IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 June 2022, accepted 2 July 2022, date of publication 6 July 2022, date of current version 14 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3188871

RESEARCH ARTICLE

Semi-GAN: An Improved GAN-Based Missing Data Imputation Method for the Semiconductor Industry

SUN-YONG LEE[1,2], TIMOTHY PAUL CONNERTON[1], YEON-WOO LEE[3], DAEYOUNG KIM[4], DONGHWAN KIM[5], AND JIN-HO KIM[6]

[1]Business School Lausanne, 1022 Chavannes, Switzerland
[2]Seoul School of Integrated Sciences & Technologies, Seoul 03767, South Korea
[3]Bae, Kim, and Lee (LLC) Seoul 03161, South Korea
[4]Research Institute for AIdentyx, San Jose, CA, USA 95134
[5]Research Institute for BISTelligence Inc., Seoul 06754, South Korea
[6]Department of AI and Big Data, Swiss School of Management, Bellinzona 6500, Switzerland

Corresponding author: Sun-Yong Lee (ip102803@gmail.com)

ABSTRACT Complete data are required for the operation, maintenance, and detection of faults in semiconductor equipment. Missing data occur frequently because of defects such as sensor, data storage, and communication faults, leading to reductions in yield, quality, and productivity. Although many attempts have been made to solve this problem in other fields, few studies have specifically addressed data imputation in the semiconductor industry. In this study, an improved generative adversarial network (GAN)-based missing data imputation for the semiconductor industry called Semi-GAN is proposed. This study introduces a machine learning approach for dealing with data imputation in the semiconductor industry. The proposed method was applied to real data and evaluated using traditional techniques. In particular, the proposed method showed excellent results compared to traditional attribution methods when all missing data ratios in the experiments were less than 20%. It was also observed to be superior when simple and repetitive patterns were omitted rather than repetitive but not simple patterns.

INDEX TERMS Data imputation, deep learning, fault classification and detection, generative adversarial networks, machine learning, missing data, semiconductor equipment.

I. INTRODUCTION

Missing data are one of the most important problems in the semiconductor industry. Missing data occur frequently because of problems such as sensor, data storage, and communication faults. Complete data play a crucial role in the operation, maintenance, and detection of faults in semiconductor equipment [1]. Data (e.g., the pressure within a chamber) play a very important role in operating and maintaining equipment in the semiconductor field, and losing data can lead to reductions in yield, quality, and productivity. These missing data have a negative effect on the stability of production systems, which can cause problems with quality and yield. The University of California Irvine (UCI) SECOM dataset [2] used in this study contains information on semiconductor production lines and includes products with errors in their test lines or undesirable properties. This dataset includes missing values, unbalanced levels, and noise functions that are similar to most semiconductor manufacturing data [3].

Figure 1 shows the missing values in the SECOM semiconductor data. In the figure, the blank (white) parts represent missing values. Therefore, data imputation is required in daily operations in the semiconductor industry.

Fault detection and classification (FDC) data must be closely managed when operating under precise process conditions to ensure semiconductor yields and quality [4]. However, some traditional missing data methods have poor accuracy, and there have been almost no studies in the semiconductor industry to resolve this problem.

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

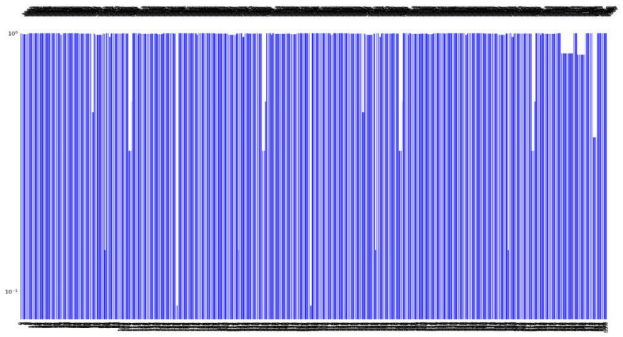

FIGURE 1. Missing data pattern of the SECOM dataset.

This paper proposes an FDC data imputation method that uses a generative adversarial network (GAN) model to resolve this problem and improve the accuracy when data are missing. GAN is an algorithm that is mainly used for image and text generation. For example, GAN learns conceptually in a way that makes the generators who create images compete with the discriminator who discriminates whether the images are real or fake. Similarly, if the missing data need to be restored, it will show good performance if the GAN method is applied. Missing data yield biased estimates, which may degrade the statistical ability of the study and lead to false conclusions. Despite various methods to replace missing data, a method using GAN has never been verified in the semiconductor industry, so we compare it with various methods. This is significant because there has been no research on imputation using the GAN algorithm in semiconductor FDC data. First, a GAN is employed as a framework to train the generative network. Deep neural networks (DNNs) are then used in both the generative and discrimination models. The data imputation model trained by the deep learning model uses several input data parameters to effectively restore data. The proposed method was verified based on FDC data from semiconductor equipment. The existing data imputation methods were used for comparison.

The precise imputation method proposed in this study could improve accuracy when analyzing correlation with yields, and it can be used not only in the semiconductor industry but also in the display, medicine, automobile, and steel industries. Through the creation of a platform and standardization, it is expected that this method will spread laterally to other industries and make significant contributions toward improving yields and increasing quality.

García-Laencina *et al.* [5] reported that the common problem of missing or unknown data is encountered in most actual field situations. For example, if a sensor fails during a production process, but sufficient information is implicitly included within the remaining sensor data, it is not necessary to halt all operations. In another example of the importance of handling missing data, 45% of the datasets in the UCI repository include missing data, which is one of the dataset collections most commonly used in benchmarking system training procedures. Generally, pattern classification with missing data is related to the problems of handling missing data and pattern classification. The technical contributions of this study are as follows.

I. A situation of missing values is assumed, which may occur in actual manufacturing data.

II. This is the first application paper applied to real semiconductor process data.

III. Through the rather simple idea of adding a loss to the validation term, meaningful results were shown in the proposed domain.

II. BACKGROUND THEORY

There are three main methods to replace missing data. The first method was based on statistical analysis. Second, it is based on machine learning. Third, it is based on deep learning [6]. Statistical analysis methods include mean imputation, regression imputation, hot deck imputation, multiple imputation, and multiple implications by chained equations (MICE). Machine-learning and deep-learning-based imputation methods are elaborate procedures, that generally generate prediction models that can estimate missing values. These approaches model missing data estimates based on the information available in the dataset. Machine learning methods [7] include k-nearest neighbor (k-NN) implementation, self-organizing map (SOM) imputation, multilayer perceptron (MLP) imputation, decision tree imputation, random forests (RFs), and support vector machines (SVMs). Deep learning methods include auto-associative neural network (AANN) implementation, NN ensemble implementation, recurrent neural networks (RNNs), and generative adversarial networks (GANs) which are intended for use in this study [8].

Figure 2 summarizes the most important missing data technologies for pattern classification and highlights their advantages and disadvantages.

