IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 10, 2020, accepted September 27, 2020, date of publication September 30, 2020,
date of current version October 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027830

A Semi-Supervised Autoencoder With an Auxiliary Task (SAAT) for Power Transformer Fault Diagnosis Using Dissolved Gas Analysis

SUNUWE KIM[1], SOO-HO JO[1], WONGON KIM[1], JONGMIN PARK[1], JINGYO JEONG[1,2], YEONGMIN HAN[2], DAEIL KIM[2], AND BYENG DONG YOUN[1,3], (Member, IEEE)

[1]Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul 08826, South Korea
[2]Department of Transmission & Substation Operation, Korea Electric Power Corporation (KEPCO), Naju 58322, South Korea
[3]OnePredict Inc., Seoul 08826, South Korea

Corresponding author: Byeng Dong Youn (bdyoun@snu.ac.kr)

This work was supported in part by the 2017 Open Research and Development Program of Korea Electric Power Corporation (KEPCO) under Grant R17tH02, and in part by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT), Korea Government under Grant 2020R1A2C3003644.

ABSTRACT This paper proposes a semi-supervised autoencoder with an auxiliary task (SAAT) to extract a health feature space for power transformer fault diagnosis using dissolved gas analysis (DGA). The health feature space generated by a semi-supervised autoencoder (SSAE) not only identifies normal and thermal/electrical fault types, but also presents the underlying characteristics of DGA. In the proposed approach, by adding an auxiliary task that detects normal and fault states in the loss function of SSAE, the health feature space additionally enables visualization of health degradation properties. The overall procedure of the new approach includes three key steps: 1) preprocessing DGA data, 2) extracting two health features via SAAT, and 3) visualizing the two health features in two-dimensional space. In this paper, we test the proposed approach using massive unlabeled/labeled Korea Electric Power Corporation (KEPCO) databases and IEC TC 10 databases. To demonstrate the effectiveness of the proposed approach, four comparative studies are conducted with these datasets; the studies examined: 1) the effectiveness of an auxiliary detection task, 2) the effectiveness of the visualization method, 3) conventional fault diagnosis methods, and 4) the state-of-the-art, semi-supervised deep learning algorithms. By examining several evaluation metrics, these comparative studies confirm that the proposed approach outperforms SSAE without the auxiliary task, existing methods, and state-of-the-art deep learning algorithms, in terms of defining health degradation performance. We expect that the proposed SAAT-based health feature space approach will be widely applicable to intuitively monitor the health state of power transformers in the real world.

INDEX TERMS Semi-supervised autoencoder, health feature space, fault diagnosis, power transformer, dissolved gas analysis.

I. INTRODUCTION

Power transformers are important components of distribution and transmission lines of power grid systems. For stable operation of transformers, insulation materials are used to prevent heat transfer and electrical discharge [1]. Although transformers are manufactured to meet reliable design conditions, uncertainties in operation can cause transformers to operate in an unexpected way. Thus, to prevent catastrophic social, economic, and energy efficiency losses, prognostics and health management techniques have attracted attention in recent decades [2]–[4].

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

Dissolved gas analysis (DGA) has been widely used to diagnose oil-filled transformers [5]. When insulation materials are continuously exposed to electrical and thermal stresses, combustible gases (e.g., H_2, C_2H_2, C_2H_4, and so on) are decomposed from the insulation materials and then dissolved in the oil [6]. Via on/offline measurement of these dissolved gases, DGA can diagnose (e.g., detect and identify) the health state of the transformers. In this study, *fault detection* refers to the binary classification of normal and fault states, *fault identification* indicates multi-classification of normal and electrical/thermal fault types.

Fault diagnosis methods using DGA are divided into two categories: rule-based methods and artificial intelligence (AI)-based methods. In rule-based methods, concentrations

and/or ratios of gases are used for fault identification based on human-experienced thresholds. Examples of rule-based methods include the IEC ratios method [7], the Rogers ratios method [8], and the Doernenburg ratios method [9]. In addition, Duval ratio methods provide two-dimensional (2D) graphics (e.g., Duval triangle and pentagon) which are intuitive for classifying fault types [6], [10], [11]. However, rule-based methods have relatively low accuracy and inconsistent diagnosis results due to insufficient mathematical computation and their empirical handcrafted thresholds [12].

In recent years, AI-techniques have been incorporated in power transformer fault diagnosis to improve accuracy. AI techniques include fuzzy logic [13]–[15], support vector machine [16], [17], artificial neural network, and multilayer perceptron [18]–[21]. To select optimal features and address imbalanced problems of DGA data, a genetic algorithm approach [22]–[25] and an adaptive over-sampling method [26], [27] have been applied, respectively. Despite some achievements using such supervised learning approaches, these studies take only labeled DGA datasets into account. In other prior work, a semi-supervised learning approach using a low-dimensional scaling was developed to consider unlabeled DGA data [28]. However, this approach has difficulty performing health feature selection for unlabeled datasets. Motivated by this challenge, several additional methods for extracting health features have been reported. A principal component analysis with fuzzy C-means method was presented as an unsupervised feature extraction method in [29], [30]. Besides, self-organizing maps (SOM) of unsupervised neural network methods extracted feature maps of several fault types [31], [32]. Further, deep learning techniques, such as by sparse autoencoder [33] and deep belief network [34], have been used to extract high-level health features by unsupervised greedy layer wise training with deep hierarchical hidden layers.

While these advances have been significant, AI-based approaches have the following three limitations. First, despite the necessity of a large amount of DGA data to represent generalized diagnosis results, it is difficult to obtain the large amount of required DGA data in real-world applications. Significant financial cost is required to periodically maintain all transformers and measure DGA data in the field. Second, most previous studies have focused on fault detection and identification features; little effort has been made to analyze the health degradation features. If degradation features are newly developed, it is worth pointing out that they enable to exhibit the monotonic health state transition from normal to fault, thus potentially estimating health states for unlabeled data or diagnosing fault states in advance. Lastly, visualization of the monotonic health state transition in 2D space has ye be addressed by other research. Since 2D graphics provide the most obvious and readable space representation for the human eye, a 2D health feature space (HFS) can intuitively show diagnosis results [35].

Thus, in this paper, we propose a novel semi-supervised autoencoder with an auxiliary task (SAAT) to extract an HSF, considering a large amount of DGA data. The proposed SAAT approach comes from a semi-supervised autoencoder (SSAE) that can simultaneously learn unsupervised and supervised tasks with shared hidden layers. Unsupervised and supervised tasks play roles in the representative health feature extraction and the fault identification, respectively. Here, by putting an auxiliary task (fault detection) in the loss function of SSAE, the trained shared parameters provide the health features, which additionally enable representation of the health degradation properties. By structuring the two nodes in the end of the shared hidden layers, two health features can be directly visualized into 2D space without an additional dimension reduction. In this paper, a large amount of DGA data, provided by Korea Electric Power Corporation (KEPCO), is considered. In addition, IEC TC 10 databases are used for validation tests. To the best of the authors' knowledge, the contributions of this work can be summarized as follows:

1. *This is the first attempt to diagnose real-world power transformers using a large amount of DGA data.*
2. *The proposed SAAT has the ability to represent health degradation properties as well as to identify normal and thermal/electrical fault types.*
3. *By directly visualizing health features without transformation or dimension reduction, the proposed 2D HFS can pictorially demonstrate the monotonic health state transition of transformers.*

The rest of paper is organized as follows. Section II describes the background of SAAT. Sections III and IV demonstrate the proposed method and experimental results, respectively. Finally, the conclusions and future works of this study are outlined in Section V.

II. BACKGROUND OF A SEMI-SUPERVISED AUTOENCODER WITH AN AUXILIARY TASK

Two basic algorithms (i.e., an autoencoder (AE) and a softmax classifier (SC)) of the proposed SAAT are described in Sections II.A and II.B, respectively. In Section II.C, SSAE is explained in terms of the AE and the SC.

A. AUTOENCODER: UNSUPERVISED FEATURE EXTRACTION

An AE, a well-known unsupervised neural network, consists of an encoder part and a decoder part with a hidden layer, as shown in Fig. 1 (a) [36]–[39]. For given training samples $\mathbf{x} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)}\}$ where N is the number of samples and $\mathbf{x}^{(m)} \in \mathbb{R}^d$ $(m = 1, 2, \cdots, N)$, an encoder function f^{en} compresses the dimension of the training samples from $\mathbb{R}^d$ to $\mathbb{R}^{d'}$ $(d > d')$ with a set of encoder parameters $\boldsymbol{\theta}^{\text{en}}$ (i.e., a weight matrix $\mathbf{W}^{\text{en}} \in \mathbb{R}^{d' \times d}$ and a bias vector $\mathbf{b}^{\text{en}} \in \mathbb{R}^{d'}$), as:

$$f^{\text{en}}\left(x_i^{(m)}\right) = h_j^{(m)} = \sigma^{\text{AE}}\left(W_{ji}^{\text{en}} x_i^{(m)} + b_j^{\text{en}}\right) \tag{1}$$

where σ^{AE} is an activation function, such as a sigmoid, a rectified linear unit (ReLU), and an exponential linear unit (ELU) that transforms $\mathbf{x}^{(m)}$ into a representative feature vector $\mathbf{h}^{(m)} \in \mathbb{R}^{d'}$ with $\boldsymbol{\theta}^{\text{en}}$. Then, in the decoder part, $\mathbf{h}^{(m)}$

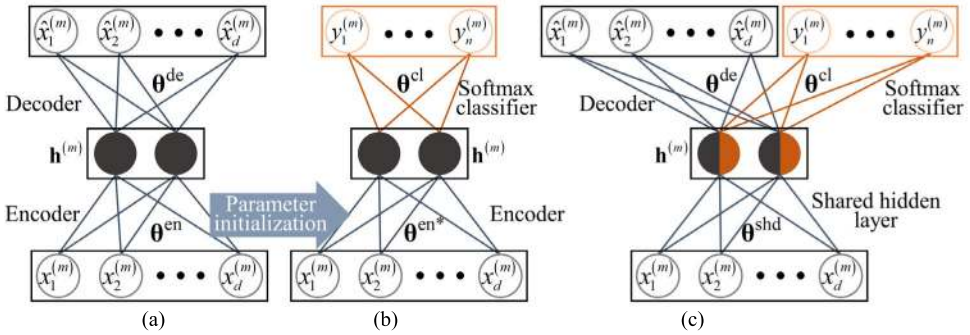

FIGURE 1. Architectures of AE, SC, and SSAE: (a) pre-training in the AE; (b) fine-tuning in the SC with initialized parameters; and (c) simultaneous learning of the supervised and unsupervised learning parts in SSAE.

is reconstructed to $\hat{\mathbf{x}}^{(m)} \in \mathbb{R}^d$ by a decoder function f^{de}, with a set of decoder parameters $\boldsymbol{\theta}^{\text{de}}$ (i.e., a weight matrix $\mathbf{W}^{\text{de}} \in \mathbb{R}^{d \times d'}$, and a bias vector $\mathbf{b}^{\text{de}} \in \mathbb{R}^d$) as:

$$f^{\text{de}}\left(h_j^{(m)}\right) = \hat{x}_k^{(m)} = \sigma^{\text{AE}}\left(W_{kj}^{\text{de}} h_j^{(m)} + b_k^{\text{de}}\right) \tag{2}$$

where σ^{AE} transforms $\mathbf{h}^{(m)}$ into $\hat{\mathbf{x}}^{(m)}$.

