Preprints of the 4th IEEE Conference on Energy Internet and Energy System Integration
October 30- November 1, 2020. Wuhan, China

A Systematic Framework Presenting Impact of Dataset Completeness on Data Driven Approaches: A Transmission Line Fault Classification Study

Yiqi Xing
School of Information Science and Technology, ShanghaiTech University
Shanghai, China

Yu Liu*
School of Information Science and Technology, ShanghaiTech University
Shanghai, China
liuyu.shanghaitech@gmail.com

Xiaodong Zheng
Department of Electrical Power Engineering, Shanghai Jiao Tong University
Shanghai, China

2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2) | 978-1-7281-9606-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/EI250167.2020.9346712

***Abstract*—Fault classification is one of the essential steps to ensure safety and power supply reliability of the transmission lines. This paper focuses on the data driven transmission line fault classification approaches and studies the impact of dataset completeness on the validity of data driven approaches. A systematic framework is proposed to verify this impact. Here three data driven approaches, including SVM, Softmax and BPNN, are considered as examples. First, a small dataset and a full dataset are established, which contain partial and complete characteristics of the fault classification problem, respectively. Afterwards, each dataset is randomly divided into training, validation and testing sets to generate the trained model. Next, an additional testing set is constructed to further validate the fault classification accuracy of the trained model. Since the additional testing set is different from both the small and the full dataset, it provides an objective criterion for the effectiveness of the trained model. Numerical experiments in a 500 kV transmission line system prove that the small dataset may result in overfitting and generalization issues, while the full dataset can fully demonstrate the effectiveness of a certain data driven approach.**

Keywords—fault classification, dataset completeness, data driven approaches, backpropagation neural network (BPNN)

I. INTRODUCTION

Transmission lines are essential components of modern power systems. Sometimes faults may occur inside transmission lines [1-2]. Transmission line fault classification is one of the important steps after the occurrence of the fault, since it can enable more reliable protection schemes and more accurate fault location schemes, and therefore can improve power system safety and power supply reliability.

Fault classification approaches utilize voltage and current measurements at one or both terminals of the transmission line of interest. Traditional fault classification approaches empirically extract the features from the available measurements. Nevertheless, these features are usually dependent on the parameters of the system with faults, including fault locations, fault resistances, fault time, system loading conditions, source impedances, etc., resulting in compromised fault classification accuracy [3].

This work is sponsored by National Nature Science Foundation of China (No. 51807119) and Shanghai Pujiang Program (No. 18PJ1408100). Their support is greatly appreciated.

Therefore, in order to solve this problem, researchers proposed data driven approaches. These methods systematically extract features from available measurements and are independent of variation of parameters of the system with faults. The methods include fuzzy logic, decision tree, Support Vector Machine (SVM), Artificial Neural Networks (ANN), etc. For example, references [4] and [5] propose an adaptive neuro-fuzzy inference system approach and a fuzzy logic based multi-criteria approach to achieve real time fault classification in transmission systems. Reference [6] proposes an ANN and SVM approach used for the fault classification in radial distribution systems. Reference [7] and [8] propose decision tree based fault identification and classification approaches.

In order to further improve the fault classification results, researchers also proposed methods that combine advanced signal processing techniques with the data driven methods. When a fault occurs inside the transmission line, the voltage and current waveforms contain rich information (sudden changes, characteristics in frequency spectrum, etc.) that can be extracted using advanced mathematical tools. The performances of the data driven approaches could be improved with the extracted information. For example, reference [9] proposes a hybrid scheme using a Fourier Linear Combiner and a fuzzy expert system to determine the types of disturbances. Reference [10] first extracts fault features using dynamic state estimation and afterwards adopt SVM for training. Reference [11] proposes a transmission line fault classification method with wavelet transform and SVM, where the features extracted by wavelet transform are sent to the data driven classifier for training. Similarly, reference [12] proposes a fault classification approach that combines wavelet transform and ANN, where the low frequency approximations obtained from the wavelet transform are sent to the ANN for training.

Nevertheless, the validity of data driven approaches highly depends on the quality of the data. If the data for training cannot fully represent the characteristics of the problem, the trained model may still encounter issues in practice even with advanced data driven approaches. In other words, if one would like to verify the effectiveness of a certain proposed data driven approach, the completeness of the training dataset is vital to avoid overfitting and generalization issues and to ensure practicability of the

method. However, this issue is not fully investigated in existing literatures.

