Uncertainty-aware Fusion of Probabilistic Classifiers for Improved Transformer Diagnostics

Jose Ignacio Aizpurua, *Member, IEEE*, Victoria M. Catterson, *Senior Member, IEEE*, Brian G. Stewart, *Member, IEEE*, Stephen D. J. McArthur, *Fellow, IEEE*, Brandon Lambert, and James G. Cross *Senior Member, IEEE*

***Abstract*—Transformers are critical assets for the reliable operation of the power grid. Transformers may fail in service if monitoring models do not identify degraded conditions in time. Dissolved gas analysis (DGA) focuses on the examination of dissolved gasses in transformer oil to diagnose the state of a transformer. Fusion of black-box classifiers, also known as an ensemble of diagnostics models, have been used to improve the accuracy of diagnostics models across many fields. When independent classifiers diagnose the same fault, this method can increase the veracity of the diagnostics. However, if these methods give conflicting results, it is not always clear which model is most accurate due to their black-box nature. In this context, the use of white-box models can help resolve conflicted samples effectively by incorporating uncertainty information and improve the classification accuracy. This paper presents an uncertainty-aware fusion method to combine black-box and white-box diagnostics methods. The effectiveness of the proposed approach is validated using two publicly available DGA datasets.**

***Index Terms*—Condition monitoring, transformer diagnosis, ensembles, classifiers, uncertainty.**

I. Introduction

TRANSFORMERS are critical assets in the power grid. The unexpected failure of a power transformer can lead to different consequences ranging from a lack of export capability to catastrophic failure [1]. Condition monitoring techniques examine the health of the system under study periodically with the aim to identify anomalies and avoid unexpected failures, e.g. [2], [3], [4]. The different components of a transformer can be monitored through different parameters [5]. This paper focuses on transformer insulation health assessment through dissolved gas analysis (DGA) [6]. The wide industrial acceptance and extended implementation of DGA monitors is the rationale and motivation to focus on DGA.

Operational and fault events generate gases which are dissolved in the oil that circulates through a transformer for cooling and insulation purposes. DGA is a mature and industry-standard method that focuses on the study of these gases [6]. The effective application of DGA enables timely diagnostics of possible insulation problems.

J. Aizpurua, V. Catterson, B. Stewart & S. McArthur are with the Inst. of Energy & Environment, Univ. of Strathclyde, Glasgow, UK (e-mail: jose.aizpurua@strath.ac.uk; vic@ieee.com; brian.stewart.100@strath.ac.uk; s.mcarthur@strath.ac.uk);

B. Lambert is with Bruce Power, Tiverton, Canada (e-mail: brandon.lambert@brucepower.com;

J. Cross is with Kinectrics Inc., Toronto, Canada (e-mail: james.cross@kinectrics.com).

There are different industry-accepted classical DGA methods including Duval's triangle, Roger's ratios or Doernenburg's ratios [6]. These techniques classify transformer faults based on the predefined range of specific fault gas ratios. However, their accuracy is limited because they assume crisp decision bounds [7]. This leads to a decreased diagnostics accuracy and conflicting diagnostics outcomes among methods which do not help engineers in the decision-making process. So as to improve the classification accuracy a number of black-box (BB) machine learning models have been proposed.

Comparisons among different DGA models are representative only when they are analysed in the same conditions. Focusing on the methods tested on the publicly available IEC TC 10 dataset [8], Mirowski and LeCun used k-nearest neighbor (kNN), support vector machine (SVM) and artificial neural network (ANN) models [9]. Wang *et al.* used deep learning methods through a continuous sparse autoencoder (CSA) [10]. The combined use of optimization and classification models has also been explored through gene programming (GP) and SVM, ANN and kNN models [11] or genetic algorithms (GA) and SVM models [12]. Table I reports the main characteristics.

TABLE I
MACHINE LEARNING MODELS TESTED ON THE IEC TC 10 DATASET.

Ref	*Machine learning model*	*Type*	*Train/Test*	*Mean accuracy*
[9]	kNN, SVM, ANN	Binary	134/33	91%, 90%, 89%
[10]	CSA	Multiclass	125/9	93.6%
[11]	GP + kNN	Multiclass	830/228	92%
[12]	GA + SVM	Multiclass	134/33	84%

The type of the classification problem has implications for decision-making purposes. Binary classifiers focus on identifying healthy or faulty samples, but they do not give more information about the type of fault present. Additionally, the number of training and testing samples directly influences the classification accuracy. The more samples that are used for training the greater will be the accuracy (e.g. [10]). However, the generalization of the diagnostics model is penalised when the testing set is much smaller than the training set.

There have been more DGA classification models tested on different proprietary datasets so as to improve the accuracy of classical methods such as fuzzy logic based DGA method [13], adaptive neuro fuzzy inference system (ANFIS) which combines ANN with fuzzy logic [14], SVM with resampling and boosting [15], differential evolution (DE) combined with extreme learning machines [16], or relevance vector machines combined with ANFIS [17].

Although the accuracy of these BB models tends to be high ($\sim$90%), there is no explainability of the results, i.e. they represent purely numerical connections and lack an interpretation of physical significance for an engineer. Additionally they do not integrate the uncertainty associated with the diagnostics outcome and assign either 100% belief to a single health state or a deterministic probability value. Therefore these techniques may be less desirable for engineering usage because there is no further information about the confidence in the result.

It is possible to specify subjective and imprecise information through fuzzy logic. However, fuzzy rules need to be specified manually based on experience and their diagnosis outcome is not a probability density function (PDF) which integrates uncertainty information. The work introduced in this paper focuses on data-driven Bayesian methods to determine decision bounds and deal with diagnostics uncertainties. Some fuzzy logic models have been designed to identify multiple fault conditions [13]. This work is focused on the identification of single fault conditions and multiple fault conditions will be considered as part of future work (see Section V).

Optimization methods along with BB models (GP in [11], GA in [12], DE in [16]) can increase the accuracy of the diagnosis model by selecting gas samples that minimize the error, or resampling the data space to generate more samples. Resampling methods generate synthetic data samples by analyzing the statistical properties of the inspection data (e.g, [11], [15]). However, this process may impact the adoption of these methods because with the extra data generation process there is a risk of losing information when undersampling and overfitting when oversampling [18]. For instance, it may have been the case that during the resampling process copies of the same data point may end up both in the training and testing set. So as to avoid any type of dependencies between the training and testing datasets this work only considers inspection data.

Ensembles of classifiers have been used to avoid the potential bias and risk of errors of individual classifiers and improve the diagnostics accuracy and prediction stability [19]. Ensemble models require post-processing the outcome of the source models so as to generate a consistent prediction. However, most of the transformer classification models have been focused on single classification algorithms and there are few works focused on ensembles, such as the fusion model in [20] which combines classical Roger, Duval, Doernenburg and IEC methods through a gating network, the hybrid approach in [21] which combines fuzzy logic with ANN through Dempster Shafer's (DS) theory, the multi-ANN approach in [22] which combines ANN models through majority voting, or the sequential combination of multiple gene expression programming models through an *if-else* process [23].

Ensemble strategies increase the veracity of the diagnostics when independent classifiers diagnose the same fault. Classical DGA and machine learning models can be combined through different methods (e.g. majority voting, weighted average, gating networks, DS). However, there is no way to further interpret the diagnostic outcome of these methods due to lack of uncertainty information associated with their outcome. Therefore, if these methods give conflicting results, it is not clear which model is most accurate, and in this situation, the engineer will not know which diagnostic conclusion to trust. Accordingly, due to the lack of uncertainty modelling of BB and classical DGA models, the application of ensembles in the field has been limited. This research addresses this and improves the selection of the correct diagnostic conclusion.

From an engineering viewpoint, the disagreements among independent classifiers are the most important situations that need to be resolved effectively because conflicting diagnoses may imply very different maintenance actions. Therefore, it is critical to analyse and quantify the strength of classifiers in the presence of conflicting data. Uncertainty quantification is very important for condition monitoring systems [24]. For instance, assume that a model has been trained to classify certain faults. So long as the test data is comprised of faults which are similar to the trained model, it should return a prediction with high confidence. However, if the model is tested on an unseen class of fault, the model should be able to quantify this with uncertainty levels, which can convey information about the confidence of the diagnosis of the model. This information is completely lost with BB models. Conversely, white-box (WB) models capture expert knowledge either as a causal model or through first-principle models. They generate the uncertainty associated with the decision-making process by quantifying the PDF of the likelihood of different diagnostics states. This function represents the strength of the model's diagnosis, i.e. the wider the variance, the lesser the confidence in the diagnostics outcome and vice-versa.

In this context, it is possible to combine WB and BB models to resolve conflicting samples effectively, assist the engineer in the decision-making process, and improve the diagnostics accuracy. To the best of the authors' knowledge this is the first approach which complements the accuracy of BB models with the uncertainty information of WB models for improved transformer diagnostics. Particularly for transformer DGA the use of ensembles has been limited. Therefore, the proposed approach aims to cover both gaps by proposing a novel ensemble classification framework and improving the accuracy of DGA diagnosis. The main contribution of this paper is thus the proposal of a novel probabilistic framework for uncertainty-aware fusion of classifiers to assist engineers in the decision-making process. The effectiveness of the framework is validated using publicly available datasets.

