Engineering Applications of Artificial Intelligence 124 (2023) 106498

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Multi-filter semi-supervised transformer model for fault diagnosis

Xuemin Tan[a], Jun Qi[a,*], John Q. Gan[b], Jianglin Zhang[a], Chao Guo[c], Fu Wan[d], Ke Wang[e]

[a] *College of Automation, Chengdu University of Information Technology, Chengdu 610225, China*
[b] *School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK*
[c] *State Grid Chengdu Power Supply Company, Chengdu 610041, China*
[d] *State Key Laboratory of Power Transmission Equipment & System Security and New Technology, Chongqing University, Chongqing 400000, China*
[e] *China Electric Power Research Institute, Beijing 100192, China*

ARTICLE INFO

Keywords:
Dissolved gas analysis
Fault diagnosis
Multi-filter semi-supervised
Feature selection
Confidence criterion

ABSTRACT

Dissolved Gas Analysis (DGA) is the most commonly used method for power transformer fault diagnosis. However, very few reliable and labeled fault DGA samples are available in the transformer substation whilst DGA data without labels is easier to obtain, which makes it difficult to train fault detectors in high-dimensional input space or select features using wrapper methods. Therefore, in order to improve the fault diagnosis accuracy using limited labeled DGA samples but more unlabeled DGA data, this paper proposes a novel multi-filter semi-supervised feature selection method for selecting optimal DGA features and building effective fault diagnosis models. A confidence criterion is also proposed for selecting high confidence unlabeled data to expand the training data set. Five filter techniques based on different evaluation criteria are employed to rank input DGA features, and a feature combination method is then applied to aggregate feature ranks by multiple filters and form a lower-dimensional candidate feature subset. The proposed method has been tested by using the IEC T10 dataset and compared with traditional supervised diagnostic models. The results show that the proposed method works well in optimizing DGA features and improving fault diagnosis accuracy significantly. Besides, the robustness of the selection of optimal feature subset is validated by testing DGA samples from the local power utility.

1. Introduction

Power transformers are important transmission and transformation equipment in the power grid, which undertake the tasks of voltage conversion and current transmission. Failure of a power transformer will lead to disconnection of the system and economic losses (Ma et al., 2021). Therefore, the study of transformer fault diagnosis not only strengthens the deep integration with the energy industry, but also will greatly improve the health level of transformers, ensure the reliable supply of clean energy across regions. Dissolved Gas Analysis (DGA) is the most commonly used method for power transformer fault diagnosis. The content of dissolved gases in the oil commonly includes H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO and CO_2. In recent years, various criteria for transformer fault diagnosis based on DGA have been proposed, such as those reported in total gases (The Institute of Electrical and Electronics Engineers, 1992), Doernenberg (The Institute of Electrical and Electronics Engineers, 1992), Rogers (Rogers, 1978), Duval Triangles Method (Mawelela et al., 2020) and IEC 60599 (Duval and Depabla, 2001). These criteria often lead to misjudgment and omission of judgment due to incomplete coding and absolute boundary. In Table 1, the advantages and disadvantages of the criteria are compared. At present, DGA has the problem of the low transformer diagnosis. A large number of offline and online DGA data are idle and not used effectively. However, traditional DGA mainly relies on manual experience and IEC method, which results in low accuracy. Therefore, it is urgent to introduce advanced AI algorithms for DAG data mining. Artificial intelligence techniques have been widely used in this field because it can establish complex nonlinear relationships between DGA content and transformer faults. Clustering based Method (CBT), Fuzzy Logic Inference System (FLIS), Artificial Neural Network (ANN), Support Vector Machine (SVM), Fisher Linear Discriminant Analysis (FLDA) and other AI techniques have been widely used in transformer fault diagnosis and achieved good results. However, CBT can only divide the fault samples into several different sub-classes rather than diagnose the fault types of DGA samples. The inference rules and fuzzy membership functions of FLIS are largely dependent on experience. ANN is prone to fall into local minimum and over-fit. The performance of SVM is greatly influenced by parameters. Among these methods, FLDA is an effective method and has the advantages of fast calculation and no need of hyperparameter adjustment.

Compared with semi-supervised method, most traditional supervised AI methods need a large number of labeled DGA data for building

[*] Corresponding author.
E-mail address: Qhhchina689@163.com (J. Qi).

https://doi.org/10.1016/j.engappai.2023.106498
Received 15 December 2022; Received in revised form 7 May 2023; Accepted 18 May 2023
Available online xxxx

Table 1
The advantages and disadvantages comparison of the different criteria.

Method	Feature sets	Benefits	Drawbacks
Total gases (The Institute of Electrical and Electronics Engineers, 1992)	H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO, CO_2.	Preliminary determination of transformer fault type	Only the characteristic gas content cannot diagnose transformer fault type effectively
Dornenburg (The Institute of Electrical and Electronics Engineers, 1992)	CH_4/H_2, C_2H_2/C_2H_4, C_2H_2/CH_4, C_2H_6/C_2H_2	The non-coding fault diagnosis method avoids the misjudgment caused by the lack of coding and the absolute coding interval	If the gas concentration is too low, transformer failure cannot be determined
Roger (Rogers, 1978)	C_2H_6/CH_4, C_2H_2/C_2H_4, CH_4/H_2, C_2H_4/C_2H_6	The mapping relationship between gas content ratio and fault type is established	Some ratios can only reflect the finite temperature range of thermal decomposition and cannot reflect discharge faults
IEC 60599 (Duval and Depabla, 2001)	C_2H_2/C_2H_4, CH_4/H_2, C_2H_4/C_2H_6	Simple coding method, less dependence on personnel experience, thus reducing the probability of misjudgment	Misjudgment between different faults near the boundary may occur because of too clear boundary,and there is no coding combination of corresponding fault type (lack of coding combination)
Duval triangle (Mawelela et al., 2020)	CH_4, C_2H_4, C_2H_2, CH_4/TT, C_2H_4/TT, C_2H_2/TT (TT=(CH_4+C_2H_4+C_2H_2))	The fault diagnosis results are complete, and the situation is not occur that real data cannot find the corresponding diagnosis or judge the failure	The use of content ratio diagnosis in low gas content situations tends to magnify accidental test errors

a fault diagnosis model and improving accuracy. Gouda et al. (2019b,a) proposed two supervised techniques which respectively consider the effect of cellulose insulation failure and C_2H_6 and H_2 concentrations for their importance in diagnosing certain types of faults, and analyze the concentration of combustible gases and interpreting their results in detecting and evaluating the incipient fault condition of oil-immersed transformers. In addition, Gouda et al. (2021) also provided a new concept using supervised artificial intelligence for enhancing the diagnostic accuracy of the conventional DGA method such as Dornenburg ratio, Rogers' ratio and IEC standard.

In fact, very few reliable and labeled fault DGA samples are available in the transformer substation. There are two reasons: (1) The fault data on-site is scarce and labeling the actual fault data may not be accurate, so it is necessary to disassemble the transformer for confirming the real fault type and obtaining reliable fault category samples, which is very costly. (2) Due to the interests of many companies, transformer fault data is not transparent, which results in limited transformer fault data with some distortion. This means that traditional AI algorithms cannot diagnose transformer faults effectively. Nevertheless, unlabeled DGA data is readily available. Semi-supervised Learning (SSL) can use a small amount of labeled DGA data as guidance and a large amount of unlabeled DGA data to improve the learning performance, which is a powerful mathematical model for transformer fault diagnosis when fault DGA samples are scarce. In the field of transformer fault diagnosis, most of the existing research work focused on Supervised Learning (SL) using a large amount of labeled data to build models, but the application of SSL in transformer fault diagnosis is very rare. Chen (2016) proposed a fault diagnosis method based on SSL, in which an SSL method based on fuzzy nearest neighbor label propagation was adopted to diagnose faults of power transformers. Mirowski and LeCun (2012) tested two SSL algorithms on DGA datasets and verified the effectiveness of low-dimensional scaling (LDS) and local linear semi-supervised regression (LLSSR) in fault diagnosis. Mao et al. (2022) proposed a fault diagnosis method based on deep neural networks and a semi-supervised transfer learning framework called Adaptive Reinforcement (AR) for solving small samples problem. However, there was not much analysis and comparison for the benefits of SSL in DGA, and the importance of feature selection to improve the fault diagnosis accuracy was also not proved and discussed. Existing SSL methods include self-training, co-training, generative model and graph regularization framework (Song et al., 2022; Wei et al., 2021). Self-training has many advantages. It does not need specific assumptions like co-training, and it also does not need to estimate parameters like generative models and construct complex graph model like graph-based methods. Self-training only needs one classification model, a small number of labeled samples and a large number of unlabeled samples to complete the complex task. Therefore, this paper proposes to combine self-training with DGA feature selection for improving the fault diagnosis accuracy iteratively. However, it is difficult to improve classifier performance in the iterative process if unlabeled samples used to update the initial model are mislabeled in SSL, and there needs a confidence criterion to find some data with high confidence from unlabeled data. Therefore, this paper also proposes a confidence criterion based multi-classification FLDA for selecting high confidence DGA samples for self-training.

In the field of fault diagnosis and detection, feature selection is very important in improving the performance of a model. Boztas and Tuncer (2022) proposed a novel multi-leveled feature extraction network, which use neighborhood component analysis and ReliefF-based 2-layered feature selector to select most discriminative features. Thomas et al. (2023) proposed a novel deep convolutional neural network transformer model to feature extraction and automatically detect the fault type in power system networks. Wang et al. (2023) combined the convolutional neural network and the long short-term memory network to apply into the fault detection. For transformer fault diagnosis, most commonly used DGA features are based on gas ratios (Jamshed et al., 2021). There are obvious differences among gas ratios used in the literature and no unified standard is available for selecting features in diagnostic models (Huang et al., 2018), so all DGA gases, gas ratios and other useful features will be used as initial features of fault diagnosis in this paper. However, redundant features tend to reduce the efficiency of data processing and prediction classification rate. Some researchers have proposed various semi-supervised feature selection methods in the past ten years, which can be categorized into two types (Sheikhpour, 2017): filter and wrapper. A filter for feature selection scores features with a ranking criterion regardless of the model for fault diagnosis (Chen et al., 2020), such as ReliefF (Khan et al., 2021), Mutual Information (MI) (Gu et al., 2022), Infinite Latent Feature Selection (ILFS) (Cai et al., 2021) and other methods for semi-supervised feature selection. ReliefF technique was used for selecting optimal feature combinations among computing statistical values in terms of weights (Khan et al., 2021). Gu et al. (2022) proposed a feature selection algorithm based on conditional mutual information for maximal relevance, minimal redundancy, and mutual information between feature sets is exploited to describe redundancy. Generally speaking, filter-based methods are fast, efficient and scalable (Chen et al., 2020; Khan et al., 2021; Gu et al., 2022; Cai et al., 2021). Ren et al. (2008) proposed a wrapper-type forward semi-supervised feature selection framework that performs supervised sequential forward feature selection on both labeled and unlabeled data. Han et al. (2011) investigated a new wrapper-type semi-supervised feature selection framework that can select a more relevant feature subset using confident unlabeled data, which employs an ensemble classifier that supports the estimation of the confidence of unlabeled data. However, such wrapper-type methods are usually time consuming for high-dimensional data. Therefore, this paper develops a novel multi-filter semi-supervised method for DGA feature selection. The method includes two stages. In the first stage, each filter method is respectively used to rank features, with all samples repeated 20 times randomly and 5-fold cross validation performed based on SSL. In the second stage, feature combination methods based on Proportion Wrapper Strategy (PWS) and Average Threshold Strategy (ATS) are used to aggregate feature ranks based on multiple filters and obtain a lower-dimensional candidate feature subset, respectively.

