Using Ensemble Random Forests for the Extraction and Exploitation of Knowledge on Gas Turbine Blading Faults Identification

Manolis Maragoudakis and Euripides Loukis
University of the Aegean, Samos, Greece
{mmarag,eloukis}@aegean.gr

Abstract

The extraction and exploitation of existing knowledge assets for supporting decision making and increasing the effectiveness of various internal and external interventions is of critical importance for the success of modern organizations. The use of advanced Operational Research based quantitative methods in combination with high capabilities information systems can be very useful for this purpose. In this paper we are investigating the use of Ensemble Random Forests for extracting, codifying and exploiting existing organizational knowledge on gas turbine blading faults identification, in the form of a large number of decision trees (called a 'forest'); each of them has internal nodes corresponding to various tests on features of signals acquired from the gas turbine and leaf nodes corresponding to classifications to the healthy condition or particular faults. Two heterogeneous kinds of inserting randomness to the development of these forest trees, based on different theoretical assumptions, have been examined (Random Input Forests and Random Combination Forests). Using data from a large power gas turbine the performance of Ensemble Random Forests has been evaluated, and also compared against other machine learning classification methods, such as Neural Networks, Classification and Regression Trees and K-Nearest Neighbor. The Ensemble Random Forests reached a level of 97% in terms of precision and recall in engine condition diagnosis from new signals acquired from the gas turbine, which was higher than the performance of all the other examined classification methods. These results provide some first evidence that Ensemble Random Forest can be an effective tool for the extraction, codification and exploitation of the technological knowledge assets of modern organizations, and contribute significantly to the improvement of organizations' decision making and interventions in this area.

1. Introduction

It is increasingly recognized that the effective management of the knowledge assets of modern organizations and their exploitation to the highest possible extent for supporting decision making and increasing the effectiveness of their internal and external interventions is of critical importance for achieving high organizational performance [1]. Organizations today possess extensive and valuable knowledge assets, which are however in forms that do not allow their full exploitation (e.g. tacit knowledge in the minds of employees, or buried in large operational datafiles). In order to exploit them it is necessary to address successfully four main challenges: i) retrieve and store this knowledge in an appropriate and directly usable form, ii) make it accessible to the employees who need it, iii) incorporate it into organizational processes and activities, and iv) use it for supporting decision making and for planning internal and external interventions. The use of advanced Operational Research (OR) based quantitative methods in combination with high capabilities information systems (IS) can be very useful for this purpose ([1], [2]), since it allows transforming this knowledge from its initial and not easily exploitable form into compact models directly meaningful to and usable by employees, and making it accessible to them through appropriate IS and networks.

One of the areas in which this approach can be applied is definitely the management and maintenance of complex equipment. Organizations today are increasingly using various types of highly complex equipment for increasing their productivity, and there is a growing reliance of their operations on such equipment; even short unscheduled losses of their availability can result in significant operational problems and economic costs. Also, their maintenance is becoming increasingly costly, due to their increasing structural and technological complexity. For these reasons there has been extensive interest and research for long time for the development of ICT-based methods and systems for 'Engine Condition Monitoring' (a good review of them is provided by [3]). The continuous monitoring of the health condition of complex equipment can offer significant benefits: avoidance of catastrophic failures, minimization of unscheduled availability losses, optimization of preventive maintenance based on the real condition

of the parts and components of the equipment (instead of performing it at predefined regular intervals, which are based on general recommendations of the manufacturer not taking into account the specific operational conditions of each particular engine), and also better planning of preventive maintenance interventions. These can result in improvements of equipment maintenance and management, and at the same time significant cost reductions. The development of ICT-based data acquisition systems was a significant driver for the development of Engine Condition Monitoring, as they allow the collection of data from various types of measurement instruments at several locations of an engine, and then their storage and processing, easily and at a low cost. The huge amount of data collected in this way (having usually the form of a large set of digitized signals from a number of measuring instruments at various time points, together with the corresponding engine condition and possibly the existing faults, as diagnosed by highly knowledgeable and experienced technical personnel) constitutes a valuable knowledge asset. This knowledge can be quite useful for exploiting better the new data to be collected in order to assess the health condition of the engine and identify faults from their early stages with higher levels of reliability; however, it is in a form that does not allow its full exploitation.

Especially the airline and power generation industries were among the first adopters of the above ideas for the condition monitoring and faults identification of the large gas turbines they are heavily using, being highly reliant on them. The development of effective gas turbine condition monitoring and fault diagnosis methods has been the target of considerable research for long time, due to the high acquisition and maintenance cost, the complexity, the sensitivity and the importance for many industries (including the abovementioned ones) of this type of engines. Most of this research is directed towards the diagnosis of gas turbine blading faults, because of the catastrophic consequences that these faults can have, if they are not diagnosed in time; even very small blading faults can rapidly grow due to the high speeds of the rotating components of gas turbines (usually at the order of magnitude of thousands of rotations per minute) and result in huge destructions ([4], [5], [6]). Blading faults' diagnosis is regarded to be a very difficult problem, because of the high levels of noise in all relevant measurements and the high interaction among the neighboring gas turbine blading rows, and also with the other components. Therefore, it is of critical importance to take advantage of advanced OR-based methods in combination with the processing power and the advanced capabilities of modern IS in order to extract, codify and exploit effectively the existing organizational knowledge on gas turbine blading faults identification, for providing fast and reliable gas turbine blading faults' identification from available measurements and in general for developing the highest possible level of intelligence and assistance to the operations and maintenance personnel.

