IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 8, 2020, accepted January 29, 2020, date of publication February 17, 2020, date of current version March 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974535

A Data-Driven Approach for Identifying Possible Manufacturing Processes and Production Parameters That Cause Product Defects: A Thin-Film Filter Company Case Study

JRJUNG LYU[1], CHIA WEN LIANG[1], AND PING-SHUN CHEN[2]
[1]Department of Industrial and Information Management, National Cheng Kung University, Tainan City 701, Taiwan
[2]Department of Industrial and Systems Engineering, Chung Yuan Christian University, Taoyuan City 32023, Taiwan

Corresponding author: Ping-Shun Chen (pingshun@cycu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2221-E-033-008.

ABSTRACT A semiconductor or photoelectric manufacturer faces a more competitive market with small quantities of many products. These products require hundreds of processes for production, thereby generating huge manufacturing data. With the help of the Internet of Things (IoT) technology, the manufacturer can collect manufacturing process data in a timely manner. Due to the massive quantities of manufacturing process data, it has become difficult for manufacturers to determine the causes of product defects, by which machine, and by what manufacturing process (or recipe) parameters. This research proposes a six-step data-driven solution to this problem. The chi-square test of independence, the Apriori algorithm, and the decision tree method identify the process that is generating the defective products and extract rules to identify the lot identification of product defects and their associated manufacturing process parameters. An empirical study was conducted at an optical thin-film filter (TFF) company in Taiwan. Based on the data of the optical TFF production lines, the coating process was identified as the source of the defective products, and the extracted rules were validated and implemented. The product defect rate decreased from 20% to 5%. Hence, the proposed data-driven solution was found to be capable of helping manufacturers enhance their product yield.

INDEX TERMS Data-driven, data mining, process improvement, association rule mining, decision tree, thin-film filter (TFF).

I. INTRODUCTION

Originally, global production of thin-film-transistor liquid-crystal displays (TFT-LCD) was mainly carried out in Taiwan, Korea, and Japan. Subsequently, global panel output experienced great changes due to demand change in industry environment and mass production in China [1]. For production of major key parts of TFT-LCD in Taiwan, as downstream application products are approaching saturation with decreasing demand and gradual slowdown in growth, and due to low technical threshold for low-order panels and low-price competition of Chinese manufacturers over recent years, survivability of upstream panel-part manufacturers in Taiwan was severely squeezed. TFT-LCD manufacturers tended to be conservative in production expansion due to the lack of market demand and difficult to earn good profits.

Crucial to the future development of TFT-LCD manufacturers is the integration of automation technology and the Internet of Things (IoT) with historical information, experience, and intelligence serving as the basis for integrating smart factories into an automated production system [2]–[8]. The objective is to allow equipment to communicate, collaborate, and make decisions. For example, equipment can quickly assess a machine's status through parameter values collected from sensors and using algorithmic rules. Warnings can be sent to process engineers that preventive maintenance should be performed before problems occur, reducing the percentage of defective products Manyika *et al.* [9]

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

reported that the manufacturing industry generates more data than any other industry. Research and development, demand, supply, and process levels are the dimensions of order placement and product design, manufacturing, and sales. In addition, data related to asynchronicity, velocity, volume, variety, and disorder are generated by machines' process sensors, image measurement systems, and information systems. Consequently, enterprises cannot perform instant Big Data analysis [10], [11], and fail to obtain timely and valuable knowledge. The cost of producing equipment in the manufacturing industry is high, and requires careful assessment; therefore, the topic of how to increase product yield by improving machine productivity warrants exploration.

In the manufacturing industry's strategic process for moving toward Industry 4.0, the IoT and Big Data are required to construct a more comprehensive technology framework or platform, thereby reducing labor cost and enhancing the flexibility and benefits of production and manufacturing processes [12]–[15]. With the gradual perfection of software and hardware, data collected from the manufacturing industry will become more diverse and accumulate rapidly, resulting in a disparity between the results of data analysis performed by process personnel and the benefits generated from the data [11].

During the manufacturing process, some engineers use their own experience to solve problems in product processing, but they seldom use data analysis to identify the source of the defects [16], [17]. This may result in variation because of differences in the engineers' technical experience. For example, optical communication products in the optoelectronics industry are extremely diverse; thus, factories manufacture small quantities of many products. Only one batch of filters in each band may be produced within a certain time period, and the process parameter settings and recipes may differ, causing large data variability than when a mass production mode is used and preventing process engineers from quickly identifying the factors affecting product yield.

Most manufacturing factories rely on process engineers' experience and simple data analysis to diagnose defects and troubleshoot problems by adopting the data derived from the detection-end and trial and error. Afterward, the process time and poor yield of a machine are estimated in an attempt to determine the possible factors affecting the yield. However, the trial-and-error method can be inaccurate. Engineers' lack of experience may cause misjudgment, preventing timely and effective improvement during product processing. In brief, this method incurs heavy labor costs and lacks efficiency and accuracy.

Due to a large amount of manufacturing data, how to identify possible causes of product defects incurred by which process, by which machine, and by what manufacturing process parameters becomes a critical problem for manufacturers wishing to improve their product yield. Researchers have applied data mining technology to identify the process and the machine that caused product defects and extract rules of types and characteristics of product defects [17]–[20]. However, few scholars have focused on extracting key manufacturing process (or recipe) parameters to avoid to defective products or produce better ones [21], [22]. This is the rationale for this study. Therefore, this study constructed a data-driven means of improving the production of small quantities of diverse products. Statistics, data mining, and domain knowledge were combined to obtain useful information hidden in large amounts of process data. Rules were established to assist engineering personnel in determining potential problems and accurately diagnosing the factors that cause the production of defective products, thereby increasing the production yield rate.

The research steps were as follows. Big Data was the basis of the data-driven method. Process yield issues were improved using the dimensions of data collection, preprocessing, and analysis.

Statistical methods, association rules, and decision trees were combined to construct a two-stage data mining architecture that categorizes and predicts issues concerning process machines, parameter settings, and yield rates. Rules were extracted to obtain valuable information.

Actual process information was applied to evaluate the two-stage data mining architecture. Several indicators were used to verify and assess the accuracy of predictions.

The rest of this article is as follows. Section II reviews the literature on the data mining technology, decision tree methods, and Big Data applications on process yield improvement. Section III proposes the six-step research methodology. Section IV uses a case study to verify the validation of the proposed methodology and presents the managerial implications. Finally, Section V concludes and suggests directions for future research.

II. LITERATURE REVIEW

This section introduces the papers related to the data mining and decision tree, and the literature on the process yield improvement by using Big Data and data mining methods.

A. DATA MINING AND DECISION TREE

Data mining problems are solved using either supervised or unsupervised data mining. The former is a top-down method that constructs specific target variables and builds associations among other variables. Two examples of supervised data mining are classification and prediction. The latter is a bottom-up method without specific target variables. Instead, there is a search for significant relationships among all variables. The importance of the variables to the target is determined only after patterns have been identified. Clustering and association rule determination are two examples of unsupervised data mining.

Data mining is a knowledge discovery process that uses artificial intelligence, machine learning, statistics, or algorithms to determine numerous data set models and the most useful models from data in large databases [10]. Fayyad *et al.* [23] described data mining as a procedure of processing, converting, exploring, and evaluating data drawn

from user needs so as to extract useful and comprehensible information that can assist in the making of business decisions. Assuncao *et al.* [24] indicated that data analysis has descriptive, predictive, and prescriptive purposes. The descriptive purpose is the use of simple methods to describe complicated phenomena and identify statuses hidden in large amounts of data, and to determine correlations, rules, patterns, or trends by analyzing the associations among data. The predictive purpose is the construction of a predictive model on the basis of the associations and rules in historical data so as to predict and evaluate future results. The prescriptive purpose is the adoption of decision-making processes to consider business objectives, demands, and limits, thereby helping analysis. According to Canito *et al.* [10], predictive analyses are receiving more research attention than before.

Berry and Linoff [25] proposed that data mining techniques consist of classification, prediction, sequential pattern, association analysis, and clustering analysis. In classification and prediction techniques, the most widely used method is the decision tree, constructed for the purposes of exploration and prediction. In exploration, the results of analyzing a decision tree's growth and formation are used to obtain information hidden in data; in prediction, the rules derived from a decision tree are adopted to predict future results, with optimal classification rules obtained from an interactive verification process incorporating training and testing data.

Lim *et al.* [26] compared the accuracy rate, complexity, and training time of commonly employed classification methods and discovered that the decision tree method has a good accuracy rate within a reasonable time and offers excellent data explanation. Decision tree methods are supervised learning methods which construct tree structures by classifying known information, with rules extracted to predict unknown samples [27], [28]. Algorithms using this method include the classification and regression tree (CART), chi-square automatic interaction detection (CHAID), and C4.5/C5.0. CART uses Gini heterogeneity as the branching indicator, with each branch generating only two child nodes. The misclassification rate of a minimized decision tree serves as a basis for managing continuous and discrete variables. CHAID uses the chi-square statistical significance test as a basis to define the branches of the decision tree, serving as an algorithm for identifying multiple branches that cannot manage continuous data. C4.5/C5.0 uses the information gain ratio as the principle of branch variables and creates decision trees with multiple branches that are employed mostly to manage categorical data.

To construct a data mining process, many scholars have used the cross-industry standard process for data mining (CRISP–DM) [29]–[36], as defined by the SPSS factory. This methodology was developed by the two factories SPSS and NCR in 1990, with the aim of using the experience of numerous successful cases to construct a cross-disciplinary standardization process [37]. The six-step process places data at the center and cycles constantly through business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

B. APPLYING BIG DATA AND DATA MINING TO ENHANCE PRODUCT YIELD

Many fields rely on Big Data analysis and data mining [15], [38]–[40]. Data are noncompetitive products in economics and can create value. The increase use of data does not diminish its value. Decision-makers can obtain different values based on their objectives, and apply the knowledge they obtain to target objects.