The rest of this paper is organised as follows. Section II introduces the datasets. Section III defines the proposed approach. Section IV presents results and finally, Section V draws conclusions.

II. Introduction to the DGA Datasets

The proposed approach is tested and validated using two real datasets. The IEC TC 10 is a standard benchmark dataset used to validate DGA methods [8]. It contains sets of seven different gases: ethane (C_2H_6), ethylene (C_2H_4), hydrogen (H_2), methane (CH_4), acetylene (C_2H_2), carbon monoxide (CO), and carbon dioxide (CO_2) sampled from different transformers, and labelled with their corresponding fault mode. Faults are classified into Normal degradation samples, Thermal faults ($T<700°C$ and $T>700°C$), Arc faults (low and high energy discharges), and partial discharge (PD) faults.

In order to generate this database, faulty equipment was removed from service, visually inspected by experienced engineers, and the fault clearly identified. The dataset also contains typical normal degradation values observed in several tens of thousands of transformers. In total, the dataset is comprised of 167 samples distributed as follows: 5.3% PD failure samples, 44.4% arcing failure samples, 20.4% thermal failure samples and 29.9% normal degradation samples.

In order to further validate the method another dataset is created comprised of C_2H_6, C_2H_4, H_2, CH_4 and C_2H_2 gas samples. This dataset is created by integrating datasets presented in [8], [25] [26] [27] and it is named Extended. In total it is comprised of 302 data samples: 3.3% PD, 40.4% arcing, 33.4% thermal and 22.9% normal degradation samples.

This work focuses on unbalanced classification problems without modifying the inspection data. So as to obtain statistically significant results Monte Carlo cross-validation (MCCV) also known as repeated random subsampling is used [28].

III. Uncertainty-Aware Ensemble Framework

The proposed framework focuses on the diagnostics of transformer faults through a supervised learning process using a dataset, DGA, comprised of n samples,

$$DGA = \{x_i, y_i\}_{i=1}^{n} \tag{1}$$

where the pair $\{x_i, y_i\}$ contains the data related to the *i*-th observation, $x_i \in X$, $y_i \in Y$. The matrix $X \in \mathbb{R}^{n \times p}$ contains the information $X = \{x_1, \ldots, x_n\}$ for p fault gases, and the vector $Y \in \mathbb{R}^{n \times 1}$ contains the information about the health state of the transformer. In a binary classification problem the set of possible states of y_i are limited to normal and fault states. However, in this case there are multiple states and the transformer state can be classified as: normal degradation, thermal fault, arc fault, and PD. Therefore, each output y_i can take the following values: $y_i = \{normal, thermal, arc, PD\}$. Multiclass classification problems are more challenging than binary classification problems, but they also generate more useful information for maintenance planning.

Fig. 1 shows the proposed generic classification framework. The ensemble classifier takes as input the deterministic probability values of each classifier ($m_{classifier_i}$) and the uncertainty parameters inferred from the WB model (m_u).

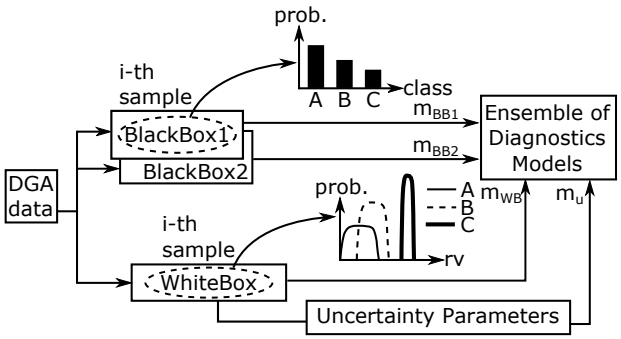


Fig. 1. Proposed uncertainty-aware ensemble diagnostics framework.

ANN and SVM models will be used as BB classification models as they have shown a high accuracy on DGA data (Table I). For WB modelling Gaussian Bayesian networks (GBNs) will be used because they are able to capture the causality among random variables and infer uncertainty information [29]. Algorithm 1 defines the implemented algorithm.

Transformer diagnostic information does not reside in absolute gas values (expressed in parts per million units, ppm) but instead in the order of magnitude. Therefore the dataset is log-normalized [9]. Firstly, the logarithm of every gas sample for all fault gases $x_i \in X$ is taken. Then each fault gas variable in the dataset is scaled to mean zero and standard deviation one. This is done for each fault gas $\{1, \ldots, p\}$ by subtracting the mean value and dividing by the standard deviation, for each sample of the fault gas variable $\{1, \ldots, n\}$ (cf. `line 2`).

MCCV is used for the quantification of the results [28]. For each trial i (cf. `line 3`), the log-normalized DGA data is randomly shuffled and then it is divided into 80% and 20% for training and testing (cf. `lines 4-5`). Then independent classifiers and ensemble models are trained and tested (cf. `lines 6-12`). The classification results of each trial i for each of the classifiers ($\overrightarrow{m}$) are evaluated with the accuracy metric (cf. `line 14`), and after repeating this process N times, the accuracy statistics are quantified (cf. `lines 18-19`) [28]:

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N} acc_i \quad sd_{\hat{a}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(acc_i - \hat{a})^2} \tag{2}$$

The repeated random subsampling process trains and tests *N* times all the models ensuring the generalization of the results.

Algorithm 1 Uncertainty-aware ensemble framework

1: i=1; m_acc=$\emptyset$; ▷ initialize variables
2: norm_data=*lognorm*(DGA); ▷ log-normalize DGA data
3: **while** $i < N$ **do**
4: rnd_dga=*shuffle*(norm_data); ▷ randomize data
5: [rnd_dga_{train}, rnd_dga_{test}]=*split_TrainTest*(rnd_dga);
6: m_{SVM}=SVM(rnd_dga_{train}, rnd_dga_{test});
7: m_{ANN}=ANN(rnd_dga_{train}, rnd_dga_{test});
8: PDF_{BN}=GBN(rnd_dga_{train}, rnd_dga_{test});
9: [m_{BN}, m_u]=Parameterization(PDF_{BN});
10: m_{DS}=DS(m_{BN}, m_{ANN}, m_{SVM});
11: m_{St}=Stacking(m_{BN}, m_{ANN}, m_{SVM});
12: m_{MDS}=MDS(m_{BN}, m_{ANN}, m_{SVM}, m_u);
13: $\overrightarrow{m} = \{m_{SVM}, m_{ANN}, m_{BN}, m_{DS}, m_{St}, m_{MDS}\}$
14: $\overrightarrow{acc} = accuracy(\overrightarrow{m})$ ▷ accuracy for all the models
15: $m\_acc[i,] = \overrightarrow{acc}$ ▷ save i-th trial accuracy results
16: $i = i + 1$ ▷ increase trial counter
17: **for** each $m_{classifier_j} \in \overrightarrow{m}$ **do** ▷ for each classifier
18: $\hat{a_j}$=mean(m_acc[, $m_{classifier_j}$])
19: $sd_{\hat{a_j}}$=sd(m_acc[, $m_{classifier_j}$])

Firstly Algorithm 1 trains and tests independent classifiers as follows (`lines 4-9`):

- `Line 4`: for each trial i, the log-normalized dataset $norm\_data$ is randomly shuffled.
- `Line 5`: the randomly shuffled dataset rnd_dga is divided into training and testing sets.

- `Lines 6-7`: SVM and ANN classifiers are trained by learning their corresponding hyperparameters. Subsequently, using the test data, their diagnostics outputs are obtained in matrix form comprised of p columns (one for each class) and $|test|$ rows. SVM and ANN classifiers generate a matrix (m_{SVM} and m_{ANN}, respectively) of deterministic probability estimates of size $|test| \times p$, where each cell specifies the diagnostics probability for each possible health state.
- `Line 8`: The GBN classifier is trained and tested. In the training process its hyperparameters are learned. In the testing process the PDF information is generated, PDF_{BN}, which includes PDFs for each health state for each test sample, i.e. a matrix of PDFs of size $|test| \times p$.
- `Line 9`: the uncertainty information is inferred from the PDF_{BN} outcome of the GBN model resulting in the maximum likelihood value matrix, m_{BN}, and the matrix of the selected uncertainty metric, m_u, such as standard deviation, entropy or kurtosis.

The test matrix (m_{SVM}, m_{ANN}, m_{BN}) can be directly used for diagnostics by assigning the most likely status among all possible faults. However, when the different classifiers diagnose different faults with different probabilities for the same gas samples, the decision-making process is complex. There are some direct solutions that can be applied, e.g. weight the training accuracy of the classifiers and then weight test data accordingly. This strategy assumes that the training data mirrors the test data. Therefore, the performance of this method is directly linked to the similarity of training and testing data, which impacts negatively on the generalization of the method. Algorithm 1 operates as follows with the adopted fusion strategies that are able to combine different classifiers:

- `Lines 10-11`: evaluate Dempster Shafer's theory and Stacking fusion strategies using the outcome of ANN and SVM models along with the maximum likelihood matrix inferred from the GBN model.
- `Line 12`: evaluate the modified Dempster Shafer's theory using the outcome of ANN and SVM models along with the maximum likelihood matrix inferred from the GBN model and the associated uncertainty information.
- `Lines 13-16`: extract and save performance metrics for the i-th trial results and prepare for the next iteration.
- `Lines 17-19`: extract performance statistics for all the classifiers using all the N results saved in `Line 15`.