In general, the loss function L_{AE} is the mean square error between $\mathbf{x}^{(m)}$ and $\hat{\mathbf{x}}^{(m)}$ as:

$$\begin{aligned} L_{\text{AE}}\left(\boldsymbol{\theta}^{\text{en}}, \boldsymbol{\theta}^{\text{de}}\right) &= \frac{1}{2N} \sum_{m=1}^{N} \left\| \hat{\mathbf{x}}^{(m)} - \mathbf{x}^{(m)} \right\|^2 \\ &= \frac{1}{2N} \sum_{m=1}^{N} L_{\text{AE}}^{(m)} \end{aligned} \tag{3}$$

where $L_{\text{AE}}^{(m)}$ represents the m-th loss function. To minimize L_{AE}, the parameters $\boldsymbol{\theta}^{\text{AE}} = \{\boldsymbol{\theta}^{\text{en}}, \boldsymbol{\theta}^{\text{de}}\}$ are updated using a backpropagation method with mini-batch gradient descent algorithms. Using chain rules, the procedure of the parameter update is organized as:

$$\theta_{kj}^{\text{de}} \leftarrow \theta_{kj}^{\text{de}} - \eta \frac{\partial L_{\text{AE}}^{(m)}}{\partial \theta_{kj}^{\text{de}}} \left(\frac{\partial L_{\text{AE}}^{(m)}}{\partial \theta_{kj}^{\text{de}}} = \delta_k^{\text{de}} \frac{\partial z_k^{(m)}}{\partial \theta_{kj}^{\text{de}}} = \delta_k^{\text{de}} h_j^{(m)} \right) \tag{4}$$

$$\theta_{ji}^{\text{en}} \leftarrow \theta_{ji}^{\text{en}} - \eta \frac{\partial L_{\text{AE}}^{(m)}}{\partial \theta_{ji}^{\text{en}}} \left(\frac{\partial L_{\text{AE}}^{(m)}}{\partial \theta_{ji}^{\text{en}}} = \delta_j^{\text{en}} \frac{\partial z_j^{(m)}}{\partial \theta_{ji}^{\text{en}}} = \delta_j^{\text{en}} x_i^{(m)} \right) \tag{5}$$

where η is a learning rate; $z_k^{(m)}$, δ_k^{de}, $z_j^{(m)}$, and δ_j^{en} are defined, respectively, as:

$$z_k^{(m)} = W_{kj}^{\text{de}} h_j^{(m)} + b_k^{\text{de}} \tag{6}$$

$$\delta_k^{\text{de}} \equiv \frac{\partial L^{(m)}}{\partial z_k^{(m)}} = \sigma^{\text{AE}'}\left(z_k^{(m)}\right) \frac{\partial L^{(m)}}{\partial x_k^{(m)}} \tag{7}$$

$$z_j^{(m)} = W_{ji}^{\text{en}} x_i^{(m)} + b_j^{\text{en}} \tag{8}$$

$$\delta_j^{\text{en}} \equiv \frac{\partial L_{\text{AE}}^{(m)}}{\partial z_j^{(m)}} = \sum_k \frac{\partial L_{\text{AE}}^{(m)}}{\partial z_k^{(m)}} \frac{\partial z_k^{(m)}}{\partial z_j^{(m)}} = \sigma^{\text{AE}'}\left(z_j^{(m)}\right) \sum_k \theta_{kj}^{\text{de}} \delta_k \tag{9}$$

δ_k^{de} and δ_j^{en} are errors in the decoder layer and the encoder layer, respectively. This process is called pre-training. Using the optimized $\boldsymbol{\theta}^{\text{AE}}$ derived through (4) to (9), AE can extract $\mathbf{h}^{(m)}$. Please note that the number of hidden layers in the encoder and the decoder can be extended.

B. SOFTMAX CLASSIFIER: SUPERVISED CLASSIFICATION

SC has been widely used for the purpose of classifying multi-classes by utilizing the extracted high-level features in AI-based algorithms [33], [34], [38]. When incorporating the SC into the AE, $\mathbf{h}^{(m)}$ can be the input data of a softmax function, as shown in Fig. 1 (b). Training samples are a set of ordered pairs $(\mathbf{x}^{(m)}, \mathbf{y}^{(m)})$ as $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(3)}), \cdots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$ where $\mathbf{y}^{(m)} \in \{1, 2, \cdots, \text{C}\}$ is a virtual discrete number of a target label that corresponds to $\mathbf{x}^{(m)}$. $\mathbf{y}^{(m)}$ is a one-hot encoding vector that has C classes, expressed as $\mathbf{y}^{(m)} = (y_1^{(m)}, y_2^{(m)}, \cdots, y_C^{(m)})$. Using the softmax function q, the probability of each element in $\mathbf{y}^{(m)}$ can be calculated with respect to $\boldsymbol{\theta}^{\text{en}*}$ and $\boldsymbol{\theta}^{\text{cl}}$ (i.e., a weight matrix $\mathbf{W}^{\text{cl}} \in \mathbb{R}^{C \times d'}$, and a bias vector $\mathbf{b}^{\text{cl}} \in \mathbb{R}^C$), as follows:

$$\begin{aligned} \hat{y}_n^{(m)} &= P\left(\mathbf{y}^{(m)} = n | f^{\text{en}}\left(\mathbf{x}^{(m)}\right); \boldsymbol{\theta}^{\text{en}*}, \boldsymbol{\theta}^{\text{cl}}\right) \\ &= q\left(z_n^{(m)}\right) = \frac{\exp\left(z_n^{(m)}\right)}{\sum_{n=1}^{C} \exp\left(z_n^{(m)}\right)} \end{aligned} \tag{10}$$

where $z_n^{(m)}$ is defined as

$$z_n^{(m)} = W_{nj}^{\text{cl}} h_j^{(m)} + b_n^{\text{cl}} \tag{11}$$

Note that n means the n-th element in $\mathbf{y}^{(m)}$, as well as the number n in $\{1, 2, \cdots, \text{C}\}$. $\hat{y}_n^{(m)}$ should satisfy $\hat{y}_n^{(m)} \in [0, 1]$ and $\sum_{n=1}^{C} \hat{y}_n^{(m)} = 1$.

For the best classification performance, it is worth noting that finding optimized parameters $\boldsymbol{\theta}^{\text{en}*}$ and $\boldsymbol{\theta}^{\text{cl}}$ is an essential procedure to match $\hat{\mathbf{y}}^{(m)}$ with $\mathbf{y}^{(m)}$. To minimize the discrepancy between $\mathbf{y}^{(m)}$ and $\hat{\mathbf{y}}^{(m)}$, the cross-entropy loss function L_{cl} has been widely used as [2]:

$$L_{\text{cl}}\left(\boldsymbol{\theta}^{\text{en}^*}, \boldsymbol{\theta}^{\text{class}}\right) = -\frac{1}{N} \sum_{m=1}^{N} \mathbf{y}^{(m)} \log\left(\hat{\mathbf{y}}^{(m)}\right) \tag{12}$$

Likewise, $\boldsymbol{\theta}^{\mathrm{en}*}$ and $\boldsymbol{\theta}^{\mathrm{cl}}$ are updated by mini-batch gradient descent algorithms as:

$$\theta_{nj}^{\mathrm{cl}} \leftarrow \theta_{nj}^{\mathrm{cl}} - \eta \frac{\partial L_{\mathrm{cl}}^{(m)}}{\partial \theta_{nj}^{\mathrm{cl}}} \left(\frac{\partial L_{\mathrm{cl}}^{(m)}}{\partial \theta_{nj}^{\mathrm{cl}}} = \delta_n^{\mathrm{cl}} \frac{\partial z_n^{(m)}}{\partial \theta_{nj}^{\mathrm{cl}}} = \delta_n^{\mathrm{cl}} h_j^{(m)} \right) \tag{13}$$

$$\theta_{ji}^{\mathrm{en}^*} \leftarrow \theta_{ji}^{\mathrm{en}^*} - \eta \frac{\partial L_{\mathrm{cl}}^{(m)}}{\partial \theta_{ji}^{\mathrm{en}^*}} \left(\frac{\partial L_{\mathrm{cl}}^{(m)}}{\partial \theta_{ji}^{\mathrm{en}^*}} = \delta_j^{\mathrm{en}^*} \frac{\partial z_j^{(m)}}{\partial \theta_{ji}^{\mathrm{en}^*}} = \delta_j^{\mathrm{en}^*} x_i^{(m)} \right) \tag{14}$$

where $z_n^{(m)}$, δ_n^{cl}, and $\delta_j^{\mathrm{en}^*}$ are defined, respectively, as:

$$z_n^{(m)} = W_{nj}^{\mathrm{cl}} h_j^{(m)} + b_n^{\mathrm{cl}} \tag{15}$$

$$\delta_n^{\mathrm{cl}} \equiv \frac{\partial L_{\mathrm{cl}}^{(m)}}{\partial z_n^{(m)}} = \sigma^{\mathrm{cl}'} \left(z_n^{(m)} \right) \frac{\partial L_{\mathrm{cl}}^{(m)}}{\partial \widehat{y}_k^{(m)}} \tag{16}$$

$$\delta_j^{\mathrm{en}^*} \equiv \frac{\partial L_{\mathrm{cl}}^{(m)}}{\partial z_j^{(m)}} = \sum_n \frac{\partial L_{\mathrm{cl}}^{(m)}}{\partial z_n^{(m)}} \frac{\partial z_n^{(m)}}{\partial z_j^{(m)}} = \sigma^{\mathrm{cl}'} \left(z_j^{(m)} \right) \sum_n \theta_{nj}^{\mathrm{cl}} \delta_n^{\mathrm{cl}} \tag{17}$$

This process is called fine-tuning. Using the feature extraction developed through the pre-training in the AE, the classification accuracy can be dramatically enhanced, as compared with SC in the absence of AE.

C. SEMI-SUPERVISED AUTOENCODER

Disjoint learning between the pre-training and the fine-tuning – by sequentially performing AE and SC – can lead to the extraction of features that are uncorrelated with the target information of the labeled data or to distortion of the underlying characteristics of the input training samples [40]. With this motivation, SSAE has been proposed, as shown in Fig. 1 (c). Compared with the previous sequentially executed training process, SSAE achieves extraction of high-level features that are highly correlated with both the input data $\mathbf{x}$ and the labeled information $\mathbf{y}$, by simultaneously optimizing $\boldsymbol{\theta}^{\mathrm{AE}}$ and $\boldsymbol{\theta}^{\mathrm{cl}}$ [35], [40]–[43].

A loss function L_{SSAE} of SSAE is a summation of the two loss functions presented in (3) and (12) with a weight α as:

$$L_{\mathrm{SSAE}} \left(\boldsymbol{\theta}^{\mathrm{shd}}, \boldsymbol{\theta}^{\mathrm{de}}, \boldsymbol{\theta}^{\mathrm{cl}} \right) = \alpha L_{\mathrm{AE}} \left(\boldsymbol{\theta}^{\mathrm{shd}}, \boldsymbol{\theta}^{\mathrm{de}} \right) + (1-\alpha) L_{\mathrm{cl}} \left(\boldsymbol{\theta}^{\mathrm{shd}}, \boldsymbol{\theta}^{\mathrm{cl}} \right) \tag{18}$$

where the shared parameters $\boldsymbol{\theta}^{\mathrm{shd}}$, which play the same role as $\boldsymbol{\theta}^{\mathrm{en}}$ in AE, are simultaneously optimized when training the representative feature extraction task of AE and the classification task of SC. For example, the procedure to update the parameters to minimize L_{SSAE} is demonstrated as:

$$\theta_{ji}^{\mathrm{shd}} \leftarrow \theta_{ji}^{\mathrm{shd}} - \eta \frac{\partial L_{\mathrm{SSAE}}^{(m)}}{\theta_{ji}^{\mathrm{shd}}} \times \left(\frac{\partial L_{\mathrm{SSAE}}^{(m)}}{\theta_{ji}^{\mathrm{shd}}} = \alpha \delta_j^{\mathrm{AE}} x_i^{(m)} + (1-\alpha) \delta_j^{\mathrm{cl}} x_i^{(m)} \right) \tag{19}$$

where δ_j^{AE} and δ_j^{cl} are equal to (9) and (16), respectively. Finally, the shared hidden layers with $\boldsymbol{\theta}^{\mathrm{shd}}$ are able to concurrently extract representative features of $\mathbf{x}$ in the unsupervised learning and the labeled information of $\mathbf{y}$ in the supervised learning. For power transformer fault diagnosis, it can be inferred that SSAE enables identification of the thermal/electrical fault types and normal state, as well as extraction of high-level features with a large amount of real-world DGA data.

III. PROPOSED METHOD

This section demonstrates the proposed SAAT method. Section III.A presents the input DGA data preprocessing approach. Section III.B describes SAAT-based fault diagnosis method, including the role of the auxiliary detection task, the architecture of SAAT, and HFS visualization. The overall procedure is demonstrated in Fig. 2.