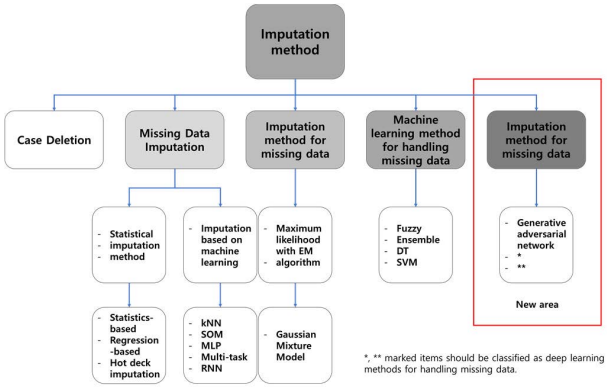

FIGURE 2. Framework of methods for pattern classification with missing data.

A. STATISTICAL BASED IMPUTATION

The mean imputation method converts incomplete data into complete data by replacing the missing values with a mean value suitable for the data obtained through observations and

experiments. Analysis was then performed by treating the complete data as if they were obtained through observations or experiments. Buck's method [9], which is an extension of the conditional mean imputation method, is widely used. Mean imputation methods are simple to use and have better efficiency than complete analysis.

Regression imputation is suitable for cases where the missing variables of interest are related to the data available for the entire sample. The missing elements were filled with the values predicted via regression analysis using the elements of the existing vectors. The regression method used varies according to the nature of the data [10].

Hot- and cold-deck imputation was performed after assigning certain suitable probability values when imputing the missing values with statistics that were estimated based on the data observed in the mean imputation method. Hot-deck imputation has the disadvantage of estimating missing data elements based on single complete vectors of datasets and ignoring general properties. Another alternative is the cold-deck imputation method. This is similar to the hot deck method, but the data source must be different from the current dataset [10]. Therefore, the cold-deck replacement method was used to replace missing values with responses from similar items in a previous sample survey.

In the MICE concept, an imputation model is specified for each variable, and the data are imputed by a variable [11]. According to a study by [12] on multiple imputation by chained equations, MICE is a special multiple imputation technique [13]. The MICE approach is a principal method for processing missing data, but it is important to recognize certain complexities and limitations. In terms of flexibility, MICE provides great advantages over other missing data techniques; however, the greatest disadvantage is that it does not have the same theoretical definitions as other imputation methods. A series of conditional distributions, as in the case where a series of regression models are used, may not correspond to a suitable joint distribution.

B. MACHINE LEARNING BASED IMPUTATION

The k-NN imputation method is a general hot-deck method that minimizes similarity measurements by selecting the nearest neighbor when the similarity measurements are complete. The nearest and most similar neighbor is discovered by minimizing the distance function [14].

The k-NN method was far superior to the other tested methods (mean imputation and imputation based on singular value decomposition), and it was robust in terms of the amount and type of missing data. The main disadvantage is that the k-NN method (in all data parts) incurs high computational costs when finding the most similar case in all datasets.

SOM imputation was originally developed to imitate the form of certain neurons in the brain [15]. In [16], this method was compared with hot-deck and MLP imputation, and it was observed that the SOM-based method performed better than the other two methods. In other models, such as MLP, incomplete observations can be used during the training stage. Consistent with this idea, data loss was missing in the tree-structured SOM (TS-SOM) created by Piela [17], which consisted of several SOMs arranged in a tree structure. The main advantages of this approach over the basic SOM are that it converges more quickly and has computational advantages when there are many input vectors. For each input vector with missing values, its image node is chosen only to measure the distances with known attributes, and each missing value is imputed based on the weights of the activation group of nodes in the incomplete attributes [5].

The basic MLP imputation method is a regression model in which only complete instances are used to train the MLP. The given input characteristics are used to learn each incomplete property from the other given properties (used as the output), which is a useful tool for recreating missing values. However, the main disadvantage of this method is that several MLP models must be designed—one for each combination of missing variables—when a missing item has several properties.

Three well-known approaches are used in the decision-tree imputation method: ID3, C4.5, and CN2. In this process, missing values can be processed for all properties of all training and test sets [18], [19]. RF is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [20]. Similarly, RF imputes the missing values. Random forests correct the tendency of decision trees to overfit their training set.

SVM imputation is a supervised learning model with associated learning algorithms that analyzes the data used for classification and regression analysis. The SVM algorithm is a popular machine -learning tool that offers solutions for both classification and regression problems.

Recently, several studies have extended the standard formulas of SVM to handle the uncertainty in input data [21], [22]. Jinbo and Zhang [22] proposed a modified risk function that considers the uncertainty in the predicted results when a value is lost. This was achieved by integrating the stochastic models of missing data. This method generalizes the mean imputation approach in linear cases, and the proposed kernel machine is reduced to a standard SVM in the absence of input values [21].

C. DEEP LEARNING BASED IMPUTATION

AANN-based imputation is a completely connected neuron set. Each neuron receives an input from and sends the output to all other neurons. Some studies have examined the use of this type of network to perform missing data imputation [23]. Missing data imputation is performed using the output unit, which learns the corresponding incomplete attributes [5].

Neural network ensemble imputation models have been used to classify incomplete data [24]. Sharpe and Solly proposed a process known as network reduction. In this method, a single set of MLPs is generated, and each MLP performs classification according to different possible combinations

of complete features to cover the entire range of properties with missing values. The main disadvantage of this method is that it requires a large number of neurons when several combinations of incomplete properties are present.

An RNN imputation is an architecture with a feedback connection in each unit. [25] proposed an RNN that uses feedback from the input units to estimate missing data. The authors showed that it is possible to improve output predictions by considering the dependencies, if any exist, between input variables. The recurrent network performed far better than a standard network with missing values imputed by mean values [26].

Although Little and Rubin [10] provided accurate mathematical and statistical background information for handling missing data in classification problems, they did not discuss machine learning methods. Our main goal was to emphasize machine learning-based approaches. Specifically, we present an imputation method for missing data that uses the new deep learning method GAN, as shown in the area within the red window in Figure 2.

Table 1 presents a comparison of imputation methods that use machine learning and deep learning. Most studies are related to the medical industry, and they were compared with energy-related studies on the wind turbine industry and semiconductor production. Specifically, there are few studies on imputation based on GANs [27]–[32]. However, there are some limitations of these studies. First, the imputation method using GAN is mostly focused on image data [27], [28]. Second, even if the source of data is not image data, it is general data, not time- series data [29]. Finally, the latest study used GAN [30]–[32]. Specifically, Yoon *et al*. [32] proposed a hint generator, which dramatically boosted performance. However, this is not a study that reflects the pattern of actual missing values occurring at the manufacturing site. In addition, since its optimization process is difficult and unstable, the computational complexity in the learning process is high [34]. Although the imputation performance may be sufficient for general data, there is still a possibility for improvement for some specific applications [35]. Several studies have been conducted in various fields to resolve this problem, but there have been almost no studies in the semiconductor industry.

TABLE 1. Comparison of imputation methods with datasets.