Numerous studies have solved problems with product yield enhancement in high-tech manufacturing industries such as the semiconductor and optoelectronics industries by using Big Data analysis and data mining. Based on the literature review, there are three categories of product yield enhancement studies. The first-category studies identify possible poor-quality processes and machines [18], [20], [41]–[43]. Rules can be extracted from manufacturing data. As a result, process engineers can use the extracted rules to find ways of improving product yield. Kittler and Wang [43] used process information to develop a data mining architecture, determined the correlation among machines by implementing the Kruskal–Wallis test, and used regression tree analysis to extract decision-making information that could serve as a reference for yield improvement, with the objective of increasing process yield. Chien *et al.* [20] integrated k-means clustering analysis with the Kruskal–Wallis test to construct a data mining architecture for diagnosing semiconductor manufacturing process problems and determining the reasons behind process variations. Tu *et al.* [42] employed tolerance control partitioning to integrate the Bayesian analysis and decision tree analysis for determining key machine indicators through comparisons of process machines, thereby monitoring process yield. Chien and Hsu [22] constructed a predictive model using model trees and artificial neural networks (ANNs) to create a design for the size of semiconductor die exposure, successfully determining and modifying abnormal areas.

The second-category studies identify and classify possible product defects [17], [44]–[49]. The purpose of these studies is to help process engineers identify the correct type of product defects in the poor-quality process and machine. The process engineers can then provide an appropriate solution for each type of product defects.

Braha and Shmilovici [46] increased chip yield by using data from an advanced chip cleaning process to construct a data mining architecture that combined a decision tree with an ANN and used a comprehensive classifier to execute categorization and prediction. Braha and Shmilovici [44] focused on data created from photolithography processes in semiconductor manufacturing and used the decision tree analysis to manage multidimensional and multiprocess parameters and explore the interactions among the parameters. Chen *et al.* [45] combined a regression tree with Canny edge detection to extract features and categorize defects.

Chen *et al.* [17] applied defect detection, feature extraction, and support vector machine approaches to classify product defect problems through the process error warnings of complementary metal-oxide-semiconductors. Chien *et al.* [48] detected root causes at the yield ramp-up stage of semiconductor manufacturing. Based on the simulation results, their approach had a higher average catching rate for identifying defect causes than random forest and linear regression methods. Nakata *et al.* [49] developed a three-stage method–failure map pattern monitoring, failure cause identification, and failure recurrence monitoring–to perform a yield analysis. They used a simple convolutional neural network (CNN). Their proposed method could discover new failure map patterns, identify the cause, and resolve them quickly. Haddad *et al.* [47] studied a defect detection and classification problem with the very small sample size of available data in semiconductor units. They developed a defect characterization framework, including feature extraction, feature subtraction, and sparse code generation, to identify defect patterns. The results of the numerical tests showed that the framework was more accurate in identifying feature types.

The third-category studies both identify poor-quality processes and machines, and extract the rules of manufacturing process parameters from manufacturing data [19], [21], [50]. The process engineers can then use their domain knowledge to extract, select, and implement the most useful rules to fulfill product yield enhancement. Casali and Ernst [50] integrated a decision tree with the chi-square test to evaluate the correlation of multiple and complex parameters for semiconductor process control. Chien *et al.* [19] adopted a multidimensional principal component analysis (PCA) to extract key indicators and subsequently used the k-nearest neighbor algorithm for grouping purposes and diagnosing and classifying process errors. Chu *et al.* [21] utilized the Kruskal–Wallis test and a decision tree to determine variation factors regarding the process failures of TFT-LCD.

Bect *et al.* [51] adopted several data mining methods, such as PCA, k-means, and canonical discriminant analysis, to detect abnormal events for helicopters. Flath and Stein [11] developed a defect prediction system for a manufacturing environment by using data mining technology. By removing non-essential features of the case study, they verified the efficiency of the proposed system to determine failure rates of manufacturing batches by station and time. Furthermore, they pointed out that the proposed defect prediction system might become inaccurate as time went by. Therefore, the proposed system requires periodic modification.

Based on this literature review, Big Data analysis and data mining techniques are determined to be capable of improving process yield. Most researchers, however, have focused either on identifying a possible process and a corresponding machine that resulted in defective products or on classifying types of defects [17]–[20]. For manufacturers that produce small quantities of diverse products, although production lots of each customer are small, products require multiple processes under different manufacturing process (or recipe) parameters of machines. How to improve product yield for manufacturers by using data mining technology when producing diverse products in small quantities has become a core competitiveness of enterprises. Therefore, this study has developed an integrated data-driven approach that can check whether or not a process was associated with product yield, identify possible lot identifications of product defects, and extract useful rules of key manufacturing process parameters from machines produced product defects.

Most studies apply such techniques to solve front-end process problems in the semiconductor industry [17]–[20], [22], [50], [52]. This study extended the research on yield improvement from the semiconductor industry to the optical communications industry by considering the optical thin-film coating process. Statistics, data mining, and data visualizations were combined to explore the process' information. Domain experts were consulted to discuss and verify the results.

III. RESEARCH METHODOLOGY

This section presents six procedures of the CRISP–DM, including defining the research problem, performing data preparation, performing data preprocessing, selecting the target variable and performing data reduction, constructing the corresponding models, and explaining analytical results.

A. RESEARCH PROCEDURES

This study consulted CRISP–DM, which is defined by the SPSS factory, and was divided into six procedures:

1) Discussions were held with domain experts to define the product yield problems, problems in determining product defects and in setting the intervals for process parameters.
2) Data for analysis was acquired from the database, and the range of data intervals was identified from the collected data types, data structures, and data were defined.
3) Data preprocessing was performed, which involved data integration, purification, and conversion, to improve the quality and reliability of follow-up analyses. The questions discussed with experts were formulated as hypotheses.
4) Target variable selection and data reduction were implemented to examine and predict whether the variables were significant. The project team identified possible factors as target variables. The selected variables were filtered through repeated assessments and fitness tests implemented by domain experts until the target variables were determined, thereby confirming the model's validity.
5) The model was constructed, and the information input into the data exploration structure. Through data measurement, the Apriori algorithm was used to search item sets with high frequency, extract relevant rules, and lower the data range and number of dimensions

(see reference [53] for the detailed Apriori algorithm). The machine setting parameters and measured data were then integrated to classify, predict, and analyze the decision tree. The correlation between process machine parameters and yields was analyzed, and the results visualized.

6) The results obtained using the rules were discussed with and modified by domain experts. Machine parameters were adjusted to verify the findings, in which systematic data searching and data mining were adopted to increase process engineers' experience in finding the problems in processes and to offer references for the engineers to consult.

The data-driven approach for identifying possible manufacturing processes and production parameters that cause product defects during the optical TFF production line is illustrated in Figure 1. The procedures are introduced in Sections III.B to III.F.

B. DEFINITION OF PROCESS PROBLEMS

In the product manufacturing process of the manufacturing industry, the factors affecting product yield may be associated with different processes. Specific machines and process recipes are involved, from ingredient import to product delivery, during which various problems of human and nonhuman (mechanical or process) origin may arise. The factors affecting product quality or causing defective products fall into seven categories, known as 7M: material, manpower, method, machine, measurement, maintenance, and management [54].

The direction of diagnosis was confirmed through problem definition and target setting. Domain knowledge was combined and relevant process data were collected or extracted based on the objective of the problems, so that appropriate methods or models could be selected for follow-up analyses.

C. DATA PREPARATION AND DATA PREPROCESSING

To identify the factors causing product defects or abnormal yields, relevant information on process sections, testing, or process setting parameters must be collected. Through data mining and Big Data analysis, useful information hidden in data was gleaned from large amounts of data.

However, human negligence, machine and equipment malfunction, and an inappropriate sampling method may result in a lack of data input, unfilled data fields, and inconsistency or contradiction in data types. Data preprocessing offers modifications before a data mining architecture is constructed, with preprocessing techniques involving data integration, cleaning, and transformation. The process involves converting different data types into forms that can be submitted for analysis.

D. TARGET VARIABLE SELECTION AND DATA REDUCTION

Based on the root case analysis of product defects, the possible factors would be candidates for improving product yield. The data of those factors were collected and identified as possible target variables. Then, selection of target variables filters out the targets requiring analysis. The predictor variables had to be confirmed after the target variables and attributes were obtained. To prevent the incorporation of variables that were irrelevant to the target or that may decrease the model's predictive power, three methods were used to select target and predictor variables.

In Method I, statistical methods were used to test whether or not a correlation existed between the predictor and the target variables; for example, the chi-square test of independence, which derives from the chi-square statistic, with a contingency table employed to calculate the degree of dependence between two variables [41], [50]. The degree of independence between two variables increased with increases in the calculated chi-square statistic of a sample. The chi-square test of independence tests the actual observed values of two categorical variables of a single sample, identifying whether a special relationship exists between them. The two variables are independent (dependent) when the chi-square statistic is nonsignificant (significant). In addition, both the chi-square test of independence [41], [50] and the PCA [19] have been used to examine the dependence between two variables in the product defect literature. Therefore, this research uses the chi-square test of independence to examine the relationship between the target variables. In Method II, discussions were held with domain experts to delete the variables that had no influence on the target variables in the data field, thereby validating the results of data mining. In Method III, the number of data dimensions was reduced, and the target variables were adopted as the standard for comparison, either using the feature selection method to delete the variables irrelevant to variable dimensions and target variables or adopting a PCA to convert the variables linearly.

To prevent deletion of crucial variables or the creation of too many irrelevant variables, a data partitioning test was conducted when the data volume was sufficiently large. Data were partitioned into training, testing, and validation, and the variables were verified using the model constructed by the training data. The misclassification measure and mean-square error (MSE) were adopted for assessment.

