output are 0.04 and 0.09 microns, respectively. These estimates have been computed based on data from 578 records involved in the construction of the decision tree. The first decision node is associated with the "*expose contact mask*" process (i.e., the first visit to the stepper). A descendant of the first node is created for each possible range of dates where lots begun that process. More specifically, three distinct periods, which best classify the training records, have been identified: {June 1, 2000–July 5, 2000}, {July 5, 2000–July 11, 2000}, and {July 11, 2000–March 11, 2001}. The 38, 46, 10 lots, respectively, that have passed through the "expose contact mask" process on those periods had a significantly different average and standard deviation for their "maximum alignment offset in x axis" output variable.

[4] The KnowledgeSEEKER commercial software (available from www.angoss.com) was used for constructing the decision trees on a Pentium II 166 MHz computer. It offers several algorithms for decision trees construction, based on slightly different tree growing and pruning principles. The Entropy method, which is most similar to the C4.5 was used in the experiments.

[5] As an exploratory research, there was no need to compute the standard deviation of the prediction accuracy

[6] Due to the commercial confidentiality, only part of the extracted knowledge has been shown.

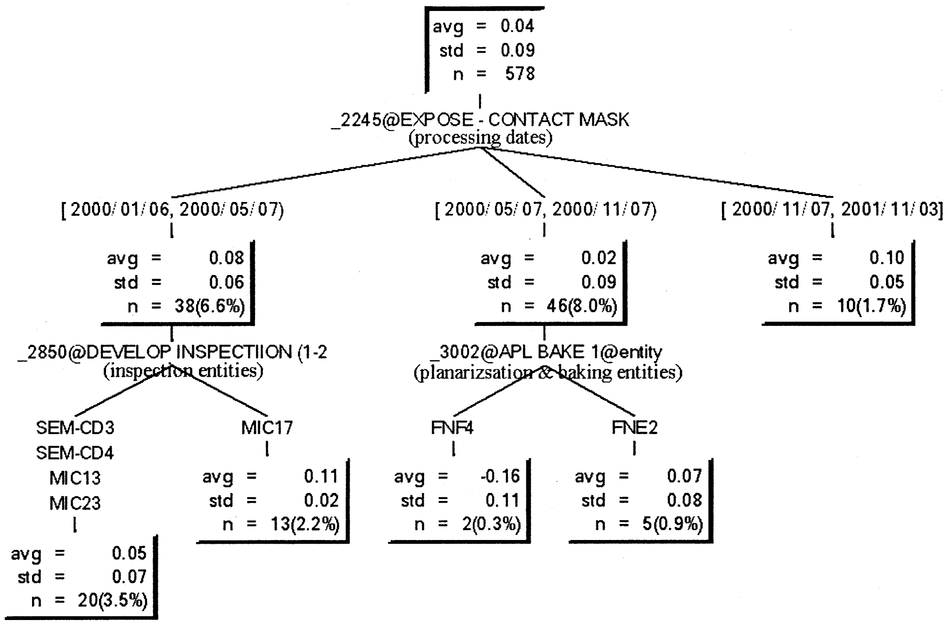

Fig. 3. Decision tree for the "maximum alignment offset in x axis" output variable.

It can be seen that the 0.08-microns average value of the output variable in the period $\{$June 1, 2000–July 5, 2000$\}$ is significantly higher than the overall average of 0.04 micron. This suggests that further analysis of the "expose contact mask" process in this period is required. Examining the second decision node shows that the lots processed in the period $\{$June 1, 2000–July 5, 2000$\}$ can be further classified by the *develop inspection* process (i.e., the inspection of the top metal layer *before* the etching process). A descendant of this node is created for each possible "entity" for the *develop inspection* process (an "entity" is the identification of a specific machine—from several alternative machines that can do a certain process). More specifically, there are two distinct groups of inspection instruments: $\{$SEM-CD3, SEM-CD4, MIC13, MIC23$\}$ and $\{$MIC17$\}$—microscopes of different technologies and maker types—which are responsible for the different values of the output attribute "*maximum* alignment offset *in x axis*." The 13 lots that have passed through the inspection device MIC17 (a microscope) at that period had a high average value of 0.11 microns. This high average value suggests that the inspection microscope MIC17 was not correctly calibrated at those processing dates; a detection that has initiated a corrective action by the process engineers. It is noted that the classification by the "APL bake" process (i.e., "third layer soft bake and planarization") is based on a small number of instances. Thus, additional instances are required if valuable decision-making information is expected based on the "third layer soft bake and planarization" process performance.

