공학박사학위논문

비표지 고장 데이터와 유중가스분석데이터를 이용한 딥러닝기반 주변압기 고장진단 연구

Deep-Learning-Based Fault Diagnosis Using Dissolved Gas Analysis for Unlabeled Fault Data of Industrial Power Transformers

2021 년 8 월

서울대학교 대학원

기계항공공학부

김 선 의

비표지 고장 데이터와 유중가스분석데이터를 이용한 딥러닝기반 주변압기 고장진단 연구

Deep-Learning-Based Fault Diagnosis Using Dissolved Gas Analysis for Unlabeled Fault Data of Industrial Power Transformers

지도교수 윤 병 동

이 논문을 공학박사 학위논문으로 제출함

2021 년 4 월

서울대학교 대학원

기계항공공학부

김 선 의

김선의의 공학박사 학위논문을 인준함

2021 년 6 월

위 원 장 : 박 종 우

부위원장 : 윤 병 동

위 원 : 김 도 년

위 원 : 조 규 진

위 원 : 권 대 일

Abstract

Deep-Learning-Based Fault Diagnosis Using Dissolved Gas Analysis for Unlabeled Fault Data of Industrial Power Transformers

Sunuwe Kim
Department of Mechanical and Aerospace Engineering
The Graduate School
Seoul National University

Due to the rapid development and advancement of today's industry, the demand for safe and reliable power distribution and transmission lines is becoming more critical; thus, prognostics and health management (hereafter, PHM) is becoming more important in the power transformer industry. Among various methods developed for power transformer diagnosis, the artificial intelligence (AI) based approach has received considerable interest from academics. Specifically, deep learning technology, which offers excellent performance when used with vast amounts of data, is also rapidly gaining the spotlight in the academic field of transformer fault diagnosis. The interest in deep learning has been especially noticed in the field of fault diagnosis, because deep learning algorithms can be applied to complex systems

that have large amounts of data, without the need for a deep understanding of the domain knowledge of the system.

However, the outstanding performance of these diagnosis methods has not yet gained much attention in the power transformer PHM industry. The reason is that a large amount of unlabeled and a small amount of fault data always restrict their deep-learning-based diagnosis methods in the power transformer PHM industry.

Therefore, in this dissertation research, deep-learning-based fault diagnosis methods are developed to overcome three issues that currently prevent this type of diagnosis in industrial power transformers: 1) the visualization of health feature space issue, 2) the insufficient data issue, and 3) the severity issue. To cope with these challenges, this thesis is composed of three research thrusts. The first research thrust develops a health feature space via a semi-supervised autoencoder with an auxiliary detection task. The proposed method can visualize a monotonic health trendability of the transformer's degradation properties. Further, thanks to the use of a semi-supervised approach, the method is applicable to situations with a large amount of unlabeled and a small amount labeled data (a situation common in industrial datasets). Next, the second research thrust proposes a new framework, that bridges the rule-based Duval method with an AI-based deep neural network (BDD). In this method, the rule-based Duval method is utilized to pseudo-label a large amount of unlabeled data. Furthermore, the AI-based DNN is used to apply regularization techniques and parameter transfer learning to learn the noisy pseudo-labelled data. Finally, the third thrust not only identifies fault types but also indicates a severity level. However, the balance between labeled fault types and the severity level is imbalanced in real-world data. Therefore, in the proposed method, diagnosis

of fault types – with severity levels – under imbalanced conditions is addressed by utilizing a generative adversarial network with an auxiliary classifier. The validity of the proposed methods is demonstrated by studying massive unlabeled dissolved gas analysis (DGA) data, provided by the Korea Electric Power Company (KEPCO), and sparse labeled data, provided by the IEC TC 10 database. Each developed method could be used in industrial fields that use power transformers to monitor the health feature space, consider severity level, and diagnose transformer faults under extremely insufficient labeled fault data.

Keywords: Fault diagnosis
Power transformer
Deep learning
Dissolved gas analysis

Student Number: 2014-22479

Table of Contents

List of Tables

List of Figures

Nomenclatures

PHM	prognostics and health management
AI	artificial intelligence
BDD	bridges the rule-based Duval method with an AI-based deep neural network
DNN	deep neural network
DGA	dissolved gas analysis
KEPCO	korea electric power company
SAAT	semi-supervised autoencoder with an auxiliary task
GANES	generative adversarial network with an embedding severity
ACGAN	auxiliary classifier generative adversarial network
GAN	generative adversarial network
SSAE	semi-supervised autoencoder
SOM	self-organizing maps
SVM	support vector machine
MLP	multilayer perceptron
D	discriminator
G	generator
z	noise vectors
p(z)	noise vector normal or uniform distribution
$G(z; \theta_g)$	fake samples via generator
θ_g	parameters of generator
x	real data vectors
θ_d	parameters of discriminator
δ_{ij}	Kronecker delta

∇	gradient operator
$\mathbb{E}$	expectation
min	minimum
max	maximum
L_{un}	unsupervised loss
L_{st}	supervised loss of diagnosing a severity level task
L_{it}	supervised loss of fault type identification task
L_{gen}	generator loss
$\mathbf{X}_{\text{un+sl}}$	input data of unlabeled fault type with the severity level
$\mathbf{X}_{\text{la+sl}}$	input data of labeled fault type with the severity level
$\mathbf{Y}_{\text{la}}$	labeled fault target data
L_{su}	supervised loss
R_i	handcrafted features
$(\mathbf{x}^{(\text{p})}, \mathbf{y}^{(\text{p})})$	training samples
p	number of DGA samples
$\mathbf{x}^{(\text{m})}$	input DGA data
$\mathbf{y}^{(\text{m})}$	one-hot encoded labeling information
$\mathbf{z}^{(m)} \in \mathbb{R}^{D'}$	activated DGA data
$\boldsymbol{\theta}$	parameters
$\mathbf{W} \in \mathbb{R}^{D' \times D}$	weight matrix
$\mathbf{b} \in \mathbb{R}^{D'}$	bias vector
$\mathbf{h}^{(m)}$	hidden unit
$q_{\text{sm}}()$	softmax function
D'	dimension of activated DGA data
C	number of class
X*	scaled DGA data

X_s^*	scaled-transformed source data
$\widetilde{\mathbf{Y}}_S$	corresponding pseudo-labeling information
$\widehat{\mathbf{X}}_{\text{end}}^*$	estimated DGA data at the end layer of pre-trained DNN
α	hyperparameter of weight between loss function
δ_n^{su}	Kronecker delta for supervised part
δ_n^{un}	Kronecker delta for unsupervised part
$\mathbf{X}_{\text{t}}^*$	transformed target training DGA data
Y_{t}	true fault states
ELU	exponential linear unit
Relu	rectified linear unit
HFS	health feature space
sm	softmax function
SC	softmax classifier
AE	autoencoder
f^{en}	encoder function
$\boldsymbol{\theta}^{\text{en}}$	encoder parameters
σ^{AE}	activation function, such as a sigmoid, a rectified linear unit (ReLU), and an exponential linear unit (ELU)
$\hat{\mathbf{x}}^{(m)}$	reconstructed data
f^{de}	decoder function
$\boldsymbol{\theta}^{\text{de}}$	decoder parameters
$L_{\text{AE}}^{(m)}$	m-th loss funtion of autoencoder corresponds to mean square error
$\boldsymbol{\theta}^{\text{AE}}$	autoencoder parameter
δ_k^{de}, δ_j^{en}	errors in the decoder and encoder layer
$\boldsymbol{\theta}^{\text{en}*}$, $\boldsymbol{\theta}^{\text{cl}}$	parameters dpdated by mini-batch gradient
$\boldsymbol{\theta}^{\text{shd}}$	shared parameters
$\boldsymbol{\theta}^{\text{SAAT}}$	parameters of SAAT

$\boldsymbol{\theta}^{\mathbf{shd},p}$, $\boldsymbol{\theta}^{\mathbf{de},q}$	parameters of p-th and q-th hidden layers in the shared network and the decoder
$\boldsymbol{\theta}^{\mathbf{iden}}$, $\boldsymbol{\theta}^{\mathbf{aux}}$	identification and auxiliary detection parameters
$\boldsymbol{\lambda}$	regularization hyperparameter
$h_i^{\mathrm{shd,end}}$	high-level features obtained at the end of the shared hidden layers
k, k', and k''	dimensions of output nodes in the first of the decoder, fault identification, and auxiliary detection tasks
β	weight hyperparameter between L_{SSAE} and L_{aux}
$\mathbf{X}_{\mathrm{un}}$, $\mathbf{X}_{\mathrm{la}}$	unlabeled, labeled DGA dataset
$\mathbf{Y}_{\mathrm{iden}}$, $\mathbf{Y}_{\mathrm{aux}}$	labeled information of fault types for identification and normal/fault for auxiliary detection task
SAE	sparse autoencoder
DBN	deep belief network
SVAE	semi-supervised variational autoencoder
SGAN	semi-supervised generative adversarial network
PPV	positive predictive value
FDR	fault detection rate
BAR	balanced accuracy rate
I-Acc	standard accuracy
Tre	trendability
Con	consistency
MCC	monotonic correlation coefficient
K and N	the number of measured time points and that of points in HFS
$\mathrm{HF1_k}$, $\mathrm{HF1_n}$	health features at the time t and those at a certain point n in HFS
SSAE-DU, SSAE-IU	‘D’, ‘I’, and ‘U’ stand for ‘fault detection task’, ‘fault identification task’, and ‘representative feature extraction task’

Chapter 1

Introduction

1.1 Motivation

As the power grid capacity continues to grow, power transformers have become crucial components of distribution and transmission lines in power 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].

Among the existing methods for diagnosing oil-filled transformers, dissolved gas analysis (DGA) is the most well-known method to determine the condition of the insulation materials found in internal transformers [5]. When insulation materials composed of hydrocarbon molecules are continuously exposed to electrical and thermal stresses, combustible gases (e.g., H_2, C_2H_2, C_2H_4) are decomposed from the insulation materials and then dissolved in the oil [6-10]. Via on/offline measurement of these dissolved gases, DGA can diagnose (i.e., detect and identify)

the fault types and estimate the fault severity level of the internal insulation health states of the transformers.

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, called handcrafted features, have been proposed; these features use experts' domain knowledge to provide fault identification that is based on human-experienced thresholds. However, rule-based methods have relatively low accuracy and inconsistent diagnosis results due to insufficient mathematical computation and their empirical handcrafted thresholds.

To overcome the underperformance of rule-based methods, AI-based methods, corroborated with data-driven methods, have been employed to improve fault diagnosis performance. In the beginning, conventional AI-based transformer fault diagnosis methods were studied as a supervised learning approach, through the use of fault-labeled DGA data [11, 12]. In addition, to increase the fault diagnosis performance, supervised approaches for feature selection techniques have been developed [13-18]. Despite some achievements of these supervised learning approaches, they have some limitations in that they use only labeled DGA datasets; these datasets are difficult to obtain in actual industrial fields. Thus, in other prior work, a few semi-supervised learning approaches have been developed to consider unlabeled data with labeled data. Furthermore, in recent years, with the help of deep learning methods, which are more advanced than conventional shallow learning approaches, the accuracy has been increased dramatically.

Although deep learning has achieved promising performance in areas such as

representation, reconstruction, and generation in image data, it is not a solution that can ultimately be directly applied to all industrial field data at once. Specifically, transformer DGA data, which is completely different from image data, has not yet received much research attention. Furthermore, deep-learning-based fault diagnosis research has not yet been focused on experts who actually manage industrial transformer maintenance. Thus, investigation of deep learning based fault diagnosis methods that examine real-world industrial issues is required so that these methods can be applied to industrial transformers.

There are currently three practical issues with transformer PHM in industrial fields. The first issue is transformer health state monitoring, which is not intuitive because conventional methods make transformer trends challenging to visualize. The second issue is the weakness of AI-based diagnosis performance, which depends on the number of labeled fault data. Although the aforementioned semi-supervised learning method has been developed, it is still not free from issues related to the number of labeled fault data and the distribution or characteristics of unlabeled data. Finally, AI-based methods so far have not been considered to determine fault severity levels. As a result, there is an issue in that industry practitioners must still maintain a comprehensive maintenance plan that uses rule-based methods to estimate the severity levels, even if AI-based methods are used to diagnose fault types.

Therefore, this thesis aims to study deep-learning-based fault diagnosis to overcome these practical issues in power transformers. After studying and developing a deep learning based fault diagnosis approach that addresses all three major issues, this thesis research achieves three major outcomes: 1) a health feature

space is enabled that can visualize the degradation of the monotonic health trendability and 2) a robust fault diagnosis method is proposed that bridges the rule-based Duval method and the deep neural network approach, and 3) feature extraction of the severity level, as well as fault identification, is enabled.

1.2 Research Scope and Overview

The goal of this dissertation is to propose deep-learning-based fault diagnosis methods for industrial issues in power transformers. Three research thrusts are proposed. First, a health feature space that can visualize the monotonic health trendability of transformer degradation via a semi-supervised autoencoder with an auxiliary task (SAAT) is developed. Next, an approach is proposed that bridges a rule-based Duval method and a deep neural network. Finally, a generative adversarial network that embeds a DGA severity level is proposed. These three thrusts are briefly described below.

Research Thrust 1: Extracting a Health Feature Space via a Semi-Supervised Autoencoder with an Auxiliary Task (SAAT)

Research thrust 1 considers a health feature space via SAAT for power transformer fault diagnosis using DGA. The health feature space generated by a semi-supervised autoencoder (SSAE) not only identifies normal and thermal/electrical fault types, it also presents the underlying characteristics of the

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 the health degradation properties. The overall procedure of the new approach includes three key steps: 1) preprocessing the DGA data, 2) extracting two health features via SAAT, and 3) visualizing the two health features in two-dimensional space. Then, 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 the 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.

Research Thrust 2: Bridging a Rule-based Duval Method and a Deep Neural Network

Research thrust 2 proposes a new framework, named BDD, that bridges Duval's method with a deep neural network (DNN) approach for power transformer fault diagnosis using dissolved gas analysis (DGA). The proposed BDD consists of the

following three key points. First, to overcome an important issue, which is that most DGA data found in real-world industrial settings is unlabeled, Duval's method is newly used to provide knowledge, which is called pseudo-labeling information, to a DNN for unlabeled DGA data. Second, motivated by the fact that the pseudo-labeled data does not always declare correct answers, a DNN architecture with an auxiliary regularization task is newly proposed; this approach is somewhat robust to the noisy labeled data. Last, a parameter transfer learning approach is applied to evolve the pre-trained DNN model, which is trained from a large amount of pseudo-labeled source data, to diagnose the sparse labeled target data. To demonstrate the effectiveness of the proposed approach, four case studies are executed: (i) a comparison with existing methods, (ii) examination of the effectiveness of parameter freezing via feature space investigation, (iii) studying the robustness of the regularization task under noisy labeled DGA, and (iv) probing the hyperparameter effects. These case studies confirm that the proposed BDD method outperforms existing methods, thanks to the Duval method's weak supervision, the regularization task, and parameter transfer.

Research Thrust 3: Embedding a Severity DGA Level into a Generative Adversarial Network

Research thrust 3 develops a generative adversarial network that embeds a severity DGA level. In actual industrial transformers, fault identification and severity estimation are essential to decide on a maintenance plan that can inform decisions about whether the system can operate normally or if repair or replacement is

necessary. However, the conventional artificial intelligence based method, which is trained with only labeled fault types, does not include the severity level. Thus, engineers must apply a different rule-based approach to estimate severity. Further, since the fault mode is challenging to obtain in industry settings, while rule-based methods simply annotate the severity, there is an unbalanced unlabeled data issue between the two states. Therefore, this research proposes a generative adversarial network with an embedding severity (GANES) DGA level. As a fundamental approach to alleviate the imbalanced problem between two classes of labeled fault types and severity levels, an auxiliary classifier of the generative adversarial network (ACGAN) was applied. To solve the unlabeled fault types that remain even with the ACGAN, this study employs a semi-supervised approach. The proposed method is demonstrated by studying massive Korea Electric Power Corporation (KEPCO) and IEC TC 10 databases. The results show that the proposed method not only outperforms conventional AI-based methods but also extracts both fault types and severity levels.

1.3 Dissertation Layout

The layout of this dissertation is as follows. Chapter 2 provides a literature review of power transformer fault diagnosis. Chapter 3 suggests extracting health feature space via semi-supervised autoencoder with an auxiliary task. Chapter 4 introduces a fault diagnosis method that bridges a rule-based Duval method and a deep neural network. Chapter 5 proposes a generative adversarial network with embedding a severity DGA level. Finally, chapter 6 concludes the dissertation by summarizing

the research and suggesting future research.

Chapter 2

Literature review

This chapter reviews the literature related to fault diagnosis of power transformer using dissolved gas analysis (DGA), specifically the review provides: (1) description of DGA, (2) an overview of rule-based fault diagnosis of transformer, and (3) an overview of conventional AI-based fault diagnosis.

2.1 A Brief Overview of Rule-Based Fault Diagnosis

Dissolved gas analysis is the most widely well-known method of diagnosing oil-filled power transformers. This is because through DGA, the internal insulation health state of the transformer can be estimated by analyzing the amounts of combustible gases and patterns generated by decomposition of insulating paper and insulating oil due to mechanical, electrical, and thermal stress.

From 1927 [19], there are a lot of rule-based methods that diagnose a fault types. Among them, five major fault diagnosis methods are as follows:

(1) Key gas method [20]: Unlike other rule-based methods, this method diagnoses a failure according to the concentration of each gas closely related

to the fault type. For example, if (1) O2 or N2 occurs, it is determined as normal. (2) Low temperature overheating of oil when CH4 and C2H6 occur, (3) high temperature overheating of oil when there is a lot of C2H4, (4) overheating of cellulose insulation when CO and CO2 are high, and (5) when H2 occurs corona, (6) C2H2 is arcing.

(2) Dornenburg ratio method [19]: Identifying faults (thermal, corona, discharge and arching)) by gas concentration ratios such as CH4/H2, C2H2/CH4, C2H4/C2H6 and C2H2/C2H4. The detailed flow chart is as follows:

(3) Rogers ratio method [21]: Compared to the Dornenburg ratio method, it uses following four gas concentration ratios such as CH4/H2, C2H6/CH4, C2H4/C2H6 and C2H2/C2H4. Rogers ratio method is more suitable to identify thermal fault than Dornenburg ratio method. The detailed flow chart is similar with as Dornenburg's flow chart:

(4) IEC ratio method [22]: The international electrotechnical commission (IEC) ratio method is similar to the Rogers ratio method, but excludes the C2H6/CH4 ratio. It identifies normal, partial discharge of low and high energy, thermal faults and electrical faults. However, it does not identify specific thermal and electrical subtypes.

(5) Duval triangle method [10]: Among rule-based methods, the Duval's method has been widely used due to its high accuracy and reliability. The basic technique is to extract three gas ratios, shown in (2.1), as handcrafted

features:

$$R_i = \mathrm{Gas}_i \Big/ \sum_{i=1}^{3} \mathrm{Gas}_i \ \text{where} \ \mathrm{Gas}_i \in \{\mathrm{C_2H_2},\ \mathrm{C_2H_4},\ \mathrm{CH_4}\}. \tag{2.1}$$

The main concern of the Duval's method is to identify seven fault types (partial discharge, high energy discharge, low energy discharge, thermal fault 1, thermal fault 2, thermal fault 3, and thermal and discharge fault).

However, all rule-based methods, features based on human experience usually underperform the diagnosis capability of AI-based methods, which are based on sufficient mathematical formulations and statistical approaches.

2.2 A Brief Overview of Conventional AI-Based Fault Diagnosis

In recent years, AI-techniques have been incorporated in power transformer fault diagnosis to improve accuracy. AI techniques include fuzzy logic [23-31], support vector machine [17, 32-39], artificial neural network, and multilayer perceptron [27, 32, 40-51]. To select optimal features and address imbalanced problems of DGA data, a genetic algorithm approach [16-18, 32, 52-54] and an adaptive over-sampling method [35, 55, 56] 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 [57]. 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 [58, 59]. Besides, self-organizing maps (SOM) of unsupervised neural network methods extracted feature maps of several fault types [58, 60-63].

Furthermore, deep learning techniques, such as sparse autoencoder [64] and deep belief network [65], have been used to pre-train the network via unsupervised greedy layer wise training with deep hierarchical hidden layers. A previous deep learning approach for transformer fault diagnosis consists of two training steps: 1) pretraining the initial network by unsupervised learning, and 2) finetuning the network with labeled information by softmax classifier.

While these advances have been developed based on academic fields, there are several practical issues in applying them to industrial power transformer fields. Among them, in this doctoral dissertation, three practical issues such as 1) visualization of health trendability, 2) insufficient data, and 3) diagnosis of fault types with severity level are studied in Chapter 4, Chapter 5, and Chapter 6.

Chapter 3

Extracting Health Feature Space via Semi-Supervised Autoencoder with an Auxiliary Task (SAAT)

Conventional 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 trendability from normal to fault, thus potentially estimating health states for unlabeled data or diagnosing fault states in advance. Lastly, visualization of the monotonic health trendability 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 [67].

Thus, in this Chapter 3, 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

The rest of section is organized as follows. In the Section 3.1 describes the background of SAAT. Section 3.2 and Section 3.3 demonstrate the proposed method and experimental results, respectively. Finally, the conclusions and future works of this study are outlined in Section V.