Subsection III-A to Subsection III-C define training and testing strategies for ANN, SVM and GBN and Subsections III-D and III-E present the fusion methods.

A. Artificial Neural Networks

Artificial neural networks (ANN) are BB models widely used for classification and regression [30]. The multilayer perceptron (MLP) feedforward model was used in this work. The MLP is a three-layer network (input, hidden, output) comprised of fully connected neurons. Each neuron performs a weighted sum of its inputs and passes the results through an activation function. All the designed ANN models use a sigmoid activation function for hidden and output nodes.

Model training is performed using a back-propagation algorithm. The goal is to learn the neuron weights so as to generate the transformer health state (network output) from DGA values (sample input), which minimizes the error with respect to the target transformer health state. Input and hidden layers may also have a bias unit analogous to intercept terms in a regression model. As part of the MCCV process, a number of networks were trained for each trial, using different gases and their ratios at the input layer and varying the number of hidden nodes. For each trial, the experiments were repeated 10 times so as to deal with the stochastic nature of ANN models [30]. Of the trained networks for each trial, the one with the highest mean accuracy was selected. For most of the trials best results were obtained with 20 hidden nodes with the inputs in Fig. 2, i.e. C_2H_6, C_2H_4, H_2, CH_4 and C_2H_2.

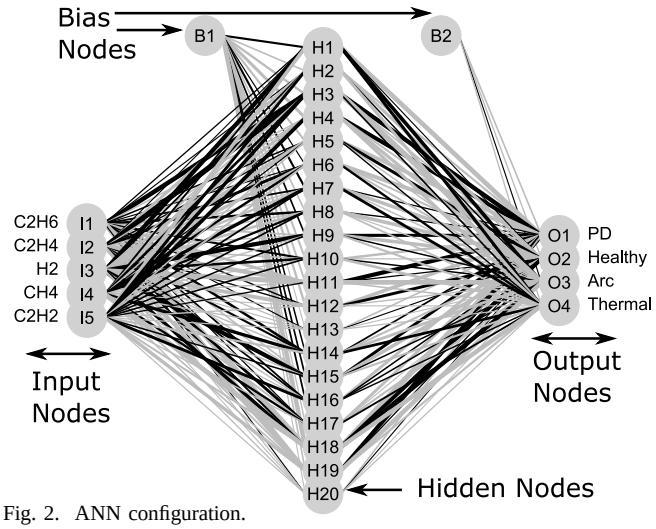


Fig. 2. ANN configuration.

Fig. 2 also shows the strength of the neuron weights with a black line for higher weights and a grey line for lower weights. Model training was performed using the R `nnet` library [31].

B. Support Vector Machines

The SVM maps input data into a space using a kernel function [32]. The SVM learns the boundary separating one transformer health state 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 transformer health state classification boundary.

The SVM is parametrized through the choice of kernel function. For a nonlinear problem, such as the transformer health state estimation, the RBF kernel is recommended [32]: $k(x, x') = exp(-\gamma||x - x'||^2)$, where γ is the RBF width, x and x' are training and testing data samples, and $||d||$ is the Euclidean norm. The SVM solves an optimization problem maximizing the distance from the transformer health classification hyperplane to the nearest DGA training point. Generally the dataset is not linearly separable and slack variables are used to allow wrongly classified samples. SVM penalizes the objective function with a cost variable c, which is a tradeoff between penalizing slack variables and obtaining a large margin for the SVM.

Therefore the SVM training consists of calculating the hyperparameters c and γ. Grid search was used to find the optimal parameters of c and γ from a grid of values. Note also that there are other optimization algorithms for parameter selection, e.g. Bayesian optimization based on Gaussian processes [33]. Namely, for each trial model training was performed using the R `e1071` package [34] and grid search was used to optimize c and γ within $c = [2^{-10}, 2^{10}]$ and $\gamma = [2^2, 2^9]$. A number of configurations were trained using all different gases and their ratios 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, which matches with the input data used for the ANN model (see Fig. 2).

C. Gaussian Bayesian Networks

Bayesian networks (BN) [29] are statistical models that represent probabilistic dependencies among random variables (RVs). In a BN model, a directed acyclic graph represents graphically the causal relation between RVs. 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 RVs using conditional probabilities. When a BN is comprised of continuous RVs a widely implemented approach adopted in this paper is the use of Gaussian BNs (GBN) [29]. In a GBN the conditional distributions are defined through linear Gaussian distributions and local distributions are modelled through Normal RVs, whose PDF is defined as:

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \tag{3}$$

where x is the variable under study, i.e. transformer health state, μ is the mean, and σ^2 is the variance, often denoted as $x \sim N(\mu, \sigma^2)$.

Local distributions are linked through linear models in which the parents, i.e. DGA samples, play the role of explanatory variables. Each node x_i which represents one specific health state of the transformer is regressed over its parent nodes which are explanatory DGA samples. Assuming that the parents of x_i are $\{u_1, \ldots, u_k\}$, then the conditional probability of each node can be expressed as $p(x_i|u_1, \ldots, u_k) \sim N(\beta_0 + \beta_1 u_1 + \ldots + \beta_k u_k; \sigma^2)$, that is:

$$p(x_i|u_1, \ldots, u_k) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{x - (\beta_0 + \beta_1 u_1 + \ldots + \beta_k u_k)}{\sigma})^2) \tag{4}$$

where β_0 is the intercept and $\{\beta_1, \ldots, \beta_k\}$ are linear regression coefficients for the parent nodes $\{u_1, \ldots u_k\}$.

So as to select the input gas variables the Normality of the fault gases was analysed and those gases which follow a Normal distribution were selected so as to match with the underlying probabilistic model and maximize the inferred information. Fig. 3 shows the GBN model comprised of nodes and arrows, where the origin of the arrow is the parent node and the destination is its child node, e.g. the parent nodes of PD are C_2H_6, C_2H_2, CH_4, C_2H_4 and H_2.

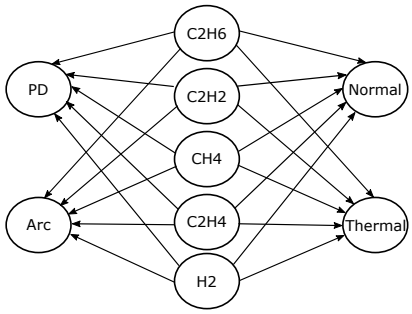


Fig. 3. GBN configuration.

The parameter estimation for GBN models is based on the maximum likelihood (ML) algorithm. The ML expression is derived from the linear Gaussian density function and the closed-form solution can be obtained (see [29] for more details). This process is used to estimate the parameters for each node in the BN model, e.g. for the PD node (Fig. 3): $P(PD|C_2H_6, C_2H_2, CH_4, C_2H_4, H_2) \sim \mathcal{N}(\beta_0 + \beta_1 C_2H_6 + \beta_2 C_2H_2 + \beta_3 CH_4 + \beta_4 C_2H_4 + \beta_5 H_2; \sigma^2)$.

After learning the parameters, the estimation of the conditional probability of nodes, i.e. probability of a specific transformer health state given input DGA data, is based on inferences using the likelihood weighting algorithm [29]. When applied to the DGA dataset, for each of the analyzed transformer health state the outcome of the inference is a set of random samples from the conditional distribution of the transformer health state node given the test DGA samples. From the random samples density values are calculated through Kernel density estimates [35]. The GBN model was implemented using the `bnlearn` R package [36].

D. Ensemble of diagnostics models

Research suggests that combining multiple classifiers can improve individual classifiers [19]. There are a number of different methods for creating ensembles.

1) Dempster Shafer's (DS) theory: DS builds beliefs of the true state of a process from distinct pieces of evidence [21]. Assuming a set of faults $\mathcal{F}$, where the i-th fault is denoted f_i, the set of possible states is called frame of discernment: $\mathcal{F} = \{f_1, \ldots, f_i, \ldots, f_{|F|}\}$. Pieces of evidence are formulated as mass functions, $m : 2^{\mathcal{F}} \longmapsto \mathbb{R}$, satisfying: $m(f_i) \geq 0$, $m(\emptyset) = 0$, and $\sum_{f_i \subseteq \mathcal{F}} m(f_i) = 1$.

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

$$m_{c_1 c_2}(f_i) = \frac{1}{1-K} \sum_{\substack{A,B \subseteq \mathcal{F} \\ A \cap B = f_i}} m_{c_1}(A) m_{c_2}(B) \tag{5}$$

$\forall f_i \subseteq \mathcal{F}, f_i \neq \emptyset$, where K is the degree of conflict between two mass functions:

$$K = \sum_{\substack{A,B \subseteq \mathcal{F} \\ A \cap B = \emptyset}} m_{c_1}(A) m_{c_2}(B) \tag{6}$$

DS theory has been successfully applied to combine independent classifiers (Section I). However, one of its criticisms is the inability to handle some conflicting situations [21].

2) Stacking: The stacking method is based on the meta-learner concept in which a stacking model learns which classifiers are reliable and which are not [19]. Instead of taking the original input variables, a stacked model takes as input the probabilistic outcomes generated from all the independent classifiers. These models are trained first and then tested with both training and testing data. The training and testing of a stacked model is based on the training and testing outcomes of the independent classifiers. Fig. 4 shows the stacking concept.