Industry	Dataset	Missing-Value Imputation	Classification
Medical	Breast Cancer Wisconsin	AC-GAN	SVM, RF, AB, GB, MLP, AC-GAN
Medical	KDDCUP 2018 PhysioNet 2012	GAN	KNN, MF, MICE, GAN
Medical	Ecoli Breast cancer	FKM, bPCA	Mean, k-NN, FKM, SVD, bPCA, MICE
Wind Turbine	Wind speed (3 machines)	GAN	ARMA, BPNN, GAN
Semiconductor manufacturing	SECOM (UCI)	In-Painting k-NN	LR, RF, SVM, k-NN

III. PROPOSED METHOD

A. DATA DESCRIPTION

In this study, we used real data and an 8-inch semiconductor etching device Lam Alliance 9400. The verification was performed through experiments using the FDC data of the semiconductor etching device. The method proposed in this paper was compared with the mean, RF, k-NN, MICE, and GAN methods [30]–[32].

In this study, we used the FDC data of four etching devices, which are 8-inch semiconductor manufacturing devices. There were 70 parameters per device, and 62 of these parameters were process-related. The FDC trace data were for one month, and the amount of data was 4.5 GB. Of the 62 parameters, 23 were selected as important parameters, and their FDC data were saved.

The types of FDC data include FDC data trends related to the top RF matcher, etching gas flow, bottom RF matcher, ESC, and endpoint. FDC data were generated once per second. The important parameters were selected because 17 parameters had fixed constant values as their data values, and six parameters had data values that were always 0, and these were confirmed to be parameters that were not used in the relevant processes.

In addition, 16 parameters were not important. These parameters are for general product classification rather than data mainly generated from facility sensors, such as Time, Lot id, Substrate id, Slot no., Recipe name, Product name, Step name, Status, etc. Therefore, the total number of parameters not required for the analysis was 39. These 39 parameters were excluded from the 62 parameters, and FDC data trends for the remaining 23 important parameters were used.

In this study, we selected a GAN for the imputation. In this study, some normal FDC data were used to supplement the missing FDC data. To test the imputation methods, actual missing data were not utilized in the experiments. Instead, full data were used. The FDC data for certain times were removed artificially. Subsequently, several imputation methods were used to restore the removed data. In these experiments, four other methods, namely, mean, RF, k-NN, and MICE, were compared with the proposed GAN method. The experimental dataset used in this case study consisted of 30 lots of fault data collected from actual semiconductor production equipment.

The experiments were conducted using the 23 major parameters selected in the previous data examination. To verify the imputation performance with respect to the missing values, the experiments generated missing data by randomly erasing entire rows, which reflects a realistic situation in which the data collection and storage from semiconductor equipment have occurred at a certain point. The method of randomly generating missing values, which has often been

used in previous studies of missing value imputation, makes assumptions that are unlikely to occur in semiconductor manufacturing. The present study takes this into account and creates a missing value dataset by randomly erasing entire rows.

B. COMPARATIVE EXPERIMENTS BETWEEN THE PROPOSED METHOD AND GAIN

The method proposed in this study is an improvement of the existing GAIN method, which is suitable for the characteristics of semiconductor data [32]. The semiconductor data are characterized such that the entire amount of data at one time is empty (a situation where data cannot be collected at a specific time). To consider this the existing GAIN method was modified, and an improved imputation was performed. The pseudocode of the proposed method is as follows:

TABLE 2. Pseudocode of the proposed method.

while training loss has not converged **do**

(1) Discriminator optimization

Draw k_D samples from the dataset $\{(\tilde{x}_{tr}(j), m_{tr}(j))\}_{j=1}^{k_D}$

Draw k_D i.i.d. samples, $\{\boldsymbol{z}(j)\}_{j=1}^{k_D}, of\ \boldsymbol{Z}$

Draw k_D i.i.d. samples, $\{\boldsymbol{b}(j)\}_{j=1}^{k_D}, of\ \boldsymbol{B}$

for $j = 1, ..., k_D$ **do**

$\bar{x}_{tr}(j) \leftarrow G\left(\tilde{x}_{tr}(j), \boldsymbol{m_{tr}}(j), \boldsymbol{z}(j)\right)$

$\hat{x}_{tr}(j) \leftarrow \boldsymbol{m}(j) \odot \tilde{x}_{tr}(j) + \left(\boldsymbol{1} - \boldsymbol{m_{tr}}(j)\right) \odot \bar{x}_{tr}(j)$

$\boldsymbol{h}(j) = \boldsymbol{b}(j) \odot \boldsymbol{m_{tr}}(j) + 0.5(\boldsymbol{1} - \boldsymbol{b}(j))$

end for

Update D using stochastic gradient descent (SGD)

$$\nabla_D - \sum_{j=1}^{k_D} L_D(\boldsymbol{m_{tr}}(j), D(\hat{x}_{tr}(j), \boldsymbol{h}(j)), \boldsymbol{b}(j))$$

(2) Generator optimization

Draw k_G samples from the dataset $\{(\tilde{x}_{tr}(j), \boldsymbol{m_{tr}}(j))\}_{j=1}^{k_G}$

Draw k_G i.i.d. samples, $\{\boldsymbol{z}(j)\}_{j=1}^{k_G}, of\ \boldsymbol{Z}$

Draw k_G i.i.d. samples, $\{\boldsymbol{b}(j)\}_{j=1}^{k_G}, of\ \boldsymbol{B}$

for $j = 1, ..., k_G$ **do**

$h(j) = b(j) \odot \boldsymbol{m_{tr}}(j) + 0.5(\boldsymbol{1} - \boldsymbol{b}(j))$

end for

Update G using SGD (for fixed D)

$$\nabla_G \sum_{j=1}^{k_G} L_G(\boldsymbol{m_{tr}}(j), \hat{\boldsymbol{m}}_{\boldsymbol{tr}}(j), \boldsymbol{b}(j)) + \alpha_1 L_M\left(x_{tr}(j), \tilde{x}_{tr}(j)\right)$$

(3) Validation optimization

Draw k_v samples from the dataset $\{(\tilde{x}_v(j), \boldsymbol{m_v}(j))\}_{j=1}^{k_v}$

$\bar{x}_v(j) \leftarrow G\left(\tilde{x}_v(j), \boldsymbol{m_v}(j), \boldsymbol{z}(j)\right)$

$\hat{x}_v(j) \leftarrow \boldsymbol{m_v}(j) \odot \tilde{x}_v(j) + \left(1 - \boldsymbol{m_v}(j)\right) \odot \bar{x}_v(j)$

Update V using SGD (for fixed G and D)

$$\nabla_V \sum_{j=1}^{k_V} L_G(\boldsymbol{m_{tr}}(j), \hat{\boldsymbol{m}}_{\boldsymbol{tr}}(j), \boldsymbol{b}(j)) + \alpha_1 L_M\left(x_{tr}(j), \tilde{x}_{tr}(j)\right) + \alpha_2 L_V(\tilde{x}_v(j), \bar{x}_v(j))$$

end while

As shown in Figure 3, the major differences between the proposed method and GAIN are the 1) validation data, and 2) loss function. First, the validation data were used to learn more patterns of semiconductor data. Among the training data used for learning, part of the data for which no missing data

