Hindawi
Complexity
Volume 2021, Article ID 5593374, 13 pages
https://doi.org/10.1155/2021/5593374

*Research Article*

# Hierarchical Factor Analysis Methodology for Intelligent Manufacturing

**Hyun Sik Sim**

*Department of Industrial & Management Engineering, Kyonggi University, Suwon 16227, Republic of Korea*

Correspondence should be addressed to Hyun Sik Sim; hssim@kgu.ac.kr

Received 21 February 2021; Accepted 25 May 2021; Published 9 June 2021

Academic Editor: Huihua Chen



To realize intelligent manufacturing, a controllable factory must be built, and manufacturing competitiveness must be achieved through the improvement of product quality and yield. The yield in the micromanufacturing process is gaining importance as a management factor used in deciding the production cost and product quality as product functions becomes more sophisticated. Because the micromanufacturing process involves manufacturing products through multiple steps, it is difficult to determine the process or equipment that has encountered failure, which can lead to difficulty in securing high yields. This study presents a structural model for building a factory integration system to analyze big data at manufacturing sites and a hierarchical factor analysis methodology to increase product yield and quality in an intelligent manufacturing environment. To improve the product yield, it is necessary to analyze the fault factors that cause low yields and locate and manage the critical processes and equipment factors that affect these fault factors. However, yield management is a difficult problem because there exists a correlation between equipment, and in the sequence of process equipment that the lot passed through, the downstream and the upstream cause complex faults. This study used data-mining techniques to identify suspected processes and equipment that affect the yield of products in the manufacturing process and to analyze the key factors of the equipment. Ultimately, we propose a methodology to find the key factors of the suspected process and equipment that directly affect the implementation of the intelligent manufacturing scheme and the yield of the product. To verify the effect of key parameters of critical processes and equipment on the yield, the proposed methodology was applied to actual manufacturing sites.

## 1. Introduction

Owing to the rapid evolution of technological environments and the gradual decrease in development periods, technological gaps in micromanufacturing processes have been gradually shrinking. In particular, in the case of semiconductor and printed circuit board (PCB) products, as customer demands diversify and demand levels increase, the process of high integration, high functionalization, and microfabrication of products becomes increasingly complex, and thus customized production is required. This complicated product structure and process increase the production cost and limit the maintenance of high yields and quality. To achieve a high product yield, quality control has been performed in the manufacturing process for a long time by introducing a statistical process control technique that checks for faults by measuring the circuit inspection of the substrate or measuring the plating thickness or line width after the product has been processed. However, it is practically impossible to inspect all production lots because it requires considerable cost and effort; thus, sample inspection is performed in the major process of the product. In the flip chip ball grid array (FCBGA) manufacturing process which is the target of this study, approximately 30 fault types were examined during inspection after the etching process. Faults discovered during the inspection process are important factors that lead to high production costs when the process progresses downstream; furthermore, they increase the overall production costs. Activities that minimize faults and maximize yield are necessary. Therefore, it is important to analyze the fault types that are the major causes of low yields and to accurately find and manage the equipment and

processes where faults occur. Each process in the FCBGA manufacturing line is set up with equal equipment and is a complex process. Therefore, it is difficult to determine which process and equipment are the main faults that cause the low yield. Furthermore, the FCBGA process is not only a suspected process that causes faults but a complex process through several processes and equipment.

First, this study analyzes the suspected processes and machines that affect the yield of the manufacturing process based on the data of equipment routing paths traversed by each manufacturing lot. Suspected machines include not only a single piece of equipment that influences the fault but also a complex group of equipment that leads to a higher level of faults as the downstream participates in the upstream. Such a problem is attributed to a phenomenon in which the possibility of faults increases because of the chemical and physical correlation between the processes. Second, this study analyzes parameters (among the various parameters of the suspected machines) that directly affect the faults of the product.

Furthermore, the analysis of big data at the manufacturing site needs to be preceded by the establishment of an environment in which the lot history of critical processes, inspection/measurement, and equipment data is gathered and fed back in real time through sensors and the Internet of things (IoT). Conversely, an environment that can collect and control the data of the manufacturing site in real time, which is the core function of a factory integration system, needs to be established first. The key to implementing a factory integration system is to construct a platform that can support the interconnection between internal and external resources in a factory based on manufacturing IoT technology, which optimizes manufacturing and services [1]. For this platform configuration, the real-time collection of production data and the analysis and application of manufacturing big data must be performed [2], and an analysis methodology for complex process structures is required [3]. In addition, the complexity and problems of big data management in the IoT field were introduced [4], and a digital design and simulation method for an automated factory were presented [5].

This study presents a factory integration architecture model of a manufacturing factory required for analyzing manufacturing big data. In section 2, the related literature is presented. Section 3 presents a factory integration system implementation plan, analyzes the suspected processes and machines, and presents a hierarchical analysis model that identifies the key factors of suspected machines. Section 4 describes the experimental and data analysis processes and the results of the proposed model. Section 5 discusses the conclusions and further research topics.

## 2. Related Research

FCBGA-PCB and semiconductor processes comprise dozens of unit processes, such as circuits, plating, and etching, and specific processes are repeated. To analyze the manufacturing process with these characteristics, various studies have been conducted on the methodology for detecting and diagnosing defects in product quality at manufacturing sites for a long time. For univariate quality control, the control charts presented by Montgomery and Douglas are commonly used; however, the increase in control variables has confronted many constraints [6].

In the case of multivariate quality control, a method of reducing dimensions using principal component analysis for numerous variables occurring in the process and monitoring product quality with multivariate statistics such as Hotelling's $T^2$ was proposed [7]. In a study on finding the equipment and equipment variables that affect the yield in multistage manufacturing processes, Ma et al. applied a statistical method to the chemical vapor deposition process to increase the yield based on important variables that affect the quality variables [8]. In addition, a methodology for monitoring and predicting equipment status by analyzing data collected from sensors [9], a method for integrated maintenance according to equipment performance reduction [10], and a reliability evaluation method for a fuzzy multistate manufacturing system based on ESFN (extended stochastic flow network) are presented [11]. In addition, an intelligent control system that monitors process parameters and detects abnormalities [12] and a framework for recognizing and obtaining big data for each product manufacturing cycle were presented [13].

However, these methods have limitations in that they analyze only a single process without considering the phenomenon that multiple machines of multiple processes simultaneously affect the yield while going through many processes. The approaches mentioned so far are all applicable methods for analyzing single processes and equipment factors.

However, Sim [14] presented a methodology to locate suspected machines by analyzing the cumulative effect of not only a single machine in a complex microfabrication process but also a number of machines in multiple processes. However, this method has a limitation in that although the suspected process or machine that affects the yield (fault) has been analyzed, the equipment factor to be managed in the actual site cannot be known.

To analyze the big data of manufacturing sites, all devices and equipment in the factory should be interconnected, and data collection and analysis should be based on such interconnectivity. Thus, functions connecting all equipment at the site and collecting and analyzing the required data can be regarded as the most basic functions of factory integration [15]. Previously, a wide variety of construction methods have been proposed for the establishment of smart factories and equipment management systems of manufacturing companies. Such existing methods are limited to implementing a smart factory using information systems and implementing individual modules required in the field. No studies on the methodology for the implementation of a practical intelligent factory by linking the big data of the manufacturing site have been reported so far.