2) Predictive Analysis: The extracted knowledge illustrated in Section VI-B1 is valuable mainly for diagnostic tasks since it involves causal interrelationships between historical processing dates and lot classes. In this section, we illustrate the use of data mining in generating knowledge that can be utilized for *prediction* purposes. Predictive models will help to improve *future* lithographic process settings. An example of such knowledge has already been presented in Fig. 2, which shows a decision tree for the strong relation between the stepper exposure time and the percentage of chrome on the mask.

In order to obtain knowledge of a predictive type, the target dataset has been re-defined in terms of the attributes of interest to the process engineers. For instance, the attributes representing the processing dates and lot identification have been removed from the dataset. Also, the data corresponding to the continuous output variables has been transformed into several discrete "*quality*" classes by employing the following method. Let μ (respectively σ) denote the estimated mean value (respectively standard deviation) of a particular output variable. Then, an output variable value is uniquely assigned to one of the following disjoint classes[7]: $[\mu-\sigma,\mu+\sigma], [\mu-2\sigma,\mu-\sigma] \cup [\mu+\sigma,\mu+2\sigma], [\mu-3\sigma,\mu-2\sigma] \cup [\mu+2\sigma,\mu+3\sigma]$, and $[-\infty,\mu-3\sigma] \cup [\mu+3\sigma,\infty]$, which will be termed A, B, C, D, respectively. This partitioning procedure has been found to be consistent with the SPC methods used by the process engineers. To illustrate, any average value of the high topography features after the etch process is classified to one of the following classes: [0.0 microns, 0.02 microns], $[-0.01, 0.0] \cup [0.02, 0.03], [-0.02, -0.01] \cup [0.03, 0.04]$, and $[-\infty, -0.02] \cup [0.04, \infty]$.

Following the data preprocessing, data mining has taken place for extracting patterns from data by generating a decision tree for each output variable.

To illustrate, consider the decision tree generated for the output attribute "*after etch CD of high features*." Table II presents the set of if–then rules, which is *equivalent* to the tree-based representation. Examining the rules reveals several

[7] Since the output CD was in the narrow range (0.69 microns, 0.72 microns) across products, it has been decided to ignore the product-related differences in σ and μ.

TABLE II
EQUIVALENT RULE-BASE REPRESENTATION FOR THE DECISION TREE GENERATED FOR THE OUTPUT ATTRIBUTE "*AFTER ETCH CD OF HIGH FEATURES*"

```
Rule_1   IF  Third layer coating & planarisation entity = {APL1/A, APL2/C,APL3/A}
            and IF Top layer metal etch cleaning entity = {MATRIX3,MATRIX4}
                  THEN class A =100.0%;
            Else  IF Top layer metal etch cleaning entity = {MATRIX1,MATRIX2}
                    THEN class A =36.4%; class B=63.6%;

Rule_2   IF  Third layer coating & planarisation entity = APL2/A
             and IF Third layer plasma etch cleaning entity = {MATRIX4,MATRIX5}
                   THEN class A=100.0%;
             Else IF Third layer plasma etch cleaning entity = MATRIX1
                     THEN class B =100.0%;

Rule_3   IF  Third layer coating & planarisation entity = {APL3, APL3/C, APL4/A ,APL4/C}
             and IF Second layer metal etch cleaning entity = {MATRIX1,MATRIX3,MATRIX4}
                   THEN class A =100.0%;
             Else IF Second layer metal etch cleaning entity = MATRIX2
                   THEN class A =33.3%; class B=66.7%;
```

