Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-020-01654-2

Approach to derive golden paths based on machine sequence patterns in multistage manufacturing process

Chang-Ho Lee[1] · Dong-Hee Lee[2] · Young-Mok Bae[3] · Seung-Hyun Choi[1] · Ki-Hun Kim[4,5] · Kwang-Jae Kim[1]

Received: 2 October 2019 / Accepted: 18 August 2020

Abstract

A multistage manufacturing process (MMP) consists of several consecutive process stages, each of which has multiple machines performing the same functions in parallel. A manufacturing path (simply referred to as path) is defined as an ordered set indicating a record of machines assigned to a product at each process stage of an MMP. An MMP usually produces products through various paths. In practice, multiple machines in a process stage have different operational performances, which accumulate during production and affect the quality of products. This study proposes a heuristic approach to derive the golden paths that produce products whose quality exceeds the desired level. The proposed approach consists of the searching phase and the merging phase. The searching phase extracts two types of machine sequence patterns (MSPs) from a path dataset in an MMP. An MSP is a subset of the path that is defined as an ordered set of assigned machines from several process stages. The two extracted types of MSPs are: (1) superior MSP, which affects the production of superior-quality products, and (2) inferior MSP, which affects the production of inferior-quality products, called inferior MSP. The merging phase derives the golden paths by combining superior MSPs and excluding inferior MSPs. The proposed approach is verified by applying it to a hypothetical path dataset and the semiconductor tool fault isolation (SETFI) dataset. This verification shows that the proposed approach derives the golden paths that exceed the predefined product quality level. This outcome demonstrates the practical viability of the proposed approach in an MMP.

Keywords Multistage manufacturing process (MMP) · Product quality · Golden path · Machine sequence pattern (MSP) · Quality engineering

Introduction

Product quality management and improvement is a key issue for competitive success in manufacturing industries. To meet customer requirements for product quality, manufacturing companies have strived for continuous improvement in product quality (Chien et al. 2014). Accordingly, the manufacturing process should be designed to produce goods that satisfy the pre-determined quality level, and the proportion of products that exceed the quality level should ideally be as high as possible (Huff 2020). Therefore, reducing the product failure rate in the early stage of the manufacturing process lifecycle is important (Du et al. 2015). In other words, detect-

✉ Kwang-Jae Kim
kjk@postech.ac.kr

Chang-Ho Lee
dlckdgh@postech.ac.kr

Dong-Hee Lee
dh@hanyang.ac.kr

Young-Mok Bae
youngmok.bae@sk.com

Seung-Hyun Choi
cshyun102@postech.ac.kr

Ki-Hun Kim
kihun@unist.ac.kr

1 Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Republic of Korea

2 Division of Interdisciplinary Industrial Studies, Hanyang University, Seoul, Republic of Korea

3 SK Hynix, Cheongju, Republic of Korea

4 Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

5 Faculty of Industrial Design Engineering, Delft University of Technology, Delft, The Netherlands

Published online: 09 September 2020

ing process failures and identifying the cause of product failure as soon as possible should be priorities to maintain the product quality within a normal range. As the manufacturing process matures, the product failure rate gradually decreases. However, improvements in product quality can still be made, because the product failure rate is still not equal to zero, and even the product quality of normal goods may vary (Oztemel and Gursev 2020; Fahey, 2020).

Thus, expanding perspectives to continuously improve the quality of products is necessary. As the manufacturing process matures and the occurrence of product failures becomes rare, a holistic approach that considers not only reducing the failure rates of products but also maximizing the product quality is needed (Psarommatis et al. 2020). However, existing studies have focused only on reducing the failure rate of products (Irrera and Vieira 2014; Wang and Wang 2016). In addition, existing studies have sought to identify the causes of product failure only for an individual or separated process stages, not for all process stages (Eger et al. 2018). Although these efforts have contributed to reducing product failures, little knowledge is available on maximizing the quality of products throughout the entire process stages.

To overcome this limitation and to expand the perspective, the characteristics of a multi-stage manufacturing process (MMP) should be considered. An MMP generally includes multiple successive process stages to produce an end product. To meet increasing demands, each process stage operates multiple machines that perform the same task. Thus, different machines could be assigned to an individual product at each process stage. Consequently, each product has its unique sequential processing record of assigned machines at each process stage (Huang and Shi 2004; Lee et al. 2015). Such a sequential record throughout the entire process stages of each product is called a manufacturing path (simply referred to as a path in this study) (Kimemia and Gershwin 1985; Jin and Liu 2013). The solid lines in Fig. 1 illustrate sample paths.

In practice, even if machines at each process stage are supposed to perform the same task, the actual performances of machines could be different. (Wuest et al. 2014; Bera and Mukherjee 2016; Taha 2019). Therefore, different performances of machines are accumulated along successive process stages in an MMP. Given that the products can have various paths, the accumulated differences could vary depending on the paths and may result in an intolerable difference in product quality (Jin and Liu 2013; Ju et al. 2015; Wuest et al. 2014; Bera and Mukherjee 2016). In this sense, a holistic approach that focuses on improvement in product quality and takes into account all process stages and the operational performances of machines at each process stage is needed (Eger et al. 2018; Huff 2020; Psarommatis et al. 2020). However, little knowledge about this research issue is available.

In this regard, this study aims to develop a heuristic approach for deriving the paths that produce products whose quality exceeds the desired level. These paths are called "golden paths." By deriving golden paths, various paths that are expected to produce superior-quality products can be provided, which contributes to increasing the proportion of products that exceed the desired quality level. In addition, the derived golden paths could contribute to reducing the quality differences of products according to the paths by being benchmarked against other paths. For example, the information of operating parameter conditions of machines included in the golden paths can be used as a basis when optimizing other machines in an MMP.

The proposed approach applies the concept of sequence pattern inspired by sequential pattern mining to derive golden paths. A sequence is an ordered set of items, and a sequence pattern is a specific subsequence existing in a set of sequences (Agrawal and Srikant 1995). The proposed approach derives the golden paths on the basis of two groups of sequence patterns extracted from the historical path dataset. The first group consists of sequence patterns that appear on the paths that produce products of a certain level or higher quality. The second group consists of sequence patterns that appear on the paths that produce products of a certain level or low quality. It is noted that the levels of product quality, which are the criteria for extracting each sequence pattern group, could vary.

Unlike existing studies that focus on individual or separated process stages, the proposed approach takes the paths into account to improve the quality of products in an MMP. Although existing studies have focused only on identifying the cause of product failure, the proposed approach extracts sequence patterns that affect the production of superior- and inferior-quality products and derives golden paths on the basis of these sequence patterns. In addition, in recent MMP, the path dataset accumulated during the operation of an MMP is easy to collect and prevalent. Accordingly, the availability of the use of the proposed approach is increasing in line with this recent trend.

This study is organized as follows. Section 2 reviews the related studies on sequential pattern mining from the viewpoint of product quality in an MMP. Section 3 proposes a heuristic approach to derive the golden paths. Section 4 applies the proposed approach to a hypothetical path dataset to judge its feasibility for deriving the golden paths. Section 5 validates the proposed approach by using the semiconductor tool fault isolation (SETFI) dataset. Section 6 concludes this study.

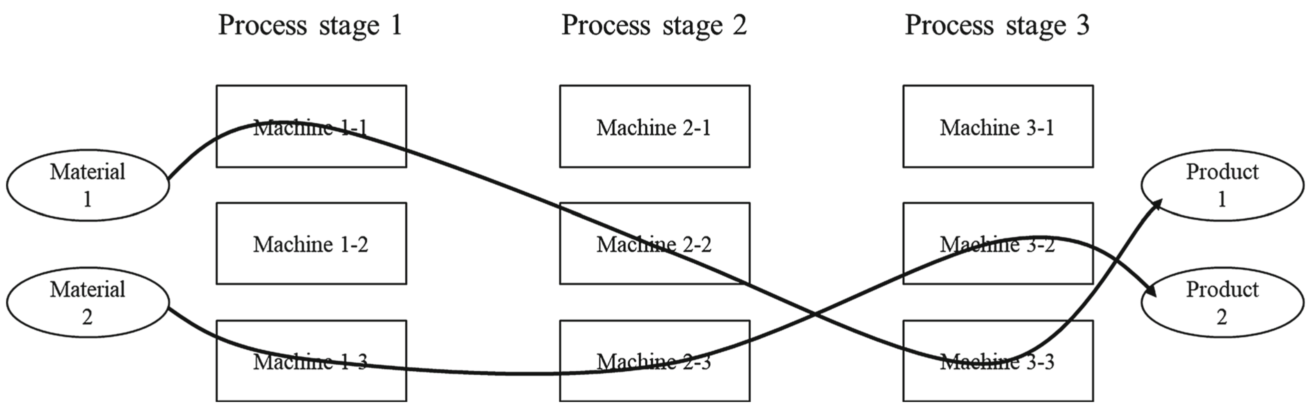


Fig. 1 An illustration of the paths

Review of related works

In this section, we review the related works that focused on improving the quality of products in an MMP on the basis of sequential pattern mining. Sequential pattern mining is defined as a method or approach to discover all subsequences that exceed a user-specified minimum frequency from a dataset (Agrawal and Srikant 1995). The determined minimum frequency is called the support threshold. The discovered subsequences are called sequence patterns (Garofalakis et al. 1999). More formally, the problem of sequential pattern mining can be explained as follows (Agrawal and Srikant 1995). Let $SD = \{s_1, s_2, s_3, \ldots s_n\}$ be a finite set of sequence and $I = \{i_1, i_2, i_3, \ldots, i_m\}$. be a set of items. Each sequence $s_j \in SD$ $(1 \leq j \leq n, j \in \mathbb{N})$ is an ordered set of itemsets $s_j = \{X_1, X_2, \ldots, X_o\}$ such that $X_1, X_2, \ldots, X_o \subseteq I$ $(n, m, o \in \mathbb{N})$. An item x is said to occur before another item y in a sequence $S_j = \{X_1, X_2, \ldots, X_o\}$ if integers $k < l$ exist, such that $x \in X_k$ and $y \in X_l$. The $\{x, y\}$ is called a subsequence of s_j. The sequential pattern mining identifies all subsequences common to several sequences from the SD such that their frequency is higher or equal to a support threshold. For example, if six transactions $T = \{a, b, c, d, e, f\}$ and four sequences of transactions are presented in the SD: $s_1 = \{a, b, d, a, c\}$, $s_2 = \{e, a, b, d, a\}$, $s_3 = \{b, a, b, f, e, a\}$, $s_4 = \{a, b, f, a\}$, then sequential pattern mining finds frequent sequence patterns that exceed a user-defined threshold. If the threshold is set to four, then the sequential pattern mining extracts the following six frequent sequence patterns from SD: $\{a\}$, $\{a, a\}$, $\{a, b\}$, $\{a, b, a\}$, $\{b\}$, $\{b, a\}$. The frequent sequence pattern $\{a, b, a\}$, for instance, indicates that once transaction a occurs, transaction b occurs before transaction a occurs again.