Therefore, this study presents a novel methodology for determining the factors of the suspected machine that affects the yield by applying the hierarchical factor analysis methodology and for building the required intelligent manufacturing scheme of the manufacturing site.

## 3. Methodology

*3.1. Factory Integration System.* The Manufacturing Enterprise Solutions Association (MESA) defines the manufacturing execution system (MES) as follows: "MES delivers information that enables the optimization of production activities from order launch to finished goods, monitors, controls, and reports factory activities with accurate real-time data" [16].

As the MES model connects the manufacturing site and the enterprise system, the ANSI/ISA-95 (2000) model, which is an enterprise control integration model proposed by MESA and ISA (Instrument Society of America), is most often used [17]. A factory integration system is an intelligent factory where information and communications technologies are applied to the equipment and machines for automated manufacturing processes and where factory automation, IoT, and big data are combined [18]. To implement such an intelligent factory, all necessary information regarding the manufacturing site should be organically connected through IoT, and predictable manufacturing should be enabled through big data analysis [19, 20].

The integration-based factory integration system model provided by MESA and ISA can be categorized into three levels, as illustrated in Figure 1. At the control level, the necessary information is collected and controlled by operating equipment and machines and managing IoT or sensors. At the management level, WIP tracking, schedule management, equipment engineering system (EES) management, and process control are performed. The analysis level serves the function of analyzing the manufacturing and equipment, processing, and inspecting data collected from the manufacturing site; it can be categorized into manufacturing analysis and big data analysis. Thus, the factory integration system can be implemented only when the equipment is controlled (see ⑦ in Figure 1) and when equipment management (see ⑤ in Figure 1) and big data analysis (see ② in Figure 1) modules are realized in addition to the existing MES functions.

In the hierarchical factor analysis stage, first, a data set is constructed by collecting data necessary for analysis such as yield, work history, and equipment parameters for each product and lot. Analysis stage 1 (Layer1) determines the suspected processes and machines that affect the product yield by using a data-mining algorithm. Stage 2 identifies the critical equipment parameters that can be managed. Stage 3 utilizes the fault detection and classification module or control function to perform real-time monitoring of the critical parameters found in Stage 2; the system is configured such that an interlock may be set in the case of anomaly detection.

This study proposes a methodology to determine the suspected processes and machines that affect the yield and to analyze the critical parameters of suspected equipment by proceeding with Stages 1 and 2. Studies on the management and control of the derived critical parameters and a big data platform will be conducted as a follow-up.

*3.2. Hierarchical Analysis*

*3.2.1. Hierarchical Analysis Methodology.* In this study, the suspected machine and critical parameters that directly affect the product yield in a factory integration environment were analyzed using two-stage layers. After identifying the fault items that cause low yields, the processes that affect the yield are identified.

In Layer 1, the suspected processes and machines that affect the yield (fault parameters) of the inspection process are searched, and in Layer 2 the study of Layer 1 is further advanced, and the relationship between the critical parameters of the suspected machines and the process parameter ($y$) is analyzed to determine the factors that cause the fault (see Figure 2).

Layer 1 uses an association analysis to preprocess data regarding the equipment trace data before finding the suspected machines that cause these faults. The equipment trace data are also called process history, which refers to the sequence of process equipment that one lot has passed. If 1 indicates that the lot has passed through a specific piece of equipment and 0 indicates otherwise, the trace can be regarded as a sequence comprising 0 s and 1 s. The partial least squares with variable importance of the projection (PLS-VIP) method is applied to equipment trace data to solve the multicollinearity existing between machines. In addition, because there are numerous machines, a number of rules are created if the association rules are applied immediately; thus, the important machines that cause the defect are first selected through PLS-VIP analysis, and the suspected machines that affect the yield are found using the association rules. In addition, not only a single machine but also the relationship that a plurality of suspected machines, such as a single machine, affects the fault was analyzed.

In Layer 2, a linear regression equation is derived using the parameters of the suspected machines, and the relationship between the suspected machine and process parameters was analyzed. The output ($y$) of the suspected process found in Layer 1 was used as the dependent variable, and the equipment variable that affected $y$ was set as the independent variable. That is, by analyzing the relationship between the process parameter ($y$) and the equipment variable ($x$), the equipment variables that affect the process parameter are found.

*3.2.2. Layer 1 Analysis.* In this section, using the PLS-VIP analysis, an important machine that causes defects is first selected, and then association analysis is used to find suspected machines that affect the yield. In addition, the cumulative effect methodology was applied in consideration of the association analysis and complexity of the process.

First, the PLS analysis method was used to solve the multicollinearity problem found in the multivariate analysis, whereas the PLS-VIP method was used to select only the machines with high contribution to defects and applied the association rule for ease of analysis. The PLS analysis, which

is commonly used, derives latent variables that simultaneously explain independent and dependent variables, enabling a more meaningful analysis. PLS is a robust model for noise and missing values, and it can be applied to a small amount of data and has the advantage of handling various types of variables, such as nominal and continuous variables. When selecting an important variable, the PLS regression analysis considers the degree of influence of the independent variable on the latent variable and the influence of the latent variable on the dependent variable simultaneously. The variable importance of projection (VIP) score of the independent variable is expressed as follows [21]:

$$\mathrm{VIP}_j = \sqrt{\frac{k\sum_{a=1}^{a^*}\left[\left(b_a^2 t_a' t_a\right)\left(w_{aj}/\left\|w_a\right\|\right)^2\right]}{\sum_{a=1}^{a^*}\left(b_a^2 t_a' t_a\right)}}, \quad j = 1, \ldots, k. \tag{1}$$

In equation (1), $k$ is the number of independent variables and $a$ is the latent variable. $a^*$ indicates the number of latent variables generated by the PL model. In equation (1), variable $w_{aj}$ is the loading weight of variable $j$ when the latent variable $a$ is used. $b_a^2 t_a' t_a$ comprises the variance represented by latent variable $a$ and y-loading ($b_a$), which can be considered the contribution of latent variable $t_a$ to the dependent variable $y$. In conclusion, $\mathrm{VIP}_j$ can be considered a measure to evaluate the importance of variable $j$ based on the variance explained by the latent variable and the importance of the independent variable constituting the latent variable.

Second, the machines that affect the yield are analyzed using association analysis for the machines selected above. Association rules help extract useful hidden rules from vast amounts of data. Rules divide the relationships between items into left-hand side (lhs) and right-hand side (rhs), and they are expressed in the {lhs⟶rhs} format. In this study, lhs refers to the process and equipment sequence and rhs is a good or bad class.

The association analysis shows different items, $a$ and $b$, in the $\{a \longrightarrow b\}$ format, where $a$ denotes the process and equipment sequence and $b$ denotes the class. The association rule strength is a measure of the support and confidence values of the rule [22]. In this study, the support is defined as the ratio of many faults that have passed through a specific equipment among all lots. Confidence is the ratio in which a and b occur together when a occurs and refers to the frequency of faults occurring when passing through certain machines.