The key contributions of this paper given as below:

• SSL is applied in DGA fault diagnosis, which is rarely used in the DGA field. It not only increases the application of SSL in DGA feature selection, but also deep the analysis and comparison for the benefits of SSL in DGA.

• A multi-filter semi-supervised fault diagnosis model based on feature selection is proposed,

which only take advantage of few labeled DGA data to obtain the Optimal Feature Combination (OFC) based on feature combination methods after SSL feature ranking based multiple filters.

• A confidence criterion for the expansion of training data is proposed. The criterion selects high confidence ones from unlabeled samples by analyzing the decision scores of FLDA multiple classifiers for avoiding unknown distribution noise and improving classifier performance.

A public DGA dataset IEC TC 10 is used to verify the effectiveness of the proposed method. Consequently, the results show that the proposed method can effectively find the OFC and significantly improve the accuracy of fault diagnose compared with traditional supervised diagnostic methods. Another dataset provided by a national power company is used to test and validate the robustness of the obtained OFC.

The rest of this paper is organized as follows. In Section 2, the Two-stage Multi-filter SSL Features Selection based on FLDA (TMSSL-FS-FLDA) method is proposed and a confidence criterion is introduced for the expansion of training data in SSL, followed by experimental study for feature selection and validation based semi-supervised and supervised learning in Session 3. Limitations are listed in Section 4, and conclusions and future directions are presented in Section 5. In addition, Appendix A lists abbreviations of term involved in the paper.

2. Selection of Optimal Feature Combination (OFC) based on multi-filter semi-supervised methods

It is necessary to remove irrelevant and redundant features, which can shorten the training time and improve the accuracy of fault diagnosis. For supervised feature selection methods, sufficient labeled data is used for feature selection, but it is difficult to obtain sufficient DGA labeled data for transformer fault diagnosis as it is time-consuming and expensive. Unsupervised feature selection methods evaluate feature relevance only with unlabeled data and ignore the value of labeled data. Semi-supervised feature selection methods can use both labeled and unlabeled data for selecting better features. This paper proposes a semi-supervised method based on filter criteria for DGA feature selection. However, it is incapable to select the optimal feature subset using a single filter criterion, while the combination of multi-filter criteria can obtain more reliable feature subsets (Yang et al., 2010).

2.1. Multi-class FLDA for transformer fault classification based on SSL

In the paper, the FLDA (Atoui and Cocquempot, 2021) is applied to solve the classification problem because of the advantages of fast calculation and no need of hyperparameter adjustment. The FLDA aims at finding a transformation matrix W which maximizes between-class scatter and minimizes within-class scatter, *i.e.*,

$$maximize \frac{tr(W^T S_b W)}{tr(W^T S_w W)} \tag{1}$$

where W is the LDA weight vector. Let $X = [x_1^1, x_1^2, \dots, x_1^{N_1}, x_2^1, x_2^2, \dots, x_2^{N_2}, \dots ..., x_1^1, x_1^2, \dots, x_d^{N_d}]$ be the data matrix of training samples and N_d be the number of samples in the dth class. The within-class scatter matrix (S_w) and the between-class scatter matrix (S_b) are defined as follows:

$$S_w = \frac{1}{N} \sum_{d=1}^{K} \sum_{i=1}^{N_d} (x_d^i - m_d)(x_d^i - m_d)^T \tag{2}$$

$$S_\mathrm{b} = \frac{1}{N} \sum_{d=1}^{K} N_d (m_d - m)(m_d - m)^T \tag{3}$$

where $m_d = \frac{1}{N_d} \sum_{i=1}^{N_d} x_d^i$ is the mean vector of the dth class and $m = \frac{1}{N} \sum_{d=1}^{K} \sum_{i=1}^{N_d} x_d^i$ is the total mean vector, N is the total number of samples, and K=*2* is the number of classes for the binary problem.

The decision score function $f\ (x)$ is defined as

$$f(x) = Wx + b \tag{4}$$

where $b = -\sum_{d=1}^{K} N_d m_d / \sum_{d=1}^{K} N_d$ is the bias, and the sign of $f\ (x)$ is used to predict the class label for a given test sample. If $f\ (x)$>0, the sample x belongs to the first class (*class1*), otherwise it belongs to the second class (*class2*).

The original FLDA is used for solving the binary classification problem, but the transformer fault classification is referred to multi-classification. Therefore, it is necessary to expand the binary FLDA to multi-class FLDA. In the paper, we select One-against-one (OAO) (Zheng et al., 2011) for multi-class FLDA because of the effectiveness in transformer fault classification. The OAO is used to train $n * (n-1)/2(n$ represent the number of class) binary FLDA classifiers with few initial available labeled samples. After the training, $n * (n-1)/2$ decision scores $f_j(x)$ $(j = 1, 2, \dots, n * (n-1)/2)$ are calculated separately based on $n * (n-1)/2$ binary LDA classifiers for all unlabeled samples. When classifying an unknown sample, the category with the most votes is the classification result of the unknown sample.

2.2. The candidate features

Transformer oil and insulating paper (board) will decompose when the transformer has an electrical or thermal fault. The decomposed contents mainly include H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO and CO_2, which are not enough as the input features of fault diagnosis algorithm for accurate diagnosis. Features derived from dissolved gases can be divided into three categories: gas concentration, gas ratio and gas relative percentage, which have discriminative power in different aspects and degrees. However, it is still unclear whether gas concentration or gas ratio or gas relative percentage is most relevant to the fault types of power transformer. According to conventional approaches and published literature (Wei et al., 2014; Koroglu and Demircali, 2016), a comprehensive feature set is created, as shown in Table 2. In Table 2, the first 11 features (number 1–11) represent gas concentration, the next 49 features (number 12–60) represent gas ratio, and the final 4 features (number 61–64) represent key gas relative percentage.

2.3. Data pre-processing

Two DGA datasets are respectively gathered to establish semi-supervised fault diagnosis model based on feature selection and test its

Table 2
The original feature set for fault diagnosis.

Number	Feature	Number	Feature	Number	Feature	Number	Feature
1	H_2	17	CO_2/C_2H_2	33	C_2H_4/H_2	49	C_2H_2/THD
2	CH_4	18	CO/C_2H_4	34	C_2H_6/H_2	50	C_2H_2/THH
3	C_2H_2	19	CO_2/C_2H_4	35	H_2/TH	51	C_2H_2/TCH
4	C_2H_4	20	CO/C_2H_6	36	H_2/THD	52	C_2H_6/C_2H_4
5	C_2H_6	21	CO_2/C_2H_6	37	H_2/THH	53	C_2H_4/TH
6	CO	22	CO/TH	38	H_2/TCH	54	C_2H_4/THD
7	CO_2	23	CO/THD	39	C_2H_2/CH_4	55	C_2H_4/THH
8	TH	24	CO/THH	40	C_2H_4/CH_4	56	C_2H_4/TCH
9	THD	25	CO/TCH	41	C_2H_6/CH_4	57	C_2H_6/TH
10	THH	26	CO_2/CO	42	CH_4/TH	58	C_2H_6/THD
11	TCH	27	CO_2/TH	43	CH_4/THD	59	C_2H_6/TH
12	CO/H_2	28	CO_2/THD	44	CH_4/THH	60	C_2H_6/TCH
13	CO_2/H_2	29	CO_2/THH	45	CH_4/TCH	61	max(key gas)
14	CO/CH_4	30	CO_2/TCH	46	C_2H_4/C_2H_2	62	$10/(C_2H_4/C_2H_2)$
15	CO_2/CH_4	31	CH_4/H_2	47	C_2H_6/C_2H_2	63	C_2H_2/THD/0.21
16	CO/C_2H_2	32	C_2H_2/H_2	48	C_2H_2/TH	64	C_2H_6/THD/0.23

TH=$H_2+CH_4+C_2H_2+C_2H_4+C_2H_6$, THD=$CH_4+C_2H_2+C_2H_4$, THH=$H_2+C_2H_2+C_2H_4$, TCH=$CH_4+C_2H_2+C_2H_4+C_2H_6$

Table 3
The number of fault samples before and after ADASYN.

Faults	LED	HED	LMT	HT	N
Before balance	26	48	16	18	50
After balance	49	48	46	49	50

performance. The public dataset, IEC TC10 dataset (Duval and Depabla, 2001), is utilized to build a fault diagnosis model based on feature selection and select optimal input feature subset. Another dataset provided by a national power company is used to test the performance of the obtained OFC. In these two DGA datasets, transformer faults are classified into six categories: Low-energy Discharge (LED), High-energy Discharge (HED), Low and Middle-temperature overheat (LMT), High Temperature overheating (HT), Partial Discharge (PD) and Normal operation (N). In our experiment the PD samples are excluded because the number of this class of samples is too small.

2.3.1. Data balance

Usually, the distributions of samples from different categories are imbalanced, and the imbalanced distributions of DGA data will cause the classification boundary shifting to the weak sample space, leading to wrongly classify weak samples as strong samples and reduce the performance of the classifier. To tackle data imbalance issue, many data balance techniques (Zhang and Li, 2022), such as Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling Algorithm (ADASYN) and Bootstraps, have been used for balancing samples. ADASYN is used to balance sample distribution in this paper. The results are shown in Table 3.

2.3.2. Data conversion

For improving fault diagnosis performance Arctangent Transformation (AT) (Li et al., 2016) and normalization are carried out after ADASYN. Let missing data be set to 0, the rule of AT for gas ratio and gas relative percentage is described as follows:

$$ratio = \left\{ \begin{array}{l} x/0 = Inf, x \neq 0 \\ 0/x = 0, x \neq 0 \\ 0/x = \mathrm{NaN}, x = 0 \end{array} \right\} \Rightarrow T_{ratio} = \left\{ \begin{array}{l} \arctan(\mathrm{inf}) \approx 1.571 \\ \arctan(0) = 0 \\ \mathrm{NaN} = 0 \end{array} \right. \quad (5)$$

Data normalization is carried out as follows to avoid data singularity and eliminate the difference in value ranges of different features:

$$x_{ij}^{'} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \quad (6)$$

where x_{ij} and $x_{ij}^{'}$ *respectively represents* the jth feature value of the ith sample before and after normalization, x_j^{max} and x_j^{min} are the maximum and minimum value of the jth feature.

2.4. Two-stage Multi-filter SSL Feature Selection based on FLDA (TMSSL-FS-FLDA) for power transformer fault diagnosis

To select important DGA features, the TMSSL-FS-FLDA algorithm for building fault diagnosis model and selecting OFC is proposed and described in this section. The algorithm includes initialization, feature ranking and feature selection. The steps of TMSSL-FS-FLDA are as follows:

Initialization (Steps 1 to 3)

Step 1: Divide DGA fault samples. DGA fault samples is divided into training set D_T and validation set D_V after pre-processing. The training set D_T is further divided into the labeled training set D_I and the unlabeled training set D_F. Thus, the sample set consists of D_I, D_F and D_V.