As described in more detail in the following section 2, there has been considerable previous research on the problem of gas turbine faults' identification, which has investigated various individual classifiers (whose learning phase exploits and codifies pre-existing relevant knowledge), and also - to much lesser extent - some forms of combination (referred to as 'fusion') of small numbers of individual classifiers. In the present paper we are investigating for the first time the combined use of large numbers of individual classifiers (in the order of hundredths) in order to exploit better the existing organizational knowledge on gas turbine blading faults identification and achieve higher levels of performance in this critical and at the same time difficult problem. Our study investigates the use of Ensemble Random Forests for extracting, codifying and exploiting this valuable knowledge. In particular:

- The initial form of this knowledge is a series of digitized signals from a number of measuring instruments at various time points, together with the corresponding engine condition (healthy or existence of a particular fault), as diagnosed by highly knowledgeable and experienced technical personnel.
- This knowledge is extracted and codified in the form of a large number of decision trees (called a 'Forest'); each of them has internal nodes corresponding to various criteria (tests) on features of signals acquired from the gas turbine (e.g. $F_n > v_n$) and leaf nodes corresponding to classifications to particular classes (e.g. C_A) corresponding to the healthy condition or particular faults (Figure 1 – providing a simplified illustration, since these threes usually have many levels), while it can also be expressed as a set of rules.

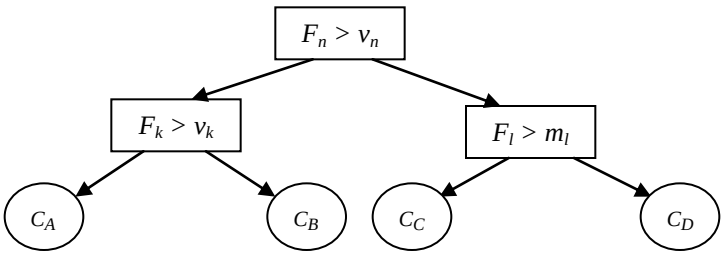

Figure 1. *Structure of a decision tree codifying organizational knowledge on gas turbine blading faults identification*

- This set of decision trees can be exploited for each new signal acquired from the gas turbine in order to assess from it the engine condition and possibly diagnose the existing fault. For this purpose initially the features are calculated for this signal, which are then for each tree subjected to its tests, starting from the top root node and proceeding downwards, leading finally to classification to one of the classes (i.e. to the healthy condition or to one of the faults); the final classification is determined through majority vote.

It should be noted that the above proposed approach is more computation-intensive than other previous approaches based on individual classifiers, or combinations of small numbers of individual classifiers, as it involves initially the construction of a large number of decision trees, and then the combined use of all of them for each new signal acquired from the gas turbine in order to produce a reliable diagnosis. Therefore its practical application relies on the use of high capabilities IS, which will provide the necessary infrastructure for performing and integrating all the above activities quickly and at a low cost: acquisition and digitization of signals from various measuring instruments, storage of them, batch processing for the extraction of the knowledge they contain on gas turbine blading faults identification and the construction of this set of decision trees, and finally online processing of each new signal acquired exploiting this codified knowledge (calculation of features and determination of class).

Two heterogeneous kinds of inserting randomness to the forest trees, based on different theoretical assumptions, have been examined (Random Input Forests and Random Combination Forests). Using data from a large power gas turbine the performance of Ensemble Random Forests in engine condition diagnosis from new signals has been evaluated, and also compared against other machine learning classification methods, such as Neural Networks, Classification and Regression Trees (CART) and K-Nearest Neighbor.

The paper is organized in 6 sections. In the following Section 2 previous relevant literature is briefly reviewed, while in Section 3 the implementation and the algorithm of the proposed Ensemble Random Forests approach is described. In Section 4 the application on which the proposed approach has been validated is described. Then in Section 5, the results are presented and discussed, followed by concluding remarks in the final Section 6.

2. Literature Review

Considerable research has been conducted on gas turbine faults' identification from various types of measurements, both static (e.g. pressure, temperature) and dynamic (e.g. vibrations). Various individual classifiers have been investigated for this purpose, such as Pattern Recognition ([7], [8], [9]), Expert Systems ([10], [11]), Fuzzy Logic ([12] - [15]), Bayesian Networks ([16]) and Neural Networks ([17] – [19]); these classifiers during their learning phase incorporate to various extents pre-existing knowledge concerning gas turbine faults identification.