Sequential pattern mining helps product and quality engineers extract veiled information from a large manufacturing dataset with various types of sequences, such as machine operation logs, logistics logs, or event logs. Accordingly, several studies have adopted sequential pattern mining to analyze the dataset accumulated from an MMP. The extracted sequence patterns from the related studies can be divided into two types: machine and event. Machine sequence patterns (MSPs) indicate that a sequence pattern consists of the information about objects that perform specific tasks on process stages. Event sequence patterns (ESPs) indicate that a sequence pattern consists of the information about signals, such as delay alarms or maintenances, collected during production.

Da Cunha et al. (2006) adopted the association rule to discern any associations between product failures and assembly sequences in the manufacturing process. They extracted MSPs to reorganize the assembly order for minimizing the occurrence of failures. Rokach et al. (2008) extended the approach of Da Cunha et al. (2006) by adopting the longest common subsequence, Teiresias algorithm, and decision tree. The developed approach showed efficient performance for analyzing long sequences and applied to classify the quality of products as “pass” or “fail.” Kerdprasop and Kerdprasop (2013, 2014) adopted the TraMineR algorithm to find MSPs that cause product failures in the semiconductor manufacturing process. They compared classification performances of failure for seven classification algorithms by learning the extracted MSPs. Nakata et al. (2017) adopted the FPGrowth algorithm to the semiconductor manufacturing process to identify MSPs that cause failure wafer map patterns. The extracted MSPs were used to support engineers’ work by constructing a diagnostic system.

On the one hand, Kamsu-Fogeum et al. (2013) adopted the TopKRules algorithm to discover the causes of different failure types in the drill manufacturing process. For each type of failure, they extracted ESPs composed of delay event logs. Through deductive reasoning, the extracted ESPs were used to determine the root cause of failure and planning the maintenance programs. Lee et al. (2015) proposed an algorithm to estimate product quality in the die attach process.

Table 1 The summary of the related studies

Related studies	Type of the sequence patterns	Meaning of the sequence patterns	Objective
Da Cunha et al. (2006)	Machine	The cause of failures	Re-organize the assembly process to minimize the occurrence of failure
Rokach et al. (2008)	Machine	The cause of failures	Construct a predictive model of failure
Kerdprasop and Kerdprasop (2013, 2014)	Machine	The cause of failures	Compare the performances of classification methods
Nakata et al. (2017)	Machine	The cause of failures	Construct a failure diagnose system
Kamsu-Foguem et al. (2013)	Event	The causes of failures	Conduct deductive reasoning of the root cause & plan the maintenance programs
Lee et al. (2015)	Event	The cause of differences on product quality	Construct a predictive model of product quality
Lim et al. (2017)	Event	The causes of failures	Construct a predictive model of failure
Sellami et al. (2019, 2020)	Event	The cause of failures	Construct a predictive model of failure
Our research	Machine	The causes of improvement & reduction in product quality	Derive golden paths

The algorithm extracts ESPs that cause a critical difference in product quality. The extracted ESPs were used to construct a tree to estimate the product quality for new event sequences. Lim et al. (2017) adopted the sequential pattern discovery by using equivalence algorithm to extract ESPs that appear frequently during the wire bonding process. The extracted ESPs were used to learn the bagged-LASSO to construct a predictive model of failure. Sellami et al. (2019, 2020) suggested an approach to estimate the occurrence and the occurrence interval of product failures on the basis of frequently appearing ESPs that were suspected to trigger failures.

Table 1 shows a summary of the reviewed studies according to three attributes: the type of sequence patterns, the meaning of sequence patterns, and the objective. This study is differentiated from the related studies according to the following aspects. First, although the reviewed studies have focused on only the sequence patterns that cause product failure (as shown in the second column in Table 1), this study focuses on the sequence patterns that cause not only inferior-quality products but also superior-quality products. In addition, the proposed approach is made from a holistic point of view of an MMP by deriving golden paths to improve the quality of products. Lastly, the proposed approach not only contributes to improving the quality of products but can also be used in various ways in the operation of MMP. For example, the derived golden paths could be used for developing dispatching rules for raw materials.

Proposed approach

Problem description

The objective of the proposed approach is to derive the golden paths from the historical path dataset. In this study, the problem situation for deriving the golden paths by applying the proposed approach is defined on the basis of the following assumptions. (1) The quality of products varies depending on the paths in an MMP. (2) The quality of products is measured on a scale of [0, 100], with 0 and 100 being the lowest and highest quality, respectively. (3) The quality of products is affected by individual machines and the combinations of machines in different process stages in an MMP. (4) Machines existing in the MMP do not have any changes in status, such as deterioration, breakdown, or maintenance, during the production.

The historical path dataset contains information on paths of produced products in an MMP and information on product quality along each path. Table 2 is a representative example of the historical path dataset used in this study. Each row of the given dataset represents the path of a product and the corresponding product quality. The N columns (Process stage 1 to N) in Table 2 represent the process stages constituting an

Table 2 An example of path dataset

Paths	Process stage							Product quality
	1	2	3	…	n	…	N	
$\boldsymbol{P}_1$	1–2	2–4	3–3	…	n –5	…	N –2	99
$\boldsymbol{P}_2$	1–1	2–4	3–5	…	n –2	…	N –2	98
$\boldsymbol{P}_3$	1–3	2–4	3–5	…	n –1	…	N –3	98
$\boldsymbol{P}_4$	1–2	2–2	3–4	…	n –3	…	N –1	96
$\boldsymbol{P}_5$	1–3	2–1	3–1	…	n –5	…	N –2	95
…	…	…	…	…	…	…	…	…
$\boldsymbol{P}_T$	1–2	2–4	3–2	…	n –1	…	N –3	90

MMP. The notation "*i–j*" on each cell in Table 2 means that the *j*th machine is assigned to conduct an operation at the *i*th process stage. The *i*th path ($\boldsymbol{P}_t$) is in the form of an ordered set, which is horizontally jointed by each cell in sequential order (e.g., $\boldsymbol{P}_1 = \{1\text{–}2, 2\text{–}4, 3\text{–}3, \ldots, n\text{–}5, \ldots, N\text{–}2\}$, where $t = 1, 2, \ldots, T$ in Table 2). The length of each path should be equal to N. The product quality along the t^{th} path (q_t) is recorded at the right-end column of the corresponding *i*th row.

Figure 2 describes a schematic overview of the research. In Fig. 2, the historical path dataset is composed of path information and product quality information as mentioned above. The historical path dataset is the input of the proposed approach, and the output of the proposed approach is the golden paths. It is noted that each derived golden path consists only of path information and no product quality information is included.

The proposed approach derives the golden paths by analyzing the dataset over two phases; namely, searching phase and merging phase. As the first step to derive the golden paths, determining which machines cause differences in product quality or contribute to the production of high-quality products from the dataset is necessary. The searching phase of the proposed approach carries out this task. This phase discovers the superior MSP set ($\boldsymbol{S}$) and the inferior MSP set ($\boldsymbol{I}$) from the dataset. Once the two sets are discovered, then the merging phase of the proposed approach derives the golden