causal interrelationships between the "third layer coating and planarization" and "top layer metal etch" processing steps that best predict the resulting "quality" class. For instance, if a certain lot had its "third layer coating and planarization" process at the "entity" APL2/A followed by the its "top layer metal etch" process at the "entity" MATRIX1, then it is predicted from rule 2 that the resulting "*after etch CD of high features*" will belong to the "quality" class B. That is, the resulting measurement is expected to fall within the range $[-0.01$ micron, 0.0 micron$]$ $\cup$ $[0.02$ micron, 0.03 micron$]$. The prediction accuracy of the extracted rules has been found to be 82%, and the frequency of false alarms (i.e., the probability of assigning a case to a wrong class) was evaluated by the confusion matrix (see Section III-C) to be less than 1.2%.

Some other findings include: 1) The interaction between the processes "resist coating and planarization of the third layer" and "third layer plasma etch cleaning" has been detected to also have influence over the "*after etch CD of low features*" output variable; 2) Daily reflectance measurements from the metal sputter have been detected to predict the "*after etch CD of low features*" output variable; 3) The mask parameter "stepper field size" has been found to affect the "average alignment offset in the x axis" output variable; 4) The mask parameter "percentage of chrome on the mask" has been found to affect the "maximum alignment offset in the x axis" output variable and interact with the process setting parameters "exposure time" and "focus setting;" 5) Two cleaning entities have been found to affect the "alignment offset in the y axis" output variables. The predictive accuracy of the generated decision trees has been within the range [73%, 100%].

These discoveries have been generated to be used by the process engineers in optimizing the machines' operating conditions (e.g., based on mask attributes), in real-time routing and scheduling of lots (e.g., skipping over machine interactions that potentially increase the process variability), and in setting the layout of future lithographic processes. Note that unlike the neural network models of [10], [21], which are "black-box" models that need to be blindly trusted by the process engineers, rule base tables such as Table II are self explanatory; thus the process engineers can understand (and sometimes reject) the cause-effect relations provided by the decision trees.

In an attempt to improve the above prediction accuracy, several composite classifier architectures have been constructed. Combining the predictions of a set of classifiers, even simple ones, has been shown to be an effective way to create composite classifiers that are more accurate than any of the component classifiers (e.g., [9], [16]–[18]). In general, our experiments show that the accuracy of the classification algorithm (even if very strong on its own) can be increased, merely by combining its predictions with those made by another classifier (even if a weak one). However, only little improvement (up to 4%) in accuracy has been observed. Since implementing composite classifiers is prohibitively expensive, it is concluded that composite classifiers may not be the proper strategy for the lithographic process under study.

V. CONCLUSION

The process of wafer fabrication is highly complex. It has been estimated [19] that up to 80% of yield losses in high volume integrated circuits production can be attributed to random, equipment related process-induced defects. In leading fabs, large volume of data is collected from the production floor under the dictum "collect now and analyze later." Yield models are developed to automate the identification of out of control manufacturing conditions. This is used as feedback to improve and confirm top quality operation.

Although standard yield improvement procedures are useful, the volume of the collected data makes it impractical to explore and detect intricate relations between the parameters of different processing steps using standard statistical procedures. Also, comprehensive manual analysis of the voluminous data becomes prohibitive, especially when on-line monitoring and control of the manufacturing process are considered.

Data mining can be seen as a supportive vehicle for determining causal relationships among "internal" factors associated with the semiconductor manufacturing processes and "external" elements related to the competitiveness of the semiconductor manufacturing company (e.g., production indices, performance parameters, yield, company goals).