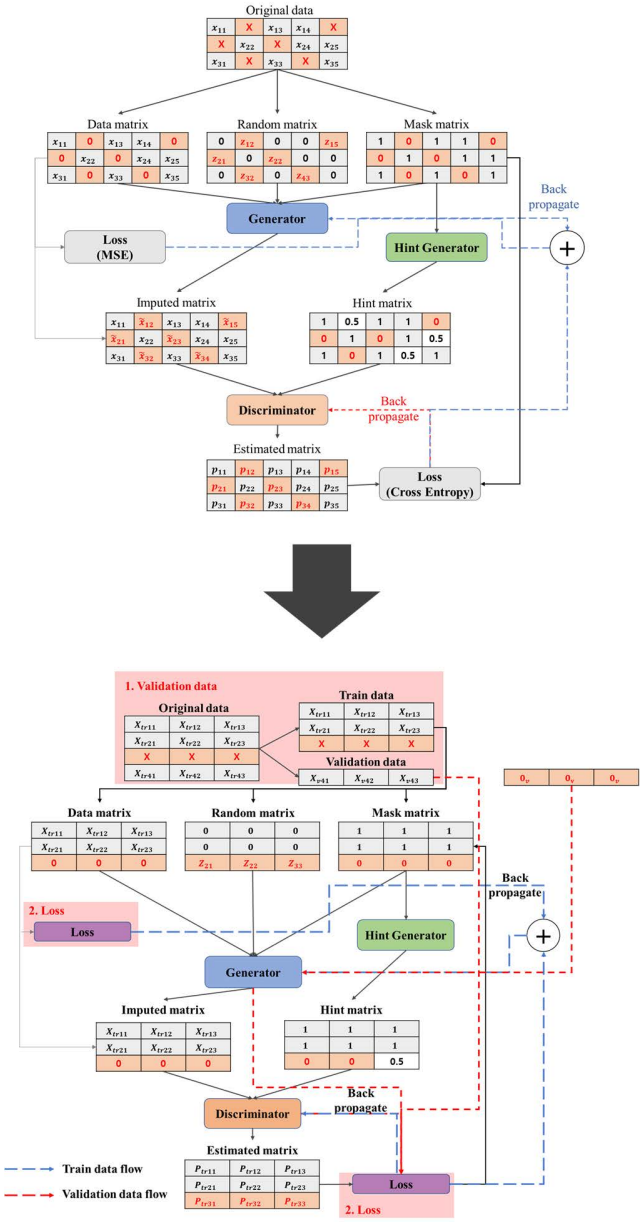

FIGURE 3. Architecture of the proposed method compared to GAIN.

were created was used as the validation data. Thus, it is possible to learn all empty data at a specific time by learning with training data and updating the learning with the difference between this and the correct answer using validation data. Second, the loss function is modified to reflect the validation process.

The validation loss was added to the existing loss function, and by defining the error of the loss function with MAE, the GAN trained the entire data distribution better than the existing MSE method.

The proposed method is similar to those proposed in [32], [33]. A detailed description of the proposed method is as follows.

The proposed method uses part of the given data X as validation data X_v and the rest of the data as training data X_{tr}. That is,

$$X = X_{tr} \cup X_v \tag{1}$$

First, the minibatch $(\tilde{x}_{tr}(j), m_{tr}(j))$ of size k_D is sampled from the dataset so that it follows *i.i.d*. Then, the discriminator D is fixed, and generator G is optimized. Here, index j refers to the minibatch's jth sample, and $m_{tr}(j)$ is a component of the masking matrix for the sampled data. Using the same method, sampling is performed on the d-dimensional noise $Z = (Z_1, Z_2, \ldots Z_d)$ and probability variables $B = (B_1, B_2, \ldots B_d) \in \{0, 1\}^d$.

Based on the sampled data, each $\bar{x}_{tr}(j), \hat{x}_{tr}(j)$ is updated, and the hint matrix component $\boldsymbol{h}(j)$ is calculated. Here, the discriminator D is updated using the stochastic gradient descent method. Here, L_D is defined as follows:

$$L_D\left(\boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{b}\right) = \sum_{i:b_i=0} \times \left[m_i \log\left(\hat{m}_i\right) + (1 - m_i) \log\left(1 - \hat{m}_i\right)\right] \tag{2}$$

Here, D is learned as follows, and $\hat{\boldsymbol{m}}(j)$ in Equation (3) is defined as $D(\hat{\boldsymbol{x}}(j), \boldsymbol{m}(j))$.

$$\min_D \sum_{j=1}^{k_D} L_D(\boldsymbol{m}(j), \hat{\boldsymbol{m}}(j), \boldsymbol{b}(j)) \tag{3}$$

Second, the minibatch $(\tilde{x}_{tr}(j), m_{tr}(j))$ of size k_G is sampled from the dataset so that it follows *i.i.d*. Then, the previously determined discriminator D is fixed, and the generator G is optimized. Thus, the d-dimensional noise Z and the probability variable B were also sampled.

The value of $\boldsymbol{h}(j)$ was calculated based on the sampled data. Here, L_G is defined as follows:

$$L_G\left(\boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{b}\right) = \sum_{i:b_i=0} (1 - m_i)\, log\left(\hat{m}_i\right) \tag{4}$$

Here, L_M is as follows.

$$L_M\left(x, x'\right) = \sum_{i=1}^{d} m_i L_M(x_i, x_i') \tag{5}$$

$$L_M\left(x_i, x_i'\right) = \begin{pmatrix} \left(x_i' - x_i\right)^2, & if\ x_i\ is\ continous, \\ -x_i\ log\left(x_i'\right), & if\ x_i\ is\ binary. \end{pmatrix} \tag{6}$$

In these definitions, L_G is applied to the missing value part $(m_i = 0)$, and L_M is applied to the part without missing values $(m_i = 1)$.

The following is a description of the proposed validation optimization method. A minibatch $(\tilde{x}_v(j), m_v(j))$ with a size k_v is input into the previously created generator G from the validation data X_v, and $\bar{x}_v(j)$ is updated. The validation data $\hat{x}_v(j)$ in X_v are updated in the same manner as in the aforementioned method. Ultimately, the validation loss is defined as the difference between $\tilde{x}_v(j)$ and $\bar{x}_v(j)$ for the fixed generator G and the discriminator D. That is, the part of the data for which missing values were not generated was set as the validation data so that the proposed method could better learn the actual data distribution [36]. Here, L_V is expressed as follows:

$$L_V(\boldsymbol{m}, \tilde{\boldsymbol{m}}, \boldsymbol{b}) = \sum_{i:b_i=0} |m_i - \tilde{m}_i| \tag{7}$$

Here, the loss function of the error is defined as MAE. This is done to use MAE, which is more robust than MSE with respect to outliers; thus, GAN learns the overall distribution of the data better than learning the outliers or one or two observed values [37].