In this paper, the impact of the dataset completeness on data driven approaches is studied via a systematic framework. Here three data driven approaches, including SVM, linear classifier with softmax loss function (Softmax) and backpropagation Neural Network (BPNN), are selected as examples. Each data driven approach is validated via a small (incomplete) dataset and a full (complete) dataset. For each dataset, the training, validation and testing sets are randomly selected to obtain the trained model. Afterwards, to further ensure that the trained model is applicable in practice, an additional testing set, which differs from both the small and the full dataset, is generated and adopted for validation. Numerical experiments in a 500 kV transmission line system prove that data driven approaches could appear to perform well with small dataset but in fact not applicable in practice. On the other hand, the validity of data driven approaches can be fully demonstrated with full dataset. In addition, numerical experiments prove that BPNN presents much higher fault classification accuracy than SVM and Softmax.

The rest of the paper is organized as follows. Section II describes the fundamentals of the three data driven classifiers. Section III introduces the influence of the dataset completeness on the validity of the data driven approaches and presents a systematic framework to demonstrate this influence. Section IV demonstrates the results of the numerical experiments. Section V draws a conclusion.

II. Classifier Fundamentals

This section reviews the fundamentals of the data driven classifiers. First, the principles of two linear classifiers, SVM and Softmax, are introduced. Afterwards, the principle and the structure of BPNN are introduced.

A. SVM

SVM is a linear classifier which tries to find the hyper-plane that separates the different classes. Fig.1 shows a two-dimensional example with two separated datasets and a hyper-plane (represented by the solid line) found by the linear classifier. The broken line in the figure is the boundary, and the distance between them is the margin. The plus and minus signs indicate two categories. In the figure, (x_i, y_i) represents one data point, and variables w and b represent the weight and the bias of the linear classifier, respectively.

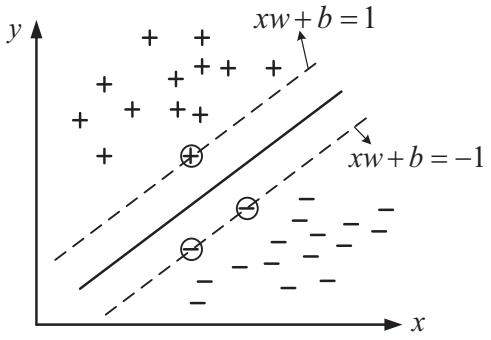

Fig. 1. Separable data set and hyper-plane

The key idea of the linear classifier is to find a hyper-plane (or w and b) that can maximize the distance between the hyper-plane and the nearest data point. Specifically, the SVM method requires that the scoring function of the correct category should be at least Δ higher than that of the wrong categories. Therefore, the SVM loss function for one data point (x_i, y_i) is defined as (hinge loss),

$$L_i = \sum_{j \neq y_i} \max(0, f(x_i; w)_j - f(x_i; w)_{y_i} + \Delta) \tag{1}$$

where f_j is the j^{th} element of the vector f, and f is,

$$f(x_i; w) = x_i w + b \tag{2}$$

Afterwards, the total loss function can be obtained by summing the loss function of the individual data points. In addition, to ensure generalization, the L2 regularization is introduced to the loss function with a hyperparameter λ. The problem can be formulated by minimizing the entire loss function, as shown in (3). Finally, the gradient descent method can be applied to solve (3).

$$\min_{w,b} L = \frac{1}{N} \sum_i L_i + \lambda \sum_k \sum_l w_{k,l}^2 \tag{3}$$

B. Softmax

Softmax is also a linear classifier that is very similar as SVM. Instead of using the hinge loss in (1), the Softmax classifier utilizes the cross-entropy loss, as follows,

$$L_i = -\log\left(e^{f_{y_i}} \Big/ \sum_j e^{f_j}\right) \tag{4}$$

where the definitions of f and f_j are the same as in SVM.

Afterwards, similar as SVM, the method formulates the optimization problem in (3) (after substituting (4) into (3)) and solves the problem using the gradient descent method.

C. BPNN

The structure of an example two-layer BPNN is shown in Fig. 2. The network includes one input layer, one hidden layer and one output layer. The number of hidden layers could be more for multi-layer BPNN. M is the number of data points, H is the number of units inside the hidden layer, C is the number of categories, and N is the latitude value after converting each input into a row vector.