A. INPUT DGA DATA PREPROCESSING

In the field of AI, normalizing raw input data and balancing imbalanced data are essential steps to avoid overfitting problems and to enable better classification performance [28]. Furthermore, from the viewpoint of power transformer fault diagnosis, handcrafted features of dissolved gas ratios, which were previously studied in rule-based methods, have been incorporated into AI-based methods to enhance the diagnosis performance [28]. Details of each preprocessing step are described as follows.

1) SCALING OF INDUSTRIAL DGA DATA

Dissolved gas concentrations have significantly skewed distributions because their concentrations tend to dramatically increase in a fault state, as compared with those in a normal state. For example, the gas concentrations changed from a few ppm (parts per million) to thousands of ppm in previous studies [28]. Thus, the input DGA data is transformed into a logarithmic scale. Further, to keep numerical operations (e.g., stochastic gradient descent) stable, the logarithmic-scaled DGA data is normalized from zero (min) to one (max).

2) BALANCING OF IMBALANCED INDUSTRIAL DGA DATA

Since real-world industrial transformers have highly imbalanced data between normal and fault states, this imbalance could disturb AI-based methods [28]. For example, if fault datasets occupy only 1 % among the training datasets, most AI-based algorithms will be more focused on the classification of major normal datasets. Thus, an accuracy of 99 % would be obtained by ignoring the minor – but critical – fault datasets and classifying all datasets as normal. To address these imbalance problems, oversampling techniques are applied into the fault datasets [28].

3) COMBINING ADDITIONAL FEATURES RELATED TO GAS RATIOS

We consider six combustible gases (i.e., H_2, C_2H_2, C_2H_4, C_2H_6, CH_4, and CO). Each of the combustible gases is denoted as DGA_i where i ranges from one to six. Normalized DGA_i in the logarithmic scale is expressed as minimax(log([DGA_i])). In rule-based methods, it is well

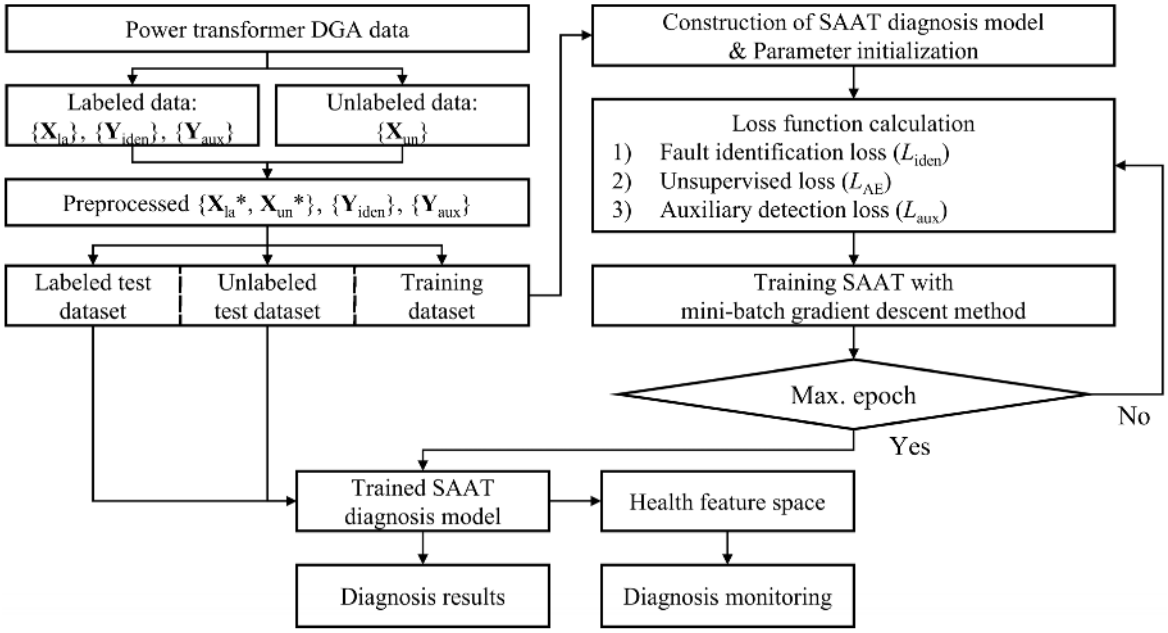

FIGURE 2. Overall procedures of the proposed SAAT-based fault diagnosis method.

known that the absolute values of gas concentrations can be useful for the fault detection; however, it is desirable to investigate the ratio-like relationships between the gas concentrations for fault identification [28]. Therefore, we consider six ratios of gas concentration DGA_i to total gas concentration $\sum_i \mathrm{DGA}_i$ in the logarithmic scale, as $\log([\mathrm{DGA}_i]/[\sum_i \mathrm{DGA}_i])$. Further, three ratios, developed by Duval triangle methods, are considered; these features are widely used in diagnosing transformer fault types [44], [45]. The total preprocessed input data lies in 15 dimensions.

B. SAAT-BASED FAULT DIAGNOSIS METHOD

The main concern of rule-based approaches is to monitor fault types. Since they do not take the normal state into account, it is difficult to visualize the overall health degradation properties. Further, in AI-based approaches, only a few prior studies have been devoted to investigating health degradation features. Since trends of measured dissolved gases present nonlinear properties over time while the health state is monotonically degraded, it is desirable to extract new health features that could also represent the monotonic health state transition from normal to fault.

Moreover, as it requires a tremendous cost to perform thorough visual inspection to recognize incipient faults every time, most DGA data in industrial fields is unlabeled. Since sparse, fault-labeled data results in limitations in the ability to confirm reliable quantitative results, additional qualitative methods have been developed, such as high-level feature visualization in 2D space using unsupervised dimension reduction algorithms (e.g., t-stochastic neighbor embedding (t-SNE) and self-organizing map (SOM)) [2], [31]. However, it is worth noting that some key information associated with fault diagnosis can be lost during the dimension reduction procedure. Moreover, since both t-SNE and SOM have the ability to cluster the neighboring data, the correlation between high-level features cannot be guaranteed [31], [32], [46].

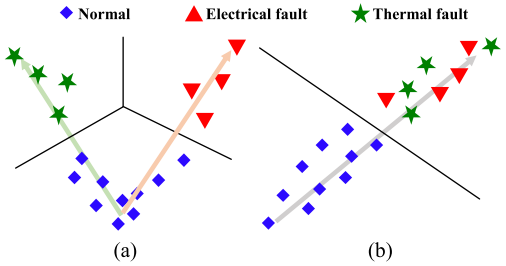

FIGURE 3. Conceptual diagrams of the health feature space: (a) fault identification task case in SSAE and (b) fault detection task case in SSAE.

Thus, we propose a SAAT that an auxiliary detection task, which is inserted into the loss function of SSAE, that can achieve health degradation feature extraction. Further, SAAT-based fault diagnosis model can directly visualize the two high-level features in 2D, called the HFS, without additional dimension reduction, while representing not only the fault identification but also the health degradation properties. Details are described as follows.

1) ROLES OF THE AUXILIARY DETECTION TASK

Since the fault identification task in the supervised learning part of SSAE recognizes the three classes as independent classes, it is not aware of whether both classes of electrical/thermal fault types are involved in fault states. Thus, the only identification task can lose the underlying characteristics of the fault detection. For example, when envisaging a 2D feature space, there could be two independent directions that represent the health state transition, as shown in Fig. 3 (a); this is against the physical phenomenon of monotonic health degradation. Here, it is important to note that the fault detection task has the potential to present the

monotonic state transition in a single direction, as shown in Fig. 3 (b). An auxiliary detection task, which can tie the two classes of electrical/thermal fault states into one fault state, is thus newly added. The proposed SAAT method has three tasks: 1) unsupervised learning to represent the input data characteristics, 2) supervised learning for fault identification, and 3) supervised learning for auxiliary detection.

The parameters $\boldsymbol{\theta}^{\mathrm{SAAT}}$ of the proposed SAAT are as:

$$\boldsymbol{\theta}^{\mathrm{SAAT}} = \left\{\boldsymbol{\theta}^{\mathrm{shd},p}, \boldsymbol{\theta}^{\mathrm{iden}}, \boldsymbol{\theta}^{\mathrm{de},q}, \boldsymbol{\theta}^{\mathrm{aux}}\right\} \tag{20}$$

where $\boldsymbol{\theta}^{\mathrm{shd},p}$, θ^{iden}, $\boldsymbol{\theta}^{\mathrm{de},q}$, and $\boldsymbol{\theta}^{\mathrm{aux}}$ are shared parameters, identification parameters, decoder parameters, and auxiliary detection parameters, respectively. Superscripts p and q stand for the p-th and q-th hidden layers in the shared network and the decoder, respectively. When training the tasks, the back-propagation method is used to optimize the parameters. In this study, this method transmits errors between key information (e.g., labeled information of electrical/thermal fault types and normal state for the identification task) and the output layer in each task, backward to each layer in the shared network. Training each task is simultaneously executed with by optimizing $\theta^{\mathrm{shd},p}$. Hence, $\theta^{\mathrm{shd},p}$ would possess all information of output layers, $\boldsymbol{\theta}^{\mathrm{iden}}$, $\theta^{\mathrm{de},q}$, and θ^{aux}.

A loss function L_{SAAT} of the proposed SAAT is defined as:

$$\begin{aligned} L_{\mathrm{SAAT}}\left(\boldsymbol{\theta}^{\mathrm{SAAT}}\right) &= \beta L_{\mathrm{SSAE}}\left(\boldsymbol{\theta}^{\mathrm{shd},p}, \boldsymbol{\theta}^{\mathrm{iden}}, \boldsymbol{\theta}^{\mathrm{de},q}\right) \\ &\quad + (1-\beta)\, L_{\mathrm{aux}}\left(\boldsymbol{\theta}^{\mathrm{shd},p}, \boldsymbol{\theta}^{\mathrm{aux}}\right) \\ &\quad + 0.5\lambda \left\|\boldsymbol{\theta}^{\mathrm{SAAT}}\right\|^2 \end{aligned} \tag{21}$$

where L_{SSAE} is similar to (18); the differences are that the number of layers are much more in (21) and $\boldsymbol{\theta}^{\mathrm{cl}}$ in (18) is changed to $\boldsymbol{\theta}^{\mathrm{iden}}$. The loss function L_{aux} of the auxiliary detection task is newly proposed in (21). A hyperparameter β is the weight between L_{SSAE} and L_{aux}. In addition, to avoid overfitting problems, a L2 regularization term $0.5\lambda||\boldsymbol{\theta}^{\mathrm{SAAT}}||^2$ is put in (21) with a hyperparameter λ [47]–[49].

SAAT can be trained by updating $\boldsymbol{\theta}^{\mathrm{SAAT}}$ to minimize L_{SAAT}. For example, in the case of $\boldsymbol{\theta}^{\mathrm{shd,end}}$ that are parameters in the end of the shared hidden layers and directly related to health feature extraction, the procedure of updating the parameters is demonstrated as:

$$\theta_{ji}^{\mathrm{shd,end}} \leftarrow \theta_{ji}^{\mathrm{shd,end}} - \eta \frac{\partial L_{\mathrm{SAAT}}^{(m)}}{\partial \theta_{ji}^{\mathrm{shd,end}}} \tag{22}$$

Similar to (19), the second term in the right-hand side of (22) can be decomposed as:

$$\begin{aligned} \frac{\partial L_{\mathrm{SAAT}}^{(m)}}{\partial \theta_{ji}^{\mathrm{shd,end}}} &= \beta \delta_j^{\mathrm{SSAE,end}} h_i^{\mathrm{shd,end}} + (1-\beta)\, \delta_j^{\mathrm{aux}} h_i^{\mathrm{shd,end}} \\ &\quad + \lambda \theta_{ji}^{\mathrm{shd,end}} \end{aligned} \tag{23}$$

where $h_i^{\mathrm{shd,end}}$ are high-level features obtained at the end of the shared hidden layers. Here, $\delta_j^{\mathrm{SSAE,end}}$ and δ_j^{aux} are expressed, respectively, as:

$$\begin{aligned} \delta_j^{\mathrm{SSAE,end}} &= \alpha \times \sigma^{\mathrm{de}'}\left(z_j\right) \sum_k \theta_{kj}^{\mathrm{de},1} \delta_k^{\mathrm{de},1} \\ &\quad + (1-\alpha) \times \sigma^{\mathrm{iden}'}\left(z_j\right) \sum_{k'} \theta_{k'j}^{\mathrm{iden}} \delta_{k'}^{\mathrm{iden}} \end{aligned} \tag{24}$$

$$\delta_j^{\mathrm{aux}} = \sigma^{\mathrm{aux}'}\left(z_j\right) \sum_{k''} \theta_{k''j}^{\mathrm{aux}} \delta_{k''}^{\mathrm{aux}} \tag{25}$$

where k, k', and k'' are dimensions of output nodes in the first layer of the decoder, fault identification, and auxiliary detection tasks, respectively. By inserting (24) and (25) into (23), $\boldsymbol{\theta}^{\mathrm{shd,end}}$ are updated as (22). Finally, high-level features obtained by the proposed SAAT could play roles in exhibiting both fault identification and health degradation.