The most common definition of data partitioning ratio is partitioning 80% of the data into the data construction mode and using the remaining 20% to validate the test mode. Additionally, k-fold cross-validation was adopted to divide data into k portions. The number $(k-1)$ was chosen for each instance of mode training, and k instances of this training were performed to ensure that every datum was trained. The average of the results was used to indicate the validity of data.

E. MODEL CONSTRUCTION

A model was constructed in this study in two stages. The first stage involved using an association rule algorithm (Apriori) to select and organize the rules with high frequency; the evaluation indicators—support, confidence, and lift—were employed to extract the rules, which were then used to reduce the data search area. The data type input from the process machines was often a mixture of categorical and continuous

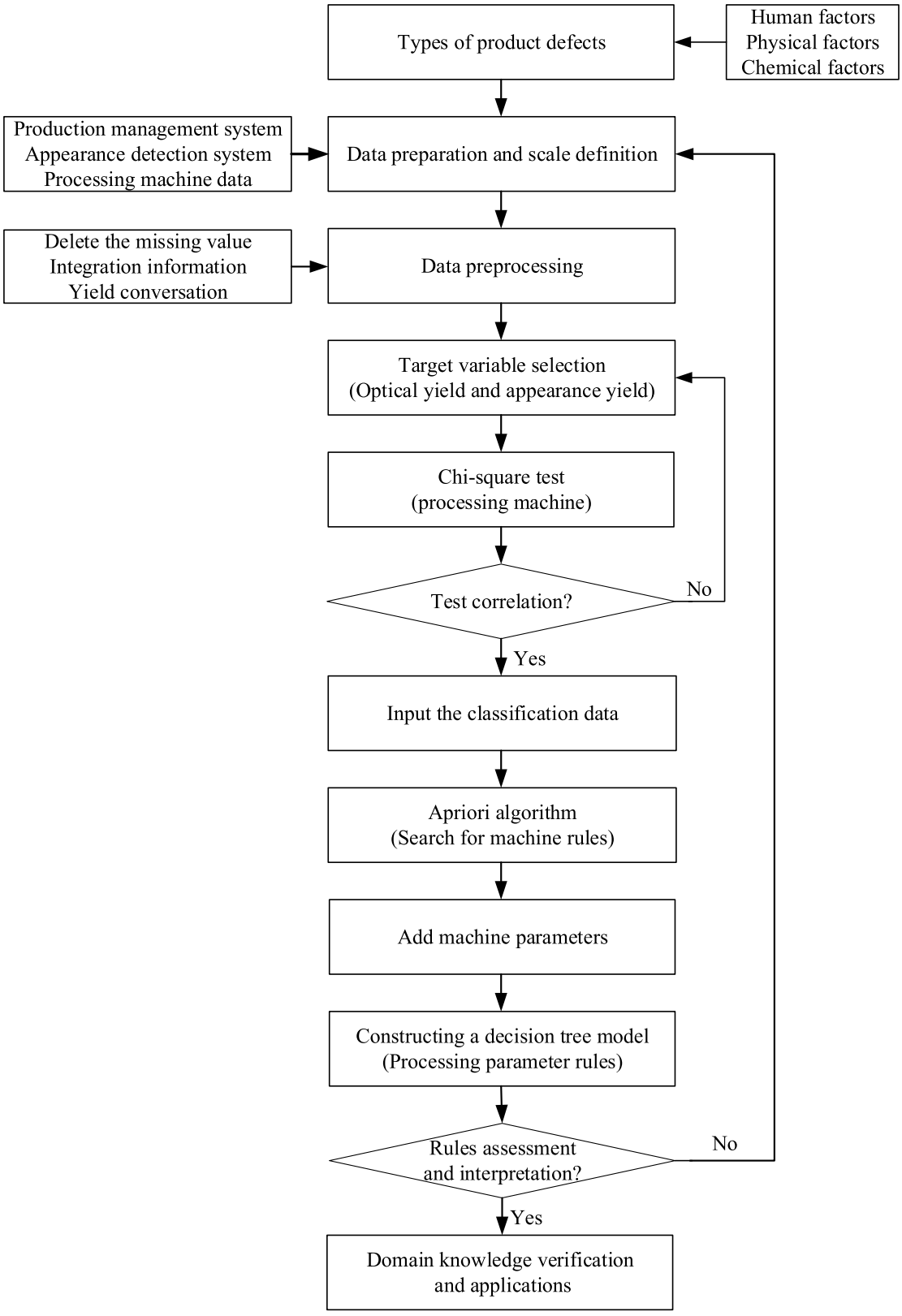

FIGURE 1. Data-driven approach for identifying possible manufacturing processes and production parameters that cause product defects.

types. Data discretization can result in loss of the meaning associated with the data and analytic errors; thus, the association rules were not suitable for data searching.

With the association rules serving as the basis in the second stage, the variables associated with the machine setting parameters and process recipes used during product processing were integrated into the data sheet. A decision tree was used to classify the process control variables and construct the prediction model, in which branch criteria determined the scale of the decision tree structure. The optimal branch results and implied information were obtained from the growth and trimming of the decision tree. The valuable information hidden in data was discovered through the complementation and application of the two data mining techniques.

1) ASSOCIATION RULES

This study used Apriori, which is the Boolean value algorithm most used for mining frequent item sets. The algorithm employs numerous data sets, and uses item sets to construct rules and calculate the frequency with which each candidate item appears. Whether the rules of candidate items are significant is determined based on the set thresholds. Association rules are constructed by executing an item set search in a horizontal direction, exploring $(k + 1)$-item sets using a combination of k-item sets, and generating a candidate item set through set joining. Support for the candidate item set must be larger than or equivalent to the predefined minimal support, and the candidate item set that satisfies this criterion is called the frequent item set. The steps of the algorithm are as follows:

Step 1. Determine the frequent 1-item set (L_1, which $k = 1$) that satisfies the minimum support.

Step 2. Obtain the candidate item set, C_{k+1}.

Join: Combine two item sets in L_k to generate the candidate item set, C_{k+1}.

Prune: Prune the subsets of C_{k+1} that do not belong to L_k to obtain C_{k+1}.

Step 3. Determine L_{k+1}.

Count: Calculate the support for pruning C_{k+1}.

Delete: Delete the support of C_{k+1} that did not achieve the minimal support, generating L_{k+1}.

$k = k + 1$.

Step 4. Repeat Steps 2 and 3 until no new frequent item set is generated; the association rules are subsequently generated from the frequent item sets.

The three evaluation indicators used were support, confidence, and lift, respectively indicating significance, correctness, and value. Discussions were held with domain experts on the settings of the support and confidence thresholds. Agrawal *et al.* [55] indicated that minimal support threshold can be calculated using Equation (1), in which $|T|$ refers to the number of the total events in the data sets and $|T(E)|$ indicates the number of event E contained in data.

$$(|T(E)|/|T|) \geq \text{Min Support}. \tag{1}$$

The value and gain of each rule were assessed. When the support and confidence of the rule were larger than the set threshold, the filtered items were advantageous for extracting the final rules. Whether lift was greater than 1 was examined. After the three indicators were calculated, the filtered rules could be arranged for the association rules, where the features and combinations of all the frequent item sets were displayed and explained. The explanation revealed that the probability of event occurrence was higher than that of the original event occurrence.

1) Support: Support indicated that the association rules, relative to all the information, must have a certain level of significance to be considered effective information. Through the setting of a minimal support threshold, the minimum data ratio that must be incorporated into the rules was controlled, and the relevant information with a relatively lower ratio was deleted. This study configured the minimal support (q) to filter the rules and retained the rule where the candidate item set support must be equal to or greater than threshold (q) to ensure that the rules had a sense of generality.
2) Confidence: Confidence indicated that when event X occurred, the degree of confidence regarding the correctness of the rule, which was inferred from the result item Y, served as the concept of conditional probability. This indicator evaluated whether the association rule contained confidence, and it must be equal to or greater than 0.8. This study configured a minimal confidence (p) to filter the rules and retain the rules which had confidence equal to or greater than threshold (p) to confirm the accuracy of the rule.
3) Lift: Lift was employed to compare an association rule's reliability and the probability of original event Y's occurrence to measure the value of a rule. Lift equal to or greater than 1 indicated that the results obtained with an association rule were superior to the original results. This study set the minimal lift as 1 to filter the rules and ensure their reliability.

2) DECISION TREE

The association rules were suitable for processing information types; however, the process setting parameters were indicated through the continuous type. Data conversion can cause the meaning of data to be lost; thus, this stage employed a decision tree to implement the second-stage data analysis, which compensated for the association rules' inadequacy when processing continuous data.

The branch-level structures of the decision tree could be employed to analyze the effects of the variations in various levels on the target variables. Based on the scale of the target variables, the decision tree classified the target variables into a classification tree of class types and classified the target variables into a regression tree of continuous types. The common decision tree techniques differ according to their different branch principles, branch approaches, and pruning

approaches. MSE, number of leaf nodes, depth of the decision tree, and number of decision tree rules were adopted to assess the model.

The decision tree algorithm adopted in this study was J48, with C4.5 serving as the core of the calculation. This algorithm selected the variables capable of reducing data disorder after branching. Ratio of misclassification was selected to represent disorder; the messages of the candidate attributes were not considered. Entropy increased with the number of the level of branch variable attributes, suggesting that the attributed branch features were nonsignificant. Selecting the attributes with lower entropy as the branch variables prevented entropy from nearing 0. On the basis of the candidate attributes that were higher than average for branching, the gain value of the average information of all the candidate attributes was calculated to determine the minimal entropy.

A true positive (*TP*) indicated that the prediction was good (*G*) and the actual result was *G*; a false positive (*FP*) that the prediction was *G* but the actual result was not good (*NG*); a true negative (*TN*) that the prediction was *NG* and the actual result was *NG*; and a false negative (*FN*) that the prediction was *NG* but the actual result was *G*. Using the confusion matrix comprising *TP*, *FP*, *TN*, and *FN*, the accuracy and classification error rates of the decision tree were calculated by Equations (2) and (3), respectively.