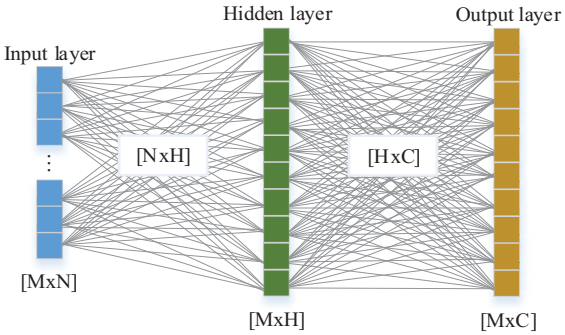

Fig. 2. The structure of an example two-layer BPNN

The structures between two consecutive layers in BPNN are similar as in linear classifiers. Nevertheless, in order to present nonlinearity in the BPNN, activation functions are usually adopted at the output of the hidden layer. Here the *ReLU* function is introduced as an example,

$$relu(x) = \max(0, x) \tag{5}$$

After calculating function f using the input x_i and the coefficients at each layer (namely w_1, b_1, w_2, b_2, etc.), similar as (3) and (4), a loss function can be established with the idea of Softmax classifier. The idea is still to use gradient descent method to solve (3). Here unlike SVM or Softmax where the gradients can be easily calculated, the calculation of gradients

for the neutral network is more challenging with the activation function. Therefore, the backpropagation algorithm can be applied to recursively calculate the gradient using the chain rule.

III. INFLUENCE OF DATASET COMPLETENESS ON THE VALIDITY OF DATA DRIVEN APPROACHES

The validity of data driven approaches highly depends on the quality of the dataset. Since the data driven approaches automatically extract features of a certain dataset through the training process, if the dataset cannot represent the entire characteristics of the problem, the data driven approaches may encounter overfitting and generalization issues, and therefore compromised results. For example, if the dataset is incomplete, the data driven approach may seem to perform well within the incomplete dataset (overfitting); however, when applied to practical systems where the scenarios are rather complete, the performance may be much degraded (lack of generalization).

Therefore, this paper presents a systematic framework to demonstrate the influence of the dataset completeness on the validity of data driven approaches. The flow chart of the systematic framework is provided in Fig. 3. Here two datasets are generated. The first one is a small (incomplete) dataset that covers only part of the characteristics of the problem, while the second one is a full (complete) dataset that covers the entire characteristics of the problem (the first dataset is generally a subset of the second dataset). Afterwards, each dataset is further randomly divided into the training set, validation set and testing set. The training set and the validation set is utilized to generate the trained model (the validation set is for tuning the hyperparameters of the corresponding data driven approach). After obtaining the trained model, the testing set is adopted to generate results with small dataset training and small data set testing. To further validate whether the trained model is applicable in practice (to consider overfitting and generalization issues), an additional testing set is generated, which is different from both the small dataset and the full dataset. The generation of this additional testing set ensures that the validity of the trained model can be tested more objectively. Finally, the testing results of the trained model with the additional testing set are recorded and analyzed.

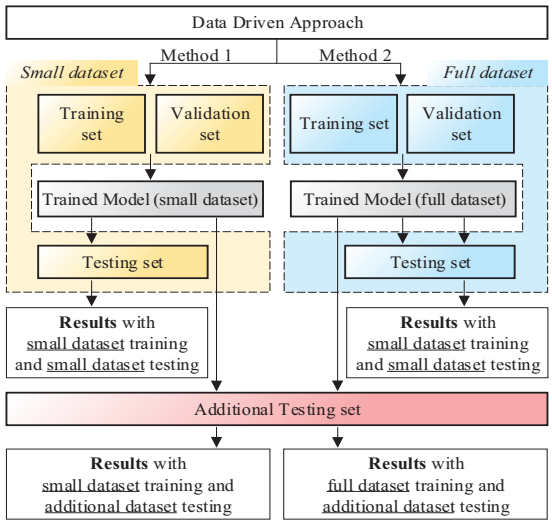

Fig. 3. Flow chart of the systematic framework to demonstrate the influence of dataset completeness on the validity of data driven approaches

Specifically, in this paper the transmission line fault classification problem is taken as an example. Transmission line faults may vary according to different fault types, fault resistances, fault location and fault time. Besides, the operating conditions of system prior to the transmission line faults may also vary, including different source impedances, and different phase angles between sources at two terminals. Additionally, there will also be noises for the measurements in practice.