Here, the final loss function is as follows.

$$\min_G \sum_{j=1}^{k_G} L_G(\boldsymbol{m_{tr}}(j), \hat{\boldsymbol{m}}_{\boldsymbol{tr}}(j), \boldsymbol{b}(j)) + \alpha_1 L_M(x_{tr}(j), \tilde{x}_{tr}(j)) + \alpha_2 L_V(\tilde{x}_v(j), \bar{x}_v(j)) \tag{8}$$

Here, α_1, α_2 are hyperparameters, and their optimal values are determined through a grid search.

C. VALIDATION

To verify the excellence of the proposed method, the existing GAN was used to perform a direct comparison with a GAIN that studied missing-value imputations. Experiments were performed with a training, validation, and testing ratio of 6:2:2. To compare the performance of the model according to the missing value ratio, the missing value ratios in the semiconductor test data were set at 0.1%, 0.2%, 0.5%, 1%, 2%, 5%, 10%, and 20%, and experiments were conducted. The general occurrence rate of missing data generated in the semiconductor process is less than 7% of the total data. However, 20% of the cases were included in the experimental conditions to confirm under extreme conditions. The model performance was evaluated using the MAE method, which can measure performance without being sensitive to outliers. The average MAE of the 30 lots was measured. Figure 4 shows the performance of the proposed method and GAIN for each missing- value ratio. The experimental results show that the performance of the proposed method is excellent for all missing value ratios, except for 20%. This shows that the proposed method has been designed to be more suitable for semiconductor equipment data than the existing GAIN method. Thus, it is concluded that the characteristics of the data must be properly considered even when using certain models such as GAN, and the method proposed in this paper better utilizes GAN.

IV. EXPERIMENTAL RESULTS

A. OVERALL PERFORMANCE OF THE PROPOSED METHOD

To verify the performance of the method proposed in this study, it was compared with other missing-value imputation methods. The experiments were performed on 10 datasets. Eight datasets were created for each lot with different missing data ratios (0.1%, 0.2%, 0.5%, 1%, 2%, 5%, 10%, and 20%), and the experiments were conducted. For comparison with existing methodologies, six models were considered: GAIN, the proposed method, and known missing value imputation

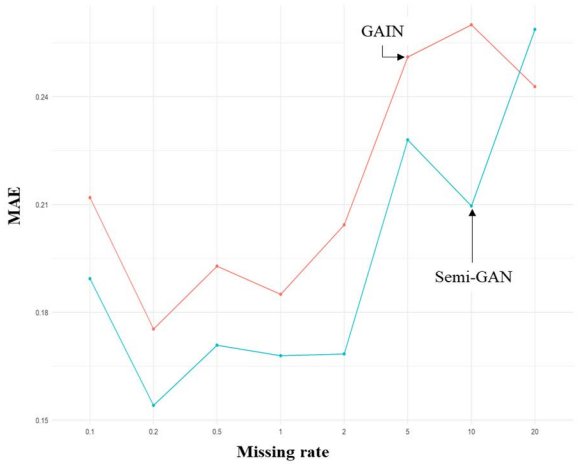

FIGURE 4. Comparison of the proposed method and GAIN.

methods such as the mean, k-NN, MICE, and RF methods. The performance of each model for each parameter was evaluated using the MAE evaluation index (the average for the eight datasets with different missing data ratios).

Table 3 shows the aggregated first rank of each model. These are the overall experimental results for the 23 parameters. The performance of each model for each parameter is shown in terms of the MAE evaluation index (the average for the eight datasets with different missing data ratios). Among the 185 results, the number of first places was compared. The results of the study were as follows: 87 RF, 60 Semi-GAN, and 31 MICE. Thus, it was confirmed that RF was ranked first, the proposed method ranked second, and MICE ranked third. It was confirmed that the proposed method and RF are valid models for semiconductor FDC data-imputation.

TABLE 3. Comparison by each model.

Model	Mean	KNN	MICE	RF	Semi-GAN
win rate	7/185	7/185	31/185	87/185	60/185
%	3.8%	3.8%	16.8%	47.0%	32.4%
Ranking	4	4	3	1	2

However, when the experimental results are viewed in terms of parameters, the missing value imputation methodology of the proposed method took first place 10 times for the MFC_CL2 parameter, and it showed good performance in all lots. However, the chamber pressure parameter did not take first place even once. Therefore, the experimental results show that there is a difference in the performance of Semi-GAN according to the lots and parameters. To perform a detailed examination of the difference in the missing value imputation analysis results by parameter, certain lots were selected and detailed analyses were performed.

B. DETAILED OVERALL PERFORMANCE OF THE PROPOSED METHOD

To examine the missing value imputation results in detail by parameter, Lot 2 was selected because it had the best results for the proposed method, and the MAE results for each parameter's model were checked. This is illustrated in Figure 5. The MAE indices were measured for each of the 23 parameters. As in the existing experimental results, the performances were compared for all models using the MAE averages for each of the eight missing data ratios. Analysis of the results showed that there was a difference in the performance of the models for each parameter. Although the overall result was not the best performance, the proposed method showed good performance in parameters that play an important role in the actual process. The proposed method showed excellent performance for nine parameters: Bottom_RF_Forward_Power, BeHe_Flow, BeHe_Pressure, EndPt_ChA, ESC_Voltage, MFC_BCL3, MFC_Cl2, MFC_N2, and MFC_SF6. Figure 5 shows the MAE results for each missing data ratio for the nine parameters that exhibited excellent performance.

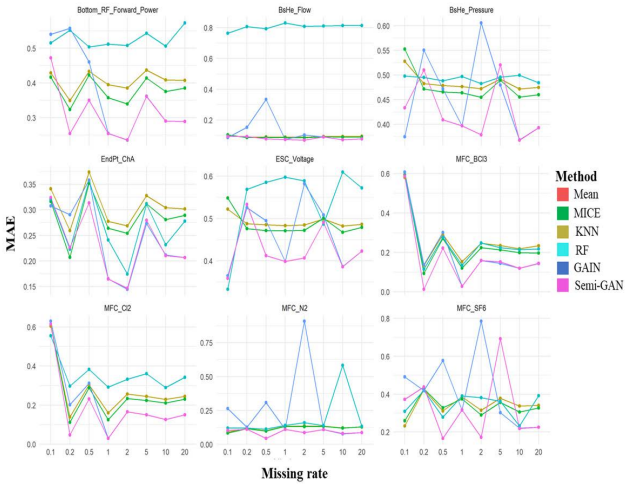

FIGURE 5. Results of MAE analysis for each parameter of Lot 2.

The proposed method showed excellent performance for all datasets with varying missing data ratios for the nine parameters shown in Figure 6. It took first place for almost all missing data ratios, and it showed particularly good results for missing data ratios of 1% and 2%. This study examined why the proposed method yielded excellent results for certain parameters, as shown in Figure 6.

It can be observed that in the case of the MFC_CL2 parameter, the proposed method took first place 10 times and showed good performance in all lots, whereas in the case of the Endpoint_Time parameter, the proposed method took first place six times and showed good performance.