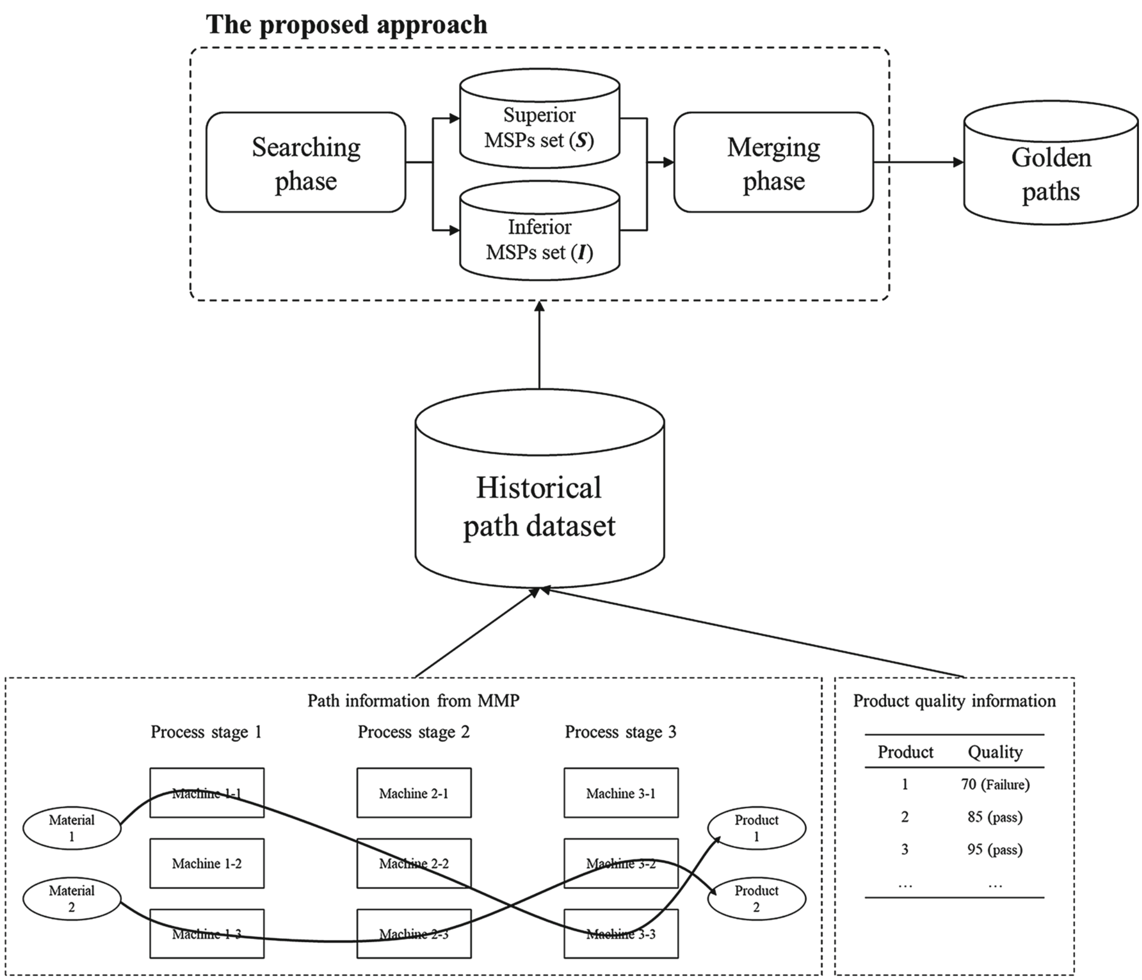


Fig. 2 A schematic overview of the research

Table 3 List of indices, parameters, and variables in this study

Indices, parameters, and variables	
N	The number of process stages in the path dataset
i	The index of ith process stage (i = 1, 2, …, N)
j_i	The number of machines at the ith process stage
j	The index of jth machines (j = 1, 2, …, j_i for each i)
M_i	Set of machines at the jth process stage, $M_i = \{i-1, i-2, \ldots, i-j_i\}$ for each i
T	The number of instances(paths) in the path dataset
$\boldsymbol{P}_t$	The t^{th} path, $\boldsymbol{P}_t \in M_1 \times M_2 \times \ldots \times M_N = \prod_{i=1}^{N} M_i$ where t = 1, 2, …, T
q_t	The product quality of the tth path
$\boldsymbol{sp}$	An MSP, $\boldsymbol{sp} \subset \boldsymbol{P}_t$ and $2 \leq \boldsymbol{sp}^* < N$
f	The minimum number of appearances of any MSPs (support threshold)
k_u	The greatest lower bound of product quality for superior MSPs
k_l	The least upper bound of product quality for inferior MSPs
$\boldsymbol{sp^s}$	A superior MSP
$\boldsymbol{sp^I}$	An inferior MSP
$\boldsymbol{S}$	The superior MSP set, $\boldsymbol{S} = \{\boldsymbol{sp}_1^s, \boldsymbol{sp}_2^s, \ldots, \boldsymbol{sp}_u^s, \ldots, \boldsymbol{sp}_U^s\}$ where U is the number of elements in $\boldsymbol{S}$
$\boldsymbol{I}$	The inferior MSP set, $\boldsymbol{I} = \{\boldsymbol{sp}_1^I, \boldsymbol{sp}_2^I, \ldots, \boldsymbol{sp}_v^I, \ldots, \boldsymbol{sp}_V^I\}$ where V is the number of elements in $\boldsymbol{I}$
$\boldsymbol{GP}$	A set of derived golden paths
$x_{\boldsymbol{sp},ij}$	1 if machine j at the i^{th} process stage of an $\boldsymbol{sp}$ is used, 0 otherwise
$y_{\boldsymbol{sp},t}$	1 if an $\boldsymbol{sp}$ is a subset of $\boldsymbol{P}_t$, 0 otherwise

* $\|a\|$ means the cardinality of a

paths on the basis of the two sets. The details are explained in Sects. 3.2.1 and 3.2.2, respectively. Table 3 lists indices, parameters, and variables used in this study.

Proposed approach

Searching phase

The objective of the searching phase is to extract a superior MSPs set ($\boldsymbol{S}$) and an inferior MSPs set ($\boldsymbol{I}$) from a historical path dataset. $\boldsymbol{S}$ and $\boldsymbol{I}$ are constructed by using three input parameters: f, k_u, and k_l. Table 3 shows the description of the three input parameters. Figure 3 shows how the searching phase is conducted according to the flowchart. First, the values of the input parameters are determined. Second, the searching phase extracts $\boldsymbol{sp}$ s. The searching procedure is iteratively conducted according to the length of $\boldsymbol{sp}$ (called n). At the beginning of the searching procedure, n is initialized to N-1. Once an $\boldsymbol{sp}$ with the length n is selected, then the procedure examines the following equation:

$$\sum_{t=1}^{T} y_{\boldsymbol{sp},t} \geq f$$

This equation expresses whether the number of appearances of an $\boldsymbol{sp}$ exceeds the support threshold f. If this equation is satisfied, then a statistical test is conducted to examine whether the $\boldsymbol{sp}$ belongs to $\boldsymbol{S}$ or $\boldsymbol{I}$. The hypothesis of the test for $\boldsymbol{S}$ ($\boldsymbol{I}$) is given as follows respectively:

$$H_0 : \mu_{\boldsymbol{sp}} = k_u (\mu_{\boldsymbol{sp}} = k_l)$$

$$H_1 : \mu_{\boldsymbol{sp}} > k_u (\mu_{\boldsymbol{sp}} < k_l),$$

where $\mu_{\boldsymbol{sp}}$ is the population mean of product quality of the paths containing $\boldsymbol{sp}$.

If H_0 is rejected at a certain significance level α, then $\boldsymbol{sp}$ is added to $\boldsymbol{S}$ ($\boldsymbol{I}$). If no more $\boldsymbol{sp}$ satisfies the equations, then n becomes $n-1$ and the same procedure is repeated to extract superior (inferior) $\boldsymbol{sp}$(s). If the search length reaches less than two, then the searching phase is terminated.

Merging phase

In this phase, the proposed approach combines superior MSPs ($\boldsymbol{sp}^{\boldsymbol{S}}$) in deriving golden paths. Figure 4 shows how the merging procedure is performed in a flowchart, where it starts with choosing two $\boldsymbol{sp}^{\boldsymbol{S}}$ s from $\boldsymbol{S}$. To merge the two $\boldsymbol{sp}^{\boldsymbol{S}}$ s, the following criteria should be satisfied.

Criteria 1. At least one common machine exists between the $\boldsymbol{sp}^{\boldsymbol{S}}$ s.

$$\bigcap_{u=1}^{2} \boldsymbol{sp}_u^{\boldsymbol{S}} \neq \emptyset$$

Criteria 2. Different machines at the same process stage within the $\boldsymbol{sp}^{\boldsymbol{S}}$ s cannot exist at the same time.

$$\sum_{j=1}^{j_i} x_{\left(\bigcup_{u=1}^{2} \boldsymbol{sp}_u^{\boldsymbol{S}}\right),ij} = 1 \text{ for all } i \text{ in } \left(\bigcup_{u=1}^{2} \boldsymbol{sp}_u^{\boldsymbol{S}}\right)$$

Criteria 3. The merged sequence pattern should not include any $\boldsymbol{sp}^{\boldsymbol{I}}$ **s** from $\boldsymbol{I}$.

$$\left(\bigcup_{u=1}^{2} \boldsymbol{sp}_u^{\boldsymbol{S}}\right) \cap \boldsymbol{sp}_v^{\boldsymbol{I}} \neq \boldsymbol{sp}_v^{\boldsymbol{I}} \text{ for all } v$$

If a merged pattern satisfies all criteria, then the merging phase checks the length of the merged pattern. If the length is equal to N, then it becomes a golden path. Otherwise, the current merged pattern is inserted to $\boldsymbol{S}$, and the merging phase

