Selecting Appropriate Machine Learning Classifiers for DGA Diagnosis

Jose Ignacio Aizpurua[1*], Victoria M. Catterson[1], Brian G. Stewart[1], Stephen D. J. McArthur[1], Brandon Lambert[2], Bismark Ampofo[2], Gavin Pereira[3], and James G. Cross[3]

[1]Institute for Energy and Environment, University of Strathclyde, Glasgow, UK

[2]Bruce Power, Kincardine, Canada

[3]Kinectrics Inc., Toronto, Canada

*Email: jose.aizpurua@strath.ac.uk

***Abstract*- Dissolved gas analysis (DGA) is a common method of assessing transformer health. There are a number of machine learning classifiers reported to give a high accuracy on specific datasets, such as Artificial Neural Networks or Support Vector Machines. When these methods reach the same conclusion about the type of fault present, this can give an increased confidence in the veracity of the diagnosis. However, it is critical to analyze and quantify the strength of these classifiers in the presence of conflicting data to test their practicality for usage in the field. This paper investigates the adequacy of different machine learning based DGA diagnosis models in the presence of conflicting data. The proposed method will aid engineers with the selection of machine learning models so as to maximize the usability and accuracy in the presence of conflicting data.**

I. INTRODUCTION

Transformers are crucial assets for power grid operation. Transformers may fail in service if monitoring models do not identify degraded conditions in time. Dissolved gas analysis (DGA) is a method that examines the dissolved gasses in transformer oil to diagnose the transformer state.

There exist a number of deterministic methods for transformer fault diagnosis based on DGA, e.g. Roger's ratios, Doernenburg's ratios, or Duval's triangle [1]. These methods classify ratios of fault gasses into predefined intervals. However, their accuracy is limited because they assume crisp gas ratio decision bounds and they assign a diagnosis with full confidence regardless of proximity to a diagnostic boundary. This may lead to conflicting situations, i.e. different methods with different diagnosis outcomes for the same input gas values. Additionally the fault types diagnosed by each method are different. In this work all diagnoses are classified into four groups: *Thermal*, *Arcing* (including high energy discharges), *partial discharge* (PD), and *Normal* degradation.

Machine learning (ML) classifiers are statistical models that overcome the limitation of crisp gas decision bounds by learning probabilistic bounds between different fault classes. IEC TC 10 is a benchmarking DGA dataset [2] and a number of ML classifiers have been tested on this dataset. Mirowski and LeCun used k-nearest neighbor (kNN), support vector machine (SVM) and artificial neural network (ANN) models for a binary classification problem obtaining 91%, 90% and 89% mean accuracy values, respectively [3]. Wang *et al.* used deep learning methods through a continuous sparse autoencoder using 125 training and 9 testing samples obtaining a mean accuracy of 93.6% [4]. The combined use of optimization and classification models have also been explored through gene programming and SVM, ANN and kNN classifiers with best mean accuracy of 92% [5] or genetic algorithms and SVM models with 84% mean accuracy [6].

The use of stochastic optimization methods along with machine learning models can increase the accuracy of the diagnosis model by selecting gas samples that minimize the error, or resampling the data space to balance the data from each fault class. Resampling methods generate data samples by analyzing the statistical properties of the inspection data. However, this process may impact the adoption of these methods in the industry because with the extra data generation process there is a risk of losing information when undersampling and overfitting when oversampling.

The number of training and testing samples directly influences the classification accuracy. The more samples that are used for training (and the less for testing), the greater will be the classifier accuracy, e.g. [4]. However, the generalization of the diagnostics model is penalized when the testing set is much smaller than the training set. This work adopts the 80% training 20% testing approach as in [3],[5],[6].

These classifiers show a high accuracy, but their usability for resolving misclassified data samples is limited because their diagnosis is deterministic and they do not generate uncertainty information, i.e. they are black-box (BB) models.

Information fusion and evidence combination methods enable the combination of different information sources to improve the diagnosis accuracy. Tang *et al.* explored the use of Dempster Shafer's (DS) theory and the analytic hierarchy process for transformer condition assessment based on expert knowledge and diagnostics results [7]. Catterson and McArthur examined the accuracy of different fusion methods tested on transformer PD data including weighted majority voting, weighted average, DS, and Bayesian networks for the combination of ANN, k-means and C5.0 (a classifier toolset) [8]. Bhalla *et al.* combined ANNs with Fuzzy logic through DS theory [9].