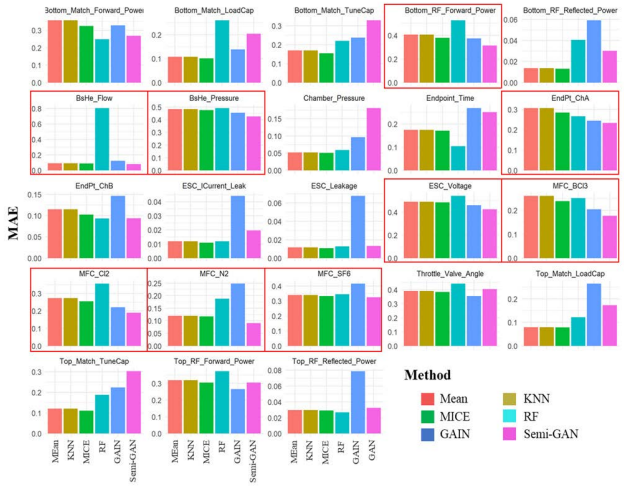

FIGURE 6. Comparison of performance of the proposed method according to MAE and parameters.

However, in the case of the chamber pressure, EndPt_ChA, Bottom_Match_TuneCap, and Top_Match_TuneCap parameters, the proposed method did not show excellent performance even once. To perform a detailed examination of these results, in which GAN performed differently for each parameter, a comparative analysis was performed on the patterns of parameters for which the proposed method was excellent and those for which it was not good, as well as their missing-value imputation characteristics.

Table 4 shows the summarized results excluding the mean in Figure 6. In all results, the proposed method showed the best performance. In particular, in the case of less than 10% of the missing rate occurring in the actual field, excellent results were obtained compared to the GAIN method. For example, in the results of the parameters Bottom_RF_Forward_Power, EndPt_CHA, MFC_BCI3, and MFC_Cl2, the proposed method exhibits much better imputation performance at a missing rate of 1% or less but shows similar results at 1% or more. Since the proposed method shows good performance when the missing ratio is less than 10%, it can be applied to solve missing problems in general manufacturing processes as well as actual semiconductor processes.

TABLE 4. Summary of the results in figure 6.

Model	MICE	KNN	RF	GAIN	Semi-GAN
Bottom_RF_Forward_Power	0.384	0.408	0.526	0.342	0.311
BsHe_Flow	0.153	0.153	0.805	0.151	0.149
BsHe_Plessure	0.475	0.468	0.481	0.442	0.423
EndPt_ChA	0.291	0.275	0.264	0.249	0.239
ESC_Voltage	0.481	0.491	0.587	0.440	0.426
MFC_BCl3	0.241	0.231	0.239	0.199	0.171
MFC_Cl2	0.245	0.256	0.353	0.220	0.191
MFC_N2	0.119	0.118	0.185	0.248	0.086
MFC_SF6	0.325	0.337	0.355	0.418	0.315

Figure 7 shows the original data, missing data, and missing data imputation results by model (MICE, Semi-GAN) for the two parameters MFC_CL2 and Endpoint_Time_signal, for which the proposed Semi-GAN method showed excellent performance in the overall analysis results. First, by observing the original data patterns, it can be observed that the patterns are repetitive owing to the characteristics of semiconductor production processes. Specifically, a simple form was repeated in the pattern. In addition, MICE's missing data imputation results included many predictions of approximately 0.5, but the Semi-GAN results showed that it accurately imputed the missing values with values on both sides near 1 and 0. The data of the repeating pattern appear simple; however, the model is difficult to learn owing to the complex relationships between the different parameters.

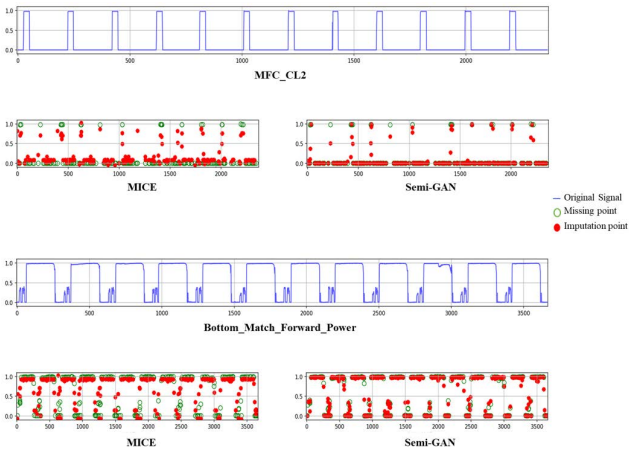

FIGURE 7. Example of parameter analysis.

In particular, it is difficult to detect minute changes in semiconductor data, and the length varies for each recipe step. Thus, the Semi-GAN method shows excellent performance in cases where the pattern is repetitive and simple. Figure 8 shows the original data, missing data, and missing data imputation results obtained by MICE and the proposed method for two parameters, the Bottom_Match_TuneCap and Endpoint_Time signals, for which the proposed method did not show excellent performance even once in the overall analysis results.

Compared with the previous two parameters, these parameters had patterns that were repetitive but not simple. The MICE methodology did not make accurate predictions, and the proposed method made predictions centered on certain values (1 or 0) and was unable to impute various missing values. Thus, it can be observed that the performance of the

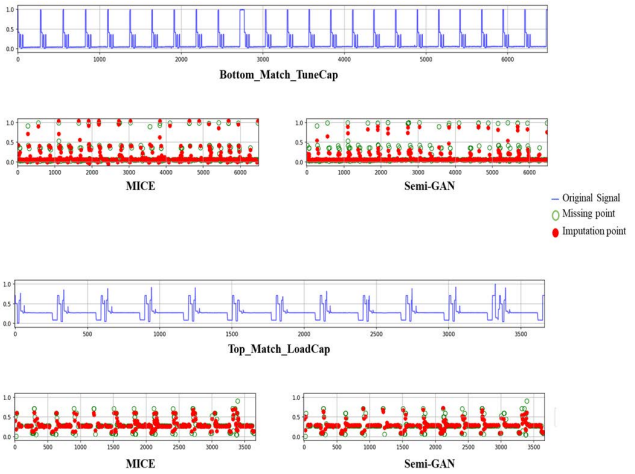

FIGURE 8. (Bottom)_Match_TuneCap, (top)_Match_LoadCap parameter analysis.

model varied according to the differences in the patterns of the parameters. Owing to its characteristics, a Semi-GAN can provide good results if a large amount of data is required for learning. Thus, it provides better results for simple and repetitive patterns.

To summarize the overall experimental results, first, when simply applying the existing GAIN method to the replacement of missing values in the data of semiconductor production facilities, the characteristics of the data are not reflected well. was able to confirm that it was suitable. In addition, through performance comparison with various other existing methodologies and the proposed method in this paper, although it did not show excellent performance in all results, it was able to prove the effectiveness of the proposed method by showing excellent performance in specific LOT and parameters. Finally, it was confirmed that the proposed method showed excellent results in a specific pattern that better reflects the characteristics of the GAN through a detailed analysis of the parameters that showed excellent performance and those that did not.