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN). \quad (2)$$

$$\begin{aligned}\text{Classification error rate} &= 1 - \text{Accuracy} \\ &= (FP + FN)/(TP + TN + FP + FN). \quad (3)\end{aligned}$$

This study assessed the observable indicators of the classification model for two types of classification and prediction abilities. The indicators were recall, sensitivity, precision, specificity, *F*-measure, Matthews correlation coefficient (MCC), receiver operating characteristic (ROC) curve, and precision-recall curve (PRC) area. Sensitivity is the percentage of the category that was correctly predicted (see Equation 4). Specificity is the percentage of the negatives that were correctly predicted as such (see Equation 5). Recall is the percentage of data in a certain category that was correctly judged as being in that category (see Equation 6). Precision is the percentage of the data that precisely pertained to that category among all the prediction types (see Equation 7). High precision indicated that the probability of data misjudgment was low.

Because recall and precision are inversely associated, these two indicators could be combined into a comprehensive indicator: the *F*-measure (see Equation 8). A high *F*-measure indicated that high precision and recall of the classification model. The MCC is the degree of imbalance in the binary classification (see Equation 9). Using the ROC curve, the area of the ROC curve can be calculated, and the judgment for classification results differs by threshold. A large (small) area under the ROC curve indicated that the classification effects were favorable (unfavorable). In general, the area under the ROC curve was greater than or equal to 0.5, where 0.5–0.7 indicates low accuracy, 0.7–0.9 indicates moderate accuracy, and greater than or equal to 0.9 indicates high accuracy. The PRC area refers to the area under the recall ratio or precision ratio curve.

$$\text{Sensitivity} = TP/(TP + FN). \quad (4)$$

$$\text{Specificity} = TN/(TN + FP). \quad (5)$$

$$\text{Recall} = TP/(TP + FN). \quad (6)$$

$$\text{Precision} = TP/(TP + FP). \quad (7)$$

$$F - \text{measure} = (2 \times \text{Recall} \times \text{Precision})/(\text{Recall} + \text{Precision}). \quad (8)$$

$$\text{MCC} = (TP \times \text{TN-FP} \times FN)/((TP+FP) \times (TP + FN) \times (TN+FP) \times (TN+FN))^{0.5}. \quad (9)$$

Pre- and post-pruning were the methods employed for decision tree pruning. C4.5/C5.0 used error-based pruning to compare the purity of a parent node and child nodes. When the attributes of the original training data were excessive or when the preferences for the branches of the decision tree algorithm differed, overfitting might easily occur, resulting in an overly complicated tree structure and preventing explanation and extraction of key rules. Error-based pruning decreases the classification error rate through pre- or post-pruning.

Based on the distribution of the target variables in the input information, examination and data cutting were implemented sequentially using root nodes. Subsequently, the branching rules and methods for the decision tree were classified, with each branch indicating the examination results and each leaf node referring to the distribution of target variables. Finally, the tree branches were displayed. The decision tree rules could be determined by identifying the path from the root node to the leaf node to determine the decision tree structures combined from the product yield, which could be calculated using the independent variables of different times, process machines, and process recipes.

F. RESULT EXPLANATION AND ASSESSMENTS

During the development of the data mining architecture, frequent discussions were held with domain experts to modify the data processing methods. Subsequently, statistical analysis and data mining were adopted to improve and modify the model. Through statistical tests, the significant correlation among the independent variables, including time, machine, process machine, and product specification variables, was selected. Key rules were extracted from the association rules and decision tree to acquire in-depth knowledge.

The association rules were filtered using the three assessment indicators—support, confidence, and lift—in which the indicator setting was configured using the experts' knowledge and after consulting the literature. Usable rules were extracted from the filtering of association rule processes.

The classification modes for the decision tree generated different algorithm classification results when the algorithms

differed. The classification effects were evaluated with an objective assessment of the favorable decision tree structure using training data results; and with consideration that the extraction of the classified rules differed with the rule explanations. Therefore, domain experts' professional knowledge was required after objective assessment to define problems and conduct a comprehensive evaluation, before selecting the most suitable decision-making rules. The input of the results could be used to assess the *TP* rate, *FP* rate, precision, specificity, recall, sensitivity, *F*-measure, MCC, ROC curve, and PRC area.

The association rules can extract the information of the defect products, such as the product identification number, production machine, production date, or lot identification. The decision tree rules can extract the information of the defect products from different process parameters. The crucial rules extracted using the two types of data mining technique were integrated to serve as the basis of the model. The experts' professional knowledge was verified to confirm which process machines were generating the defective products, and historical data were employed to trace the data of the poorly performing process machines. This study examined the control parameters of machines producing defective products, and the possible reasons were identified and organized. These problems were described to engineering personnel to improve the manufacturing process and enable preventive regulations. The machine parameters and events that could result in problems were adjusted, decreasing the frequency of defective products being generated.

IV. RESULTS AND DISCUSSIONS

This section introduces the case study background, verifies the validation of the proposed methodology in Section III by using the case study method, and presents the managerial implications.

A. CASE STUDY

The research motivation of this study was to increase the yield of actual products manufactured by a factory. The factory produced thin-film filters (TFFs) that were used in fiber-optic communication devices and precision optics. Big Data analysis and data mining were conducted based on test data and process data recorded by the product management system. The case company was engaged in the design, production, and sales of active and passive components used for optical communication, such as the processing of optical interference filters and filters for optical transceiver modules used in dense wavelength division multiplexing.

The case company's core technology involved TFFs. The company recently invested in the precision optical coating technology to manufacture products, such as optical fibers, lens coatings, LED reflective film coatings, and transparent conductive films. Coatings and films can be evenly produced on glass substrate boards using vapor deposition. With this method, hundreds of layers are coated onto a substrate, and the wavelength can attain 250–1700 nm. The company

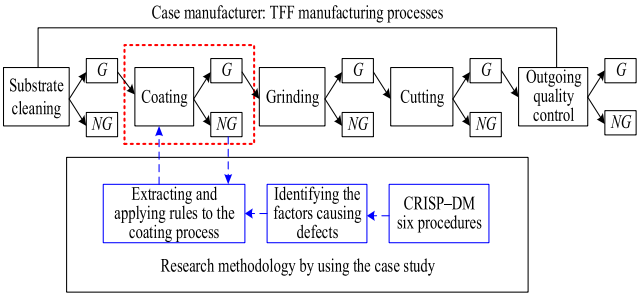

FIGURE 2. The framework of the case study.

manufactured the low- and high-end products and produced customized products, offering multiple product lines to satisfy the needs of customers.

The case company must reply on machine equipment to offer optical film coating services, and it mainly produced boards with substrate areas of 4, 6, and 9 inches. In the TFF production process, a quality check was performed after each stage was completed. When defects were detected, inspectors made judgments based on the size of the defect. If a defect was small enough for the product's usable area to meet the lowest standard, the product would advance to the next step. Product yield was statistically calculated after the cutting was completed. The statistical method involved a comprehensive examination of the estimated usable area; subsequently, the product lot was rated as *G* or *NG*.

Because product defects have physical, chemical, and human causes, the product defects resulting from process steps (e.g., cleaning, coating, grinding, and cutting) were classified. The TFF manufacturer produced diverse types of products in small quantities. An optical microscope was used to test the quality of the manufacturing processes of products with substrate area of 4, 6, and 9 inches to collect the data. Data collection lasted six months, and then the data were classified.

After discussions with domain experts on the problems encountered in the data, it was determined that the parameters set by the initial machines led to plating during the conversion between layers because of the machines' uncontrollable factors, thereby generating unknown fluctuations. Both new and old machines were used for grinding and cutting. The new machines generated their own data files, but data had to be collected manually from the old ones. The analysis quality could be increased only through data preprocessing. This study employed data from the case company and converted it during preprocessing so that it could be used in subsequent analyses, as shown in Figure 2.

B. DATA SELECTION AND COLLECTION

Six months of data on the optical coating process were collected from automatic and semiautomatic process equipment such as coating machines, the appearance inspection system, and the production management system. The data comprised machine parameter records, appearance models, optical measurement records, defective product styles, manufacturing

TABLE 1. Data preprocessing for the coating machines' data.

Variables	Scale
Layer	Ordinal scale
Progress time [sec]	Ordinal scale
Pressure [%]	Ratio scale
Flow rate [sccm]	Interval scale
EB1 emission [mA]	Interval scale
EB2 emission [mA]	Interval scale
Rate [A/sec]	Ordinal scale
Thickness [KA]	Ordinal scale
Light value [%]	Interval scale
Beam current [mA]	Interval scale
ACC current [mA]	Interval scale
RF source forward [W]	Interval scale
RF source reflect [W]	Interval scale
Neut.forward [W]	Interval scale
Neut.reflect [W]	Interval scale
Emission current [mA]	Interval scale

order numbers, coating machine models, product numbers, substrate types, sizes, usable areas and numbers, process recipes, and machine parameters.