Step 2: Train an initial classifier. Use the candidate features shown in Table 1 and corresponding labels of set D_I to train the initial FLDA multi-class classifier based on OAO, and then predict the labels of the samples in D_F and D_V.

Step 3: Calculate the decision scores. Calculate the decision scores of FLDA multi-class classifier based on OAO in D_F, denoted as *Scores*(m), where m represents the number of iterations and is equal to 0 in initialization.

The first stage — feature ranking based on multi-filter semi-supervised method (Steps 4 to 8)

Step 4: Update the training set. In the mth iteration, according to the confidence criteria shown in Section 2.5, select samples with high confidence from the unlabeled training set D_F and predict their labels, forming a set of extended training samples denoted as Q_m. Therefore, a new training set P_m ($P_\mathrm{m} = D_I \cup Q_\mathrm{m}$) is constituted and the labels of P_m are denoted as $y_m(.)$.

Step 5: Retrain the classifier. Retrain FLDA multi-class classifier based on OAO by using the new training set P_m and their corresponding labels $y_m(.)$, and then perform classification on D_F and D_V.

Step 6: Re-calculate the decision scores. Calculate the decision scores of each sample in D_F, denoted as *Scores*(m) in the mth iteration.

Step 7: Find out the number of samples from D_F with different predicted labels in the mth and (m-1)th iterations when the number of iteration is greater than 1.

$$t(m) = number(find(y_m(.) \neq y_{m-1}(.)))(m > 1) \quad (7)$$

where *number*(.) represents the number of the samples satisfying the condition in the parenthesis.

Step 8: Check the termination criterion. If $t(m)$=0 or $m = m_0$ (the preset maximum number of iterations) the algorithm terminates. Otherwise, go to Step 4 to perform the (m+1)th iteration. After termination of the algorithm, the final accuracy on D_V is obtained, and the candidate

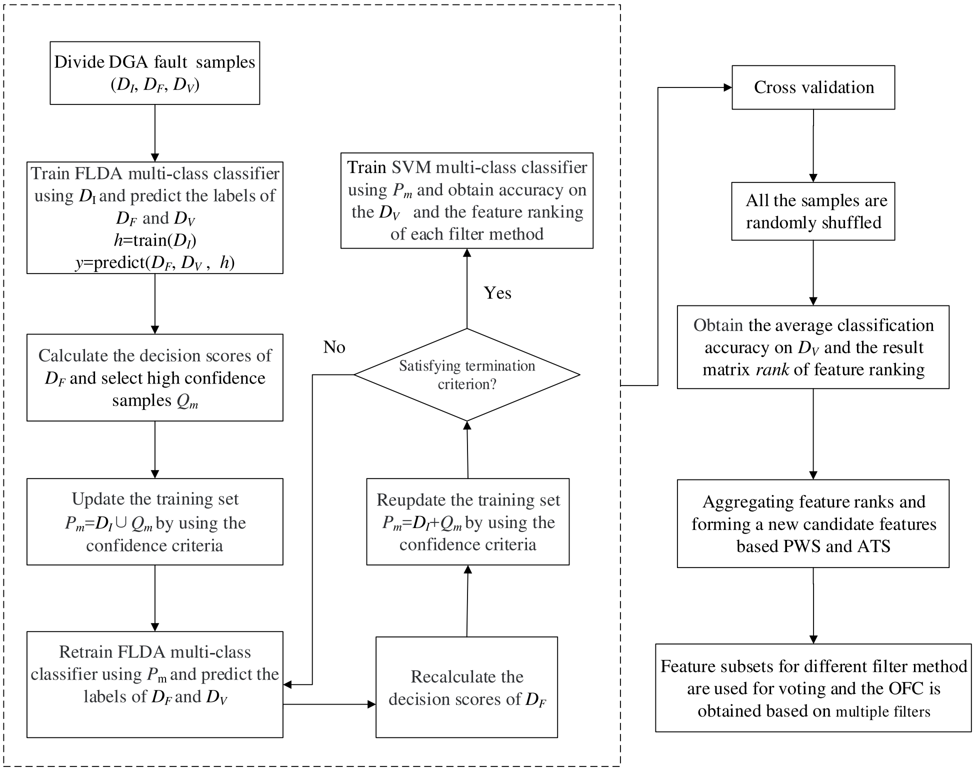

Divide DGA fault samples (D_I, D_F, D_V)
Train FLDA multi-class classifier using D_I and predict the labels of D_F and D_V h=train(D_I) y=predict(D_F, D_V , h)
Calculate the decision scores of D_F and select high confidence samples Q_m
Update the training set P_m=D_I∪Q_m by using the confidence criteria
Retrain FLDA multi-class classifier using P_m and predict the labels of D_F and D_V
Recalculate the decision scores of D_F
Reupdate the training set P_m=D_I+Q_m by using the confidence criteria
Satisfying termination criterion?
No
Yes
Train SVM multi-class classifier using P_m and obtain accuracy on the D_V and the feature ranking of each filter method
Cross validation
All the samples are randomly shuffled
Obtain the average classification accuracy on D_V and the result matrix *rank* of feature ranking
Aggregating feature ranks and forming a new candidate features based PWS and ATS
Feature subsets for different filter method are used for voting and the OFC is obtained based on multiple filters

Fig. 1. The flowchart of TMSSL-FS-FLDA.

features are ranked by each filter listed in Section 2.4.1 according to the candidate features and final labels of Q_m after the mth iteration. (In this paper, m_0 is set to 10 because $t(m)$ has become stable after 10 iterations)

The second stage — feature selection (Steps 9 to 11)

Step 9: After k-fold cross-validation, the average accuracy on D_V is calculated and a *Rank* matrix with results from k times ranking based on different filter methods is obtained (The matrix size is $k*c$, with c representing the number of candidate features, which is set to 64 in this paper).

Step 10: DGA fault samples are repeated r times in validation, with all the samples randomly shuffled each time and r different data ensembles are obtained. The average accuracy on D_V is then calculated. For each filter method, the size of the resulted matrix *Rank* is $(k*r)*c$, with $(k*r)$ and c respectively representing the number of rows and number of columns of the matrix.

Step 11: Feature combination method is used to select OFC based on the result matrix (See Section 2.4.2 for details). Aggregating feature ranks of each filter method according to the matrix *Rank* by using PWS and ATS. No matter PWS or ATS, the OFC and the corresponding average iteration accuracy for each filter method are obtained. After that, feature subsets from different filter methods are used for voting and the OFC is obtained.

In order to better understand the TMSSL-FS-FLDA algorithm, its flowchart is shown in Fig. 1. The pseudo-code is shown in Fig. 2.

2.4.1. Feature ranking based on SSL

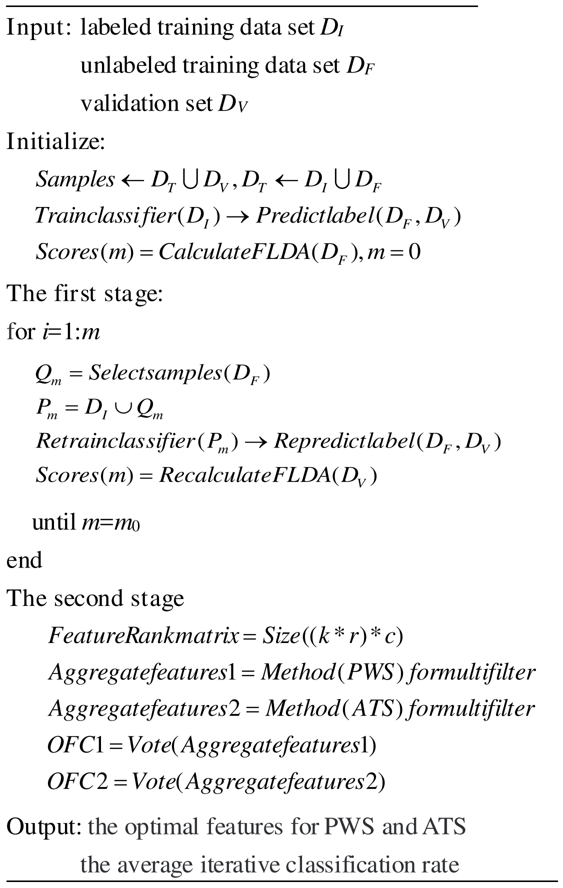
Five popular filter techniques with different evaluation criteria (Jovic et al., 2015), including ReliefF, Mutual Information (MI), Infinite Latent Feature Selection (ILFS), Max-Relevance and Min-Redundancy (MRMR), Least absolute shrinkage and selection operator (Lasso) are adopted to rank all features. The details of the ranking procedure are described in Steps 4–8 in the TMSSL-FS-FLDA algorithm. All samples are repeated 20 times randomly and 5-fold cross-validation is performed.

2.4.2. Feature combination method for selecting OFC

Feature combination method is used to select OFC based on the result matrix of feature ranking. Firstly, we proposed the two methods of Proportion Wrapper Strategy (PWS) and Average Threshold Strategy (ATS), which are respectively applied to aggregate feature ranks from each filter method and obtain a dimension-reduced feature subset. When using the PWS method to analyze the result matrix of each filter method, the summation of each row of the result matrix is calculated and sorted from smallest to largest, and then a feature subset which can provide the best average accuracy is selected as OFC after the features with different proportions are used to execute SSL based on 20 × 5 cross-validation. By the way, the average iteration accuracy based on different feature proportions are also obtained. For ATS, the features whose rank order summation is less than average summation value are kept as OFC after calculating the summation of each row of the result matrix of each filter method. Therefore, no matter PWS or ATS is adopted, the OFC from each filter method is obtained. Secondly, feature subsets for different filter methods are used for voting. These features with voting frequency >2 are retained and used to form a new candidate feature subset respectively for PWS and ATS, otherwise discarded.

2.5. Confidence criterion for training data expansion based on Multi-class Nearest Average Distance (MNAD) in SSL

In the semi-supervised learning process, there are two reasons for preventing the improvement of the classifier: using small labeled training set cannot obtain a reliable initial model and the unlabeled samples used to update initial model are not informative or have no discriminative power. Therefore, it is critical to find some data with high confidence from unlabeled data. In this paper, the MNAD criterion is proposed to select high confidence ones from unlabeled samples, and only those unlabeled samples with high confidence are used to expand training dataset. The criterion is implemented by analyzing the decision scores of FLDA classifier for unlabeled samples. The OAO



```
Input: labeled training data set $D_l$
       unlabeled training data set $D_F$
       validation set $D_V$
Initialize:
    $Samples \leftarrow D_T \bigcup D_V, D_T \leftarrow D_l \bigcup D_F$
    $Trainclassifier(D_l) \rightarrow Predictlabel(D_F, D_V)$
    $Scores(m) = CalculateFLDA(D_F), m = 0$
The first stage:
for $i$=1:$m$
    $Q_m = Selectsamples(D_F)$
    $P_m = D_l \cup Q_m$
    $Retrainclassifier(P_m) \rightarrow Repredictlabel(D_F, D_V)$
    $Scores(m) = RecalculateFLDA(D_V)$
  until $m$=$m_0$
end
The second stage
    $FeatureRankmatrix = Size((k*r)*c)$
    $Aggregatefeatures1 = Method(PWS) formultifilter$
    $Aggregatefeatures2 = Method(ATS) formultifilter$
    $OFC1 = Vote(Aggregatefeatures1)$
    $OFC2 = Vote(Aggregatefeatures2)$
Output: the optimal features for PWS and ATS
        the average iterative classification rate
```

Fig. 2. The pseudo-code of TMSSL-FS-FLDA.

strategy involves a parallel architecture made up of $n*(n-1)/2$ binary LDA classifiers, and n-1 binary classifiers for each class are defined by one class against another class. In the proposed MNAD technique, we use the OAO strategy to train $n*(n-1)/2$ binary FLDA classifiers with an initial set of labeled samples. After the initial training, $n*(n-1)/2$ decision scores $f_j(x)$ (j=$1, 2, \ldots, n*(n-1)/2$) are calculated separately based on $n*(n-1)/2$ binary LDA classifiers for each of unlabeled samples. The confidence of each unlabeled sample depends on the votes received, which means that we only consider samples with a full vote (each category receive a maximum of n-1 votes), and then these samples satisfying the following formula are selected for expanding training dataset at each iteration.

$$S = Vote(X)_{full}(X \in D_F) \tag{8}$$

$$value = mean(scores(S))(S \in D_S) \tag{9}$$

$$selectsamples = \{value > h * mean(value)\} \tag{10}$$

where S denote the samples with a full vote from unlabeled samples, $value$ and h respectively represents the average value of the decision scores and a weighting parameter.