Finally, in this study, the cumulative effect algorithm was used in consideration of the association analysis and complexity of the process. The core of this analysis is not only to discover a single suspected machine but also to grasp the extent to which the downstream affects faults along with the upstream and simultaneously manage the suspected machines to increase the yield. Conversely, the accuracy of the upstream and the accuracy when the downstream is included in the upstream must be compared. If the accuracy ratio increases upon the participation of the downstream in the process, then compared with the accuracy when the downstream does not participate (i.e., the accuracy of the upstream only), the accuracy is above a fixed level, which means that the rule is a cumulative factor. In this study, this ratio is called the cumulative effect, and the cumulative effect is expressed as follows:

$$\text{CE value } (\%) = \frac{\text{accuracy of downstream} - \text{accuracy of upstream}}{\text{accuracy of upstream}} \times 100\%. \tag{2}$$

The cumulative effect is measured in rules with a length of two or more. In this process, the rules are expressed in a tree form to easily understand the relationship between upstream and downstream. In the tree, which shows the inclusion relationships between rules, a rule is placed on the upper layer of the tree as its length increases. In this study, this was defined as an upper rule. The subsets constituting the upper rule are called lower rules, and the lower rules naturally have smaller lengths than the upper rules. The author followed the methodology of Sim [14] in Layer 1 and extended it one step further and applied it to the failure mode ($Y_1$) of the FCBGA products.

Figure 3 shows a relationship tree model expressed by the rules generated using the Apriori algorithm [23] when minimum confidence and minimum support are 0.05, and the minimum lift is set to a value greater than 1. Rules of length 1 in the relationship tree represent a single factor. Therefore, the single factors in Figure 3 are the rules ($x3 : e3$) and ($x10 : a10$). In Figure 3, the numerical value expressed on the right side of the rule constituting the tree refers to the number of good and bad products found when passing through the equipment represented by the rule. In Figure 3, ($x3 : e3$, $x10 : a10$) [4, 46] indicates that when the lot passed through the equipment ($e3$) in the $x3$ process and the machine ($a10$) in the $x10$ process, the normal four times and 46 faults occurred. Therefore, the accuracy of this rule is 0.92. The number on the line connecting both rules indicates the rate of increase in accuracy between the upper and lower rules, and a positive value indicates that the accuracy increases when moving from the lower rule to the upper rule immediately above. In Figure 3, to examine the cumulative effect of the downstream ($x10 : a10$) on upstream ($x3 : e3$), the accuracies of the rules ($x3 : e3$, $x10 : a10$) and ($x3 : e3$) are used. The cumulative effect represents the ratio of the accuracy of the upstream and the accuracy increases upon the participation of the downstream. In Figure 3, the cumulative effect between the two rules is 17.5% (=0.137/0.783 × 100%). Because this value is greater than the minimum cumulative effect threshold, the rule ($x3 : e3$, $x10 : a10$) becomes a cumulative factor.

*3.2.3. Layer 2 Analysis.* The previous section described a method for analyzing the suspected processes and equipment that affected the quality variables. This section identifies the relationship between the output of the suspected process described in the previous section and the relevant equipment parameters. The process and equipment parameters are linearly related, and a regression model is selected as the analysis method for determining the equipment parameters that affect the process parameters (output) of the suspected processes [24]. To describe the dependent variable in the regression analysis, the relationship with the independent variable that affects it is expressed as a functional expression and is mainly used to predict the change in the dependent variable based on the change in the independent variable [25]. This study employs a regression model in the case of two or more independent variables; thus, the model is referred to as a multiple regression model [26]. This results in the following equation:

$$\begin{aligned} y_i &= \beta_0 + \beta_i x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \\ \varepsilon_i &= y_i - \beta_0 - \beta_i x_i, \end{aligned} \tag{3}$$

where $y_i$ denotes the process parameter, $x_i$ is the equipment variable, and $\varepsilon_i$ is the random error term. Additionally, $\beta_0$ denotes the intercept of the regression equation and $\beta_i$ is the slope, which can be estimated using $\beta_0$ and $\beta_i$ [27]. The least squares method is an estimation approach that minimizes the error between the actual value $y$ and the predicted value $\widehat{y}$. It is widely used to estimate the regression coefficients $\beta_0$ and $\beta_i$. The main reason for calculating the sum of squares of the error term is that even if a severe error occurs, the calculation result may indicate that almost no error is caused by the errors of the (+) and (−) values canceling each other out. The sum of squares for error (SSE), which represents the SSE terms, is expressed as follows [28]:

$$\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_i x_i\right)^2. \tag{4}$$

Herein, the estimated values of $\beta_0$ and $\beta_i$, that is, $\widehat{\beta}_0$ and $\widehat{\beta}_i$, can be derived using the least squares method. The condition of minimizing the SSE is that the partial derivatives of the SSE with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_i$ should satisfy 0. Conversely, it can be obtained by satisfying $(\text{dSSE}/\text{d}\widehat{\beta}_0) = 0$ and $(\text{dSSE}/\text{d}\widehat{\beta}_i) = 0$. This results in the following equation:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_i x_i + \widehat{\varepsilon}_i. \tag{5}$$

The coefficient of determination ($R^2$) for verifying the fitness of the model serves as a coefficient that indicates the contribution of the independent variable to describing the dependent variable in the regression equation. That is, it shows the extent to which the independent variable describes the change in the dependent variable. This results in the following equation:

$$\begin{aligned} R^2 &= \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \\ \text{SST} &= \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2, \\ \text{SSE} &= \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2. \end{aligned} \tag{6}$$

SST indicates the total variation, and it is expressed as SST = SSR + SSE. The sum of squares for regression (SSR) is a variation of a regression equation, and a variation amount can be explained by an estimated regression equation. If the SSR exceeds the SST, the regression equation can be used to explain the dependent variable. SSE represents the variation caused by the error. If the value of the SSE decreases, the variation decreases, indicating a strong statistical significance of the regression equation.

The regression analysis algorithm is executed through the following five stages [29]:

*Stage 1.* Prediction model selection and data definition: A multiple regression model was selected, and dependent and independent variables, as well as data properties, were defined.

*Stage 2.* Selection of critical variables using a variable selection method: The optimal value is selected using a stepwise variable selection method.

*Stage 3.* Model optimization: An optimal model was selected based on the validation data from the models generated by the training data after dividing the predefined training and validation data.

*Stage 4.* Verification of the statistical significance of the variables: To verify the significance of individual variables, a variable with a $p$ value of 0.05 is selected.