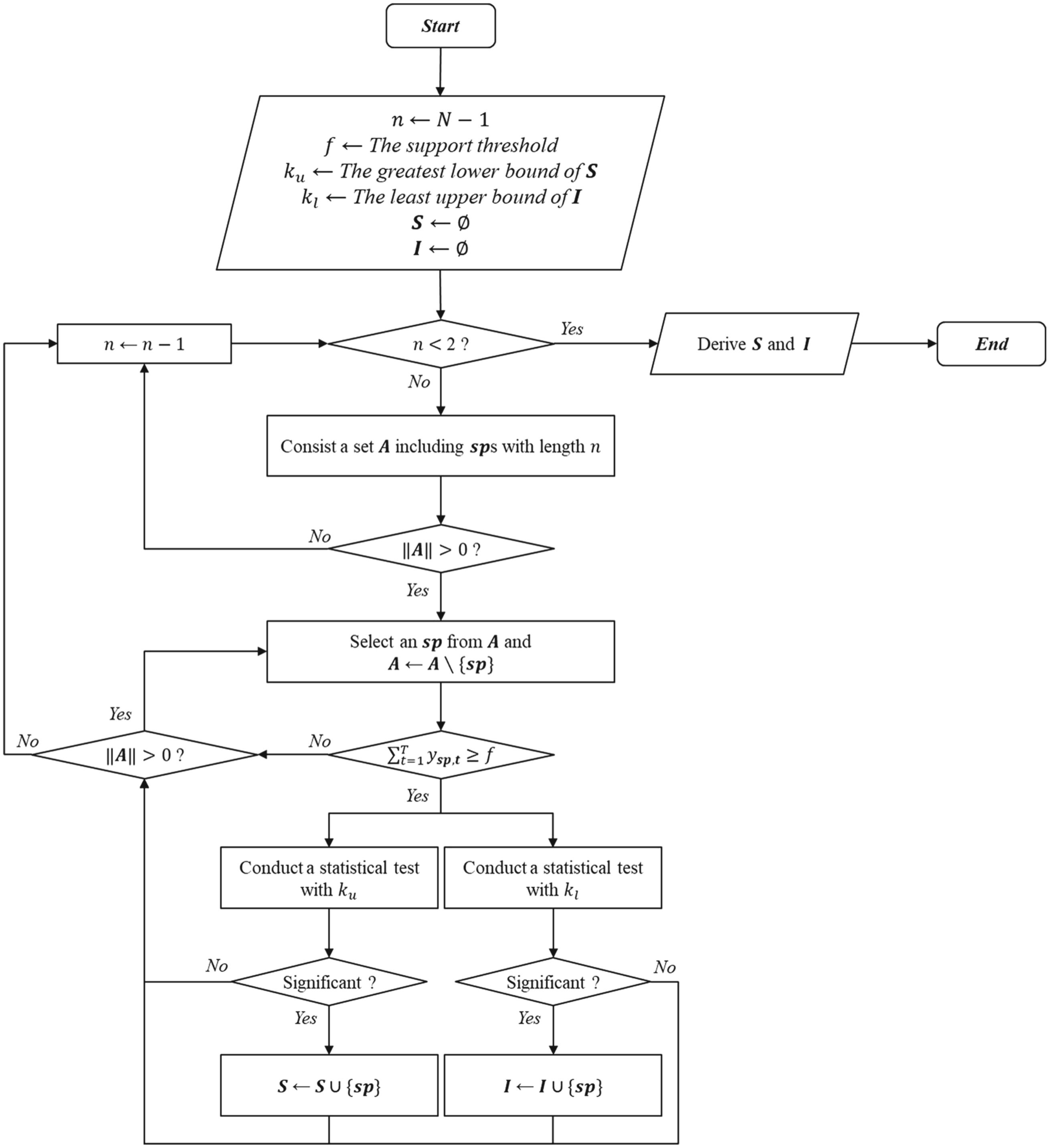


Fig. 3 A flowchart of the searching phase

repeats the merging procedure according to the flowchart in Fig. 4 until no more golden path is derived.

The following description outlines how the three criteria are applied in the merging phase. For example, in Table 2, suppose that the following $\boldsymbol{sp}^{\boldsymbol{S}}$ s belong to $\boldsymbol{S}$: {1–2, 2–4}, {2–4, 3–5}, {1–2, 2–2, *n*–3}, and {*n*–5, …, *N*–2} and the following $\boldsymbol{sp}^{\boldsymbol{I}}$ belongs to $\boldsymbol{I}$: {*n*–1, *N*–3}. On the one hand, if we merge {1–2, 2–4} and {2–4, 3–5}, then all the criteria are satisfied. The two patterns have a common element {2–4} (*Criteria* 1). No different machines exist in the same process stage within the two patterns (*Criteria* 2), and the merged pattern {1–2, 2–4, 3–5} does not include inferior MSP (*Criteria* 3). On the other hand, if we merge {1–2, 2–4} and {1–2, 2–2, *n*–3}, then merging is not possible because *Criteria* 2 is

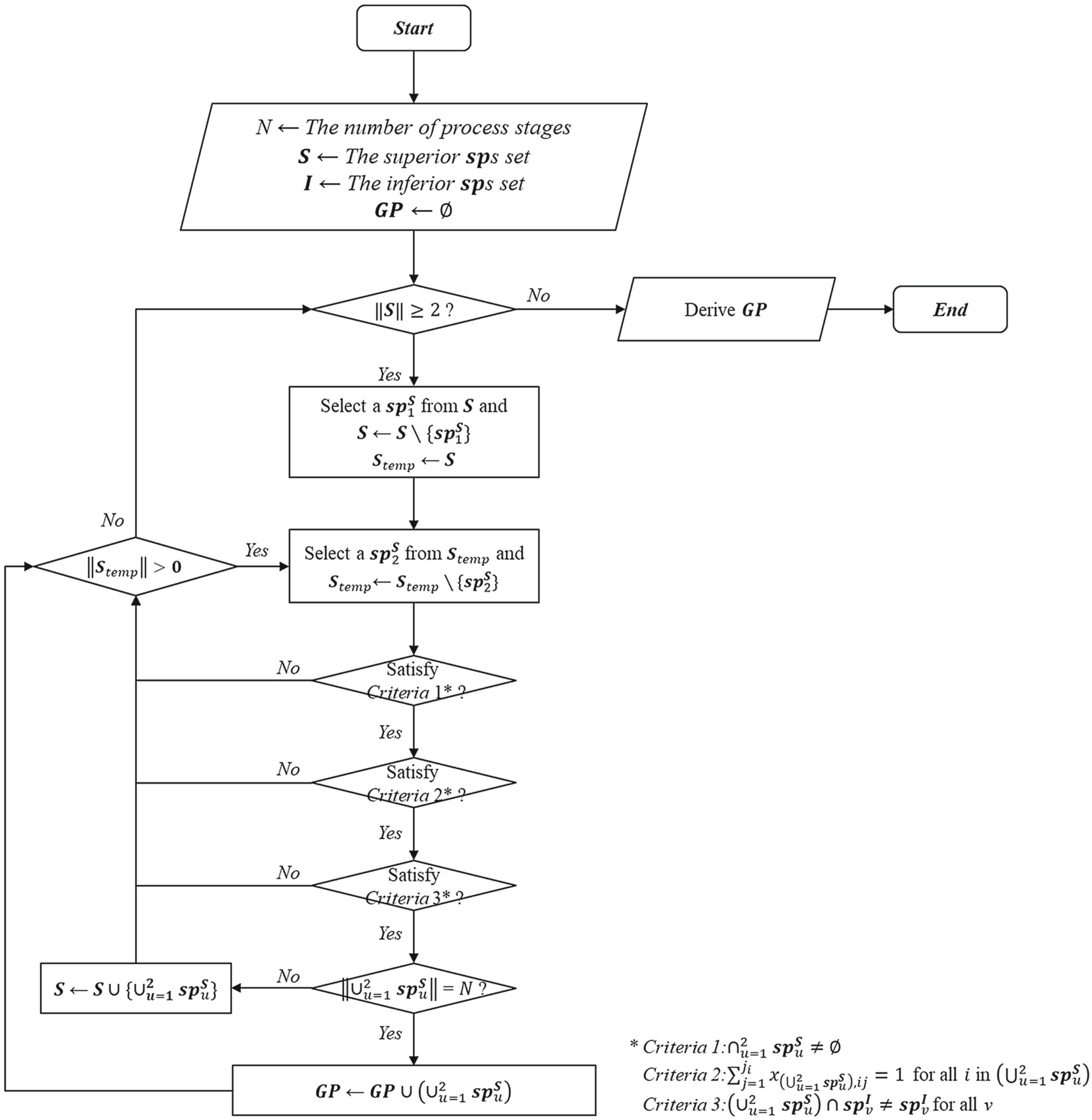


Fig. 4 A flowchart of the merging phase

not satisfied even if *Criteria 1* is satisfied because {2–4} of {1–2, 2–4} and {2–2} of {1–2, 2–2, *n*–3} are incompatible.

Illustrative example

To understand how the proposed approach works, the entire procedure is described with a simple example (Table 4). Suppose that an MMP consists of four process stages ($N = 4$), and each process stage has three machines. Sixteen paths and each quality of the product are given as the historical path dataset ($T = 16$). Each cell under the columns of the process stage represents the assigned machine at each process stage. To derive the golden paths, the input parameters f, k_u, and k_l are determined by 2, 71, and 69, respectively. One sample T test is employed to test the following hypothesis with a significance level of 0.3 ($\alpha = 0.3$).

$$H_0: \ \mu_{\boldsymbol{sp}} = 71, \ \mu_{\boldsymbol{sp}} = 69 \ \text{(for } \boldsymbol{S} \text{ and } \boldsymbol{I}\text{, respectively)}$$

Table 4 A path dataset for the illustrative example

Paths	Process stage				Quality
	1	2	3	4	
$\boldsymbol{P}_1$	1–3	2–1	3–1	4–2	66.14
$\boldsymbol{P}_2$	1–2	2–1	3–1	4–2	71.97
$\boldsymbol{P}_3$	1–3	2–1	3–2	4–2	68.43
$\boldsymbol{P}_4$	1–2	2–1	3–2	4–1	76.61
$\boldsymbol{P}_5$	1–2	2–3	3–3	4–2	76.54
$\boldsymbol{P}_6$	1–2	2–3	3–3	4–3	76.97
$\boldsymbol{P}_7$	1–2	2–1	3–3	4–1	80.27
$\boldsymbol{P}_8$	1–1	2–1	3–1	4–2	67.68
$\boldsymbol{P}_9$	1–1	2–3	3–3	4–1	74.88
$\boldsymbol{P}_{10}$	1–3	2–3	3–2	4–3	70.20
$\boldsymbol{P}_{11}$	1–1	2–3	3–1	4–2	68.88
$\boldsymbol{P}_{12}$	1–1	2–2	3–2	4–2	71.79
$\boldsymbol{P}_{13}$	1–3	2–1	3–1	4–3	70.57
$\boldsymbol{P}_{14}$	1–2	2–1	3–1	4–3	76.80
$\boldsymbol{P}_{15}$	1–3	2–1	3–2	4–3	72.76
$\boldsymbol{P}_{16}$	1–1	2–3	3–3	4–2	72.64

$$H_1:\ \mu_{\boldsymbol{sp}} > 71,\ \mu_{\boldsymbol{sp}} < 69\ \ (\text{for } \boldsymbol{S} \text{ and } \boldsymbol{I}, \text{respectively})$$