3. Results and analysis

3.1. Data preparation

The public dataset IEC TC 10 (Duval and Depabla, 2001), which consists of 158 samples from 6 classes of faults (Partial Discharge class was excluded in this study), was utilized to build a fault diagnosis model and select the OFC. Another dataset provided by a national power company was used to test the performance of the obtained OFC. The ADASYN method was applied to balance the sample distribution by increasing the number of samples from 158 to 242, and the balanced results is shown in Table 3. After that, the balanced samples with 64 candidate features were further processed by Arctangent Transformation and normalization for improving fault diagnosis performance.

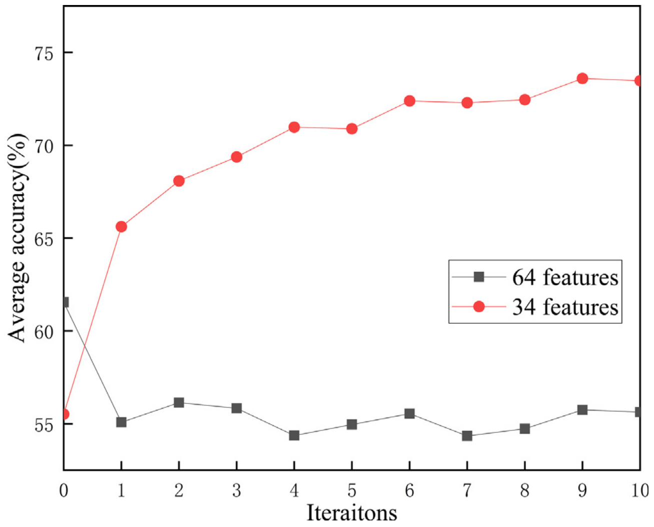


Fig. 3. Comparison of iterative average accuracy based on 34 features and 64 features respectively.

3.2. Feature ranking based on the multi-filter semi-supervised method

In the stage of feature ranking for TMSSL-FS-FLDA, 242 DGA fault samples (with 64 features) were divided into training set and validation set, and experiments based on 5-fold cross-validation and 20 repeated runs were conducted for obtaining average accuracy and feature ranks. An initial labeled training dataset with 75 samples and an unlabeled training dataset with 119 samples constitute the training dataset, and the remaining samples are in the validation dataset. Firstly, the average iteration accuracy based on 64 and 34 features after 20 × 5-fold cross-validation is shown in Fig. 3. The average accuracy for 64 and 34 features respectively decreases and increase with increasing number of iterations, which denotes that redundant features degrade the performance of the classifier with the SSL process.

The average feature ranks obtained by different filter methods are presented in Fig. 4. For most methods, the number of average feature rank is generally decreasing with the feature number from 1 to 64, which means the ranking is increasing. It is obvious that the DGA gases and the related gas ratios with CO and CO_2 (these features are numbered 1–30 in Table 2) are ranked lower by most methods except for Lasso. 34 highly ranked features were selected in the first stage of TMSSL-FS-FLDA, which are listed in Table 4. For comparison purpose, the average iteration accuracy based on these 34 features and 20 × 5 cross-validation is also shown in Fig. 3. Unlike the case based on 64 features, the average accuracy based on 34 selected features generally increases with increasing number of iterations, which demonstrates that with the selected 34 features the SSL process generally improve the performance of the classifier although the classification rate may not reach the optimal.

3.3. Feature selection based on feature combination methods

In the stage of feature ranking, the best accuracy achieved is 73.48% based on 34 features. However, the selected feature set could be further optimized using feature combination method based on PWS and ATS.

Table 4
The selected 34 features for fault diagnosis.

Number	Feature	Number	Feature	Number	Feature
1	CH_4/H_2	12	CH_4/TH	23	C_2H_4/TH
2	C_2H_2/H_2	13	CH_4/THD	24	C_2H_4/THD
3	C_2H_4/H_2	14	CH_4/THH	25	C_2H_4/THH
4	C_2H_6/H_2	15	CH_4/TCH	26	C_2H_4/TCH
5	H_2/TH	16	C_2H_4/C_2H_2	27	C_2H_6/TH
6	H_2/THD	17	C_2H_6/C_2H_2	28	C_2H_6/THD
7	H_2/THH	18	C_2H_2/TH	29	C_2H_6/TH
8	H_2/TCH	19	C_2H_2/THD	30	C_2H_6/TCH
9	C_2H_2/CH_4	20	C_2H_2/THH	31	max(key gas)
10	C_2H_4/CH_4	21	C_2H_2/TCH	32	$10/(C_2H_4/C_2H_2)$
11	C_2H_6/CH_4	22	C_2H_6/C_2H_4	33	C_2H_2/THD/0.21
				34	C_2H_6/THD/0.23

TH=$H_2+CH_4+C_2H_2+C_2H_4+C_2H_6$, THD=$CH_4+C_2H_2+C_2H_4$, THH=$H_2+C_2H_2+C_2H_4$, TCH=$CH_4+C_2H_2+C_2H_4+C_2H_6$

Table 5
Best features selected with supervised and semi-supervised algorithms based on PWS or ATS.

Algorithms		Selected features
TMSSL-FS-FLDA	PWS (based on ILFS)	32, 33, 31, 30, 29, 27, 28, 34, 21, 26
	ATS (based on MRMR)	18, 22, 24, 25, 26, 29, 30, 31, 32, 33, 21, 27, 28, 34
TMSTC-FS-FLDA	PWS (based on ReliefF)	30, 31, 28, 29, 32, 27, 26, 34, 25, 33
	ATS (based on ILFS)	12, 32, 30, 27, 28, 34, 13, 16, 25, 33, 15, 18, 21, 24, 29, 31
TMSUC-FS-FLDA	PWS (based on ReliefF)	29, 30, 26, 28, 27, 21, 31, 25, 20, 22
	ATS (based on ReliefF)	12, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34

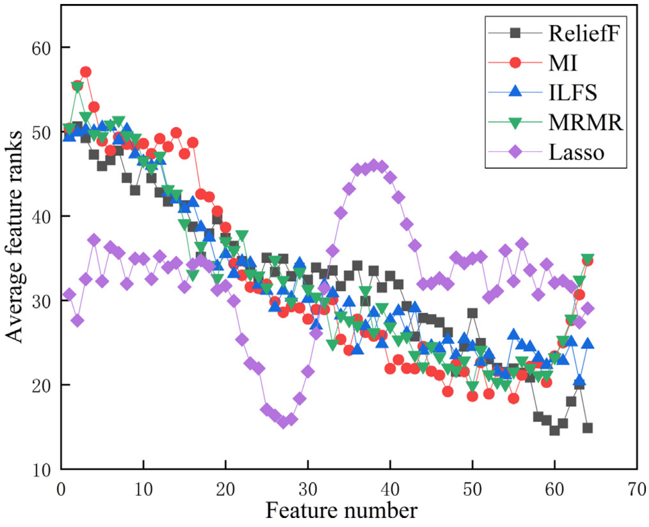


Fig. 4. Average feature ranks based on different filter methods.

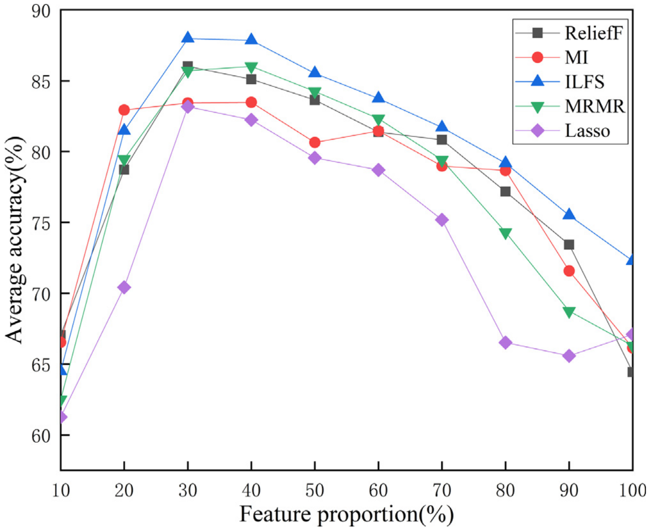


Fig. 5. Comparison of average accuracy of TMSSL-FS-FLDA with different proportions of features selected and using PWS for OFC.

For selecting OFC based on PWS, the variation of fault diagnosis accuracy with different proportions features selected according to feature ranking is presented in Fig. 5. The results are obtained based on 20 × 5-fold semi-supervised cross-validation after PWS. It is shown that the best accuracy corresponds to certain feature proportion for each filter method. Respectively, for the five different filter methods, using the first 30%, 40%, 30%, 40% and 30% of the ranked features can obtain the best fault diagnosis accuracy. Besides, it is found that fault diagnosis accuracy increases firstly and then decreases when the selected features exceed certain proportion, which shows the importance of feature selection in fault diagnosis. Fig. 6 shows the average accuracy after each iteration by selecting the best proportion of features for each filter method. The average iteration accuracy for each filter method is generally increasing and ILFS can obtain the best accuracy among other filter methods for TMSSL-FS-FLDA. The selected features based on ILFS for TMSSL-FS-FLDA are listed in Table 5. After aggregating feature ranks and voting for all filter methods, the OFC based PWS is obtained and shown in Table 6.

It can be seen from Fig. 6 that the DGA features selected by ILFS for PWS produced the best fault diagnosis accuracy among the tested filter methods. Fig. 7 shows the average accuracy after each iteration by selecting the best proportion of features and using ATS for OFC. It can be seen that MRMR achieved the best fault diagnosis accuracy. The selected features based on MRMR for TMSSL-FS-FLDA are listed in Table 5. After aggregating feature ranks and voting for all filter methods, the OFC based ATS is obtained shown in Table 6.

For comparing the performance of PWS and ATS in optimizing feature combinations, Fig. 8 shows the average accuracy of the last iteration of TMSSL-FS-FLDA after completing PWS or ATS based on different filter methods. For PWS, the best accuracy based on certain proportion of features for different filter methods is shown in Fig. 5, and 30% or 40% of the features can obtain higher accuracy. It is noteworthy that PWS method can obtain better iteration accuracy than ATS for all filter methods.