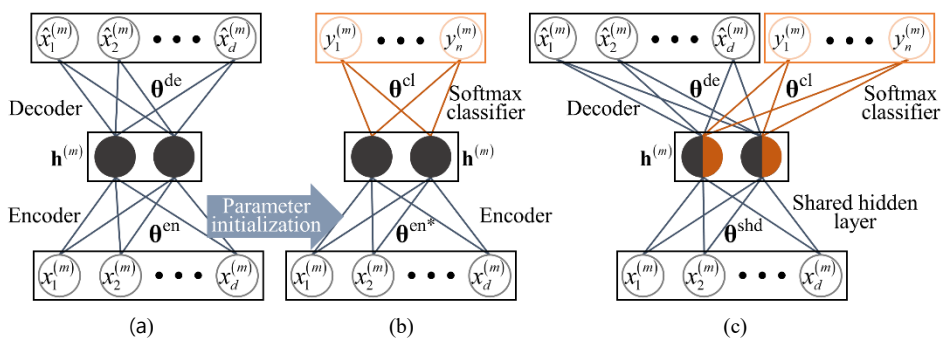


Figure 3-1 Architectures of AE, SC, and SSAE: (a) pre-training in the AE;

(b) fine-tuning in the SC with initialized parameters; and (c) simultaneous learning of the supervised and unsupervised learning parts in SSAE.

3.1 Backgrounds of Semi-supervised autoencoder (SSAE)

Two basic algorithms (i.e., an autoencoder (AE) and a softmax classifier (SC)) of the proposed SAAT are described in 3.1.1 and 3.1.2, respectively. In Section 3.1.3, SSAE is explained in terms of the AE and the SC.

3.1.1 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 Figure 3-1 (a) [68-71]. 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, ⋯ , 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:

$$f^{\text{en}}\left(x_i^{(m)}\right) = h_j^{(m)} = \sigma^{\text{AE}}\left(W_{ji}^{\text{en}} x_i^{(m)} + b_j^{\text{en}}\right) \tag{3.1}$$

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, $\mathbf{h}^{(m)}$ 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:

$$f^{\text{de}}\left(h_j^{(m)}\right) = \hat{x}_k^{(m)} = \sigma^{\text{AE}}\left(W_{kj}^{\text{de}} h_j^{(m)} + b_k^{\text{de}}\right) \tag{3.2}$$

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:

$$L_{\text{AE}}\left(\boldsymbol{\theta}^{\text{en}},\ \boldsymbol{\theta}^{\text{de}}\right) = \frac{1}{2N}\sum_{m=1}^{N}\left\|\hat{\mathbf{x}}^{(m)} - \mathbf{x}^{(m)}\right\|^2 = \frac{1}{2N}\sum_{m=1}^{N} L_{\text{AE}}^{(m)} \tag{3.3}$$

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:

$$\theta_{kj}^{\text{de}} \leftarrow \theta_{kj}^{\text{de}} - \eta \frac{\partial L_{\text{AE}}^{(m)}}{\partial \theta_{kj}^{\text{de}}}\left(\frac{\partial L_{\text{AE}}^{(m)}}{\partial \theta_{kj}^{\text{de}}} = \delta_k^{\text{de}} \frac{\partial z_k^{(m)}}{\partial \theta_{kj}^{\text{de}}} = \delta_k^{\text{de}} h_j^{(m)}\right) \tag{3.4}$$

$$\theta_{ji}^{\text{en}} \leftarrow \theta_{ji}^{\text{en}} - \eta \frac{\partial L_{\text{AE}}^{(m)}}{\partial \theta_{ji}^{\text{en}}}\left(\frac{\partial L_{\text{AE}}^{(m)}}{\partial \theta_{ji}^{\text{en}}} = \delta_j^{\text{en}} \frac{\partial z_j^{(m)}}{\partial \theta_{ji}^{\text{en}}} = \delta_j^{\text{en}} x_i^{(m)}\right) \tag{3.5}$$

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

$$z_k^{(m)} = W_{kj}^{\text{de}} h_j^{(m)} + b_k^{\text{de}} \tag{3.6}$$

$$\delta_k^{\text{de}} \equiv \frac{\partial L^{(m)}}{\partial z_k^{(m)}} = \sigma^{\text{AE}}{}'\left(z_k^{(m)}\right)\frac{\partial L^{(m)}}{\partial x_k^{(m)}} \tag{3.7}$$

$$z_j^{(m)} = W_{ji}^{\text{en}} x_i^{(m)} + b_j^{\text{en}} \tag{3.8}$$

$$\delta_j^{\text{en}} \equiv \frac{\partial L_{\text{AE}}^{(m)}}{\partial z_j^{(m)}} = \sum_k \frac{\partial L_{\text{AE}}^{(m)}}{\partial z_k^{(m)}}\frac{\partial z_k^{(m)}}{\partial z_j^{(m)}} = \sigma^{\text{AE}}{}'\left(z_j^{(m)}\right)\sum_k \theta_{kj}^{\text{de}}\delta_k \tag{3.9}$$

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 (3.3) to (3.9), AE can extract $\mathbf{h}^{(m)}$. Please note that the number of hidden layers in the encoder and the decoder can be extended.

3.1.2 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 [64, 65, 70]. When incorporating the SC into the AE, $\mathbf{h}^{(m)}$ can be the input data of a softmax function, as shown in Figure 3-1 (b). Training samples are a set of ordered pairs ($\mathbf{x}^{(m)}$, $\mathbf{y}^{(m)}$) as {($\mathbf{x}^{(1)}$, $\mathbf{y}^{(1)}$), ($\mathbf{x}^{(2)}$, $\mathbf{y}^{(2)}$), ⋯ , ($\mathbf{x}^{(N)}$, $\mathbf{y}^{(N)}$)} where $\mathbf{y}^{(m)} \in \{1, 2, \cdots, C\}$ 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 $\mathbf{y}^{(m)}$=($y_1^{(m)}$, $y_2^{(m)}$, ⋯ , $y_C^{(m)}$). 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:

$$\hat{y}_n^{(m)} = P\left(\mathbf{y}^{(m)} = n \mid f^{\text{en}}\left(\mathbf{x}^{(m)}\right); \boldsymbol{\theta}^{\text{en}*}, \boldsymbol{\theta}^{\text{cl}}\right) = q\left(z_n^{(m)}\right) = \frac{\exp\left(z_n^{(m)}\right)}{\sum_{n=1}^{C} \exp\left(z_n^{(m)}\right)} \tag{3.10}$$

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

$$z_n^{(m)} = W_{nj}^{\text{cl}} h_j^{(m)} + b_n^{\text{cl}} \tag{3.11}$$

Note that *n* means the *n*-th element in $\mathbf{y}^{(m)}$, as well as the number *n* in {1, 2, ⋯ , 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]:

$$L_{\text{cl}}\left(\boldsymbol{\theta}^{\text{en}*},\ \boldsymbol{\theta}^{\text{class}}\right) = -\frac{1}{N}\sum_{m=1}^{N}\mathbf{y}^{(m)}\log\left(\hat{\mathbf{y}}^{(m)}\right) \tag{3.12}$$

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

$$\theta_{nj}^{\text{cl}} \leftarrow \theta_{nj}^{\text{cl}} - \eta\frac{\partial L_{\text{cl}}^{(m)}}{\partial\theta_{nj}^{\text{cl}}}\left(\frac{\partial L_{\text{cl}}^{(m)}}{\partial\theta_{nj}^{\text{cl}}} = \delta_n^{\text{cl}}\frac{\partial z_n^{(m)}}{\partial\theta_{nj}^{\text{cl}}} = \delta_n^{\text{cl}}h_j^{(m)}\right) \tag{3.13}$$

$$\theta_{ji}^{\text{en}*} \leftarrow \theta_{ji}^{\text{en}*} - \eta\frac{\partial L_{\text{cl}}^{(m)}}{\partial\theta_{ji}^{\text{en}*}}\left(\frac{\partial L_{\text{cl}}^{(m)}}{\partial\theta_{ji}^{\text{en}*}} = \delta_j^{\text{en}*}\frac{\partial z_j^{(m)}}{\partial\theta_{ji}^{\text{en}*}} = \delta_j^{\text{en}*}x_i^{(m)}\right) \tag{3.14}$$

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

$$z_n^{(m)} = W_{nj}^{\text{cl}}h_j^{(m)} + b_n^{\text{cl}} \tag{3.15}$$

$$\delta_n^{\text{cl}} \equiv \frac{\partial L_{\text{cl}}^{(m)}}{\partial z_n^{(m)}} = \sigma^{\text{cl}}{}'\left(z_n^{(m)}\right)\frac{\partial L_{\text{cl}}^{(m)}}{\partial y_k^{(m)}} \tag{3.16}$$

$$\delta_j^{\text{en}*} \equiv \frac{\partial L_{\text{cl}}^{(m)}}{\partial z_j^{(m)}} = \sum_n\frac{\partial L_{\text{cl}}^{(m)}}{\partial z_n^{(m)}}\frac{\partial z_n^{(m)}}{\partial z_j^{(m)}} = \sigma^{\text{cl}}{}'\left(z_j^{(m)}\right)\sum_n\theta_{nj}^{\text{cl}}\delta_n^{\text{cl}} \tag{3.17}$$

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.

3.1.3 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 [72]. With this motivation, SSAE has been proposed, as shown in Figure 3-1 (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 **x** and the labeled information **y**, by simultaneously optimizing $\boldsymbol{\theta}^{\mathrm{AE}}$ and $\boldsymbol{\theta}^{\mathrm{cl}}$ [67, 72-75].

A loss function L_{SSAE} of SSAE is a summation of the two loss functions presented in (3.3) and (3.12) with a weight α as:

$$L_{\mathrm{SSAE}}\left(\boldsymbol{\theta}^{\mathrm{shd}},\ \boldsymbol{\theta}^{\mathrm{de}},\ \boldsymbol{\theta}^{\mathrm{cl}}\right)=\alpha L_{\mathrm{AE}}\left(\boldsymbol{\theta}^{\mathrm{shd}},\ \boldsymbol{\theta}^{\mathrm{de}}\right)+\left(1-\alpha\right)L_{\mathrm{cl}}\left(\boldsymbol{\theta}^{\mathrm{shd}},\ \boldsymbol{\theta}^{\mathrm{cl}}\right) \tag{3.18}$$

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:

$$\theta_{ji}^{\mathrm{shd}} \leftarrow \theta_{ji}^{\mathrm{shd}}-\eta\frac{\partial L_{\mathrm{SSAE}}^{(m)}}{\theta_{ji}^{\mathrm{shd}}}\quad\left(\frac{\partial L_{\mathrm{SSAE}}^{(m)}}{\theta_{ji}^{\mathrm{shd}}}=\alpha\delta_j^{\mathrm{AE}}x_i^{(m)}+\left(1-\alpha\right)\delta_j^{\mathrm{cl}}x_i^{(m)}\right) \tag{3.19}$$

where δ_j^{AE} and δ_j^{cl} are equal to (3.9) and (3.16), respectively. Finally, the shared hidden layers with $\boldsymbol{\theta}^{\mathrm{shd}}$ are able to concurrently extract representative features of **x** in the unsupervised learning and the labeled information of **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.

3.2 Input DGA Data Preprocessing

This 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 [57]. 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 [57]. 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 [57]. 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 [57]. 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 [57].

(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 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 [57]. Therefore, we consider six ratios of gas concentration DGA_i to total gas concentration $\sum_i DGA_i$ in the logarithmic scale, as log([DGA_i]/[$\sum_i DGA_i$]). Further, three ratios, developed by Duval triangle methods, are considered; these features are widely used in diagnosing transformer fault types [49, 76]. The total preprocessed input data lies in 15 dimensions.

3.3 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 trendability 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, 63]. 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 [61, 63, 77].

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.

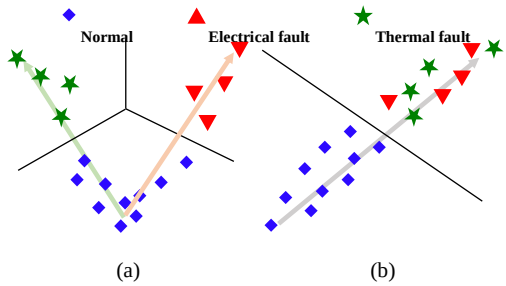


Figure 3-2 Conceptual diagrams of the health feature space: (a) fault identification task case in SSAE and (b) fault detection task case in SSAE

3.3.1 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 trendability, as shown in Figure 3-2 (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 health trendability in a single direction, as shown in Figure 3-2 (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:

$$\boldsymbol{\theta}^{\mathrm{SAAT}} = \left\{ \boldsymbol{\theta}^{\mathrm{shd},p}, \boldsymbol{\theta}^{\mathrm{iden}}, \boldsymbol{\theta}^{\mathrm{de},q}, \boldsymbol{\theta}^{\mathrm{aux}} \right\} \tag{3.20}$$

where $\boldsymbol{\theta}^{\mathrm{shd},p}$, $\boldsymbol{\theta}^{\mathrm{iden}}$, $\boldsymbol{\theta}^{\mathrm{de},q}$, 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 backpropagation 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 $\boldsymbol{\theta}^{\mathrm{shd},p}$. Hence, $\boldsymbol{\theta}^{\mathrm{shd},p}$ would possess all information of output layers, $\boldsymbol{\theta}^{\mathrm{iden}}$, $\boldsymbol{\theta}^{\mathrm{de},q}$, and $\boldsymbol{\theta}^{\mathrm{aux}}$.

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

$$\begin{aligned} L_{\mathrm{SAAT}}\left(\boldsymbol{\theta}^{\mathrm{SAAT}}\right) = {} & \beta L_{\mathrm{SSAE}}\left(\boldsymbol{\theta}^{\mathrm{shd},p}, \boldsymbol{\theta}^{\mathrm{iden}}, \boldsymbol{\theta}^{\mathrm{de},q}\right) \\ & + \left(1-\beta\right) L_{\mathrm{aux}}\left(\boldsymbol{\theta}^{\mathrm{shd},p}, \boldsymbol{\theta}^{\mathrm{aux}}\right) + 0.5\lambda \left\| \boldsymbol{\theta}^{\mathrm{SAAT}} \right\|^2 \end{aligned} \tag{3.21}$$

where L_{SSAE} is similar to (3.18); the differences are that the number of layers are much more in (3.21) and $\boldsymbol{\theta}^{\mathrm{cl}}$ in (3.18) is changed to $\boldsymbol{\theta}^{\mathrm{iden}}$. The loss function L_{aux} of the auxiliary detection task is newly proposed in (3.21). 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 (3.1) with a hyperparameter λ [78-80].

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:

$$\theta_{ji}^{\text{shd,end}} \leftarrow \theta_{ji}^{\text{shd,end}} - \eta \frac{\partial L_{\text{SAAT}}^{(m)}}{\partial \theta_{ji}^{\text{shd,end}}} \tag{3.22}$$

Similar to (3.19), the second term in the right-hand side of (3.22) can be decomposed as:

$$\frac{\partial L_{\text{SAAT}}^{(m)}}{\partial \theta_{ji}^{\text{shd,end}}} = \beta \delta_j^{\text{SSAE,end}} h_i^{\text{shd,end}} + (1-\beta) \delta_j^{\text{aux}} h_i^{\text{shd,end}} + \lambda \theta_{ji}^{\text{shd,end}} \tag{3.23}$$

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

$$\delta_j^{\text{SSAE,end}} = \alpha \times \sigma^{\text{de}\,\prime}\left(z_j\right) \sum_k \theta_{kj}^{\text{de},1} \delta_k^{\text{de},1} + (1-\alpha) \times \sigma^{\text{iden}\,\prime}\left(z_j\right) \sum_{k'} \theta_{k'j}^{\text{iden}} \delta_{k'}^{\text{iden}} \tag{3.24}$$

$$\delta_j^{\text{aux}} = \sigma^{\text{aux}\,\prime}\left(z_j\right) \sum_{k''} \theta_{k''j}^{\text{aux}} \delta_{k''}^{\text{aux}} \tag{3.25}$$

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 (3.24) and (3.25) into (3.23), $\boldsymbol{\theta}^{\text{shd,end}}$ are updated as (3.22). Finally, high-level features obtained by the proposed SAAT could play roles in exhibiting both fault identification and health degradation.

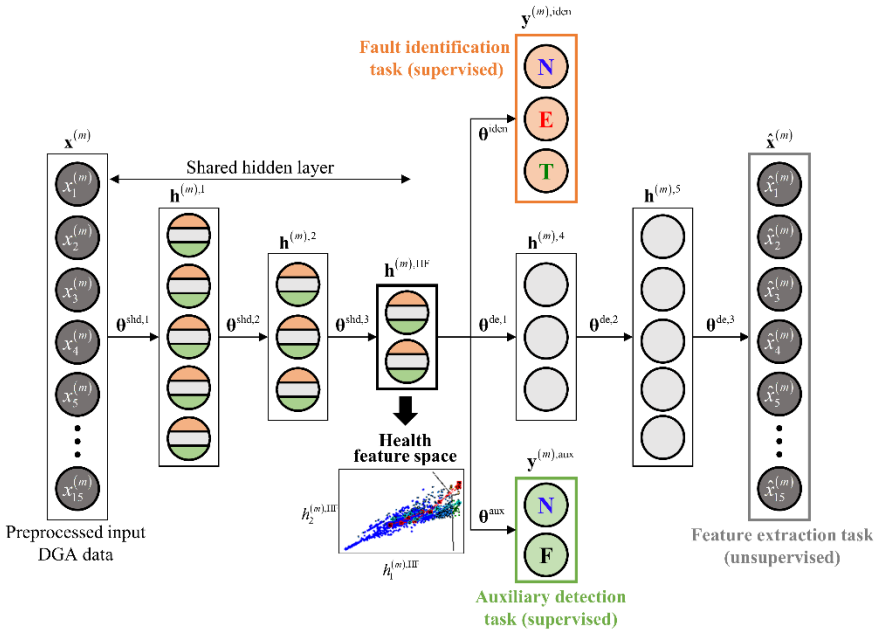


Figure 3-3 Architecture of the proposed SAAT: colors with orange, gray, and green in the shared hidden layers stand for the features related to the fault identification, representative characteristics of DGA data, and health trendability

Table 3-1 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

3.3.2 Architecture of the Proposed SAAT

As shown in Figure 3-3, 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 [81-83] 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 3-1. 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 $\mathbf{h}^{HF} \in \mathbb{R}^2$ 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.

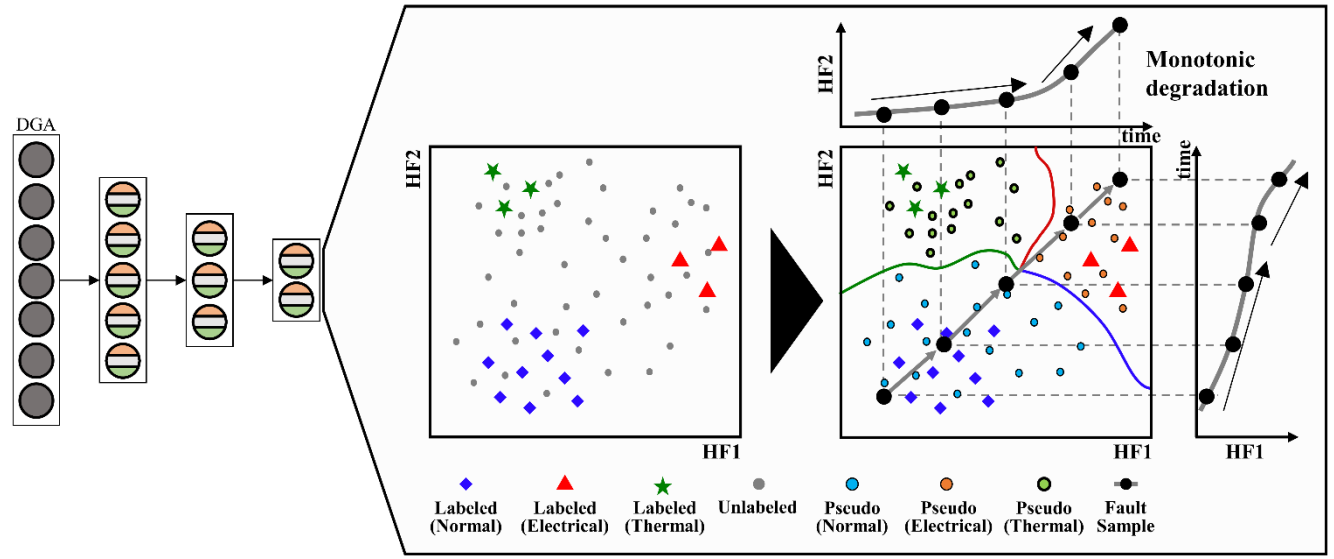


Figure 3-4 Visualization scheme of HFS with labeled and unlabeled data

3.3.3 Health Feature Space Visualization

Figure 3-4 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 $\mathbf{h}^{HF}$ for the training/test datasets can be visualized with a set of four dots, as shown in Figure 3-4. 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.

Moreover, thanks to the auxiliary detection task, the monotonic health trendability 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 trendability.

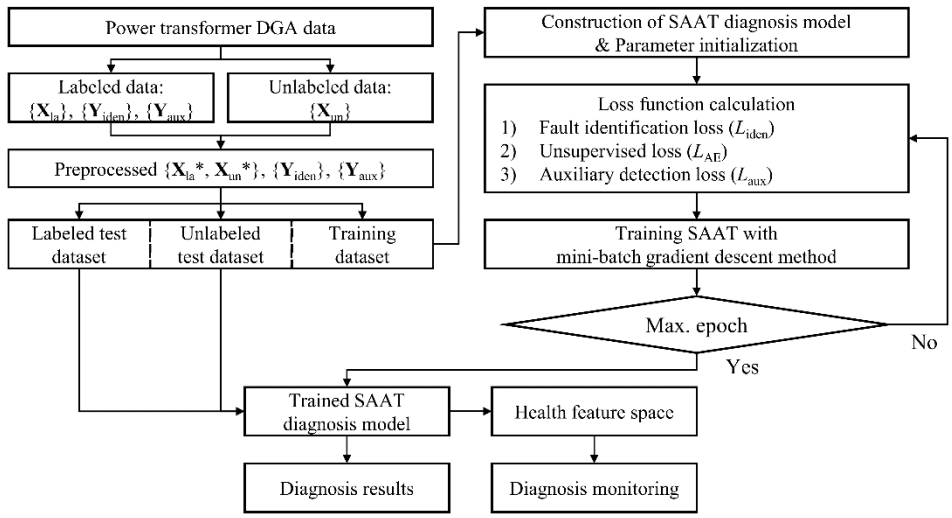


Figure 3-5 Overall procedures of the proposed SAAT-based fault diagnosis method

3.3.4 Overall Procedure of the Proposed SAAT-based Fault Diagnosis

Figure 3-5 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 $\{\mathbf{X}_{un}\}$, a labeled DGA dataset $\{\mathbf{X}_{la}\}$, and labeled information datasets $\{\mathbf{Y}_{iden}\}$ and $\{\mathbf{Y}_{aux}\}$ for the supervised tasks. After preprocessing, the input DGA datasets are denoted as $\{\mathbf{X}_{un}^*\}$ and $\{\mathbf{X}_{la}^*\}$. To train SAAT model and evaluate its performance, datasets, $\{\mathbf{X}_{un}^*\}$, $\{\mathbf{X}_{la}^*\}$, $\{\mathbf{Y}_{iden}\}$ and $\{\mathbf{Y}_{aux}\}$ 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, 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

Table 3-2 KEPCO maintenance standards for power transformer

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	-	-

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 $\mathbf{h}^{HF}$ in 2D space, the diagnosis results can be pictorially monitored.