*Stage 5.* Target value prediction:

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \varepsilon_i. \tag{7}$$

The model can be regarded as valid when the estimated regression equation does not deviate by more than 0.05, with respect to the threshold value. Therefore, the key variables that affect the process parameters have a significant influence on the possibility of faults. In the above algorithm, the variable selection method employs a stepwise approach that supplements the drawbacks of the forward and backward methods. Although there are various methods, the reason for selecting a stepwise method is to minimize the number of variables and to select only good variables efficiently [30]. The stepwise method checks whether the existing variables can be eliminated at each step of adding a variable when the importance of each existing variable is lowered because of a

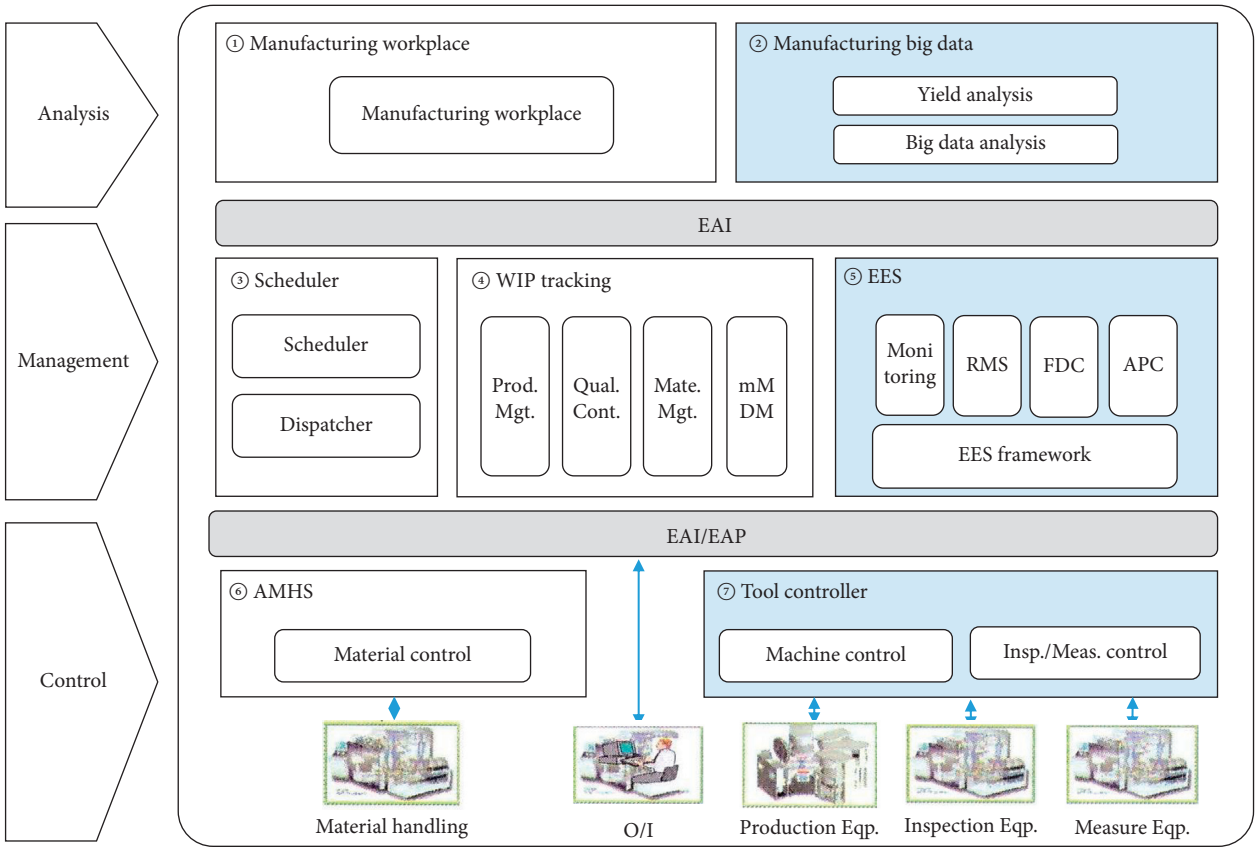


Figure 1: Factory integration system architecture.

newly added variable [31]. Thus, this study selected key equipment variables based on a stepwise variable selection method. Stepwise regression analysis estimates the regression coefficient using the least squares method and calculates the coefficient of determination to display the extent to which the regression model explains the given data.

## 4. A Case Study

### 4.1. Setup

*4.1.1. Layer 1 Setup.* The example in this section is the result of analyzing the suspected processes and machines that affect the failure mode ($Y_1$) of FCBGA products. Here, the failure mode ($Y_1$) refers to a defect item in the test process and is referred to as a quality variable. The FCBGA-PCB manufacturing line considered in the case study comprises 10 processes and 33 machines (see Table 1). The detailed process comprised nine processes in addition to the plating process ($x3$), which is a Layer 2 analysis process. Because the 10 processes comprise several machines for each process, the number of all possible combinations of trace types was calculated to be approximately 90,000. If we reorganize this combination by trace type, we can find approximately 300 trace types. In this study, the number of representative faults was calculated as the average value of all faults generated when the machines specified in the trace passed. After preprocessing, $x_{ij}$ constituting the trace for the 300 trace types was defined as the independent variable, and the number of representative faults was defined as the dependent variable.

*4.1.2. Layer 2 Setup.* Layer 2 analyzes the equipment variables that affect the process parameters. In this section, the target processes and equipment for the analysis are selected, and the equipment variables that affect the process parameters are identified. The analysis results for Layer 1 revealed that processes $x5$ and $x6$ were critical suspected processes, and that process $x5$ affected six $Y$ parameters. Although process $x6$ was included as a critical process that affected the $Y_1$ parameter, it was excluded from the experiment because the lot traceability of the data could not be secured. The regression model was constructed using the critical parameters of the suspected machine ($a5$) of process $x5$, and the relationship between the process parameters and equipment parameters was identified. The typical process parameters of process $x5$ include thickness, width, and space. Table 2 presents the target process, process parameters, and equipment parameters that affect them.

*4.2. Analysis Results for Layer 1.* In this section, the suspected machine is selected using the PLS-VIP value, and the single and cumulative effects are analyzed. First, based on the VIP value, we look at the degree of importance that the machines constituting the trace $\{x_{11}, x_{12}, \ldots, x_{ij}\}$ have on the fault of the quality variable $Y_1$. In the PLS regression analysis, the

