Discovery of Resource-Oriented Transition Systems for Yield Enhancement in Semiconductor Manufacturing

Minsu Cho, Gyunam Park, Minseok Song, *Member, IEEE*, Jinyoun Lee, Byeongeon Lee, and Euiseok Kum

***Abstract*—In semiconductor manufacturing, data-driven methodologies have enabled the resolution of various issues, particularly yield management and enhancement. Yield, one of the crucial key performance indicators in semiconductor manufacturing, is mostly affected by production resources, i.e., equipment involved in the process. There is a lot of research on finding the correlation between yield and the status of resources. However, in general, multiple resources are engaged in production processes, which may cause multicollinearity among resources. Therefore, it is important to discover resource paths that are positively or negatively associated with yield. This article proposes a systematic methodology for discovering a resource-oriented transition system model in a semiconductor manufacturing process to identify resource paths resulting in high and low yield. The proposed method is based on the model-based analysis (i.e., finite state machine mining) in process mining and statistical analyses. We conducted an empirical study with real-life data from one of the leading semiconductor manufacturing companies to validate the proposed approach.**

***Index Terms*—Process mining, resource paths, semiconductor manufacturing, statistical analysis, transition systems, yield enhancement.**

I. Introduction

EVIDENCE-BASED decision making through data analysis has a vital role in a manufacturing company [1], since plenty of data has been collected in a manufacturing environment [2]. In the semiconductor manufacturing industry, many efforts have been made to manage and improve yield with the use of data, such as identifying factors that affect the yield [3].

Manuscript received October 18, 2020; revised November 22, 2020, November 23, 2020, and December 7, 2020; accepted December 15, 2020. Date of publication December 18, 2020; date of current version February 3, 2021. This work was supported by the Research Program from Samsung Electronics Company Ltd. *(Corresponding author: Minseok Song.)*

Minsu Cho is with the Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang 37673, South Korea, and also with the Research Institute of Industry and SME Strategy, Korea Institute of Science and Technology, Seoul 06211, South Korea.

Gyunam Park and Minseok Song are with the Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang 37673, South Korea (e-mail: mssong@postech.ac.kr).

Jinyoun Lee, Byeongeon Lee, and Euiseok Kum are with the Mechatronics Research and Development Center, Samsung Electronics Company Ltd., Hwaseong 18381, South Korea.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSM.2020.3045686.

Digital Object Identifier 10.1109/TSM.2020.3045686

Production resources, i.e., equipment involved in manufacturing processes, are one of the critical factors that affect yields [4]. Thus, a handful of research works have been conducted on identifying significant resources that affect yield and analyzing the status of the important resources in order to obtain better productivity [5]. However, most of the works are limited to investigating the relationship between a single resource and yield. Since the resources involved in the production are interconnected and influenced by each other, resource paths (i.e., the combination of resources in manufacturing processes) should be considered.

As shown in Fig. 1, semiconductor manufacturing has lengthy and complicated processes [3], [6]. All wafers pass through hundreds of process steps, each of which has several alternative equipment. Thousands of machines are involved in a whole manufacturing process, and each equipment is composed of multiple chambers performing the same process step; thus, there are countless combinations of machines at the chamber level [3]. Therefore, we should deal with the problem of multicollinearity among the combinations of resources [7], [8]. For example, in Fig. 1, $chamber_{i,1,1}$ is identified as a critical resource that results in a high-yield wafer (i.e., it has a substantial impact on high-yield). However, the corresponding machine at the same time belongs to another path which produces a wafer of low yield. To this end, it is important to investigate an association of yield with the sequence of resources rather than with a single resource involved in the production.

In this article, we propose a new methodology for discovering a resource-oriented transition system (TS) model in a semiconductor manufacturing process. Specifically, it aims to identify the resource paths resulting in the high and low yield, which are utilized to predict the future outcomes of given wafers. The proposed methodology consists of four phases in total: data preparation, resource-oriented TS model construction, model simplification, and evaluation and interpretation. In a nutshell, we first generate a TS model from manufacturing data, and then the integrated model is derived by visualizing the best and worst resource paths on the model. After that, the integrated model is simplified based on the indispensable resources derived from statistical analyses that are significant to yield. Lastly, the discovered model is evaluated and interpreted.

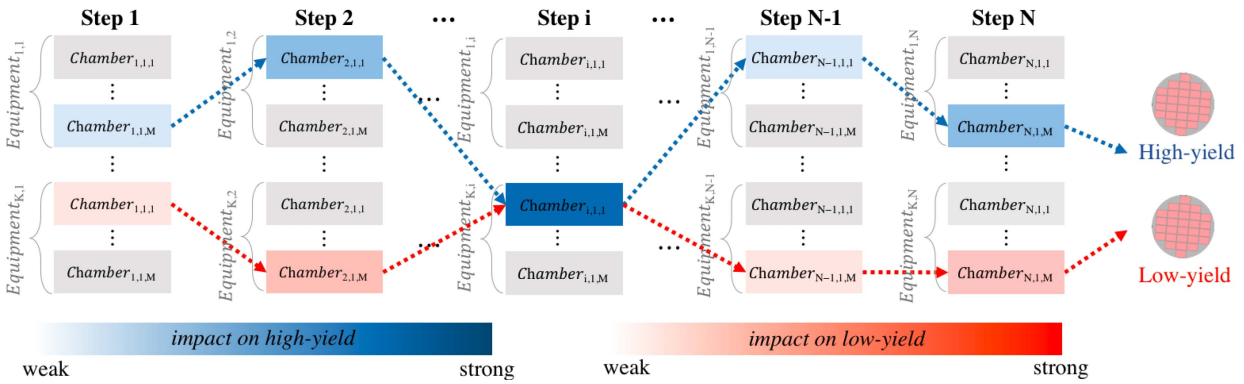

Fig. 1. An example of the complicated semiconductor manufacturing process.

To show the usefulness of the proposed methodology, we provide an empirical study in a large Korean manufacturing corporation in depth.

II. Background

In the semiconductor manufacturing process, one of the multistage manufacturing processes with multiple process steps, there have been many efforts to improve yield management. Among several studies, two research streams are relevant to our research: (1) physical modeling approaches that derive a mathematical model based on principles that characterize manufacturing processes and (2) data-driven modeling research that investigates insightful patterns using a massive historical manufacturing data collected from information systems including manufacturing execution systems (MESs). Therefore, we reviewed existing approaches engaged in these two streams.

Since the 1990s, multistage manufacturing process physical modeling is one of the disciplines that many researchers have devoted to yield enhancement. The main focus was on anomaly (i.e., variation) propagation that could appear in the manufacturing process and developing a mathematical model for error propagation between process steps. To this end, the most popular model was the stream-of-variation method using the concept of a state-space model. In [9], the researchers firstly have proposed this method for sheet metal assembly modeling, and the relevant principle is as follows: $x_i = A_{i-1}x_{i-1} + B_i u_i + w_i$ and $y_i = C_i x_i + v_i$, where the quality index y_i of the step i is derived by C_i which maps the state vector x_i into the quality measurement, and here the state vector x_i consists of the quality deviation propagation from the step $i-1$ to step i (i.e., $A_{i-1}x_{i-1}$) and the anomaly quality deviation in the step i (i.e., $B_i u_i$) [10].

Until now, stream-of-variation methods have been applied, focusing on two fields: assembly processes [11] and machining processes [12]. In building a concrete mathematical model, these methods require the first principles of the relevant manufacturing processes. Specifically, these principles include the possible variations, i.e., definitions of errors, and the associated parameters, i.e., coefficients (e.g., A_i, B_i, and C_i in the above formula). For example, most stream-of-variation models for assembly processes have employed fixture locator errors and common errors between parts [11], whereas fixture errors, datum errors, and machining errors have been utilized for modeling of machining processes [12]. In addition to these approaches, some research has been devoted to modelling a partial sequence of the semiconductor manufacturing process. In [13], [14], the authors presented the stream-of-variation model for multilayer overlay lithography processes in consideration of overlay errors. Specifically, they employed objective errors and grid errors as the first principles, and the physical model for the relevant process was derived.

Although these existing studies use physical modeling for yield enhancement, they do not provide a visualization to penetrate the whole manufacturing process at a glance. In other words, there is a limitation in modeling and visualizing the flow of the process. Also, stream-of-variation methods require a high dependency on the first principles. To this end, a couple of researchers have attempted to derive the relevant information (e.g., coefficients) using measurement data [15]. However, they suffer from low applicability because of large amounts of data and excessive resources and complexity to modeling.

Massive manufacturing data has enabled development of data-driven modeling approaches, including data mining and process control. Specifically, there have been attempts to achieve a specific goal, such as yield enhancement as well as process improvement, demand forecasting, and cycle time reduction [3], [16], [17], [18], [19]. Regarding quality management, there was a method to distinguish abnormal process stages using the Kruskal-Wallis test and decision tree [16]. In [17], the researchers proposed the design-of-experiment data mining method, which resolves inefficiencies of data mining with extensive data, and analyzed the relationship between a specific machine and yield. Also, a study proposed a clustering method for wafer map analysis to identify defect clusters and their connectivity to the yield [19], whereas there was a framework using Bayesian inference and Gibbs sampling to identify abnormal devices, i.e., fault detection [3].

In addition to these data mining approaches, there have been numerous researches to developing process control methods and the relevant systems for yield enhancement [20], [21], [22]. Among them, some researchers proposed a novel approach that enables manufacturing process control with effective data visualization, i.e., geometry process control [20]. It was an impactful approach to identify

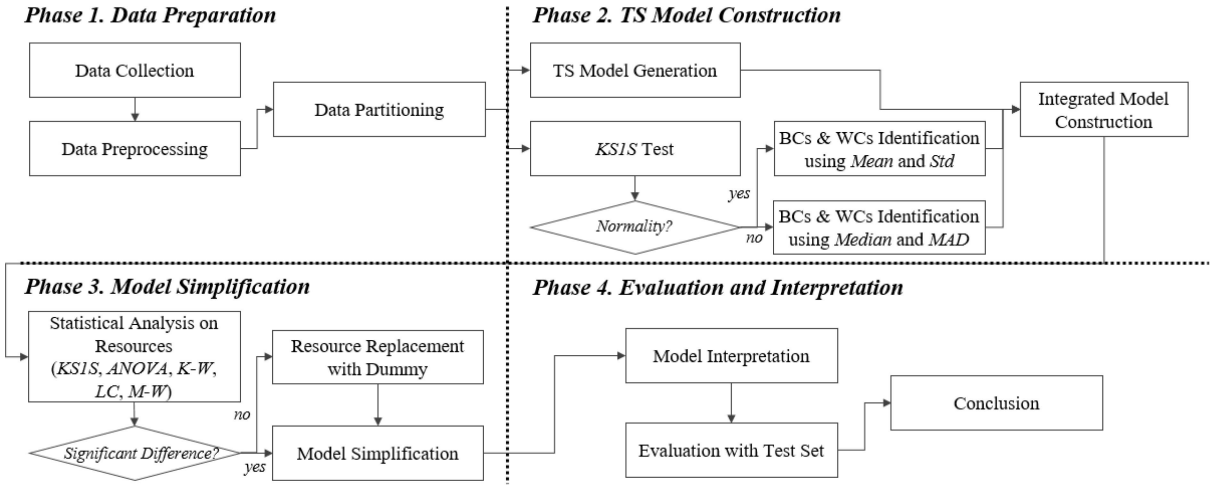

Fig. 2. Overview of the proposed research framework.

the relationship between n-dimensional operational variables and quality measurements.

Despite their contributions to yield enhancement, all these techniques only focused on identifying a model between a critical variable (e.g., a process step or equipment) and the quality. However, they have not dealt with multicollinearity existing in multistage manufacturing processes. Thus, they require the extension to the sequence or path of variables, a limitation which we resolve in this article.

In summary, the limitations of the reviewed papers for two research streams are as follows: (1) none of the studies deal with effective visualization by combining the semiconductor manufacturing process flows and the quality-related measurements, (2) poor usability results from heavy dependence on first principles and requirement for excessive data and resources, and (3) data-driven approaches that deal with multicollinearity are insufficient. To overcome these limitations, we suggest a resource-oriented transition system, which visualizes the quality measurements. Instead of first principles, we use manufacturing data and build a sound system using an abstraction method that simplifies operations. Furthermore, our approach focuses on deriving resource paths rather than specific resources for improving yields, and by doing so, we resolve all of the above limitations.

III. A Framework for Constructing a Resource-Oriented TS Model for Yield Enhancement

This section describes our approach to derive a resource-oriented TS model for yield enhancement based on manufacturing data.

Fig. 2 provides the proposed research framework in this article. Our method includes four phases: data preparation, resource-oriented TS model construction, model simplification, and evaluation and interpretation. First, manufacturing data is prepared with a series of steps, including data collection, preprocessing, and partitioning. Second, based on the prepared data, the TS model is generated, and the cases (i.e., wafers) recorded that are relatively superior or inferior to others are identified. After that, these two results are combined into one network model. The discovered model, however, may be remarkably complicated because plenty of resources are engaged in the manufacturing process. Therefore, we need further steps to simplify the obtained results. In the third phase, we perform a series of statistical analyses on the performance of resources. It aims to determine highly-impacting resources on the quality of products. Then, the model is simplified based on the result of the statistical analysis. In the last phase, the discovered model is interpreted by discussing with domain experts and evaluated with the test set to identify whether the model has a predictive capability on yield or not.

A. Phase 1. Data Preparation

The data preparation phase aims to construct sufficiently structured data for obtaining valuable analysis results. This research takes the format of event logs utilized in process mining [23]. Definition 1 provides formal definition of event logs. It consists of a series of records for events e of process instances c (i.e., wafers) that are executed by whom (i.e., resources) at what time (i.e., timestamps). Also, it includes the yield as an attribute of process instances $\psi(c)$. Table I provides an example of manufacturing event logs. After collecting event logs for a specific period, the data preprocessing step is performed to improve the accuracy and effectiveness of the data analysis. It includes removing noisy data, identifying outliers, and handling incomplete or error data. In the last step of this phase, the preprocessed data is partitioned into the training and test dataset. Then, the training set is employed to discover a resource-oriented TS model for predicting yield, while the test set is utilized for evaluation of the discovered model from the training set.

Definition 1 (Event, Attribute, Case, Event Log, and Yield): Let E be the event universe. Events may have several attributes (e.g., step, originator, timestamp), and let AN be a set of attribute names. Hence, for any event $e \in E$ and any attribute name $an \in AN : \pi_{an}(e)$ is the value of the attribute an for event e. Let $C = \{c_1, c_2, c_3, \ldots, c_k\}$ be the set of cases and mapped

TABLE I
AN EXAMPLE OF EVENT LOGS

CaseID	EventID	Step	Resource	Timestamp	Yield
Case 1	E1	A	M1	2018-01-01 10:30	0.9
	E2	B	M2	2018-01-01 12:00	
	E3	C	M5	2018-01-01 17:00	
	E4	D	M7	2018-01-02 09:00	
Case 2	E5	A	M1	2018-01-04 09:00	0.8
	E6	B	M3	2018-01-04 12:00	
	E7	C	M6	2018-01-04 15:00	
	E8	D	M7	2018-01-05 13:00	
Case 3	E9	A	M1	2018-01-06 09:00	0.7
	E10	B	M4	2018-01-06 14:00	
	E11	C	M6	2018-01-07 13:00	
	E12	D	M7	2018-01-07 18:00	

into a trace σ, i.e., a finite sequence of events ($\sigma \in E^*$). An event log L is a collection of possible cases. Let $\psi : C \Rightarrow Y$ is a function that obtains yield values recorded for a case. Hence, $\psi(c) \Rightarrow Y$ signifies to obtain the corresponding yield $y \in Y$ for a case c.

B. Phase 2. TS Model Construction

The second phase aims at generating a resource-oriented TS model for yield enhancement based on manufacturing event logs. First of all, a process mining discovery technique is employed to construct a TS model from event logs [23]. The TS model is formalized in Definition 2. The model $TS = (S, E, T)$ is composed of states S, events E, and transitions T. Here, states represent the status in which an event in a process is performed, and transitions refer to the relationship that changes a state as an event is performed in a particular state. This research focuses on resource behaviors in semiconductor manufacturing. As such, the model is generated with states and events representing the resources of process steps. From the model, we are able to identify what resources (i.e., chambers) are involved in a process step or what connection between resources does exist in the model. Fig. 3 gives an example model generated from the sample log in Table I. As shown in the figure, all resource behaviors discovered in the log can be represented as a single network model.

Definition 2 (State Representation Function, Event Representation Function, Transition System): Let L be an event log. Let E^* and E be a set of possible traces and events. A prefix function hd^k is a function that returns the sequence consisting of first k elements from a trace, t. A state representation function $rep^s \in E^* \rightarrow R^s$ is a function that gives a representation from a trace t, where R^s is the set of possible state representations. An event representation function $rep^e \in E \rightarrow R^e$ is a function that gives a representation from an event $e \in E$, where R^e is the set of possible event representations. A transition system TS is defined as a triplet (S, E, T), where $S = \{rep^s(hd^k(t)) | t \in L \wedge 0 \leq k \leq |t|\}$, $E = \{rep^e(t(k)) | t \in L \wedge 1 \leq k \leq |t|\}$, and $T \subseteq S \times E \times S$ with $T = \{rep^s(hd^k(t)), rep^e(t(k+1)), rep^s(hd^{k+1}(t)) | t \in L \wedge 0 \leq k \leq |t|\}$ is the set of states, the set of events, and the set of transitions.

The second step of this phase is to identify the high-yield cases (HCs), and the low-yield cases (LCs). First, we verify

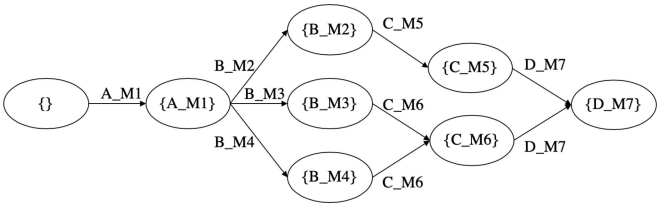

Fig. 3. An example of transition systems based on the example log in Table I.

the normality of the empirical distribution for yield values by employing Kolmogorov-Smirnov one sample test (KS1S) [24]. It investigates whether the empirical yield distribution from the log and the normal distribution are statistically the same or not. The test statistic D is $sup_x|F_n(X) - F(X, \mu, \sigma)|$, where $F_n(X)$ is the cumulative empirical distribution and $F(X, \mu, \sigma)$ is the cumulative normal distribution. Also, the hypothesis test is conducted as follows.

$$H_0 : F_n(X) = F(X, \mu, \sigma); \quad H_1 : F_n(X) \neq F(X, \mu, \sigma)$$

If the case satisfies the normality condition, HCs and LCs are determined based on the mean (μ) and the standard deviation (σ) of the yield distribution. Conversely, when the normality condition is unsatisfactory, the median (M) and the median absolute deviation (MAD) [25] are utilized to establish HCs and LCs. Based on these values, we can create a certain range that represents just a mediocre quality, as provided in (1) and (2). Here, w is the user-specified weight. Based on the acceptable range, the yield values of all cases are evaluated whether they are within or outside the range. If the yield values are higher than the maximum value of the range, the corresponding cases are classified as HCs. On the other hand, if cases have a lower yield than the minimum value of the range, they belong to LCs.

$$\mu - w \times \sigma \leq yield \leq \mu + w \times \sigma \tag{1}$$

$$M - w \times MAD \leq yield \leq M + w \times MAD \tag{2}$$

Suppose the normality condition is satisfied for the cases in Table I. Given $w = 1$, the acceptable range becomes $[0.79, 0.81]$ since $\mu = 0.8$ and $\sigma = 0.01$. Thus, the *Case 1* belongs to the high-yield case since its yield, 0.9, is higher than the maximum value of the range. On the other hand, *Case 3* is a low-yield case as its yield, 0.7, is lower than the minimum value of the range.

Lastly, the TS model and statistical analysis results are combined into one single model. As such, the high-yield and low-yield paths are visualized on the discovered model. Firstly, the high-yield case ratio (HR) and low-yield case ratio (LR) are computed for every arc in the TS model, where HR is the ratio of HCs to the total number of cases on an arc a, and LR is the ratios of LCs to the total number of cases on a. Suppose the total number of cases passing through a is 1000 and 400 of them are HCs and 200 of them are LCs. Then, HR and LR on a are 0.4 and 0.2, respectively.

After that, the summation of these two values is evaluated if it is greater than the predefined threshold to determine whether the flow is paramount in the model. Then, we measure the difference between HR and LR for critical paths. As such, if

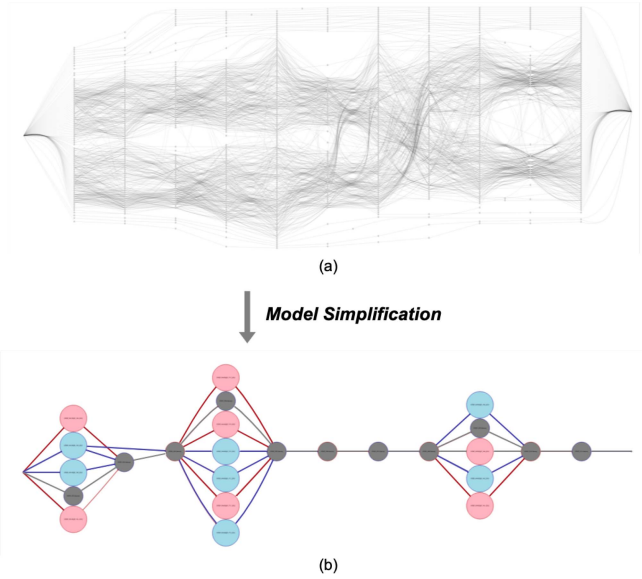

Fig. 4. An example of model simplification from (a) the integrated model to (b) the simplified model.

HR is higher than *LR*, the relevant arc is determined as the part of the paths that lead to the high yield. In the opposite case, it is considered as a piece of the low yield paths. Given a predefined threshold of 0.5, the arc a is evaluated as a critical one since the summation of *HR* and *LR*, 0.6, is more significant than 0.5 and classified as a high-yield path since *HR* is more significant than *LR*.

C. Phase 3. Model Simplification

As depicted in Fig. 4(a), the integrated model has a limitation in that it has considerable complexity as the process gets more extended or more machines are engaged in activities. To overcome the limitation, we need to identify mainline states that have a statistical significance to yield and simplify the model based on them, as shown in Fig. 4(b).

As such, the third phase aims at simplifying the model with the results from the statistical analysis. To this end, we employ a series of statistical techniques. Fig. 5 provides the flow of statistical analysis.

First, the normality of distributions of all resources involved in the same process step is tested with Kolmogorov-Smirnov one sample test (KS1S) [24], and the same test statistic and hypothesis testing are applied as in Phase 2. Following the normality testing result, further steps are divided into two different directions. More specifically, they exploit different statistical techniques; however, each step has the same objectives: (i) it is first analyzed to identify whether the states engaged in the same step have a discrepancy with each other and (ii) after that, if the difference exists, it is determined what states make the quality of products better or worse.

In the case where the distributions of all states follow the normal distribution, analysis of variance (ANOVA) [26], is applied to identify whether all states have the indifferent yield value. The following are the null and alternative hypotheses for ANOVA.

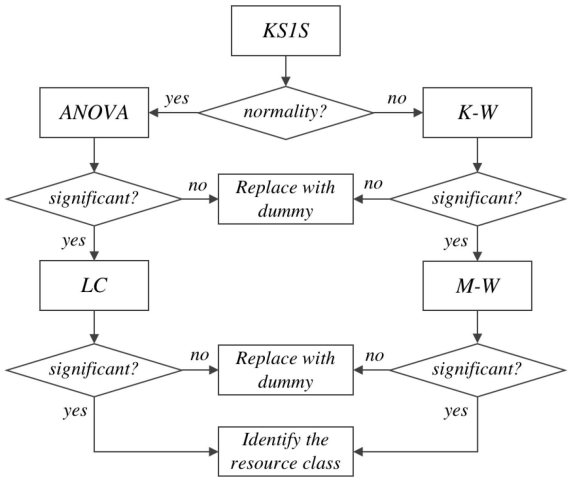

Fig. 5. A flowchart of statistical analysis.

H_0 : The means of all groups under consideration are equal

H_1 : The means are not all equal

If the null hypothesis is accepted, all relevant resources are replaced with dummy resources whereas it is connected to the next phase when rejected. For example, if the null hypothesis for the resources in step *B* in Fig. 3 is accepted, the resources in this step (i.e., *B_M2*, *B_M3*, *B_M4*) are replaced by a dummy resource *B_Dummy*. Otherwise, they proceed to the next step.

The following step is linear contrast (LC) [27] that is effective to test whether the distribution of a particular group is different from other groups. Here, contrast signifies the linear combination of variables and coefficients that sum to 0. Assume that we build a contrast using four sample groups, i.e., X_1, X_2, X_3, and X_4. Then, the coefficient of the target group ($\overline{X_1}$) becomes 3 whereas the others have -1 (e.g., $\overline{X_2}$, $\overline{X_3}$, $\overline{X_4}$) as coefficients. That is, the contrast L is defined such that $L = 3\overline{X_1} - \overline{X_2} - \overline{X_3} - \overline{X_4}$. The hypothesis testing of linear contrast is defined as follows.

$$H_0 : L = 0; H_1 : L \neq 0$$

If H_0 is accepted, the corresponding resource is replaced with the dummy resource. In the opposite case, i.e., H_0 is rejected, we conduct further analysis to identify whether the average of yields resulting from the resource is greater or less than the average of the other resources. As a result, if greater, the relevant resource is classed as high-yield resource; otherwise, it is labeled as low-yield resource. For instance, if the null hypothesis test for *B_M2* and other resources in step *B* is rejected and the average of yields by *B_M2* is greater than the average of yields by the other resources, we classify it as a high-yield resource.

If any of the states do not have normality, it is necessary to employ Kruskal-Wallis H-Test (K-W) [28] instead of ANOVA. The overall approach is quite similar to ANOVA testing, and the null (H_0) and the alternative (H_1) hypothesis are defined as follows.

H_0 : All populations have the same distribution

H_1 : Not all populations have the same distribution

After that, as a substitute for linear contrast, the Mann-Whitney U Test (M-W) [28] is applied, which has the null and the alternative hypothesis as follows.

$$H_0 : \text{Two populations are equal}$$
$$H_1 : \text{Two populations are not equal}$$

Unlike from the linear contrast, the distribution of the target state and the others are compared with the rank-based approach.

Note that all statistical analyses need user-defined significance level thresholds.

The final step of phase 3 is the model simplification based on the integrated model from phase 2 and the resource properties from the previous analysis result. In a nutshell, the dummy resources that belong to the same step are merged into one. As merging methods, we can consider three different types. The first type of merging is to combine all dummy resources. For example, if some of the resources that are relevant to the same step are classified as the dummy, those are straightforwardly combined into one moderate state. This approach, however, neglects the performance of arcs. Thus, the second or third method considers both dummy resources and the corresponding arc performances. In the case of the second approach, only dummy nodes that have the same incoming event property are merged into one. Suppose *B_M2* and *B_M3* are dummy resources. They are merged if all the incoming arcs to *B_M2* and *B_M3* belong to the same class (i.e., high-yield arcs or low-yield arcs). On the other hand, the third approach conducts merging when the outgoing event property is the same. For example, the dummy resources (e.g., *B_M2* and *B_M3*) are merged if all the outgoing arcs are the same class (i.e., high-yield arcs or low-yield arcs).

D. Phase 4. Evaluation and Interpretation

In the last phase, the discovered model from phase 3 is interpreted and evaluated with the test set partitioned in phase 1. First of all, the discovered model is interpreted by a discussion with domain experts. It focuses on identifying the root causes of yield. That is, it understands which flows have impacts on yield and tries to find out the reasons for them. As such, these interpretation results can be the connecting link with yield enhancement.

As far as the evaluation is concerned, we employ the data mining approach for evaluating the model generated from classification algorithms, i.e., confusion matrix and accuracy measure. Fig. 6 depicts the confusion matrix. All cases in the test set are classified as high-yield, low-yield, and moderate cases from model-based analysis, where the resource path of each case is mapped into the reference patterns derived from the obtained model. Then, these predicted results ($\mathcal{P}$) are compared with the actual results ($\mathcal{A}$) of cases.

Based on the confusion matrix, we can derive the accuracy measurement in (3). The accuracy reflects the fraction of cases that are correctly classified.

$$accuracy(\mathcal{A}, \mathcal{P}) = \frac{BB + MM + WW}{N} \tag{3}$$

		Predicted class			
		Best	*Moderate*	*Worst*	
Actual class	*Best*	*BB*	*BM*	*BW*	*B*
	Moderate	*MB*	*MM*	*MW*	*M*
	Worst	*WB*	*WM*	*WW*	*W*
		B'	*M'*	*W'*	*N*

Fig. 6. A confusion matrix for evaluation.

IV. An Empirical Study

To validate the proposed methodology, we conducted a case study with real-life manufacturing data. The goal of this evaluation is to diagnose whether the sound or poor resource paths in a manufacturing process can be determined using our approach. Moreover, the case study will also evaluate the capability of the discovered model that predicts the yield of these cases.

A. Context

The case study presented in this article was conducted in one of the largest semiconductor manufacturers in Korea. The company has produced a variety of semiconductor products, and we applied our approach to one of the production processes. Because of confidentiality agreements, we cannot reveal a detailed explanation of the process. Alternatively, we provide a couple of features for the corresponding process. The process was composed of more than 500 steps, including wafer fabrication, such as oxidation, photo, and etching. It had a stereotypical flow; in other words, most of the wafers were processed using a sequence of steps in the same order. Different from the systematic control-flow, however, a number of different resources were involved in a single step. For this reason, i.e., the lengthy process and the huge number of resources involved, all the wafers were represented by different resource paths, which made it difficult to identify the best and worst resource paths and predict the yield of wafers.

B. Data Preparation

We collected manufacturing data from the manufacturing execution system (MES) that supports the execution of the corresponding process. Then, we conducted some data preprocessing, including handling data error (e.g., removing redundant events and eliminating multiple yield values) and imperfect data (e.g., removing cases that did not cover the whole steps). The preprocessed data was composed of approximately 13,000 wafers that have manufactured over four months. Also, out of the whole process, 74 steps were set up to be analyzed with the help of domain experts, and only relevant events were extracted. It was made up of approximately 950,000 events. Furthermore, more than 1,500 resources were involved in 74 steps. Also, the gathered data included yield

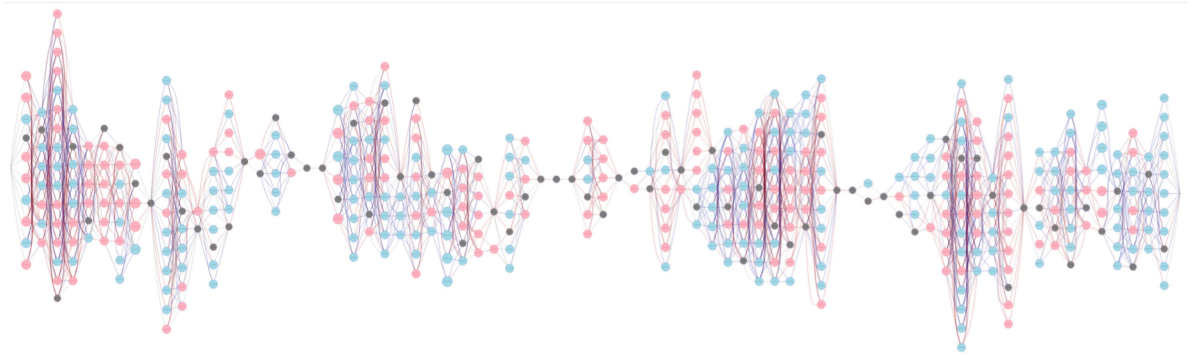

Fig. 7. The discovered model from the real-life log.

for each case. The yield values had a range from 0.0 to 1.0, but we utilized the transformed yield values that have a range from −20 to 10 because of the confidentiality. It was identified that cases have an average yield of 7.75. For the purpose of the evaluation, we partitioned the data to 90% for the training set and 10% for the test set. That is, approximately 11,500 wafers were utilized for constructing a resource-oriented TS model for yield enhancement, while the rest of them were used to evaluate the model.

C. Model Construction and Simplification

First, we discovered the TS model from the manufacturing log. After that, we identified the best and worst cases based on yields of cases. In this case study, we used the user-specified weight of 0.5 to create a range that represents the mediocre quality (see (1) and (2)). As a result, we obtained 30% of the best and worst cases, respectively, which lie outside the range. Then, the integrated model was generated based on these two results. In order to simplify the model, a series of statistical analyses with all significance level thresholds set to 0.05 was performed, and the simplification step was conducted using the integrated model with the first merging type.

The simplified model in Fig. 7 consisted of 74 vertical lines. Each line includes multiple nodes (i.e., production resources). Note that the corresponding nodes have statistically significant effects on yields compared to others. For example, the first step in the process, i.e., the leftmost vertical line, is composed of 10 nodes with three classes of resources. The blue and red nodes represent individual resources that produced high and low yields, respectively. On the other hand, the gray nodes indicate the composite of resources insignificant to the yield. Also, blue arcs represent the resource paths connected to high yields, while red arcs illustrate the low quality. That is, we can identify that 27% and 36% of the whole arcs from the first line are related to the better and poor quality, respectively.

D. Evaluation and Interpretation

In the discovered model, 97 paths contribute to the high-yield cases, while 136 paths contribute to the low-yield cases. As far as the model's predictive competency was concerned, the test set from phase 1 was applied to the discovered model for measuring accuracy. As a result of validating the model with the test set, approximately 83.3% of cases were accurately classified. Out of 393 actual best cases, 318 cases (i.e., 80.9%) were correctly predicted, while the worst cases had an accuracy of 85.7%. Also, we noticed that almost no occurrences were reversed entirely, i.e., the predicted best cases in actual worst ones or the predicted worst cases in actual best ones. Fig. 8 depicts the confusion matrix of the test set.

		Predicted class			
		Best	Moderate	Worst	
Actual class	Best	318 (BB)	70 (BM)	5 (BW)	393 (B)
	Moderate	43 (MB)	438 (MM)	45 (MW)	526 (M)
	Worst	8 (WB)	48 (WM)	335 (WW)	391 (W)
		369 (B')	556 (M')	385 (W')	1310 (N)

Fig. 8. The confusion matrix of the test set.

The domain experts in the company confirmed that finding a high-yield or low-yield run path is useful in practices. They can understand the combinations of resources that influence products' quality and utilize the insights to improve process operations. The crucial factor for the domain experts to identify the root causes of the product quality is the reduced number of possible resource paths (i.e., 97 for high yield wafers and 136 for low yield wafers). In other words, approaches that produce the discovered model without abstraction or simplification do not enable domain experts to analyze the product quality's root cause. Comparing the models from phase 2 and 3, we can figure out how essential it is to simplify the discovered models based on the knowledge of critical resources. The discovered model from phase 2 (i.e., a model without simplification) results in thousands of paths for high-yield and low-yield cases, which are uninterpretable to domain experts.

Concerning the predictive power of the discovered model, the proposed method can predict the high yield and low yield cases by considering the resource path. Naïve approaches using hypothetical tests to identify low-yield resources do not suggest reliable predictive capabilities. In practice, cases processed by the same log-yield resource show different outputs according to the resource path they have passed in the previous

steps. For instance, a naive technique predicts that all the cases from "*EQP_380*" are low-yield cases, while the resource generated 17 high-yield cases and 134 low-yield cases depending on the resources where it has been processed.

Furthermore, regarding the prediction results, the domain experts commented that we should conduct the post hoc analysis for cases predicted as moderate but involved in worst, i.e., false-negative cases. It is essential to identify the worst cases and remove them to maintain an acceptable quality of products.

V. Conclusion

This article suggested a research framework to identify the best and worst resource paths using yield-enhanced manufacturing data. To this end, we employed process mining, data mining, and statistical techniques. Also, the proposed framework was validated using a real-life manufacturing event log. The case study results represented that the proposed approach can help understand the current behaviors and predict the performance (i.e., yield) of the manufacturing process.

This study still needs to be improved. First, this study considers the offline setting for deriving models based on historical manufacturing data. Many researchers have recently attempted to utilize real-time data, and the proposed approach should be extended for real-time monitoring. Besides, this study does not cover the state variables of a particular resource with the sensor data. However, it is believed that yield enhancement is improved with more sophisticated modeling using additional data.

Therefore, for future works, we will extend our approach to an online-fashioned method based on real-time data that estimates yield and recommend the optimized resource paths when manufacturing processes are running. We will also consider a technique for constructing a model that quantifies interaction effects among resources and reflects the resources' status based on the sensor-data for yield enhancement.

References

[1] A. Kusiak, "Smart manufacturing," *Int. J. Prod. Res.*, vol. 56, pp. 508–517, 2017, doi: 10.1080/00207543.2017.1351644.

[2] J. Lee, H.-A. Kao, and S. Yang, "Service innovation and smart analytics for industry 4.0 and big data environment," *Procedia CIRP*, vol. 16, pp. 3–8, 2014, doi: 10.1016/j.procir.2014.02.001.

[3] M. Khakifirooz, C. F. Chien, and Y.-J. Chen, "Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0," *Appl. Soft Comput.*, vol. 68, pp. 990–999, Jul. 2018, doi: 10.1016/j.asoc.2017.11.034.

[4] Y. Meidan, B. Lerner, G. Rabinowitz, and M. Hassoun, "Cycle-time key factor identification and prediction in semiconductor manufacturing using machine learning and data mining," *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 2, pp. 237–248, May 2011, doi: 10.1109/TSM.2011.2118775.

[5] C.-M. Yu and C.-J. Kuo, "Data mining approaches to optimize the allocation of production resources in semiconductor wafer fabrication," in *Proc. Int. Symp. Semicond. Manuf.*, 2016, pp. 1–4, doi: 10.1109/ISSM.2016.7934507.

[6] J. Iskandar, J. Moyne, K. Subrahmanyam, P. Hawkins, and M. Armacost, "Predictive maintenance in semiconductor manufacturing," in *Proc. Adv. Semicond. Manuf. Conf.*, 2015, pp. 384–389, doi: 10.1109/ASMC.2015.7164425.

[7] Y. C. Chang and C. Mastrangelo, "Addressing multicollinearity in semiconductor manufacturing," *Qual. Rel. Eng. Int.*, vol. 27, no. 6, pp. 843–854, 2011, doi: 10.1002/qre.1173.

[8] S. F. Lee and C. J. Spanos, "Prediction of wafer state after plasma processing using real-time tool data," *IEEE Trans. Semicond. Manuf.*, vol. 8, no. 3, pp. 252–261, Aug. 1995, doi: 10.1109/66.400999.

[9] J. Jin and J. Shi, "State space modeling of sheet metal assembly for dimensional control," *J. Manuf. Sci. Eng.*, vol. 121, no. 4, pp. 756–762, 1999, doi: 10.1115/1.2833137.

[10] D. Djurdjanovic and J. Ni, "Stream-of-variation (SOV)-based measurement scheme analysis in multistation machining systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 3, no. 4, pp. 407–422, Oct. 2006, doi: 10.1109/TASE.2006.876617.

[11] F. Yang, S. Jin, and Z. Li, "A comprehensive study of linear variation propagation modeling methods for multistage machining processes," *Int. J. Adv. Manuf. Technol.*, vol. 90, nos. 5–8, pp. 2139–2151, 2017, doi: 10.1007/s00170-016-9490-7.

[12] S. Zhou, Q. Huang, and J. Shi, "State space modeling of dimensional variation propagation in multistage machining process using differential motion vectors," *IEEE Trans. Robot. Autom.*, vol. 19, no. 2, pp. 296–309, 2003, doi: 10.1109/TRA.2003.808852.

[13] H. Fuyun and Z. Zhang, "State space model and numerical simulation of overlay error for multilayer overlay lithography processes," in *Proc. 2nd Int. Conf. Image Vis. Comput.*, 2017, pp. 1123–1127, doi: 10.1109/ICIVC.2017.7984728.

[14] F. He and Z. Zhang, "An empirical study-based state space model for multilayer overlay errors in the step-scan lithography process," *RSC Adv.*, vol. 5, no. 126, pp. 103901–103906, 2015, doi: 10.1039/C5RA07164J.

[15] J. Li and J. Shi, "Knowledge discovery from observational data for process control using causal Bayesian networks," *IIE Trans.*, vol. 39, no. 6, pp. 681–690, 2007, doi: 10.1080/07408170600899532.

[16] C.-F. Chien, W.-C. Wang, and J.-C. Cheng, "Data mining for yield enhancement in semiconductor manufacturing and an empirical study," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 192–198, 2007, doi: 10.1016/j.eswa.2006.04.014.

[17] C.-F. Chien, K.-H. Chang, and W.-C. Wang, "An empirical study of design-of-experiment data mining for yield-loss diagnosis for semiconductor manufacturing," *J. Intell. Manuf.*, vol. 25, no. 5, pp. 961–972, 2014, doi: 10.1007/s10845-013-0791-5.

[18] C.-F. Chien, C.-Y. Hsu, and P.-N. Chen, "Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence," *Flexible Service Manuf. J.*, vol. 25, no. 3, pp. 367–388, 2013, doi: 10.1007/s10696-012-9161-4.

[19] J. Y. Hwang and W. Kuo, "Model-based clustering for integrated circuit yield enhancement," *Eur. J. Oper. Res.*, vol. 178, no. 1, pp. 143–153, 2007, doi: 10.1016/j.ejor.2005.11.032.

[20] R. Brooks, J. Wilson, and R. Thorpe, "Geometry unifies process control, production control and alarm management," *Comput. Control Eng. J.*, vol. 15, no. 1, pp. 22–27, Feb./Mar. 2004, doi: 10.1049/cce:20040105.

[21] J. Moyne and B. Schulze, "Yield management enhanced advanced process control system (YMeAPC)—Part I: Description and case study of feedback for optimized multiprocess control," *IEEE Trans. Semicond. Manuf.*, vol. 23, no. 2, pp. 221–235, May 2010, doi: 10.1109/TSM.2010.2041294.

[22] T. Tsuda *et al.*, "Advanced semiconductor manufacturing using big data," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 3, pp. 229–235, Aug. 2015, doi: 10.1109/TSM.2015.2445320.

[23] W. M. P. van der Aalst, M. H. Schonenberg, and M. Song, "Time prediction based on process mining," *Inf. Syst.*, vol. 36, no. 2, pp. 450–475, 2011, doi: 10.1016/j.is.2010.09.001.

[24] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: A guide for non-statisticians," *Int. J. Endocrinol. Metab.*, vol. 10, no. 2, pp. 486–489, 2012, doi: 10.5812/ijem.3505.

[25] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013, doi: 10.1016/j.jesp.2013.03.013.

[26] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical J.*, vol. 50, no. 3, pp. 346–363, 2008, doi: 10.1002/bimj.200810425.

[27] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statist. Appl. Genet. Mol. Biol.*, vol. 3, no. 1, pp. 1–25, 2004, doi: 10.2202/1544-6115.1027.

[28] G. W. Corder and D. I. Foreman, *Nonparametric Statistics for Non Statisticians: A Step-by-Step Approach*. Hoboken, NJ, USA: Wiley, 2011, pp. 100–106.