3.4 Performance Evaluation of SAAT

This section is devoted to performance evaluation of the proposed SAAT method. Section 3.4.1 presents a description of datasets provided by KEPCO and implementation of the proposed method. In Section 3.4.2, the experimental setup is demonstrated. Lastly, the experimental results and discussion are covered in Section IV.C.

3.4.1 Data Description and Implementation

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 3-3. 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 [65] and 3,000 DGA dataset in [84]). 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 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.

Table 3-3 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

3.4.2 An Outline of Four Comparative Studies and Quantitative 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 $\mathbf{h}^{HF}$ in SAAT are visualized, we investigate whether the monotonic health trendability 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) [59], sparse autoencoder (SAE) [64], and deep belief network (DBN) [65]. 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 [85]: positive predictive value (PPV), fault detection rate (FDR), and balanced accuracy rate (BAR). For the fault identification task [57], standard accuracy (I-Acc) is considered. With the confusion matrix presented in Table 3-4, these four metrics can be mathematically expressed as:

$$\text{PPV}=\sum_{i=1}^{2}\sum_{j=1}^{2}C_{ij}\Big/\sum_{i=1}^{2}\sum_{j=1}^{3}C_{ij} \tag{3.26}$$

$$\text{FDR}=\sum_{i=1}^{2}\sum_{j=1}^{2}C_{ij}\Big/\sum_{i=1}^{3}\sum_{j=1}^{2}C_{ij} \tag{3.27}$$

$$\text{BAR}=0.5\left(\sum_{i=1}^{2}\sum_{j=1}^{2}C_{ij}\Big/\sum_{i=1}^{3}\sum_{j=1}^{2}C_{ij}+C_{33}\Big/\sum_{i=1}^{3}\text{C}_{i3}\right) \tag{3.28}$$

$$\text{I-Acc}=\sum_{i=1}^{2}C_{ii}\Big/\sum_{i=1}^{2}\sum_{j=1}^{2}C_{ij} \tag{3.29}$$

Table 3-4 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}

In addition, as the quantitative evaluation metrics of health degradation performance in HFS, the following three metrics are under consideration [86]: 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:

$$\mathrm{Tre}=\frac{K\sum_{k=1}^{K}\mathrm{HF}_k t_k-\sum_{k=1}^{K}\mathrm{HF}_k\sum_{k=1}^{K}t_k}{\sqrt{K\sum_{k=1}^{K}\mathrm{HF}_k^2-\left(\sum_{k=1}^{K}\mathrm{HF}_k\right)^2}\sqrt{K\sum_{k=1}^{K}t_k^2-\left(\sum_{k=1}^{K}t_k\right)^2}} \tag{3.30}$$

$$\mathrm{Con}=\frac{\sum_{k=1}^{K}\left(\mathrm{HF1}_k-\overline{\mathrm{HF1}}_{\mathrm{Con}}\right)\left(\mathrm{HF2}_k-\overline{\mathrm{HF2}}_{\mathrm{Con}}\right)}{\sqrt{\sum_{k=1}^{K}\left(\mathrm{HF1}_k-\overline{\mathrm{HF1}}_{\mathrm{Con}}\right)^2\left(\mathrm{HF2}_k-\overline{\mathrm{HF2}}_{\mathrm{Con}}\right)^2}} \tag{3.31}$$

$$\mathrm{MCC}=\frac{\sum_{n=1}^{N}\left(\mathrm{HF1}_n-\overline{\mathrm{HF1}}_{\mathrm{MCC}}\right)\left(\mathrm{HF2}_n-\overline{\mathrm{HF2}}_{\mathrm{MCC}}\right)}{\sqrt{\sum_{k=1}^{N}\left(\mathrm{HF1}_n-\overline{\mathrm{HF1}}_{\mathrm{MCC}}\right)^2\left(\mathrm{HF2}_n-\overline{\mathrm{HF2}}_{\mathrm{MCC}}\right)^2}} \tag{3.32}$$

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 trendability) 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.

3.4.3 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 3-5 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

Table 3-5 Fault diagnosis performance of SSAE-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

(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, Figure 3-6 (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 Figure 3-6 (a) cannot illustrate the monotonicity between health features and monotonic health trendability, as we expected in Figure 3-2 (a). To support this interpretation, Figure 3-6 (a) and (d) show the trends of health features for four samples, which are presented in Table 3-3, in HFS, and in the time domain, respectively. As shown Figure 3-6 (a), two independent ways for the health trendability are observed. Moreover, Figure 3-6 (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, Figure 3-6 (b) depicts the monotonic health trendability,

as well as the high linearity between health features, as we expected in Figure 3-2 (b). Further, from Figure 3-6 (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 3-5, 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 Figure 3-6 (c), enables pictorial visualization not only of the health identification results but also of the slightly separated monotonic health trendability from normal to each fault type. Furthermore, from four samples in Figure 3-6 (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 3-6 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

Table 3-6 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. Figure 3-6. In addition, MCCs of 0.96 and 0.88 for SSAE-DU and SAAT are much

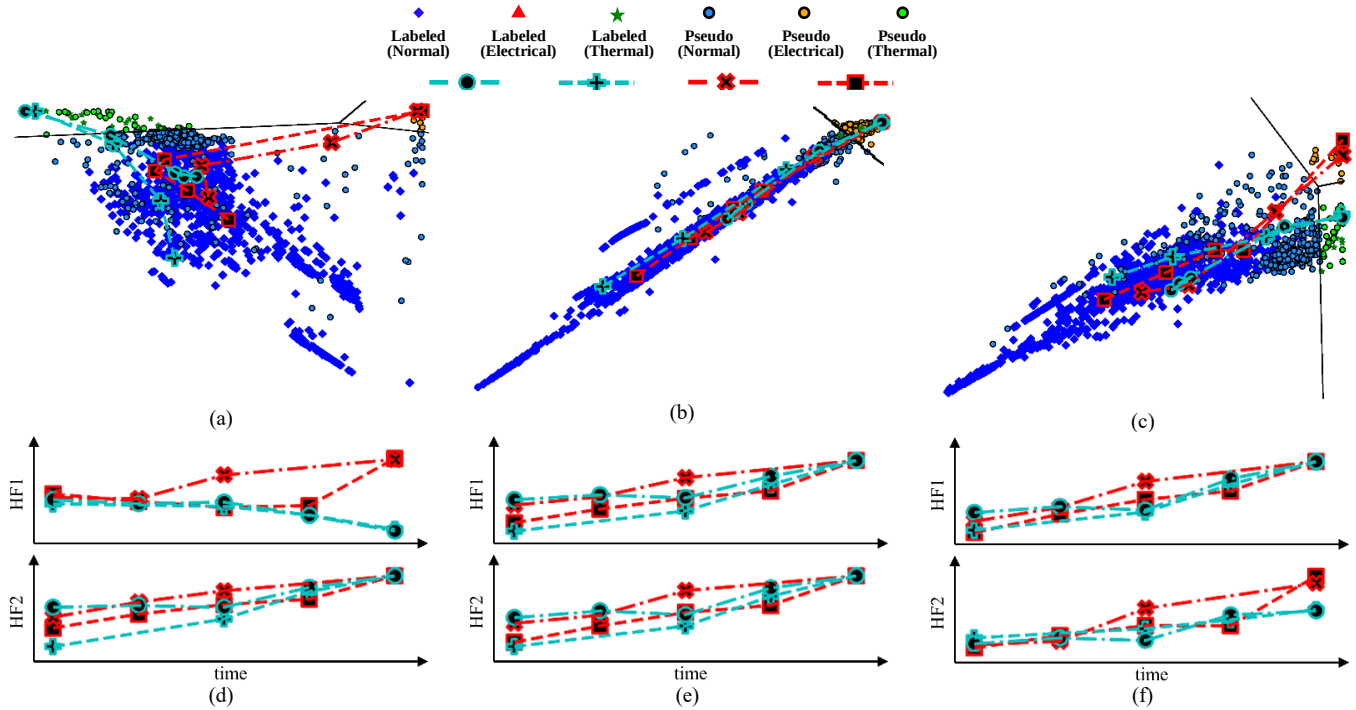


Figure 3-6 Results of comparative study 1: HFSs in (a) SSAE-IU, (b) SSAE-DU, and (c) the proposed SAAT; the trends of two health features with time for four samples in (d) SSAE-IU, (e) SSAE-DU, and (f) the proposed SAATverall procedures of the proposed SAAT-based fault diagnosis method

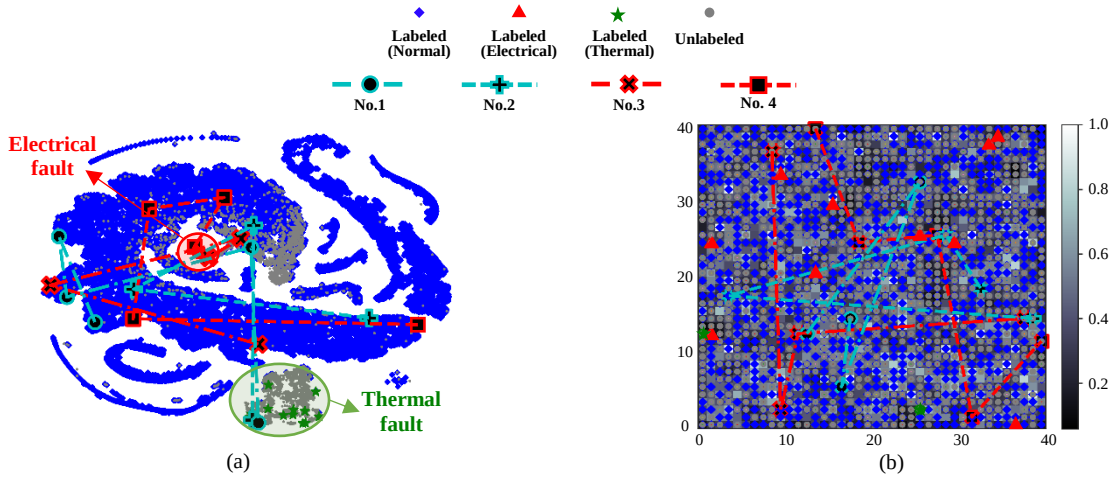


Figure 3-7 Results of comparative study 2: HFSs in (a) t-SNE and (b) SOM

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 trendability 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 Figure 3-6 (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. Figure 3-7 (a) and (b) illustrate the obtained feature spaces that correspond to t-SNE and SOM, respectively. In Figure 3-7 (a), both electrical and thermal faults are well clustered. However, it can be confirmed that the monotonic health trendability 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 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 Figure 3-7 (b), the color map presents the results of the clustering. Since SOM has the ability

Table 3-7 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

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 trendability; 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.

Figure 3-8 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 [64, 65]. Second, although the first result satisfies the existing performance, since SAE 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.

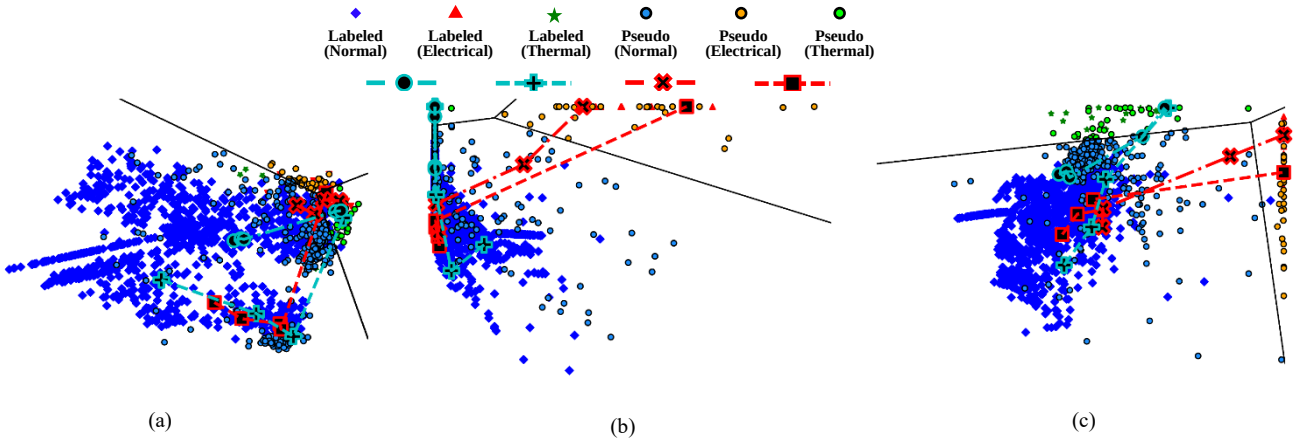


Figure 3-8 Results of comparative study 3: HFSs in (a) PCA, (b) SAE, and (c) DBN

For qualitative results, Figure 3-8 (a) to (c) present HFSs that correspond to PCA, SAE, and DBN, respectively. Figure 3-8 (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 Figure 3-8 (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 3-7. 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 3-7 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, Figure 3-9 (a) and (b) present HFSs that correspond to SVAE and SGAN, respectively. In Figure 3-9 (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 Figure 3-9 (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 3-7. 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.

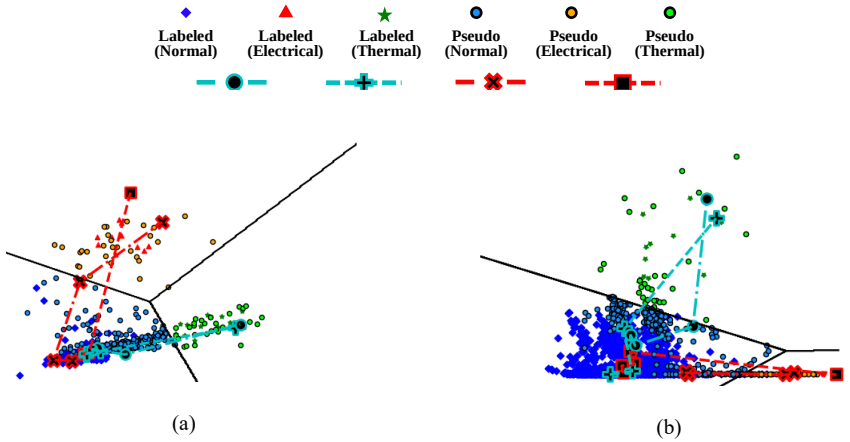


Figure 3-9 Results of comparative study 4: HFSs in (a) SVAE and (b) SGAN

3.5 Summary and Discussion

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.

Sections of this chapter have been published or submitted as the following journal articles:

1) **Sunuwe Kim,** Soo-Ho Jo, Wongon Kim, Jongmin Park, Jingyo Jeong, Yeongmin Han, Daeil Kim, and Byeng D. Youn, "A Semi-Supervised Autoencoder with an Auxiliary Task (SAAT) for Power Transformer Fault Diagnosis Using Dissolved Gas Analysis," *IEEE ACCESS*, Published, 2020.

Chapter 4

Learning from Even a Weak Teacher: Bridging Rule-based Duval Weak Supervision and a Deep Neural Network (BDD) for Diagnosing Transformer

The prerequisite for stable and reliable results for conventional AI-based fault diagnosis is that sufficient labeled datasets must be available for the training process. Unfortunately, as thorough visual inspection requires tremendous cost and time to consistently recognize incipient faults, most massive DGA datasets are unlabeled.

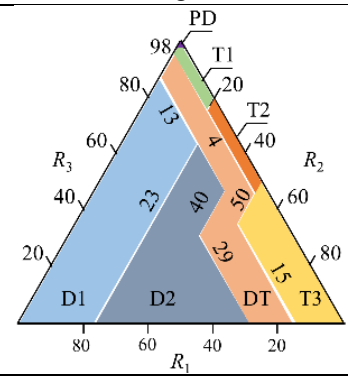
Emerging research in computer vision and image recognition has also examined real-world industrial settings, where data is overwhelmingly unlabeled [87]. In these fields, rather than developing a new complex model, there have been several attempts to integrate existing pre-trained models to make use of the advantages of each model, while minimizing the disadvantages [2, 88-92]. It implies that the combination of one model's weak supervision with the other deep-learning-approach can be one promising solution. For fault diagnosis, it is reasonable to combine the

advantages of rule-based methods, which enable to identify the unlabeled data, with those of deep learning-based methods, which do not require handcrafted features. Despite this physical insight, only a few studies have worked to bridge these two different methods for use in industrial applications [93, 94]. In the field of transformer fault diagnosis, this approach has yet to be explored.

Thus, in this Chapter 4, we propose a new framework, called BDD, which bridges Duval's rule-based weak supervision with a deep neural network (DNN) for transformer fault diagnosis using DGA. Key points in BDD are Duval's method, DNN, and parameter transfer method. The Duval's method (the teacher) virtually pseudo-labels health states for massive unlabeled data. Although the teacher does not always provide correct answers, it paves the way for AI (the student) to take expert knowledge into account. To learn the teacher's knowledge, as well as to reduce the effects of answers that might be overfitted, a DNN model with an auxiliary unsupervised task is used to both train and regularize the pseudo-labeled source data. Then, the pre-trained DNN model is transferred to sparse, but similar, labeled target data. Here, since the size of the target data is much less than that of the source data, a parameter-freezing technique is used to update the pre-trained DNN model, resulting in a re-trained DNN model [2]. The validity of BDD is demonstrated by massive unlabeled source data, provided by Korea Electric Power Corporation (KEPCO), and sparse target data, provided by IEC TC 10 database

The rest of this section is organized as follows. Section 4.1 outlines backgrounds of Duval's method and of parameter transfer of DNN. Sections 4.2 and 4.3 demonstrate the proposed method and experimental results, respectively. Finally, the conclusions of this work are provided in Section 4.4.

Table 4-1 Fault identification of the Duval triangle method

R1	R2	R3	Faults	Duval's triangle coordinate
0.00-0.02	0.98-1.00	0.00-0.02	PD	
0.00-0.04	0.46-0.80	0.20-0.50	T1	
	0.76-0.98	0.02-0.20	T2	
0.00-0.15	0.00-0.50	0.50-1.00	T3	
0.04-0.13	0.47-0.96	0.00-0.40		
0.13-0.29	0.21-0.56	0.40-0.50	DT	
0.15-0.29	0.00-0.35	0.50-0.85		
0.13-0.29	0.31-0.64	0.23-0.40	D1	
0.29-0.77	0.00-0.48	0.23-0.71	D2	

4.1 Backgrounds of BDD

4.1.1 Rule-based method: Duval Method

Among rule-based methods, the Duval's method has been widely used due to its high consistency and reliability [10]. The basic technique is to extract gas ratios, shown in (4.1), as handcrafted features:

$$R_i = \mathrm{Gas}_i \Big/ \sum_{i=1}^{3} \mathrm{Gas}_i \ \text{where}\, \mathrm{Gas}_i \in \{\mathrm{C_2H_2},\ \mathrm{C_2H_4},\ \mathrm{CH_4}\} \tag{4.1}$$

The main concern of the Duval's method is to identify seven fault types by using given thresholds, as presented in Table 4-1. Here, the thresholds were heuristically determined by pervious humans' experience and laboratory-level experiments. From the thresholds, the fault identification results can be intuitively depicted on the triangular coordinate system (Table 4-1) in terms of gas ratios. For example, when (R_1, R_2, R_3) is equal to (0.13, 0.4, 0.57), the health state is identified to T3. However, the rule-based method usually underperforms the AI-based method due to a lack of sufficient mathematical formulations and statistical approaches.

4.1.2 Deep learning Based Method: Deep Neural Network

A DNN is a hypernym of convolutional neural network (CNN) or recurrent neural network. However, since the DGA data is not a sort of images or time-series data, we limited the definition of the DNN to deeply stacked hidden layers consisting of nodes and activation functions that relate input and output responses in nodes [95]. A training sample is a set $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \cdots, (\mathbf{x}^{(p)}, \mathbf{y}^{(p)})\}$; p is the number of DGA samples. Input DGA data $\mathbf{x}^{(m)}$ and one-hot encoded labeling information $\mathbf{y}^{(m)}$ are in the D- and C-dimensions ($\mathbf{x}^{(m)} \in \mathbb{R}^{D}$ and $\mathbf{y}^{(m)} \in \mathbb{R}^{C}$), respectively.