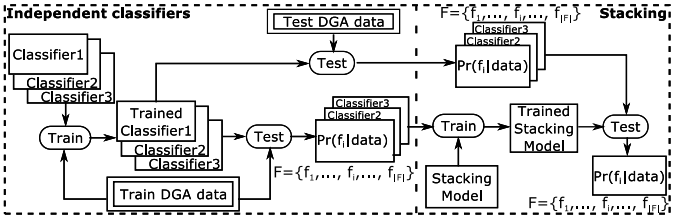


Fig. 4. Stacking configuration.

ANN and SVM models generate a deterministic probability value for each health state. The GBN model generates a PDF for each health state, and the maximum likelihood of each PDF is used in the stacking configuration.

As opposed to DS theory, in the stacking configuration a learning model is trained. An ANN model has been used in this work as a stacking model to aggregate independent classifiers. As part of the MCCV process, for each trial, a number of stacking models are trained varying the number of hidden nodes to select the one with the best performance. In most of the cases the best ANN model is comprised of 10 hidden nodes. The activation function is the sigmoid function.

E. Reasoning under uncertainty with ensemble models

The methods outlined in Subsection III-D have been used for the fusion of black-box classifiers. However, they ignore any uncertainty information which may be generated by the classifiers. There is potential for this information to improve the performance of the ensemble, especially on conflicting samples.

Fixsen and Mahlen proposed the Modified DS (MDS) framework by merging DS theory and Bayesian approaches [37]. In this work the MDS framework is adapted for the particular case of evidence combination of different faults to integrate the uncertainty information generated by WB probabilistic classifiers.

Namely, assuming a set of faults $\mathcal{F}$ with a prior probability π_i for each fault ($1 \leq i \leq |\mathcal{F}|$), the fusion of different classifiers for each fault taking into account the prior information is calculated as follows:

$$m_{c_1c_2}(f_i|\overrightarrow{\pi}) = \frac{m_{c_1}(f_i).m_{c_2}(f_i).\prod_{j=1}^{|\mathcal{F}|\setminus f_i}\pi_j}{\sum_{k=1}^{|\mathcal{F}|}(m_{c_1}(f_k).m_{c_2}(f_k)\prod_{l=1}^{|\mathcal{F}|\setminus f_k}\pi_l)} \quad (7)$$

where $|\mathcal{F}|$ is the cardinality of the set of faults and $\overrightarrow{\pi}$ is the set of priors for each fault $\overrightarrow{\pi} = \{\pi_1, \pi_2, \ldots, \pi_{|F|}\}$.

The strength of the proposed reasoning framework is highlighted with conflicting data samples which are incorrectly classified by independent classifiers. In this situation, the prior information is critical to weight the probabilities and decide which is the real cause of the fault. For example, for two faults f_1 and f_2, and two classifiers c_1 and c_2, (7) reduces to

$$m_{c_1c_2}(f_1|\overrightarrow{\pi}) = \frac{m_{c_1}(f_1).m_{c_2}(f_1).\pi_{f_2}}{m_{c_1}(f_1).m_{c_2}(f_1)\pi_{f_2} + m_{c_1}(f_2).m_{c_2}(f_2)\pi_{f_1}} \quad (8)$$

In the extreme case that both classifiers give the same probabilistic output for both faults, (8) reduces to

$$m_{c_1c_2}(f_1|\overrightarrow{\pi}) = \frac{1}{1+\pi_{f_1}/\pi_{f_2}} \quad (9)$$

From (9) one can observe that the probability mass of fault f_1 is dependent on the ratio between π_{f_2} and π_{f_1}. Namely, the greater the uncertainty of f_2 with respect to f_1, the greater the assigned probability mass to f_1 and the lower the assigned probability mass to f_2. Usually the probability mass values of different faults and different classifiers are not equal, but the same reasoning process is generally applicable for all cases to reason under uncertainty. Therefore (7) creates a suitable framework to integrate uncertainty information in the ensemble of diagnostics classifiers.

The key assumption of this method is that the fusion method accepts a common prior for different mass values. That is, the uncertainty information inferred from a single classification method will be used to influence the combination of different classifiers. Therefore, the generation of representative uncertainty information will be critical. In the set of classifiers analyzed in this work, only the GBN model is able to generate uncertainty information from the classification output. Therefore, uncertainty parameters will be extracted from the density functions inferred by the GBN model so as to reason under uncertainty.

1) Uncertainty parameters: There are different metrics that can be used in order to extract uncertainty information from density functions such as standard deviation, kurtosis or entropy. Depending on the metric, the effect of the prior on the final accuracy is different. Best results were obtained with the standard deviation and weighted log-likelihood, wll, defined as follows:

$$wll = -\frac{1}{M}\sum_{i=1}^{M}(w_i.p_i + log(w_i.p_i)) \quad (10)$$

where M denotes the total number of Kernel density samples, p_i is the diagnosis probability of the fault i, and w_i is the weight assigned to this probability.

IV. Case Studies

The proposed approach is tested on the datasets introduced in Section II. So as to validate and generalize the result all the models and ensemble strategies have been examined N=10^3 times using the MCCV strategy. For each trial, firstly the dataset is shuffled, then it is divided into training and testing sets, and finally training and testing steps are completed. After randomly shuffling the dataset, different training and testing data proportions and data split strategies have been tested (cf. Algorithm 1, `line 5`):

- *80%-20% global*: all the dataset is divided into 80% and 20% for training and testing, respectively. In the testing set there is always at least one sample of each state.
- *80%-20% class-by-class*: each health state is divided into 80% and 20% for training and testing, respectively.
- *70%-30% class-by-class*: each health state is divided into 70% and 30% for training and testing, respectively.

The *80%-20% global* strategy reflects closely the real transformer operation. However, this strategy affects the number of samples for each health state in the testing set. The *class-by-class* strategies ensure the same amount of randomly sampled data samples per each group for each trial.

Generally there are four possible outcomes for a classifier. True positive (TP) when there is a fault and it is correctly diagnosed, false positive (FP) when there is no fault, but the classifier diagnoses a fault, true negative (TN) when there is no fault and the classifier does not diagnose any fault, and false negative (FN) when there is a fault and it is not correctly diagnosed. In addition to the accuracy indicator (cf. Algorithm 1, `line 14`), which quantifies the percentage of correct predictions over the total number of predictions, four complementary classification metrics have been analysed.

- Positive preditive value (PPV): $PPV = \frac{TP}{TP+FP}$
- Negative preditive value (NPV): $NPV = \frac{TN}{TN+FN}$
- False Positive rate (FPR): $FPR = \frac{FP}{FP+TN}$
- F1 score (F1): $F1 = \frac{2TP}{2TP+FP+FN}$

PPV and NPV quantify respectively the proportions of positive and negative results in diagnostics tests. PPV is different from accuracy because it considers only TP and FP events. The PPV is also known as precision and its complement is the false discovery rate. The complement of the NPV is the false omission rate. The complement of the FPR is the specificity. The F1 score is the harmonic mean of PPV and recall, which is commonly used for unbalanced classification problems.

For multiclass classification problems, the classifier outcomes and metrics are counted per class, and then they are averaged according to the prevalence of each class.

A number of independent classifiers and ensemble strategies have been examined:

#1 Gaussian Bayesian Networks.
#2 Support Vector Machines.
#3 Artificial Neural Networks.
#4 Stacking with ANN, SVM and GBN models aggregated with an ANN model.
#5 Dempster-Shafer with ANN, SVM, and GBN models.
#6 Modified DS with ANN, SVM, and GBN models using the standard deviation of GBN results as a prior.
#7 Modified DS with ANN, SVM, and GBN models using the weighted log-likelihood of GBN results as a prior.

A. Results & Discussion

The accuracy results for the IEC TC 10 and Extended datasets are displayed in Table II. The best performing results with highest mean accuracy and lowest deviation are highlighted in bold.

Table II confirms that the overall accuracy of the proposed novel configurations (#6, #7) are higher than other fusion (#4, #5) and machine learning methods (#1-#3) for both datasets. The order of the accuracy improvement of the proposed configurations with respect to other fusion and machine learning methods remains the same for both datasets, which confirms the validity and consistency of the proposed approach.

As for the data training and testing strategies, it is possible to see that the accuracy decreases for all the configurations across both datasets when decreasing the size of the training set from 80% to 70%. Additionally, the *class-by-class* strategy reduces the standard deviation (SD) of the results by imposing a predefined number of samples in the testing set. For PD samples the SD is bigger compared with the rest of the states because the accuracy values for most of the trials are concentrated at one value with few outliers.

As for the comparison between datasets, in general the overall accuracy improves with the Extended dataset. This is due to an improved capability to detect Thermal and Arc faults, purely because the training dataset contains more examples of these fault types. Conversely, the accuracy of PD faults decreases with the Extended dataset. The trend of the PD samples on the IEC TC 10 dataset is predictable (H_2≃[10000-80000], CH_4≃[1000-18000], C_2H_6≃[100-2000], C_2H_2≃[1-25], C_2H_4≃[1-25], all in ppm). However, with the Extended dataset the PD is more complex to diagnose due to the introduced additional data samples for all fault types. For instance, another PD sample is added (H_2=980, CH_4=73, C_2H_6=58, C_2H_2=0.1, C_2H_4=1.2, all in ppm) [27], which is more complex to diagnose and therefore, the PD accuracy decreases.