2) ARCHITECTURE OF THE PROPOSED SAAT

As shown in Fig. 4, the proposed SAAT consists of three shared hidden layers, three decoder hidden layers, and one hidden layer for each supervised task. Activation functions of all hidden layers, except for the supervised tasks, are ELUs; this function has the advantages of not only increasing computational learning speeds in deep neural networks [50]–[52] but also achieving robust optimization in backpropagation methods. Activation functions of the output hidden layers in cases of fault identification and auxiliary detection tasks are the SC and the logistic regression for binary classification, respectively. Detailed parameters in SAAT architecture are summarized in Table I. Both the number of epochs and batch size are set as 200. α, β, λ, and η are set as 0.25, 0.4, 0.0001, and 0.001, respectively.

Note that we consider a compressed-type structure in the shared hidden layers. For the purpose of extracting only two high-level health features $\mathbf{h}^{\mathrm{HF}} \in \mathbb{R}^2$ that could be directly visualized in the 2D space, the end of the shared hidden layer is set as having two nodes. These two nodes are connected with the three tasks.

3) HEALTH FEATURE SPACE VISUALIZATION

Fig. 5 depicts interpretation schemes for HFS. HFS is directly visualized into 2D space (x-y plane); the features are denoted as 'Health Feature 1 (HF1)' and 'Health Feature 2 (HF2)', respectively. x- and y-axes correspond to HF1 and HF2, respectively. Here, to show the degree of health degradation, the extracted health features are arranged to increase over time.

It is expected that $\mathbf{h}^{\mathrm{HF}}$ for the training/test datasets can be visualized with a set of four dots, as shown in Fig. 5. Further, from the fault identification task, the identification decision boundaries can be obtained and visualized. It is important to emphasize that the decision boundaries in 2D HFS have the following merits: 1) health states or fault types can be determined for the labeled data and 2) the classes for the unlabeled data can be predicted (pseudo-labeled) by investigating to which health state region the unlabeled data belongs.

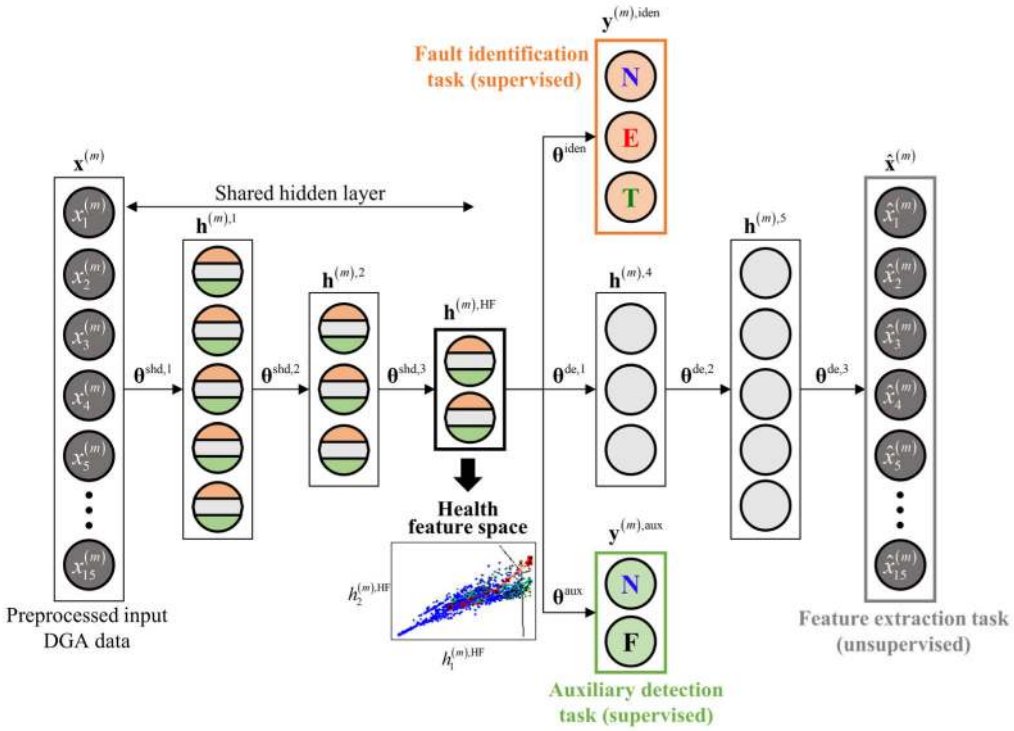

FIGURE 4. Architecture of the proposed SAAT: colors with orange, gray, and green in the shared hidden layers stand for the features related to the fault identification, representative characteristics of DGA data, and health degradation (or health state transition).

TABLE 1. Parameters in the architecture of the proposed SAAT.

Layer	Activation	Node #	Parameter #
Input	-	15	-
Shared layer 1	ELU	10	160
Shared layer 2	ELU	6	66
Shared layer 3	ELU	2	14
Decoder1	ELU	6	18
Decoder2	ELU	10	70
Output1 (Representative feature extraction task)	ELU	15	165
Output2 (Fault identification task)	Softmax	3	9
Output3 (Auxiliary detection task)	Sigmoid	1	3

Moreover, thanks to the auxiliary detection task, the monotonic health state transition from normal to fault will be observed in 2D HFS. In real-world applications, normal transformers gradually degrade as time passes. Then, one of the thermal/electrical fault types will occur at a certain point. From this physical interpretation, the monotonic trend of the two health features in 2D HFS can be shown up to a certain point; it tends to be slightly separated into one of two ways toward the thermal or electrical fault regions, which are divided by the decision boundaries. Therefore, it is worth noting that the proposed 2D HFS also enables intuitive visualization of the historical health degradation information in terms of 1) the monotonicity between the health features and 2) the monotonic health state transition.

4) OVERALL PROCEDURES OF THE PROPOSED SAAT-BASED FAULT DIAGNOSIS METHOD

Fig. 2 illustrates the flowchart of the proposed SAAT-based fault diagnosis method. The first step is to organize the collected DGA data into four groups: an unlabeled DGA dataset $\{\mathbf{X}_{\text{un}}\}$, a labeled DGA dataset $\{\mathbf{X}_{\text{la}}\}$, and labeled information datasets $\{\mathbf{Y}_{\text{iden}}\}$ and $\{\mathbf{Y}_{\text{aux}}\}$ for the supervised tasks. After preprocessing, the input DGA datasets are denoted as $\{\mathbf{X}_{\text{un}}*\}$ and $\{\mathbf{X}_{\text{la}}*\}$. To train SAAT model and evaluate its performance, datasets, $\{\mathbf{X}_{\text{un}}*\}$, $\{\mathbf{X}_{\text{la}}*\}$, $\{\mathbf{Y}_{\text{iden}}\}$ and $\{\mathbf{Y}_{\text{aux}}\}$ are randomly separated into training datasets and test datasets.

The next step is to construct and stabilize SAAT-based fault diagnosis model using the training datasets. Parameters in SAAT are randomly initialized. For given parameters,

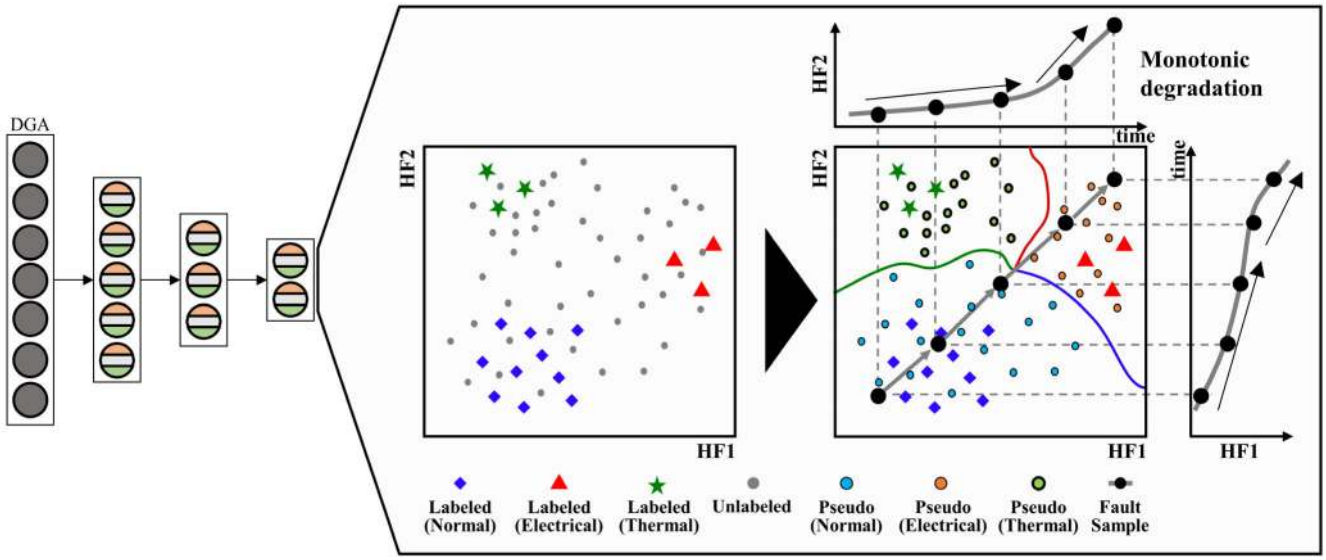

FIGURE 5. Visualization scheme of HFS with labeled and unlabeled data.

L_{iden}, L_{AE}, and L_{aux} are calculated. With the given batch size, the backpropagation method in the mini-batch gradient descent method can train SAAT model by repetitively updating parameters. In addition, loss function calculation and parameter updates are iteratively implemented until satisfying the given maximum epoch.

After completing the training process, the health states of the unlabeled test datasets are pseudo-labeled by the decision boundaries obtained in the fault identification task. Furthermore, several evaluation metrics are calculated as diagnosis results for the labeled and pseudo-labeled test datasets. Finally, by directly visualizing $\mathbf{h}^{\text{HF}}$ in 2D space, the diagnosis results can be pictorially monitored.

IV. PERFORMANCE EVALUATION OF THE PROPOSED METHOD

This section is devoted to performance evaluation of the proposed SAAT method. Section IV.A presents a description of datasets provided by KEPCO and implementation of the proposed method. In Section IV.B, the experimental setup is demonstrated. Lastly, the experimental results and discussion are covered in Section IV.C.

A. DATA DESCRIPTION AND IMPLEMENTATION

DGA data provided by KEPCO was obtained for three decades, from 1980 to 2018. KEPCO measured nine gases (i.e., H_2, C_2H_2, C_2H_4, C_2H_6, CH_4, C_3H_8, CO, CO_2 and N_2). Among them, six combustible gases (H_2, C_2H_2, C_2H_4, C_2H_6, CH_4 and CO) were studied in this work. Note that these gases are also included in the IEC TC 10 database, which is one representative open data set of dissolved gases [53]. KEPCO's DGA data can be divided into unlabeled data and labeled data. Health states of the transformers are defined from Table II, which summarizes the human-experienced thresholds of gases used in KEPCO. The company labeled the transformers as normal when the concentrations of all gases were less than corresponding thresholds. On the other hand, electrical/thermal fault types were labeled after visual inspection when actual failures occurred.