A non-linear activation function *f*, such as a rectified linear unit and an exponential linear unit, linearly compresses input DGA data $\mathbf{x}^{(m)}$ into activated DGA data $\mathbf{z}^{(m)}$ ($\mathbf{z}^{(m)} \in \mathbb{R}^{D'}$) with parameters $\boldsymbol{\theta}$ (i.e., a weight matrix $\mathbf{W} \in \mathbb{R}^{D' \times D}$ and a bias vector $\mathbf{b} \in \mathbb{R}^{D'}$) and then converts it into a hidden unit $\mathbf{h}^{(m)}$ as follows:

$$\mathbf{h}^{(m)} = f\left(\mathbf{z}^{(m)}\right) = f\left(\mathbf{W}\mathbf{x}^{(m)} + \mathbf{b}\right) \tag{4.2}$$

If *f* is defined as the softmax function, q_{sm} is presented as:

$$q_{\text{sm}}\left(\mathbf{z}_n^{(m)}\right) = \exp\left(\mathbf{z}_n^{(m)}\right) \Big/ \sum_{n=1}^{C} \exp\left(\mathbf{z}_n^{(m)}\right) \tag{4.3}$$

where the dimension (D') of $\mathbf{z}^{(m)}$ is equal to that (C) of $\mathbf{y}^{(m)}$. Then, $\mathbf{h}^{(m)}$ becomes a one-hot encoded vector $\hat{\mathbf{y}}^{(m)}$ that contains the probability of $\mathbf{y}^{(m)}$. When DNN has N hidden layers, $\hat{\mathbf{y}}_{\text{end}}^{(m)}$ at the last hidden layer can be expressed as

$$\hat{\mathbf{y}}_{\text{end}}^{(m)} = q_{\text{sm}}\left(\mathbf{W}_N f_{N-1}\left(\cdots f_{N-1}\left(\mathbf{W}_1\mathbf{x}^{(m)} + \mathbf{b}_1\right)\cdots\right) + \mathbf{b}_N\right) \tag{4.4}$$

To match $\hat{\mathbf{y}}_{\text{end}}^{(m)}$ with $\mathbf{y}^{(m)}$, the parameters $\boldsymbol{\theta}$ in each layer need to be optimized by minimizing the loss function $L(\mathbf{y}, \hat{\mathbf{y}}_{\text{end}})$, which represents the discrepancy between $\mathbf{y}$ and $\hat{\mathbf{y}}_{\text{end}}$. Here, the cross-entropy loss function has been widely used as:

$$L\left(\mathbf{y},\ \hat{\mathbf{y}}_{\text{end}}\right) = -\frac{1}{p}\sum_{m=1}^{p}\mathbf{y}^{(m)}\log\left(\hat{\mathbf{y}}_{\text{end}}^{(m)}\right) \tag{4.5}$$

Thanks to the backpropagation method with mini-batch gradient descent algorithms, the parameters $\boldsymbol{\theta}$ are updated. For example, in the case of $\boldsymbol{\theta}^N$, which are parameters in the last hidden layer, the procedure to update $\boldsymbol{\theta}^N$ is organized as:

$$\theta_{nj}^{N} \leftarrow \theta_{nj}^{N} - \eta\frac{\partial L^{(m)}}{\partial \theta_{nj}^{N}}\left(\frac{\partial L^{(m)}}{\partial \theta_{nj}^{N}} = \delta_n^L\frac{\partial z_n^{(m)}}{\partial \theta_{nj}^{N}} = \delta_n^L h_j^{(m)}\right) \tag{4.6}$$

where η is a learning rate; an error δ_n^N is defined as:

$$\delta_n^N \equiv \frac{\partial L^{(m)}}{\partial z_n^{(m)}} = q_{\text{sm}}\left(z_n^{(m)}\right)\frac{\partial L^{(m)}}{\partial \hat{y}_k^{(m)}} \tag{4.7}$$

After parameter optimization, the DNN is able to identify the labels and extract high-level features that relate input DGA $\mathbf{x}^{(m)}$ to labeling information $\mathbf{y}^{(m)}$.

4.1.3 Parameter Transfer

One important issue in fault diagnosis research is that fault data or labeled data are typically insufficient due to the tremendous maintenance cost that would be required to collect it. Deep learning works well under the general assumption that both training and test data are drawn from the same distribution. However, this assumption fails in many real-world engineering applications. Parameter transfer learning is one promising solution to address this issue. Briefly, parameter transfer seeks to store knowledge (e.g., optimized parameters) obtained in an engineering problem (called the source data) and transfer it to a different, but related, problem (called the target data). Trained models in the source and target data are called pre-trained and re-trained models, respectively. Transfer learning plays a vital role in

achieving a dramatic improvement in fault diagnosis performance in the target data by reusing the pre-trained model. Note that several parameter transfer approaches exist, depending on the target data size: freezing, partial freezing, fine-tuning, and selective parameter freezing. If the source data is similar, but much larger than the target data, the freezing method has been mainly used [2, 88-91, 96, 97]

4.2 BDD Based Fault Diagnosis

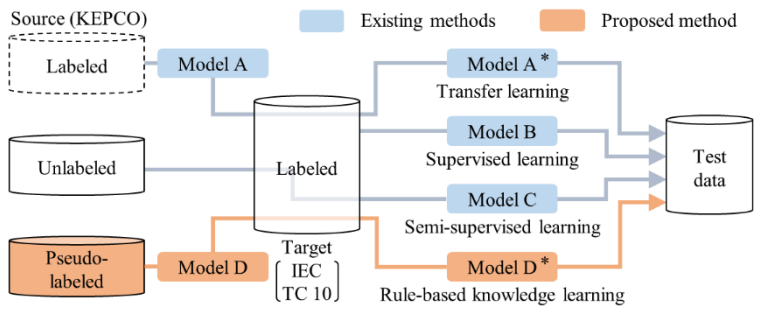


Figure 4-1 Problem statement of conventional and proposed method

4.2.1 Problem Statement

As shown in Figure 4-1, in the early days of research that has incorporated shallow learning techniques and handcrafted features into fault diagnosis, supervised (Model B) or semi-supervised (Model C) learning approaches were prevalent. Recently, to apply a fault diagnosis method that was developed for one system (source) into another but a similar, system (target), several deep learning studies that utilize transfer learning (e.g., parameter transfer (Models A and A*)) have been frequently reported. Here, the prerequisite of the parameter transfer is that a little labeled source data should be given. However, it should be emphasized that it is struggling to obtain

the labeled DGA in industrial power transformers. For example, although KEPCO has stored DGA data from thousands of transformers in South Korea for three decades, most of them are recorded as unlabeled data. Thus, in order to investigate the parameter transfer, when label data is absent, pseudo-labeling is performed based on prior knowledge of the rule-based Duval method. This is why the proposed approach attempts to bridge rule-based and deep learning methods.

However, it is worth pointing out that pseudo-labeling via rule-based methods also has limitations. Rule-based methods can convey some prior pseudo-labeled data; however, since the Duval's method is not always correct, there will be wrong or noisy labeled data. Thus, it is necessary to develop a deep-learning-based fault diagnosis model that is somewhat robust to noisy labeled data. A further step is required to update pre-trained models to reflect true labeling information. This is why we consider both DNN with a regularization task and parameter transfer learning.

4.2.2 Framework of the Proposed BDD

(1) Scaling input industrial DGA

The underlying characteristics of the measured DGA data are highly skewed distributions of gas concentrations. When a transformer suffers from a fault state, the concentrations suddenly rise to amounts hundreds or thousands of times the amounts of the prior concentration [57]. Motivated by this, DGA data is converted to a logarithmic scale. Normalizing the log-scale DGA data into a range [0, 1] helps to stabilize the numerical operations [57]. In Figure 4-2, the input DGA data without

and with scaling is denoted as X and X*, respectively. Unlike previous studies, we use only raw DGA data, without any augmented handcrafted features. This is in contrast to previous studies, which generally added a numerical combination of gas ratios and concentrations in input DGA data [64, 65].

(2) Pseudo-labeling unlabeled source DGA data by Duval weak supervision

A strong advantage of Duval's method is to identify fault states of unlabeled data from three gas ratios. This infers that Duval's method enables pseudo-labeling of unlabeled data. However, it is important to note that Duval's method includes two steps: (i) classification of the normal and fault states in advance using a rule-based method, called IEC 60599 (Table 4-2), and (ii) identification detailed fault states using the Duval's triangle (Table 4-1) [10].

The first step is to annotate the normal or fault states of given unlabeled source data using the several heuristic criteria of gas concentrations provided in IEC 60599 [66]. The unlabeled source data is classified as the fault state. The second step is to transform the gas concentrations of three gases (C_2H_2, C_2H_4, and CH_4) into gas ratios as presented (4.1). The last step is to pseudo-label the detailed fault states by spanning the transformed gas ratios into the Duval's triangle and identifying the fault states, as shown in Figure 4-2. $\mathbf{X}_s^*$ and $\widetilde{\mathbf{Y}}_s$ stand for the scale-transformed source data and corresponding pseudo-labeling information, respectively.

(3) Pre-training the DNN with an auxiliary unsupervised regularization task

Table 4-2 Thresholds for normal values in IEC 60599

Gas	H_2	C_2H_2	C_2H_4	C_2H_6	CH_4
Threshold [ppm]	60~150	3~50	60~280	50~90	40~110

With $\mathbf{X}_{\mathrm{s}}^{*}$ and $\tilde{\mathbf{Y}}_{\mathrm{s}}$, DNN needs to understand and follow up on the knowledge of the teacher (i.e., the Duval's method). This learning process can be regarded as parameter updating in the direction of minimizing the loss function L_{su} of the supervised task. As presented in (4.3), L_{su} is a cross-entropy loss function that represents the discrepancy between pseudo-labeled fault states $\tilde{\mathbf{Y}}_{\mathrm{s}}$ by the Duval's method and the estimated faults states $\hat{\mathbf{Y}}_{\mathrm{end}}$ by DNN. Here, the pseudo-labeled data is considered to be the 100% correct answer from the viewpoint of the DNN; however, it is not guaranteed that it is always true in reality. Therefore, to achieve robust diagnosis performance under noisy labeling problems, this paper newly adds an auxiliary unsupervised task term L_{un} in (4.2). L_{un} is a cross-entropy loss function that represents the discrepancy between the given DGA data $\mathbf{X}_{\mathrm{s}}^{*}$ and estimated DGA data $\hat{\mathbf{X}}_{\mathrm{end}}^{*}$ at the end layer of DNN. A well-known unsupervised autoencoder takes charge of extracting representative features of input source data by learning itself. Thus, L_{un} can work as a regularization effect to avoid overfitting problems of supervised learning [72, 98]. Finally, the loss function L_{DNN} can be expressed as:

$$L_{\mathrm{DNN}}\left(\boldsymbol{\theta}\right)=\left(1-\alpha\right)L_{\mathrm{su}}\left(\tilde{\mathbf{Y}}_{\mathrm{s}},\ \hat{\mathbf{Y}}_{\mathrm{end}}\right)+\alpha L_{\mathrm{un}}\left(\mathbf{X}_{\mathrm{s}}^{*},\ \hat{\mathbf{X}}_{\mathrm{end}}^{*}\right) \tag{4.8}$$

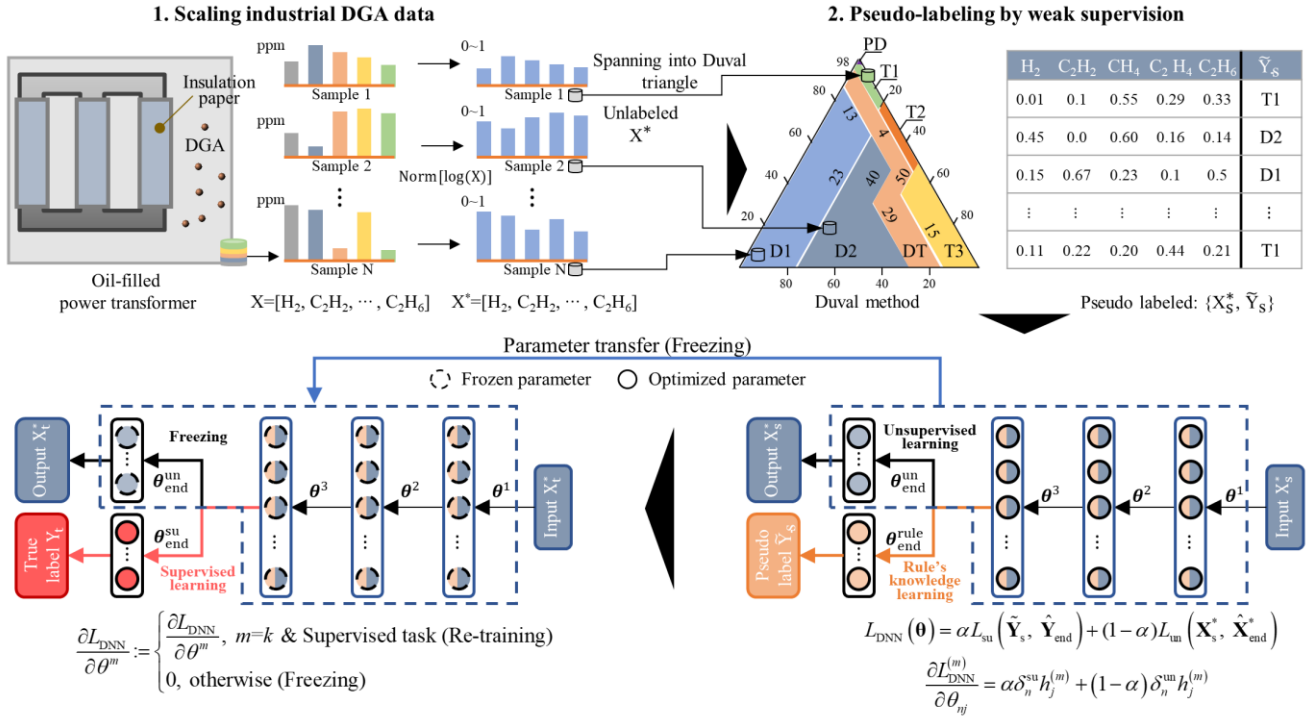


Figure 4-2 A framework of the proposed BDD

where the hyperparameter α is the weight between L_{su} and L_{un}. Therefore, DNN pre-trains not only the pseudo-labeled source data but also the input DGA source data to simultaneously extract both estimated labeled data (L_{su}) and representative features (L_{un}). Although this approach has been widely used in computer vision and image recognition [99-103], little attention has been paid in the field of transformer fault diagnosis.

A process to optimize the parameters is as follows:

$$\theta_{nj} \leftarrow \theta_{nj} - \eta\left((1-\alpha)\delta_n^{\text{su}} h_j^{(m)} + \alpha\delta_n^{\text{un}} h_j^{(m)}\right) \tag{4.9}$$

where δ_n^{su} and δ_n^{un} are defined, respectively, as:

$$\delta_n^{\text{su}} \equiv \frac{\partial L_{\text{su}}^{(m)}}{\partial z_n^{(m)}} = f_{\text{su}}'\left(z_n^{(m)}\right)\frac{\partial L_{\text{su}}^{(m)}}{\partial \hat{\mathbf{Y}}_{\text{end}}^{(m)}}, \tag{4.10}$$

$$\delta_n^{\text{un}} \equiv \frac{\partial L_{\text{un}}^{(m)}}{\partial z_n^{(m)}} = f_{\text{un}}'\left(z_n^{(m)}\right)\frac{\partial L_{\text{un}}^{(m)}}{\partial \hat{\mathbf{X}}_{\text{end}}^{*(m)}} \tag{4.11}$$

Eqs. (4.3) and (4.4) represent the updating process to learn $\widetilde{\mathbf{Y}}_{\text{s}}$ and $\mathbf{X}_{\text{s}}^{*}$, respectively. Finally, with the optimized parameters of the supervised and unsupervised tasks presented in (4.5), the pre-trained DNN model enables both regularizing and extracting the labeling information for fault diagnosis.

(4) Re-training the DNN with parameter transfer

With the transformed target training DGA data $\mathbf{X}_{\text{t}}^{*}$ and corresponding true fault states $\mathbf{Y}_{\text{t}}$, the pre-trained DNN model must be updated via parameter transfer. Recall the research backgrounds presented in Section II.B and the problem statement in

Section III.A. When the source data is similar, but much larger than the target data, the freezing method has been mainly used. Therefore, the parameters in all layers of the pre-trained model, other than the last layer, are frozen. Furthermore, since the main focus of the re-training is to reflect the true labeled target data as much as possible, the parameters in the unsupervised regularization task are also frozen. This parameter freezing approach can be mathematically expressed as:

$$\frac{\partial L_{DNN}}{\partial \theta^m} = \begin{cases} \frac{\partial L_{DNN}}{\partial \theta^m}, & m=N \,\&\, \text{Supervised task (Re-training)} \\ 0, & \text{otherwise (Freezing)} \end{cases} \tag{4.12}$$

where m is the m-th hidden layer and N is the number of total layers. Similar to the procedure in(4.9) to (4.12), the parameters in the supervised task can be re-

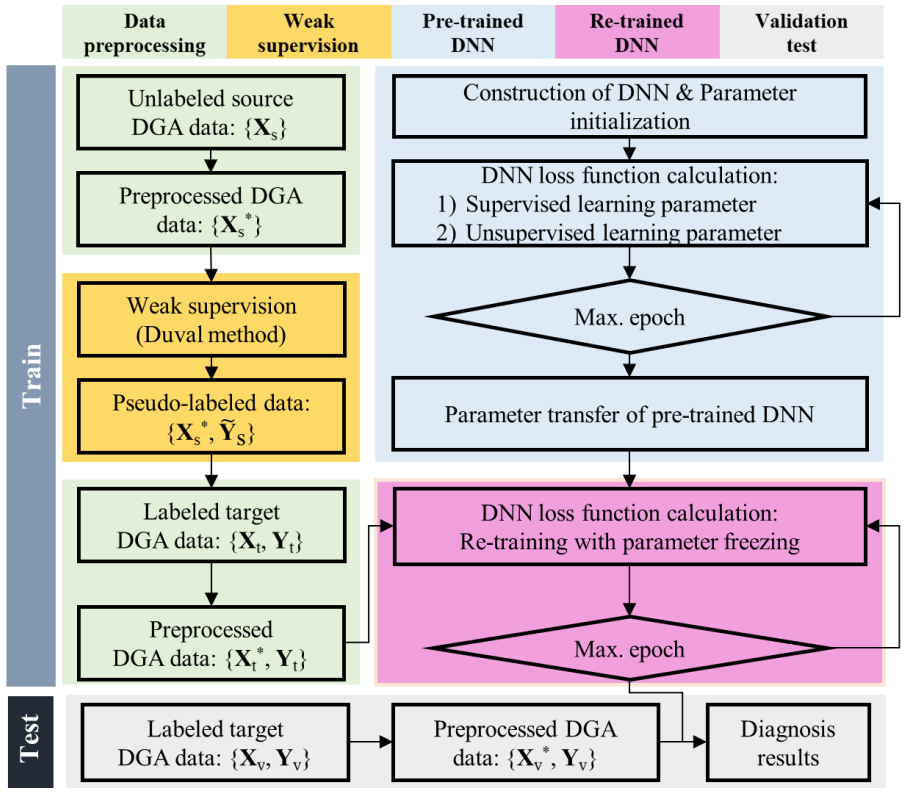


Figure 4-3 Overall procedure of the proposed BDD-based fault diagnosis-

optimized. Thus, the proposed BDD can consider both characteristics of the massive unlabeled source and sparse labeled target data. This approach is distinguished from conventional studies, which were conducted under ideal conditions of well-organized, labeled data [64].

4.2.3 Overall Procedure of BDD-based Fault Diagnosis

Figure 4-3 illustrates the overall procedure of the proposed BDD-based fault diagnosis. The first step is to transform a scale of both unlabeled source {Xs} and labeled target data {Xt}. After preprocessing, each input data is denoted as $\{X_s^*\}$ and

$\{X_t^*\}$, respectively. The next step is to weakly supervise {Xs*} by pseudo-labeling as $\widetilde{Y}_s$, using the Duval's method. The third step is to construct and stabilize the DNN-based fault diagnosis model using {Xs*, $\widetilde{Y}_s$}. For randomly initialized parameters, Lsu, and Lun are calculated. For a given batch size, the backpropagation method with a mini-batch gradient descent method trains the BDD-based model. The last step is to re-train the pre-trained DNN model from $\{X_t^*$, Yt} with parameter freezing. Here, only parameters in the supervised task are updated until the epoch reaches the given maximum value, while minimizing LDNN. To evaluate the effectiveness of the BDD-based fault diagnosis method, the preprocessed target test data is used and fault diagnosis performance can be finally calculated. Besides, the diagnosis results can be visualized in two-dimensional (2D) space with the help of t-stochastic neighboring embedding (t-SNE) to depict diagnosis results.

4.3 Performance Evaluation of the BDD

4.3.1 Description of Data and the DNN Architecture

DGA data used in this study was provided by KEPCO. Due to maintenance and visual inspection costs, 4,000 KEPCO datasets are unlabeled; thus, they are defined as the source data. KEPCO has measured five combustible gases (i.e., H_2, C_2H_2, C_2H_4, C_2H_6, and CH_4) once a year, when the transformers are in a normal state. In abnormal or urgent situations, the gases have been measured once a month or once a week. The target data is IEC TC 10, which has 117 datasets; this is a unique and official open DGA dataset [66]. Despite the different specifications and operating periods of transformers in the KEPCO and IEC TC 10, there is a similarity between them in that the scale and distribution of gas concentrations are comparable. These properties

Table 4-3 Parameters in the DNN

Layer	Activation	Node #	Parameter #
Input	-	5	-
Shared layer 1	ELU	30	35
Shared layer 2	ELU	20	630
Shared layer 3	ELU	15	320
Output1 (supervised)	SM	6	96
Output 2 (unsupervised)	ELU	5	80

support the validity of our approach to make use of the parameter freezing. The IEC TC 10 includes five fault states: PD, T12 (T1 & T2), T3, D1, and D2. Here, it should be noted that the lack of labeling information in KEPCO leads to utilizing these datasets as training data. On the other hand, the IEC TC 10 data, which has true labeling, is divided into two folds: training data (80%) and test data (20%).

Table 4-3 summarizes the DNN architecture. Unlike a convolutional neural network [88], there is a lack of standardized guidelines of DNN architecture. Referring to previous studies of transformers [64, 65, 98], the considered DNN consists of three shared hidden layers and one end layer with two tasks. Each shared hidden layer has 30, 20, and 15 nodes, respectively. Supervised and unsupervised tasks in the end layer has six and five nodes, respectively. Six nodes are for one normal and five fault states. Five nodes are for five dissolved gases. The activation function of the supervised task is the softmax function, presented in (4.6), while the activation function of other layers – including the unsupervised task – is the exponential linear unit for robust and stable computation. The batch size is 200. Epochs in pre-training and re-training are 200 and 20, respectively.