The performance evaluation of these models is based on accuracy indicators tested on datasets comprised of conflicting and non-conflicting samples. However, the strength and validity of a fusion method is assessed in the presence of conflicting data, i.e. when different classifiers diagnose

different faults for the same input gas samples. In these situations it is not clear which model is most accurate. BB models suffer from lack of explicability and uncertainty management, whereas white-box methods generate more useful diagnostic information for decision-making under uncertainty due to their transparency and uncertainty information, but generally their accuracy is lower.

The main contribution of this paper is the analysis of different fusion methods for transformer DGA analysis and their quantification for decision-making under uncertainty.

The rest of this paper is organized as follows. Section II presents the proposed framework for the analysis of ML classifiers. Section III presents results obtained on the IEC TC 10 database and finally, Section IV presents conclusions and future goals.

II. A FRAMEWORK FOR PERFORMANCE ANALYSIS OF MACHINE LEARNING CLASSIFIERS

Fig. 1 shows the adopted framework for the assessment of ML classifiers based on Gaussian Bayesian networks (GBN), SVM, and ANN source classifiers. The framework is divided into cross-validation, data pre-processing, model training and testing, and evidence combination stages.

The goal of the cross-validation is to validate the results and assess how they will generalize to an independent dataset. To this end, Monte Carlo cross-validations are implemented as follows [10]: (i) initialize the trial counter, *trials=0*; (ii) randomly shuffle the dataset and execute preprocessing, training, testing, and evidence combination steps, and store the results separately for all samples and only conflicting samples; (iii) if *trials*<*Max_trials* iterate from the previous step and increase the trial counter by 1; (iv) otherwise extract mean and standard deviation values of the stored diagnosis results. For each trial, the random shuffle and the train/test steps generate different training and testing datasets, and therefore, this process evaluates the framework with *Max_trials* different training and testing datasets (*Max_trials*=10^3). As a result this process generates repeatable and consistent diagnosis results.

The data preprocessing stage starts by applying a *log-scale* step because diagnostic information resides in the order of magnitude [3]. Firstly the logarithm of every gas sample in the dataset is taken and then each variable in the dataset is scaled to mean zero and standard deviation one. This is done for each gas within the dataset, by subtracting the mean value and dividing by the standard deviation, for each sample of the variable. Then a *feature selection* step is applied to select relevant input data and maximize the diagnosis accuracy. This can be done through engineering knowledge or using stochastic optimization methods. In this case engineering knowledge is used. For all the models five key fault gasses are used: ethane (C_2H_6), ethylene (C_2H_4), hydrogen (H_2), methane (CH_4), acetylene (C_2H_2). Subsequently, the dataset is divided into train and test datasets using 80% and 20% of the randomly shuffled dataset, respectively.

The model *training & testing* stage is specific to each source classifier introduced in the next subsection. Subsequently, Subsections II.B and II.C introduce the *evidence combination* and accuracy quantification steps.

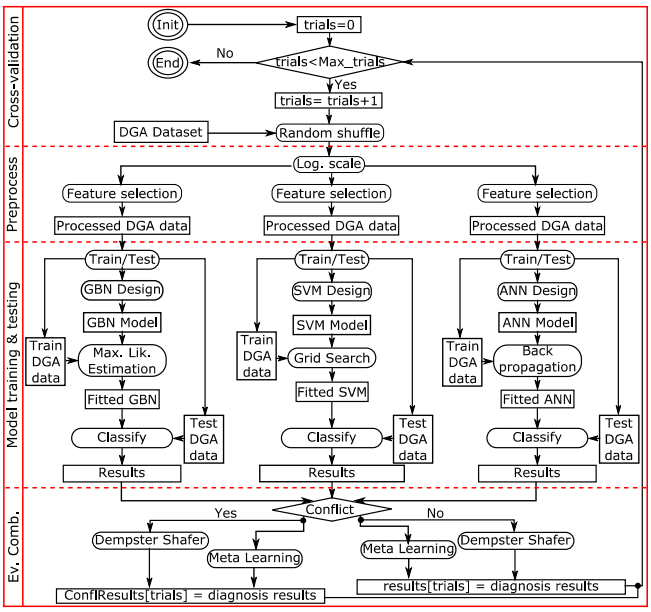

Fig. 1. Performance analysis framework.

A. Source Classifiers

1) Gaussian Bayesian Networks (GBN): Bayesian networks (BN) [11] are statistical models that capture probabilistic dependencies among random variables. Graphically, these variables are represented through nodes and they are linked through edges to reflect dependencies between variables. Statistically, dependencies are quantified through conditional probabilities. BNs are a compact representation of joint probability distributions. In probability theory, the chain rule permits the calculation of any member of the joint distribution of a set of random variables using conditional probabilities.