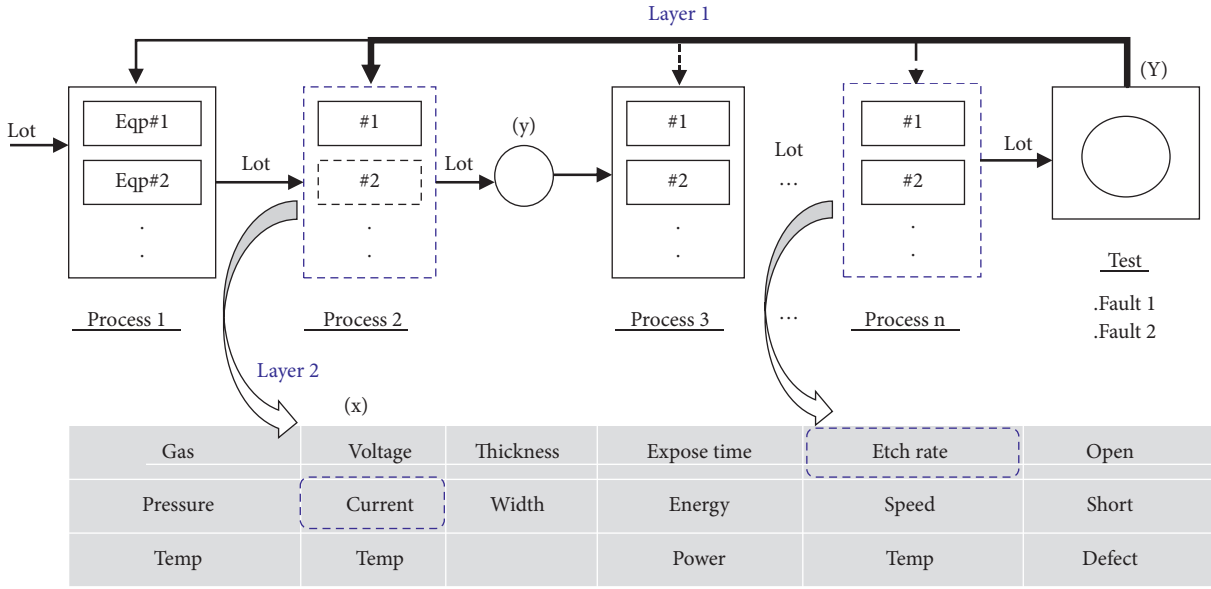


Figure 2: Hierarchical factor analysis framework.

number of latent variables was selected through five-fold cross validation, which is widely used to estimate prediction errors. The VIP value for the quality variable $Y_1$ can be determined from Figure 4. Generally, the mean of the square of the VIP value is 1; therefore, an independent variable greater than 1 is selected as a meaningful variable. Following the results of a study that showed good performance when the VIP value was between 0.8 and 1.2, in this experiment, a variable with a value of 0.8 or more was selected [32].

Figure 4 shows the VIP values of the 33 independent variables for quality variable $Y_1$. The (x8:b8) variable had the lowest VIP values of (0.10), and ($x5 : c5$) had the highest value of (2.14). Table 3 presents the suspected machine candidates selected for quality variable $Y_1$. Here, $Y_1$ denotes a major item among the defective items in the inspection process.

The association rule applies to suspected machines and machine groups that affect the yield of the quality variable selected above. First, to apply the association rule, the minimum confidence and minimum support parameters must be set. In this study, the experiment was conducted with both minimum confidence and minimum support set at 0.05. The value is set to such a low level because even a single fault can be a significant loss from the perspective of a company in an environment where the technological changes introduced above and the technology level between competitors are similar. Moreover, it might result in a suspected machine or machine group that causes potential faults beyond the limit and accumulates data even if it currently shows a low frequency. Under the support and confidence conditions set here, as a result of selecting rule sets with a min-lift value greater than 1, 19 rule sets were found.

Consequently, out of the 19 rules found, 15 rules have a confidence value of 0.8 or higher and three rules show a confidence value of 1. When rule generation is completed, the machine that affects the fault independently and the machine group that affects the fault together with the upstream and downstream are obtained from the generated rule based on the previously suggested algorithm. Figures 5 and 6 show the tree shape, composed of upper and lower rules based on the rule length to find the single factor and cumulative factor in quality variable $Y_1$.

To discover the cumulative factor in the relationship tree, we set the minimum cumulative effect threshold to 5%. That is, the cumulative factor was chosen by selecting the rules that showed a cumulative effect of 5% or more based on the accuracies before and after the downstream participation. Based on the results in Figure 5, the rule ($x1 : b1$, $x6 : a6$) refers to the upstream of the upper layer rule ($x1 : b1$, $x6 : a6$, $x9 : a9$). The cumulative effect of downstream ($x9 : a9$) is calculated to be 7.7% (=0.066/0.857 × 100%), which is greater than the minimum cumulative effect threshold; hence, rule ($x1 : b1, x6 : a6, x9 : a9$) becomes a cumulative factor. Figure 6 shows the relationship tree of parameter $y$ when the length of the rule is 2. From the figure, it is evident that if the lot goes through equipment $a5$ in process $x5$ and then through equipment $a6$ in process $x6$, then 100.0% of the faults will be found out of the total lot, and it is 10.1% higher than the fault detection performance by a single factor ($x5 : a5$). This implies that there is a performance. Table 4 presents the cumulative factors for the quality variable, accuracy, and cumulative effect values indicated by the cumulative factor.

There are six cumulative factors that cause faults in the quality variable $Y_1$, and the cumulative effect of these factors is distributed from 5.3% to 12.9%. An accuracy that indicates a relatively high cumulative factor can be observed, and the cumulative factors discovered in this experiment have an average accuracy of 87.7%. The cumulative factor ($x5 : a5$, $x6 : a6$) in Table 4 shows that faults are found in 100.0% of all the lots that go through equipment $a5$ in process $x5$ and then

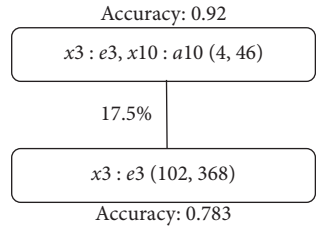


Figure 3: Relationship tree between upstream and downstream process.

Table 1: FCBGA-PCB processes and machines configuration.

| $x1$ | $x2$ | $x3$ | $x4$ | $x5$ | $x6$ | $x7$ | $x8$ | $x9$ | $x10$ |
|---|---|---|---|---|---|---|---|---|---|
| $a1$ | $a2$ | $a3$ | $a4$ | $a5$ | $a6$ | $a7$ | $a8$ | $a9$ | $a10$ |
| $b1$ | $b2$ | $b3$ | $b4$ | $b5$ | $b6$ | $b7$ | $b8$ | $b9$ | $b10$ |
| $c1$ | | $c3$ | $c4$ | $c5$ | $c6$ | $c7$ | $c8$ | $c9$ | |
| | | $d3$ | $d4$ | | | $d7$ | | | |
| | | $e3$ | | | | $e7$ | | | |

Table 2: Target process, process, and equipment parameters.

| Process 1 | Control variable | Process 2 | Control variable | Process 3 | Process parameter |
|---|---|---|---|---|---|
| Plating | Temp1 Temp2<br>Voltage 1 Voltage 2<br>Current 1<br>Current 2<br>Flux 1 | Etching | Temp 1 Temp 2<br>Speed pressure<br>Current 1<br>Current 2<br>Concentration | Measurement | Thickness<br>Width<br>Space |

through equipment $a6$ in process $x6$. Furthermore, the cumulative factor shows a 10.1% higher performance than the fault detection performance of a single factor ($x5 : a5$).

*4.3. Analysis Results for Layer 2.* This section analyzes the critical suspected processes ($x5$) and equipment ($a5$) identified in the previous section. The analysis model employed a multiple regression model and was used for plating. The criterion variables ($y$) are the width ($y1$), thickness ($y2$), and space ($y3$), which were divided into 200 (57 variables) and 137 lots (200 variables) based on six conditions. The input variable ($x$) is selected from the temperature, voltage, current, and flux, and the effect of the input variable on the process parameter was analyzed using stepwise regression analysis. To verify the analysis, the data were divided into training (70%) and validation sets (30%). Subsequently, the optimal model was selected based on validation data from the models generated using the training data. Herein, for each parameter by the selected factor, the variable satisfying a $p$ value of 0.05 is deemed significant. Finally, the criterion variable value was predicted by setting a regression equation using the selected parameters.