4.3.2 Experimental Results and Discussion

When the Duval's weak supervision is incorporated into other AI-based methods, the first case study aims to demonstrate the effectiveness of BDD, as compared with conventional shallow and deep learning methods. The shallow learning methods include linear support vector machine (L-SVM), SVM with radial basis function (R-SVM), K-nearest neighbors (KNN) algorithm, and a neural network with one hidden layer (1-NN). The deep learning methods include deep autoencoder (DeA) and DNN. For DNN, there are four cases, namely DNN_{FT}^{Non}, DNN_{PF}^{Non}, DNN_{FT}^{Aux}, and the proposed BDD. Superscripts 'Non' and 'Aux' stand for 'without the auxiliary task' and 'with the auxiliary task,' respectively. Subscripts 'PF' and 'FT' stand for 'parameter freezing' and 'fine tuning,' respectively. When the auxiliary regularization term is given, the second case study is proposed to validate the effectiveness of the freezing approach in BDD with respect to the feature space and confusion matrix. In light of the Duval's weak supervision, the third case study is to validate the robustness of the auxiliary regularization task under various percentages of noisy pseudo-labeled source data. The last case study investigates how the fault diagnosis performance of the BDD is sensitive to hyperparameters (learning rate η and a weight α).

(1) Case Study 1. Comparison with Existing Methods

Table 4-4 summarizes the fault diagnosis accuracy of the BDD approach and several AI-based methods with respect to three aspects. To clearly figure out the effects of parameter freezing and auxiliary task, Figure 4-4 pictorially describes the results of DNN_{FT}^{Non}, DNN_{PF}^{Non}, DNN_{FT}^{Aux}, and BDD. The first point investigates the diagnosis performance according to the amount of target training data X. In Table

4-4, 4~80% presents the percentage of target training data in the entire IEC TC 10. The number of each data is 5, 24, 45, 60, and 94, respectively. The result implies that while other AI methods are vulnerable to the amount of labeled data, BDD can be relatively robust even in extremely rare cases for labeled data. This is because AI-algorithms

Table 4-4 Evaluation of the fault diagnosis accuracy

Learning approaches	Methods	Ratio (%) of labeled training data					Unlabeled
		5	10	30	60	80	
Supervised learning	L-SVM	50.0	54.5	63.6	68.2	68.2	X
	R-SVM	40.9	40.9	45.5	50.0	54.5	
	KNN	40.9	40.9	59.1	72.7	72.7	
	1-NN	50.0	40.9	50.0	50.0	59.1	
	DeA	13.6	18.2	22.7	77.2	86.4	
	DNN	50.0	63.6	68.2	81.8	90.9	
Semi-supervised learning	L-SVM	40.9	45.5	63.6	63.6	59.1	O
	R-SVM	50.0	63.6	72.7	77.3	68.2	
	KNN	45.5	50.0	59.1	63.6	68.2	
	1-NN	40.9	63.6	63.6	63.6	72.7	
	DeA	50.0	54.5	68.1	77.2	86.4	
	DNN	50.0	63.6	72.7	77.3	86.4	
Rule-based knowledge learning	L-SVM	45.5	45.5	45.5	50.0	50.0	△
	R-SVM	63.6	72.7	77.3	81.8	81.8	
	KNN	54.5	54.5	59.1	63.6	63.6	
	1-NN	59.1	63.6	63.6	77.3	72.7	
	DeA	63.6	68.2	77.3	81.8	81.8	
	BDD	**86.4**	**90.9**	**95.4**	**95.4**	**95.4**	

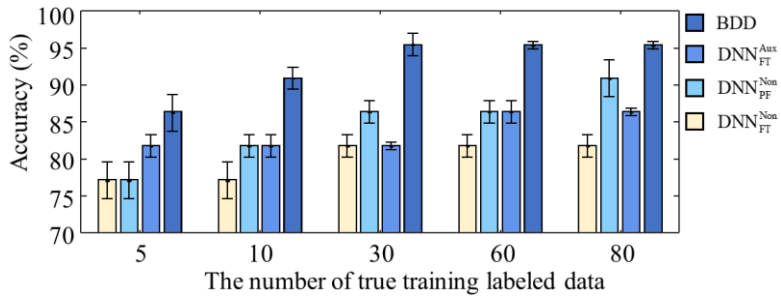


Figure 4-4 Evaluation of the fault diagnosis accuracy in terms of DNN with a transfer learning and an auxiliary task.

usually tend to lose generalized diagnosis performance if an insufficient amount of true labeled data is provided. However, since BDD takes rule-based knowledge in advance, it seems to overcome the limitation of the amount of labeled data.

The second point examines the effects of the learning approach. Performance evaluation is executed for three cases: supervised learning (Model B in Figure 4-1) without unlabeled $\mathbf{X}_s$, semi-supervised learning (Model C in Figure 4-1) with unlabeled $\mathbf{X}_s$, and rule-based knowledge learning (Models D and D* in Figure 4-1) with unlabeled $\mathbf{X}_s$. Here, the last learning is the proposed method that bridges the Duval's pseudo-labeling and AI-algorithms. Since the transfer learning (Models A and A* in Figure 4-1) cannot operate directly in the absence of labeled source data, it is not included in this paper. For AI-based methods other than DNN, the proposed rule-based knowledge learning improves diagnosis performance to some extent; however, that is not always guaranteed. According to previous studies, these algorithms react sensitively to what features are newly added at the stage of preprocessing [17]; thus, only raw DGA data does not yield consistent results. On the contrary, all cases ($\mathrm{DNN}_{\mathrm{FT}}^{\mathrm{Non}}$, $\mathrm{DNN}_{\mathrm{PF}}^{\mathrm{Non}}$, $\mathrm{DNN}_{\mathrm{FT}}^{\mathrm{Aux}}$, and BDD) in rule-based knowledge learning outperform DNN in supervised and semi-supervised learning. It

implies that the combination of rule-based knowledge learning and an auxiliary regularization task or parameter freezing can significantly enhance the diagnosis performance.

The third point is the effects of the auxiliary regularization task and parameter freezing in DNN. In the case of the regularization task, the diagnosis performance of both BDD and DNN_{FT}^{Aux} are larger than those of DNN_{PF}^{Non} and DNN_{FT}^{Non}, respectively, for all amounts of target data. In the case of the parameter freezing, the diagnosis performances of BDD and DNN_{PF}^{Non} are larger than those of DNN_{FT}^{Aux} and DNN_{FT}^{Non}, respectively. We confirm that BDD, has both regularization task and parameter freezing, shows the best diagnosis performance of 95.4%, when 80% of target data is given.

Here, there are two things that should be emphasized. First, the proposed BDD method exhibits the best diagnosis performance of 86.4%, as compared with other algorithms, even if an extremely sparse amount (5%) of labeled training DGA data is used. This is a situation in which only one data sample for each fault (PD, D1, D2, T12, and T3) is given. Second, the diagnosis performance quickly reaches the maximum value, obtained from 80% of true labeled data, even with sparse (30%) true labeled data. Comprehensive analyses of these points infers that the proposed BDD approach can be sufficiently applicable even in engineering situations where only a few labeled DGA data points are provided.

(2) Case Study 2. The effectiveness of parameter freezing via feature space investigation

To deeply understand the effectiveness of parameter freezing approach, 2D feature space is analyzed where hidden features (i.e., estimated labeled data), obtained from the supervised task, are projected via t-SNE. Figure 4-5 illustrates the feature space results of the following three cases: (i) a pre-trained DNN model before re-training (Figure 4-5 (a)), (ii) a re-trained BDD model with parameter freezing (Figure 4-5 (b), the proposed method), and (iii) a re-trained DNN model with a fine tuning approach (Figure 4-5 (c)). To support the qualitative results, the quantitative results of the confusion matrix for each case are presented in Figure 4-6.

In Figure 4-5 (a) and Figure 4-6 (a), three samples in T12 (star-purple) are classified into PD, T12, and T3, respectively; thus, the fault identification of T12 does properly work. In Figure 4-5 (c) and Figure 4-6 (c), only one of them is misdiagnosed into T3; however, two of nine in D2 (plus-red) are diagnosed in D1, which was not observed in both Figure 4-5 (a) and Figure 4-6 (a). For the proposed parameter freezing approach (Figure 4-5 (b) and Figure 4-6 (b)), all samples in T12 (star-purple) and D2 (plus-red) are well classified into corresponding fault states. Therefore, the parameter freezing approach can exhibit much better fault diagnosis performance.

When the parameter freezing approach is adopted, there are two things that should be emphasized: (i) different properties become farther apart and (ii) similar properties become closer together in Figure 4-5. In detail, the fault zone of PD becomes farther from the thermal fault zone ($dt_1 < dt_1'$) and the thermal fault zone of T3 (highlighted with solid-pink) becomes farther from the electrical fault zone ($dt_3 < dt_3'$). On the other hand, the zone of PD (highlighted with solid-orange) becomes closer to the electrical fault zone ($dt_2 > dt_2'$). It is a fact in the field of

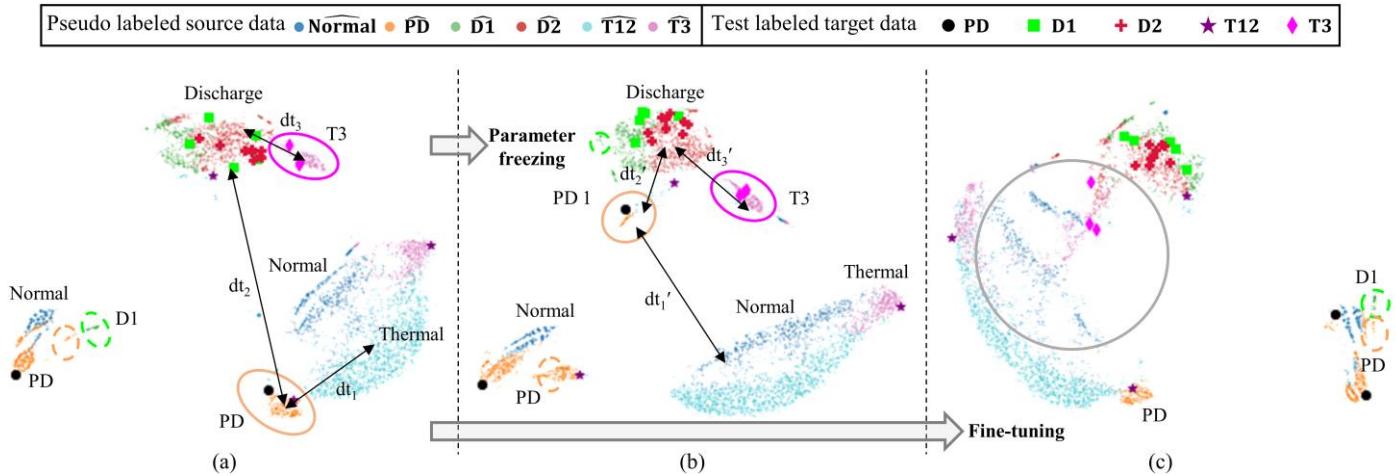


Figure 4-5 2D Feature space using t-SNE of the following three cases: (a) pre-trained DNN model before re-training, (b) re-trained BDD model with parameter freezing (the proposed method), and (c) re-trained DNN model with fine tuning approach.

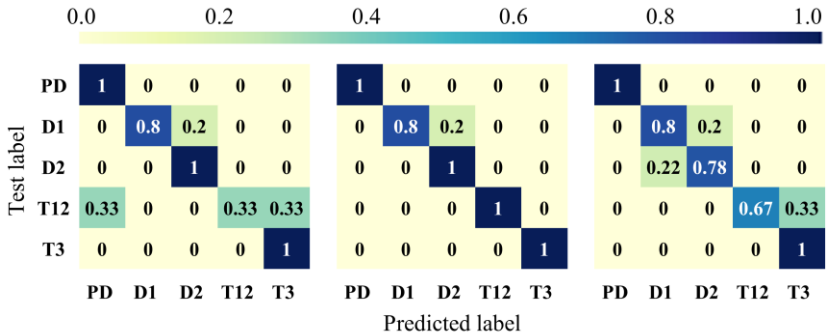


Figure 4-6 Confusion matrix results for the following three cases: (a) pre-trained DNN model, (b) re-trained BDD model with parameter freezing (the proposed method), and (c) re-trained DNN model with fine tuning approach.

transformers that the fault properties of electrical faults and partial discharge are considerably similar [104]. Furthermore, it can be seen that the remote zones of D1 (highlighted with dashed-green) and PD (highlighted with dashed-orange) go back and are clustered into corresponding fault zones. In addition, unlike the parameter freezing approach, in which the fault zones are clearly classified as a whole, several overlapped fault zones (highlighted with solid-gray) are observed for the fine-tuning approach. Therefore, it is implied that the parameter freezing approach in BDD enables it to present better fault identification and clustering performances.

(3) Case Study 3. The effectiveness of the auxiliary regularization term under noisy labeled ratios

In preliminary experiments, it was found that Duval's method had a fault diagnosis accuracy of 76% for the given data. Therefore, we should investigate how the BDD approach is robust to the noisy labeled data. Since the diagnosis performance of the BDD converges to the maximum value from 30% of the true

Table 4-5 Robustness performance under noisy labeled ratios

Learning approaches	Methods	Ratio (%) of the noisy labeled data				
		30	40	50	70	90
Rule-based knowledge learning	L-SVM	45.5	45.5	45.5	54.5	54.5
	R-SVM	77.3	77.3	77.3	63.6	54.5
	KNN	59.1	59.1	59.1	54.5	59.1
	1-NN	63.6	59.1	59.1	59.1	59.1
	DeA	81.8	77.2	77.2	63.6	54.5
	DNN_{FT}^{Non}	81.8	77.3	77.2	63.6	63.6
	DNN_{FT}^{Aux}	81.8	81.8	77.2	77.2	63.6
	DNN_{PF}^{Non}	81.8	81.8	72.7	63.6	54.5
	BDD	**90.9**	**86.3**	**81.8**	**81.8**	**63.6**

labeled data, as shown in Table 4-4, the amount of true labeled data is fixed to 30%. Table 4-5 summarizes the diagnosis results from 30% to 90% of the noisy labeled data in the case of the rules-based knowledge learning. Here, the N% of noisy label means that (N-24) % of pseudo-labeled samples are randomly selected and the health state of the corresponding sample is randomly changed to the remaining health states.

In shallow learning, when the noise is small enough, the accuracy is considerably low; however, it can be seen that the diagnosis performance is not sensitive enough to the change, even though the degree of noise increases. This is because shallow learning needs handcrafted features to avoid the overfitting problem [17, 98]; thus, it already suffers from overfitting with raw data, whether noise is involved or not. These results are similar to the idea that poor students generally have poor grades no matter the teacher's ability. On the other hand, deep-learning-based methods show relatively high performance; however, it can be seen that the accuracy decreases as the noise level increases. When the noise reaches 90%, the performance of both shallow and deep learning looks similar. However, it should be noted that when the noise is up to 70%, DNN_{FT}^{Aux} and the proposed BDD, which have the auxiliary regularization term, show a high accuracy (>75%) and outperform DNN_{FT}^{Non} and DNN_{PF}^{Non}. These results support that the auxiliary regularization term makes the fault diagnosis model less sensitive to the noisy labeled data.

(4) Case Study 4. Effects of Hyperparameters on Fault Diagnosis

The main concern of this case study is to evaluate how the fault diagnosis performance of the proposed BDD varies with hyperparameters (i.e., the learning rate η and weight α). In the gradient descent algorithm, the learning rate η indicates

how much the parameters move to certain minimum points while decreasing the slopes. If an improper learning rate is given, the cost function can deviate from the minimum value, so-called overshooting. The weight α determines the ratio between the supervised and unsupervised tasks. This has an important meaning, as the unsupervised task is used to solve the overfitting problem that may be induced by the Duval's weak supervision. Table 4-6 summarizes the calculated accuracy. Here, depending on the given amount of labeled data, three cases of 10%, 30%, and 60%, are under consideration. As shown in Table 4-6, depending on the learning rate η and weight α, it can be seen that the accuracies at 10%, 30%, and 60% exhibit non-linear trends. It should be noted that maximum accuracies are obtained when η and α are 0.0005 and 0.01, respectively, via Bayesian optimization. Therefore, case studies 1 to 3 were performed with these values.

4.4 Summary and Discussion

A framework for power transformer fault diagnosis, called BDD, was newly proposed to bridge the Duval's rule-based weak supervision and the deep neural network (DNN) approach using dissolved gas analysis (DGA). BDD overcomes problems found in real-world industrial settings, where a large amount of DGA data is unlabeled, and an extremely small size of data is labeled. In this paper, we tested the proposed approach using massive unlabeled Korea Electric Power Corporation (KEPCO) databases and sparse-labeled IEC TC 10 databases. The proposed BDD approach achieved an accuracy of 95.4%, outperforming existing methods. It should be noted that BDD exhibited high accuracy even under situations in which extremely

small labeled target or noisy pseudo-labeled source data were given.

Table 4-6 Effects of hyperparameters on accuracy of BDD

η	0.00005	0.0001	**0.0005**	0.001	0.005	0.01
$Acc_{60\%}$	95.4	90.9	**95.4**	86.4	86.4	86.4
$Acc_{30\%}$	90.9	86.4	**95.4**	86.4	81.8	81.8
$Acc_{10\%}$	86.4	86.4	**90.9**	86.4	81.8	72.7
α	0.001	0.05	**0.01**	0.1	0.5	0.9
$Acc_{60\%}$	90.9	95.4	**95.4**	90.9	86.4	81.8
$Acc_{30\%}$	90.9	90.9	**95.4**	95.4	90.9	81.8
$Acc_{10\%}$	90.9	90.9	**90.9**	90.9	86.4	81.8

Chapter 5

Generative Adversarial Network with Embedding Severity DGA Level

Chapter 5 is dedicated to diagnose the fault severity level as well as the fault types. Although the conventional AI-based approaches, which are described in Chapter 1 and Chapter2, have been done in power transformer fault diagnosis, there are two limitations as follows. The first problem is the absence of severity level for AI-based power transformer diagnosis. The transformer severity level was diagnosed according to the threshold of the DGA concentration, which is estimated through the power utility company and in the academic field. This transformer severity suggests to the field engineer when to measure the DGA data. For example, if the state is caution 1, the next measurement will be measured after 12 months, but in an abnormal condition, DGA data is measured after a month. In addition, most failure modes are diagnosed except for the usual case, using the rule-based or AI-based method, which is described in the previous chapter is utilized. Thus, the existing fault diagnosis methods first obtained DGA data, diagnosed the severity level based on rules, and then diagnosed the fault mode, which could be a cumbersome method. The second drawback is that even if it has a chance to devise a method that can

perform two different tasks at the same time (multi-task learning), there is a problem that the information of the two classes is imbalanced. Specifically, since the severity level is diagnosed based on rules, the severity level information is always annotated in most DGA data, while the fault type is challenging to obtain. This is because it is very difficult to diagnose the complex internal systems of the transformer to identify specific fault types, and it cost a huge economic loss to shut down the transformer for visual inspection.

To address the two different diagnostic tasks simultaneously in an imbalanced condition between severity level and fault type, we propose an auxiliary dual classifier of generative adversarial network for diagnosing fault severity levels and types (GAST). We devise a dual classifier for diagnosing fault types and severity levels by training two labeled DGA dataset. The industrial DGA data, however, that should contain two different label information, usually only has a severity level, and fault types are unlabeled, so the two tasks cannot be trained simultaneously with a conventional supervised learning approach. For such an imbalanced problem in transformer fault diagnosis, sampling techniques were developed, but it is known that these sampling techniques only generate similar data and it is difficult to generate various data. Therefore, we tackle two imbalanced conditions through generative adversarial network (GAN), which achieved great performance in today's data generation.

The rest of this paper is organized as follows. Section 5.1 outlines backgrounds of neural network and generative adversarial network. Sections 5.2 and 5.3 demonstrate the proposed method and experimental results, respectively. Finally, the conclusions of this work are provided in Section 5.4.

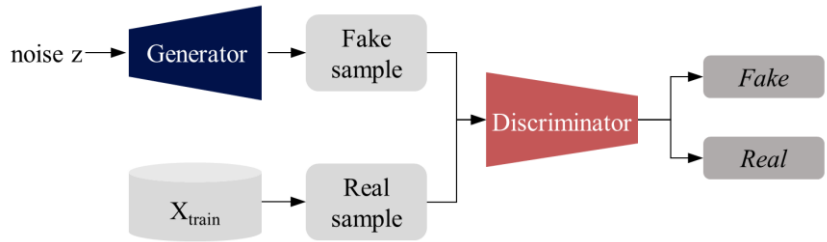


Figure 5-1 Architecture of GAN

5.1 Backgrounds of Generative Adversarial Network

GAN is a framework devised by Ian Goodfellow to generate data [105]. Instead of single generative network, it contains a generative network G and a discriminative network D forming an adversarial framework, as shown in Figure 5-1. The term "adversarial" is that the generator is designed to confuse the discriminator by making the fake data as similar as real data. In contrast, the discriminator is designed to distinguish between fake and real data correctly.