When a BN is comprised of continuous random variables GBNs capture dependencies through linear Gaussian distributions and variable distributions are modelled through Normal random variables. Local distributions are linked through linear models in which the parents play the role of explanatory variables. Each node x_i is regressed over its parent nodes. Assuming that the parents of x_i are $\{u_1,\ldots, u_k\}$, then the conditional probability of each node can be expressed as $p(x_i \mid u_1, \ldots,u_k) \sim N(\beta_0 + \beta_1 u_1 + \ldots + \beta_k u_k; \sigma^2)$, where β_0 is the intercept and $\{\beta_1, \ldots, \beta_k\}$ are linear regression coefficients for the parent nodes $\{u_1,\ldots, u_k\}$. Fig. 2 shows the GBN model.

The parameter estimation for GBN models is based on the maximum likelihood algorithm which estimates the corresponding parameters for each node in the BN model, e.g. for the Arc node (Fig. 2): $Pr(Arc \mid C_2H_6, C_2H_2, CH_4, C_2H_4, H_2) \sim N(\beta_0 + \beta_1 C_2H_6 + \beta_2 C_2H_2 + \beta_3 CH_4 + \beta_4 C_2H_4 + \beta_5 H_2; \sigma^2)$.

After learning the parameters the estimation of the conditional probability of nodes is based on inferences. In this case the likelihood weighting algorithm is implemented, which fixes the test DGA gas samples (evidence) and uses the likelihood of the evidence to weight samples [11]. When applied to the DGA dataset, for each of the analyzed faults the

outcome of the inference is a set of random samples from the conditional distribution of the fault node given the evidence.

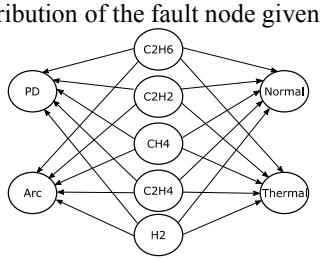

Fig. 2. GBN model.

Density values of the inference outcomes can be calculated through Kernel density estimates [14]. The GBN model was implemented using the bnlearn R package.

2) Support Vector Machines (SVMs): SVMs map input data into a space using a kernel function [12]. The SVM learns the boundary separating one class from another with maximum distance. The kernel function aims to translate a problem that is nonlinearly separable into a feature space, which is linearly separable by a hyperplane. The hyperplane represents the classification boundary. The SVM is parametrized through the choice of kernel function. For a problem which may be nonlinear, the RBF kernel is recommended.

The SVM solves an optimization problem maximizing the distance from the hyperplane to the nearest training point. Generally, the dataset is not linearly separable and slack variables are used to correct incorrectly classified samples. Cost variables are used to penalize the objective function, which is a tradeoff between penalizing slack variables and obtaining a large margin for the SVM.

The SVM training consists of calculating the cost and kernel parameters. Model training was performed using the R e1071 package. The RBF was chosen as the kernel, and grid search was used to optimize the parameters. All available gas data was used as input to the SVM. Of the trained SVMs, the one with the highest accuracy from the test data was selected as the choice for that output.

3) Artificial Neural Networks (ANN): ANNs are popular black-box models used for classification and regression [13]. The multilayer perceptron (MLP) feedforward ANN model was used in this work. The MLP model is a three-layer network (input, hidden and output layer) comprised of fully connected neurons. Each neuron performs a weighted sum of its inputs and passes the results through an activation function. A sigmoid activation function is used for hidden and output neurons.

Model training is performed using the back-propagation algorithm. The goal is to learn the neuron weights so as to generate the network output from the sample input, which minimizes the error with respect to the target output. Input and hidden layers may also have a bias unit analogous to intercept terms in a regression model. A number of networks were trained, using all the gases at the input layer and varying by the number of hidden nodes. Of the trained networks, the one with the highest accuracy was selected with 20 hidden nodes. Model training was performed using the R nnet library.

B. Evidence Combination Methods

1) Dempster Shafer's (DS) Theory: DS builds beliefs of the true state of a process from distinct pieces of evidence, e.g. see [7]-[9]. Assuming a set of faults $F=\{f_1,\ldots,f_i,\ldots,f_{|F|}\}$ the set of possible states is called the frame of discernment, F. Pieces of evidence are formulated as mass functions, m: $2^F \rightarrow$ satisfying: $m(f_i)\geq 0$, $m(\varnothing)=0$, and $\sum_{f_i \subseteq F} m(f_i) = 1$.