Table 6
Candidate feature subsets obtained using the proposed feature combination methods in comparison with wrapper methods.

Algorithms		Selected features
TMSSL-FS-FLDA	PWS	25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	ATS	22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	FW-SemiFS	5, 8, 9, 14, 15, 19, 20, 25, 27, 32, 33, 34
TMSTC-FS-FLDA	PWS	27, 28, 29, 30, 31, 32, 33, 34
	ATS	12, 13, 15, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	SFFS-STC	1, 2, 6, 11, 12, 13, 15, 16, 26, 29, 30, 34
TMSUC-FS-FLDA	PWS	12, 15, 21, 22, 23, 25, 26, 27, 29, 30, 31, 32, 33, 34
	ATS	12, 13, 14, 15, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34
	SFFS-SUC	1, 2, 10, 11, 12, 13, 14, 15, 25, 32, 33, 34

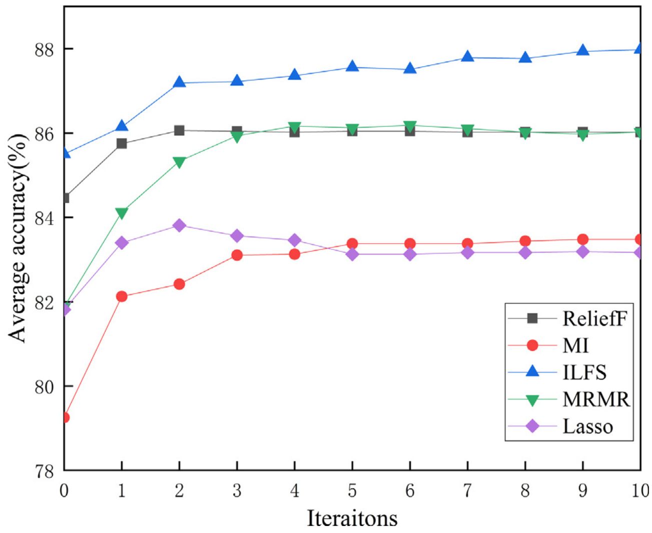


Fig. 6. Comparison of average accuracy after each iteration by selecting the best proportion of features selected and using PWS for OFC.

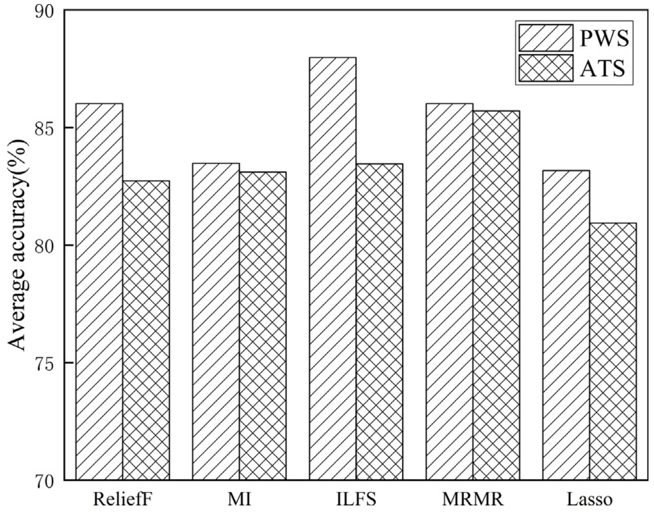


Fig. 8. Comparison of average accuracy of TMSSL-FS-FLDA using PWS or ATS for OFC.

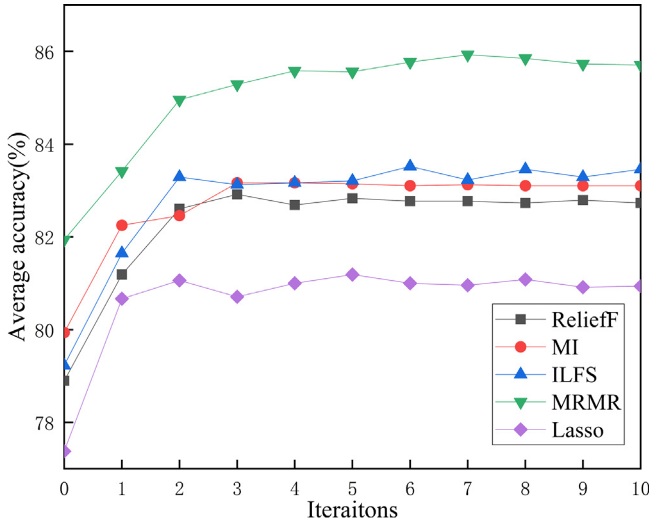


Fig. 7. Comparison of average accuracy after each iteration by selecting the best proportion of features selected and using ATS for OFC.

3.4. *Effectiveness of using threshold in selecting unlabeled data for expanding training dataset*

In SSL for classification, an initial training dataset is used to train a standard FLDA classifier at first, and then unlabeled samples predicted with high confidence are used to expand the training dataset and update the standard FLDA classifier iteration by iteration. In each iteration, a part of the unlabeled data are classified and those samples with predicted labels of high confidence are added to the training dataset, and the standard FLDA classifier is then updated (retrained) using the expanded training dataset and tested on the validation dataset. If we select all samples of the available unlabeled data, which could cause some unreliable samples to be added into the training dataset in the previous iteration and deteriorate the performance of the classifier. In this paper, MNAD criterion (Section 2.5 for details) is proposed to select high confidence ones from unlabeled samples. From Fig. 8, it can be seen that using ILFS method can achieve the best accuracy using PWS for OFC (the proportion of selected features is 30%). Therefore, based on the features selected by ILFS, Fig. 9 compares the average results of using MNAD and without using MNAD (no threshold) after 20 × 5-fold cross-validation, and it is obvious that the average iteration accuracy of using MNAD is better than the average accuracy without using threshold, which denotes the effectiveness of MNAD in selecting high confidence samples.

3.5. *Performance metrics*

In addition to the classification accuracy, F-score (Baldi et al.) and Matthews correlation coefficient (MCC) (Baldi et al.) are used to performance metrics. The metrics results are the average value based OAO multi-class. F-score considers both the precision and recall measures to analyze the accuracy of binary classification. See Eq. (11).

$$F = (1+\beta^2)\frac{precision \cdot recall}{(\beta^2 precision) + recall} \tag{11}$$

When β is equal to 1 the measure is called balanced F-score (F1 score) which is the harmonic mean of precision and recall and takes both precision and recall into account equally.

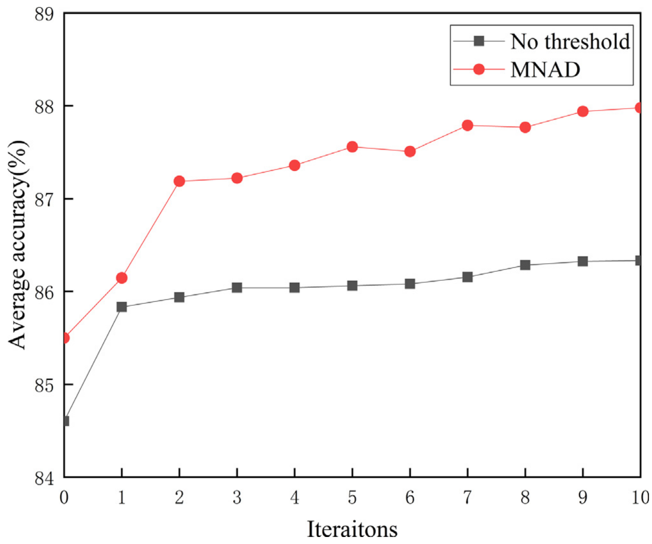


Fig. 9. Comparison of average accuracy using and without using MNAD.

Table 7
Comparison of the metrics.

Method	The classification accuracy (%)	F1 score	MCC
TMSSL-FS-FLDA based PWS	88.49	0.68	0.62
TMSSL-FS-FLDA based ATS	84.56	0.53	0.59

MCC interprets the correlation between the target and prediction in a two class classification. The value of MCC shows the classification ability or total conflict between prediction and target. See Eq. (12).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{12}$$

Where TP represents the True Positive, FP represents the False Positive, TN represents the True Negative and FN represents the False Negative.

Therefore, we compared different metrics in the proposed method based PWS and ATS in Table 7. The results confirm that TMSSL-FS-FLDA based PWS has better classification ability.

3.6. Supervised learning for DGA based fault diagnosis

To further evaluate the performance of TMSSL-FS-FLDA, which uses 75 labeled samples only in the initial training set, in this section two supervised learning methods are used to replace the SSL in TMSSL-FS-FLDA, leading to two new models: Two-stage Multi-filter Static Classification Feature Selection based on FLDA (TMSTC-FS-FLDA) and Two-stage Multi-filter Supervised Classification Feature Selection based on FLDA (TMSUC-FS-FLDA). The supervised learning models use all the available labeled samples as the training set that will not be expanded, and they use the same two-stage multi-filter method for feature selection. The average accuracy of TMSTC-FS-FLDA and TMSUC-FS-FLDA is calculated on 48 validation samples over 20 × 5-fold cross-validation. The major difference between them is that TMSTC-FS-FLDA only uses the initial labeled dataset (75 samples) for training the classifier, TMSUC-FS-FLDA uses the whole training dataset (194 samples), and TMSSL-FS-FLDA uses 75 labeled samples as initial training set and 119 unlabeled samples for expanding the training set.

Fig. 10 and Fig. 11 respectively show the fault diagnosis accuracy with different proportion of selected features. The new ranks for the 34 features based on TMSTC-FS-FLDA and TMSUC-FS-FLDA are given in Appendix B respectively, and the number in Appendix B corresponds to the feature number in Table 4. Fig. 10 shows that the average best accuracy can be achieved by selecting 30% of features based on TMSTC-FS-FLDA for all filter methods. The performance drops significantly when more features are selected. However, the performance of TMSUC-FS-FLDA peaks by selecting 30% of features and only slightly drops with more features selected, as shown in Fig. 11. This demonstrates that redundant features have little influence on the performance of TMSUC-FS-FLDA, but have great influence on TMSTC-FS-FLDA and intermediate impact on TMSSL-FS-FLDA as shown in Fig. 5. Therefore, it can be concluded that the disturbance degree of redundant features to the model is related to the number of training samples. The average iteration accuracy of TMSTC-FS-FLDA and TMSUC-FS-FLDA after completing PWS (the best accuracy based on certain proportion for each filter method) and ATS based on different filter methods are presented in Fig. 12 and Fig. 13, respectively. It is noteworthy that PWS method can obtain better iteration accuracy than ATS for TMSTC-FS-FLDA based all filter methods, and PWS method can achieve better iteration accuracy than ATS for TMSUC-FS-FLDA based the majority of filter methods. In addition, ReliefF based on PWS achieved the best accuracy of 85.02% among tested filter methods for TMSTC-FS-FLDA and ReliefF based on ATS achieved the best accuracy of 89.21% among tested filter methods for TMSUC-FS-FLDA. The selected features based on ReliefF for TMSTC-FS-FLDA after PWS and based on ReliefF for TMSUC-FS-FLDA after ATS are listed in Table 5. After aggregating feature ranks and voting for all filter methods, the OFC based PWS and ATS is obtained and shown in Table 6.