C. DATA PREPROCESSING

Three steps for data preprocessing are as follows:

1) Data integration: The collected data originated from different information systems, and the data formats thus differed, as shown in Table 1. This study integrated data on the basis of domain experts' definitions of the production process. An item with inconsistent product numbers and manufacturing order numbers was selected. Missing values were found in the columns of appearance yield and optical yield in different data sheets. The corresponding data were deleted, or valuable data were obtained through reasonable judgment.
2) Data cleaning: The data that were not input manually and that had columns with missing yield records were deleted. The variables of the columns were deleted according to domain engineers' experience. Because this study analyzed the factors affecting the yield of the coating process, only coating machine information was kept for follow-up statistical tests.
3) Data conversion: The historical data collected from optical TFF production were discussed with domain experts and engineers; afterward, the test data and machine process parameters were converted. The collected test section information did not include the database columns of whether the shipped products were free-of-defects products; thus, products were categorized based on the thresholds of optical yield and appearance yield. Based on the discussions with the experts, the qualification threshold of spectral yield was selected to be higher than 70%. Therefore, this study set the threshold value (spectral yield at least 70%) for passed as *G* through the conversion of yield calculation. Product yield lower than the threshold was viewed as *NG*. Regarding the conversion of process data, the format of data output from the process machine was a log file, and the recorded data did not mark the switch points for layer conversion. The code conversion method was adopted to convert data into the Excel format, with the column 'Total' serving as the indicator, as shown in Table 2. Based on Table 2 of data variables compiled after data pretreatment, the online measurement data that were originally fetched comprised data on 528 products. A total of 461 test data remained after the columns without yield record or with a datum of 0 were deleted after data preprocessing. After data were cleaned and filtered based on the appearance yield and optical yield thresholds, data were categorized into batches. The batch results revealed that the number of *G* is 265 and the number of *NG* is 196. Table 3 shows a summary of the 461 products' (by lot) quality data after being processed by coating machines. These products were produced by the coating machines, P01, P05, P09, P10, and P19. For example, P01 coating machine has manufactured 105 good product lots, 51 defective product lots, and a total of 156 product lots.

D. EMPIRICAL ANALYSIS

During the analysis, the parameters were set based on literature and the domain experts' suggestions. The association rule threshold was set based on the test results. The support threshold was the first threshold layer, which was excessively high and could cause the deletion of effective rules. This study specified that support was required to be at least 5%, confidence at least 80%, and lift at least 1, where the data in the training and testing sets were 80% and 20%, respectively, of the total data set. The level of significance of the chi-square test was set as 0.05, and the Cramer's *V* coefficient was required to be at least 0.4, indicating a high correlation between the dependent and independent variables [56].

1) CHI-SQUARE TEST OF INDEPENDENCE

Contingency table analysis was employed to analyze the cross-classification data. Using the chi-square statistic, the contingency table analysis calculated the degree of dependence between two variables. This degree was high when the calculated sample of the chi-square statistic was large. The chi-square test of independence tested the actual observation values of two categorical variables in the same sample to determine whether special relevance existed. Two variables were independent (dependent) when the chi-square statistic was nonsignificant (significant).

When the given α value attained the level of significance, a calculated p-value smaller than α indicated that the results were significant; the null hypothesis (H_0) was rejected, and

TABLE 2. Data preprocessing for the coating machines' data.

Layer	Progress time [sec]	Pressure [%]	Flow rate [sccm]	EB1 emission [mA]	EB2 emission [mA]	Rate [A/sec]	Thickness [KA]	Light value [%]	Beam current [mA]	ACC current [mA]	RF source forward [W]	RF source reflect [W]	Neut. forward [W]	Neut. reflect [W]	Emission current [mA]	Total
1	1	66.8	0	325	0	2.9	-0.001	90.00097	135	1	130	0	35	1	266	1
1	2	66.9	0	325	0	6.4	0.003	90.00072	136	1	129	0	35	1	270	2
1	3	66.8	0	334	0	9.2	0.013	90.00063	135	1	129	0	35	0	272	3
1	4	66.7	0	335	0	10.5	0.021	90.00071	135	1	130	0	35	0	272	4
1	5	66.6	0	334	0	11.4	0.034	90.00130	134	1	129	0	35	0	262	5
1	6	66.6	0	331	0	11.7	0.044	90.00254	134	1	130	0	35	1	267	6
1	7	66.5	0	328	0	11.7	0.054	90.00425	134	1	130	0	35	0	270	7
1	8	66.5	0	324	0	11.5	0.068	90.00784	134	1	129	0	35	0	270	8
1	9	66.4	0	321	0	11.8	0.079	90.01082	135	1	129	0	35	0	270	9
1	10	66.4	0	316	0	11.5	0.093	90.01433	134	1	129	0	35	0	268	10
1	11	66.3	0	313	0	11.8	0.104	90.01627	133	1	129	0	35	0	268	11
1	12	66.3	0	310	0	11.1	0.117	90.01783	133	1	129	0	35	0	270	12
1	13	66.3	0	307	0	10.7	0.126	90.01877	135	1	130	0	35	0	270	13
1	14	66.3	0	306	0	10.9	0.136	90.01846	135	1	130	0	35	0	270	14
1	15	66.3	0	304	0	10.0	0.149	90.01763	134	1	130	0	35	0	270	15
1	16	66.3	0	302	0	10.6	0.157	90.01877	134	1	130	0	35	0	270	16
1	17	66.2	0	301	0	9.9	0.170	90.02341	134	1	130	0	35	0	270	17

TABLE 3. Observed products' quality data after being processed by coating machines.

No. of coating machine	Number of *G* product lots	Number of *NG* product lots	Total
P01	105	51	156
P05	12	10	22
P09	77	63	140
P10	30	23	53
P19	41	49	90
Total	265	196	461

the column and line classifications were independent. Subsequently, the independent variables and the results causing *G*/*NG* were judged as significantly related. Otherwise, the results were determined to be irrelevant. Variables determined as nonsignificant were directly filtered, and follow-up analyses were not conducted for them. After the data were preprocessed, the coating machine variable was tested using the chi-square test of independence. The number of *G*/*NG* classified products generated by each coating machine was employed to construct the chi-square contingency table. The test was executed with the level of significance (α) set as 0.05. Based on the data of Tables 3 and 4, the chi-square statistic was discovered to be 11.3742, and the p value 0.02267, which is lower than 0.05; thus, H_0 was rejected, and the coating machine was identified as significantly affecting the yield. The extracted association rules are described in the next section. This case study validated the chi-square test of independence. Therefore, manufacturers can apply the chi-square test of independence to determine whether or not there is a relationship between processes (or machines) and product defects.

2) EXTRACTION OF ASSOCIATION RULES

The chi-square test of independence revealed that coating machine is correlated with the product lot with *G* or *NG*. Association rule analyses were performed by using product number to classify the independent variables including major items, manufacturing time (divided according to months), and coating machine serial number, as well as the dependent variable columns *G*-lot and *NG*-lot. Discussions were held

TABLE 4. Expected products' quality data after being processed by coating machines.

No. of coating machine	Number of *G* product lots	Number of *NG* product lots	Total
P01	89.6746	66.3254	156
P05	12.6464	9.5358	22
P09	80.4772	59.5228	140
P10	30.4664	22.5336	53
P19	51.7354	38.2646	90
Total	265	196	461

TABLE 5. The association rules generated by the apriori algorithm.

No.	Rule	Support	Confidence	Lift
1	TIME = 201608 => *Y* = *NG* => LOT ID = DWF-MBD7D	5%	91%	1.75
2	LOT ID = DWF-1AQ7D => *Y* = *NG* => TIME = 201610	5%	87%	4.46

with domain experts to filter the association rules on the basis of the following: the first layer threshold regarding support (q) equal to 0.05, the second layer threshold regarding confidence (p) equal to 0.8, and the lift threshold at least 1.

Regarding the results of the Apriori algorithm, rules were selected based on these thresholds, and the following items were retained: support greater than or equal to 5%, confidence greater than or equal to 80%, and lift greater than or equal to 1. Subsequently, the rules were filtered according to domain experts' assessments, with rules with meaningless explanations deleted. Finally, two results with effective rules of $Y = NG$ were retained for follow-up assessment, observation, and analysis (see Table 5). In Table 5, Rule 1 indicated that when the major product number was under the condition LOT ID = DWF-MBD7D, the occurrence frequency of $Y = NG$ was especially high in August 2016. Rule 2 indicated that when the major product number was under the condition LOT ID = DWF-1AQ7D, the occurrence frequency of $Y = NG$ was especially high in October 2016. Therefore, this study validated the Apriori algorithm to be able to identify possible defective product LOT IDs, such as LOT ID = DWF-MBD7D in August 2016 and LOT ID = DWF-1AQ7D in October 2016.

Based on these results, the data of the rules leading to *NG* products were explored and backtracked. In Rule 1, the major product number MBD7D was classified by a coating machine, and the optical yield did not attain the standardized manufacturing serial number in August 2016. Among the different coating machine classifications for August, the product yield data of product number MBD7D-201608-P01 showed that 10 of the 21 product lots manufactured by P01 coating machine failed to meet the standard; that of MBD7D-201608-P09 demonstrated that 10 of the 16 product lots manufactured by P09 coating machine did not meet the standard; that of MBD7D-201608-P10 revealed that 8 of the 15 product lots manufactured by P10 coating machine did not meet the standard; and that of MBD7D-201608-P19 showed that two of the five product lots manufactured by P19 coating machine did not meet the standard. In total, 10, 10, 8, and 2 products did not meet the coating standards when processed by the P01, P09, P10, and P19 coating machines, respectively. The percentage of classified products processed by P01, P09, and P10 that did not meet the yield standards in the same month was nearly 50% of that month's total production. Therefore, the rules determined in this study exhibited a certain level of credibility. Thirty *NG* manufacturing serial numbers were collected to organize coating machine data and compare the process data.

The same analysis was applied to Rule 2. Among the coating machine classifications for October, the product yield chart of product number 1AQ7D-201610-P01 showed that seven of the 18 product lots manufactured by P01 coating machine did not meet the standard; that of 1AQ7D-201610-P05 demonstrated that 10 of the 22 product lots manufactured by P05 coating machine did not meet the standard; and that of 1AQ7D-201610-P09 revealed that 10 of the 15 product lots manufactured by P09 coating machine did not meet the standard. In total, 7, 10, and 10 products did not meet the coating standards when processed by the P01, P05, and P09 coating machines, respectively. The percentage of classified products processed by P05 and P09 that did not meet the yield standards in the same month was nearly 50% of the month's total production. Therefore, the rules determined in this study exhibited a certain level of credibility. Twenty-seven *NG* manufacturing serial numbers were organized to collect coating machine data to compare the process data.