The combined probability mass for the i-th fault, f_i, of two classifiers, denoted c_1 and c_2, is defined as

$$m_{C1C2}(f_i) = \frac{1}{1-K} \sum_{\substack{A,B \subseteq F \\ A \cap B = f_i}} m_{c1}(A) m_{c2}(B) \quad (1)$$

$\forall\, f_i \subseteq F$, $f_i = \varnothing$, where K is the degree of conflict between two mass functions:

$$K = \sum_{\substack{A,B \subseteq F \\ A \cap B = \varnothing}} m_{c1}(A) m_{c2}(B) \quad (2)$$

2) Meta-learning methods: a meta-learning method learns which classifiers are reliable and which are not. The stacking method is based on this concept [15]. Instead of taking the original input variables, a stacked model takes as input the probabilistic outcomes generated from the source classifiers. These models are trained first and then tested with both training and testing data. The training and testing of the stacked model is based on the training and testing outcomes of the independent classifiers. Fig. 3 shows the stacking concept.

As opposed to DS theory, in the stacking configuration a learning model is trained. ANN and SVM models have used as a stacking model to aggregate independent classifiers.

C. Accuracy Metrics

The raw accuracy figure takes into account all testing data samples and evaluates which are the samples that match the true fault category. Additionally the proposed framework evaluates the accuracy taking into account only conflicting samples in which the outcome of at least one out of three classifiers is different. This accuracy value will highlight the effectiveness of different methods for resolving conflicting data samples.

III. RESULTS

Table I displays the classification results for the analyzed source classifiers tested on the IEC TC 10 database and Table II displays the accuracy of fusion methods assessing their performance on a broad database of samples (raw accuracy) and on difficult cases (conflicting sample accuracy).

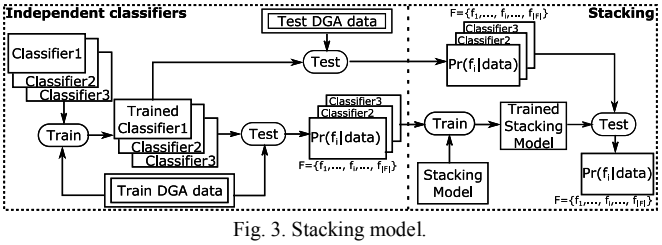

Fig. 3. Stacking model.

TABLE I
SOURCE CLASSIFIER RESULTS.

Class.	Overall	Thermal	PD	Arc	Normal
GBN	81.9% ± 6.5%	67.9% ± 17.7%	**83.6% ± 35.5%**	93.7% ± 6.7%	72.9% ± 15.1%
SVM	86.3% ± 5.9%	71.2% ± 17.9%	76.5% ± 37.8%	92.9% ± 7.1%	87.2% ± 11.8%
ANN	**89.4% ± 5.1%**	**78.1% ± 17.4%**	76.5% ± 37.9%	**95.2% ± 5.7%**	**89.2% ± 10.2%**

TABLE II
FUSION STRATEGY RESULTS.

Fusion	Raw Acc.	Confl. sample Acc.
Stacking ANN	89.7% + 5.2%	73.5%+19.4%
Stacking SVM	89.7%+5.3%	74.1% + 19.7%
Dempster Shafer	**90.1%+5.3%**	**75.4%+19.3%**

The conflicting case accuracy is lower than the raw accuracy because all the data samples diagnosed consistently have been removed and this affects the effectiveness of the methods. Additionally, the difference among fusion methods is greater for conflicting samples and this indicates the strength of the studied methods in the presence of conflicting data. For instance, the stacking SVM model performs better with conflicting data samples than the stacking ANN model.

The obtained results identify the performance of some fusion methods for DGA diagnosis. This performance also depends on the performance of the source classifiers and their capability to handle uncertainty information. For example, Fig. 4 shows the outcome of ANN, SVM and GBN models for the following gas values: C_2H_6 = 35 ppm, C_2H_2 = 434 ppm, CH_4 = 530 ppm, C_2H_4 = 383 ppm, H_2 = 1900 ppm. The observed fault according to IEC TC 10 is an Arc fault.

According to the ANN and SVM classifiers this is a Normal fault with 0.84 and 0.8 occurrence probability respectively. The x-axis of the GBN model shows the fault occurrence probabilities and the peak density indicates the maximum likelihood value. Arc and Normal faults have high likelihood, but the density of the Arc fault is slightly narrower than the Normal fault, which means that the GBN model is slightly more confident in the classification of the Arc fault.