To verify the conditional regression equation for the criterion variables $y1$, $y2$, and $y3$, the regression models for $y1$ (137 Lot) and $y2$ (137 Lot) were selected as the optimal models (see Table 5). The regression model was diagnosed after setting the explanatory power to more than 0.7; to enhance the model fitness, the root mean squared error and the SSE were derived to be close to 70 : 30 (training: validation). The criterion variables were analyzed by prioritizing $y2$ between variables $y1$ and $y2$. For the equipment variables that affect the process parameter ($y2$), a significant variable with a $p$ value of less than 0.05 was selected using the stepwise variable selection method. The selected equipment variables were Rect124_vtg, Rect125_vtg, Rect150_vtg, c_temp_003, and a_col 143. Based on these variables, it was determined that the plating thickness of the PCB is affected by the temperature, voltage, and electric current of the plating equipment.

To verify the conditional regression equation for the criterion variables $y1$, $y2$, and $y3$, the regression models for $y1$ (137Lot) and $y2$ (137Lot) were selected as the optimal models (see Table 5). The regression model was diagnosed after setting the explanatory power to more than 0.7; to enhance the model fitness, the root mean squared error and the SSE were derived to be close to 70 : 30 (training: validation). The criterion variables were analyzed by prioritizing $y2$ between variables $y1$ and $y2$. For the equipment variables that affect the process parameter ($y2$), a significant variable with a $p$ value of less than 0.05 was selected using the stepwise variable selection method. The selected equipment parameters were Rect124_vtg, Rect125_vtg, Rect150_vtg, c_temp_003, and a_col 143. Based on these variables, it was determined that the plating thickness of the PCB is affected by the temperature, voltage, and electric current of the plating equipment.

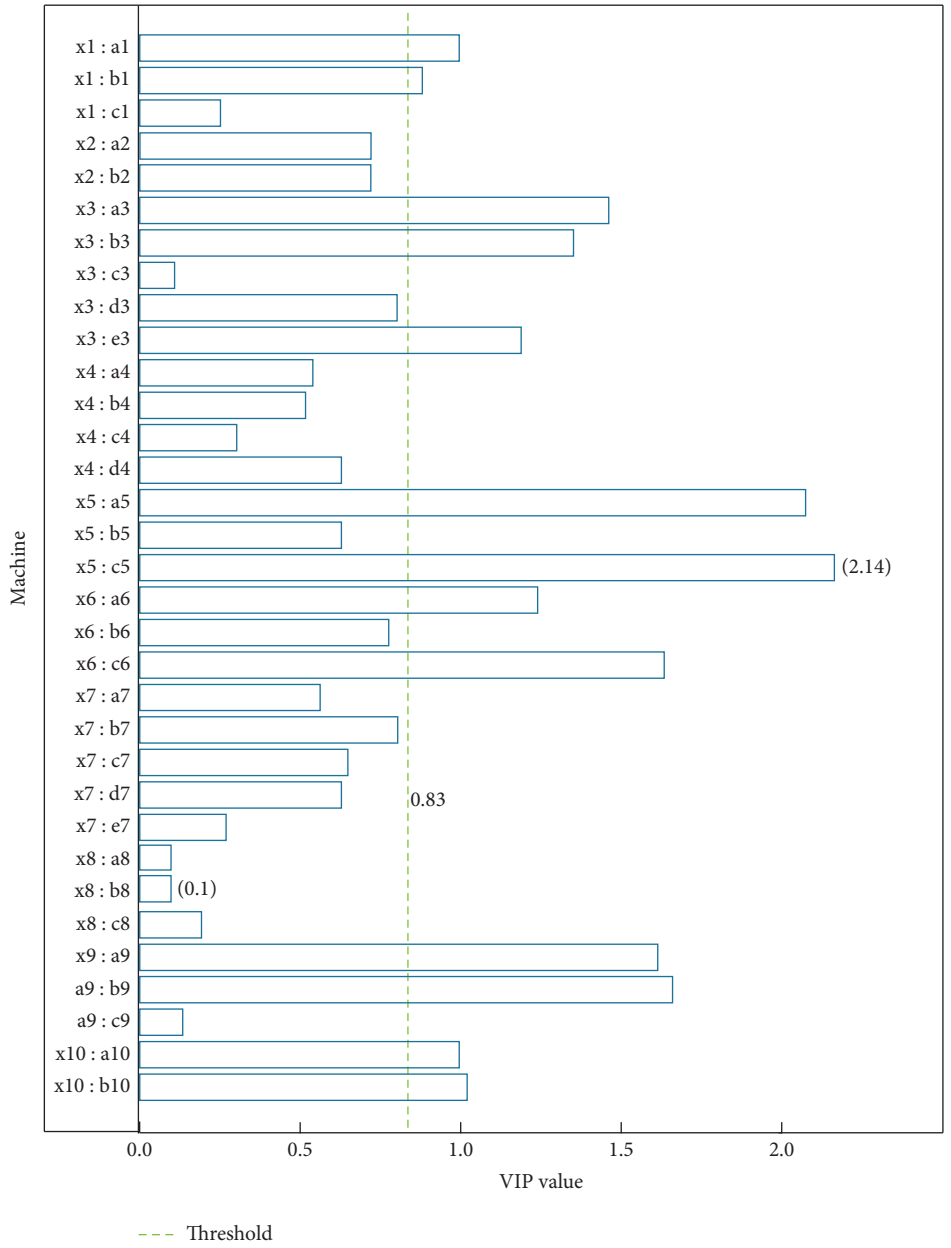


Figure 4: VIP values for the quality variable ($Y_1$).

The ANOVA test results on the criterion variable ($y2$) are as follows.

Table 6 shows that the $p$ value of the model is less than 0.0001. This indicates that the $p$ value of the regression equation is less than 0.05, thereby confirming the statistical significance. The value of $R^2$($R$-square) was found to be 0.7614, which indicates that the estimated regression line can describe more than 76.14% of the actual sample. Because the $p$ values of the five selected variables are smaller than 0.05, the variables of c_temp_003, Rect124_vtg_00, Rect125_vtg_00, Rect150_vtg_00, and a_col143 can be considered statistically significant (see Table 6).

$$y\,(\text{thickness}) = -86.9 + 3.4\,\text{c\_temp} + 3.9\,\text{Rect124\_vtg} + 8.7\,\text{Rect125\_vtg} - 3.5\,\text{Rect150\_vtg} - 10.1\,\text{a\_col143}. \quad (8)$$

*4.4. Verification.* The experimental results of this study were verified in the field using a theoretical approach. The plating process provides decorative esthetics, corrosion resistance, and electrical conductivity by forming a metal film on the surface of a metal or nonmetal. There are two main types of plating. Electrical plating employs a method

Table 3: Machine factors selected as VIP value.

| Quality variable | Selected machines (independent variable) |
|---|---|
| $Y_1$ | $x_1$: $a_1$, $x_1$: $b_1$, $x_3$: $a_3$, $x_3$: $b_3$, $x_3$: $e_3$, $x_5$: $a_5$, $x_5$: $c_5x_6$: $a_6$, $x_6$: $c_6$, $x_9$: $a_9$, $x_9$: $b_9$, $x_{10}$: $a_{10}$, $x_{10}$: $b_{10}$ |

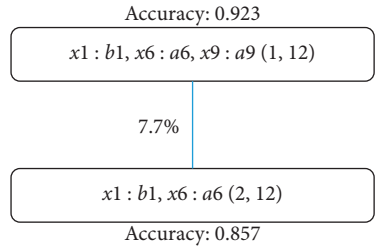


Figure 5: Tree of rule length 3 for quality variable $Y_1$.