The generator consists of input and output, where random noise vectors z (usually normal or uniform distribution, p(z)) are imported into the input and output fake samples via generator $G(z;\theta_g)$ where $\theta_g$indicates the parameters of generator. And, the discriminator D is inputted by real data *x* or G(*z*) to distinguish real from fake data by $D(x;\theta_d)$ or $D(G(z);\theta_d)$ where θ_d denotes the parameters of discriminator. Besides, it can be interpreted that D(x)=1 when x~p(x) and D(x)=0 when x was generated from G. More formally, these objective loss function can be trained by minimax two-player game expressed as:

$$\min_G \min_D \mathbb{E}_{x\sim p_{data}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim p_{noise}(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right] \qquad (5.1)$$

Eq (5.1) is solved by optimizing each parameter, θ_d and θ_g, by gradient updates as follows:

$$\nabla\theta_d \frac{1}{m}\sum_{i=1}^{m}\left[\log D\left(x^{(i)}\right)+\log\left(1-D\left(G\left(z^{(i)}\right)\right)\right)\right] \tag{5.2}$$

$$\nabla\theta_g \frac{1}{m}\sum_{i=1}^{m}\left[\log\left(1-D\left(G\left(z^{(i)}\right)\right)\right)\right] \tag{5.3}$$

where (5.2) and (5.3) used a minibatch algorithms, *m* indicates a minibatch size, and the first term of (5.1) was ignored in (5.3) because G only takes noise vector z, not a real data x.

Goodfellow et al. [105] described that if D and G are guaranteed to have sufficient parameters and computational amount, (5.1) can find the global minimum value to generate realistic data. However, many research areas have shown that GAN is unstable in many applications.

5.2 GANES based Fault Diagnosis

In this section, we propose a diagnosis technique for power transformer that uses the auxiliary classifier with generative adversarial network for embedding severity level and fault types (GANES). This approach is designed to extract features and fault diagnosis by simultaneously learning severity level and fault types under imbalanced dataset. As shown in Figure 5-2, the details of the loss function and architecture are described in next section.

5.2.1 Training Strategy of GANES

(1) Training of the Discriminator: Embedding Severity DGA Level for Supervised and Unsupervised Learning

Although ACGAN's discriminator is known to be able to classify classes and distinguish between fakes and reals, another task, semi-supervised learning, is not attempted in fault diagnosis and has not been noticed, especially in transformer fault diagnosis. Moreover, the classification of multiclass for two tasks (severity level and fault types) is challenging, rather than the classification of multiclass for single task. Therefore, the training of GANES's discriminator is specifically described.

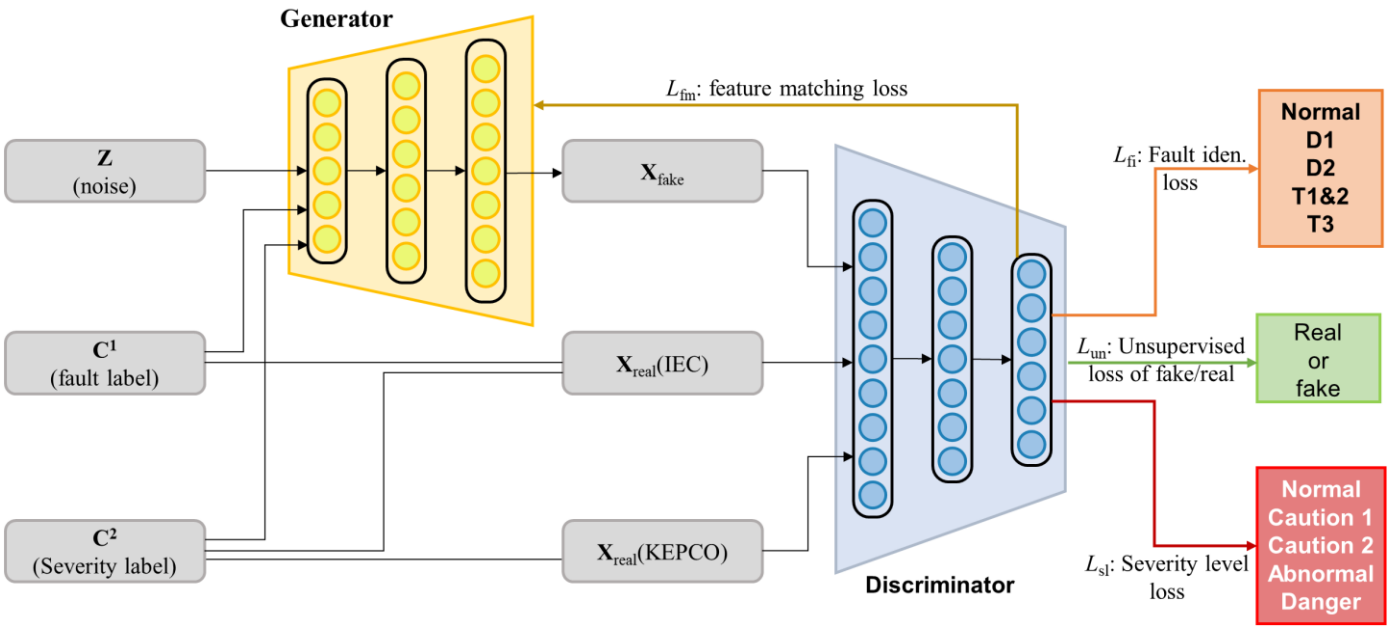


Figure 5-2 Architecture of GANES

Firstly, the most important role in the discriminator is to identify the $f(x)$ or logits that has been transformed into a high-level feature through the hidden layer of the input X to match the true fault type, where $f(x)$ is organized with discriminator's parameter θ_{dis}. To conduct identification $p_{\text{model}}(y = j \mid x)$, we applied softmax function to the f(x) for transforming K (normal, PD, DL, DH, TL, TH) dimensional vector into probability vector as follows:

$$p_{\text{model}}\left(y = j \mid x\right) = \frac{\exp\left(f\left(x\right)_j\right)}{\sum_{k=1}^{K} \exp\left(f\left(x\right)_k\right)} \tag{5.4}$$

where j is the one of the health states and to minimize the discrepancy between $p_{\text{model}}(y = j \mid x)$ and true label Y, the cross-entropy loss function is applied for supervised learning of the identification task as follows:

$$L_{\text{it}} = -\mathbb{E}_{X,Y \sim p_{data}(x,y)} \log p_{\text{model}}\left(y \mid x, y < K+1\right) \tag{5.5}$$

here, instead of $p_{\text{model}}(y = j \mid x)$, we have to consider the unlabeled data, so if it is less than K+1 that extends one dimension, we identify as $p_{\text{model}}(y = j \mid x,\ y < K{+}1)$.

The second role of the discriminator for diagnosing a severity level is newly added in this study. The second task of classifying a severity level of normal, cautious 1,2, abnormal, and danger is similar to the identification task but plays an essential role in generating various DGA data. The reason is that most of the DGA concentration value (ppm) of the actual failure transformer (IEC TC 10 database) is almost lies at a dangerous level in terms of the severity level, so it is challenging to generate various fault types in the remaining severity levels, also known as mode collapse. A detailed model is as follows:

$$p_{\text{model}}\left(y=l\mid x\right)=\frac{\exp\left(f\left(x\right)_{l}\right)}{\sum_{s=1}^{S}\exp\left(f\left(x\right)_{s}\right)} \tag{5.6}$$

where l indicates the one of the severity levels, S represents the number of severity levels, and $f(x)$ is same as the outcome of the identification task, which means it shares the same parameters. The loss function of diagnosing a severity level task is also cross-entropy as follows:

$$L_{\text{st}}=-\mathbb{E}_{X,Y\sim p_{data}(x,y)}\log p_{\text{model}}\left(y\mid x\right) \tag{5.7}$$

here, the condition of $y<S+1$ is omitted because the severity level is labeled in all DGA data.

The last task is the principal role of the discriminator, which distinguish between real and fake data. We sampled the real data $X \sim p_{\text{data}}(x)$ from KEPCO and IEC TC 10. The goal of the loss function is to minimize the following function:

$$\begin{aligned}L_{\text{un}}=&-\mathbb{E}_{X\sim p_{data}(x)}\log\left[1-p_{\text{model}}\left(y=K+1\mid x\right)\right]\\&-\mathbb{E}_{Z\sim p_{noise}(z)}\log\left[p_{\text{model}}\left(y=K+1\mid x\right)\right]\end{aligned} \tag{5.8}$$

where we defined as unsupervised loss L_{un} because it does not consider the labeled information, the first term represents the negative log-likelihood of real data x belongs to any other labeled data and the second term indicates the negative log-likelihood of fake data belongs to "generated" class of $y = K+1$. In addition, $1-p_{\text{model}}(y=K+1|x)$ corresponds to $D(x)$ in the original GAN framework and it can be substituted as follows:

$$L_{\text{un}}=-\left\{\mathbb{E}_{X\sim p_{data}(x)}\log D\left(x\right)+\mathbb{E}_{z\sim noise}\log\left(1-D\left(G\left(z\right)\right)\right)\right\} \tag{5.9}$$

Through the Eq. (5.9) of the discriminator, the imbalanced two-class problem is balanced from the generator by generating a real-like fake data with different

severity including fault mode.

(2) Training of the generator

The purpose of the GAN's generator for the transformer fault diagnosis is to generate virtually real data (fake) that could overcome the imbalanced issues between fault types and severity levels in the real industry. To train the generator, the original GAN's loss function of (5.3) could be used, but to extract health features of fault types and severity levels, we modified generator's loss term in according to the feature matching loss term [106]. Originally, the feature matching loss term is developed to stabilize the mode collapse of GAN. Additionally, we expected that the health feature, which includes both fault types and severity levels, can be visualized in a low-dimensional space more clearly identified and explained. Specifically, the objective of feature matching loss function is defined as:

$$L_{\text{gen}} = \left\| \mathbb{E}_{X \sim p_{data}(x)} f(x) - \mathbb{E}_{Z \sim p_{noise}(z)} f(G(z)) \right\|_2^2 \tag{5.10}$$

by matching the high-level features of real and fake data, L_{gen} could regularize the generator to find the underline distribution of the real data [106].

5.2.2 Overall procedure of GANES

Figure 5-3 illustrates the flowchart of the proposed GANES-based fault diagnosis method. There first step is the data acquisition which obtains the real dataset and samples a noise data from a multi-normal distribution. An unlabeled fault type with the severity level $\{\mathbf{X}_{\text{un+sl}}\}$, a labeled fault type with the severity level $\{\mathbf{X}_{\text{la+sl}}\}$, labeled fault target data $\{\mathbf{Y}_{\text{la}}\}$, severity levels with unlabeled fault target data {Y},

and severity levels with labeled fault target data {} is described. After defining the dataset, preprocess is applied in DGA data {}.

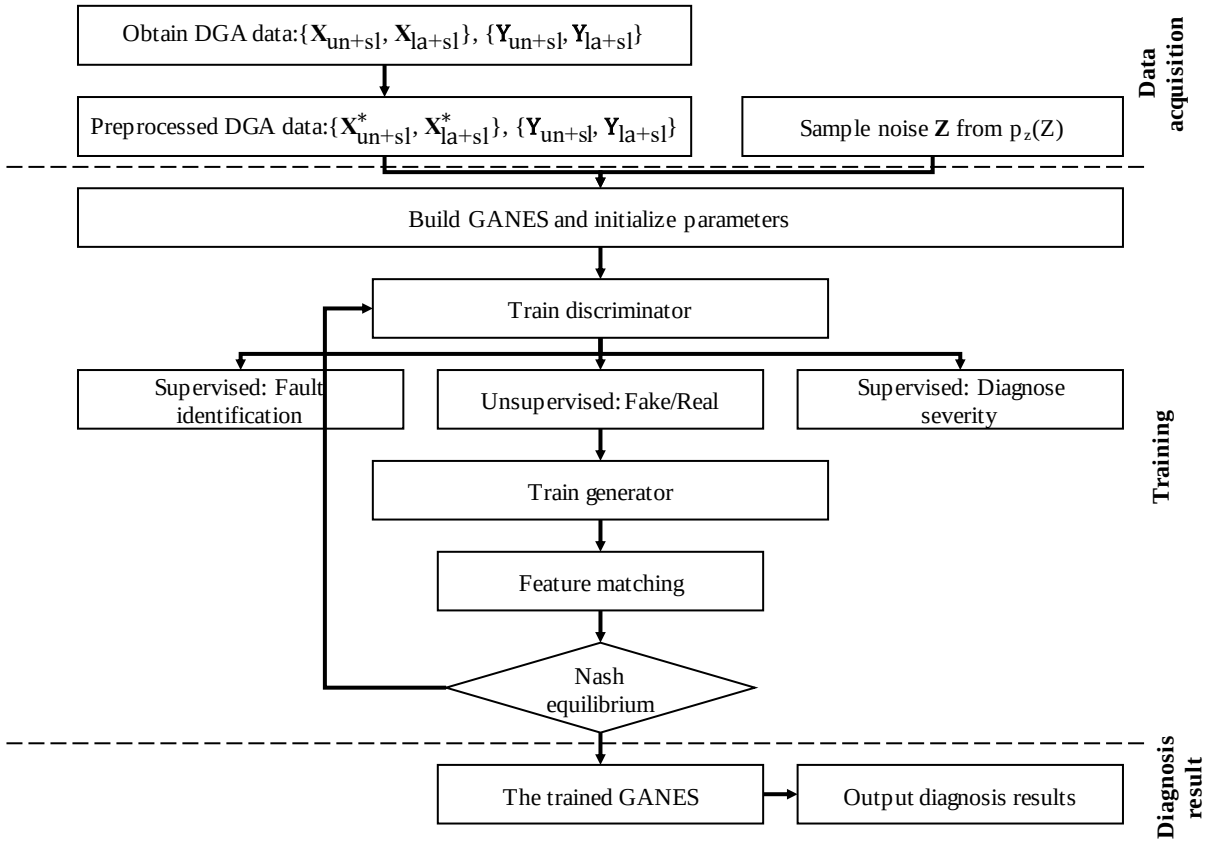


Figure 5-3 Flowchart of the GANES

Table 5-1 Parameters in the architecture of GANES

Discriminator			
Layer	Activation	Node #	Parameter #
Input	-	5	-
Shared layer 1	Softplus	30	35
Shared layer 2	Softplus	20	630
Shared layer 3	Softplus	15	320
Identification layer	Softmax	6	96
DGA lever layer	Softmax	5	80
Generator			
Layer	Activation	Node #	Parameter #
Input	-	7	-
Layer 1	Softplus	10	35
Layer 2	Softplus	10	630
Layer 3	Softplus	5	320

The next step is the training session where it optimizes the GANES-based fault diagnosis model. Firstly, parameters in GANES, as described in Table 5-1, are randomly initialized. For discriminator's loss function it consists three tasks: supervised learning-based fault identification, severity diagnosis, and unsupervised learning-based discriminating real and fake data. Notably, it can be expressed as follows:

$$L_{\text{dis}} = \frac{1}{2}\left(L_{\text{su}} + L_{\text{un}}\right) \tag{5.11}$$

where L_{un} is same as (5.9), and L_{su} is a summation of fault identification task and a severity level task as follows:

$$L_{su} = \frac{1}{2}\left(L_{\text{it}} + L_{\text{st}}\right) \tag{5.12}$$

in order to balance a number of data between unlabeled and labeled data, we augmented noise to the original fault labeled data, $\mathbf{X} = \mathbf{X} + \epsilon$, so changing half of our labeled data set consist of labeled fault data with the severity level and half of it is unlabeled with the severity level. Then, the generator trains according to the

feature matching loss function L_{gen}. When the loss function reaches the Nash equilibrium iteration will be finished. Finally, to test the result, GANES-based model runs a test data to investigate the diagnosis accuracy and qualitative result is also derived.

5.3 Performance Evaluation of GANES

5.3.1 Description of Data

DGA data used in this study was provided by KEPCO. The company has stored DGA data from numerous transformers in South Korea for three decades. KEPCO has measured five combustible gases (i.e., H2, C2H2, C2H4, C2H6, and CH4) once a year, when the transformers are in a normal state. In abnormal or urgent situations, the gases have been measured once a month or once a week. Due to maintenance and visual inspection costs, 4,000 KEPCO datasets are unlabeled, however, with the help of KEPCO's severity plan we could annotate DGA level on each sample. We use IEC TC 10, which has 117 datasets and we ignored 20 datasets of communicated OLTC which has different spec on KEPCO's transformer. The IEC TC 10 includes five fault states: PD, T12 (T1 & T2), T3, D1, and D2.

5.3.2 Outlines of Experiments

Since GAN, which is difficult to learn and stabilize, is applied for the first time in transformer fault diagnosis, we first investigated preliminary experiments whether it is possible to learn DGA data. The preliminary experiment was divided into two experiments according to the fault label information. Firstly, we tested various

objective function optimization based GANs such as Wassestein GAN (WGAN), WGAN with gradient penalty (WGAN-GP), least square GAN (LSGAN) and adversarial autoencoder (AAE) to investigate a stabilization and generative performance in a large amount of unlabeled DGA data. Specifically, these various GANs are summarized in Table 5-3 with their objective loss function and parameters. It should be noted that parameters are kept same as the proposed method. Secondly, we experimented a various condition-based objective function optimization GANs (condition-based GAN (CGAN), semi-supervised GAN (SGAN), and auxiliary classifier GAN (ACGAN)) with a small amount of labeled fault data (IEC TC 10). Detailed parameters and the architectures are summarized in Table 5-4. In addition, their conceptual architecture are also shown in figure [].

Table 5-2 Comparative ACGAN for GANES

Methods	Discriminator loss	Generator loss
$\text{ACGEN}_{\text{OG}}^{\text{FI}}$	$L_{\text{dis}} = L_{\text{FI}} + L_{\text{un}}$	$L_{\text{gen}} = L_{\text{gan}}$
$\text{ACGAN}_{\text{OG}}^{\text{FI+SL}}$	$L_{\text{dis}} = \alpha L_{\text{FI}} + (1-\alpha) L_{\text{SL}} + L_{\text{un}}$	$L_{\text{gen}} = L_{\text{gan}}$
$\text{ACGAN}_{\text{FM}}^{\text{FI}}$	$L_{\text{dis}} = L_{\text{FI}} + + L_{\text{un}}$	$L_{\text{gen}} = L_{\text{FM}}$
GANES	$L_{\text{dis}} = \alpha L_{\text{FI}} + (1-\alpha) L_{\text{SL}} + L_{\text{un}}$	$L_{\text{gen}} = L_{\text{FM}}$

Table 5-3 Various GANs for unlabeled DGA data

Method	Parameter	Gen.	Dis.	Model	Objective function
AAE	# Layers	2	3	5	$L_D^{AAE} = -\mathbb{E}_{x\sim p_{data}}\left[\log D(x)\right] - \mathbb{E}_{z\sim p(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]$
	# Neurons	15/10	15/15	15/10/15/15	$L_G^{AAE} = -\mathbb{E}_{z\sim p(z)}\left[\log D\left(G(z)\right)\right]$
GAN	# Layers	2	3	5	$L_D^{GAN} = -\mathbb{E}_{x\sim p_{data}}\left[\log D(x)\right] - \mathbb{E}_{z\sim p(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]$
	# Neurons	15/10	15/15	15/10/15/15	$L_G^{GAN} = \mathbb{E}_{z\sim p(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]$
LSGAN	# Layers	2	3	5	$L_D^{GAN} = \frac{1}{2}\mathbb{E}_{x\sim p_{data}}\left[\left(D(x)-b\right)^2\right] + \frac{1}{2}\mathbb{E}_{z\sim p(z)}\left[\left(D\left(G(z)\right)-a\right)^2\right]$
	# Neurons	15/10	15/15	15/10/15/15	$L_G^{GAN} = \frac{1}{2}\mathbb{E}_{z\sim p(z)}\left[\left(D\left(G(z)\right)-c\right)^2\right]$
WGAN	# Layers	2	3	5	$L_G^{WGAN} = -\mathbb{E}_{x\sim p_{data}}\left[D(x)\right] + \mathbb{E}_{z\sim p(z)}\left[D\left(G(z)\right)\right]$
	# Neurons	15/10	15/15	15/10/15/15	$L_G^{WGAN} = -\mathbb{E}_{z\sim p(z)}\left[D\left(G(z)\right)\right]$
WGAN GP	# Layers	2	3	5	$L_D^{WGANGP} = L_D^{WGAN} + \lambda\mathbb{E}_{(x,z)\sim p(x,z)}\left[\left(\left\|\nabla D\left(ax-\left(1-\alpha G(z)\right)\right)\right\|-1\right)^2\right]$
	# Neurons	15/10	15/15	15/10/15/15	$L_G^{WGANGP} = L_G^{WGAN}$
DRAGAN	# Layers	2	3	5	$L_D^{GAN} = -\mathbb{E}_{x\sim p_{data}}\left[\log D(x)\right] - \mathbb{E}_{z\sim p(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]$
	# Neurons	15/10	15/15	15/10/15/15	$L_G^{GAN} = \mathbb{E}_{z\sim p(z)}\left[\log\left(1-D\left(G(z)\right)\right)\right]$

Table 5-4 Various supervised GANs

Method	Objective function	Parameter	Gen.	Dis.	Model
ACGAN	$L_D^{ACGAN} = L_D^{GAN} - \mathbb{E}_{x\sim p_{data}}\left[p\left(y=k \mid x\right)\right] - \mathbb{E}_{z\sim p(z)}\left[p\left(y=k \mid G\left(z\right)\right)\right]$	# Layers	2	4	5
	$L_G^{ACGAN} = L_G^{GAN} - \mathbb{E}_{z\sim p(z)}\left[p\left(y=k \mid G\left(z\right)\right)\right]$	# Neurons	15/10	15/15/5/2	15/10/15/15/5/2
SGAN	$L_D^{SGAN} = L_D^{WGAN} - \mathbb{E}_{x\sim p}\left[p\left(y=k \mid x, k<c+1\right)\right]$	# Layers	2	3	5
	$L_G^{SGAN} = L_G^{WGAN} + \left\|E_{x\sim p} f\left(x\right) - \mathbb{E}_{z\sim p} f\left(G\left(z\right)\right)\right\|^2$	# Neurons	15/10	15/15	15/10/15/15
CGAN	$L_D^{GAN} = -\mathbb{E}_{x\sim p_{data}}\left[\log D\left(x,c\right)\right] - \mathbb{E}_{z\sim p(z)}\left[\log\left(1-D\left(G\left(z\right),c\right)\right)\right]$	# Layers	2	3	5
	$L_G^{GAN} = \mathbb{E}_{z\sim p(z)}\left[\log\left(1-D\left(G\left(z\right),c\right)\right)\right]$	# Neurons	15/10	15/15/5	15/10/15/15/5

After we demonstrate the stability of various GANs, the proposed method of embedding severity DGA level experiments were conducted with four case studies. The first comparative study aims to validate the effectiveness of the auxiliary severity diagnosis task for stabilizing the performance in SGAN-based fault diagnosis model. We consider the following three models: 1) $SGAN_{FM}$, 2) $SGAN_{SL}$ and 3) SGAN. Notations 'FM' stands for 'feature matching' which substitute the generator's loss function, 'SL' stands for 'severity level' that puts an auxiliary classifier task at the discriminator, and SGAN indicates that none of the above techniques are implemented. A detailed description is shown in Table 5-2. Next, the diagnosis performance of identifying the fault types is evaluated in the second study. Here, we also compared the above three methods and additionally compared with the existing fault diagnosis methods such as semi-supervised based Linear-SVM (LSVM), RBF-SVM (RSVM), KNN, one-hidden layer of neural network (1-NN), deep autoencoder (DeA). Besides, we investigate the robustness under extremely small amount of labeled data. In order to investigate thoroughly the performance degradation according to the number of label data, not only the semi-supervised method but also the supervised learning method were compared. Finally, to give more clarity of the severity DGA level effects, we visualize the features by projecting into low dimensional space by using tSNE.