In this paper, a decision tree induction algorithm has been proposed as a means of mitigating the above mentioned limitations, and a successful application to an actual photolithographic process has been presented. The developed models, which have been validated by the process engineers, have been useful in 1) detecting production periods and specific machines (from several alternative machines) where the variances of the key quality characteristics (i.e., CD and overlay metrologies) are high; thus, suggesting directions for improving the process (e.g., recalibrating the semiconductor process equipment, establishing equipment routing preferences); 2) predicting the probability distribution of key quality characteristics (summarized by the average and standard deviation measures) by determining the causal relationships among "internal" factors associated with the lithographic manufacturing process; 3) optimizing the recipes of specific processes; e.g., setting the stepper exposure time based on the percentage of chrome on the mask; and 4) allocating defect budgets (see [20]) to semiconductor process equipment.

It is expected that with larger datasets and faster computing capabilities, it would be possible to semi-automate the data mining process, so that the process engineer can benefit from more comprehensive interaction models. The ultimate goal of our research is to develop an integrated yield management system, which utilizes data mining methodologies as a supportive vehicle for closed-loop system-level control. To illustrate, the process recipe of later processing stages may be modified in order to offset drifts of key quality characteristics in earlier stages of production; e.g., modifying the etch recipe based on measurements of the develop inspection.

An effective utilization of data mining methodologies should also handle false-alarms, mis-calibrations and drifts in the training data. Indeed, most statistical and machine-learning algorithms assume that the training data is a random sample drawn from a stationary distribution. The process engineer can reduce the probability of false alarms by aggressive tree pruning (at the cost of possibly missing some rare interactions), or by considering the accuracy of each activated rule. Unfortunately, the underlying processes generating the data in a semiconductor manufacturing environment are not truly stationary and change over time. Consequently, algorithms for extracting decision trees from time-changing data streams from very large databases are required (see [23] for review). The incorporation of these algorithms is expected to improve the effectiveness of the data mining process, and will be the subject of future research.

System-level control encompasses many tools and processes, and can involve human intervention or be entirely automated; depending on the reliability and potential impact of the decision-making process. Achieving system-level control would benefit from larger datasets (generated by powerful data acquisition systems); faster computing capabilities; and a collaborative effort of process engineers, information technology professionals, and data mining experts. Although system-level control is far down the road in terms of current technical capabilities, it can be realized if proper adaptive learning methodologies, such as data mining, are utilized.

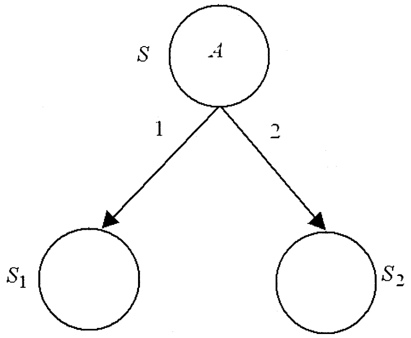

Fig. 4. Node with the inhomogeneous instance set S is replaced by a test node (attribute A) that divides the inhomogeneous instance set S into minimally inhomogeneous subsets S_1 and S_2, according to the information gain criterion presented in (A.2).

Appendix A
A Basic Algorithm for Constructing a Decision Tree [13], [14]

The ID3 and its successor C4.5 follow an information theoretic measure, the information gain, for deciding which attribute to split on at any given point in building a decision tree. The information gain measure is based on the notion of entropy. Given training examples in the dataset S, each of which is assigned to one of c classes, the entropy of the dataset is given by

$$E(S) = -\sum_{c} \left(\frac{n_c}{n}\right) \log_2 \left(\frac{n_c}{n}\right) \tag{A.1}$$

where n is the number of instances in S, and n_c is the number of instances of class c. Notice that (n_c/n) is an estimate for the prior probability that an instance is assigned to class n_c. The entropy measures the impurity of a set of instances; that is, the entropy equals zero when the set is perfectly homogeneous (i.e., all instances are assigned to the same class) and equals one when the set is perfectly inhomogeneous.