3) DECISION TREE ANALYSIS

The association rules were adopted to determine the rules of the frequent item set for the process machines causing defective products. On the basis of the product items extracted from

TABLE 6. The confusion matrix input for product ID: 1AQ7D-201610.

Real \ Predict	*G*	*NG*
G	27	0
NG	1	27

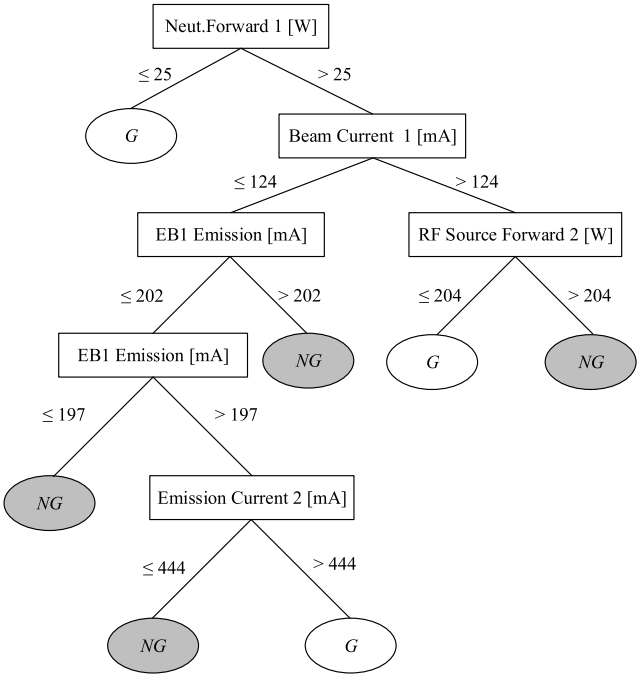

FIGURE 3. Structure chart of the decision tree of product id with 1AQ7D-201610.

the rules, this stage explored the influences of process control parameters on yield factors. Through data backtracking and discussions with domain experts, the data sheet that was initially analyzed was incorporated with new process control parameters—such as process duration, target, highly reflective film material, electric gun power, and coating rate—to determine whether effective rules were generated.

Because the case company manufactured small quantities of diverse products, the types of process data generated during the production were numerous, and the control variations were in large numbers. Therefore, process engineers could not clearly deduce which factors affected the yield. However, during actual data collection, numerous variables of both categorical and numerical types were generated. The target variables were set as *G* or *NG*, cooperation, product number classification, process machine, process machine parameters, and other input variables. Different process control variables were repeatedly input into the decision tree algorithm for calculation to determine reliable classification rules.

The decision tree algorithm adopted in this study was J48, with C4.5 serving as the core of calculation, and was suitable for managing both categorical and continuous data types. The algorithm has favorable classification accuracy and information explanation ability, and information gain ratio served as the basis of the branches. The following parameters were adopted to construct the decision tree model: the value of confidence factor is set as 0.25; 80% of data was training data and 20% testing data; the value of seed is set as 1; Unpruned was defined as False (indicating decision tree pruning); and reduced error pruning was defined as True (indicating reduction of the pruning of wrong branches).

This study organized the process-setting parameters of the major product number classification of 1AQ7D and used the decision tree to construct the classification prediction model. In Figure 3, seven leaf nodes were generated in the structure chart of the decision tree for product ID: 1AQ7D-201610. The tree size was 13, and the seven rules were divided, with three rules being positive rules about good products and four rules being negative rules about defective products. For example, a positive rule is that, if Neut.Forward 1 is less than or equal to 25 W, the product lot will be classified as *G*; a negative rule is that, if Neut.Forward 1 is greater than 25 W, Beam Current is greater than 124 mA, and RF Source Forward 2 is greater than 204 W, the product lot will be classified as *NG*. Process engineers can use their domain knowledge to determine useful rules from the extracted seven rules in Section IV.E.

In Table 6, the confusion matrix input from the decision tree revealed that 27 counts were actually *G* and accurately predicted as *G* and that 27 counts were actually *NG* and accurately predicted as *NG*. Regarding the prediction results classified by the input of the decision tree (see Table 7), the classification of good products indicated the following: *TP* Rate = 1.000, *FP* Rate = 0.036, precision = 0.966, recall = 1.000, *F*-measure = 0.982, MCC = 0.965, ROC curve = 0.991, and PRC area = 0.981. The classification of defective products indicated the following: *TP* Rate = 0.964, *FP* Rate = 0.000, precision = 1.000, recall = 0.964, *F*-measure = 0.982, MCC = 0.965, ROC curve = 0.991, and PRC area = 0.988. The *F*-measure of good and defective products was 0.982 and 0.982, respectively, revealing that recall and precision reached a desirable level. ROC curve, which explains the classification models, was 0.991.

In Figure 4, four leaf nodes were generated in the structure chart of the decision tree for product ID: MDB7D-201608. The tree size was seven, and the four rules were divided, with two rules being positive rules about good products and two rules being negative rules about defective products. For example, a positive rule is that, if Neut. Forward 2 is greater than 59 W, the product lot will be classified as *G*; a negative rule is that, if Neut.Forward 2 is less than or equal to 59 W

TABLE 7. Prediction results classified by the input of the decision tree (1AQ7D).

Class	*TP* rate	*FP* rate	Precision	Recall	*F*-measure	MCC	ROC area	PRC area
G	1.000	0.036	0.966	1.000	0.982	0.965	0.991	0.981
NG	0.964	0.000	1.000	0.964	0.982	0.965	0.991	0.988
Weighted Average	0.982	0.018	0.983	0.982	0.982	0.965	0.991	0.985

TABLE 8. The confusion matrix input for product ID: MDB7D-201608.

Real \ Predict	*G*	*NG*
G	10	18
NG	5	24

TABLE 9. Prediction results classified by the input of the decision tree (MBD7D).

Class	*TP* rate	*FP* rate	Precision	Recall	*F*-measure	MCC	ROC area	PRC area
G	0.357	0.172	0.674	0.357	0.467	0.209	0.626	0.571
NG	0.828	0.643	0.563	0.828	0.670	0.209	0.626	0.606
Weighted Average	0.596	0.412	0.618	0.596	0.570	0.209	0.626	0.589

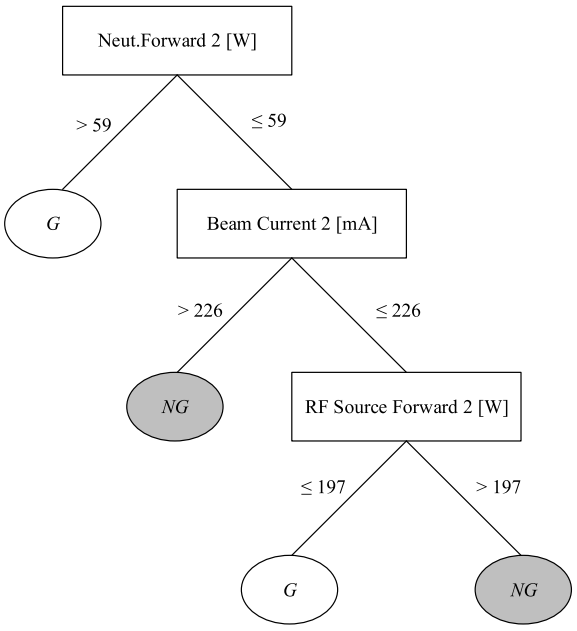

FIGURE 4. Structure chart of the decision tree of product ID: MBD7D-201608.

and Beam Current 2 is greater than 226 mA, the product lot will be classified as *NG*.

In Table 8, the confusion matrix input from the decision tree revealed that 10 counts were actually *G* and accurately predicted as *G* and that 24 counts were actually *NG* and accurately predicted as *NG*. Regarding the prediction results classified by the input of the decision tree (see Table 9), the classification of good products indicated the following: *TP* Rate = 0.357, *FP* Rate = 0.172, precision = 0.674, recall = 0.357, *F*-measure = 0.467, MCC = 0.209, ROC curve = 0.626, and PRC area = 0.571. The classification of defective products indicated the following: *TP* Rate = 0.828, *FP* Rate = 0.643, precision = 0.563, recall = 0.828, *F*-measure = 0.670, MCC = 0.209, ROC curve = 0.626, and PRC area = 0.606. The accuracy of product MBD7D was not less significant than that of product 1AQ7D; however, TP Rate = 0.828 when the model was used to predict defective product classification. Most of the indicators had superior values to those obtained for good products. ROC curve was 0.626.

E. EXPLANATION AND ASSESSMENTS

The chi-square test of independence was used to assess the variables process machine and yield. Cramer's *V* coefficient did not exceed 0.4, even though the results were significant. A product's final yield was determined based on the processing combination of several process machines. Defects in a product can be caused during cutting, testing, shipping, or in another process. Relevant problems are defined in the section of the research limitations and scope.

In the calculation of the association rules in the first stage, six rules were extracted after filtering based on the set thresholds, such as minimal support is equal to 5%, minimal confidence is equal to 80%, and lift is equal to 1. Subsequently, the association rules were offered to and discussed by the domain experts. After four rules that were identified as repeated or meaningless were deleted, two

association rules were selected. Among the products categorized by product number, the percentage of *NG* product lots with product numbers of MBD7D-201608 and 1AQ7D-201610 was particularly high in some months. However, process machines that generated defective products were not extracted from the rules. Searching and backtracking of the historical records of the process machines indicated that the number of *NG* product lots accounted for at least 50% of the products manufactured each month. Consequently, the data obtained by this study exhibited considerable reliability.

After the Apriori algorithm was used to obtain the frequent item set of *NG* product lots, this study explored the relationship between the process setting parameters and yields. The process setting parameters for two months were compiled, and classification and prediction models were constructed based on the decision tree algorithm. Seven positive and negative rules were obtained from the decision tree model of product 1AQ7D, and four negative rules were finally extracted for the *NG* product lots, as shown in Figure 3.