5.3.3 Preliminary Experimental Results of Various GANs

(1) Stabilization Analysis 1. Unlabeled DGA Data (KEPCO)

Fig. 4 indicates the quantitative results of various unsupervised GAN's

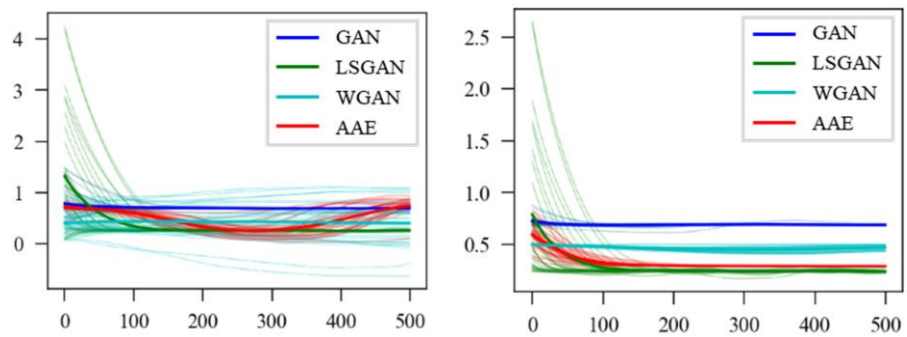


Figure 5-4 Unsupervised various GANs loss

discriminator and generator loss function, respectively. Although there is no absolute metrics to evaluate discriminator, WGAN and WGAN-GP show the best performance since those of discriminator loss function closely reaches up to 0.5. The discriminator loss of WGAN and WGAN-GP that reaches to 0.5, indicates that the probability of *x* being real is 1/2 [], when the optimized discriminator is

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} = \frac{1}{2}, \ \left(\text{when } p_d(x) = p_g(x)\right) \tag{5.13}$$

However, the discriminator's loss value that achieved a 0.5 does not validate that the GAN's performance is good because the generator can generate the same DGA data to deceive the discriminator (mode collapse). LSGAN is the closest to 0, followed by WGAN, the second-lowest in the generator loss function. For WGAN-GP and GAN, the generator's loss function is slightly higher but exhibits an unfluctuating and stable loss pattern, as LSGAN or WGAN shown. Despite the fact that the GANs generator loss is stable or close to zero, like the discriminator loss, it does not also guarantee the performance of the GANs.

Therefore, due to the lack of quantitative indicators, the DGA data and actual

Table 5-5 Generated DGA samples from various GANs and real DGA data

Generated/sampled	H_2	C_2H_2	C_2H_4	C_2H_6	CH_4
Real	100	20	100	20	100
	5	5	5	5	5
	30	6	30	6	30
	4	5	4	5	4
AAE	10	20	100	20	100
	16	512	5	51	2
	346	0	30	833	303
	1	225	4	5	45
GAN	100	20	100	20	100
	5	5	5	5	5
	30	63	30	6	30
	4	0	4	5	4
LSGAN	100	20	100	20	100
	5	5	5	5	5
	30	6	30	6	30
	4	52	4	5	4
WGAN	100	310	36	420	10
	515	4	5	5	5
	2	1003	20	6	30
	46	3	6	55	42
WGAN GP	100	0	100	20	100
	54	523	0	53	5
	30	2	30	612	350
	4	5	4	5	4
DRAGAN	100	23	1003	20	1050
	54	5	3	6	5
	30	6	0	100	30
	6	19	523	0	4

data generated from the generator are displayed in Table 5-5. Although we confirmed that the virtually generated data were similar to the actual DGA data, it is difficult to determine which method performs the best. Still, it can be seen from the table that WGAN learns faster than other discriminator's methods and has achieved the

minimum loss in the generator.

(2) Stabilization Analysis 2: Labeled DGA Data (Fault types of IEC TC 10 and Severity Level with KEPCO)

Quantitative results of various supervised GAN discriminator and generator loss functions are depicted in Figure 5-5. For discriminator loss, training ACGAN is higher than CGAN, and SGAN. After 300 around epochs later, ACGAN is less than other GANs, representing that ACGAN's generator is more likely to generate virtual fake DGA data. Moreover, in the generator's loss function, ACGAN is gradually stabilized as generator optimized, as shown in Figure 5-5 (b), whereas SGAN and CGAN keep the same loss, which could be interpreted as the generator is not working or generating similar data. In the case of CGAN's generator, as it diverges, it may be considered that training is not being optimized. To investigate the generated fake DGA data according to the fault labeled information, table 6 reveals several fake samples compared with an IEC TC 10 database. It should be noted that since the generated fake data keeps generating the same or similar fault data, which is occurring a mode collapse, supervised learning-based GAN failed to learn a small number of fault labeled data.

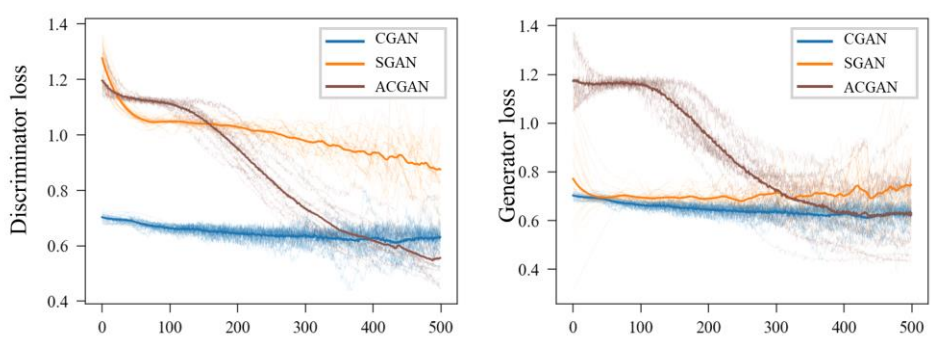


Figure 5-5 Supervised various GANs loss of IEC TC 10 data

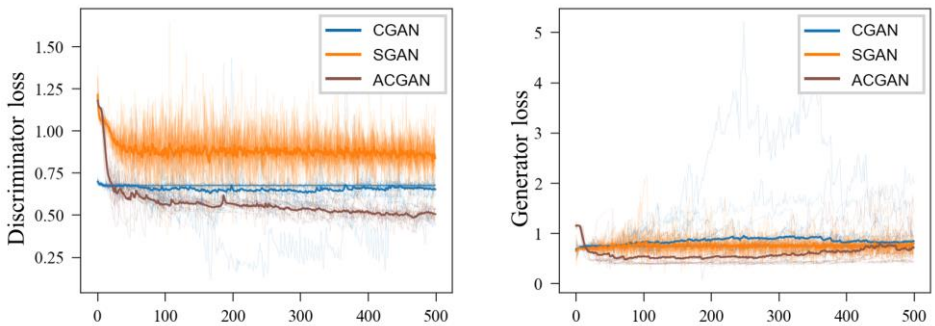


Figure 5-6 Supervised various GANs loss of severity level informed in KEPCO

On the other hand, although there is no fault types in KEPCO, the result of learning a large amount of KEPCO data given a severity is shown in the figure. The result indicates that if certain label information is assigned to a large amount of data, GAN can be more stable for learning; in particular, ACGAN outperforms. excellent.

5.3.4 Experiments for the Effectiveness of Embedding Severity DGA Level

From the previous preliminary experiments, it was found that there is a limitation to evaluating the performance of GAN through the loss function or the generated DGA data. Therefore, to evaluate the GANES proposed in this study, the accuracy of fault diagnosis was investigated. However, as an extension of the previous experiment, we first performed a stabilization experiment according to the feature matching that constitutes GANES and the severity level additionally entered into the discriminator, and then the accuracy and characteristics were demonstrated.

(1) Case Study 1: Stabilization

Figure 5-7 shows the quantitative results of the $SGAN_{FM}$, $SGAN_{SL}$, SGAN and

GANES, to investigate the effectiveness of feature matching and auxiliary task of severity level that we proposed. Figure 5-7 depicts three major loss function: (a) represents a loss function of generator, (b) indicates the unsupervised loss function which discriminates the fake and real, and (c) investigates the supervised loss function of fault identification. At about 150 epochs, the result of Figure 5-7 (a) shows the stabilized generator's loss function of GANES. On the other hand, the generator's loss function of $SGAN_{FM}$, $SGAN_{SL}$, and SGAN keep increasing when they are trained. We could expect that the generator of GANES may more stable than other comparison methods. Moreover, for Figure 5-7 (b) is the discriminator loss function of $SGAN_{FM}$, $SGAN_{SL}$, and SGAN, which are below than 0.5, while GANES closely reaches to the 0.5. It implies that the distribution of GANES generator is more likely to follow the real DGA data. Finally, as shown in Figure 5-7 (c) of the supervised loss function of fault identification, we could confirm that the GANES achieves the best performance than other comparison methods. A detailed fault identification accuracy is described in following section.

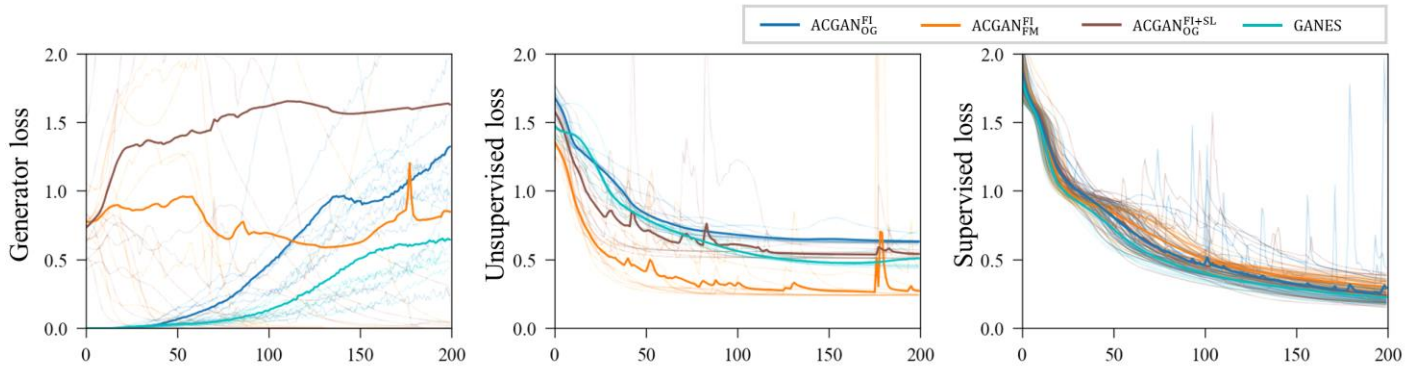


Figure 5-7 Multi-task ACGAN loss

(2) Case Study 2: Fault Diagnosis Accuracy

Figure 5-8 summarizes the fault type diagnosis performance of GANES, ACGAN1, ACGAN2, and ACGAN3. The accuracy of ACGAN2 and ACGAN3 is the same at 89.7, but the diagnosis result of the fault type is different. It is certain that each method improves diagnostic performance, but it is difficult to interpret which one is superior. In addition, the accuracy of severity level estimation for GANES and ACGAN3 is 95.6% and 93.6%, respectively. Furthermore, the accuracy of GANES with both FM and SL is 93.1%, which is superior to the other three methods. Therefore, we could expect that SL and FM complement each other for GANES and improve the diagnosis performance.

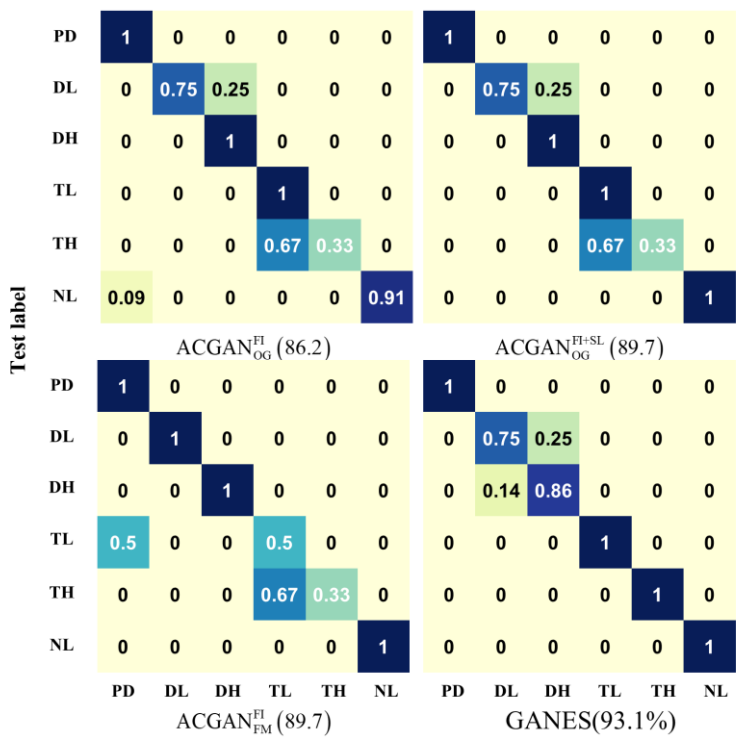
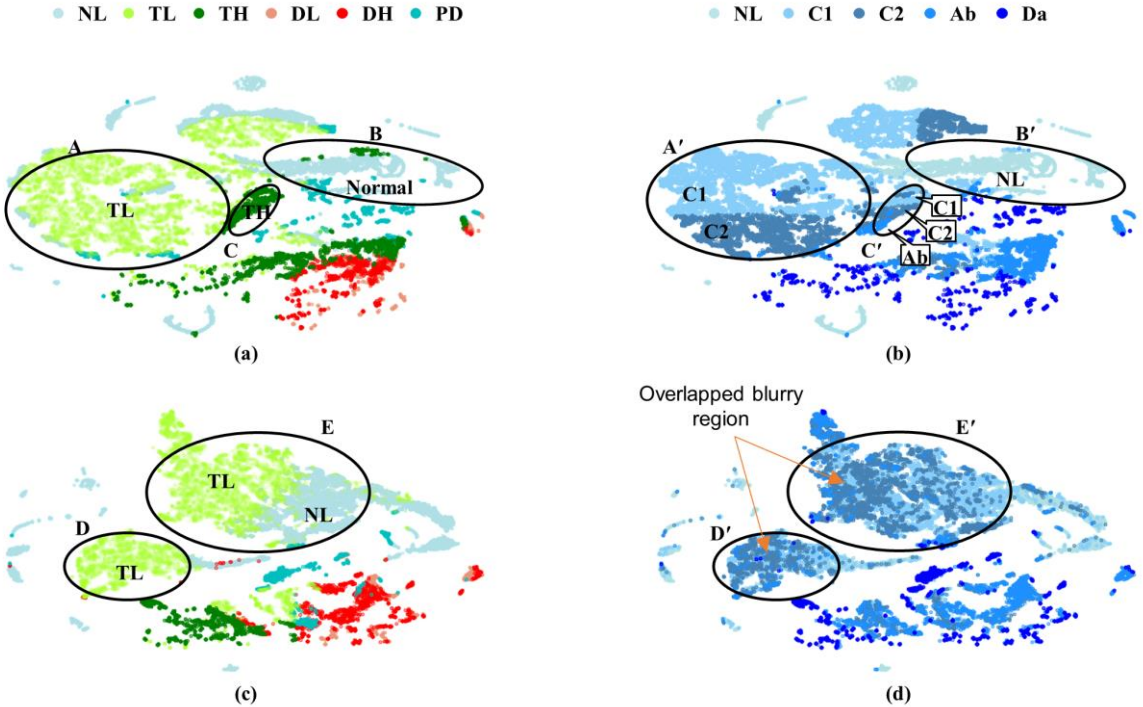


Figure 5-8 Fault diagnosis accuracy of fault types

(3) Case Study 3: Visualization of Expected Features

In the previous two case studies, it was examined whether the loss function was quantitatively stabilized or the fault diagnosis performance for identifying fault types was improved by additionally learning the severity DGA level. In this study, we investigate the qualitative effectiveness of inserting severity DGA level for diagnosing the fault types. Figure [] represents that GANES or ACGAN3 method could distinguish not only the failure mode but also the severity level features simultaneously. Figure 8 (a) is represented by failure mode, and (b) is represented by severity. Both are low-dimensional projections of features extracted by the GANES method and are expressed according to their respective viewpoints. On the other hand, Figures (c) and (d) are SGANs that have not been studied severity, and each plot indicates the fault mode and severity, respectively.


NL TL TH DL DH PD
A
TL
B
Normal
TH
C
(a)
NL C1 C2 Ab Da
A′
C1
C2
B′
NL
C1
C2
C′
Ab
(b)
E
TL
NL
D
TL
(c)
Overlapped blurry region
E′
D′
(d)

5.4 Summary and Discussion

In this study, we develop a generative adversarial network that includes severity DGA levels. In actual industrial transformers, fault identification and severity estimation are essential to decide on a maintenance plan that can inform decisions about whether the system can operate normally or if repair or replacement is necessary. However, traditional artificial intelligence-based methods trained only by labeled fault types do not include severity levels. Therefore, engineers must apply different rule-based approaches to estimate severity. Moreover, because fault modes are difficult to obtain in an industrial environment, rule-based methods simply annotate severity, resulting in unbalanced labeled data problems between the two states. Therefore, this research proposes a generative adversarial network with an embedding severity (GANES) DGA level. As a fundamental approach to alleviating the imbalanced problem between two classes of labeled fault types and severity levels, an auxiliary classifier of the generative adversarial network (ACGAN) was applied. The proposed method is demonstrated by studying massive Korea Electric Power Corporation (KEPCO) and IEC TC 10 databases. The results show that the proposed method not only outperforms conventional AI-based methods but also extracts both fault types and severity levels.

Sections of this chapter have been published or submitted as the following journal articles:

1) **Sunuwe Kim,** Soo-Ho Jo, Heonjun Yoon, Yong Chang Shin, and Byeng D. Youn, "A Graded Phononic Crystal with Decoupled Double Defects for Broadband Energy Localization," International Journal of Mechanical Sciences, Vol. 183, pp. 105833, 2020.

Chapter 6

Conclusion

6.1 Contributions and Significance

The proposed research in this doctoral dissertation aims at overcoming practical issues in industrial power transformers via deep learning based methods. This doctoral dissertation is composed of three research thrusts: (1) semi-supervised autoencoder with auxiliary detection task to extract health feature space; (2) bridging a rule-based Duval method and deep learning-based DNN; and (3) a generative adversarial network with embedding severity level. It is expected that the proposed research offers the following potential contributions and broader impacts in industrial power transformer fault diagnosis.

Contribution 1: Extraction of the Health Feature Space to Visualize a Degradation Trendability

This doctoral dissertation suggests 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). This is the first attempt to diagnose real-world

power transformers using a large amount of DGA data. By using the industrial DGA dataset, the proposed SAAT extracts the health degradation properties as well as to identify normal and thermal/electrical fault types. In addition, by directly visualizing health features without transformation or dimension reduction, the proposed 2D HFS can pictorially demonstrate the monotonic health trendability of transformers.

Contribution 2: Learning from Even a Weak Teacher via Bridging Rule-based Duval Weak Supervision and a Deep Neural Network for Diagnosing Transformers

This doctoral dissertation provides a new framework, named BDD, that bridges Duval method with a deep neural network approach for transformer fault diagnosis. The main concept of our approach – incorporating a rule-based method into an AI-based method – is newly proposed in the field of transformer fault diagnosis. Besides, an auxiliary unsupervised loss task is added to regularize the rule-based method's partially incorrect knowledge. After that, a parameter transfer learning approach is incorporated into DNN to deliver rule-based knowledge from the pseudo-labeled source data to the labeled target data. The results indicates that

Contribution 3: Improvement of Transformers Fault Diagnosis by Elucidating Fault Types with Severity Levels

This doctoral dissertation aims to diagnose not only fault types but also severity

levels of transformers. To the best of authors' knowledge, the unique contributions of this study are two-fold. First, since labeled information of severity level and fault types are imbalanced, an auxiliary classifier of the generative adversarial network is newly proposed for balancing as well as diagnosing two different conditions. Second, in the low dimensional space, the extracted health features elucidate severity levels as well as fault types properties.

6.2 Suggestions for Future Research

Although the technical advances proposed in this doctoral dissertation successfully address practical issues in the industrial field of power transformer fault diagnosis, there are still several research topics that further investigations and developments are required to bring deep learning-based fault diagnosis method into an alternative solution for industrial transformer PHM. Specific suggestions for future research are listed as follows.

Suggestion 1: Enhancement of the Health Feature Space

For research thrust 1, 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.