As for the diagnostics capacity of specific models, it can be seen that the GBN has a good performance for identifying PD faults. Then, the classification outputs of the GBN model also have less uncertainty for this fault, which in turn leads to improving the ensemble models when including the prior, e.g. see PD diagnostics accuracy for the IEC TC 10 dataset. The improvements for Arc, Normal, and Thermal faults are similar for all the fusion methods, with slight improvements when including the uncertainty information in the ensemble.

For the Extended dataset the GBN model has a decreased accuracy for the Normal state. This affects the fusion strategies as the prior becomes less informative and the accuracy of the proposed fusion strategies for the Normal state becomes less accurate. In contrast, the GBN model has an increased accuracy for the Thermal state for the same dataset. In this case, this benefits the fusion strategies because the prior becomes more informative and the accuracy of the proposed fusion strategy for Thermal faults becomes more accurate.

Table III displays more performance metrics. For the overall metrics, the best models in terms of F1, PPV, NPV and FPR are the proposed fusion strategy results #6 and #7.

The PPV improvement of the proposed strategy results are in the same order of improvement as the accuracy results. The only difference with respect to the accuracy results in Table II is the increased percentage value of PPV results due to the definition of PPV, i.e. it only considers TP and FP events and no FN events. The NPV is very high for all the tested configurations. That is, these models are able to correctly detect when a fault class has not occurred. The FPR

TABLE II
CLASSIFICATION ACCURACY OF INDEPENDENT & ENSEMBLE MODELS.

A^*	*80%-20% global*										*80%-20% class-by-class*										*70%-30% class-by-class*									
	Overall		*Thermal*		*PD*		*Arc*		*Normal*		*Overall*		*Thermal*		*PD*		*Arc*		*Normal*		*Overall*		*Thermal*		*PD*		*Arc*		*Normal*	
	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$
#1	82.1	6.3	68.3	17.9	97.3	9.4	93.6	6.7	72.9	14.9	82.2	6.1	68.3	15.8	**100**	**0**	93.7	6.4	73.1	14.5	82.3	5	67.4	13.3	97	9.8	94.2	5.4	71.9	11.9
#2	86.6	6	71.6	18.3	93.1	17	92.8	7.3	87.6	11.2	86.6	5.4	70.9	16	99.4	7.7	93	7.1	87.5	11	86	4.6	70.5	14	87.9	19.8	92.6	6.1	86.5	9
#3	89.4	5.3	78.5	16.4	91.6	18.5	95.4	5.7	88.7	10.7	89.5	4.8	78.6	14.8	94.7	15.9	95.2	5.6	88.9	10.1	88.8	4	76.6	13	88.6	16.5	95	4.6	87.9	8.6
#4	89.7	5.3	**79.1**	**16.3**	91.5	20.6	95.5	5.6	88.5	10.9	89.8	4.8	**79.1**	**14.8**	89.8	30.2	95.5	5.6	88.8	10	89	4	77.2	13	90.2	16.1	95.1	4.6	87.8	8.8
#5	90.2	5.4	77.5	16.6	94.5	15.1	96.2	5.4	**89.9**	**10.3**	90.2	4.9	77.4	15.1	97.8	14.7	96	5.4	**90.4**	**9.5**	89.4	3.9	75.9	12.9	91.4	16.2	95.6	4.6	**89**	**8.1**
#6	**90.7**	**5.2**	78.5	16.5	**99**	**6**	**96.3**	**5.4**	89.5	10	**90.7**	**4.9**	79.1	14.9	99.9	3	**96.2**	**5.3**	89.6	9.9	89.9	4	**77.4**	**12.7**	98.2	7.9	**95.8**	**4.5**	87.9	8.6
#7	**90.7**	**5.2**	78.5	16.5	**99**	**6**	**96.3**	**5.4**	89.5	10	90.6	4.8	78.6	14.9	**100**	**0**	**96.2**	**5.3**	89.7	9.8	**90**	**3.9**	77	12.7	**99.1**	**5**	**95.8**	**4.5**	88.4	8.4

B^*	*80%-20% global*										*80%-20% class-by-class*										*70%/30% class-by-class*									
	Overall		*Thermal*		*PD*		*Arc*		*Normal*		*Overall*		*Thermal*		*PD*		*Arc*		*Normal*		*Overall*		*Thermal*		*PD*		*Arc*		*Normal*	
	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$
#1	79.9	4.6	91	6.1	73.1	29.7	90.1	5.4	50.8	12.8	80.6	4.1	91.1	6.1	73.3	25.7	90.4	5.5	51.4	12.7	80.3	3.3	90.9	4.8	65.4	24.4	89.8	4.3	51.4	10
#2	88.9	3.6	90.3	6.6	81.7	23.8	93.4	5.1	82.4	11.3	89.3	3.5	90.6	6.5	77.4	25.3	93.6	5	82.7	10.8	88.7	3.3	90.4	5.5	73.6	22.3	93.1	4.4	80.5	10
#3	90.9	3.8	91.3	6.4	87.4	20.9	94.8	4.7	85.2	10	91.1	3.5	91.5	5.9	83.6	23.6	95	5	85	9.3	90.6	2.9	90.9	5.1	83.6	19.3	94.5	4.4	84.2	8.3
#4	90.9	3.4	91	6.5	76.7	31.4	**95.1**	**4.4**	**85.9**	**9.7**	91.1	3.4	91.2	6	78.3	27.9	**95.4**	**4.6**	**85.6**	**9.2**	90.7	2.9	90.8	5.3	79.9	20.4	**94.9**	**3.9**	**84.5**	**3.9**
#5	91.4	3.4	92.6	5.8	85.2	22.3	95	4.6	85.4	9.9	91.6	3.4	92.9	5.6	80.8	24.6	95.3	4.7	85.4	9.3	91.1	3	92.2	4.9	80.7	19.8	94.8	4.1	84.2	8.5
#6	**91.9**	**3.5**	**93.3**	**5.5**	92.8	16.2	95	4.7	84.6	10.3	**92.1**	**3.4**	**93.8**	**5.2**	90.7	19.4	95.2	4.8	84.6	9.6	**91.5**	**2.9**	**92.9**	**4.6**	91	15	94.8	4.1	83.3	8.7
#7	91.8	3.5	93.2	5.6	**93**	**15.7**	95.1	4.6	84.3	10.4	92	3.4	93.5	5.3	**91**	**19.2**	95.3	4.8	84.4	9.6	91.4	2.9	92.8	4.6	**91.4**	**14.9**	94.9	4.1	83.1	8.7

* **A**: IEC TC 10 dataset, **B**: Extended dataset. #1: GBN, #2: SVM, #3: ANN, #4: Stacking, #5: DS, #6: MDS with SD, #7: MDS with WLL.

TABLE III
PERFORMANCE METRICS OF INDEPENDENT & ENSEMBLE MODELS.

A^*	*80%-20% global*								*80%-20% class-by-class*								*70%-30% class-by-class*							
	F1		*PPV*		*NPV*		*FPR*		*F1*		*PPV*		*NPV*		*FPR*		*F1*		*PPV*		*NPV*		*FPR*	
	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
#1	82.6	6.6	83.9	5.9	91.8	3.7	9.7	4	81.1	6.5	83.5	6.3	92.5	3.2	10.4	4	81.7	5.3	83	5.1	92.9	2.5	9.9	3.1
#2	86.5	5.9	88.9	5.3	93.9	3.3	6.5	3.2	86.6	5.5	89.3	4.5	94.2	3.2	7	3	85.8	4.6	87.1	4.3	94.3	2.3	6.9	2.3
#3	89.4	5.4	90.6	4.9	95.4	2.8	5.2	2.9	89.8	4.7	91.1	4.2	95.6	2.3	5.5	2.7	88.6	4.1	89.6	3.8	95.6	1.8	5.5	2.1
#4	89.6	5.4	90.9	4.9	95.5	2.8	5.1	2.9	90.1	4.8	91.2	4.3	95.8	2.3	5.4	2.7	88.8	4.2	89.8	3.4	95.6	1.8	5.4	2.1
#5	90.1	5.4	91.4	4.8	95.8	2.8	4.9	2.9	90	4.9	91.6	4.3	96.1	2.4	5.1	2.9	89.2	4.2	90.2	3.7	95.9	1.8	5.3	2.1
#6	90.6	5.4	**91.9**	**4.8**	**95.9**	**2.8**	**4.7**	**2.9**	**90.6**	**5**	**91.9**	**4.3**	**96.2**	**2.3**	**5**	**2.8**	89.7	4.1	90.7	3.8	95.9	1.8	**5.1**	**2.1**
#7	**90.7**	**5.4**	**91.9**	**4.8**	**95.9**	**2.8**	**4.7**	**2.9**	90.5	5	91.8	4.2	96.1	2.4	5.1	2.8	**89.8**	**4**	**90.8**	**3.7**	**96**	**1.8**	**5.1**	**2.1**