1) If Neut.Forward 1 is greater than 25 W, Beam Current 1 is less than or equal to 124 mA, EB1 Emission is less than or equal to 202 mA, and EB1 Emission is less than or equal to 197 mA, *Y* will be classified as *NG*.
2) If Neut.Forward 1 is greater than 25 W, Beam Current 1 is less than or equal to 124 mA, EB1 Emission is less than or equal to 202 mA, EB1 Emission is greater than 197 mA, and Emission Current 2 is less than or equal to 444 mA, *Y* will be classified as *NG*.
3) If Neut.Forward 1 is greater than 25 W, Beam Current 1 is less than or equal to 124 mA, and EB1 Emission is greater than 202 mA, *Y* will be classified as *NG*.
4) If Neut.Forward 1 is greater than 25 W, Beam Current 1 is greater than 124 mA, and RF Source Forwards 2 is greater than 204 W, *Y* will be classified as *NG*.

The decision tree model for product MBD7D contained four positive and negative rules, and two negative rules were finally extracted for the *NG* product lots, as shown in Figure 4.

1) If Neut.Forward 2 is less than or equal to 59 W and Beam Current 2 is greater than 226 mA, *Y* will be classified as *NG*.
2) If Neut.Forward 2 is less than or equal to 59 W, Beam Current 2 is less than or equal to 226 mA, and RF Source Forward 2 is greater than 197 W, *Y* will be classified as *NG*.

The results input from the binary confusion matrix revealed that the accuracy rate of the classified *G*-lot and *NG*-lot was higher than 0.7. However, the quality of a classification model was determined by its accuracy rate. The importance associated with the categories differed when the percentage of a category was low and when that category received more attention than another. In terms of the process yield, the ratios of good to defective products generated from the product lines were inconsistent, and employing only the accuracy rate could determine the rule of product yield. For process engineers, the information and rules causing the production of defective products warrants more attention and deeper analysis; thus, several other indicators were adopted to assess the models. After the indicators were assessed, the indicators of the various items for defective products were all greater than 0.7. As a result of the company's adjustment, the defect rate decreased from 20% to 5%; finally, a data mining architecture for the optical TFF production line was constructed based on the number of cases.

The contributions of this study are both academic and practical. From the viewpoint of academic research, most of the product yield enhancement research belongs to first- or second-category studies (see Section II.B), but this research falls into the third. Regardless of identifying the poor-quality process and machine or classifying types of product defects, this research has not only identified the poor-quality process and machine and extracted rules of manufacturing process parameters, but also consulted with domain experts to extract the useful rules and so improve product yield. From the viewpoint of practical research, this research is an empirical study in which the case manufacturer faced a competitive market with diverse products in small quantities and validated a data-driven approach. After the case manufacturer implemented the approach, the product defect rate decreased significantly.

In conclusion, this study has developed a six-step data-driven approach and validated by an empirical case. The approach resulted in an improved product yield. In practice, complicated process problems cannot be solved with a single analytic method because each industry has its own production modes and issues Kamsu-Foguem *et al.* [57] reported that complicated process problems can be diagnosed and analyzed through a combination of the knowledge of domain experts and data mining techniques. Alonso *et al.* [58] categorized the knowledge of domain experts in terms of their ability to remove inaccurate tests, ranges, and noise before data mining is employed, their ability to select and verify related modes among the candidate modes for the data mining system, and guidance during model generation once the required population has been selected. Consequently, this study adopted domain experts' professional knowledge and data mining to solve process problems that occurred during flexible production modes, consisting of small quantities of diverse products. Furthermore, an empirical study has been validated. Hence, manufacturers under the environment of diverse products produced in small quantities can use the approach developed in this study to improve their product yield.

V. CONCLUSION AND FUTURE RESEARCH

This study has made three contributions. First, it designed a data-driven approach that is applicable to the production of diverse products in small quantities. This production mode requires manufacturers to fulfill customers' orders by the customers' due date. Quantities may be produced in small batches by using identical or non-identical machines. The approach described here integrated the processes of data collection, preprocessing, analysis, and evaluation. The

manufacturing industry can apply these methods to diagnose process problems, improve yields, and prevent accidents.

Second, domain knowledge and process engineers' experience were combined to collect and select manufacturing process (or receipt) parameters as key variables. The clear and definite process problems could assist enterprises in defining the scope of a required data analysis in advance. The approaches to and explanations for data preprocessing in this study enabling enterprises to execute data integration, cleaning, and conversion, thereby improving the quality of the analyzed data.

Finally, actual case data were imported into this study to construct a data mining architecture and verify the research model. Several filters and indicators were used to classify the association rules and decision tree, which was used to predict the model. Through rule extraction, malfunctioning machines were detected and the factors causing the errors, traditionally determined by trial and error by an experienced engineer, were revealed, increasing the search speed and accurately offering analytic results to process engineers. The data can serve as a reference for process engineers to modify and offer feedback on process control variables. Applications at the case company resulted in an increase in products' final yields.

This study concludes with three suggestions for future studies. First, a database model can be incorporated into data collection and preprocessing to integrate all the product defects. Information on machines was classified and connected to enterprises' internal online databases. The production status of machines was updated in a timely manner according to the products waiting to be manufactured, based on historical data. If historical data indicate that a product yield was low in the past, the data mining model can be activated to search the optimal intervals for process parameters, enabling process engineers to make beneficial engineering modifications.

In addition, this study relied on product yield to judge the target variables associated with process problems and considered only the yield derived from a single process. However, several factors affect product yield. Future studies can incorporate the machine information of different process sections into the research scope to determine whether interactive relationships exist among different process machines.

Finally, in the future, when the focus will be on fully intelligent and automated manufacturing, more emphasis should be placed on time-series data to combine automatization with analytic models. The aim is to analyze automatically generated real-time process data and set the warning boundaries of each process section, thereby reducing the need for product remanufacturing. China's access to global TFT-LCD industry has led to changes in international market shares. Therefore, it is recommended incorporating China, Taiwan, Japan, and Korea, in analysis, in order to understand the trend and competition of global panel industry, and provide companies with more forward looking suggestions for improving corporate competitiveness.

REFERENCES

[1] Z. Y. Lee, G. T. R. Lin, and S. J. Lee, "Measuring dynamic operation efficiency for universal top 10 TFT-LCDs by improved data envelopment analysis," *J. Sci. Ind. Res. India*, vol. 77, no. 8, pp. 447–450, 2018.

[2] D. Preuveneers and E. Ilie-Zudor, "The intelligent industry of the future: A survey on emerging trends, research challenges and opportunities in industry 4.0," *J. Ambient Intell. Smart Environ.*, vol. 9, no. 3, pp. 287–298, Apr. 2017.

[3] K. Kwon, D. Kang, Y. Yoon, J.-S. Sohn, and I.-J. Chung, "A real time process management system using RFID data mining," *Comput. Ind.*, vol. 65, no. 4, pp. 721–732, May 2014.

[4] M. Pinzone, F. Albè, D. Orlandelli, I. Barletta, C. Berlin, B. Johansson, and M. Taisch, "A framework for operative and social sustainability functionalities in human-centric cyber-physical production systems," *Comput. Ind. Eng.*, vol. 139, Jan. 2020, Art. no. 105132.

[5] Y. Liu, L. Wang, X. V. Wang, X. Xu, and L. Zhang, "Scheduling in cloud manufacturing: State-of-the-art and research challenges," *Int. J. Prod. Res.*, vol. 57, nos. 15–16, pp. 4854–4879, Aug. 2019.

[6] P. Štefanič, M. Cigale, A. C. Jones, L. Knight, I. Taylor, C. Istrate, G. Suciu, A. Ulisses, V. Stankovski, S. Taherizadeh, G. F. Salado, S. Koulouzis, P. Martin, and Z. Zhao, "SWITCH workbench: A novel approach for the development and deployment of time-critical microservice-based cloud-native applications," *Future Gener. Comput. Syst.*, vol. 99, pp. 197–212, Oct. 2019.

[7] L. Zhou, L. Zhang, L. Ren, and J. Wang, "Real-time scheduling of cloud manufacturing services based on dynamic data-driven simulation," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5042–5051, Sep. 2019.

[8] K. Evans, J. Trnkoczy, G. Suciu, V. Suciu, P. Martin, J. Wang, Z. Zhao, A. Jones, A. Preece, F. Quevedo, D. Rogers, I. Spasić, I. Taylor, V. Stankovski, and S. Taherizadeh, "Dynamically reconfigurable workflows for time-critical applications," in *Proc. 10th Workshop Workflows Support Large-Scale Sci. (WORKS)*, 2015, pp. 1–10.

[9] J. Manyika, J. Sinclair, R. Dobbs, G. Strube, L. Rassey, J. Mischke, and J. Remes, *Manufacturing the Future: The Next Era of Global Growth and Innovation*. New York, NY, USA: McKinsey Global Institute, 2012.

[10] J. Canito, P. Ramos, S. Moro, and P. Rita, "Unfolding the relations between companies and technologies under the big data umbrella," *Comput. Ind.*, vol. 99, pp. 1–8, Aug. 2018.

[11] C. M. Flath and N. Stein, "Towards a data science toolbox for industrial analytics applications," *Comput. Ind.*, vol. 94, pp. 16–25, Jan. 2018.

[12] J. Yan, Y. Meng, L. Lu, and L. Li, "Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance," *IEEE Access*, vol. 5, pp. 23484–23491, 2017.

[13] J. Wan and M. Xia, "Cloud-assisted cyber-physical systems for the implementation of industry 4.0," *Mobile Netw. Appl.*, vol. 22, no. 6, pp. 1157–1158, Dec. 2017.