TABLE 2. KEPCO maintenance standards for power transformers.

Gas \ Cond.	Normal	Caution		Abnormal	Danger (>ppm)
		I	II		
H_2	<200	201~400	400~800	>800	-
C_2H_2	<10	11~20	21~60	61~120	>120
C_2H_4	<100	100~200	201~500	>500	-
C_2H_6	<150	151~250	251~750	>750	-
CH_4	<200	201~350	351~750	>750	-
CO	<800	801~1200	1200	-	-

We obtain 110,000 normal data, categorized into 73 thermal fault data, and 48 electrical fault data as similar to IEC TC 10 fault types. As an example, historical DGA data for four samples of KEPCO is listed in Table III. Next, unlabeled data was obtained from cases where some gas concentrations were over the threshold values but visual inspection was not executed. The number of unlabeled data is 24,405. Note that the amount of DGA data used in this study is much larger than that used in previous studies (e.g., 4,642 DGA dataset in [34] and 3,000 DGA dataset in [54]). To validate the effectiveness of the proposed SAAT, two test datasets are examined: 1) 20% of KEPCO datasets and 2) IEC TC 10 datasets. It should be noted that 100 electrical/thermal faults were selected in the IEC TC 10 databases. Even though the transformer specifications of the IEC TC 10 and KEPCO datasets are different, the scale of DGA data in the KEPCO databases is comparable to that in the IEC TC 10 databases. The difference between the two datasets is that only DGA data for fault states is provided in the IEC TC 10 databases.

The implementation of the proposed approach was executed on a desktop computer equipped with an Intel Core i7-6700K processor (4.00 GHz), 32 gigabytes of RAM, and an NVIDIA GeForce GTX 1080 graphics card (3072 CUDA

TABLE 3. Historical DGA data of four samples provided by KEPCO.

Sample	Year	H_2	C_2H_2	C_2H_4	C_2H_6	CH_4	CO	Health State
No.1	1999	0	0	2	2	6	172	N
	2000	0	0	13	9	25	282	N
	2001	0	0	37	31	35	163	N
	2002	0	0	28	85	44	209	N
	2003	251	1064	256	123	139	269	E
No.2	2011	10	0	2	5	7	57	N
	2012	13	0	3	26	11	71	N
	2013	48	14	12	63	24	214	N
	2015	335	1123	1324	150	246	105	E
No.3	2000	0	0	5	0	1	91	N
	2002	0	0	11	14	7	169	N
	2003	0	0	150	99	64	169	N
	2004	218	7	1743	264	744	371	T
No.4	2000	5	0	4	9	44	802	N
	2001	6	0	10	9	42	858	N
	2002	6	0	12	10	44	617	N
	2003	7	0	12	10	56	900	N
	2004	628	2.8	1873	351	1381	805	T

cores, 24 gigabytes of GDDR5 memory). The training of the proposed SAAT was conducted with the NVIDIA graphics card, while the other tasks (e.g., DGA data loading, fault classification and identification, and HFS extraction) were conducted with the Intel processor. The computer was controlled by Windows 10 and Python version 3.7. Computational times for each step were as follows: 1) loading the 110,000 DGA and preprocessing the dataset took 20 sec with the Intel processor, 2) training the proposed method SAAT consumed 61 sec, and 3) extracting the HFS took 15 sec. Thus, the overall computational time took 96 sec.

B. A BRIEF OUTLINE OF FOUR COMPARATIVE STUDIES AND QUANTATITIVE EVALUATION METRICS

The first comparative study aims to validate the effectiveness of the auxiliary detection task in SSAE-based fault diagnosis model. We consider the following two models: 1) SSAE-DU and 2) SSAE-IU. Notations 'D', 'I', and 'U' stand for 'fault detection task', 'fault identification task', and 'representative feature extraction task', respectively. Here, SSAE-DI is not considered, since a large portion of DGA data is unlabeled. Next, the validity of the proposed visualization method is elucidated in the second study. The following comparative methods are considered: 1) t-SNE and 2) SOM. Depending on how the high-level features $\mathbf{h}^{\mathrm{HF}}$ in SAAT are visualized, we investigate whether the monotonic health state transition can be represented in each method. In the third comparative study, we compared SAAT with existing methods to demonstrate the superior diagnosis performance of the proposed SAAT approach. Here, existing methods that can perform the unsupervised task were considered, such as principal component analysis (PCA) [29], sparse autoencoder (SAE) [33], and deep belief network (DBN) [34]. Finally, the diagnosis performance of state-of-the-art, semi-supervised deep learning algorithms – such as a semi-supervised variational autoencoder (SVAE) and semi-supervised generative adversarial network (SGAN) – are described in the last comparative study. To perform a one-to-one comparison, SGAN and the SVAE have the same three tasks as the proposed SAAT. We set parameters in SAE, DBN, SVAE, and SGAN, such as hyperparameters, layer and node sizes, activation functions in each layer, and the regularization terms, to be the same as those in the proposed SAAT.

When the given data suffers from imbalanced problems (e.g. the amount of data from the normal state is more than 1000 times that of the fault state, as in this study), several metrics are required to investigate the fault detection and identification performance. For the detection task, the following three metrics are under consideration [55]: positive predictive value (PPV), fault detection rate (FDR), and balanced accuracy rate (BAR). For the fault identification task [28], standard accuracy (I-Acc) is considered. With the confusion matrix presented in Table IV, these four metrics can be mathematically expressed as:

$$\mathrm{PPV} = \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} \Big/ \sum_{i=1}^{2}\sum_{j=1}^{3} C_{ij} \tag{26}$$

$$\mathrm{FDR} = \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} \Big/ \sum_{i=1}^{3}\sum_{j=1}^{2} C_{ij} \tag{27}$$

$$\mathrm{BAR} = 0.5\left(\sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} \Big/ \sum_{i=1}^{3}\sum_{j=1}^{2} C_{ij} + C_{33} \Big/ \sum_{i=1}^{3} C_{i3}\right) \tag{28}$$

$$\text{I-Acc} = \sum_{i=1}^{2} C_{ii} \Big/ \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} \tag{29}$$

In addition, as the quantitative evaluation metrics of health degradation performance in HFS, the following three metrics

TABLE 4. A confusion matrix for fault detection and identification evaluation metrics.

Predicted \ True	Thermal fault	Electrical fault	Normal state
Thermal fault	C_{11}	C_{12}	C_{13}
Electrical fault	C_{21}	C_{22}	C_{23}
Normal state	C_{31}	C_{32}	C_{33}

are under consideration [56]: 1) the trendability (Tre) of each health feature in terms of time, 2) the consistency (Con) between health features in HFS, and 3) the monotonic correlation coefficient (MCC) between health features in HFS. These metrics can be mathematically expressed as:

$$\mathrm{Tre} = \frac{K\sum_{k=1}^{K}\mathrm{HF}_k t_k - \sum_{k=1}^{K}\mathrm{HF}_k \sum_{k=1}^{K} t_k}{\sqrt{K\sum_{k=1}^{K}\mathrm{HF}_k^2 - \left(\sum_{k=1}^{K}\mathrm{HF}_k\right)^2}\sqrt{K\sum_{k=1}^{K} t_k^2 - \left(\sum_{k=1}^{K} t_k\right)^2}} \tag{30}$$

$$\mathrm{Con} = \frac{\sum_{k=1}^{K}\left(\mathrm{HF1}_k - \overline{\mathrm{HF1}}_{\mathrm{Con}}\right)\left(\mathrm{HF2}_k - \overline{\mathrm{HF2}}_{\mathrm{Con}}\right)}{\sqrt{\sum_{k=1}^{K}\left(\mathrm{HF1}_k - \overline{\mathrm{HF1}}_{\mathrm{Con}}\right)^2\left(\mathrm{HF2}_k - \overline{\mathrm{HF2}}_{\mathrm{Con}}\right)^2}} \tag{31}$$

$$\mathrm{MCC} = \frac{\sum_{n=1}^{N}\left(\mathrm{HF1}_n - \overline{\mathrm{HF1}}_{\mathrm{MCC}}\right)\left(\mathrm{HF2}_n - \overline{\mathrm{HF2}}_{\mathrm{MCC}}\right)}{\sqrt{\sum_{k=1}^{N}\left(\mathrm{HF1}_n - \overline{\mathrm{HF1}}_{\mathrm{MCC}}\right)^2\left(\mathrm{HF2}_n - \overline{\mathrm{HF2}}_{\mathrm{MCC}}\right)^2}} \tag{32}$$

where K and N are the number of measured time points and that of points in HFS, respectively; $\mathrm{HF1}_k$ (or $\mathrm{HF2}_k$) and $\mathrm{HF1}_n$ ($\mathrm{HF2}_n$) are health features at the time t_k and those at a certain point n in HFS, respectively; $\overline{\mathrm{HF1}}_{\mathrm{Con}}$ (or $\overline{\mathrm{HF2}}_{\mathrm{Con}}$) and $\overline{\mathrm{HF1}}_{\mathrm{MCC}}$ (or $\overline{\mathrm{HF2}}_{\mathrm{MCC}}$) are mean values of the health features at all times and those at all points in HFS, respectively. For one given sample, Tre aims at investigating the health degradation properties (or monotonic health state transition) in the time domain and Con shows the correlation between health features. On the other hand, MCC represents the degree of the linearity between two health features for all samples, which are scattered in HFS. These metrics are bounded from -1 to 1; these bounds in Tre and Con mean that the features are the strongest negative or positive linear correlation with time, respectively; those in MCC mean the highest monotonicity in the space. Please note that our IEC TC 10 datasets are only used for the I-Acc, since they do not have any historical information or normal state data.

C. EXPERIMENTAL RESULTS AND DISCUSSION

1) COMPARATIVE STUDY 1: EFFECTIVENESS OF THE AUXILIARY DETECTION TASK

The first comparative study is to investigate the effectiveness of the auxiliary detection task in SSAE-based fault diagnosis model. Table V summarizes the quantitative results of the fault detection and identification for SAAT, SSAE-DU, and SSAE-IU. For PPVs, SAAT shows the best fault detection performance, which reaches up to 92.8%, as compared with the others. FDRs of both SAAT and SSAE-IU are 100%, while that of SSAE-DU is 97.9%. For BARs, three diagnosis models exhibit more than 99%. It can be found that SAAT and SSAE-IU show better fault detection performance than SSAE-DU, although SAAT and SSAE-IU use the fault identification task that does not recognize whether the classes of the electrical/thermal fault types belong to the fault state. This can be interpreted from the number of classes; since SAAT and SSAE-IU have more classes to identify the fault types, they have more opportunities to impose more weights into the two classes (electrical/thermal fault types) in the fault identification task than one class (fault state) in the fault detection task. In the case of the fault identification performance, both SAAT and SSAE-IU show I-Acc of 100% for KEPCO datasets. It is worth pointing out that SSAE-DU cannot calculate I-Acc due to the lack of fault type information. For the IEC TC 10 datasets, SAAT presents a slightly better performance of 95.7% than that of SSAE-IU.

In terms of qualitative results, Figs. 6 (a) and (b) present HFSs that correspond to SSAE-IU and SSAE-DU, respectively. With the obtained decision boundaries, the results of the fault detection and/or identification can be visualized. However, it should be emphasized that Fig. 6 (a) cannot illustrate the monotonicity between health features and monotonic health state transition, as we expected in Fig. 3 (a). To support this interpretation, Figs. 6 (a) and (d) show the trends of health features for four samples, which are presented in Table III, in HFS, and in the time domain, respectively. As shown in Fig. 6 (a), two independent ways for the health state transition are observed. Moreover, Fig. 6 (d) presents that HF1s of the thermal faults (No. 3 and 4) tend to decrease, while HF2s gradually increases. Since these opposite trends are contradictory to the physical phenomenon, it is difficult for the two health features of SSAE-IU to represent the health degradation. For SSAE-DU, Fig. 6 (b) depicts the monotonic health state transition, as well as the high linearity between health features, as we expected in Fig. 3 (b). Further, from Fig. 6 (e), it can be found that both health features steadily increase. This implies that the fault detection task has the ability to present the health degradation features; however, as presented in Table V, the fault identification performance cannot be evaluated.