Suggestion 2: Extract Health Feature Space to Visualize Fault Types and Severity

In this doctoral dissertation, GANES was newly proposed in transformers fault diagnosis to diagnose fault types and severity. However, it should be examined more experiments and interpretation for health feature space to demonstrate degradation property. Future work is suggested as follows. First, more clear health feature space need to be extracted for visualizing degradation properties of severity and fault types. Second, more case studies need to be performed for various GANs methods.

Suggestion 3: Implementation of the BDD Framework in Other Fields

In this doctoral dissertation, a new framework of bridging a rule-based and AI-based method was investigated. However, it should be examined more experiments and applied to other applications to demonstrate BDD performance. Future work is suggested as follows. First, more case studies need to be performed for various rule-based methods. Second, more techniques for handling the noisy labeled problems, used in other research fields, should be further studied, in place of the regularization task used here. Finally, in-depth investigation of parameter transfer should be conducted by adjusting the hyperparameters (e.g., learning rate in re-training).

References

[1] A. Christina, M. Salam, Q. Rahman, F. Wen, S. Ang, and W. Voon, "Causes of transformer failures and diagnostic methods–A review," *Renewable and Sustainable Energy Reviews,* vol. 82, pp. 1442-1456, 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 *et al.*, "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.

[6] J. J. Kelly, "Transformer fault diagnosis by dissolved-gas analysis," *IEEE Transactions on Industry Applications,* no. 6, pp. 777-782, 1980.

[7] F. W. Heinrichs, "The Impact of Fault-Detection Methods and Analysis on

the Transformer Operating Decision," *IEEE Transactions on Power Delivery,* vol. 2, no. 3, pp. 836-842, 1987.

[8] M. Duval, "Dissolved gas analysis: It can save your transformer," *IEEE Electrical Insulation Magazine,* vol. 5, no. 6, pp. 22-27, 1989.

[9] C. E. Lin, J.-M. Ling, and C.-L. Huang, "An expert system for transformer fault diagnosis using dissolved gas analysis," *IEEE Transactions on Power Delivery,* vol. 8, no. 1, pp. 231-238, 1993.

[10] M. Duval, "A review of faults detectable by gas-in-oil analysis in transformers," *IEEE electrical Insulation magazine,* vol. 18, no. 3, pp. 8-17, 2002.

[11] Z. Wang, "Artificial intelligence applications in the diagnosis of power transformer incipient faults," Virginia Tech, 2000.

[12] T. K. Saha, "Review of modern diagnostic techniques for assessing insulation condition in aged transformers," *IEEE transactions on dielectrics and electrical insulation,* vol. 10, no. 5, pp. 903-917, 2003.

[13] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab1996.

[14] A. Malhi and R. X. Gao, "PCA-based feature selection scheme for machine defect classification," *IEEE Transactions on Instrumentation and Measurement,* vol. 53, no. 6, pp. 1517-1525, 2004.

[15] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2006, pp. 507-514.

[16] M. Mittal, M. Bhushan, S. Patil, and S. Chaudhari, "Optimal feature selection for SVM based fault diagnosis in power transformers," *IFAC Proceedings Volumes,* vol. 46, no. 32, pp. 809-814, 2013.

[17] 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 Transactions on Dielectrics and Electrical Insulation,* vol. 23, no. 2, pp. 1198-1206, 2016.

[18] T. Kari *et al.*, "Hybrid feature selection approach for power transformer fault diagnosis based on support vector machine and genetic algorithm," *IET Generation, Transmission & Distribution,* vol. 12, no. 21, pp. 5672-5680, 2018.

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

[20] E. Engineers and I. Board, "IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers," *IEEE: Piscataway, NJ, USA,* 2009.

[21] R. Rogers, "IEEE and IEC codes to interpret incipient faults in transformers, using gas in oil analysis," *IEEE transactions on electrical insulation,* no. 5, pp. 349-354, 1978.

[22] P. CODE and C. PRIX, "mineral oil-impregnated electrical equipment in service–guide to the interpretation of dissolved and free gases analysis," 2008.

[23] H. Yann-Chang, Y. Hong-Tzer, and H. Ching-Lien, "Developing a new transformer fault diagnosis system through evolutionary fuzzy logic," *IEEE Transactions on Power Delivery,* vol. 12, no. 2, pp. 761-767, 1997.

[24] Y. Hong-Tzer and L. Chiung-Chou, "Adaptive fuzzy diagnosis system for dissolved gas analysis of power transformers," *IEEE Transactions on Power Delivery,* vol. 14, no. 4, pp. 1342-1350, 1999.

[25] 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 Transactions on Power Systems,* vol. 15, no. 2, pp. 593-598, 2000.

[26] S. M. Islam, T. Wu, and G. Ledwich, "A novel fuzzy logic approach to transformer fault diagnosis," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 7, no. 2, pp. 177-186, 2000.

[27] Y. Hong-Tzer, L. Chiung-Chou, and C. Jeng-Hong, "Fuzzy learning vector quantization networks for power transformer condition assessment," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 8, no. 1, pp. 143-149, 2001.

[28] W. Mang-Hui, "A novel extension method for transformer fault diagnosis," *IEEE Transactions on Power Delivery,* vol. 18, no. 1, pp. 164-169, 2003.

[29] D. R. Morais and J. G. Rolim, "A hybrid tool for detection of incipient faults in transformers based on the dissolved gas analysis of insulating oil," *IEEE Transactions on Power Delivery,* vol. 21, no. 2, pp. 673-680, 2006.

[30] V. Duraisamy, N. Devarajan, D. Somasundareswari, A. A. M. Vasanth, and S. Sivanandam, "Neuro fuzzy schemes for fault detection in power transformer," *Applied Soft Computing,* vol. 7, no. 2, pp. 534-539, 2007.

[31] R. Naresh, V. Sharma, and M. Vashisth, "An Integrated Neural Fuzzy Approach for Fault Diagnosis of Transformers," *IEEE Transactions on Power Delivery,* vol. 23, no. 4, pp. 2017-2024, 2008.

[32] A. Shintemirov, W. Tang, and Q. H. Wu, "Power Transformer Fault Classification Based on Dissolved Gas Analysis by Implementing Bootstrap and Genetic Programming," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 39, no. 1, pp. 69-79, 2009.

[33] K. Bacha, S. Souahlia, and M. Gossa, "Power transformer fault diagnosis based on dissolved gas analysis by support vector machine," *Electric power systems research,* vol. 83, no. 1, pp. 73-79, 2012.

[34] F. Gieseke, A. Airola, T. Pahikkala, and O. Kramer, "Sparse Quasi-Newton Optimization for Semi-supervised Support Vector Machines," in *ICPRAM (1)*, 2012, pp. 45-54.

[35] H. Ma, C. Ekanayake, and T. K. Saha, "Power transformer fault diagnosis

under measurement originated uncertainties," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 19, no. 6, pp. 1982-1990, 2012.

[36] 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 Generation, Transmission & Distribution,* vol. 6, no. 2, pp. 142-151, 2012.

[37] A. D. Ashkezari, H. Ma, T. K. Saha, and C. Ekanayake, "Application of fuzzy support vector machine for determining the health index of the insulation system of in-service power transformers," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 20, no. 3, pp. 965-973, 2013.

[38] Y. Cui, H. Ma, and T. Saha, "Improvement of power transformer insulation diagnosis using oil characteristics data preprocessed by SMOTEBoost technique," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 21, no. 5, pp. 2363-2373, 2014.

[39] Z. B. Sahri and U. T. Malaysia, "Support vector machine-based fault diagnosis of power transformer using k nearest-neighbor imputed DGA dataset," *Journal of Computer and Communications,* vol. 2, no. 09, p. 22, 2014.

[40] Y. Zhang, X. Ding, Y. Liu, and P. J. Griffin, "An artificial neural network approach to transformer fault diagnosis," *IEEE Transactions on Power Delivery,* vol. 11, no. 4, pp. 1836-1841, 1996.

[41] W. Zhenyuan, L. Yilu, and P. J. Griffin, "A combined ANN and expert system tool for transformer fault diagnosis," *IEEE Transactions on Power Delivery,* vol. 13, no. 4, pp. 1224-1229, 1998.

[42] J. L. Guardado, J. L. Naredo, P. Moreno, and C. R. Fuerte, "A comparative study of neural network efficiency in power transformers diagnosis using dissolved gas analysis," *IEEE Transactions on Power Delivery,* vol. 16, no. 4, pp. 643-647, 2001.

[43] Y. Huang, H. Yang, and K. Huang, "Abductive network model-based diagnosis system for power transformer incipient fault detection," *IEE Proceedings - Generation, Transmission and Distribution,* vol. 149, no. 3, pp. 326-330, 2002.

[44] H. Yann-Chang and H. Chao-Ming, "Evolving wavelet networks for power transformer condition monitoring," *IEEE Transactions on Power Delivery,* vol. 17, no. 2, pp. 412-416, 2002.

[45] M. Wang, "Extension neural network for power transformer incipient fault diagnosis," *IEE Proceedings - Generation, Transmission and Distribution,* vol. 150, no. 6, pp. 679-685, 2003.

[46] H. Yann-Chang, "Evolving neural nets for fault diagnosis of power transformers," *IEEE Transactions on Power Delivery,* vol. 18, no. 3, pp. 843-848, 2003.

[47] A. R. G. Castro and V. Miranda, "Knowledge discovery in neural networks

with application to transformer failure diagnosis," *IEEE Transactions on Power Systems,* vol. 20, no. 2, pp. 717-724, 2005.

[48] H. Wu, X. Li, and D. Wu, "RMP neural network based dissolved gas analyzer for fault diagnostic of oil-filled electrical equipment," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 18, no. 2, pp. 495-498, 2011.

[49] 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," *International Journal of Electrical Power & Energy Systems,* vol. 43, no. 1, pp. 1346-1353, 2012.

[50] F. Zakaria, D. Johari, and I. Musirin, "Artificial neural network (ANN) application in dissolved gas analysis (DGA) methods for the detection of incipient faults in oil-filled power transformer," in *2012 IEEE International Conference on Control System, Computing and Engineering*, 2012, pp. 328-332: IEEE.

[51] V. Miranda, A. R. G. Castro, and S. Lima, "Diagnosing Faults in Power Transformers With Autoassociative Neural Networks and Mean Shift," *IEEE Transactions on Power Delivery,* vol. 27, no. 3, pp. 1350-1357, 2012.

[52] M. A. B. Amora, O. M. Almeida, A. P. S. Braga, F. R. Barbosa, L. A. C. Lisboa, and R. S. T. Pontes, "Improved DGA method based on rules extracted from high-dimension input space," *Electronics Letters,* vol. 48, no. 17, pp. 1048-1049, 2012.

[53] H. Ma, T. K. Saha, C. Ekanayake, and D. Martin, "Smart Transformer for Smart Grid—Intelligent Framework and Techniques for Power Transformer Asset Management," *IEEE Transactions on Smart Grid,* vol. 6, no. 2, pp. 1026-1034, 2015.

[54] Y. Zhang *et al.*, "A Fault Diagnosis Model of Power Transformers Based on Dissolved Gas Analysis Features Selection and Improved Krill Herd Algorithm Optimized Support Vector Machine," *IEEE Access,* vol. 7, pp. 102803-102811, 2019.

[55] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE transactions on neural networks and learning systems,* vol. 29, no. 8, pp. 3573-3587, 2018.

[56] V. Tra, B. Duong, and J. Kim, "Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 26, no. 4, pp. 1325-1333, 2019.

[57] P. Mirowski and Y. LeCun, "Statistical machine learning and dissolved gas analysis: a review," *IEEE Transactions on Power Delivery,* vol. 27, no. 4, pp. 1791-1799, 2012.

[58] L. Zheng, H. Yuan, X. Wang, and H. Yin, "Fault diagnosis of transformer based on principal component analysis and self-organizing map neural network," in *2016 IEEE International Conference on High Voltage*

Engineering and Application (ICHVE), 2016, pp. 1-4: IEEE.

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

[60] Y. Hong-Tzer and H. Yann-Chang, "Intelligent decision support for diagnosis of incipient transformer faults using self-organizing polynomial networks," *IEEE Transactions on Power Systems,* vol. 13, no. 3, pp. 946-952, 1998.

[61] 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 Transactions on Power Delivery,* vol. 18, no. 4, pp. 1241-1248, 2003.

[62] K. Meng, Z. Y. Dong, D. H. Wang, and K. P. Wong, "A Self-Adaptive RBF Neural Network Classifier for Transformer Fault Analysis," *IEEE Transactions on Power Systems,* vol. 25, no. 3, pp. 1350-1360, 2010.

[63] 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.

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

[65] 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 Transactions on Dielectrics and Electrical Insulation,* vol. 24, no. 5, pp. 2828-2835, 2017.

[66] M. Duval and A. DePabla, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," *IEEE Electrical Insulation Magazine,* vol. 17, no. 2, pp. 31-41, 2001.

[67] D. C. Ferreira, F. I. Vázquez, and T. Zseby, "Extreme Dimensionality Reduction for Network Attack Visualization with Autoencoders," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1-10: IEEE.

[68] B. Qi, Y. Wang, P. Zhang, C. Li, and H. Wang, "A novel deep recurrent belief network model for trend prediction of transformer DGA data," *IEEE Access,* vol. 7, pp. 80069-80078, 2019.

[69] 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.

[70] 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.

[71] 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.

[72] W. Haiyan, Y. Haomin, L. Xueming, and R. Haijun, "Semi-supervised autoencoder: A joint approach of representation and classification," in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, 2015, pp. 1424-1430: IEEE.

[73] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096-1103: ACM.

[74] 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 *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 2018, pp. 708-715: IEEE.

[75] F. Zhuang, D. Luo, X. Jin, H. Xiong, P. Luo, and Q. He, "Representation learning via semi-supervised autoencoder for multi-task learning," in *2015 IEEE International Conference on Data Mining*, 2015, pp. 1141-1146: IEEE.

[76] M. Yang and L. Hu, "Intelligent fault types diagnostic system for dissolved gas analysis of oil-immersed power transformer," *IEEE Transactions on*

Dielectrics and Electrical Insulation, vol. 20, no. 6, pp. 2317-2324, 2013.

[77] X. Wu, Y. He, and J. Duan, "A Deep Parallel Diagnostic Method for Transformer Dissolved Gas Analysis," *Applied Sciences,* vol. 10, no. 4, p. 1329, 2020.

[78] F. Zhang, J. Yan, P. Fu, J. Wang, and R. X. Gao, "Ensemble sparse supervised model for bearing fault diagnosis in smart manufacturing," *Robotics and Computer-Integrated Manufacturing,* vol. 65, p. 101920, 2020.

[79] 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.

[80] 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.

[81] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289,* 2015.

[82] 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.

[83] H. Shao, H. Jiang, Y. Lin, and X. Li, "A novel method for intelligent fault

diagnosis of rolling bearings using ensemble deep auto-encoders," *Mechanical Systems and Signal Processing,* vol. 102, pp. 278-297, 2018.

[84] 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 Generation, Transmission & Distribution,* vol. 11, no. 10, pp. 2721-2729, 2017.

[85] S. M. Frank *et al.*, "Metrics and Methods to Assess Building Fault Detection and Diagnosis Tools," National Renewable Energy Lab.(NREL), Golden, CO (United States)2019.

[86] 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," *Mechanical Systems and Signal Processing,* vol. 104, pp. 799-834, 2018.

[87] J. Jiao, M. Zhao, and J. Lin, "Unsupervised adversarial adaptation network for intelligent fault diagnosis," *IEEE Transactions on Industrial Electronics,* 2019.

[88] X. Li, W. Zhang, Q. Ding, and X. Li, "Diagnosing rotating machines with weakly supervised data using deep transfer learning," *IEEE Transactions on Industrial Informatics,* vol. 16, no. 3, pp. 1688-1697, 2019.

[89] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Transactions on Industrial Informatics,* vol. 15, no. 4, pp. 2446-2455, 2018.

[90] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Transactions on Industrial Electronics,* vol. 66, no. 9, pp. 7316-7325, 2018.

[91] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717-1724.

[92] C. H. Park, H. Kim, J. Lee, G. Ahn, M. Youn, and B. D. Youn, "A Feature Inherited Hierarchical Convolutional Neural Network (FI-HCNN) for Motor Fault Severity Estimation Using Stator Current Signals," *International Journal of Precision Engineering and Manufacturing-Green Technology,* pp. 1-14, 2020.

[93] G. B. Goh, C. Siegel, A. Vishnu, and N. Hodas, "Using rule-based labels for weak supervised learning: a ChemNet for transferable chemical property prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 302-310.

[94] H. Zuo, G. Zhang, J. Lu, and W. Pedrycz, "Fuzzy rule-based transfer learning for label space adaptation," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017, pp. 1-6: IEEE.

[95] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv*

preprint arXiv:1803.08375, 2018.

[96] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 443-449.

[97] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geoscience and Remote Sensing Letters,* vol. 13, no. 1, pp. 105-109, 2015.

[98] S. Kim *et al.*, "A Semi-Supervised Autoencoder With an Auxiliary Task (SAAT) for Power Transformer Fault Diagnosis Using Dissolved Gas Analysis," *IEEE Access,* vol. 8, pp. 178295-178310, 2020.

[99] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 5049-5059.

[100] B. Han *et al.*, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8527-8537.

[101] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in neural information processing systems*, 2013, pp. 1196-1204.

[102] A. Tarvainen and H. Valpola, "Mean teachers are better role models:

Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195-1204.

[103] H. Song, M. Kim, D. Park, and J.-G. Lee, "Learning from Noisy Labels with Deep Neural Networks: A Survey," *arXiv preprint arXiv:2007.08199,* 2020.

[104] S. Singh, D. Joshi, and M. Bandyopadhyay, "Software implementation of Duval triangle technique for DGA in power transformers," *International Journal of Electrical Engineering,* vol. 4, no. 5, pp. 529-540, 2011.

[105] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014.

[106] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234-2242.

국문 초록

비표지 고장 데이터와 유중가스분석데이터를 이용한 딥러닝기반 주변압기 고장진단 연구

서울대학교 대학원
기계항공공학부
김 선 의

오늘날 산업의 급속한 발전과 고도화로 인해 안전하고 신뢰할 수 있는 전력 계통에 대한 수요는 더욱 중요해지고 있다. 따라서 실제 산업 현장에서는 주변압기의 안전한 작동을 위해 상태를 정확하게 진단할 수 있는 prognostics and health management (PHM)와 같은 기술이 필요하다. 주변압기 진단을 위해 개발된 다양한 방법 중 인공지능(AI) 기반 접근법은 산업과 학계에서 많은 관심을 받고 있다. 더욱이 방대한 데이터와 함께 높은 성능을 달성하는 딥 러닝 기술은 주변압기 고장 진단의 학자들에게 높은 관심을 갖게 해줬다. 그 이유는 딥 러닝 기술이 시스템의 도메인 지식을 깊이 이해할 필요 없이 대량의 데이터만 주어진다면 복잡한 시스템이라도 사용자의 목적에 맞게 그 해답을 찾을

수 있기 때문에 딥 러닝에 대한 관심은 주변압기 고장 진단 분야에서 특히 두드러졌다.

그러나, 이러한 뛰어난 진단 성능은 아직 실제 주변압기 산업에서는 많은 관심을 얻고 있지는 못한 것으로 알려졌다. 그 이유는 산업현장의 비표지데이터와 소량의 고장데이터 때문에 우수한 딥러닝기반의 고장 진단 모델들을 개발하기 어렵다.

따라서 본 학위논문에서는 주변압기 산업에서 현재 대두되고 있는 세가지 이슈를 연구하였다. 1) 건전성 평면 시각화 이슈, 2) 데이터 부족 이슈, 3) 심각도 이슈 들을 극복하기 위한 딥 러닝 기반 고장 진단 연구를 진행하였다. 소개된 세가지 이슈들을 개선하기 위해 본 학위논문은 세 가지 연구를 제안하였다.

첫 번째 연구는 보조 감지 작업이 있는 준지도 자동 인코더를 통해 건전성 평면을 제안하였다. 제안된 방법은 변압기 열하 특성을 시각화 할 수 있다. 또한, 준지도 접근법을 활용하기 때문에 방대한 비표지데이터 그리고 소수의 표지데이터만으로 구현될 수 있다. 제안방법은 주변압기 건전성을 건전성 평면과 함께 시각화하고, 매우 적은 소수의 레이블 데이터만으로 주변압기 고장을 진단한다.

두 번째 연구는 규칙 기반 Duval 방법을 AI 기반 deep neural network (DNN)과 융합(bridge)하는 새로운 프레임워크를 제안하였다. 이 방법은 룰기반의 Duval을 사용하여 비표지데이터를 수도 레이블링한다 (pseudo-labeling). 또한, AI 기반 DNN은 정규화 기술과 매개 변수 전이 학습을 적용하여 노이즈가 있는 pseudo-label 데이터를 학습하는데 사용된다. 개발된 기술은 방대한양의 비표지데이터를 룰기반으로 일차적으로 진단한 결과와 소수의 실제

고장데이터와 함께 학습데이터로 훈련하였을 때 기존의 진단 방법보다 획기적인 향상을 가능케 한다.

끝으로, 세 번째 연구는 고장 타입을 진단할 뿐만 아니라 심각도 또한 진단하는 기술을 제안하였다. 이때 두 상태의 레이블링된 고장 타입과 심각도 사이에는 불균일한 데이터 분포로 이루어져 있다. 그 이유는 심각도의 경우 레이블링이 항상 되어 있지만 고장 타입의 경우는 실제 주변압기로부터 고장 타입 데이터를 얻기가 매우 어렵기 때문이다. 따라서, 본 연구에서 세번째로 개발한 기술은 오늘날 데이터 생성에 매우 우수한 성능을 달성하고 있는 generative adversarial network (GAN)를 통해 불균형한 두 상태를 균일화 작업을 수행하는 동시에 고장 모드와 심각도를 진단하는 모델을 개발하였다.

주요어: 고장진단
주변압기
딥러닝
유중가스분석

학 번: 2014-22479