V. DISCUSSION

Most semiconductor manufacturing data contain missing values, imbalances, or noise. Missing data are one of the most common and important problems in the semiconductor industry. Missing data occur frequently because of problems such as sensor faults, communication problems, and data storage problems. Complete data play an important role in the operation, maintenance, and detection of faults in semiconductor equipment because data such as the pressure within a chamber play an important role in the operation and maintenance of equipment in the semiconductor field. Therefore, losing these data can lead to significant losses in semiconductor yield, quality, and productivity, which ultimately leads to corporate insolvency.

Therefore, missing data, overall, have a negative effect on production system stability, which can create several problems with quality and yield rates. In the end, missing data can yield biased estimates, reducing the statistical ability of the study and leading to false conclusions. Firms in the semiconductor industry have sought solutions through data imputation to resolve various problems caused by missing data. Data imputation has become an integral and daily operation in the semiconductor industry, serving to enhance productivity by reducing data loss. Accordingly, FDC data have been closely managed by operating under precise process conditions to ensure semiconductor yields and quality.

Although many efforts have been made to solve such problems stemming from missing data in the semiconductor industry, traditional missing data processing methods suffer from poor accuracy. However, almost no studies have been conducted in the semiconductor industry to resolve these problems. Therefore, this study examined various methods to resolve these problems and improve the accuracy of missing data processing. By solving the abovementioned problems, the company can save money by eliminating product scraps removing defects, improving facility operation rates, and reducing engineers' work. Because these methods have never been verified in the semiconductor industry, we apply and analyze statistical analysis methods, machine learning methods, and deep learning methods as alternatives to FDC missing data in semiconductor facilities to find the best solution.

The methodologies compared in this study demonstrate that the AI-embedded approach is sophisticated and advanced. It is suitable for generating fake data to fill the gap in missing data and improve the deficiency caused by missing data. The findings from this research have strong economic implications for industries that can be disrupted by inaccuracies caused by missing data, such as medicine and health, along with other high-precision manufacturing industries, offering business transformation practices for imputation techniques in such fields.

VI. CONCLUSION

The precise imputation method proposed in this study can improve accuracy when analyzing correlations with yields, and it can be used not only in the semiconductor industry but also in the display, medicine, automobile, and steel industries. Through the creation of a platform and standardization, it is expected that this method will spread laterally to other industries and make great contributions toward improving yields and increasing quality. Since it is almost the first case of applying the deep learning imputation method to actual manufacturing data, it will be the basis for such research in the future. In particular, FDC data have not been studied publicly well until now because of data security problems. Over time, actual semiconductor manufacturing process data are being disclosed. Therefore, increasing the number of research cases applied to actual manufacturing data will be of great help not only to the industry but also to the research of feasible and applicable methodologies.

This study was limited to one type of model for semiconductor etching equipment, which is one of the many pieces of semiconductor equipment, and it will be necessary to perform follow-up studies on different manufacturing equipment. The RF and Semi-GAN methods are applied not only to the semiconductor industry but also to other industries such as the display, medical, automobile, and steel industries. The validity of the missing data imputation method proposed in this study must be verified by applying it to other industries through the creation of a platform and standardization.

In this study, the improved proposed Semi-GAN method showed better results than the existing GAIN method, but it did not show better results than other existing methods for all parameter cases. Deep learning-based proposal methods require high-quality data. Therefore, many domain-based high-quality data are needed to improve the performance of the proposed method. In addition, the method of selecting variables can be chosen because dozens and hundreds of process parameters are complicated. Additionally, since the data patterns vary depending on the recipe, research according to the recipe should be conducted. Therefore, the model must be developed for further improvements. In addition, it will be necessary to apply the method developed in this study to an actual work site and confirm improvements in operational efficiency, product quality, etc.

REFERENCES

[1] Y.-J. Chen, B.-C. Wang, J.-Z. Wu, Y.-C. Wu, and C.-F. Chien, "Big data analytic for multivariate fault detection and classification in semiconductor manufacturing," in *Proc. 13th IEEE Conf. Autom. Sci. Eng. (CASE)*, Aug. 2017, pp. 731–736.

[2] M. Salem, S. Taheri, and J. S. Yuan, "An experimental evaluation of fault diagnosis from imbalanced and incomplete data for smart semiconductor manufacturing," *Big Data Cogn. Comput.*, vol. 2, no. 4, pp. 1–20, 2018.

[3] M. McCann, Y. Li, L. Maquire, and A. Johnston, "Causality challenge: Benchmarking relevant signal components for effective monitoring and process control," *J. Mach. Learn. Res. Work. Proc.*, vol. 6, pp. 277–288, Jan. 2010.

[4] C. F. Chien, A. C. Diaz, and Y. B. Lan, "A data mining approach for analyzing semiconductor MES and FDC data to enhance overall usage effectiveness (OUE)," *Int. J. Comput. Intell. Syst.*, vol. 7, no. 2, pp. 52–65, 2014.

[5] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, 2010.

[6] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative robust semi-supervised missing data imputation," *IEEE Access*, vol. 8, pp. 90555–90569, 2020.

[7] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020.

[8] Q. Li, H. Tan, Y. Wu, L. Ye, and F. Ding, "Traffic flow prediction with missing data imputed by tensor completion methods," *IEEE Access*, vol. 8, pp. 63188–63201, 2020.

[9] S. F. Nielsen, "Nonparametric conditional mean imputation," *J. Stat. Planning Inference*, vol. 99, no. 2, pp. 129–150, Dec. 2001.

[10] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2002.

[11] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annu. Rev. Psychol.*, vol. 60, pp. 549–576, Jan. 2009.

[12] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?" *Int. J. Methods Psychiatric Res.*, vol. 20, no. 1, pp. 40–49, Mar. 2011.

[13] S. van Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Stat. Methods Med. Res.*, vol. 16, no. 3, pp. 219–242, Jun. 2007.

[14] G. E. Batista and M. C. Monard, "Experimental comparison of K-nearest neighbor and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data," Univ. São Paulo, 2003, vol. 34.

[15] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Heidelberg, Germany: Springer, 2001.

[16] F. Fessant and S. Midenet, "Self-organising map for data imputation and correction in surveys," *Neural Comput. Appl.*, vol. 10, no. 4, pp. 300–310, 2002.

[17] P. Piela, "Introduction to SOM modelling for imputation—Techniques and technology," *Res. Off. Statist.*, vol. 2, pp. 5–19, Jan. 2002.

[18] J. R. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed. San Mateo, CA, USA: Morgan Kaufmann, 1994.

[19] G. I. Webb, "The problem of missing values in decision tree grafting," in *Advanced Topics in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 1502, G. Antoniou and J. Slaney, Eds., 1998.

[20] D. Kim, S. H. Park, and J. Baek, "A kernel Fisher discriminant analysis-based tree ensemble classifier: KFDA forest," *Int. J. Ind. Eng.*, vol. 25, no. 5, pp. 569–579, 2018.