[14] Y. Kang and L. Zhou, "RubE: Rule-based methods for extracting product features from online consumer reviews," *Inf. Manage.*, vol. 54, no. 2, pp. 166–176, Mar. 2017.

[15] R. Y. Zhong, S. T. Newman, G. Q. Huang, and S. Lan, "Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives," *Comput. Ind. Eng.*, vol. 101, pp. 572–591, Nov. 2016.

[16] S. Kang, E. Kim, J. Shim, S. Cho, W. Chang, and J. Kim, "Mining the relationship between production and customer service data for failure analysis of industrial products," *Comput. Ind. Eng.*, vol. 106, pp. 137–146, Apr. 2017.

[17] Y.-J. Chen, C.-Y. Fan, and K.-H. Chang, "Manufacturing intelligence for reducing false alarm of defect classification by integrating similarity matching approach in CMOS image sensor manufacturing," *Comput. Ind. Eng.*, vol. 99, pp. 465–473, Sep. 2016.

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

[19] C.-F. Chien, C.-Y. Hsu, and P.-N. Chen, "Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence," *Flexible Services Manuf. J.*, vol. 25, no. 3, pp. 367–388, Sep. 2013.

[20] C.-F. Chien, W.-C. Wang, and J.-C. Cheng, "Data mining for yield enhancement in semiconductor manufacturing and an empirical study," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 192–198, Jul. 2007.

[21] P. C. Chu, C. F. Chien, and C. C. Chen, "Analyzing TFT-LCD array big data for yield enhancement and an empirical study of TFT-LCD manufacturing in Taiwan," *Int. J. Ind. Eng.-Theory*, vol. 23, no. 5, pp. 318–331, 2016.

[22] C.-F. Chien and C.-Y. Hsu, "Data mining for optimizing IC feature designs to enhance overall wafer effectiveness," *IEEE Trans. Semicond. Manuf.*, vol. 27, no. 1, pp. 71–82, Feb. 2014.
[23] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996.
[24] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big data computing and clouds: Trends and future directions," *J. Parallel Distrib. Comput.*, vols. 79–80, pp. 3–15, May 2015.
[25] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, 2nd ed. Danvers, MA, USA: Wiley, 2004.
[26] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Mach. Learn.*, vol. 40, no. 3, pp. 203–228, 2000.
[27] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[28] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *J. Artif. Intell. Res.*, vol. 4, pp. 77–90, Mar. 1996.
[29] C. Cobos, J. Zuniga, J. Guarin, E. Leon, and M. Mendoza, "CMIN—A CRISP-DM-based case tool for supporting data mining projects," *Ingeniería Investigación*, vol. 30, no. 3, pp. 45–56, 2010.
[30] A. Afify, "A novel algorithm for fuzzy rule induction in data mining," *Proc. Inst. Mech. Eng. C, J. Mech. Eng. Sci.*, vol. 228, no. 5, pp. 877–895, Apr. 2014.
[31] C. C. Lien, "The application of crisp and fuzzy decision trees to monitor insurance customer database," *Int. Inf. Inst.*, vol. 15, no. 9, pp. 3871–3876, 2012.
[32] A. A. Afify, "A fuzzy rule induction algorithm for discovering classification rules," *J. Intell. Fuzzy Syst.*, vol. 30, no. 6, pp. 3067–3085, Apr. 2016.
[33] S. C. Chen and M. Y. Huang, "Constructing credit auditing and control & management model with data mining technique," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5359–5365, 2011.
[34] M. Bohanec, M. Robnik-Šikonja, and M. K. Borštnar, "Decision-making framework with double-loop learning through interpretable black-box machine learning models," *Ind. Manage. Data Syst.*, vol. 117, no. 7, pp. 1389–1406, Aug. 2017.
[35] Y. Chen, Y. Chen, and A. Oztekin, "A hybrid data envelopment analysis approach to analyse college graduation rate at higher education institutions," *INFOR, Inf. Syst. Oper. Res.*, vol. 55, no. 3, pp. 188–210, Jul. 2017.
[36] G. Ulutagay, F. Ecer, and E. Nasibov, "Performance evaluation of industrial enterprises via fuzzy inference system approach: A case study," *Soft Comput.*, vol. 19, no. 2, pp. 449–458, Feb. 2015.
[37] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 step-by-step data mining guide 2000," SPSS Inc., Chicago, IL, USA, 2013, pp. 1–78.
[38] F. Rousseaux, "BIG DATA and data-driven intelligent predictive algorithms to support creativity in industrial engineering," *Comput. Ind. Eng.*, vol. 112, pp. 459–465, Oct. 2017.
[39] M. Subramaniyan, A. Skoogh, H. Salomonsson, P. Bangalore, and J. Bokrantz, "A data-driven algorithm to predict throughput bottlenecks in a production system based on active periods of the machines," *Comput. Ind. Eng.*, vol. 125, pp. 533–544, Nov. 2018.
[40] R. Addo-Tenkorang and P. T. Helo, "Big data applications in operations/supply-chain management: A literature review," *Comput. Ind. Eng.*, vol. 101, pp. 528–543, Nov. 2016.
[41] C.-F. Chien, S. Lin, and J.-C. Cheng, "Construct fuzzy decision tree for mining interrelated semiconductor manufacturing data for yield enhancement," *J. Qual.*, vol. 15, no. 3, pp. 193–210, 2008.
[42] K. K.-W. Tu, J. C.-S. Lee, and H. H.-S. Lu, "A novel statistical method for automatically partitioning tools according to Engineers' tolerance control in process improvement," *IEEE Trans. Semicond. Manuf.*, vol. 22, no. 3, pp. 373–380, Aug. 2009.
[43] R. Kittler and W. Wang, "Data mining for yield improvements," in *Proc. Int. Conf. Model. Anal. Semiconductor Manuf.*, 2000, pp. 270–277.
[44] D. Braha and A. Shmilovici, "On the use of decision tree induction for discovery of interactions in a photolithographic process," *IEEE Trans. Semicond. Manuf.*, vol. 16, no. 4, pp. 644–652, Nov. 2003.
[45] Y.-J. Chen, T.-H. Lin, K.-H. Chang, and C.-F. Chien, "Feature extraction for defect classification and yield enhancement in color filter and microlens manufacturing: An empirical study," *J. Ind. Prod. Eng.*, vol. 30, no. 8, pp. 510–517, Dec. 2013.
[46] D. Braha and A. Shmilovici, "Data mining for improving a cleaning process in the semiconductor industry," *IEEE Trans. Semicond. Manuf.*, vol. 15, no. 1, pp. 91–101, Aug. 2002.
[47] B. M. Haddad, S. Yang, L. J. Karam, J. Ye, N. S. Patel, and M. W. Braun, "Multifeature, sparse-based approach for defects detection and classification in semiconductor units," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 1, pp. 145–159, Jan. 2018.
[48] C.-F. Chien, C.-W. Liu, and S.-C. Chuang, "Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement," *Int. J. Prod. Res.*, vol. 55, no. 17, pp. 5095–5107, Sep. 2017.
[49] K. Nakata, R. Orihara, Y. Mizuoka, and K. Takagi, "A comprehensive big-data-based monitoring system for yield enhancement in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 4, pp. 339–344, Nov. 2017.
[50] A. Casali and C. Ernst, "Discovering correlated parameters in semiconductor manufacturing processes: A data mining approach," *IEEE Trans. Semicond. Manuf.*, vol. 25, no. 1, pp. 118–127, Feb. 2012.
[51] P. Bect, Z. Simeu-Abazi, and P.-L. Maisonneuve, "Identification of abnormal events by data monitoring: Application to complex systems," *Comput. Ind.*, vol. 68, pp. 78–88, Apr. 2015.
[52] A. Chen and A. Hong, "Sample-efficient regression trees (SERT) for semiconductor yield loss analysis," *IEEE Trans. Semicond. Manuf.*, vol. 23, no. 3, pp. 358–369, Aug. 2010.
[53] J. D. E. S. Carvalho, F. M. Santoro, and K. Revoredo, "A method to infer the need to update situations in business process adaptation," *Comput. Ind.*, vol. 71, pp. 128–143, Aug. 2015.
[54] W. A. Schewhart, *Economic Control of Quality of Manufactured Product*. New Providence, NJ, USA: Bell Telelphone Laboratories, 1931.
[55] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 1993, vol. 22, no. 2, pp. 207–216.
[56] P. Crewson, *Applied Statistics Handbook*. Winter Garden, FL, USA: AcaStat Software, 2006.
[57] B. Kamsu-Foguem, G. Tchuenté-Foguem, L. Allart, Y. Zennir, C. Vilhelm, H. Mehdaoui, D. Zitouni, H. Hubert, M. Lemdani, and P. Ravaux, "User-centered visual analysis using a hybrid reasoning architecture for intensive care units," *Decis. Support Syst.*, vol. 54, no. 1, pp. 496–509, Dec. 2012.
[58] F. Alonso, L. Martínez, A. Pérez, and J. P. Valente, "Cooperation between expert knowledge and data mining discovered knowledge: Lessons learned," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 7524–7535, Jun. 2012.

JRJUNG LYU received the Ph.D. degree from the Department of Industrial Engineering, University of Iowa, USA. He is currently a Professor with the Department of Industrial and Information Management, National Cheng Kung University, Taiwan. He has participated in many projects, public services, and reviewing committees. He has published more than 80 journal articles and earned the National Quality Award, in 2002. He has worked on research projects in the areas of new product development, collaborative design, and supply chain management.

CHIA WEN LIANG received the M.S. degree from the Department of Industrial and Information Management, National Cheng Kung University, Taiwan. His research interest includes data mining technology for improving product yield.

PING-SHUN CHEN received the Ph.D. degree from the Department of Industrial and Systems Engineering, Texas A&M University, USA. He is currently a Professor with the Department of Industrial and Systems Engineering, Chung Yuan Christian University, Taiwan. His research areas focus on supply chain management, system simulation, and healthcare simulation applications.