B^*	*80%-20% global*								*80%-20% class-by-class*								*70%-30% class-by-class*							
	F1		*PPV*		*NPV*		*FPR*		*F1*		*PPV*		*NPV*		*FPR*		*F1*		*PPV*		*NPV*		*FPR*	
	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
#1	79	8	81.3	4.3	92.1	2.1	9.7	2.4	79.6	4.5	81.5	4.5	92.6	1.9	9.4	2.2	79.7	3.5	81.1	3.6	92.5	1.6	10.5	1.8
#2	88.9	3.9	89.7	3.7	95.1	1.8	5.1	2	89.4	3.9	90	3.8	95.6	2.7	5	1.9	88.6	3.3	89.1	3.2	95.2	2.6	5.4	1.7
#3	90.9	3.6	91.6	3.4	95.9	1.8	4	1.8	91.1	3.4	91.7	3.3	96.2	1.6	4	1.7	90.6	2.9	91.1	2.8	96	1.4	4.2	1.4
#4	90.9	3.5	91.7	3.3	96	1.7	4	1.8	91.2	3.3	91.8	3.2	96.3	1.5	4	1.6	90.7	2.9	91.1	2.8	96.1	1.4	4.3	1.4
#5	91.4	3.5	92	3.3	96.2	1.7	4	1.8	91.6	3.4	92.1	3.2	96.5	1.6	4	1.7	91.1	3	91.5	2.9	96.2	1.4	4.2	1.5
#6	**91.9**	**3.5**	**92.4**	**3.4**	**96.3**	**1.7**	**3.7**	**1.8**	**92**	**3.4**	**92.5**	**3.2**	**96.6**	**1.6**	**3.7**	**1.7**	**91.5**	**3**	**91.8**	**2.9**	**96.3**	**1.4**	**3.9**	**1.4**
#7	91.8	3.5	92.3	3.4	**96.3**	**1.7**	3.8	1.8	91.9	3.5	92.4	3.3	**96.6**	**1.6**	3.8	1.7	91.4	3	91.7	2.9	**96.3**	**1.4**	4	1.4

* **A**: IEC TC 10 dataset, **B**: Extended dataset. #1: GBN, #2: SVM, #3: ANN, #4: Stacking, #5: DS, #6: MDS with SD, #7: MDS with WLL.

improvement of the proposed strategy is in the same order of improvement as the accuracy results. This is caused by the reduced number of FP events and increased number of TN events as confirmed by the NPV results. Finally, the F1 score is very similar to the accuracy results both in the order of improvement and absolute values. The F1 score is a combined metric of precision and recall and therefore it includes the number of correctly classified instances as well as FP and FN events.

As for the effect of different training and testing strategies on the performance results, the *class-by-class* strategy reduces the SD of the results as happened with the accuracy results in Table II. Concerning the effect of the size of the dataset, a decrease in the training set causes a decrease of F1 and PPV scores and an increase of the FPR score indicating a decreased accuracy and an increased false positive rate respectively, while the NPV score remains high for all the configurations. Finally, with respect to the performance comparison across datasets, the results are again consistent with the accuracy results in Table II. That is, F1, PPV and NPV scores increase and FPR decreases with the Extended dataset due to the extended number of samples per health state.

Tables II and III agree that the accuracy and the performance of the proposed fusion strategies (#6, #7) are superior to other machine learning (#1-#3) and fusion (#4, #5) models. The main factor which makes a difference among these models is the post-processing and integration of the uncertainty information in the ensemble of classifiers. This is dependent on the used WB approach and the post-processed uncertainty information in the form of uncertainty metrics. These metrics along with the combination of machine learning methods enable the resolution of conflicting samples. As demonstrated in the next section, the order of improvement of the proposed method with respect to existing fusion methods is correlated with the amount of conflicting diagnostics samples. That is, the more conflicting samples the better the accuracy and performance of the proposed approach.

B. Decision Making Under Uncertainty

From an engineering viewpoint the data samples which create disagreement among the source classifiers are the most important cases. Table IV displays the accuracy results considering only conflictive samples, i.e. data samples which create disagreements among GBN, ANN and SVM models.

TABLE IV
CLASSIFICATION ACCURACY FOR CONFLICTIVE DATA SAMPLES.

Strategy	*Dataset*	*#4*		*#5*		*#6*		*#7*	
		$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$
80%-20% whole dataset	IEC TC 10	73.6	19.4	75.7	19.3	**78.2**	**18.5**	78	18.4
	Extended	75.9	11.7	77.3	12	**78.8**	**12**	78.5	12
80%-20% class by class	IEC TC 10	74	19.1	76.1	18.9	**78.4**	**18.6**	77.8	18.8
	Extended	75.77	12	77.4	11	**78.6**	**11.8**	78.3	12
70%-30% class-by-class	IEC TC 10	70.9	16.3	72.9	15.5	75	15.5	**75.5**	15
	Extended	74.3	10.1	76	10.4	77	10.3	**76.9**	10.3

Results in Table IV are in agreement with the results in Table II. However, the overall accuracy is lower because consistently diagnosed data samples are removed, and the differences in the accuracy of the fusion methods are higher because only conflictive cases are taken into account. Under conflicting situations, the proposed uncertainty-aware fusion strategy is more effective due to the accuracy improvements for all health states (cf. Table II). This accuracy is a critical value for any ensemble approach because the strength of the method is highlighted when independent classifiers diagnose different faults and it is able to reason under uncertainty.

The proposed model is able to assist the engineer in the decision-making process. For instance, consider that after training the classifiers they are tested for the following absolute gas values [8]: H_2 = 26788 ppm, C_2H_4 = 27 ppm, C_2H_6 = 2111 ppm, C_2H_2 = 1 ppm, CH_4 = 18342 ppm and the observed fault type is PD. Table V displays probabilistic results for different classifiers, $m_{classifiers}$.

TABLE V
EXAMPLE A: DIAGNOSTICS RESULTS OF SOURCE CLASSIFIERS.

ID	*Pr(Normal)*	*Pr(Thermal)*	*Pr(Arc)*	*Pr(PD)*
#1	0.23	0.28	0.18	**0.31**
#2	0.08	**0.45**	0.07	0.4
#3	3.9E-2	**0.5**	4.8E-6	0.46

The results of the independent classifiers highlight their disagreement. BB models do not generate uncertainty information, but observing the output of the GBN the PDFs for different faults and the associated uncertainty can be inferred. Fig. 5 shows the GBN's output for the considered example. That is, ID #1 in Table V without normalising probabilities.

The x-axis in Fig. 5 denotes random samples drawn from the conditional distribution of the node given the evidence, $P(f_i|C_2H_6, C_2H_4, H_2, CH_4, C_2H_2)$. The x-axis value of the peak density indicates the maximum likelihood value. The greater the peak density value, the narrower the variance, and the higher the confidence of the GBN model in the diagnostics.

For instance, the density function of the PD fault shows a narrow function with a high peak density value with a maximum likelihood value located at 0.7. This suggests that GBN is very confident that PD is the type of fault present for these test gas values. Thermal fault has a maximum likelihood value of 0.65, but its standard deviation is greater than the PD fault, which indicates the decreased confidence of the GBN that this is the true fault. The density functions for the rest of faults located at lower x-axis probability values, indicate that they are not the cause of this fault.

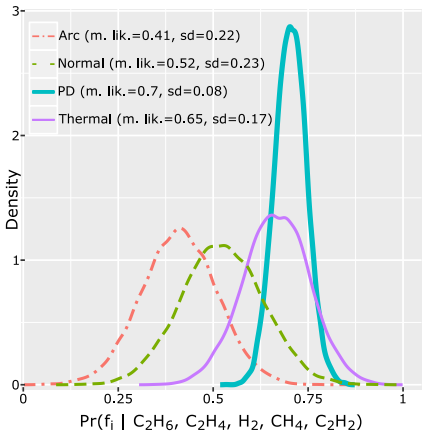


Fig. 5. Example A: GBN diagnostics output.

It is possible to evaluate different uncertainty metrics in Fig. 5 and use them as priors in (7) so as to influence the fusion strategy. Using the standard deviation [Eq. (2)] and weighted log-likelihood [Eq. (10)] as the prior, Table VI displays the results of the analyzed ensemble strategies.

TABLE VI
EXAMPLE A: DIAGNOSTICS RESULTS OF ENSEMBLE MODELS.

ID	*Pr(Normal)*	*Pr(Thermal)*	*Pr(Arc)*	*Pr(PD)*
#4	0.027	**0.59**	0.023	0.36
#5	0.0061	**0.522**	0.0019	0.47
#6	0.0019	0.223	1.3E-7	**0.775**
#7	0.0005	0.107	2.9E-8	**0.892**

The fusion methods stacking and DS (#4, #5 Table VI) do not identify the actual fault. However, the proposed approach (#6, #7) which uses the uncertainty information inferred from the GBN model is effective in resolving conflictive samples.

The crucial point of this method is the accuracy of the WB model and conflictive cases. The GBN model has a good performance for identifying PD faults. Therefore, this leads to improving the ensemble models when including the prior, because the uncertainty associated with the PD fault is lower. However, note also that the deterministic probability values of different classifiers count in the ensemble [cf. Eq. (7)], and therefore, the fusion is not biased by the potential poor performance of the GBN model. For instance, the GBN performs worse than ANN or SVM for Normal and Thermal faults, but the ensemble strategy improves the final accuracy.

In another test, the classifiers are tested for the following absolute gas values [8]: H_2 = 290 ppm, CH_4 = 966 ppm, C_2H_2 = 57 ppm, C_2H_4 = 1810 ppm, C_2H_6 = 299 ppm and the observed fault type is a Thermal fault. Table VII displays the classification results for different source classifiers.