The proposed approach begins the searching phase to discover $\boldsymbol{sp}^{\boldsymbol{S}}$ s and $\boldsymbol{sp}^{\boldsymbol{I}}$ s. The searching length n is initialized to N-1(= 3), and no $\boldsymbol{sp}$ s are discovered at $n = 3$. When n shortens to 2, five $\boldsymbol{sp}$ s are discovered as $\boldsymbol{sp}^{\boldsymbol{S}}$: {1–2, 2–1}, {1–2, 3–3}, {1–3, 4–3}, {2–3, 3–3}, and {2–1, 4–3}. For example, in the case of paths that contain an $\boldsymbol{sp}$ = {1–2, 2–1}, This $\boldsymbol{sp}$ appears four times from the given dataset (the 2nd, 4th, 7th, and 14th rows of Table 4). The q_t containing this $\boldsymbol{sp}$ is 71.97, 76.61, 80.27, and 76.80, respectively. The statistical test rejects the null hypothesis of $\boldsymbol{S}$ at the significance level $\alpha = 0.3$. Therefore, we put this $\boldsymbol{sp}$ into $\boldsymbol{S}$. On the other hand, two $\boldsymbol{sp}$ s are discovered as $\boldsymbol{sp}^{\boldsymbol{I}}$: {2–1, 4–2} {3–1, 4–2}. For example, in the case of paths that contain an $\boldsymbol{sp}$ = {2–1, 4–2}, this $\boldsymbol{sp}$ appears four times from the given dataset (the 1st, 2nd, 3rd, and 8th rows of Table 4). The q_t containing this $\boldsymbol{sp}$ is 66.14, 71.97, 68.43, and 67.68, respectively. The statistical test rejects the null hypothesis of $\boldsymbol{I}$ at the significance level $\alpha = 0.3$. Therefore, we put this $\boldsymbol{sp}$ into $\boldsymbol{I}$.When n shortens to 1, the searching procedure meets the termination condition, and the merging phase begins. Five $\boldsymbol{sp}^{\boldsymbol{S}}$ s are present in $\boldsymbol{S}$(= {1–2, 2–1}, {1–2, 3–3}, {1–3, 4–3}, {2–3, 3–3}, and {2–1, 4–3}) and two $\boldsymbol{sp}^{\boldsymbol{I}}$ s are present in $\boldsymbol{I}$(= {2–1, 4–2}, {3–1, 4–2}). Suppose that {1–2, 2–1} and {2–1, 4–3} are selected from $\boldsymbol{S}$ in deriving a golden path. The element {2–1} is common in the two selected patterns (*Criteria* 1), and an incompatible element does not exist between the two patterns (*Criteria* 2). Lastly, the merged pattern {1–2, 2–1, 4–3} does not include any $\boldsymbol{sp}^{\boldsymbol{I}}$ s from $\boldsymbol{I}$ (*Criteria* 3). Thus, we can merge the two $\boldsymbol{sp}^{\boldsymbol{S}}$ s as {1–2, 2–1, 4–3}. Given that the length of the merged pattern is not equal to N, the same procedure is repeated to the current merged pattern. If another pattern {1–2, 3–3} is selected, then a common element exists ({1–2}) and no incompatible element exists. Moreover, the merged pattern {1–2, 2–1, 3–3, 4–3} does not include any $\boldsymbol{sp}^{\boldsymbol{I}}$ s from $\boldsymbol{I}$. After merging, the length of the merged pattern is equal to N. Hence, one golden path(= {1–2, 2–1, 3–3, 4–3}) is derived in this example.

In Sects. 4 and 5, the experimental results are provided to demonstrate the performance of the proposed approach. In Sect. 4, the performance of the proposed approach is evaluated with sensitivity analysis by applying it to a hypothetical path dataset. In Sect. 5, the practical viability of the proposed approach is validated by applying it to the SETFI dataset.

Experiment with hypothetical example

Hypothetical data generation

To generate the hypothetical path dataset, we assume that there are six process stages. For each process stage, we arrange 5, 3, 7, 3, 5, and 7 machines respectively. Therefore, a total of 11,025 paths are possible. On the basis of Lee et al. (2015), the product quality according to each path is calculated as follows:

$$q_t = Y_b \times \prod_{i=1}^{N} Y_{m_{ij}}$$

where Y_b is the base product quality and $Y_{m_{ij}}$ is the operational performance of the jth machine at the ith process stage. Y_b is given by a random value at each q_t drawn from the uniform distribution of [0.99, 1], reflecting the natural randomness. Each machine in a process stage may have a different performance of the operation. Thus, each $Y_{m_{ij}}$ is given by a random value drawn from the uniform distribution of [0.97, 1]. Once all paths are generated, they are divided into a learning set and a test set in a ratio of 1:1, respectively. The learning set is constructed via a random sampling without replacement from the 11,025 possible paths. The test set consists of the remaining paths.

Experimental design

The performance of the proposed approach is evaluated by the accuracy and the ranking of the derived golden paths. The accuracy is determined by how many among the paths derived by the proposed approach meet the product quality standard of the golden paths. The ranking of the derived golden paths indicates how superior the quality of products produced according to the derived golden paths is within the population. For example, if a path is ranked at 5%, then the product quality produced by this path is approximately the 551st highest quality out of the 11,025 paths in the hypothetical path dataset. Therefore, if the quality of products of the derived golden paths is higher than k_u, then the usefulness of the proposed approach is demonstrated.

The proposed approach is applied to the learning set generated in Sect. 4.1. To analyze the performance of the proposed approach, sensitivity analysis is conducted on the basis of various values of the three input parameters (f, k_u, and k_l). Each input parameter value consists of three levels; f has 2, 4, and 8; k_u has the values corresponding to the top 10%, 15%, and 20% levels of the values of product quality included in the learning set; k_l has the values corresponding to the bottom 10%, 15%, and 20% levels of the values of product quality included in the learning set. A full factorial design ($= 3^3$) is used to test the statistical significance of an effect that the three input parameters exert on the performance of the proposed approach. Accordingly, 27 experimental cases are built and 100 repeated experiments are conducted for each case. Each iterative experiment involves the construction of a learning set and a test set. The repeated measure ANOVA (Weinfurt 2000) is conducted to verify the statistical significance of the performance of the proposed approach according to the levels of input parameters. To give more insight into the statistical results, the Bonferroni test is conducted as a post hoc analysis.

Results and discussion

The derived golden paths are divided into two types. The first type of golden path, called EG, refers to the paths that match within the learning set. The second type of golden path, called NG, refers to the paths that match within the test set. Figure 5 summarizes the description of the structure in the derived golden paths. The ranking of the derived golden paths is compared according to the type of the derived golden paths.

Table 5 summarizes the experimental results according to the levels of the three input parameters. The accuracy is approximately 95% on average in the entire experiments (Table 5-a). The sensitivity analysis has shown that if each value of f, k_u, and k_l increases, then the accuracy of the proposed approach is improved. The highest accuracy ($= 98.812\%$) is achieved when f is 8, k_u is the top 10% value of the product quality in the learning set, and k_l is the bottom 20% value of the product quality in the learning set. Under all the conditions of the input parameters, the accuracy showed an average of above 90%. Table 5-b and 5-c show the ranking of the derived golden paths according to

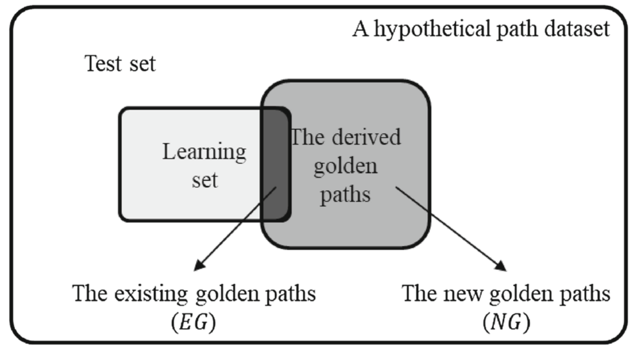


Fig. 5 The structure diagram of the types of the derived golden paths

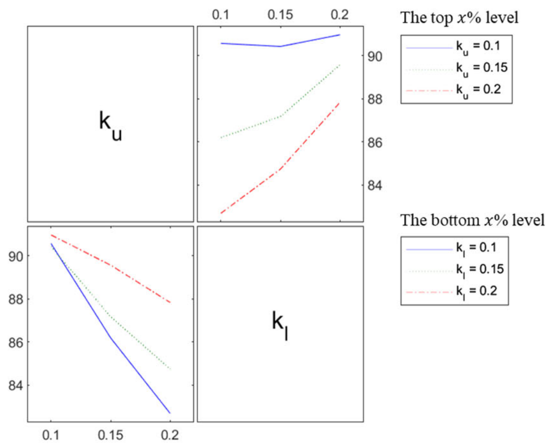
EG and NG, respectively. The sensitivity analysis results for the ranking of the derived golden paths for EG and NG show a tendency similar to the accuracy. The ranking of the derived golden paths increases when f, k_u, and k_l increase, respectively. It is noted that k_u means the lowest quality level that a path becomes the golden path. Therefore, the ranking of the derived golden paths inevitably depends on the value of k_u.

Table 6 shows the results of repeated-measure ANOVA for the performance of the proposed approach at a significance level $\alpha = 0.05$. In the case of accuracy, f and $(k_u \times k_l)$ have significant effects (Table 6a). The Bonferroni test is performed to examine the mean differences among the levels of each parameter. According to the test result, the accuracy at $f = 8$ is significantly higher than those at $f = 2$ and $f = 4$. Figure 6-a shows the interaction plot of k_u and k_l from the Bonferroni test. The accuracy increases when k_u changes from the top 20% level to the top 10% level or when k_l changes from the bottom 20% level to 10% level. In the case of the ranking of the derived golden paths, the results show that $(f \times k_u)$ and $(k_u \times k_l)$ are statistically significant in EG and NG (Table 6b and c). Figure 6b and c show the interaction plots of $(f \times k_u)$ and $(k_u \times k_l)$ for EG and NG, respectively. These interaction plots indicate that the ranking of the derived golden paths improves when f changes from 2 to 8, k_u changes from the top 20% level to 10% level, and k_u changes from the bottom 10% level to 20% level, respectively.