In summary, HFSs of SSAE-IU and SSAE-DU indicate that SSAE-IU can extract adequate health identification features, while SSAE-DU can extract adequate health degradation features. Therefore, by adding the auxiliary detection task into the loss function of SSAE-IU, HFS of SAAT, shown in Fig. 6 (c), enables pictorial visualization not only of the health identification results but also of the slightly separated monotonic health state transition from normal to each fault type. Furthermore, from four samples in

TABLE 5. Fault diagnosis performance of SAAE-IU, SSAE-DU, and the proposed SAAT.

Methods	Fault detection (%)			Fault identification (%)	
	KEPCO			KEPCO	IEC TC 10
	PPV	FDR	BAR	I-Acc	I-Acc
SSAE-IU	85.4±0.02	100	99.9±0.00	100	94.3±0.00
SSAE-DU	80.3±0.01	97.9±0.01	99.1±0.67	-	-
SAAT	92.8±0.02	100	99.9±0.00	100	95.7±0.01

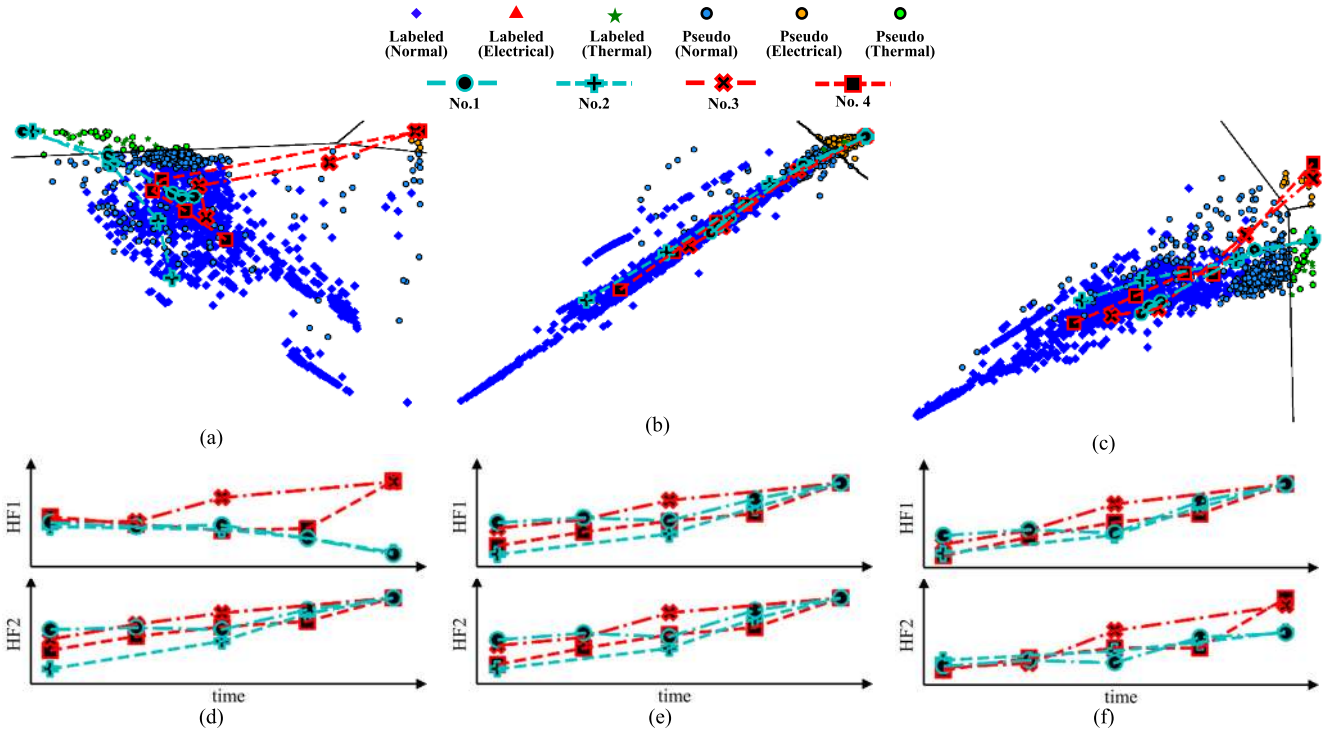

FIGURE 6. Results of comparative study 1: HFSs in (a) SSAE-IU, (b) SSAE-DU, and (c) the proposed SAAT; the trends of two health features with time for four samples in (d) SSAE-IU, (e) SSAE-DU, and (f) the proposed SAAT.

Figs. 6 (c) and (f), it can be seen that SAAT can successfully realize the representation of the health degradation properties in HFS. We devise a strict meaning of HFS as 2D space that can provide important information about both the health identification and health degradation.

Table VI summarizes the quantitative results of the health degradation. In the case of SSAE-IU, it can be confirmed that Tres of HF1 for the thermal fault have a negative sign, despite the health degradation properties. Therefore, unlike the results of SSAE-DU and SAAT, Cons for the electrical fault in SSAE-IU become the negative sign. These results are consistent with the intuitive interpretation from Fig. 6. In addition, MCCs of 0.96 and 0.88 for SSAE-DU and SAAT are much closer to 1 than that of the 0.69 result for SSAE-IU. Thus, MCC, which stands for the monotonicity between health features, can indirectly represent the health degradation performance of the health state transition in the time domain. Thus, it can be concluded that the auxiliary detection task significantly improves the health degradation performance that would otherwise be a challenge for SSAE-IU to represent.

2) COMPARATIVE STUDY 2: EFFECTIVENESS OF THE VISUALIZATION METHOD

The second comparative study is to investigate the effectiveness of the visualization method in the proposed SAAT approach. Here, there are two important points of emphasis. First, the feature spaces of t-SNE and SOM are obtained from the same values of HF1 and HF2 that were used when obtaining HFS in Fig. 6 (c). Second, since two high-level features obtained from two nodes are visualized in 2D, issues of the dimension reduction do not exist in t-SNE or SOM.

Figs. 7 (a) and (b) illustrate the obtained feature spaces that correspond to t-SNE and SOM, respectively. In Fig. 7 (a), both electrical and thermal faults are well clustered. However, it can be confirmed that the monotonic health state transition from normal to fault is not observed. The results of the samples (No. 1 to 4) do not show any specific trend. These observations are attributed to the characteristics of t-SNE. t-SNE converts similarities between the given high-level features into joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities

TABLE 6. Health degradation performance of SSAE-IU, SSAE-DU and the proposed SAAT.

Fault type	Dataset	Evaluation metrics	SAAE-IU	SAAE-DU	SAAT
Electrical fault	No.1	Tre (HF1)	0.52	0.97	0.98
		Tre (HF2)	0.98	0.97	0.89
		Con	0.67	0.99	0.95
	No.2	Tre (HF1)	0.94	0.98	0.98
		Tre (HF2)	0.98	0.97	0.97
		Con	0.92	0.99	0.99
Thermal fault	No.3	Tre (HF1)	-0.89	0.97	0.97
		Tre (HF2)	0.98	0.97	0.98
		Con	-0.91	0.99	0.99
	No.4	Tre (HF1)	-0.90	0.90	0.91
		Tre (HF2)	0.89	0.91	0.91
		Con	-0.98	0.99	0.98
Test dataset		MCC	0.69	0.96	0.88

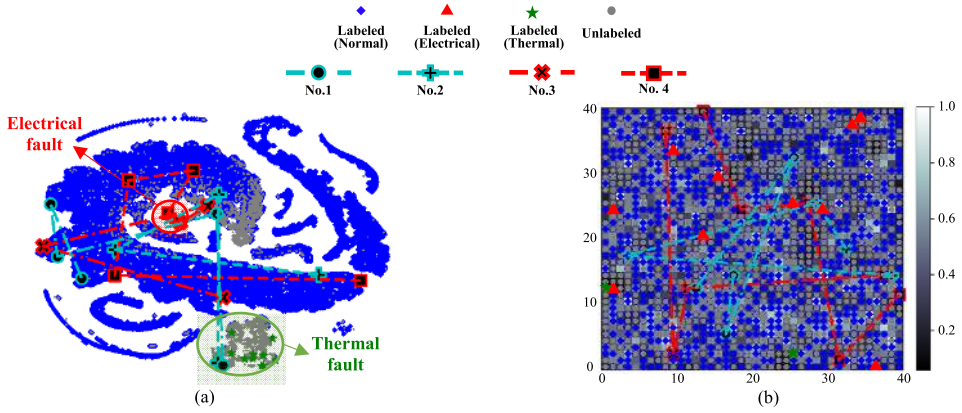

FIGURE 7. Results of comparative study 2: HFSs in (a) t-sNE and (b) SOM.

of the original features and converted features. During this process, the historical health degradation information in features can be significantly lost or distorted; thus, t-SNE is not suitable for representing the health degradation properties. In Fig. 7 (b), the color map presents the results of the clustering. Since SOM has the ability to map an ordered pair of the given high-level features HF1 and HF2 into a grid space, a certain point in the grid space can represent a grouping of similar features. The color close to one (white), indicates that the grid region consists of distinguishable features. On the other hand, the color close to zero (black), means that the grid region is clustered with similar features. It can be seen that in the feature space for SOM it is difficult to distinguish the fault states from the normal state. SOM is not suitable even for fault detection and identification before investigating the health degradation characteristics of the transformers. Therefore, it can be concluded that the proposed direct visualization method enables depiction of both fault diagnosis results and monotonic health state transition; it is otherwise a challenge for t-SNE and SOM to represent these results.

3) COMPARATIVE STUDY 3: CONVENTIONAL FAULT DIAGNOSIS METHODS

Next, we compare the fault diagnosis performance of conventional methods with those of the proposed SAAT. PCA, SAE, and DBN consider SC in the fault identification task. For PCA, extracted features from the unsupervised PCA algorithms are used to obtain diagnosis results. For SAE and DBN, sequential learning approaches are used; the methods of Restricted Boltzmann Machines and AE are under consideration in the pre-training part of SAE and DBN, respectively.

Table VII presents the quantitative results of fault detection and identification for PCA, SAE and DBN. It can be seen that PCA exhibits the worst diagnosis performance among the four models. Unlike other conventional and proposed methods, PCA is based on a fully unsupervised learning approach. The lack of labeled information makes it difficult to guarantee that the extracted features have correlation and consistency with the target labeling, thus worsening the detection and identification performance. Except for PPV, it can be seen that SAAT, SAE, and DBN show quite similar diagnosis performance; however, PPV of 92.8% in SAAT is much higher than those of 86.6% and 55.7% for SAE and DBN, respectively. These results indicate two important findings. First, from the viewpoint of fault identification results, it can be regarded that SAE and DBN were trained correctly in this study, because the results show reasonably high performance, as presented in previous studies [33], [34]. Second, although the first result satisfies the existing performance, since SAE

TABLE 7. Fault diagnosis and health degradation performance for conventional methods and state-of-the-art methods.