In this case all the classifiers consistently diagnose a normally degrading transformer. Examining the output of the GBN model in Fig. 6 (i.e. #1 in Table VII, normalised), it is possible to see the uncertainty information of the diagnosis.

TABLE VII
EXAMPLE B: DIAGNOSTICS RESULTS OF SOURCE CLASSIFIERS.

ID	*Pr(Normal)*	*Pr(Thermal)*	*Pr(Arc)*	*Pr(PD)*
#1	**0.31**	0.29	0.24	0.16
#2	**0.47**	0.44	0.08	0.01
#3	**0.54**	0.45	0.0082	0.0018

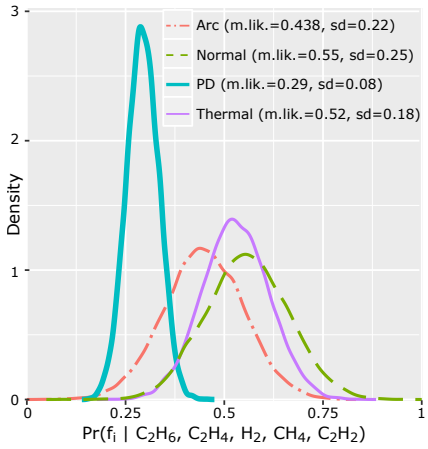


Fig. 6. Example B: GBN diagnostics output.

Although the Normal fault has the highest maximum likelihood value among all faults, the GBN's diagnostics for the Thermal fault has higher confidence with a slightly lower maximum likelihood value. Using the uncertainty information of the GBN model, Table VIII displays the fusion results.

TABLE VIII
EXAMPLE B: DIAGNOSTICS RESULTS OF ENSEMBLE MODELS.

ID	*Pr(Normal)*	*Pr(Thermal)*	*Pr(Arc)*	*Pr(PD)*
#4	**0.57**	0.39	0.02	0.02
#5	**0.578**	0.42	2.4E-4	5.4E-7
#6	0.431	**0.568**	2.36E-4	3.25E-6
#7	0.448	**0.55**	2.1E-4	9.2E-7

Stacking and DS (#4, #5 in Table VIII) do not identify the actual fault. However, the proposed fusion strategy (#6, #7 in Table VIII) again is effective in resolving conflictive samples.

Consider the classifiers are tested for the following values [8]: H_2 = 250 ppm, CH_4 = 150 ppm, C_2H_2 = 150 ppm, C_2H_4 = 150 ppm, C_2H_6 = 250 ppm and the observed health state is Normal. Table IX displays the classification results for the source classifiers.

TABLE IX
EXAMPLE C: DIAGNOSTICS RESULTS OF SOURCE CLASSIFIERS.

ID	*Pr(Normal)*	*Pr(Thermal)*	*Pr(Arc)*	*Pr(PD)*
#1	0.296	0.195	**0.33**	0.179
#2	**0.6**	0.04	0.34	0.02
#3	0.38	5E-4	**0.61**	1e-4

GBN and ANN diagnose an Arc fault (#1, #3 in Table IX), while SVM diagnoses a Normal transformer (#2 in Table IX). Uncertainty information of the GBN's diagnosis is inferred from the GBN output in Fig. 7 (#1 in Table IX, normalised).

The Arc fault has the highest maximum likelihood value and the Normal state has slightly higher confidence with a slightly lower maximum likelihood value. Using the uncertainty information of the GBN model, Table X displays fusion results.

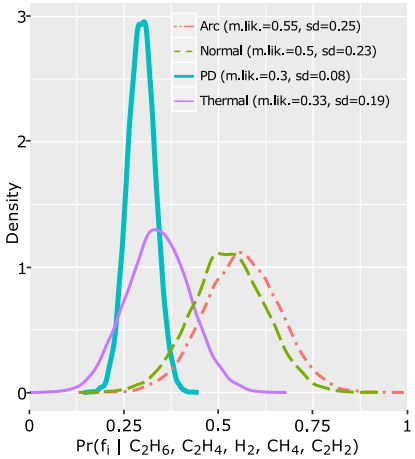


Fig. 7. Example C: GBN diagnostics output.

TABLE X
EXAMPLE C: DIAGNOSTICS RESULTS OF ENSEMBLE MODELS.

ID	*Pr(Normal)*	*Pr(Thermal)*	*Pr(Arc)*	*Pr(PD)*
#4	0.195	0.01	**0.78**	0.015
#5	0.49	2.5E-5	**0.5**	2.76e-6
#6	**0.54**	3.9E-5	0.45	2e-6
#7	**0.54**	2E-5	0.45	4.1e-6

The proposed fusion strategy effectively diagnoses the Normal state and this justifies why despite the accuracy of the GBN being lower, the fusion improves the final diagnosis accuracy (Table II). Note that the GBN diagnosis results in Figs. 5-7 show the non-normalized probabilities corresponding to different Monte Carlo trials and this results in different SD values.

Accordingly, results in Table IV report the accuracy of the ensemble taking into account only conflictive diagnostics of source classifiers and the presented examples focus on conflicts among the source classifiers. These examples can be individually analysed with classical DGA methods. For instance, in Fig. 5 the Duval's triangle correctly identifies a PD fault, Roger indicates normal degradation and Doernenburg does not give a diagnostics or in Fig. 7, the Duval's triangle incorrectly identifies an Arc fault, and Roger and Doernenburg do not give a diagnostics. Even if there is a correct diagnostics by some of the classical methods, their overall diagnostics accuracy is lower. There are other cases where classical methods do not diagnose the correct fault and all the analysed models consistently diagnose the correct fault and this causes the difference in the overall accuracy. Additionally, note that the classical methods are not probabilistic models [7], which makes it difficult to solve conflicts (see Subsection IV-C).

Note also that the density functions generated by the GBN model (e.g., Figs 5-7) do not only help to improve the accuracy of the ensemble, but they also represent a more intuitive visualization for understanding the conflicts. This representation should help to increase the trust of the engineer in the technique as opposed to deterministic probability values inferred from black-box models.

C. Comparison to other methods

The results obtained by the proposed fusion framework are better than other models tested in the same conditions (in this paper) and very close to results obtained with the same dataset but tested in different conditions (reported in the literature). This demonstrates that despite the challenging conditions (multiclass, imbalanced inspection data), the performance is comparable to binary classifiers and to the techniques which use resampling methods (see Table I).

Results displayed in Table II confirm that the proposed fusion strategy improves the accuracy compared with other fusion methods (Dempster Shafer, Stacking) and classifiers (ANN, GBN, SVM). Table XI displays the accuracy of the classical methods using the *80%-20% global* sampling strategy. There is no need to train classical models, but for direct comparisons with Table II, the same testing data samples have been used for machine learning and classical methods.

TABLE XI
COMPARISON WITH CLASSICAL METHODS.

Dataset	***Method***	*Overall*		*Thermal*		*PD*		*Arc*		*Normal*	
		$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$	$\hat{a}$	$sd_{\hat{a}}$
TC 10	Rogers	42.1	7.7	58.9	18.6	0	0	66	11.6	**4**	**5.7**
	Doern.	55.6	7.7	79.4	15.3	46.7	37.6	83.6	9.2	0	0
	Duval	**67.8**	**7.2**	**88.4**	**12.2**	**100**	**0**	**100**	**0**	0	0
Extend.	Rogers	47.7	5.8	74.9	8.5	0	0	54.2	9.2	**4.3**	**5.2**
	Doern.	59.4	5.7	89.2	6.3	47.7	35	69.5	8.6	0	0
	Duval	**70.9**	**5.2**	**94**	**5**	**90.5**	**21.3**	**90.1**	**5.6**	0	0

The overall accuracy results of the proposed approach (cf. Table II) are better compared with the classical methods for both datasets. This is mainly caused by the detection of normally degrading transformers. Duval has an excellent accuracy for PD and Arc faults tested on the IEC TC 10 dataset. However, the overall accuracy is negatively affected because it is not able to diagnose normally degrading transformers. When testing the Extended dataset, the accuracy of the Duval's triangle for PD and Arc faults decreases and for the Thermal fault increases. This occurs because the boundaries between diagnostic regions of the triangle are fixed, as opposed to statistical learning strategies which can adapt to training data. The performance of Rogers and Doernenburg models is lower for both datasets compared with the Duval's triangle.

V. CONCLUSIONS

Transformers are key assets for the reliable and cost-effective operation of the power grid and DGA is an industry-accepted standard method used to monitor transformers. However, the use of classical DGA models or black-box classifiers may generate conflicting diagnostics outputs which are difficult to resolve due to the lack of uncertainty information generated by these models. This situation complicates the decision-making process for engineers.

In order to increase the confidence of the engineer in the decision-making process this paper presents a novel method which takes into account uncertainty information when integrating the output of different classifiers. Using the proposed method for DGA, the accuracy with respect to other fusion methods has improved and the model shows that it is effective for correcting conflictive samples when the prior information inferred from probability density functions is informative.

The results obtained in this paper can be used as a benchmark to other techniques because the used datasets are publicly available. So as to extract general accuracy statistics the models were cross-validated using Monte Carlo cross validation and different proportions and sampling strategies for dividing training and testing strategies have been tested.