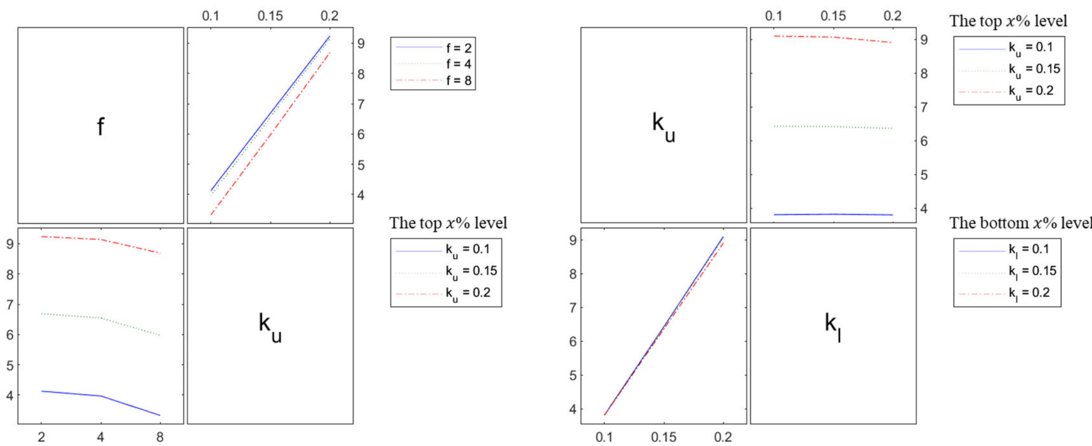
The findings from the experimental results with sensitivity analysis are as follows. The accuracy of the proposed approach is approximately 95% on average, implying that it derives the golden paths reasonably well. In terms of the ranking of the derived golden paths, when the minimum product quality level to become a golden path (i.e., k_u) is given, according to the repeated-measure ANOVA results, the proposed approach derives superior golden paths, in which the larger the value of f and the higher the value of k_l, respectively. As the value of the input parameters increases, the number of derived golden paths could decrease.

Table 5 Sensitivity analysis results of the performance of the proposed approach for the three input parameters

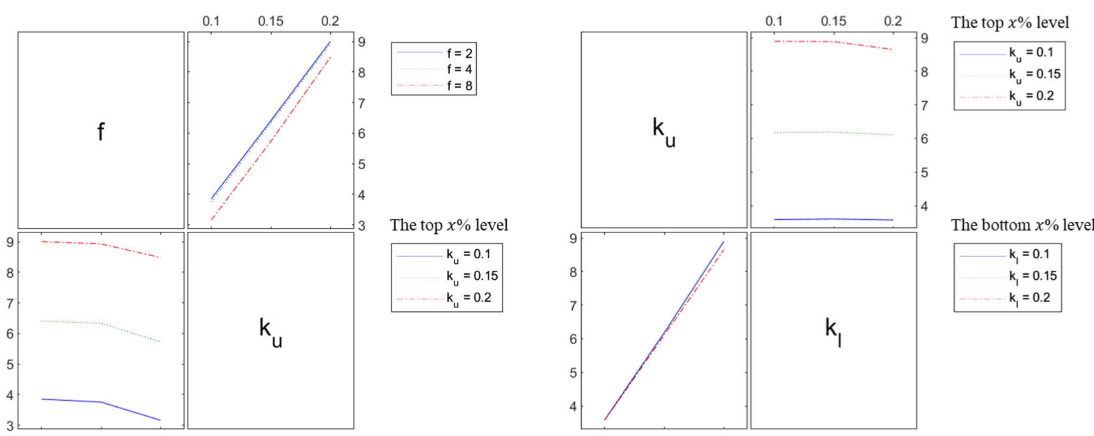
f	k_u	k_l			Overall mean
		Bottom 10%	Bottom 15%	Bottom 20%	
a. Accuracy					
2	Partial Sum.	94.229	94.61	95.483	94.774
	Top 10%	97.439	96.678	96.939	97.019
	Top 15%	93.776	94.75	95.487	94.671
	Top 20%	91.471	92.4	94.022	92.631
4	Partial Sum.	93.909	94.841	95.582	94.777
	Top 10%	97.066	97.305	97.573	97.315
	Top 15%	93.766	94.944	95.371	94.693
	Top 20%	90.895	92.274	93.801	92.323
8	Partial Sum.	95.607	95.514	95.989	95.703
	Top 10%	98.43	98.111	98.812	98.451
	Top 15%	96.522	95.37	95.961	95.951
	Top 20%	91.868	93.061	93.193	92.707
Total		94.581	94.988	95.684	95.085
b. Ranking of the derived golden paths belonging to EG					
2	Partial Sum.	5.577	5.611	5.564	5.584
	Top 10%	3.105	3.249	3.153	3.169
	Top 15%	5.589	5.581	5.563	5.577
	Top 20%	8.038	8.004	7.975	8.006
4	Partial Sum.	**5.335**	5.233	5.118	5.228
	Top 10%	2.961	2.756	2.723	2.813
	Top 15%	5.089	5.125	5.028	5.081
	Top 20%	7.954	7.817	7.602	7.791
8	Partial Sum.	4.519	4.593	4.707	4.607
	Top 10%	2.269	2.091	2.029	2.130
	Top 15%	4.136	4.485	4.859	4.493
	Top 20%	7.152	7.204	7.235	7.197
Total		5.144	5.146	5.130	5.140
c. Ranking of the derived golden paths belonging to NG					
2	Partial Sum.	4.819	4.921	4.868	4.869
	Top 10%	2.421	2.656	2.531	2.536
	Top 15%	4.812	4.817	4.828	4.819
	Top 20%	7.224	7.292	7.243	7.253
4	Partial Sum.	4.806	4.642	4.652	4.700
	Top 10%	2.475	2.197	2.300	2.324
	Top 15%	4.590	4.503	4.542	4.545
	Top 20%	7.355	7.226	7.116	7.232
8	Partial Sum.	4.138	4.294	4.372	4.268
	Top 10%	2.004	1.870	1.746	1.873
	Top 15%	3.665	4.166	4.550	4.127
	Top 20%	6.745	6.847	6.821	6.804
Total		4.588	4.619	4.631	4.613



(a) The interaction plot of k_u and k_l for the accuracy



(b) The interaction plots of f and k_u(left) and k_u and k_l(right) for the ranking of the derived golden paths belonging to EG



(c) The interaction plots of f and k_u(left) and k_u and k_l(right) for the ranking of the derived golden paths belonging to NG

Fig. 6 **a** The interaction plot of k_u and k_l for the accuracy. **b** The interaction plots of f and k_u(left) and k_u and k_l(right) for the ranking of the derived golden paths belonging to EG. **c** The interaction plots of f and k_u(left) and k_u and k_l(right) for the ranking of the derived golden paths belonging to NG

Table 6 The repeated-measure ANOVA result results of the performance of the proposed approach for the three input parameters

Source	**A**. Accuracy					**B**. Ranking of the derived golden paths belonging to EG					**C**. Ranking of the derived golden paths belonging to NG				
	Sum. Sq.	DF	Mean Sq.	F	P-value	Sum. Sq.	DF	Mean Sq.	F	P-value	Sum. Sq.	DF	Mean Sq.	F	P-value
f	144.76	2	72.38	33.07	0.000*	239.52	2	119.76	5416.41	0.000*	211.08	2	105.54	4,102.93	0.000*
k_u	14,009.42	2	7,004.71	2,896.46	0.000*	12,237.39	2	6,118.7	264,742.8	0.000*	12,245.44	2	6,122.74	182,809.6	0.000*
k_l	4,144.88	2	2,072.44	806.26	0.000*	4.3	2	2.15	120.34	0.000*	7.63	2	3.82	135.75	0.000*
$f \times k_u$	7.35	4	1.84	0.81	0.521	5.4	4	1.35	57.24	0.000*	3.4	4	0.85	29.29	0.000*
$f \times k_l$	5.73	4	1.43	0.64	0.631	0.13	4	0.03	1.51	0.197	0.1	4	0.03	0.87	0.481
$k_u \times k_l$	1,742.41	4	435.6	172.84	0.000*	2.98	4	0.75	34.45	0.000*	5.28	4	1.32	43.33	0.000*
$f \times k_u \times k_l$	7.58	8	0.95	0.41	0.917	0.12	8	0.02	0.71	0.679	0.09	8	0.01	0.38	0.932
Error (f)	433.36	198	2.19	1	0.5	4.38	198	0.02	1	0.5	5.09	198	0.03	1	0.5
Error (k_u)	478.84	198	2.42	1	0.5	4.58	198	0.02	1	0.5	6.63	198	0.03	1	0.5
Error (k_l)	508.95	198	2.57	1	0.5	3.54	198	0.02	1	0.5	5.57	198	0.03	1	0.5
Error ($f \times k_u$)	901.59	396	2.28	1	0.5	9.34	396	0.02	1	0.5	11.48	396	0.03	1	0.5
Error ($f \times k_l$)	881.29	396	2.23	1	0.5	8.42	396	0.02	1	0.5	11.63	396	0.03	1	0.5
Error ($k_u \times k_l$)	998.04	396	2.52	1	0.5	8.57	396	0.02	1	0.5	12.07	396	0.03	1	0.5
Error ($f \times k_u \times k_l$)	1,847.86	792	2.33	1	0.5	17.29	792	0.02	1	0.5	23.73	792	0.03	1	0.5

*Indicates that p value is < 0.05

Table 7 The illustration of the SETFI dataset

Paths	Process stage						Quality level
	1	2	3	4	…	300	
$\boldsymbol{P}_1$	1–2	2–5	3–9	4–1	…	300–2	2,953.627
$\boldsymbol{P}_2$	1–5	2–4	3–1	4–4	…	300–3	2,777.102
$\boldsymbol{P}_3$	1–4	2–1	3–6	4–5	…	300–5	2,770.209
$\boldsymbol{P}_4$	1–5	2–2	3–10	4–3	…	300–4	2,727.662
$\boldsymbol{P}_5$	1–10	2–5	3–4	4–3	…	300–10	2,788.332
$\boldsymbol{P}_6$	1–2	2–3	3–10	4–2	…	300–5	2,794.530
…	…	…	…	…	…	…	…
$\boldsymbol{P}_{4000}$	1–6	2–3	3–6	4–3	…	300–1	2,921.653

Thus, an appropriate choice of the input parameter values is required. The experimental results based on the hypothetical path dataset demonstrate that the proposed approach can be adopted and applied in practice. In addition, it should be noted that the ranking of the derived golden paths belonging to NG meets the standard of golden paths. This result indicates that the proposed approach is also capable of discovering new golden paths that do not exist in a given path dataset.