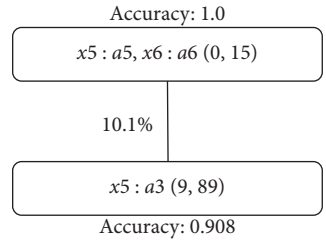


Figure 6: Tree of rule length 2 for quality variable $Y_1$.

of plating using electrolysis by flowing electricity to the anode and cathode, whereas chemical plating employs a method of plating using Cu ions as a catalyst and a precipitating metal (Cu) as a reducing agent. Faraday's law of electrolysis states that the amount of substances generated on electrodes by electrolysis in an aqueous solution is directly proportional to the amount of electricity charged (current × time). When a certain amount of electricity is provided, the amount of substance precipitated on the electrode in the aqueous solution is directly proportional to the chemical equivalent (atomic weight/valence). This results in the following equation:

$$\text{plating thickness } (\mu\text{m}) = \frac{\text{electric current (A)} \times \text{time (s)} \times \text{1g equivalent (g)} \times 10,000}{96,500\,\text{C} \times \text{surface area}\left(\text{dm}^2\right) \times 100 \times \text{density}\left(\text{g/cm}^3\right)},$$

$$^{*}\,\text{Faraday constant (F)} = 96,500\,\text{C}, \tag{9}$$

$$^{*}\,1\,\text{dm}^2 = 100\,\text{cm}^2.$$

Table 4: Cumulative factors for quality variable $Y_1$.

| Cumulative factor | Accuracy (%) | Cumulative effect (%) |
|---|---|---|
| ($x1$: $b1$, $x9$: $a9$) | 80.1 | 5.3 |
| ($x3$: $e3$, $x9$: $a9$) | 85.3 | 6.4 |
| ($x1$: $b1$, $x6$: $a6$) | 85.7 | 12.9 |
| ($x5$: $a5$, $x6$: $a6$) | 100.0 | 10.1 |
| ($x1$: $a1$, $x3$: $e3$, $x5$: $c5$) | 83.2 | 7.0 |
| ($x1$: $b1$, $x6$: $a6$, $x9$: $a9$) | 92.3 | 7.7 |

Table 5: Summary of the target model.

| Variable | | Target | | | | | |
|---|---|---|---|---|---|---|---|
| | | $y1$ (width) | | $y2$ (thickness) | | $y3$ (space) | |
| Data set | Lot size | 137 | 200 | 137 | 200 | 137 | 200 |
| | Variable | 200 | 57 | 200 | 57 | 200 | 57 |
| | Training validation | 70 : 30 | 70 : 30 | 70 : 30 | 70 : 30 | 70 : 30 | 70 : 30 |
| Variable | Pump flux | .Pump5, 6, 12_flux | .Pump1_flux | .c_temp_003 .a_col 143 | .Pump6_flux | .Pump12_flux | .Pump1_flux |
| | Rectifier voltage | **.Rect119**, 124, 149, 62, 86_vtg | **.Rect27_vtg** **.Rect36_vtg** | **.Rect124_vtg** .Rect125_vtg .Rect150_vtg | **.Rect13_vtg** .Rect45_vtg | **.Rect124_vtg** .Rect135_vtg | **.Rect1, 20**, 27, 36 .Rect3 |
| Coefficient of determination | R-square | 0.83 | 0.36 | 0.76 | 0.29 | 0.52 | 0.41 |
| Validation (training: validation) | RMSE | 1.22 : 1.28 | 0.90 : 0.91 | 0.87 : 1.04 | 0.84 : 0.76 | 0.93 : 1.0 | 0.90 : 1.08 |
| | SSE | 130.9 : 65.5 | 111.5 : 49.9 | 68.8 : 43.6 | 95 : 34.7 | 80.1 : 40.6 | 107.2 : 70.1 |

Table 6: Analysis of variance.

| Source | DF | Sum of Sq | Mean Sq | $F$ value | Pr (> \|F\|) |
|---|---|---|---|---|---|
| Model | 5 | 31.6454 | 6.3290 | 22.98 | <.0001 |
| Error | 36 | 9.9145 | 0.2754 | | |
| Corrected total | 41 | 41.5599 | | | |
| $R$-square | | 0.7614 | | Adj R-Sq | 0.7283 |
| AIC | | −48.6340 | | BIC | |
| SBC | | −38.2080 | | $C(p)$ | |
| Variables | DF | Estimate | Standard error | $T$ value | Pr > \|$t$\| |
| Intercept | 1 | −86.9945 | 45.8186 | −1.90 | 0.0656 |
| c_temp_003 | 1 | 3.4276 | 1.2003 | 2.86 | 0.0071 |
| Rect124_vtg_00 | 1 | 3.9345 | 0.9224 | 4.27 | 0.0001 |
| Rect125_vtg_00 | 1 | 8.7680 | 1.9205 | 4.57 | <0.0001 |
| Rect150_vtg_00 | 1 | −3.5887 | 0.9794 | −3.66 | 0.0008 |
| a_col143 | 1 | −10.1484 | 1.8758 | −5.41 | <0.0001 |

The plating thickness is directly proportional to the amount of electricity applied to the rectifier ($P = \mathrm{V} \times \mathrm{I}$). Conversely, the control of the plating thickness is affected by the temperature of the plating equipment and the applied voltage.

## 5. Conclusion

The purpose of this study is to find processes and machines that affect the yield of micromanufacturing processes and to secure corporate competitiveness by improving product yield and productivity through the analysis of equipment parameters of suspected machines. Consequently, by analyzing the fault data and equipment parameters of the manufacturing line, the process affecting the yield and the suspected machine that significantly affects the fault were determined by analyzing the machine that processed the product by the process. The experimental results revealed that the factors that cause faults are not only the single process variables but also the cumulative factor in which the downstream and upstream contribute to the faults. From the experimental results, the cumulative factor ($x5 : a5$, $x6 : a6$) suggested that 100.0% of the faults were found in all lots that went through equipment $a5$ in process $x5$ and equipment $a6$

in process $x6$. Furthermore, it was demonstrated that the cumulative effect had a 10.1% higher performance than the fault detection performance by a single factor ($x5 : a5$). Stepwise analysis of the process parameters (thickness) and equipment parameters of the $x5$ ($a5$) process—the critical suspected process found in Layer 1—helped identify four equipment parameters, in addition to c_temp, as significant parameters. The proposed methodology might significantly improve product yield and quality by identifying the cause of product faults in manufacturing enterprises. Meanwhile, processes, machines, and critical parameters classified as critical factors should be managed thoroughly by collecting opinions from field engineers. Furthermore, to perform big data analysis for such manufacturing sites, it is necessary to establish an environment wherein the history of the critical processes and the data of inspection/measurement and manufacturing equipment are gathered and fed back in real time through sensors and IoT. Conversely, an environment that can collect and control the manufacturing site data in real time, which are the core functions of intelligent manufacturing, needs to be established. Therefore, in this study, we propose a factory integration system and system architecture of a PCB line to realize a practical intelligent factory in connection with the analysis of big data at the manufacturing site.