At any given point in building the decision tree, a decision is made regarding the attribute to split on. Assume that attribute A is selected as shown in Fig. 4.[8] In Fig. 4, S represents the current set of instances, and $S_v (v = 1, 2)$ represents the set of instances for which the attribute A receives the value v. The information gain $\mathrm{Gain}(S, A)$ is defined as follows:

$$\mathrm{Gain}(S, A) = E(S) - \sum_{v \in \mathrm{values}(A)} \frac{|S_v|}{|S|} \mathrm{Entropy}(S_v). \tag{A.2}$$

The information gain is computed for each attribute, and the attribute with the highest information gain is selected. Equivalently, the attribute for which the average entropy after splitting is the lowest [second term in (A.2)] is chosen.

[8] For simplicity of presentation, we assume that each attribute has only two values.

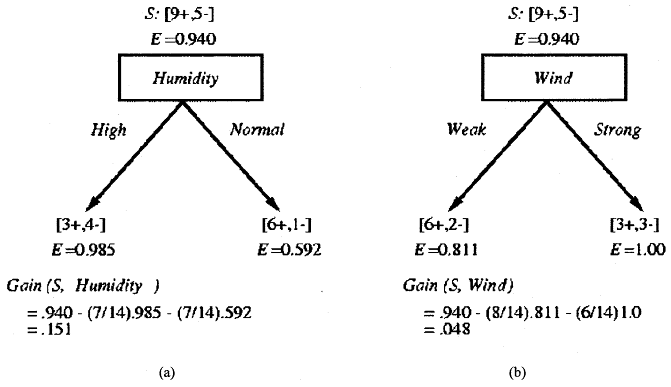

Fig. 5. Selecting the "best" attribute to split on according to the information gain criterion (adapted from [13]). (a) Split on the attribute *Humidity*. (b) Split on the attribute *Wind*.

The above procedure is repeated until each leaf node is populated by as homogeneous an instance set as possible. More precisely, a leaf node with an inhomogeneous instance set is selected, and is replaced by a test node (an attribute) that divides the inhomogeneous instance set into minimally inhomogeneous subsets, according to the information gain criterion presented above.

Example A.1: Assume that the current set of instances includes nine instances that are classified as positive and five instances that are classified as negative. Fig. 5(a) shows the tree that is constructed if we decide to split on the attribute Humidity. Fig. 5(b) shows the constructed tree if a decision to split on the attribute Wind is made. The calculations of the respective information gains are also specified in the figure. It can be seen that the attribute "Humidity" achieves a higher information gain, and therefore is selected for splitting.

ACKNOWLEDGMENT

The authors would like to thank the engineers at Tower Semiconductor Inc. for their useful discussions and helpful comments regarding the manufacturing process, and for sharing their production datasets.

REFERENCES

[1] P. V. Zant, *Microchip Fabrication: A Practical Guide to Semiconductor Processing*, 3rd ed. New York: McGraw-Hill, 1997.

[2] S. P. Cunningham, C. J. Spanos, and K. Voros, "Semiconductor yield improvement: Results and best practices," *IEEE Trans. Semiconduct. Manufact.*, vol. 8, pp. 103–109, 1995.

[3] R. C. Leachman and D. A. Hodges, "Benchmarking semiconductor manufacturing," *IEEE Trans. Semiconduct. Manufact.*, vol. 2, pp. 158–169, May 1996.

[4] W. Kuo and T. Kim, "An overview of manufacturing yield and reliability modeling for semiconductors products," *Proc. IEEE*, vol. 87, pp. 1329–1344, Aug. 1999.

[5] S. S. Anand and A. G. Buchner, *Decision Support Using Data Mining*. New York: Pitman, 1997.

[6] R. Brachman and T. Anand, "The process of knowledge discovery in databases: A human-centered Approach," in *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, S. P. Amith, and R. Uthurusamy, Eds. Cambridge, MA: MIT Press, 1996.

[7] D. Braha, Ed., *Data Mining for Design and Manufacturing: Methods and Applications*. Boston, MA: Kluwer Academic, 2001.