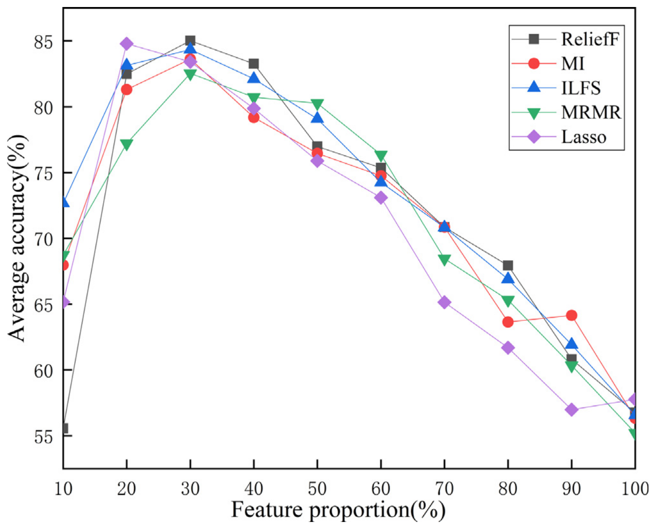


Fig. 10. Comparison of average accuracy of TMSTC-FS-FLDA with different proportions of features selected and using PWS for OFC.

3.7. Comparison of supervised and semi-supervised learning for DGA based fault diagnosis

From Figs. 5, 10 and 11, it can be seen that ILFS combined with PWS achieved the best accuracy for TMSSL-FS-FLDA, ReliefF combined with PWS and ReliefF combined with ATS worked the best for TMSTC-FS-FLDA and TMSUC-FS-FLDA, respectively.

In the stage of the feature ranking, Fig. 14 compares the average accuracy of supervised and semi-supervised learning algorithms using 20 × 5-fold cross-validation based on 34 features selected in the first stage for feature ranking. It is clear that TMSSL-FS-FLDA outperformed TMSTC-FS-FLDA, but achieved lower accuracy than TMSUC-FS-FLDA when more unlabeled samples were used to train classifier. The p values between TMSSL-FS-FLDA and TMSTC-FS-FLDA, TMSSL-FS-FLDA and TMSUC-FS-FLDA are 0.2594 and 0.6224, as shown in Table 8, demonstrating no statistically significant difference.

In order to compare the fault diagnosis accuracy of supervised and semi-supervised learning algorithms under different situations for feature selection, Fig. 15 compares the performance of TMSSL-FS-FLDA,

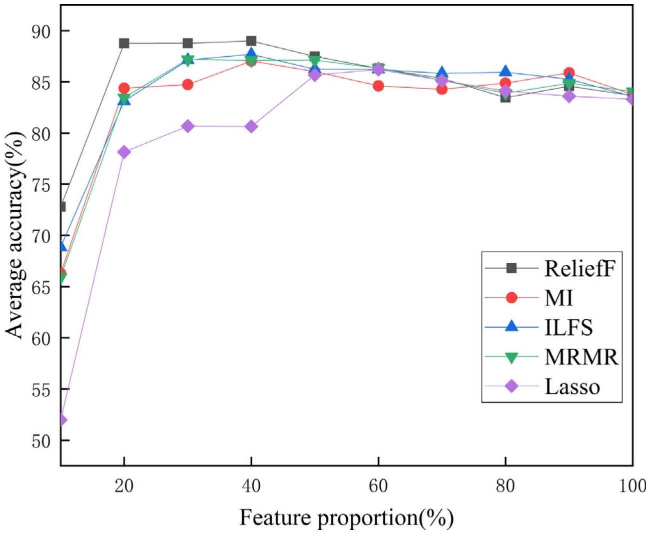


Fig. 11. Comparison of average accuracy of TMSUC-FS-FLDA with different proportions of features selected and using PWS for OFC.

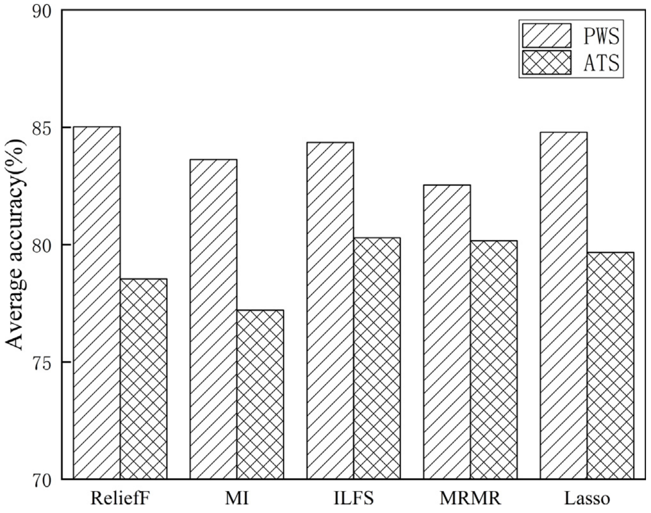


Fig. 12. Comparison of average accuracy of TMSTC-FS-FLDA using PWS or ATS for OFC.

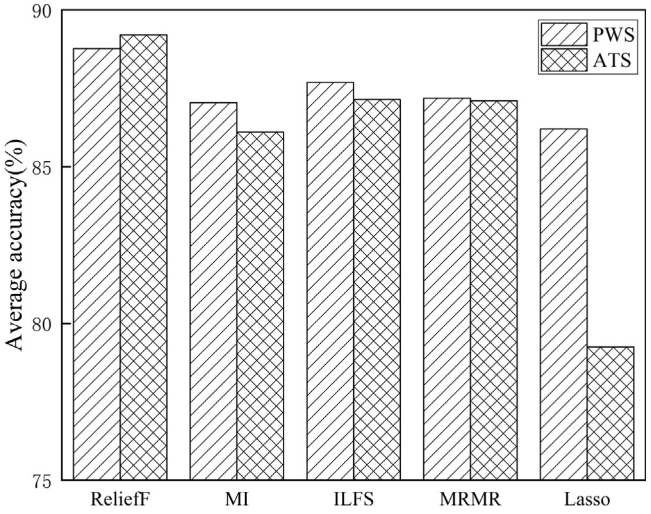


Fig. 13. Comparison of average accuracy of TMSUC-FS-FLDA using PWS or ATS for OFC.

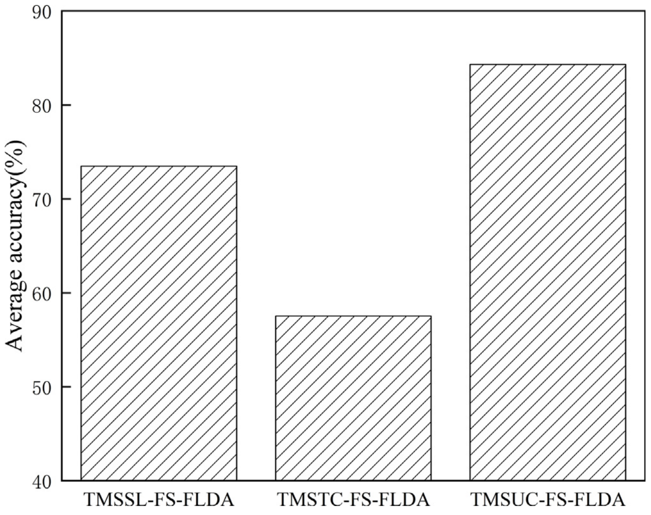


Fig. 14. Comparison of average accuracy of supervised and semi-supervised learning algorithms based on 34 features after 20 × 5-fold cross-validation.

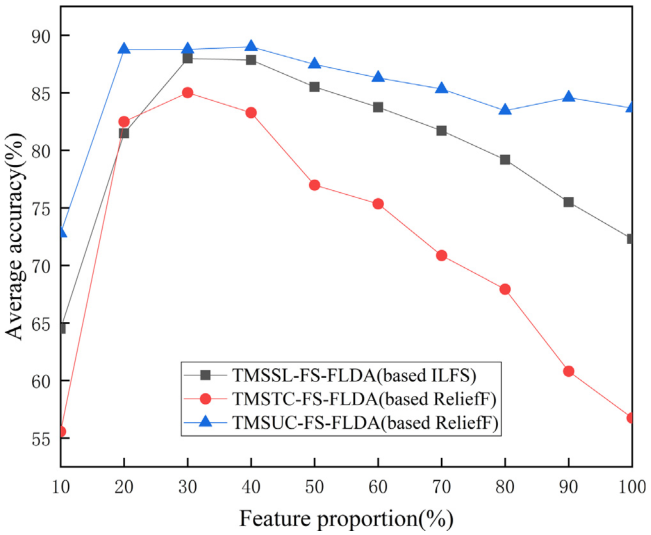


Fig. 15. Comparison of the performance of TMSSL-FS-FLDA, TMSTC-FS-FLDA and TMSUC-FS-FLDA with different proportions of features selected and using PWS for OFC.

TMSTC-FS-FLDA and TMSUC-FS-FLDA with different proportions of features selected and using PWS for OFC. It can be observed that TMSUC-FS-FLDA achieved the best accuracy consistently with different proportions of features selected. The accuracy achieved by TMSSL-FS-FLDA is close to that of TMSUC-FS-FLDA when 30% of highly ranked features were used to build the fault diagnosis model.

Fig. 16 compares the best accuracy of supervised and semi-supervised algorithms using PWS or ATS for OFC. It is obvious that TMSUC-FS-FLDA achieved the highest fault diagnosis accuracy because of using more labeled samples. However, TMSSL-FS-FLDA selected better features(especially with PWS)and its accuracy is close to TMSUC-FS-FLDA with much fewer labeled samples used for training the classifier. It outperformed TMSTC-FS-FLDA in the case of using the same number of labeled samples. Generally methods using PWS can achieve better accuracy than using ATS, which is related to the selected features as shown in Table 5. It is worth noticing that features numbered 26–34 were selected by most algorithms using PWS or ATS, which demonstrates that these features are critical for improving the fault diagnosis accuracy no matter for semi-supervised or supervised method. However, too many features can deteriorate the performance of the

Table 8
Comparison of different feature combinations in terms of p-values.

Features	p1	p2
34 features	0.2594	0.6224
After PWS	0.0299*	0.9904
After ATS	0.0422*	0.5724
After using feature combination method based on PWS	0.0058*	0.8635
After using feature combination method based on ATS	0.0063*	0.0884

*The statistically significant p-values (<0.05).
More details about p-values can be found in Goodman (1999). p1 and p2, respectively, represents the p-values between TMSSL-FS-FLDA and TMSTC-FS-FLDA, TMSSL-FS-FLDA and TMSUC-FS-FLDA over 20 × 5-fold cross-validation.