Experiment with SETFI dataset

Data description and preprocessing

The proposed approach is applied to the SETFI dataset (AA&YA and Intel, 2008) to validate its practical viability. The SETFI dataset is representative of an MMP and closely mimics the actual semiconductor manufacturing process (Kerdprasop and Kerdprasop 2014; Moldovan et al. 2018). This dataset contains 4000 observations of the wafer fabrication process, which has 300 process stages, and each wafer goes through a sequence of operations. Two to 20 machines are deployed in each process stage. Table 7 describes the form of this dataset.

A statistical test is adopted in this study to select significant process stages before applying the proposed approach owing to the following two reasons. First, the number of observations in the SETFI dataset is much smaller than that of possible paths. Even if we suppose that two machines at every process stage exist, the number of possible paths becomes 2^{300}. Therefore, the number of observations needed to find frequent MSPs is insufficient. Second, although the operational performance of the machines could be different at any process stage, not all process stages may affect the product quality significantly. Thus, significant process stages need to be screened. Thus, the Kruskal–Wallis test is adopted to screen significant process stages. The Kruskal–Wallis test is

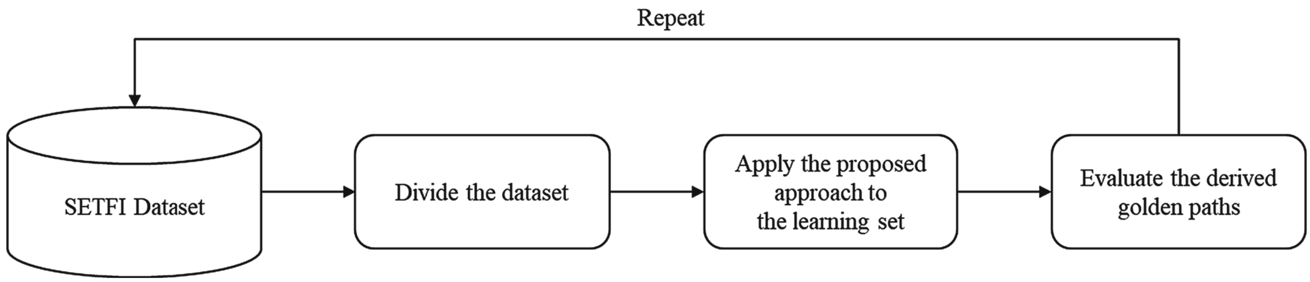


Fig. 7 The overall procedure of the experiment

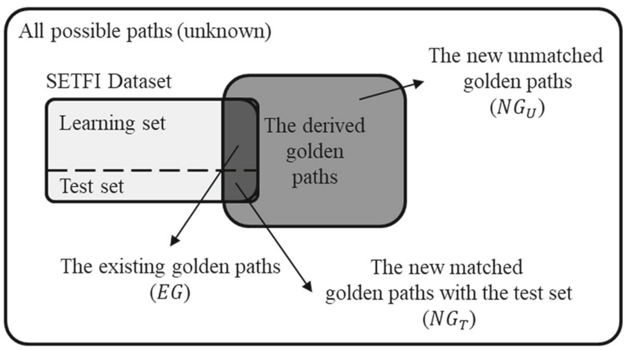


Fig. 8 The structure diagram of the types of the derived golden paths from the SETFI dataset

a widely used non-parametric test to check whether the medians of all sample groups are equal (Chien et al. 2014). In this sense, the Kruskal–Wallis test examines whether machines at each process stage have significant effects on product quality. As a result, six process stages (the 15th, 25th, 133th, 147th, 198th, and 246th process stage) show significant differences in the product quality among their machines at the 0.05 significance level. The number of machines in the selected process stages is 7, 8, 6, 7, 8, and 5, respectively. The total number of possible paths is 94,080 ($= 7 \times 8 \times 6 \times 7 \times 8 \times 5$) from the selected six process stages.

Experimental design

Figure 7 illustrates the overall procedure of the experiment performed on the SETFI dataset. As in Sect. 4.2, the SETFI dataset of the selected process stages is divided into two groups. By random sampling without replacement, half of the total observations in the SETFI dataset consist of the learning set, and the remaining observations comprise the test set. The proposed approach is applied to the learning set to derive the golden paths. Figure 8 illustrates the types of the derived golden paths from the SETFI dataset. Similarly with Sect. 4, if a derived golden path matches within the learning set, it is classified as EG. Otherwise, two types of derived golden paths are possible. If the path matches within the test set, then it is classified as NG_T. If the path matches within neither the learning set or test set, it is classified as NG_U. The performance of the proposed approach is evaluated by the paths belonging to EG and NG_T, because only the quality level of products of those paths could be known. This procedure is repeated a hundred times.

To derive golden paths, f is set to 2, and k_u and k_l are set correspondingly to the top 10% and bottom 20% values, respectively, of q_t from the learning set. According to the normality test result of q_t, the distribution of q_t does not follow a normal distribution. Thus, the Wilcoxon-signed rank test is adopted as the statistical test at the searching phase of the proposed approach. The Wilcoxon signed-rank test is one of the widely used nonparametric tests for analyzing the matched-pair and one-sample problem. It can be used for any distribution and is much less sensitive to outliers than the paired t-test (Woolson 2008; Ma et al. 2010). The significance level α is determined as 0.2 for the Wilcoxon-signed rank test.

Results and discussion

Table 8 summarizes the experimental results with the SETFI dataset. During the iterative implementations, the number of derived golden paths is 56 on average. Among the derived golden paths, 5 golden paths belong to EG and the remaining 51 golden paths are the new ones that do not exist in the learning set. Among the new golden paths, on average, 3 golden paths belong to NG_T, and the remaining 48 golden paths belong to NG_U. The performance of the proposed approach is evaluated in terms of the accuracy and the ranking of the derived golden paths, as in Sect. 4.3. It should be noted that the performance of the proposed approach is evaluated only for the case of EG and NG_T. The average accuracy of the proposed approach for EG and NG_T is 100% and 99.8%, respectively. In addition, in the case of the ranking of the derived golden paths, on average, EG is 2.351% and NG_T is 2.140%, which satisfies the standard of the golden paths. These experimental results demonstrate that the proposed approach works well with deriving golden paths. Given that the SETFI dataset closely represents the practical MMP circumstance, these experimental results imply that the proposed approach is able to derive the golden paths in practice.

Table 8 The experimental results of the SETFI dataset

	The number of derived paths			Accuracy			Ranking of the derived golden paths		
	Mean	Min.	Max.	Mean	Min.	Max.	Mean	Min.	Max
EG	5 (3.087)	1	12	100	100	100	2.351	8.975	0.025
NG_T	3 (1.686)	0	6	99.8	99	100	2.140	8.086	0.033
NG_U	48 (37.784)	3	181	–			–	–	–

*The value in parentheses means a standard deviation

**Rounding down number

The quality level of the golden paths belonging to NG_U is unknown. One of the possible reasons is that the number of paths in the learning set (= 2000) is extremely smaller than that of possible paths from the six process stages (= 94,080). However, on the basis of the average ranking of the derived golden paths belonging NG_T and the experimental results under the same input parameters condition in Sect. 4, the remaining golden paths belonging to NG_U are also expected to have a similar ranking with NG_T

In practice, managers usually want to maximize the productivity of an MMP by utilizing all possible paths. However, the quality of products along the paths is not the same, and paths without existing product quality information are relatively risky to utilize in production because the quality of the products along these paths is unknown. Thus, the proposed approach can provide a rationale for screening the paths to produce superior-quality products. Thus, if the desired productivity level is given, then managers could be able to plan operational scheduling of an MMP to maximize the quality of products being produced while maintaining the desired productivity level by utilizing the golden paths derived from the proposed approach.

Concluding remarks

This study proposes a heuristic approach to derive golden paths from an MMP to improve product quality. It takes into account all process stages and different performances of the machines in MMP when deriving the golden paths. The derived golden paths could contribute to maximizing product quality. The performance and practical viability of the proposed approach are demonstrated by experimental results using a hypothetical path dataset and the SETFI dataset, respectively.