Follow-up studies will be conducted on the methods of critical parameter management and control for processes and equipment and on manufacturing big data platforms.

## Data Availability

The data used to support this study are used by companies to provide "research data" for research and paper publishing and are also included within the article.

## Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This research work was supported by the National Research Foundation of Korea (no. 2021-0090).

## References

[1] J. Y. Lee, J. S. Yoon, and B.-H. Kim, "A big data analytics platform for smart factories in small and medium-sized manufacturing enterprises: an empirical case study of a die casting factory," *International Journal of Precision Engineering and Manufacturing*, vol. 18, no. 10, pp. 1353–1361, 2017.

[2] J.-m. Park, "Technology and issue on embodiment of smart factory in small-medium manufacturing business," *The Journal of Korean Institute of Communications and Information Sciences*, vol. 40, no. 12, pp. 2491–2502, 2015.

[3] H. Sim, D. Choi, and C. O. Kim, "A data mining approach to the causal analysis of product faults in multi-stage PCB manufacturing," *International Journal of Precision Engineering and Manufacturing*, vol. 15, no. 8, pp. 1563–1573, 2014.

[4] D. Gil, M. Johnsson, H. Mora, and J. Szymański, "Review of the complexity of managing big data of the internet of things," *Complexity*, vol. 2019, Article ID 4592902, 12 pages, 2019.

[5] X. Zhang and X. Ming, "An implementation for Smart Manufacturing Information System (SMIS) from an industrial practice survey," *Computers & Industrial Engineering*, vol. 151, Article ID 106938, 2021.

[6] D. C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley & Sons, Hoboken, NJ, USA, 8 edition, 2020.

[7] R. Dunia, S. J. Qin, T. F. Edgar, and T. J. McAvoy, "Identification of faulty sensors using principal component analysis," *AIChE Journal*, vol. 42, no. 10, pp. 2797–2812, 1996.

[8] M. D. Ming-Da Ma, D. S.-H. Wong, S. S. Sheng-Tsaing Tseng, and S. T. Tseng, "Fault detection based on statistical multivariate analysis and microarray visualization," *IEEE Transactions on Industrial Informatics*, vol. 6, no. 1, pp. 18–24, 2010.

[9] C.-F. Chien and C.-C. Chen, "Data-Driven framework for tool health monitoring and maintenance strategy for smart manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 4, pp. 644–652, 2020.

[10] Y. Zhao, Y. He, D. Zhou et al., "Functional risk-oriented integrated preventive maintenance considering product quality loss for multistate manufacturing systems," *International Journal of Production Research*, vol. 59, no. 4, 2021.

[11] Y. He, Z. Chen, Y. Zhao, X. Han, and D. Zhou, "Mission reliability evaluation for fuzzy multistate manufacturing system based on an extended stochastic flow network," *IEEE Transactions on Reliability*, vol. 69, no. 4, pp. 1239–1253, 2020.

[12] Z. Wang, W. Feng, J. Ye et al., "A study on intelligent manufacturing industrial internet for injection molding industry based on digital twin," *Complexity*, vol. 2021, Article ID 8838914, 16 pages, 2021.

[13] A. Majeed, Y. Zhang, S. Ren et al., "A big data-driven framework for sustainable and smart additive manufacturing," *Robotics and Computer Integrated Manufacturing*, vol. 67, Article ID 102026, 2021.

[14] H. S. Sim, "Big data analysis methodology for smart manufacturing systems," *International Journal of Precision Engineering and Manufacturing*, vol. 20, no. 6, pp. 973–982, 2019.

[15] Special Report, "Smart factory," *Dong-A Business Review*, vol. 227, pp. 67-68, 2017.

[16] MESA International, *MES Harmonization in a Multi-site, Multi-country, and Multi-cultural Environment: Case Study of a Plant to Enterprise Solution*, MESA International White Paper, 2007.

[17] MESA International, *MES Explained: A High Level Vision*, MESA International White Paper, 1997.

[18] Y. S. Jeong, "Linking algorithm between IoT devices for smart factory environment of SMEs," *The Journal of Cases on Information Technology*, vol. 8, no. 2, pp. 233–238, 2018.

[19] Y.-H. Choi, S. H. Choi, and S. H. Choi, "A study on the factors influencing the competitiveness of small and medium companies applied with smart factory system," *Information Systems Review*, vol. 19, no. 2, pp. 95–113, 2017.

[20] D. B. Ko and J. M. Park, "A study on the visualization of facility data using manufacturing data collection standard," *The Journal of the Institute Internet, Broadcasting and Communication*, vol. 18, no. 3, pp. 159–166, 2018.

[21] T. Mehmood, H. Martens, S. Sæbø, J. Warringer, and L. Snipen, "A Partial Least Squares based algorithm for parsimonious variable selection," *Algorithms for Molecular Biology: AMB*, vol. 6, no. 1, pp. 27–12, 2011.

[22] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, Burlington, MA, USA, Second edition, 2006.
[23] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proceedings of 20th International Conference on Very Large Data Bases (VLDB)*, pp. 487–499, Santiago, Chile, September 1994.
[24] H. Amos and C. Argon, "Piecewise regression model construction with sample efficient regression tree(SERT) and applications to semiconductor yield analysis," *Journal of Process Control*, vol. 22, no. 7, pp. 1307–1317, 2012.
[25] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, Wiley & Sons, Hoboken, NJ, USA, 4 edition, 2017.
[26] S. Weisberg, *Applied Linear Regression*, Wiley & Sons, Hoboken, NJ, USA, 3 edition, 2005.
[27] D. M. Hawkins, "Multivariate quality control based on regression-adjusted variables," *Technometrics*, vol. 33, no. 1, pp. 61–75, 1991.
[28] B. K. Kim and B. J. Yum, "Development of virtual metrology models in semiconductor manufacturing using genetic algorithm and partial least square regression," *IE Interface*, vol. 23, no. 3, pp. 229–238, 2010.
[29] A. M. Krieger, M. Pollak, and B. Yakir, "Surveillance of a simple linear regression," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 456–469, 2003.
[30] L. Debbie and V. Hans, *Applied Multivariate Statistical Concepts*, Routledge, Milton, UK, 2017.
[31] R. D. Cook, "Detection of influential observations in linear regression," *Technometrics*, vol. 42, no. 1, pp. 65–68, 2012.
[32] I. G. Chong and C. H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and Intelligent Laboratory Systems*, vol. 78, no. 1, pp. 103–112, 2005.