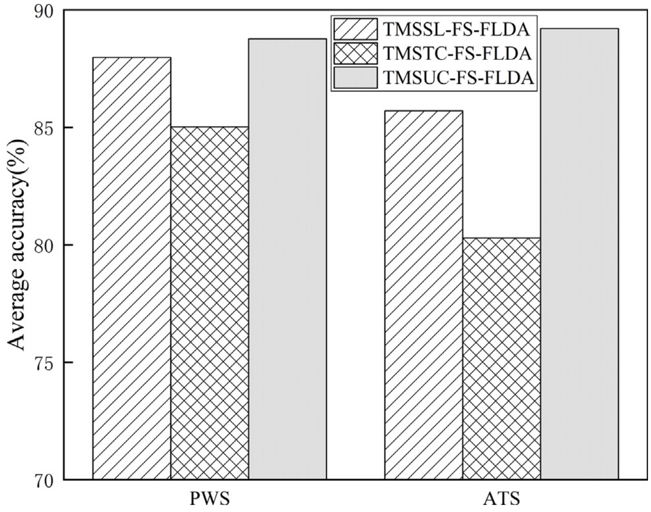


Fig. 16. Comparison of the best average accuracy of supervised and semi-supervised algorithms using PWS or ATS for OFC.

classifier. TMSTC-FS-FLDA based on ATS selected more features but achieved the lowest accuracy. In addition, from the statistical test results as shown in Table 8, the performance of TMSSL-FS-FLDA is significantly better than TMSTC-FS-FLDA, and the performance of TMSUC-FS-FLDA is not significantly better than TMSSL-FS-FLDA, which demonstrates the validity of the proposed SSL algorithm for feature selection.

Table 9 shows the candidate feature subsets obtained using the proposed feature combination methods in comparison with wrapper methods. The number of features selected by PWS and ATS are 10 and 13 for TMSSL-FS-FLDA, 8 and 14 for TMSTC-FS-FLDA, 14 and 17 for TMSUC-FS-FLDA, respectively. Features selected by PWS are also selected by ATS, and ATS selected more features.

Table 10 compares the results of TMSSL-FS-FLDA, TMSTC-FS-FLDA, and TMSUC-FS-FLDA using the candidate feature subsets obtained by the proposed SSL for feature ranking combined with feature combination methods. It is worth noticing that TMSSL-FS-FLDA always achieved better classification accuracy than TMSTC-FS-FLDA after using feature combination method based on PWS or ATS. The p-values in Table 8 show that the difference is significant, however the difference between the performance of TMSUC-FS-FLDA and TMSSL-FS-FLDA is not statistically significant.

It is obvious that the multi-filter semi-supervised feature selection approach plays a significant role in feature selection. To further evaluate the proposed method, it is compared with a semi-supervised wrapper method called FW-SemiFS (Forward Semi-Supervised Feature Selection) (Ren et al., 2008) and a supervised wrapper method called SFFS (Supervised Sequential Forward Feature Selection) (Ren et al., 2008). Table 11 shows the average accuracy of different methods in pairs and the corresponding p-values. It is shown that the performance of the multi-filter methods is better than that of the wrapper methods no matter for semi-supervised or supervised learning. The OFC based supervised and semi-supervised wrapper methods are obtained and shown in Table 6.

In addition, a computational time complexity about TMSSL-FS-FLDA is O($(2n/C)^3$)≈O$(n)^3$, which is followed by O(n^2) of TMSUC-FS-FLDA, and TMSTC-FS-FLDA has the lowest time complexity (O(n)). Where n and C are the number of training samples and class, respectively. This demonstrates the semi-supervised method has more higher complexity than supervised method because of using more unlabeled samples and spending more time for model training. However, the semi-supervised method also gets the corresponding return in the correct rate.

3.8. Robustness of the proposed method for feature subset selection

Another dataset provided by a national power company is used to test the performance of the feature subsets obtained by the proposed methods, especially their robustness. Features selected by the conventional methods, such as Total gases, Doernenberg, Rogers, IEC 60599 and Duval Triangles, provide different power transformer fault diagnosis performance. Table 12 shows the fault diagnosis performance based on semi-supervised and supervised learning with different feature subsets, including the features selected by the conventional methods as shown in Table 13 those selected by the proposed algorithms using PWS and by wrapper methods as shown in Table 5. 5-fold cross-validation was carried out 20 times with randomly split training and validation data, and the average validation accuracy is used to evaluate the fault diagnosis performance. It can be concluded from Table 12 that the Dornenburg method achieved the worst fault diagnosis performance, while the proposed approach achieved the highest fault diagnosis accuracy overall no matter for semi-supervised or supervised learning. This is also demonstrated in Fig. 17, from which it can be observed that the accuracy of the proposed SSL method improves with more iterations and can outperform the wrapper approach, with feature subsets selected from one dataset applied to another new dataset.

Table 9
Candidate feature subsets obtained using the proposed feature combination methods in comparison with wrapper methods.

Algorithms		Selected features
TMSSL-FS-FLDA	PWS	25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	ATS	22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	FW-SemiFS	5, 8, 9, 14, 15, 19, 20, 25, 27, 32, 33, 34
TMSTC-FS-FLDA	PWS	27, 28, 29, 30, 31, 32, 33, 34
	ATS	12, 13, 15, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	SFFS-STC	1, 2, 6, 11, 12, 13, 15, 16, 26, 29, 30, 34
TMSUC-FS-FLDA	PWS	12, 15, 21, 22, 23, 25, 26, 27, 29, 30, 31, 32, 33, 34
	ATS	12, 13, 14, 15, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34
	SFFS-SUC	1, 2, 10, 11, 12, 13, 14, 15, 25, 32, 33, 34

Table 10
Comparison of results of using the proposed feature combination methods.

Algorithms		After using feature combination method(%)	The best results after PWS and ATS (%)	The mean results of multi-filter after PWS and ATS (%)
TMSSL-FS-FLDA	PWS	88.49	87.86	85.13
	ATS	84.56	85.71	82.99
TMSTC-FS-FLDA	PWS	85.85	85.02	84.07
	ATS	80.14	80.29	79.18
TMSUC-FS-FLDA	PWS	88.98	88.77	87.38
	ATS	87.48	89.21	85.76

Table 11
Comparison of multi-filter and wrapper methods.

Algorithms		Average accuracy	p3
SSL	TMSSL-FS-FLDA	88.49	0.0178*
	FW-SemiFS (Wei et al., 2014)	85.56	
STC	TMSTC-FS-FLDA	85.85	0.0421*
	SFFS-STC	82.29	
SUC	TMSUC-FS-FLDA	88.98	0.9309
	SFFS-SUC	88.22	

p3 represents the *p*-value between multi-filter and wrapper method for semi-supervised and supervised algorithms.

*The statistically significant *p*-value (< 0.05).

Table 12
Comparison of fault diagnosis performance based on different methods.

	Total gases	Dornenburg	Roger	IEC 60599	Duval triangle	The features from wrapper method	The features from our proposed methods based on PWS	p4
SSL	57.73	54.15	65.06	73.12	54.00	85.46	88.19	0.0556
STC	59.88	58.94	67.62	77.62	61.06	81.33	82.98	0.6522
SUC	64.04	57.12	70.87	78.85	62.88	91.75	89.83	0.0853

p4 represents the p-value between multi-filter and wrapper method for semi-supervised and supervised algorithms.

Table 13
Feature sets selected by conventional methods.

Method	Feature sets
Total gases	H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO, CO_2.
Dornenburg	CH_4/H_2, C_2H_2/C_2H_4, C_2H_2/CH_4, C_2H_6/C_2H_2
Roger	C_2H_6/CH_4, C_2H_2/C_2H_4, CH_4/H_2, C_2H_4/C_2H_6
IEC 60599	C_2H_2/C_2H_4, CH_4/H_2, C_2H_4/C_2H_6
Duval triangle	CH_4, C_2H_4, C_2H_2, $CH_4/(CH_4+C_2H_4+C_2H_2)$, $C_2H_4/(CH_4+C_2H_4+C_2H_2)$, $C_2H_2/(CH_4+C_2H_4+C_2H_2)$

4. Limitations

Although iterative training can improve the fault diagnose accuracy in proposed TMSSL-FS-FLDA algorithm, some limitations still exist. The model cannot correct its own mistakes, and the wrong predictions from unlabeled samples will destroy the entire model. In addition, the similarity between labeled and unlabeled samples is high, which makes it difficult to great improve the classifier performance through further iteration. Therefore, the better threshold strategy should be explored for restraining noise from unlabeled samples and the active learning is introduced in combination with semi-supervised learning to select informative samples for improve the classifier performance in the future work section.

5. Conclusion and future directions

This paper proposes a Two-stage Multi-filter SSL Feature Selection method based on FLDA (TMSSL-FS-FLDA) for building fault diagnosis model and selecting optimal DGA features with limited labeled data samples. The method includes two stages in which feature ranking and aggregating feature ranks for obtaining a candidate feature subset are performed respectively. The validity of the TMSSL-FS-FLDA has been proved by experiments on dataset IEC TC 10. The results show that: (1) Features referred to CO and CO_2 are not important in the process of feature selection. (2) TMSSL-FS-FLDA has much stronger ability to select OFC and improve fault diagnosis accuracy than TMSTC-FS-FLDA when using the same number of labeled samples, and considerable ability when compared with TMSUC-FS-FLDA by using fewer labeled samples. (3) PWS method can obtain better iteration accuracy than ATS for almost all filter criteria. In addition, the ability of feature combination using PWS is also superior to ATS no matter semi-supervised or supervised learning is adopted. (4) The effectiveness of MNAD for selecting high-confident samples in the progress of SSL is proved. (5) The features ranging from number 26 to number 34 (including the partial gas radio and gas relative percentage) are critical for improving fault diagnosis accuracy no matter semi-supervised or supervised method is adopted. (6) The robustness of the obtained optimal feature subset is validated by a test using DGA samples provided by a national power company compared with features selected by conventional and wrapper methods. In the future work, more DGA data should be investigated and collected for algorithm training and testing. New semi-supervised feature selection methods and threshold criteria need to be further studied. In addition, more informative features and other semi-supervised methods should be explored to improved DGA fault diagnose classification in further.

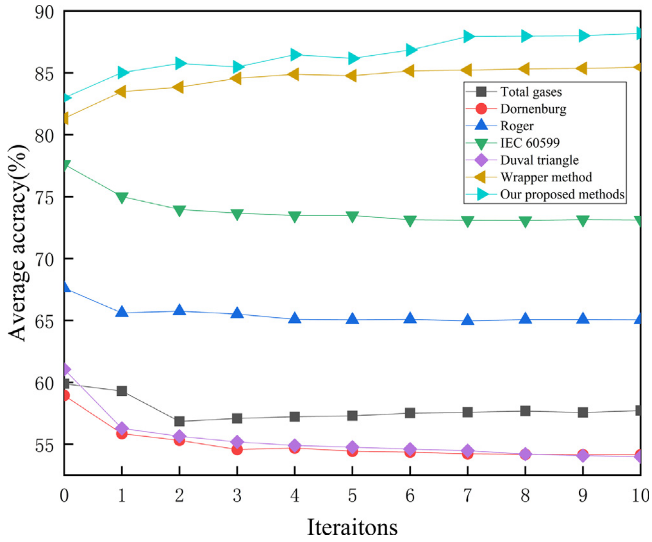


Fig. 17. Comparison of average accuracy after each iteration with different feature combinations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

These data were derived from the following resources available in the public domain: M. Duval, A. Depabla, “Interpretation of gas-in-oil” analysis using new IEC publication 60599 and IEC.

Acknowledgments

This work was supported by Research Initiation Project of Introducing Talents of Chengdu University of Information Technology, China (KYTZ201902 and 376252), Sichuan Science and Technology Program, China (2021YFG0215, 2021YFSY0052, 2023YFQ0029 and 2023YFSY0036), and Natural Science Foundation, China (51977017).

Appendix A

See Table A.14.

Appendix B

See Table B.15.

Table A.14
The abbreviation of terms.