Future work can address the integration of other white-box methods or the extension of the approach to combine prior information from multiple sources. This extension may be able to create a more informative prior distribution by combining, e.g. uncertainty information with different fault gas indicators. Ultimately, this enhanced model may open the way for the identification of multiple simultaneous fault conditions.

REFERENCES

[1] M. J. Heathcote, *J & P Transformer Book*, 13rd ed. Oxford: Newnes, 2007.

[2] D. Codetta-Raiteri and L. Portinale, "Dynamic bayesian networks for fault detection, identification, and recovery in autonomous spacecraft," *IEEE Trans. Syst., Man, and Cybern., Syst.*, vol. 45, no. 1, pp. 13–24, Jan 2015.

[3] L. Jiao, T. Denoeux, and Q. Pan, "A hybrid belief rule-based classification system based on uncertain training data and expert knowledge," *IEEE Trans. Syst., Man, and Cybern., Syst.*, vol. 46, no. 12, pp. 1711–1723, Dec 2016.

[4] N. Daroogheh, A. Baniamerian, N. Meskin, and K. Khorasani, "Prognosis and health monitoring of nonlinear systems using a hybrid scheme through integration of PFs and neural networks," *IEEE Trans. Syst., Man, and Cybern., Syst.*, vol. 47, no. 8, pp. 1990–2004, Aug 2017.

[5] Y. Han and Y. H. Song, "Condition monitoring techniques for electrical equipment-a literature survey," *IEEE Trans. Pow. Del.*, vol. 18, no. 1, pp. 4–13, Jan 2003.

[6] 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.

[7] J. I. Aizpurua, V. M. Catterson, B. G. Stewart, S. D. J. McArthur, B. Lambert, A. Bismark, G. Pereira, and J. Cross, "Improving the accuracy of transformer DGA diagnosis in the presence of conflicting evidence," in *Proc. of IEEE Electr. Ins. Conf.*, Baltimore, USA, 2017.

[8] M. Duval and A. dePablo, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," *IEEE Electr. Ins. Mag.*, vol. 17, no. 2, pp. 31–41, March 2001.

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

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

[11] 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, no. 1, pp. 69–79, Jan 2009.

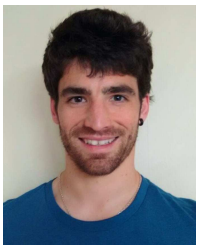
[12] J. Li, Q. Zhang, K. Wang, J. Wang, T. Zhou, and Y. Zhang, "Optimal dissolved gas ratios selected by genetic algorithm for power transformer fault diagnosis based on support vector machine," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 2, pp. 1198–1206, April 2016.

[13] Q. Su, L. L. Lai, and P. Austin, "A fuzzy dissolved gas analysis method for the diagnosis of multiple incipient faults in a transformer," *IEEE Trans. Pow. Sys.*, vol. 15, no. 2, pp. 593–598, May 2000.

[14] S. A. Khan, M. D. Equbal, and T. Islam, "A comprehensive comparative study of DGA based transformer fault diagnosis using fuzzy logic and ANFIS models," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 22, no. 1, pp. 590–596, Feb 2015.

[15] H. Ma, T. K. Saha, C. Ekanayake, and D. Martin, "Smart transformer for smart grid - intelligent framework and techniques for power transformer asset management," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 1026–1034, March 2015.

[16] S. Li, G. Wu, B. Gao, C. Hao, D. Xin, and X. Yin, "Interpretation of DGA for transformer fault diagnosis with complementary SaE-ELM and arctangent transform," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 1, pp. 586–595, February 2016.
[17] J. Fan, F. Wang, Q. Sun, F. Bin, F. Liang, and X. Xiao, "Hybrid RVM-ANFIS algorithm for transformer fault diagnosis," *IET Generation, Transmission & Distrib.*, vol. 11, pp. 3637–3643(6), September 2017.
[18] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358 – 3378, 2007.
[19] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 1st ed. Chapman & Hall/CRC, 2012.
[20] M. Allahbakhshi and A. Akbari, "Novel fusion approaches for the dissolved gas analysis of insulating oil," *IJST Transactions of Electrical Engineering*, vol. 35, no. E1, p. 13, 2011.
[21] 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, 2013.
[22] S. S. M. Ghoneim, I. B. M. Taha, and N. I. Elkalashy, "Integrated ANN-based proactive fault diagnostic scheme for power transformers using dissolved gas analysis," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 3, pp. 1838–1845, June 2016.
[23] H. Malik and S. Mishra, "Application of gene expression programming (GEP) in power transformers fault diagnosis using DGA," *IEEE Trans. Industry Applications*, vol. 52, no. 6, pp. 4556–4565, Nov 2016.
[24] S. Shankaraman and K. Goebel, "Uncertainty in prognostics and systems health management," *International Journal Prognostics and Health Management*, vol. 6, no. 10, p. 14, 2015.
[25] L. Ganyun, C. Haozhong, Z. Haibao, and D. Lixin, "Fault diagnosis of power transformer based on multi-layer SVM classifier," *Electric Power Systems Research*, vol. 74, no. 1, pp. 1 – 7, 2005.
[26] X. Z. Wang, M. Z. Lu, and J. B. Huo, "Fault diagnosis of power transformer based on large margin learning classifier," in *IEEE Int. Conf. on Machine Learning and Cybernetics*, Aug 2006, pp. 2886–2891.
[27] S. Seifeddine, B. Khmais, and C. Abdelkader, "Power transformer fault diagnosis based on dissolved gas analysis by artificial neural network," in *IEEE Int. Conf. Renewable Energies and Vehicular Technology*, March 2012, pp. 230–236.
[28] Q.-S. Xu and Y.-Z. Liang, "Monte carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.
[29] R. E. Neapolitan, *Learning Bayesian Networks*. Prentice Hall, 2004.
[30] M. R. G. Meireles, P. E. M. Almeida, and M. G. Simoes, "A comprehensive review for industrial applicability of artificial neural networks," *IEEE Trans. Ind. Electron.*, vol. 50, no. 3, pp. 585–601, June 2003.
[31] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0.
[32] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov*, vol. 2, no. 2, pp. 121–167, 1998.
[33] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.
[34] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-25." 2011.
[35] J. Kim and C. Scott, "Robust kernel density estimation," in *2008 IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2008, pp. 3381–3384.
[36] M. Scutari, "Learning bayesian networks with the bnlearn R package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010.
[37] D. Fixsen and R. P. S. Mahler, "The modified Dempster-Shafer approach to classification," *IEEE Trans. Syst. Man, Cybern. A: Syst., Humans*, vol. 27, no. 1, pp. 96–104, Jan 1997.



Jose Ignacio Aizpurua (M'17) is a Research Associate within the Institute for Energy and Environment at the University of Strathclyde, Glasgow, Scotland. He received his Eng., M.Sc., and Ph.D. degrees from Mondragon University (Basque Country, Spain) in 2010, 2012, and 2015 respectively. He was a visiting researcher in the Dependable Systems Research group at the University of Hull (UK) in 2014. His research interests include prognostics and health management, reliability, availability, maintenance and safety (RAMS) analysis and systems engineering for power engineering applications.

Victoria M. Catterson (M'06-M'12) was a Senior Lecturer within the Institute for Energy and Environment at the University of Strathclyde, Scotland, UK. She received her B.Eng. (Hons) and Ph.D. degrees from the University of Strathclyde in 2003 and 2007 respectively. Her research interests include condition monitoring, diagnostics, and prognostics for power engineering applications

Brian G. Stewart (M'08) is is Professor within the Institute of Energy and Environment at the University of Strathclyde, Glasgow, Scotland. He graduated with a BSc (Hons) and PhD from the University of Glasgow in 1981 and 1985 respectively. He also graduated with a BD (Hons) in 1994 from the University of Aberdeen, Scotland. His research interests are focused on high voltage engineering, electrical condition monitoring, insulation diagnostics and communication systems. He is currently an AdCom Member within the IEEE Dielectrics and Electrical Insulation Society.

Stephen D. J. McArthur (M'93-SM'07-F'15) received the B.Eng. (Hons.) and Ph.D. degrees from the University of Strathclyde, Glasgow, U.K., in 1992 and 1996, respectively. He is a Professor and co-Director of the Institute for Energy and Environment at the University of Strathclyde. His research interests include intelligent system applications in power engineering, covering condition monitoring, diagnostics and prognostics, active network management and wider smart grid applications.

Brandon Lambert is a Design Engineering Manager within Bruce Power. He received his B.Eng. degree from Lakehead University, Thunder Bay, Canada in 2012 and his P.Eng. from the Professional Engineers of Ontario in 2015. His design interests include large power transformers, high voltage transmission systems, as well as dielectric and insulating materials.

James Cross (M '79) is currently Director of Transformer Services at Kinectrics, In. in Toronto, Canada. After graduating from the University of Manitoba with a B.Sc. in Electrical Engineering, he worked for 18 years at Carte International, a transformer manufacturer in Winnipeg, Canada as Vice-President, Technology. He then worked as a Project Engineer at Pauwels Canada, a manufacturer of large power transformers up to 500 kV class. Most recently, he worked for 18 years at Weidmann Electrical Technology in St. Johnsbury, Vermont serving as Manager of R&D/Innovation and Manager of Technical Services. He has co-authored several papers in the area of electrical insulating materials and testing, and transformer diagnostics. He is a former Chairperson of the IEEE Winnipeg Section.