Here a small dataset, a full dataset and an additional testing set are generated in an example 50Hz, 500kV/200km three phase transmission line system. The system consists of two power sources at each terminal of the transmission line. Three phase instantaneous current measurements are installed at one terminal of the line. The sampling rate is selected according to IEC61850-9-2 standard as 80 samples/cycle (4k samples/second for the 50Hz system). The available time window is one cycle after the occurrence of the fault, corresponding to 243 values in total (3 phases times 81 samples for each phase).

The small dataset considers different fault types, fault resistances, fault locations and measurement noises, but with fixed fault time, source impedances and phase angle difference between the two sources. The full dataset considers different fault types, fault resistances, fault locations, measurement noises, fault time, source impedances and phase angle differences. The additional testing set considers additional events, including 6 additional phase angle differences, 6 additional fault locations, and 6 additional fault resistances (other variables remain unchanged). Note that events with high fault resistances of 100, 200, 300, 400 and 500 ohms are only considered for single phase to ground faults. Different events in small dataset, full dataset and additional testing sets are shown in Table. I, Table. II, and Table III, with 1188 events, 109620 events and 1884 events in total, respectively.

TABLE I. EVENTS IN SMALL DATASET

Variables	Values
Fault type	AG, BA, CG, AB, AC, BC, ABG, ACG, BCG, ABC
Fault resistance	0.01 Ω, 1 Ω, 10 Ω, 100 Ω, 500 Ω
Fault location	5km, 20km, 40km, 60km, 80km, 100km, 120km, 140km, 160km, 180km, 195km
Fault time	0.5s
Source impedance	1+j5 Ω
Phase angle difference	10°
Gaussian noise	0%, 2%, 5%

TABLE II. EVENTS IN FULL DATASET

Variables	Values
Fault type	AG, BA, CG, AB, AC, BC, ABG, ACG, BCG, ABC
Fault resistance	0.01 Ω, 1 Ω, 10 Ω, 100 Ω, 500 Ω
Fault location	5km, 20km, 40km, 60km, 80km, 100km, 120km, 140km, 160km, 180km, 195km
Fault time	0.5s, 0.502s, 0.504s, 0.506s, 0.508s, 0.51s, 0.512s, 0.514s, 0.516s, 0.518s
Source impedance	1+j5 Ω, 3+j15 Ω, 10+j50 Ω
Phase angle difference	10°, 30°, 50°
Gaussian noise	0%, 2%, 5%

TABLE III. EVENTS IN ADDITIONAL TESTING SET

Variables	Values		
Descriptions of additional testing set	Phase angle difference additional testing set: 15°, 25°, 35°, 45°, 55°, 60°	Fault location additional testing set: 25km, 65km, 105km, 115km, 145km, 175km	Fault resistance additional testing set: 15 Ω, 30 Ω, 50 Ω, 200 Ω, 300 Ω, 400 Ω
Fault type	AG, BA, CG, AB, AC, BC, ABG, ACG, BCG, ABC		
Fault resistance	0.01 Ω, 1 Ω, 10 Ω	0.01 Ω, 1Ω, 10 Ω	/
Fault location	50km	/	100km
Source impedance	1+j5Ω, 3+j15Ω, 10+j50 Ω	1+j5Ω, 3+j15Ω	1+j5 Ω, 3+j15 Ω
Phase angle difference	/	10°	10°
Fault time	0.5s, 0.502s	0.5s, 0.502s	0.5s, 0.502s

IV. NUMERICAL EXPERIMENTS

In this section, we adopt the small data set, the full dataset, and the additional testing set defined in section III to verify the influence of the dataset completeness on the validity of data driven approaches. The three data driven approaches under test are SVM, Softmax and BPNN, respectively. Prior to the training procedure, both the small dataset and the full dataset need to be randomly divided into three part: training set, validation set and testing set, with the number of events shown in Table IV and Table V, respectively. According to the definitions of the fault classification problem and available data, there are 243 inputs and 10 outputs for each data driven approach (N = 243 and C = 10 in Fig. 2). Specifically for BPNN, a two-layer network is utilized, where the dimension of the hidden layer is set as 10 (H = 10 in Fig. 2). Next, the experiments are conducted according to Fig. 3.