Term	Abbreviation	Term	Abbreviation
Dissolved Gas Analysis	DGA	Low-energy Discharge	LED
Clustering based Method	CBT	High-energy Discharge	HED
Fuzzy Logic Inference System	FLIS	Low and Middle-temperature overheat	LMT
Artificial Neural Network	ANN	High Temperature overheating	HT
Support Vector Machine	SVM	Partial Discharge	PD
Fisher Linear Discriminant Analysis	FLAD	Normal	N
Semi-supervised Learning	SSL	Synthetic Minority Oversampling Technique	SMOTE
Supervised Learning	SL	Adaptive Synthetic Sampling Algorithm	ADASYN
Low-dimensional Scaling	LDS	Arctangent Transformation	AT
Local Linear Semi-supervised Regression	LLSSR	Max-Relevance and Min-Redundancy	MRMR
Adaptive Reinforcement	AR	Least absolute shrinkage and selection operator	Lasso
Mutual Information	MI	Multi-class Nearest Average Distance	MNAD
Infinite Latent Feature Selection	ILFS	Matthews Correlation Coefficient	MCC
Proportion Wrapper Strategy	PWS	Two-stage Multi-filter Static Classification Feature Selection based on FLDA	TMSTC-FS-FLDA
Average Threshold Strategy	ATS	Two-stage Multi-filter Supervised Classification Feature Selection based on FLDA	TMSUC-FS-FLDA
Optimal Feature Combination	OFC	Forward Semi-Supervised Feature Selection	FW-SemiFS
Two-stage Multi-filter SSL Features Selection based on FLDA	TMSSL-FS-FLDA	Supervised Sequential Forward Feature Selection	SFFS
One-against-one	OAO		

Table B.15
Feature ranks obtained by different filter methods.

Feature ranks	TMSSL-FS-FLDA					TMSTC-FS-FLDA					TMSUC-FS-FLDA				
	ReliefF	MI	ILFS	MRMR	Lasso	ReliefF	MI	ILFS	MRMR	Lasso	ReliefF	MI	ILFS	MRMR	Lasso
1	31	32	32	32	1	30	31	29	32	1	29	32	31	32	1
2	30	33	33	31	34	31	32	30	31	34	30	33	30	33	13
3	32	34	31	33	15	28	30	32	30	2	26	34	32	34	14
4	29	31	30	30	14	29	33	31	33	13	28	20	33	31	34
5	28	30	29	34	16	32	29	33	29	14	27	30	29	22	15
6	33	29	27	29	13	27	28	28	28	15	21	31	28	23	12
7	27	23	28	27	2	26	27	24	34	33	31	15	13	15	16
8	26	11	34	28	33	34	23	21	26	12	25	22	12	30	23
9	23	27	21	24	12	25	12	34	24	16	20	14	27	12	10
10	25	22	26	26	9	33	34	27	14	11	22	21	34	21	2
11	34	25	23	22	25	24	25	25	25	26	34	25	26	24	25
12	20	13	22	25	32	18	10	18	22	29	33	24	11	16	32

(*continued on next page*)

Table B.15 (*continued*).

Feature ranks	TMSSL-FS-FLDA					TMSTC-FS-FLDA					TMSUC-FS-FLDA				
	ReliefF	MI	ILFS	MRMR	Lasso	ReliefF	MI	ILFS	MRMR	Lasso	ReliefF	MI	ILFS	MRMR	Lasso
13	15	26	25	18	10	19	26	15	27	32	24	23	7	26	11
14	22	12	20	21	11	21	11	16	15	9	12	12	22	25	27
15	24	24	24	13	17	20	24	13	13	10	32	26	10	14	24
16	14	14	18	14	3	17	22	12	12	3	15	10	14	13	26
17	21	21	19	17	6	22	21	20	21	28	19	11	6	29	30
18	17	20	8	23	26	23	13	17	23	24	14	13	9	27	17
19	9	28	12	15	4	15	14	22	10	4	11	27	8	11	29
20	16	15	11	12	23	12	15	26	11	8	18	5	5	9	6
21	19	16	6	20	31	16	20	19	16	7	10	28	23	4	33
22	13	10	4	19	8	13	5	7	9	23	17	29	25	1	28
23	8	5	17	16	28	9	6	23	1	27	13	1	21	28	31
24	7	18	15	3	29	14	1	10	3	17	9	6	2	10	9
25	11	1	3	1	5	11	8	14	20	30	23	4	19	20	22
26	18	9	10	11	27	3	16	6	19	6	16	9	20	6	7
27	10	7	14	4	30	5	9	8	4	31	4	8	24	5	5
28	6	8	1	10	24	2	7	11	18	25	5	7	16	3	8
29	12	4	7	5	22	4	4	9	5	5	6	16	15	8	3
30	5	6	9	8	7	6	19	5	17	22	3	3	4	2	4
31	2	19	13	6	18	8	18	3	8	18	2	18	18	19	21
32	3	3	16	7	19	10	2	4	7	19	8	2	17	17	18
33	4	2	2	9	21	7	17	2	6	21	7	17	1	18	19
34	1	17	5	2	20	1	3	1	2	20	1	19	3	7	20

References

Atoui, M.A., Cocquempot, V., 2021. New decision rules for Fisher discriminant analysis: applied to fault diagnosis. In: 2021 European Control Conference. pp. 2298–2303.

Baldi, P., Brunak, S., Chauvin, Y., et al., Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16 (5), 412–424.

Boztas, G., Tuncer, T., 2022. A fault classification method using dynamic centered one-dimensional local angular binary pattern for a PMSM and drive system. Neural Comput. Appl. 34, 1981–1992.

Cai, L.H., et al., 2021. Improved cancer biomarkers identification using network-constrained infinite latent feature selection. Plos One 16 (2).

Chen, S.P., 2016. A kind of semi-supervised classifying method research for power transformer fault diagnosis. In: 2016 7th IEEE International Conference on Software Engineering and Service Science. ICSESS, Beijing, pp. 1013–1016.

Chen, X., Yuan, G., Nie, F., Ming, Z., 2020. Semi-supervised feature selection via sparse rescaled linear square regression. IEEE Trans. Knowl. Data Eng. 32 (1), 165–176.

Duval, M., Depabla, A., 2001. Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases. IEEE Electr. Insul. Mag. 17 (2), 31–41.

Goodman, S.N., 1999. Toward evidence-based medical statistics. 1: the P value fallacy. Annal. Internal Med. 130 (12), 995–1004.

Gouda, O.E., El-Hoshy, S.H., El-Tamaly, H.H., 2019a. Condition assessment of power transformers based on dissolved gas analysesx. IET Gener. Trans. Distrib. 13 (12), 2299–2310.

Gouda, O.E., El-Hoshy, S.H., El-Tamaly, H.H., 2019b. Proposed heptagon graph for DGA interpretation of oil transformersx. IET Gener. Trans. Distrib. 12 (2), 490–498.

Gouda, O.E., El-Hoshy, S.H., Ghoneim, S.S.M., 2021. Enhancing the diagnostic accuracy of DGA techniques based on IEC-TC10 and related databases. IEEE Access 9, 118031–118041.

Gu, X.Y., et al., 2022. Conditional mutual information-based feature selection algorithm for maximal relevance minimal redundancy. Appl. Intell. 52 (2), 1436–1447.

Han, Y.K., Park, K., Lee, Y.K., 2011. Confident wrapper-type semi-supervised feature selection using an ensemble classifier. In: 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Deng Feng, China. pp. 4581–4586.

Huang, X.G., et al., 2018. A novel fault diagnosis system on polymer insulation of power transformers based on 3-stage GA-SA-SVM OFC selection and ABC-SVM classifier. Polymers 10 (10).

Jamshed, A., Chatterjee, K., Haque, N., 2021. Random forest classifier based dissolved gas analysis for identification of power transformer faults using gas ratio data. In: 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India. pp. 1–5.

Jovic, A., Brkic, K., Bogunovic, N., 2015. A review of feature selection methods with applications. Mipro Proc. 1–6.

Khan, S.M., Umair Alam, M., Khan, A.A., Farooq, O., 2021. A preliminary study on relieff based feature ranking for classification of myoelectric signals. In: 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII), Chennai, India. pp. 1–5.

Koroglu, S., Demircali, A., 2016. Diagnosis of power transformer faults based on multi-layer support vector machine hybridized with optimization methods. Electr. Power Compon. Syst. 44 (19), 2172–2184.

Li, S.B., Wu, G.N., Gao, B., et al., 2016. Interpretation of DGA for transformer fault diagnosis with complementary SaE-ELM and arctangent transform. IEEE Trans. Dielectr. Electr. Insul. 23 (1), 586–595.

Ma, G.M., et al., 2021. Optical sensors for power transformer monitoring: A review. High Volt. 6 (3), 367–386.

Mao, W., et al., 2022. Fault diagnosis for power transformers through semi-supervised transfer learning. Sensors 22 (12).

Mawelela, T.U., Nnachi, A.F., Akumu, A.O., Abe, B.T., 2020. Fault diagnosis of power transformers using duval triangle. In: 2020 IEEE PES/IAS PowerAfrica, Nairobi, Kenya. pp. 1–5.

Mirowski, P., LeCun, Y., 2012. Statistical machine learning and dissolved gas analysis: A review. IEEE Trans. Power Deliv. 27 (4), 1791–1799.

Ren, J., et al., 2008. Forward semi-supervised feature selection. In: Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. pp. 970–976.

Rogers, R.R., 1978. IEEE and IEC codes to interpret incipient faults in transformers, using gas in oil analysis. IEEE Trans. Electr. Insul. 13 (5), 349–354.

Sheikhpour, R., 2017. A survey on semi-supervised feature selection methods. Pattern Recognit. 64, 141–158.

Song, Z., Yang, X., Xu, Z., et al., 2022. Graph-based semi-supervised learning: A comprehensive review. IEEE Trans. Neural Netw. Learn. Syst. 1–21.

The Institute of Electrical and Electronics Engineers, 1992. IEEE guide for the interpretation of gases generated in oil-immersed transformers. IEEE Stand. C 57, 104–1991.

Thomas, B., Chaudhari, S.G., K.V., Shihabudheen, Verma, N.K., 2023. CNN-based transformer model for fault detection in power system networks. IEEE Trans. Instrum. Meas. 72, 1–10.

Wang, T., Zhang, L., Wang, X.F., 2023. Fault detection for motor drive control system of industrial robots using CNN-LSTM-based observer. CES Trans. Elect. Mach. Syst. 1–9.

Wei, C., Guo, C., Yan, W., 2021. Forest fire risk forecast method with pseudo label based on semi-supervised learning. In: 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China. pp. 36–39.

Wei, C.H., Tang, W.H., Wu, Q.H., 2014. Dissolved gas analysis method based on novel feature prioritisation and support vector machine. IET Electr. Power Appl. 8 (8), 320–328.

Yang, P.Y., Zhou, B.B., Zhang, Z.L., et al., 2010. A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. BMC Bioinf. 11, 1–12.

Zhang, Z., Li, J., 2022. Synthetic minority oversampling technique based on adaptive local mean vectors and improved differential evolution. IEEE Access 10, 74045–74058.

Zheng, H.B., et al., 2011. Fault diagnosis of power transformers using multi-class least square support vector machines classifiers with particle swarm optimisation. IET Electr. Power Appl. 5 (9), 691–696.