Methods		Fault detection (%)			Fault identification (%)		Health degradation
		KEPCO			KEPCO	IEC TC 10	KEPCO
		PPV	FDR	BAR	I-Acc	I-ACC	MCC
Conventional	PCA	2.00±0.00	55.0±0.04	76.5±1.78	38.3±4.00	69.6±0.02	0.00
	SAE	86.6±0.04	93.2±0.03	97.1±0.67	94.6±1.73	94.8±0.01	0.41
	DBN	55.7±0.01	100	99.7±0.00	100	92.3±0.01	0.42
State-of-the-art	SVAE	92.6±0.01	94.9±0.02	97.5±0.82	95.0±0.02	93.7±0.01	0.44
	SGAN	6.10±0.05	100	98.7±0.63	100	94.9±0.01	0.05

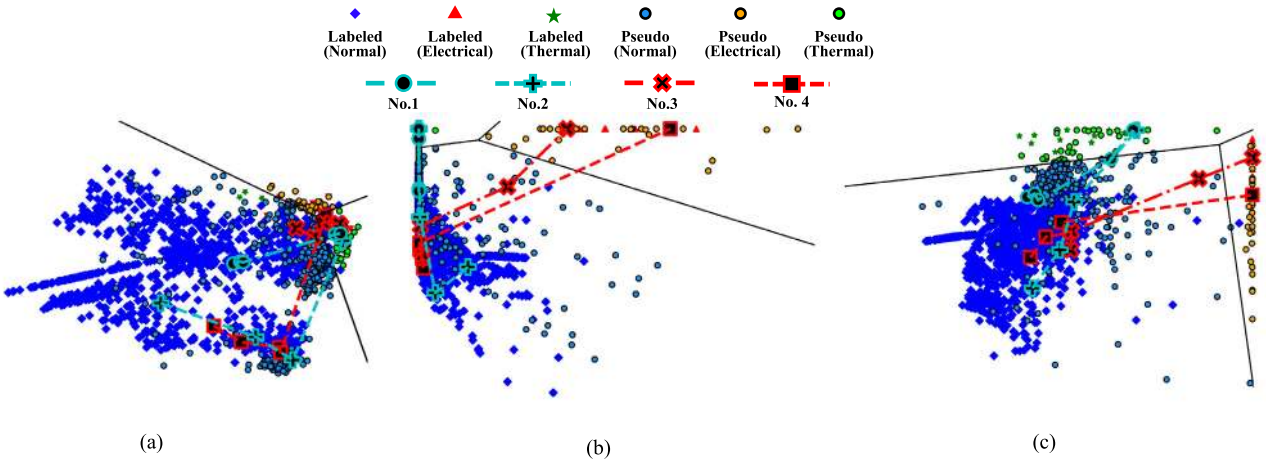

FIGURE 8. Results of comparative study 3: HFSs in (a) PCA, (b) SAE, and (c) DBN.

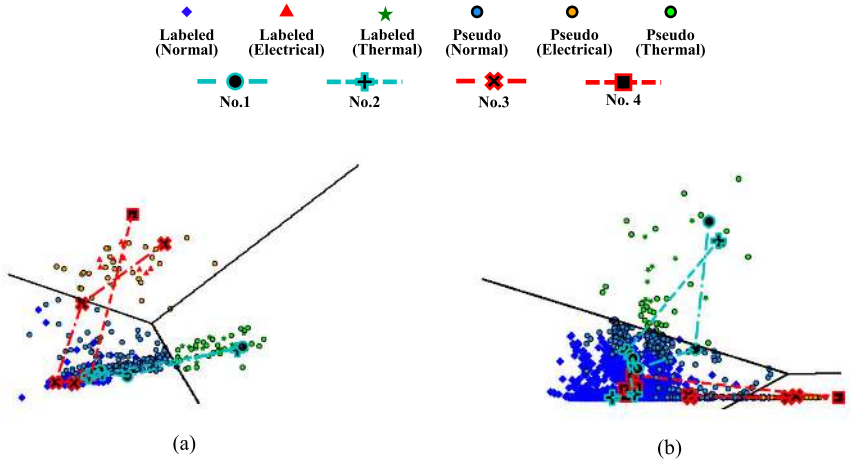

FIGURE 9. Results of comparative study 4: HFSs in (a) SVAE and (b) SGAN.

and DBN are prone to Type I error (i.e., estimating truly normal data as a fault), they could frequently raise a false alarm, which would be a vulnerability in terms of fault detection performance.

For qualitative results, Figs. 8 (a) to (c) present HFSs that correspond to PCA, SAE, and DBN, respectively. Fig. 8 (a) depicts that several normal points are misdiagnosed into fault regions; thus, the poor diagnosis performance of PCA can be confirmed. This is consistent with the quantitative results of fault detection and identification. In Figs. 8 (b) and (c), it can be seen that SAE and DBN can well classify the three classes; however, it is worth noting that they have difficulty representing the overall monotonicity between health features. The directions from the normal to the two fault regions are independent. This interpretation can be strengthened through the quantitative results of the health degradation, as shown in Table VII. MCC of 0.88 in SAAT is much closer to 1 than those of 0.00, 0.41 and 0.42 in PCA, SAE and DBN, respectively. Therefore, it can be concluded that the proposed SAAT approach outperforms conventional methods, with respect to the representation of health degradation in HFS.

4) COMPARATIVE STUDY 4: STATE-OF-THE-ART SEMI-SUPERVISED DEEP LEARNING

Lastly, we investigate whether the auxiliary detection task can be useful not only for SSAE method but also with other state-of-the-art, semi-supervised deep learning methods. The auxiliary detection task is added to the classifier part in SVAE and to the discriminator part in SGAN, respectively. Table VII presents the quantitative results of fault detection and identification for SVAE and SGAN. Except for PPV, it can be seen that SVAE, SGAN, and SAAT show quite similar diagnosis performance; however, PPVs of 92.8% in SAAT and 92.6% in SVAE are much higher than that of 6.10% in SGAN. This indicates the following two messages: 1) SGAN is prone to Type I error, since it could be unstable when optimizing parameters under an adversarial learning process, and 2) SVAE with the auxiliary detection task exhibits the best performance for fault detection and identification.

As qualitative results, Figs. 9 (a) and (b) present HFSs that correspond to SVAE and SGAN, respectively. In Fig. 9 (a), it can be seen that SVAE can well classify the three classes; it is worth pointing out that it is difficult to represent the overall monotonicity between health features, since the distribution of the latent space of SVAE follows the Gaussian distribution. The directions from normal to the two fault regions are independent. In Fig. 9 (b), it can be seen that SGAN misdiagnoses the normal points in the fault regions; thus, the poor diagnosis performance of SGAN can be confirmed and monotonicity between health features is not observed due to the unstable parameter optimization procedure. The quantitative results of the health degradation are summarized in Table VII. MCC of 0.88 in SAAT is much closer to 1 than those of 0.44 and 0.05 in SVAE and SGAN, respectively. Therefore, it can be concluded that the auxiliary detection task can be well executed only for SSAE-based fault diagnosis model.

V. CONCLUSION

In this study, a semi-supervised autoencoder with an auxiliary task (SAAT) was newly proposed to diagnose industrial power transformers using dissolved gas analysis (DGA). The method was tested using a large amount of DGA datasets provided by Korea Electric Power Corporation (KEPCO). The proposed idea consists of three main steps: 1) preprocessing DGA data, 2) extracting two health features by SAAT method, and 3) visualizing the two health features into two-dimensional space, a so-called health feature space (HFS). We evaluated the fault diagnosis and health degradation performance of the proposed approach in four comparative studies. The first study investigated the effectiveness of the auxiliary detection task in a semi-supervised autoencoder (SSAE)-based fault diagnosis model. The quantitative results of the fault detection and identification show that SAAT achieves over 90% performance in all metrics. Qualitative results of HFS show that SAAT represented the integrated characteristics of fault identification features in SSAE-IU and health degradation features in SSAE-DU. In the second comparative study, the proposed method of directly visualizing heath features without transformation or dimension reduction intuitively illustrates the health degradation properties as compared with conventional visualization methods (t-stochastic neighbor embedding (t-SNE) and self-organizing map (SOM)). In the third study, SAAT outperformed all conventional fault diagnosis methods (principal component analysis (PCA), sparse autoencoder (SAE), and deep belief network (DBN)) in terms of both quantitative and qualitative results of the health degradation performance. The last study investigated whether the auxiliary detection task can be useful not only for SSAE method but also for other state-of-the-art, semi-supervised deep learning methods (semi-supervised variational autoencoder (SVAE) and semi-supervised generative adversarial network (SGAN)). It was found that the auxiliary detection task can be well executed only for SSAE-based fault diagnosis model. Therefore, these experimental results examining real-world DGA datasets confirm that the auxiliary detection task in SSAE provides the opportunity to investigate not only fault identification but also health degradation; further, HFS helps to intuitively monitor the health state of power transformers.

Future research is suggested, as follows. First, the prediction of health state and/or remaining useful life of industrial power transformers should be performed using the proposed SAAT and its performance should be evaluated. Second, the proposed SAAT method should be verified with other systems where the health degradation is an important issue, (e.g., batteries and rotary machinery). Finally, more detailed fault types should be investigated, such as partial discharge faults, electrical faults of low and high discharge, and thermal faults of low, medium and high level.

ACKNOWLEDGMENT

(Sunuwe Kim and Soo-Ho Jo co-first authors.)

REFERENCES

[1] C. Aj, M. A. Salam, Q. M. Rahman, F. Wen, S. P. Ang, and W. Voon, "Causes of transformer failures and diagnostic methods—A review," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 1442–1456, Feb. 2018.

[2] H. Kim and B. D. Youn, "A new parameter repurposing method for parameter transfer with small dataset and its application in fault diagnosis of rolling element bearings," *IEEE Access*, vol. 7, pp. 46917–46930, 2019.

[3] S.-H. Jo, B. Seo, H. Oh, B. D. Youn, and D. Lee, "Model-based fault detection method for coil burnout in solenoid valves subjected to dynamic thermal loading," *IEEE Access*, vol. 8, pp. 70387–70400, 2020.

[4] M. Dong, H. Zheng, Y. Zhang, K. Shi, S. Yao, X. Kou, G. Ding, and L. Guo, "A novel maintenance decision making model of power transformers based on reliability and economy assessment," *IEEE Access*, vol. 7, pp. 28778–28790, 2019.

[5] E. Li, L. Wang, and B. Song, "Fault diagnosis of power transformers with membership degree," *IEEE Access*, vol. 7, pp. 28791–28798, 2019, doi: 10.1109/ACCESS.2019.2902299.

[6] M. Duval, "A review of faults detectable by gas-in-oil analysis in transformers," *IEEE Elect. Insul. Mag.*, vol. 18, no. 3, pp. 8–17, May 2002.

[7] *Mineral Oil-Filled Electrical Equipment in Service. Guidance on the Interpretation of Dissolved and Free Gases Analysis [Electronic Resource]* B. EN, document 60599: 2016, 2016.

[8] *IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers*, standard C57.104-1991, Piscataway, NJ, USA, 2009.

[9] E. Dornenburg and W. Strittmatter, "Monitoring oil-cooled transformers by gas-analysis," *Brown Boveri Rev.*, vol. 61, no. 5, pp. 238–247, 1974.

[10] M. Duval, "The duval triangle for load tap changers, non-mineral oils and low temperature faults in transformers," *IEEE Elect. Insul. Mag.*, vol. 24, no. 6, pp. 22–29, Nov. 2008.

[11] D.-E.-A. Mansour, "Development of a new graphical technique for dissolved gas analysis in power transformers based on the five combustible gases," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 5, pp. 2507–2512, Oct. 2015, doi: 10.1109/TDEI.2015.004999.

[12] V. Miranda and A. R. G. Castro, "Improving the IEC table for transformer failure diagnosis with knowledge extraction from neural networks," *IEEE Trans. Power Del.*, vol. 20, no. 4, pp. 2509–2516, Oct. 2005, doi: 10.1109/TPWRD.2005.855423.

[13] R. Naresh, V. Sharma, and M. Vashisth, "An integrated neural fuzzy approach for fault diagnosis of transformers," *IEEE Trans. Power Del.*, vol. 23, no. 4, pp. 2017–2024, Oct. 2008, doi: 10.1109/TPWRD.2008.2002652.

[14] H.-T. Yang, C.-C. Liao, and J.-H. Chou, "Fuzzy learning vector quantization networks for power transformer condition assessment," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 8, no. 1, pp. 143–149, Mar. 2001, doi: 10.1109/94.910437.

[15] Q. Su, C. Mi, L. L. Lai, and P. Austin, "A fuzzy dissolved gas analysis method for the diagnosis of multiple incipient faults in a transformer," *IEEE Trans. Power Syst.*, vol. 15, no. 2, pp. 593–598, May 2000, doi: 10.1109/59.867146.

[16] K. Bacha, S. Souahlia, and M. Gossa, "Power transformer fault diagnosis based on dissolved gas analysis by support vector machine," *Electr. Power Syst. Res.*, vol. 83, no. 1, pp. 73–79, Feb. 2012.

[17] R. J. Liao, J. P. Bian, L. J. Yang, S. Grzybowski, Y. Y. Wang, and J. Li, "Forecasting dissolved gases content in power transformer oil based on weakening buffer operator and least square support vector machine-Markov," *IET Gener., Transmiss. Distrib.*, vol. 6, no. 2, pp. 142–151, 2012, doi: 10.1049/iet-gtd.2011.0165.