There is a conflict among the output of these classifiers. The high deterministic probability values of SVM and ANN models encourage the engineer to assume that it is a normally degrading transformer. However, the GBN output shows that there is a conflict between Normal and Arc faults. This example shows how GBN models are able to generate uncertainty information which can aid the engineer in decision-making.

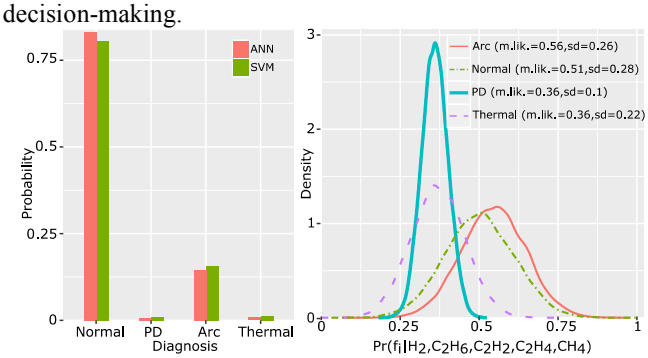

Fig. 4. Diagnosis output: ANN, SVM (left), and GBN (right) models.

Note that for the fusion strategy results displayed in Table II GBN outputs are estimated from normalized maximum likelihood values.

IV. CONCLUSIONS

This paper has examined the performance of machine learning classifiers and fusion methods focusing on conflicting data samples. The obtained results identify the strengths and drawbacks of some of the key machine learning methods for DGA diagnosis. Suitable metrics for assessing the performance of the techniques have been discussed, with the intention of highlighting the performance on a broad database of samples and on particularly difficult cases.

The results obtained in this paper can be used as a benchmark to other techniques because the IEC TC 10 dataset is publicly available. Under the presented conditions the best fusion method obtains 90.1%±5.3% raw accuracy and 75.4%±19.3% accuracy for conflicting data samples.

Future work will address the implementation of a sound evidence combination framework which is able to include uncertainty information and resolve more conflicting cases.

REFERENCES

[1] IEEE Power and Energy Society, "IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers," IEEE Std C57.104-2008, pp. 1-36, 2009.

[2] M. Duval and A. de Pablo, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," IEEE Electrical Insul. Mag., vol. 17, pp. 31-41, 2001.

[3] P. Mirowski and Y. LeCun, "Statistical Machine Learning and Dissolved Gas Analysis: A Review," IEEE Trans. Pow. Del., vol. 27, pp. 1791-1799, 2012.

[4] L. Wang, X. Zhao, J. Pei, and G. Tang, "Transformer fault diagnosis using continuous sparse autoencoder," SpringerPlus, vol. 5, p. 448, 2016.

[5] A. Shintemirov, W. Tang, and Q. H. Wu, "Power Transformer Fault Classification Based on Dissolved Gas Analysis by Implementing Bootstrap and Genetic Programming," IEEE Trans. Systems, Man, and Cybern. C, vol. 39, pp. 69-79, 2009.

[6] J. Li, Q. Zhang, K. Wang, J. Wang, T. Zhou, and Y. Zhang, "Optimal dissolved gas ratios selected by genetic algorithm for power transformer fault diagnosis based on support vector machine," IEEE Trans. Dielectr. Electr. Insul., vol. 23, pp. 1198-1206, 2016.

[7] W. H. Tang, K. Spurgeon, Q. H. Wu, and Z. J. Richardson, "An evidential reasoning approach to transformer condition assessments," IEEE Trans. Pow. Del., vol. 19, pp. 1696-1703, 2004.

[8] V. M. Catterson and S. D. J. McArthur, "Using evidence combination for transformer defect diagnosis," Int. J. of Innovations in Energy Systems and Power, vol. 1, 2006.

[9] D. Bhalla, R. K. Bansal, and H. O. Gupta, "Integrating AI based DGA fault diagnosis using Dempster–Shafer Theory," Electrical Power & Energy Systems, vol. 48, pp. 31-38, 6, 2013.

[10] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," Chemometrics and Intelligent Laboratory Systems, vol. 56, pp. 1-11, 4/16/ 2001.

[11] R. E. Neapolitan, Learning Bayesian networks: Prentice Hall, 2004.

[12] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Min. Knowl. Discov., vol. 2, pp. 121-167, 1998.

[13] M. Anthony and P. L. Bartlett, Neural network learning: Theoretical foundations: Cambridge university press, 2009.

[14] J. Kim and C. Scott, "Robust kernel density estimation," in 2008 IEEE Int. Conf. Acoustics, Speech and Signal Process., 2008, pp. 3381–3384.

[15] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms: Chapman & Hall/CRC, 2012.