TABLE IV. SMALL DATASET GROUPING

Method	training set	validation set	testing set
SVM	1000	68	120
Softmax	1000	68	120
BPNN	1000	68	120

TABLE V. FULL DATASET GROUPING

Method	training set	validation set	testing set
Svm	90000	6920	10000
Softmax	90000	6920	10000
BPNN	90000	6920	10000

A. Small Dataset Results

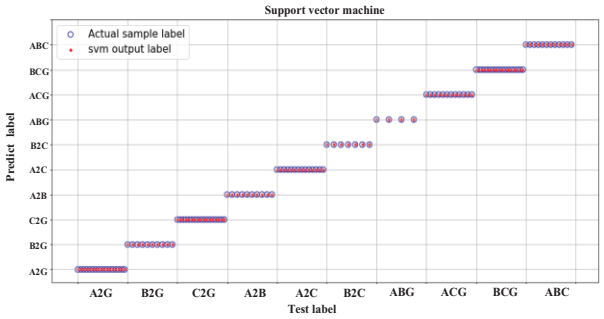

Fig. 4. The result of SVM method using small dataset

With the small dataset, Fig. 4 to Fig. 6 demonstrate the classification results with SVM, Softmax and BPNN, respectively. One can observe that the results are with very high fault classification accuracies for all three data driven approaches. The fault classification accuracies for each fault type and each data driven approach are further summarized in Table VI. The average classification accuracies are 100%, 98.3% and 100% for SVM, Softmax and BPNN, respectively.

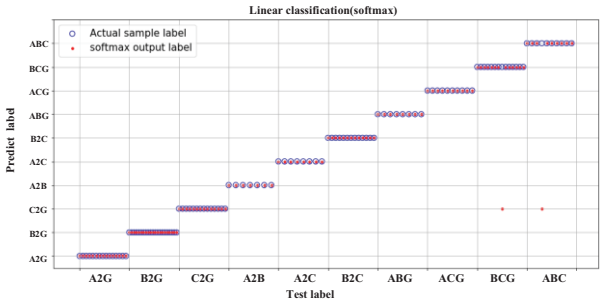

Fig. 5. The result of softmax method using small dataset

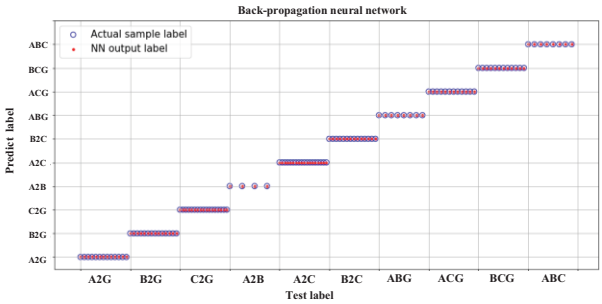

Fig. 6. The result of back-propagation neural network method using small dataset

TABLE VI. RESULTS WITH SMALL DATASET FOR TRAINING AND TESTING

Fault types	Accuracy		
	SVM	Softmax	BPNN
A2G	1.000	1.000	1.000
B2G	1.000	1.000	1.000
C2G	1.000	1.000	1.000
A2B	1.000	1.000	1.000
A2C	1.000	1.000	1.000
B2C	1.000	1.000	1.000
ABG	1.000	1.000	1.000
ACG	1.000	1.000	1.000
BCG	1.000	0.929	1.000
ABC	1.000	0.900	1.000
Total	1.000	0.983	1.000

To further validate whether the trained model is applicable in practice, the trained model is examined via the additional testing set. The results are summarized in Table VII to IX. One can observe the following facts. a) The greater the difference between the training data and the additional testing data, the lower the accuracy is. For example, the training set of the small dataset only covers 10° phase angle difference; the accuracy of BPNN is 91.7% and 53.9%, with 15° and 60° phase angle difference, respectively. b) The accuracies of BPNN are slightly higher than the other two classifies. For example, in the case where the phase angle difference is 60°, the accuracy of SVM, Softmax and BPNN are 43.9%, 52.2%, 53.9%, respectively. In the case where the phase angle difference is 15°, the accuracy of SVM, Softmax and BPNN are 82.2%, 81.6%, 91.7%, respectively. The above results verify that the BPNN has a stronger ability to extract features.

Nevertheless, one can observe that the testing results with the additional testing set are significantly degraded. This is because the small dataset is incomplete and cannot fully represent the characteristics of the problem. In fact, this proves that the trained model has overfitting issue and is lack

of generalization, and is not applicable to practical systems.