[18] H. Wu, X. Li, and D. Wu, "RMP neural network based dissolved gas analyzer for fault diagnostic of oil-filled electrical equipment," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 18, no. 2, pp. 495–498, Apr. 2011, doi: 10.1109/TDEI.2011.5739454.

[19] M. H. Wang, "Extension neural network for power transformer incipient fault diagnosis," *IEE Proc. Gener., Transmiss. Distrib.*, vol. 150, no. 6, pp. 679–685, Nov. 2003, doi: 10.1049/ip-gtd:20030901.

[20] Y.-C. Huang, "Evolving neural nets for fault diagnosis of power transformers," *IEEE Trans. Power Del.*, vol. 18, no. 3, pp. 843–848, Jul. 2003, doi: 10.1109/TPWRD.2003.813605.

[21] Z. Wang, Y. Liu, and P. J. Griffin, "A combined ANN and expert system tool for transformer fault diagnosis," *IEEE Trans. Power Del.*, vol. 13, no. 4, pp. 1224–1229, 1998, doi: 10.1109/61.714488.

[22] T. Kari, W. Gao, A. Tuluhong, Y. Yaermaimaiti, and Z. Zhang, "Mixed kernel function support vector regression with genetic algorithm for forecasting dissolved gas content in power transformers," *Energies*, vol. 11, no. 9, p. 2437, Sep. 2018.

[23] J. Li, Q. Zhang, K. Wang, J. Wang, T. Zhou, and Y. Zhang, "Optimal dissolved gas ratios selected by genetic algorithm for power transformer fault diagnosis based on support vector machine," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 2, pp. 1198–1206, Apr. 2016.

[24] T. Hiroyasu, T. Shiraishi, T. Yoshida, and U. Yamamoto, "A feature transformation method using genetic programming for two-class classification," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, Dec. 2014, pp. 234–240.

[25] A. Shintemirov, W. Tang, and Q. H. Wu, "Power transformer fault classification based on dissolved gas analysis by implementing bootstrap and genetic programming," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 39, no. 1, pp. 69–79, Jan. 2009.

[26] V. Tra, B.-P. Duong, and J.-M. Kim, "Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 26, no. 4, pp. 1325–1333, Aug. 2019, doi: 10.1109/TDEI.2019.008034.

[27] Y. Cui, H. Ma, and T. Saha, "Improvement of power transformer insulation diagnosis using oil characteristics data preprocessed by SMOTE-Boost technique," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 21, no. 5, pp. 2363–2373, Oct. 2014, doi: 10.1109/TDEI.2014.004547.

[28] P. Mirowski and Y. LeCun, "Statistical machine learning and dissolved gas analysis: A review," *IEEE Trans. Power Del.*, vol. 27, no. 4, pp. 1791–1799, Oct. 2012.

[29] R. M. A. Velásquez and J. V. M. Lara, "Principal components analysis and adaptive decision system based on fuzzy logic for power transformer," *Fuzzy Inf. Eng.*, vol. 9, no. 4, pp. 493–514, Dec. 2017.

[30] T. Kari and W. Gao, "Power transformer fault diagnosis using FCM and improved PCA," *J. Eng.*, vol. 2017, no. 14, pp. 2605–2608, Jan. 2017.

[31] S. Misbahulmunir, V. K. Ramachandaramurthy, and Y. H. M. Thayoob, "Improved self-organizing map clustering of power transformer dissolved gas analysis using inputs pre-processing," *IEEE Access*, vol. 8, pp. 71798–71811, 2020.

[32] K. F. Thang, R. K. Aggarwal, A. J. McGrail, and D. G. Esp, "Analysis of power transformer dissolved gas data using the self-organizing map," *IEEE Trans. Power Del.*, vol. 18, no. 4, pp. 1241–1248, Oct. 2003.

[33] L. Wang, X. Zhao, J. Pei, and G. Tang, "Transformer fault diagnosis using continuous sparse autoencoder," *SpringerPlus*, vol. 5, no. 1, p. 448, Dec. 2016.

[34] J. Dai, H. Song, G. Sheng, and X. Jiang, "Dissolved gas analysis of insulating oil for power transformer fault diagnosis with deep belief network," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 24, no. 5, pp. 2828–2835, Oct. 2017, doi: 10.1109/TDEI.2017.006727.

[35] D. C. Ferreira, F. I. Vazquez, and T. Zseby, "Extreme dimensionality reduction for network attack visualization with autoencoders," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–10.

[36] X. Li, H. Jiang, K. Zhao, and R. Wang, "A deep transfer nonnegativity-constraint sparse autoencoder for rolling bearing fault diagnosis with few labeled data," *IEEE Access*, vol. 7, pp. 91216–91224, 2019.

[37] H. Shao, H. Jiang, H. Zhao, and F. Wang, "A novel deep autoencoder feature learning method for rotating machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 95, pp. 187–204, Oct. 2017.

[38] Y. Qi, C. Shen, D. Wang, J. Shi, X. Jiang, and Z. Zhu, "Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery," *IEEE Access*, vol. 5, pp. 15066–15079, 2017.

[39] J. Dai, H. Song, G. Sheng, and X. Jiang, "Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders," *IEEE Access*, vol. 5, pp. 22863–22870, 2017.

[40] W. Haiyan, Y. Haomin, L. Xueming, and R. Haijun, "Semi-supervised autoencoder: A joint approach of representation and classification," in *Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Dec. 2015, pp. 1424–1430.

[41] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[42] M. Chen, Y. Yao, J. Liu, B. Jiang, L. Su, and Z. Lu, "A novel approach for identifying lateral movement attacks based on network embedding," in *Proc. IEEE Intl Conf Parallel Distrib. Process. Appl., Ubiquitous Comput. Commun., Big Data Cloud Comput., Social Comput. Netw., Sustain. Comput. Commun. (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, Dec. 2018, pp. 708–715.

[43] F. Zhuang, D. Luo, X. Jin, H. Xiong, P. Luo, and Q. He, "Representation learning via semi-supervised autoencoder for multi-task learning," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 1141–1146.

[44] M.-T. Yang and L.-S. Hu, "Intelligent fault types diagnostic system for dissolved gas analysis of oil-immersed power transformer," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 20, no. 6, pp. 2317–2324, Dec. 2013, doi: 10.1109/TDEI.2013.6678885.

[45] S. Souahlia, K. Bacha, and A. Chaari, "MLP neural network-based decision for power transformers fault diagnosis using an improved combination of rogers and doernenburg ratios DGA," *Int. J. Electr. Power Energy Syst.*, vol. 43, no. 1, pp. 1346–1353, Dec. 2012.

[46] X. Wu, Y. He, and J. Duan, "A deep parallel diagnostic method for transformer dissolved gas analysis," *Appl. Sci.*, vol. 10, no. 4, p. 1329, Feb. 2020.

[47] F. Zhang, J. Yan, P. Fu, J. Wang, and R. X. Gao, "Ensemble sparse supervised model for bearing fault diagnosis in smart manufacturing," *Robot. Comput.-Integr. Manuf.*, vol. 65, Oct. 2020, Art. no. 101920.

[48] S. Nagpal, M. Singh, R. Singh, and M. Vatsa, "Regularized deep learning for face recognition with weight variations," *IEEE Access*, vol. 3, pp. 3010–3018, 2015.

[49] F. Li, J. M. Zurada, Y. Liu, and W. Wu, "Input layer regularization of multilayer feedforward neural networks," *IEEE Access*, vol. 5, pp. 10979–10985, 2017.

[50] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: http://arxiv.org/abs/1511.07289
[51] L. Yang, W. Chen, W. Liu, B. Zha, and L. Zhu, "Random noise attenuation based on residual convolutional neural network in seismic datasets," *IEEE Access*, vol. 8, pp. 30271–30286, 2020.
[52] H. Shao, H. Jiang, Y. Lin, and X. Li, "A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders," *Mech. Syst. Signal Process.*, vol. 102, pp. 278–297, Mar. 2018.
[53] M. Duval and A. dePabla, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," *IEEE Elect. Insul. Mag.*, vol. 17, no. 2, pp. 31–41, Mar. 2001.
[54] M. Noori, R. Effatnejad, and P. Hajihosseini, "Using dissolved gas analysis results to detect and isolate the internal faults of power transformers by applying a fuzzy logic method," *IET Gener., Transmiss. Distrib.*, vol. 11, no. 10, pp. 2721–2729, Jul. 2017, doi: 10.1049/iet-gtd.2017.0028.
[55] S. M. Frank, G. Lin, X. Jin, R. Singla, A. Farthing, L. Zhang, and J. Granderson, "Metrics and methods to assess building fault detection and diagnosis tools," Nat. Renew. Energy Lab., Golden, CO, USA, Tech. Rep., 2019, pp. 11–14.
[56] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, May 2018.

SUNUWE KIM received the B.S. degree from Korea University, Seoul, South Korea, in 2014. He is currently pursuing the Ph.D. degree in mechanical and aerospace engineering from Seoul National University, Seoul. His research interest includes prognostics and health management for power transformers. He received an award as the PHM Society Data Challenge Competition Winner, in 2015.

SOO-HO JO received the B.S. degree from Seoul National University, Seoul, South Korea, in 2016, where he is currently pursuing the Ph.D. degree in mechanical and aerospace engineering. His current research interests include piezoelectric vibration energy harvesting and elastic metamaterials. He was a recipient of the Bronze Prize from the KSME-SEMES Open Innovation Challenge, in 2016, the Best Paper Award from the KSME, in 2018 and 2020, the 2nd Place Winner in the Student Paper Competition of the KSME, in 2018, the Best Paper Award from the KSNVE, in 2019, and the 2nd Place Winner in the PHM Society Data Challenge Competition, in 2019.

WONGON KIM received the B.S. degree from Hanyang University, Seoul, South Korea, in 2015. He is currently pursuing the Ph.D. degree in mechanical and aerospace engineering with Seoul National University, Seoul. His current research interests include model verification and validation (V&V) and digital twins.

JONGMIN PARK received the B.S. degree from Seoul National University, Seoul, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in mechanical and aerospace engineering. His current research interests include prognostics and health management for rolling element bearings and gas insulated switchgear. He received an award as the PHM Asia Pacific Data Challenge Winner, in 2017 and 2019.

JINGYO JEONG received the B.S. and M.S. degrees from Hanyang University, Seoul, South Korea, in 1998 and 2004, respectively. He is currently pursuing the Ph.D. degree in mechanical engineering with Seoul National University, Seoul. He is also a Senior Manager with the Department of Transmission & Substation Operation, Korea Electric Power Corporation (KEPCO), Naju, South Korea. His current research interest includes PHM for power transmission systems.

YEONGMIN HAN received the B.S. degree in electrical engineering from Konkuk University, Seoul, South Korea, in 2004, and the M.S. degree in electrical engineering from Korea University, Seoul, in 2018. He is a Senior Manager with the Department of Transmission & Substation Operation, Korea Electric Power Corporation (KEPCO), Naju, South Korea.

DAEIL KIM received the B.S. degree from Dong Seoul University, Seoul, South Korea. He is a Senior Manager with the Department of Transmission & Substation Operation, Korea Electric Power Corporation (KEPCO), Naju, South Korea.

BYENG DONG YOUN (Member, IEEE) received the B.S. degree in mechanical engineering from Inha University, Incheon, South Korea, in 1996, the M.S. degree in mechanical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1998, and the Ph.D. degree in mechanical engineering from The University of Iowa, Iowa City, IA, USA, in 2001. He is currently a Full Professor of Mechanical Engineering with Seoul National University (SNU), and the Founder and CEO of OnePredict, Inc. His current research interests include prognostics and health management (PHM), engineering design under uncertainty, and energy harvester design. His dedication and efforts in research have garnered substantive peer recognition, resulting in many notable awards, including the Commendation of Prime Minister, in 2019, the Shin Yang Academic Award from Seoul National University, in 2017, the IEEE PHM Competition Winner, in 2014, and the PHM Society Data Challenge Winners, in 2014, 2015, 2017, and 2019.