[21] K. Pelckmans, J. D. Brabanter, J. A. Suykens, and B. De Moor, "Handling missing values in support vector machine classifiers," *Neural Netw.*, vol. 18, nos. 5–6, pp. 684–692, Aug. 2005.

[22] J. Bi and T. Zhang, "Support vector classification with input data uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1–8.

[23] T. Marwala and S. Chakraverty, "Fault classification in structures with incomplete measured data using auto associative neural networks and genetic algorithm," *Current Sci.*, vol. 90, no. 4, pp. 542–548, 2006.

[24] K. Jiang, H. Chen, and S. Yuan, "Classification for incomplete data using classifier ensembles," in *Proc. Int. Conf. Neural Netw. Brain*, 2005, pp. 559–563.

[25] Y. Bengio and F. Gingras, "Recurrent neural networks for missing or asynchronous data," in *Proc. 8th Int. Conf. Neural Inf. Process. Syst.*, 1995, pp. 395–401.

[26] J. Zhao, Y. Nie, S. Ni, and X. Sun, "Traffic data imputation and prediction: An efficient realization of deep learning," *IEEE Access*, vol. 8, pp. 46713–46722, 2020.

[27] U. Hwang, S. Choi, H.-B. Lee, and S. Yoon, "Adversarial training for disease prediction from electronic health records with missing data," 2017, *arXiv:1711.04126*.

[28] Y. Luo, X. Cai, Y. Zhang, J. Xu, and X. Yuan, "Multivariate time series imputation with generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1596–1607.

[29] S. P. Mandel J, "A comparison of six methods for missing data imputation," *J. Biometrics Biostatist.*, vol. 6, no. 1, pp. 1–6, 2015.

[30] F. Qu, J. Liu, X. Hong, and Y. Zhang, "Data imputation of wind turbine using generative adversarial nets with deep learning models," in *Neural Information Processing*, vol. 1. Cham, Switzerland: Springer, 2018.

[31] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[32] J. Yoon, J. Jordon, and M. Van Der Schaar, "Gain: Missing data imputation using generative adversarial nets," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, vol. 13, 2018, pp. 9052–9059.

[33] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[34] D. T. Neves, J. Alves, M. G. Naik, A. J. Proença, and F. Prasser, "From missing data imputation to data generation," *J. Comput. Sci.*, vol. 61, May 2022, Art. no. 101640.

[35] Z. Yao and C. Zhao, "FIGAN: A missing industrial data imputation method customized for soft sensor application," *IEEE Trans. Autom. Sci. Eng.*, early access, Dec. 8, 2021, doi: 10.1109/TASE.2021.3132037.

[36] X. Zhang, R. R. Chowdhury, J. Shang, R. Gupta, and D. Hong, "ESC-GAN: Extending spatial coverage of physical sensors," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, Feb. 2022, pp. 1347–1356.

[37] D. Kim, S. Lee, and D. Kim, "An applicable predictive maintenance framework for the absence of run-to-failure data," *Appl. Sci.*, vol. 11, no. 11, p. 5180, Jun. 2021.

SUN-YONG LEE was born in Seoul, South Korea, in 1958. He received the B.S. degree in electronics from Kwangwoon University, Seoul, in 1980, the M.B.A. degree in big data from the Business School Lausanne (BSL), Switzerland, in 2018, the Doctorate of Business Administration (DBA) degree from the BSL, and the Ph.D. degree in business administration from the Seoul School of Integrated Sciences and Technologies (aSSIST), Seoul, in 2020. From 1984 to 2016, he worked with Samsung Electronics, specializing in semiconductor and display manufacturing system design. He is currently a Professor of semiconductor and display manufacturing process with the Swiss School of Management. His research interests include mass product management, innovation by technology, and big data analysis with data mining, machine learning, deep learning, computational intelligence, pattern recognition, and signal processing.

TIMOTHY PAUL CONNERTON received the Doctorate in Business Administration (DBA) degree from the Grenoble Ecole de Management, France, with a focus on the influences of leadership characteristics on employee motivation and organizational behavior. He advises companies and teaches at the Business School Lausanne (BSL) and other universities on the subjects of strategy, change management, and leadership/organizational behavior and supervises Ph.D., master's, and bachelor's theses. In addition to being an accomplished academic, he has over 35 years of practical experience in project management, marketing and business development, strategic planning, and managing multibillion dollar construction-related and industrial manufacturing businesses worldwide. As the President and the CEO of major divisions with Norsk Hydro/Hydro Aluminum, Alcoa, and Alghanim Industries, he traversed the globe to lead extensive networks of companies in Europe, Asia, USA, Africa, and the Middle East for sustainable growth, operational performance excellence, strategic realignment, and organizational renewal through empowerment to revitalize stagnant and troubled businesses. Having taught throughout his career, created many corporate training programs with Harvard Business School, IMD, and HEC Paris, and engaged in developing his people through mentoring and coaching in his businesses, he believes in lifelong learning. Having also co-published several articles, delivered inspirational speeches, and executive seminars, and given a TEDx talk.

YEON-WOO LEE is currently the Lead Expert at South Korea's Top Law Firm, Bae, Kim, and Lee (LLC). Prior to her current job, she has researched and lectured on international business management, strategy, and competitiveness at prestigious universities for over the past 15 years in South Korea. Her main contributions involve providing advice and consultations on the topics of sustainability, the global environment, and change management for multinational corporations and government bodies. She has a broad scope of knowledge and experiences in connecting the areas of strategic management and international business in deriving cooperative innovative models that better capture the constant change that occurs in technology, law and ethics, and MZ generation in today's volatile, uncertain, complex, and ambiguous (VUCA world). Her main research interests include industries, such as the energy, media and entertainment, health care, and ICT industries.

DAEYOUNG KIM received the degrees in industrial engineering from Seoul National University, Seoul, South Korea. He worked at BISTel and BISTelligence as a Senior Data Scientist for seven years. He is currently a Global Product Manager of asset performance management (APM) solutions at AIdentyx. He is working on product design and algorithm research for smart manufacturing solutions. In particular, he is interested in machine learning, deep learning, and other artificial intelligence (AI) algorithms for the manufacturing industry. He mainly focuses on developing APM solutions using deep-learning algorithms for time-series data.

DONGHWAN KIM received the B.S. degree in industrial engineering from Konkuk University, Seoul, South Korea, in 2014, and the Ph.D. degree in industrial engineering from Korea University, Seoul, in 2022. Since 2017, he has been a Senior Data Scientist at BISTel and BISTelligence Inc. His research interests include machine learning, artificial intelligence algorithms for predictive maintenance manufacturing solutions, and an ensemble method for general manufacturing.

JIN-HO KIM received the Ph.D. degree from the Wharton School. He is the author of ten books published in South Korea. He is currently a Program Director of the MBA/DBA in AI Big Data Programs, Swiss School of Management, and the Director of the SSM AI Big Data Research Center. He coauthored *Keeping Up With the Quants: Your Guide to Understanding and Using Analytics* published in USA. He has developed and run programs for building individuals' analytical capabilities. His current research interest includes the application of machine learning methods to address various issues in business and society.