TABLE VII. RESULTS OF SVM WITH ADDITIONAL TESTING SET, SMALL DATASET FOR TRAINING

Parameters	Phase Angle Difference (°)					
	15	25	35	45	55	60
Event Number	180	180	180	180	180	180
Wrong Number	32	41	59	68	95	101
Accuracy	0.822	0.722	0.672	0.622	0.472	0.439
Parameters	**Fault Location (km)**					
	25	65	105	115	145	175
Event Number	108	108	108	108	108	108
Wrong Number	17	12	8	8	14	11
Accuracy	0.843	0.889	0.926	0.926	0.870	0.898
Parameters	**Fault Resistance (Ω)**					
	15	30	50	200	300	400
Number	40	40	40	12	12	12
Wrong	8	10	12	4	4	4
Accuracy	0.800	0.725	0.700	0.666	0.666	0.666

TABLE VIII. RESULTS OF SOFTMAX WITH ADDITIONAL TESTING SET, SMALL DATASET FOR TRAINING

Parameters	Phase Angle Difference (°)					
	15	25	35	45	55	60
Event Number	180	180	180	180	180	180
Wrong Number	33	41	67	75	80	86
Accuracy	0.816	0.722	0.628	0.583	0.556	0.522
Parameters	**Fault Location (km)**					
	25	65	105	115	145	175
Event Number	108	108	108	108	108	108
Wrong Number	13	8	10	13	10	12
Accuracy	0.833	0.926	0.907	0.880	0.907	0.888
Parameters	**Fault Resistance (Ω)**					
	15	30	50	200	300	400
Number	40	40	40	12	12	12
Wrong	10	10	13	4	4	4
Accuracy	0.750	0.750	0.675	0.666	0.666	0.666

TABLE IX. RESULTS OF BPNN WITH ADDITIONAL TESTING SET, SMALL DATASET FOR TRAINING

Parameters	Phase Angle Difference (°)					
	15	25	35	45	55	60
Event Number	180	180	180	180	180	180
Wrong Number	15	32	43	57	79	83
Accuracy	0.917	0.822	0.761	0.683	0.561	0.539
Parameters	**Fault Location (km)**					
	25	65	105	115	145	175
Event Number	108	108	108	108	108	108
Wrong Number	10	8	8	8	8	8
Accuracy	0.907	0.926	0.926	0.926	0.926	0.926
Parameters	**Fault Resistance (Ω)**					
	15	30	50	200	300	400
Number	40	40	40	12	12	12
Wrong	4	8	8	2	2	2
Accuracy	0.900	0.800	0.800	0.833	0.833	0.833

B. *Full Dataset Results*

With the full dataset, Fig. 7 to Fig. 9 demonstrate the classification results with SVM, Softmax and BPNN, respectively. The fault classification accuracies for each fault type and each data driven approach are further summarized in Table X. The average classification accuracies are 19.5%, 20.5% and 98.6% for SVM, Softmax and BPNN, respectively. One can observe that with the full dataset, the results of SVM and Softmax are with very low accuracies while the results of BPNN are still with high accuracies. This further proves that BPNN has a stronger ability to extract features.

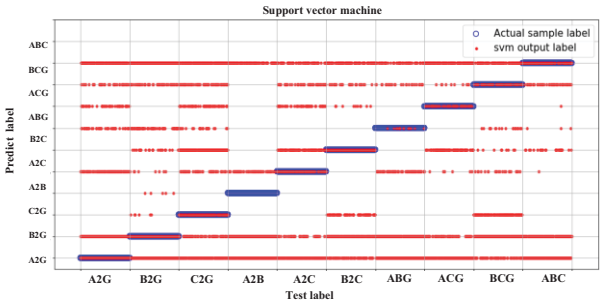

Fig. 7. The result of SVM method using full dataset

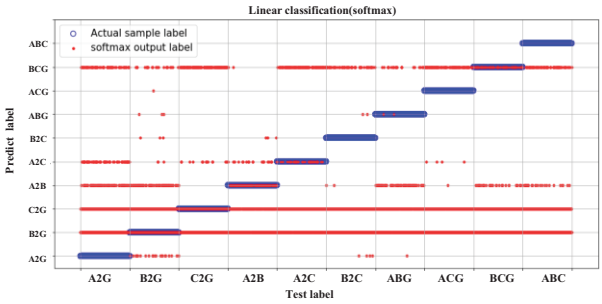

Fig. 8. The result of linear classifier method using full dataset

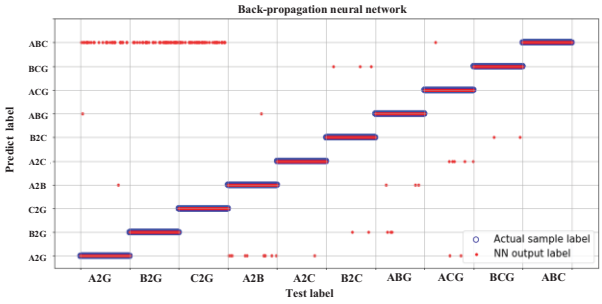

Fig. 9. The result of back-propagation neural network method using full dataset

TABLE X. RESULTS WITH FULL DATASET FOR TRAINING AND TESTING

Fault types	Accuracy		
	SVM	Softmax	BPNN
A2G	0.114	0.304	0.978
B2G	0.157	0.091	0.974
C2G	0.024	0.155	0.983
A2B	0.359	0.058	1.000
A2C	0.000	0.269	0.986
B2C	0.375	0.000	0.998
ABG	0.350	0.242	0.995
ACG	0.286	0.344	0.993
BCG	0.326	0.336	0.988
ABC	0.147	0.292	1.000
Total	0.195	0.205	0.986

To further validate whether the trained model is applicable in practice, the trained model is examined via the additional testing set. The results are summarized in Table XI to XIII. One can observe that with the additional testing set, the results of SVM and Softmax are still with low accuracies. On the other hand, BPNN performs very well even with the additional testing set, which proves the importance of completeness of the dataset: the trained model with the full dataset successfully captures the actual features embedded inside the instantaneous values of the three phase currents. The trained model does not have overfitting issues and is with good generalization properties, and therefore is applicable in practice to accurately classify the transmission line faults.

TABLE XI. RESULTS OF SVM WITH ADDITIONAL TESTING SET, FULL DATASET FOR TRAINING

Parameters	Phase Angle Difference (°)					
	15	25	35	45	55	60
Event Number	180	180	180	180	180	180
Wrong Number	126	129	135	137	136	141
Accuracy	0.300	0.283	0.250	0.239	0.244	0.217
Parameters	**Fault Location (km)**					
	25	65	105	115	145	175
Event Number	108	108	108	108	108	108
Wrong Number	71	71	71	71	71	74
Accuracy	0.333	0.333	0.333	0.333	0.333	0.315
Parameters	**Fault Resistance (Ω)**					
	15	30	50	200	300	400
Number	40	40	40	12	12	12
Wrong	28	30	32	8	8	8
Accuracy	0.300	0.250	0.200	0.333	0.333	0.333

TABLE XII. RESULTS OF SOFTMAX WITH ADDITIONAL TESTING SET, FULL DATASET FOR TRAINING

Parameters	Phase Angle Difference (°)					
	15	25	35	45	55	60
Event Number	180	180	180	180	180	180
Wrong Number	158	151	147	147	144	143
Accuracy	0.122	0.161	0.183	0.183	0.200	0.206
Parameters	**Fault Location (km)**					
	25	65	105	115	145	175
Event Number	108	108	108	108	108	108
Wrong Number	93	94	94	94	94	92
Accuracy	0.139	0.130	0.130	0.130	0.130	0.148
Parameters	**Fault Resistance (Ω)**					
	15	30	50	200	300	400
Number	40	40	40	12	12	12
Wrong	34	32	32	12	12	12
Accuracy	0.150	0.200	0.200	0.000	0.000	0.000

TABLE XIII. RESULTS OF BPNN WITH ADDITIONAL TESTING SET, FULL DATASET FOR TRAINING

Parameters	Phase Angle Difference (°)					
	15	25	35	45	55	60
Event Number	180	180	180	180	180	180
Wrong Number	0	0	0	0	0	0
Accuracy	1.000	1.000	1.000	1.000	1.000	1.000
Parameters	**Fault Location (km)**					
	25	65	105	115	145	175
Event Number	108	108	108	108	108	108
Wrong Number	0	0	1	1	0	0
Accuracy	1.000	1.000	0.990	0.990	1.000	1.000
Parameters	**Fault Resistance (Ω)**					
	15	30	50	200	300	400
Number	40	40	40	12	12	12
Wrong	0	0	1	0	0	0
Accuracy	1.000	1.000	0.975	1.000	1.000	1.000

V. DISCUSSION

From the aforementioned numerical experiments, one can observe that a complete dataset for training is required to ensure the practicability of the fault classifier. In practical power systems, since faults do not usually occur, the field data may not be adequate for a complete dataset. In this case, power system fault simulation software could help generate a complete dataset. Nevertheless, the effectiveness of this generated dataset is based on accurate simulation models of the transmission line as well as the rest of the power system. Therefore, the gap between the simulation models and the practical power systems could potentially generate fault classification errors. In addition, advanced structures of the neural networks could potentially improve the fault classification accuracy as well as the robustness of the algorithm. These issues will be studied in future publications.

VI. CONCLUSION

This paper proposes a systematic framework to demonstrate the influence of the dataset completeness on the validity of the data driven approaches. Specifically, transmission line fault classification problem with SVM, Softmax and BPNN data driven approaches is utilized as an example. First, a small (incomplete) dataset and a full (compete) dataset are constructed respectively. Second, each dataset is randomly separated into training, validation and testing sets, and the trained model is generated for each dataset. An additional testing set, which differs from both the small and the full dataset, is established, to further validate the effectiveness of the trained model. Numerical experiments prove that the performance on the additional testing set is much degraded if the model is trained using the small dataset. On the other hand, the trained model using full dataset performs well with the additional testing set, demonstrating no overfitting issues and good generalization properties.

REFERENCES

[1] A. P. S. Meliopoulos et al., "Dynamic State Estimation-Based Protection: Status and Promise," *IEEE Trans Power Del.*, vol. 32, no. 1, pp. 320-330, Feb. 2017.

[2] Y. Liu, A. P. S. Meliopoulos, Z. Tan, et al., "Dynamic state estimation-based fault locating on transmission lines," *IET Gener. Transm. Distrib.*, vol 11, no. 17, pp. 4184-4192, Nov. 2017.

[3] J. A. Jiang, C. S. Chen, and C. W. Liu, "A new protection scheme for fault detection, direction, discrimination, classification, and location in transmission lines," *IEEE Trans. Power Del.*, vol. 18, no. 1, pp. 34–42, Jan. 2003.

[4] M. J. Reddy, D. K. Mohanta, "Adaptive-neuro-fuzzy inference system approach for transmission line fault classification and location incorporating effects of power swings,"*IET Gener. Transm. Distrib.*, vol. 2, no. 2, pp. 235-244, March 2008.

[5] O. A. S. Youssef, "Combined fuzzy-logic wavelet-based fault classification technique for power system relaying," *IEEE Trans. Power Del.*, vol. 19, no. 2, pp. 582-589, April 2004.

[6] D. Thukaram, H. P. Khincha, H. P. Vijaynarasimha, "Artificial neural network and support vector Machine approach for locating faults in radial distribution systems," *IEEE Trans. Power Del.*, vol. 20, no. 2, pp. 710-721, April 2005.

[7] S. R. Samantaray, "Decision tree-based fault zone identification and fault classification in flexible AC transmissions-based transmission line," *IET Gener. Transm. Distrib.*,vol. 3, no. 5, pp. 425-436, May 2009.

[8] A. Jamehbozorg, S. M. Shahrtash, "A Decision Tree-Based Method for Fault Classification in Double-Circuit Transmission Lines," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2184-2189, Oct. 2010.

[9] P. K. Dash, S. Mishra, M. A. Salama, A. C. Liew, "Classification of power system disturbances using a fuzzy expert system and a Fourier

linear combiner," *IEEE Trans. Power Del.*, vol. 15, no. 2, pp. 472-477, April 2000.

[10] J. Xie, A. P. S. Meliopoulos and B. Xie, "Transmission Line Fault Classification Based on Dynamic State Estimation and Support Vector Machine," in *North American Power Symposium (NAPS)*, Fargo, ND, 2018, pp. 1-5.

[11] U. B. Parikh, B. Das, R. P. Maheshwari, "Combined Wavelet-SVM Technique for Fault Zone Detection in a Series Compensated Transmission Line," *IEEE Trans. Power Del.*, vol. 23, no. 4, pp. 1789-1794, Oct. 2008.

[12] N. Zhang, M. Kezunovic, "Transmission Line Boundary Protection Using Wavelet Transform and Neural Network," *IEEE Trans. Power Del.*,vol. 22, no. 2, pp. 859-869, April 2007.