The proposed approach could provide several managerial implications. First, the proposed approach is easy to use. With the aid of the advancement of information technologies such as IoT (Internet of things) technology, collecting and storing historical path data of products is now common among recent manufacturing companies (Lee et al. 2019). Therefore, this recent trend makes the proposed approach more accessible to managers, because it only uses historical path data of products. Second, the proposed approach can be flexibly used according to a manager's purpose. The values of the input parameters of the proposed approach could be determined by a manager's circumstances. For example, if the value of f is increased, then paths for stably producing a product with a certain level of quality could be derived. Moreover, the proposed approach can be applied universally to various types of MMP, such as machining, assembly, or additive manufacturing process, because it is designed by considering the basic characteristics of an MMP. Third, the proposed approach could be used in various ways in the operation of an MMP. For example, in designing the dispatching rules of raw materials, the golden paths derived by the proposed approach could be able to optimize product quality and productivity simultaneously. If a target level of productivity is given, then a dispatching rule for maximizing product quality by assigning the golden paths could be designed. Conversely, if a target level of product quality is given, then a dispatching rule for maximizing productivity could be designed by adjusting the proportion of the use of the derived golden paths.

Some limitations of the proposed approach require further investigation. Given that the proposed approach lacks the modeling between product quality and the operational performances of machines in an MMP, it does not estimate the individual effects of machines on product quality. To relax this limitation, combining the paths with operation settings of the machines when extracting the superior and MSPs at the searching phase of the proposed approach could be helpful. The information about the operation setting of the machines could assist to estimate the difference between the superior and inferior MSPs. Thus, utilizing the operation settings of the machines could complement the reliability of the extracted MSPs. Next, the proposed approach is unable to provide real-time information because it is based on the historical path dataset. Hence, it cannot take into account that the operational performance of the machines may vary in real-time during the production process. On-line quality management methods, such as real-time inspection and statistical process monitoring, could be employed together to alleviate this limitation. By combining on-line quality management methods and the proposed approach, changes in the operational performance of the machines could be detected in a timely manner and corresponding golden paths can be derived. Lastly, although the SETFI dataset used in this study closely mimics the real semiconductor manufacturing process, the proposed approach has not been tested on a real industrial case. A real case study needs to be conducted as future research to validate the practical usefulness of the proposed approach.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2019R1A2C1007834). This work was supported under the framework of international cooperation program managed by the National Research Foundation of Korea (NRF-2016K2A9A2A11938390).

References

Advanced Analytics—Intel, SETFI: Manufacturing data: Semiconductor tool fault isolation. *Causality Workbench Repository*. Retrieved June 8, 2020, from http://www.causality.inf.ethz.ch/repository.php.

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*. https://doi.org/10.1109/ICDE.1995.380415.

Bera, S., & Mukherjee, I. (2016). A multistage and multiple response optimization approach for serial manufacturing system. *European Journal of Operational Research, 248*(2), 444–452. https://doi.org/10.1016/j.ejor.2015.07.018.

Chien, C.-F., Chang, K.-H., & Wang, W.-C. (2014). An empirical study of design-of-experiment data mining for yield-loss diagnosis for semiconductor manufacturing. *Journal of Intelligent Manufacturing, 25*(5), 961–972. https://doi.org/10.1007/s10845-013-0791-5.

Cunha, C. D., Agard, B., & Kusiak, A. (2006). Data mining for improvement of product quality. *International Journal of Production Research, 44*(18–19), 4027–4041.

Du, S., Yao, X., & Huang, D. (2015). Engineering model-based Bayesian monitoring of ramp-up phase of multistage manufacturing process. *International Journal of Production Research, 53*(15), 4594–4613. https://doi.org/10.1080/00207543.2015.1005247.

Eger, F., Reiff, C., Brantl, B., Colledani, M., & Verl, A. (2018). Correlation analysis methods in multi-stage production systems for reaching zero-defect manufacturing. *Procedia CIRP, 72,* 635–640. https://doi.org/10.1016/j.procir.2018.03.163.

Fahey, W., Jeffers, P., & Carroll, P. (2020). A business analytics approach to augment six sigma problem solving: A biopharmaceutical manufacturing case study. *Computers in Industry, 116,* 103153. https://doi.org/10.1016/j.compind.2019.103153.

Garofalakis, M., Rastogi, R., & Shim, K. (1999). SPIRIT: Sequential pattern mining with regular expression constraints. In *Proceedings of the 25th International Conference on Very Large Data Bases*, 223–234.

Huang, Q., & Shi, J. (2004). Stream of variation modeling and analysis of serial-parallel multistage manufacturing systems. *Journal of Manufacturing Science and Engineering, 126*(3), 611–618. https://doi.org/10.1115/1.1765149.

Huff, M. (2020). *Yield analysis and quality assurance and control methods used in microsystems manufacturing (Process Va)*. Cham: Springer. https://doi.org/10.1007/978-3-030-40560-1_9.

Irrera, I., & Vieira, M. (2014). A practical approach for generating failure data for assessing and comparing failure prediction algorithms. *Proceedings of IEEE Pacific Rim International Symposium on Dependable Computing, PRDC*. https://doi.org/10.1109/PRDC.2014.19.

Jin, R., & Liu, K. (2013). Multimode variation modeling and process monitoring for serial-parallel multistage manufacturing processes. *IIE Transactions, 45*(6), 617–629. https://doi.org/10.1080/0740817X.2012.728729.

Ju, F., Li, J., Xiao, G., Arinez, J., & Deng, W. (2015). Modeling, analysis, and improvement of integrated productivity and quality system in battery manufacturing. *IIE Transactions (Institute of Industrial Engineers), 47*(12), 1313–1328. https://doi.org/10.1080/0740817X.2015.1005777.

Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). Mining association rules for the quality improvement of the production process. *Expert Systems with Applications, 40*(4), 1034–1045. https://doi.org/10.1016/j.eswa.2012.08.039.

Kerdprasop, K., & Kerdprasop, N. (2013). Performance analysis of complex manufacturing process with sequence data mining technique. *International Journal of Control and Automation, 6*(3), 301–311.

Kerdprasop, K., & Kerdprasop, N. (2014). Tool sequence analysis and performance prediction in the wafer fabrication process. *International Journal of Systems Applications, Engineering & Development, 8,* 268–276.

Kimemia, J., & Gershwin, S. B. (1985). Flow optimization in flexible manufacturing systems. *International Journal of Production Research, 23*(1), 81–96. https://doi.org/10.1080/00207548508904692.

Lee, H., Kim, C. O., Ko, H. H., & Kim, M. (2015). Yield prediction through the event sequence analysis of the die attach process. *IEEE Transactions on Semiconductor Manufacturing, 28*(4), 563–570. https://doi.org/10.1109/TSM.2015.2487540.

Lee, D.-H., Lee, C.-H., Choi, S.-H., & Kim, K.-J. (2019). A method for wafer assignment in semiconductor wafer fabrication considering both quality and productivity perspectives. *Journal of Manufacturing Systems, 52,* 23–31. https://doi.org/10.1016/j.jmsy.2019.05.006.

Lim, H. K., Kim, Y., & Kim, M. (2017). Failure prediction using sequential pattern mining in the wire bonding process. *IEEE Transactions on Semiconductor Manufacturing, 30*(3), 285–292. https://doi.org/10.1109/TSM.2017.2721820.

Ma, M., Wong, D. S., Jang, S., & Tseng, S. (2010). Fault detection based on statistical multivariate analysis and microarray visualization. *IEEE Transactions on Industrial Informatics, 6*(1), 18–24.

Moldovan, D., Chifu, V., Pop, C., Cioara, T., Anghel, I., & Salomie, I. (2018). Chicken Swarm Optimization and deep learning for manufacturing processes. In *2018 17th RoEduNet conference: networking in education and research (RoEduNet)*. (pp. 1–6). IEEE.

Nakata, K., Orihara, R., Mizuoka, Y., & Takagi, K. (2017). A comprehensive big-data-based monitoring system for yield enhancement in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing, 30*(4), 339–344. https://doi.org/10.1109/TSM.2017.2753251.

Oztemel, E., & Gursev, S. (2020). Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing, 31*(1), 127–182. https://doi.org/10.1007/s10845-018-1433-8.

Psarommatis, F., May, G., Dreyfus, P. A., & Kiritsis, D. (2020). Zero defect manufacturing: State-of-the-art review, shortcomings and future directions in research. *International Journal of Production Research, 58*(1), 1–17. https://doi.org/10.1080/00207543.2019.1605228.

Rokach, L., Romano, R., & Maimon, O. J. J. (2008). Mining manufacturing databases to discover the effect of operation sequence on the product quality. *Journal of Intelligent Manufacturing, 19*(3), 313–325. https://doi.org/10.1007/s10845-008-0084-6.

Sellami, C., Miranda, C., Samet, A., Bach Tobji, M. A., & de Beuvron, F. (2020). On mining frequent chronicles for machine failure prediction. *Journal of Intelligent Manufacturing, 31*(4), 1019–1035. https://doi.org/10.1007/s10845-019-01492-x.

Sellami, C., Samet, A., & Bach Tobji, M. A. (2019). Frequent Chronicle Mining: Application on Predictive Maintenance. In *Proceedings—17th IEEE international conference on machine learning and applications*, *ICMLA 2018*, 1388–1393. https://doi.org/10.1109/ICMLA.2018.00226.

Taha, K. (2019). An effective approach for identifying defective critical fabrication path. *Cogent Engineering*, *6*(1), 1575636. https://doi.org/10.1080/23311916.2019.1575636.

Wang, Z., & Wang, P. (2016). Accelerated failure identification sampling for probability analysis of rare events. *Structural and Multidisciplinary Optimization, 54*(1), 137–149. https://doi.org/10.1007/s00158-016-1405-6.

Weinfurt, K. P. (2000). Repeated measures analysis: ANOVA, MANOVA, and HLM. In *Reading and understanding MORE multivariate statistics*. (pp. 317–361). American Psychological Association.

Woolson, R. F. (2008). Wilcoxon singed-rank test. *Wiley encyclopedia of clinical trials*. https://doi.org/10.1002/9780471462422.eoct979.

Wuest, T., Irgens, C., & Thoben, K.-D. (2014). An approach to monitoring quality in manufacturing using supervised machine learning on product state data. *Journal of Intelligent Manufacturing, 25*(5), 1167–1180. https://doi.org/10.1007/s10845-013-0761-y.