[8] B. S. Kang, J. H. Lee, C. K. Shin, S. J. Yu, and S. C. Park, "Hybrid machine learning system for integrated yield management in semiconductor manufacturing," *Expert Syst. with Applicat.*, vol. 15, pp. 123–132, 1998.

[9] D. Braha and A. Shmilovici, "Data mining for improving a cleaning process in the semiconductor industry," *IEEE Trans. Semiconduct. Manufact.*, vol. 15, pp. 91–101, Feb. 2002.

[10] T. S. Kim and G. S. May, "Intelligent control of via formation by photosensitive BCB for MCM-L/D applications," *IEEE Trans. Semiconduct. Manufact.*, vol. 12, pp. 503–515, Nov. 1999.

[11] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, S. P. Amith, and R. Uthurusamy, Eds. Cambridge, MA: MIT Press, 1996, pp. 1–36.

[12] M. J. Berry and G. Linoff, *Data Mining Techniques*. New York: Wiley, 1997.

[13] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[14] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[15] P. Adriaans and D. Zantinge, *Data Mining*. New York, NY: Addison-Wesley, 1996.

[16] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 66–75, Jan. 1994.

[17] H. Drucker, "Improving regressors using boosting techniques," in *Proc. Int. Conf. Machine Learning*, 1997, pp. 107–115.

[18] D. Wolpert and W. Macready, "An efficient method to estimate bagging's generalization error," *Machine Learning*, vol. 35, no. 1, pp. 41–55, 1999.

[19] V. Sankaran, C. M. Weber, K. W. Tobin Jr., and F. Lakhani, "Inspection in semiconductor manufacturing," in *Webster's Encyclopedia of Electrical and Electronic Engineering*. New York: Wiley, 1999, vol. 10, pp. 242–262.

[20] D. L. Dance, D. Jensen, and R. Collica, "Developing yield modeling and defect budgeting for 0.25 mm and beyond," *Micro*, vol. 16, no. 3, pp. 51–61, 1998.

[21] E. A. Rietman, A. Whitlock, M. Beachy, A. Roy, and T. L. Willingham, "A system model for feedback control and analysis of yield: A multistep process model of effective gate length, poly line width, and IV parameters," *IEEE Trans. Semiconduct. Manufact.*, vol. 14, pp. 32–47, Jan. 2001.

[22] J. R. Quinlan, *C4.5 Programs for Machine Learning*. Los Altos, CA: Morgan Kaufman, 1993.

[23] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proc. Seventh Int. Conf. Knowledge Discovery and Data Mining*. San Francisco, CA, 2001, pp. 97–106.

Dan Braha is an affiliate of the New England Complex Systems Institute (NECSI), and a senior engineering faculty member at Ben-Gurion, University, Beer-Sheva, Israel. He has been a Visiting Professor at the MIT Center for Innovation in Product Development (CIPD), and a Research Associate in the Department of Manufacturing Engineering at Boston University. One of his primary areas of research is engineering design and manufacturing. His research within engineering design focuses on developing methods to help the designer move from the conceptual phase to the realization of the physical device. He has developed a mathematical theory—the Formal Design Theory (FDT). He has published extensively, including a book on the foundations of engineering design with Kluwer Academic Publishers, and an edited book on data mining in design and manufacturing; also with Kluwer. He serves on the editorial board of AI EDAM, and was the editor of several special journal issues. He has also served on executive committees and as chair in several international conferences. Currently, he aims to advance the understanding of Complex Engineered Systems (CES); arrive at their formal analysis; as well as facilitate their application.

Armin Shmilovici received the B.Sc. and M.Sc. degrees in electrical and electronics engineering and the Ph.D. degree in industrial engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1986,1991,1997, respectively.

In 1998, he joined the faculty of Ben-Gurion University, Beer-Sheva, Israel, and he is currently with the Faculty of the Department of Information Systems Engineering. His research interests include fuzzy systems, information systems, and control of production processes. He published more than 30 articles in international conference proceedings and in publications such as Fuzzy Sets and Systems, and IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING.