

Development of path discovery method for yield enhancement in semiconductor manufacturing

Minsu Cho, Gyunam Park, Minseok Song, and Jinyoun Lee

Abstract—In semiconductor manufacturing, data-driven methodologies have enabled to resolve a variety of issues, in particular, yield management and enhancement. Yield, one of the important key performance indicators in semiconductor manufacturing, is mostly affected by production resources, i.e., equipment involved in the process. As such, there has been a tendency to find 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 paper proposes a systematic methodology for discovering a resource-oriented finite state machine model in a semiconductor manufacturing process to identify resource paths resulting in high and low yield. The proposed methodology is based on the model-based analysis (i.e., finite state machine mining) in process mining and a couple of statistical analyses. To validate our approach, we conducted an empirical study with real-life data from one of the leading semiconductor manufacturing companies.

Index Terms—Yield enhancement, process mining, semiconductor manufacturing, resource paths, statistical analysis

I. INTRODUCTION

EVIDENCE-BASED decision making through data analysis has a vital role in a modern manufacturing corporation [1]–[3], since plenty of data has been collected in a manufacturing environment [4]–[6]. 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 [7]–[10].

Production resources, i.e., equipment involved in manufacturing processes, are one of the critical factors that affect yields [11]. 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 [12]. However, most of the works are limited to investigate the relation 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 [9], [13]. All wafers pass

through hundreds of process steps, each of which has a number of 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 [9]. Therefore, we should deal with the problem of multicollinearity among the combinations of resources [14], [15]. For example, in Fig. 1, $chamber_{i,1,1}$ is identified as a key resource that results in the high-yield wafer (i.e., it has a strong impact on high-yield). However, the corresponding machine at the same time belongs to another path which produces the wafer of poor yield. To this end, it is taken for granted to investigate an association of yield with the sequence of resources rather than with a single resource involved in the production.

In this paper, we propose a new methodology for discovering a resource-oriented Finite State Machine (FSM) model in a semiconductor manufacturing process. More in detail, 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 FSM model construction, model simplification, and evaluation and interpretation. In a nutshell, we first generate an FSM 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 in yield. Lastly, the discovered model is evaluated and interpreted.

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

II. A FRAMEWORK FOR CONSTRUCTING A RESOURCE-ORIENTED FSM MODEL FOR YIELD ENHANCEMENT

This section describes our approach to derive a resource-oriented FSM model for yield enhancement based on manufacturing data. Fig. 2 provides the proposed research framework in this paper. Our method includes four phases: data preparation, resource-oriented FSM 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 FSM model is generated, and the cases (i.e., wafers) recorded relatively superior or inferior to

This work was supported by ...

M. Cho, G. Park and M. Song are with the Department of Industrial and Management Engineering, Pohang University of Science and Technology, 77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, 37673, Korea (e-mail: minsucho@postech.ac.kr; gnpark@postech.ac.kr; mssong@postech.ac.kr).

J. Lee is with the Mechatronics R&D Center, 10, Samsung 1-ro 6-gil, Hwaseong-si, Gyeonggi-do, 18381, Korea (e-mail: jinyoun0lee@samsung.com).

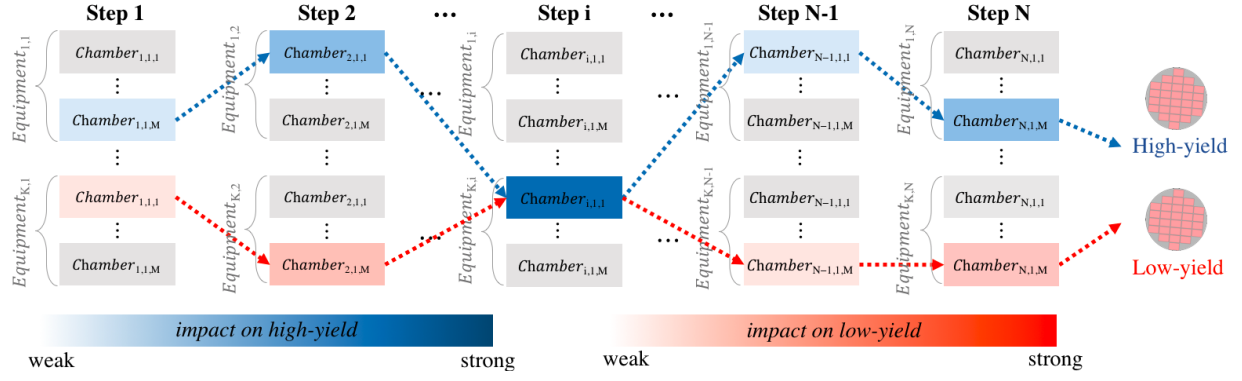


Fig. 1. An example of the complicated semiconductor manufacturing process

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 at determining 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 at constructing the sufficiently structured data for obtaining valuable analysis results. This research takes the format of event logs utilized in process mining [16]. Definition 1 provides formal definition of yield-enhanced 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 $\pi_y(c)$. 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 FSM 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, Yield-enhanced event log): Let E be the event universe. Events may have several attributes (e.g., step, originator, timestamp). Let AN be a set of attribute names. Let $C = \{c_1, c_2, c_3, \dots, c_k\}$ be the set of cases. For any case $c \in C$ and any attribute name $an \in AN$: $\pi_{an}(c)$ is the value of the attribute an for case c . Every case has a mandatory attribute, yield y : $\pi_y(c) \in \mathcal{B}(\mathcal{R})$. A trace σ is a finite sequence of events $\sigma \in E^*$. Also, for any case $c \in C$: $\pi_y(c)$ is the yield value for case c . A yield-enhanced event log L is a collection of possible cases.

TABLE I
AN EXAMPLE OF YIELD-ENHANCED EVENT LOG

CaseID	EventID	Step	Resource	Timestamp	Yield
Case 1	E1	A	M1	2018-01-01 10:30	0.95
	E2	B	M2	2018-01-01 12:00	
	E3	C	M4	2018-01-01 17:00	
	E4	D	M6	2018-01-02 09:00	
Case 2	E5	A	M1	2018-01-04 09:00	0.85
	E6	B	M3	2018-01-04 12:00	
	E7	C	M5	2018-01-04 15:00	
	E8	D	M6	2018-01-05 13:00	
Case 3	E9	A	M1	2018-01-06 09:00	0.9
	E10	B	M2	2018-01-06 14:00	
	E11	C	M5	2018-01-07 13:00	
	E12	D	M6	2018-01-07 18:00	

B. Phase 2. FSM Model Construction

The second phase aims at generating a resource-oriented FSM model for yield enhancement based on manufacturing event logs. First of all, a process mining discovery technique is employed to construct a FSM model from event logs [16]. The FSM model is formalized in Definition 2. The model $FSM = (S, E, T)$ 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 that 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, Finite State Machine): Let L be a yield-enhanced event log. Let E^* and E be a set of possible traces and events. A state representation function $rep^s \in E^* \rightarrow R^s$ is a function that gives a representation from a trace σ , 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

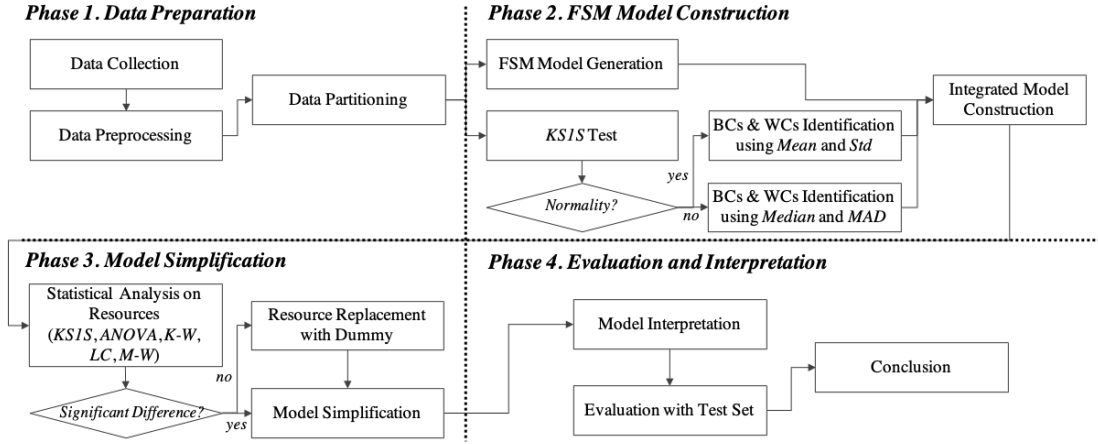


Fig. 2. Overview of the proposed research framework

possible event representations. A finite state machine FSM is defined as a triplet (S, E, T) , where $S = \{rep^s(hd^k(\sigma)) | \sigma \in L \wedge 0 \leq k \leq |\sigma|\}$, $E = \{rep^e(\sigma(k)) | \sigma \in L \wedge 1 \leq k \leq |\sigma|\}$, and $T \subseteq S \times E \times S$ with $T = \{rep^s(hd^k(\sigma)), rep^e(\sigma(k+1)), rep^s(hd^{k+1}(\sigma)) | \sigma \in L \wedge 0 \leq k \leq |\sigma|\}$ is the set of states, the set of events, and the set of transitions.

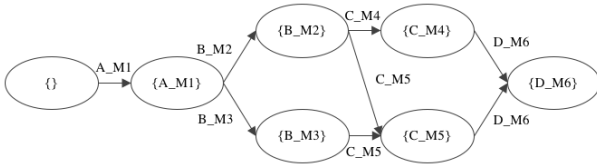


Fig. 3. An example of transition systems based on the example log in Table I

The second step of this phase is to identify the best cases, i.e., BCs , and the worst cases, i.e., WCs . First, we verify the normality of the empirical distribution for yield values by employing *Kolmogorov-Smirnov one sample test* (i.e., $KSIS$) [17]. 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.

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

If the case satisfies the normality condition, BCs and WCs 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) [18] are utilized to establish BCs and WCs . Based on these values, we can create a certain range that represents just a mediocre quality as provided in Eq. (1) and (2). Here, w is the user-specified weight. On the basis of 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

BCs . On the other hand, if cases have a lower yield than the minimum value of the range, they belong to WCs .

$$\mu - w \times \sigma \leq yield \leq \mu + w \times \sigma \quad (1)$$

$$M - w \times MAD \leq yield \leq M + w \times MAD \quad (2)$$

Lastly, the FSM model and statistical analysis results are combined into one single model. As such, the best and worst paths are visualized on the discovered model. Here, firstly, the best and worst case ratios are calculated for every arc in the FSM model. After that, it is evaluated that the summation of these two values is greater than the predefined threshold. It determines whether the flow is paramount in the model. Then, we measure the difference between the best and worst case ratio for important paths. As such, if the best case ratio is greater than the worst case ratio, the relevant arc is determined as the part of the best paths that lead to the high yield. In the opposite case, it is considered as a piece of the worst paths.

C. Phase 3. Model Simplification

The integrated model has a limitation that it has a considerable complexity as the process gets more extended or more machines are engaged in activities. To overcome the limitation, we need to find out mainline states that have a statistical significance to yield and simplify the model on the basis of them. The third phase aims at simplifying the model with the results from the statistical analysis. To this end, we employ the a series of statistical techniques. Fig. 4 provides the flow of statistical analysis.

First, the normality of distributions of all resources involved in the same process step is testified with *Kolmogorov-Smirnov one sample test* (i.e., $KSIS$) [17]. As with *Phase 2*, the same test statistic and hypothesis testing are applied. Following the normality testing result, further steps are divided into two different directions. Though they exploit different statistical techniques, each step of them has common objectives. First, it is analyzed to identify whether the states engaged in the same step have a discrepancy with each other. After that, if the difference exists, it is determined what states make the quality of products better or worse.

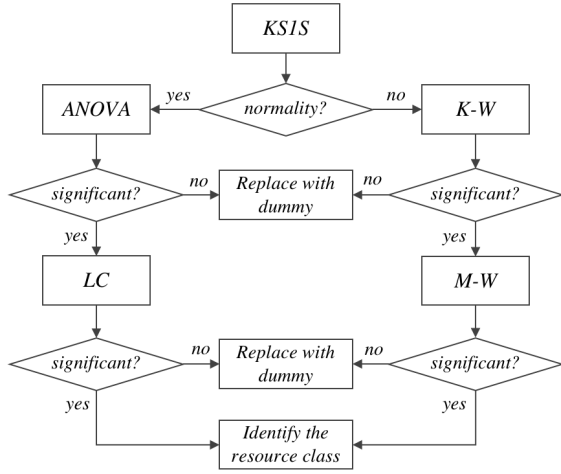


Fig. 4. A flowchart of statistical analysis

In case the distributions of all states follow the normal distribution, *Analysis of variance*, i.e., *ANOVA* [19], is applied to identify whether all states have the indifferent yield value. Followings are the null and alternative hypothesis for ANOVA.

$$H_0: \text{The means of all groups under consideration are equal}$$

$$H_1: \text{The means are not all equal}$$

If the null hypothesis is accepted, all relevant resources are replaced with *dummy resources*, whereas it is connected to the next phase when rejected.

The following step is *Linear contrast* (i.e., *LC*) [20] that is effective to testify whether the distribution of a particular group is different from other groups. Here, contrast signifies the linear combination of variables and coefficients of them sums 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 ($\bar{X}_1$) becomes 3, whereas the others have -1 (e.g., $\bar{X}_2, \bar{X}_3, \bar{X}_4$) as coefficients. That is, the contrast L is defined such that $L = 3\bar{X}_1 - \bar{X}_2 - \bar{X}_3 - \bar{X}_4$. 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 for identifying whether the average of 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-class*; otherwise, it is labeled as *low-class*.

If any of the states do not have the normality, it is required to employ *Kruskal-Wallis H-Test* (i.e., *K-W*) [21] 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.

$$H_0: \text{All populations have the same distribution}$$

$$H_1: \text{Not all populations have the same distribution}$$

After that, as a substitute for linear contrast, *Mann-Whitney U Test* (i.e., *M-W*) [21] is applied, which has the null and the alternative hypothesis as follows.

$$H_0: \text{Two populations are equal}$$

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

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

Note that all statistical analysis 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. On the other hand, the third approach conducts merging when the outgoing event property is the same.

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 on 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. 5 depicts the confusion matrix. All cases in the test set are classified as *best*, *worst*, 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 measure, and Eq. (3) presents it. The accuracy reflects the fraction of cases that are correctly classified.

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

III. AN EMPIRICAL STUDY

To demonstrate the usefulness of the proposed methodology, we conducted a case study with real-life manufacturing data. The goal of this evaluation is to diagnose whether the sound

		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. 5. A confusion matrix for evaluation

or poor resource paths in a manufacturing process can be determined based on our approach. Moreover, the case study will also evaluate the capability of the discovered model that predicts the yield of cases.

A. Context

The case study presented in this paper 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, oxidation, photo, and etching. It had a stereotypical flow; in other words, most of the wafers were performed in 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 yields of wafers.

B. Data Preparation

We collected a manufacturing data from the manufacturing execution systems (MES) supporting the execution of the corresponding process. Then, we conducted a couple of data preprocessing including data error handling (e.g., removing redundant events and eliminating multiple yield values) and imperfect data handling (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 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 FSM model for yield enhancement, while the rest of them were used to evaluate the model.

C. Model Construction and Simplification

First, we discovered the FSM model from the manufacturing log. After that, we identified the best and worst cases based on yields of cases. In this case study, we set the user-defined weight as 0.5; as such, we obtained 30% of the best and worst cases, respectively. Then, the integrated model was generated based on these two results. To simplify the model, a series of statistical analysis with all significance level thresholds as 0.05 was performed, and the simplification step was conducted to the integrated model with the first merging type.

The simplified model in Fig. 6 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 other ones. 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 the 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 poor quality. That is, we are able to identify that 27% and 36% of the whole arcs from the first line are related to the sound and poor quality, respectively.

D. Evaluation and Interpretation

As far as the evaluation of the model was concerned, the test set was applied to the 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 the 85.7% of accuracy. Also, we noticed that there were almost no occurrences that were reversed entirely, i.e., the predicted best cases in actual worst ones or the predicted worst cases in actual best ones.

The company's domain experts were provided the case study results including the resource paths and combinations for the high or low yield and the important resources that had a significant effect on the product quality. They were indeed interested in the results and expressed their willingness to consider utilizing them in order to improve process operation. Also, regarding the prediction results, they commented that we should conduct the post-hoc analysis for cases predicted as *moderate* but actually involved in *worst*, i.e., false-negative cases. This is because it is important to identify the worst cases and remove them to maintain an acceptable quality of products.

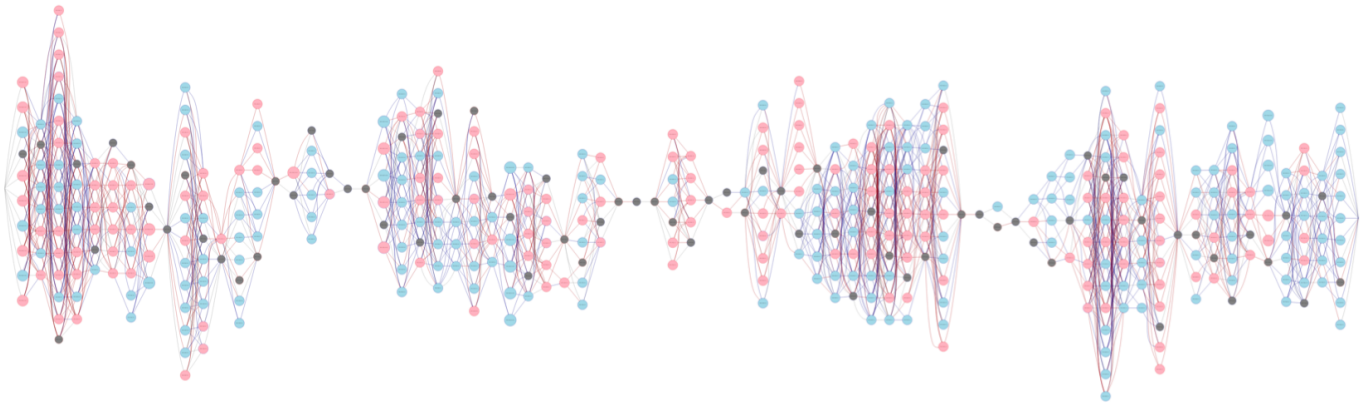


Fig. 6. The discovered model from the real-life log

IV. CONCLUSION

This paper suggested a research framework to identify the best and worst resource paths using yield-enhanced manufacturing data. To this end, we employed a variety of process mining and data mining techniques. Also, the proposed framework was validated using a real-life manufacturing event log. The results of the case study represented that the proposed approach can be effective to understand the current behaviors and predict the performance (i.e., yield) of the manufacturing process. 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 at the moment the process is running. Also, we will consider an approach for constructing a model that reflects the status of the resources based on the sensor-data for yield enhancement.

ACKNOWLEDGMENT

This work was supported by ...

REFERENCES

- [1] A. Kusiak, "Smart manufacturing," *International Journal of Production Research*, vol. 0, no. 0, pp. 1–10, 2017.
- [2] J. Lee, B. Bagheri, and C. Jin, "Introduction to cyber manufacturing," *Manufacturing Letters*, vol. 8, pp. 11–15, 2016.
- [3] C.-F. Chien, Y.-J. Chen, and J.-T. Peng, "Manufacturing intelligence for semiconductor demand forecast based on technology diffusion and product life cycle," *International Journal of Production Economics*, vol. 128, no. 2, pp. 496–509, 2010.
- [4] 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.
- [5] J. Moyne and J. Iskandar, "Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing," *Processes*, vol. 5, no. 3, 2017.
- [6] Y.-t. Kao, S.-c. Chang, S. Dauzere-Peres, and J. Blue, "Opportunity for improving fab effectiveness by predictive overall equipment effectiveness (poe)," in *e-Manufacturing and Design Collaboration Symposium (eMDC)*, 2016. IEEE, 2016, pp. 1–4.
- [7] K. B. Lee, S. Cheon, and C. O. Kim, "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 30, no. 2, pp. 135–142, 2017.
- [8] C.-F. Chien, C.-Y. Hsu, and P.-N. Chen, "Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence," *Flexible Services and Manufacturing Journal*, vol. 25, no. 3, pp. 367–388, 2013.
- [9] 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," *Applied Soft Computing*, 2017.
- [10] K. Nakata, R. Orihara, Y. Mizuoka, and K. Takagi, "A comprehensive big-data-based monitoring system for yield enhancement in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 30, no. 4, pp. 339–344, 2017.
- [11] 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 Transactions on Semiconductor Manufacturing*, vol. 24, no. 2, pp. 237–248, 2011.
- [12] C. M. Yu and C. J. Kuo, "Data mining approaches to optimize the allocation of production resources in semiconductor wafer fabrication," *2016 International Symposium on Semiconductor Manufacturing (ISSM)*, pp. 1–4, 2016.
- [13] J. Iskandar, J. Moyne, K. Subrahmanyam, P. Hawkins, and M. Armacost, "Predictive maintenance in semiconductor manufacturing," in *2015 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2015, pp. 384–389.
- [14] Y. C. Chang and C. Mastrangelo, "Addressing multicollinearity in semiconductor manufacturing," *Quality and Reliability Engineering International*, vol. 27, no. 6, pp. 843–854, 2011.
- [15] S. F. Lee and C. J. Spanos, "Prediction of wafer state after plasma processing using real-time tool data," *IEEE Transactions on Semiconductor Manufacturing*, vol. 8, no. 3, pp. 252–261, 1995.
- [16] W. van der Aalst, M. Schonenberg, and M. Song, "Time prediction based on process mining," *Information Systems*, vol. 36, no. 2, pp. 450–475, 2011.
- [17] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: a guide for non-statisticians," *International journal of endocrinology and metabolism*, vol. 10, no. 2, pp. 486–489, 2012.
- [18] 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," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.
- [19] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical journal*, vol. 50, no. 3, pp. 346–363, 2008.
- [20] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, pp. 1–25, 2004.
- [21] G. W. Corder and D. I. Foreman, "Nonparametric statistics for non statisticians: A step-by-step approach," 2011.



Minsu Cho received the Ph.D. degree in management engineering at UNIST (Ulsan National Institute of Science and Technology), Ulsan, South Korea in 2018. He is currently a postdoctoral research associate at POSTECH (Pohang University of Science and Technology).

His research interest includes business process management, process mining, manufacturing process analysis, clinical pathway analysis, simulation analysis based on process mining, and process mining case studies. Also, his work has been published in

Decision Support Systems and International Journal of Medical Informatics as well as at BPM conference.



Gyunam Park received the B.S. degree in business administration and computer science from UNIST (Ulsan National Institute of Science and Technology), Ulsan, South Korea, in 2013. He is currently a M.S student Industrial & Management Engineering at POSTECH (Pohang University of Science and Technology).

His research interest includes process mining, manufacturing process analysis, business process management, machine learning, deep learning, and big data.



Minseok Song received his Ph.D. degree in industrial & management engineering at POSTECH (Pohang University of Science and Technology) in 2006. He is now an associate professor at the Department of Industrial & Management Engineering at POSTECH. Prior to this, he stayed at the Information Systems department of the Technology Management of Eindhoven University of Technology as a post-doctoral researcher from 2006 until 2009. Also, he was an assistant/associate professor at UNIST (Ulsan National Institute of Science and

Technology).

His research interest includes business process management, process mining, business analytics, simulation, and social network analysis. He has published more than 60 scientific papers in several top-level venues such as Decision Support Systems, Information Systems, Journal of Information Technology, International Journal of Medical Informatics, etc.

Jinyoun Lee To be updated.

PLACE
PHOTO
HERE