One research stream in this area investigates the use of Pattern Recognition techniques for the above purpose. Loukis et al. [7] developed a method for automated diagnosis of gas turbine compressor blade faults from dynamic measurement data (casing vibrations, unsteady pressure inside the casing, sound), based on the principles of statistical pattern recognition; also, they propose a method for formulating the optimal discriminants that can be calculated from the available data, which maximizes the discrimination between condition classes. Similarly, Aretakis and Mathioudakis [8] use pattern-recognition techniques for the identification of various faults (inlet obstruction, obstruction in a diffuser passage, variation of impeller tip clearance and impeller fouling) in a radial compressor from casing vibration and sound emission. The problem of identification of sensor faults in turbofan engines is addressed by Aretakis et al. [9] using three different alternative pattern recognition techniques (geometric, statistical and statistical using optimal directions), in combination with an adaptive performance analysis algorithm, which calculates a set of component performance modification factors.

Another research stream in this area is dealing with the exploitation of Expert Systems and Fuzzy Logic for the identification of various types of gas turbine malfunctions. Breese et al [10] developed an expert system for the diagnosis of efficiency problems in large gas turbines. The system relies on a model-based approach which combines experts' probabilistic assessments with statistical data and thermodynamic analysis; it employs a causal probabilistic graph in order to update the probabilities of alternative faults given information. In the same direction DePold and Gass [11] propose the development of a new generation of diagnostic systems for gas turbines, which use artificial intelligence methods in order to automate diagnosis to the highest possible extent and to improve its quality; this advanced systems should combine neural networks (for trend change detection and classification to diagnose performance change) with expert systems (to diagnose, provide alerts and to rank maintenance action recommendations). Siu et al [12] presents an expert system for the diagnosis of vibration problems in turbomachinery, which is based on incremental forward chaining, and employs both fuzzy logic based approximate reasoning and traditional certainty factor techniques to deal with uncertainty; also, a simple case-based reasoning component was incorporated into the system to provide more accurate diagnoses when similar past experience can be applied. In the same direction Ganguli et al [13] developed a fuzzy system for gas path performance diagnostics,

which can automatically develop its rule base using a linearized performance model of the gas turbine; the measurements used are deviations in exhaust gas temperature, low rotor speed, high rotor speed and fuel flow from a base line 'good engine'; this fuzzy system in order to provide more reliable diagnosis it is combined with a genetic algorithm (used for tuning the fuzzy sets), and a radial basis neural network (used for measurements' pre-processing and noise reduction) (a more detailed description of it is provided by Verma et al [14]). Ogaji et al [15] propose a method of setting up an 'intelligent' fuzzy-logic process for the diagnosis of degradations of single engine-component in military turbofan engines, using gas-path measurements; the fuzzy rules have been produced by running an engine performance model for various degraded conditions.

A third research stream investigates the use of various techniques from the area Data Mining for faults identification in gas turbines. Romessis and Mathioudakis [16] present a method for diagnosis of performance problems in jet engine gas turbines based on a probabilistic approach through the use of a bayesian belief network, which has been built using information provided by an engine performance model (so that there is no need of acquiring flight data of different faulty operations of the engine). Angelakis et al [17] investigate the use of various neural network architectures, such as multi-layer perceptron (MLP), learning vector quantization (LVQ), modular multi-layer perceptron and radial basis function (RBF), for gas turbine blading faults diagnosis from dynamic measurement data (casing vibrations, unsteady pressure inside the casing, sound), coming to positive results. In the same direction Romesis et al [18] focus on the probabilistic neural networks, and study the effect of several parameters related to their structure and training, the noise level of measurements, the operating conditions and the severity of fault on the diagnostic performance for turbofan engines. Joly et al [19] propose a more complex diagnostic structure consisting of three layers of neural networks for the identification of deteriorations in an aircraft turbofan engine; the top level distinguishes between single-component and double component faults, while the middle level identifies particular components, or component pairs, which are faulty, and the final bottom level determines the extent of deterioration.

Palme at al [20] present a method for evaluating gas turbine sensor accuracy, based on training neural networks as classifiers to recognize sensor drifts, and evaluate it on two types of gas turbines (one single-shaft and one twin-shaft machine) with positive results.

Although satisfactory diagnostic performance have been achieved using the above individual classifiers, there has been some research investigating the use of various forms of combination of small numbers of individual classifiers (usually 2 or 3), which is called 'fusion', for achieving higher diagnostic performance. Volponi et al [21] developed an information fusion system for fault diagnostics and health management of an aircraft engine, which combines various types of data (e.g. gas path measurements, vibration signals, oil debris analysis data). Dewallef et al [22] examined the combination of a bayesian belief network with a soft-constrained Kalman filter for aircraft engine fault diagnosis using deviations of gas path data. In this study the Kalman filter uses a priori information derived by a bayesian belief network at each time step, in order to derive estimations of the unknown health parameters. The resulting algorithm has improved identification capability in comparison to both the stand-alone Kalman filter and the bayesian belief network. Kyriazis and Mathioudakis [23] proposed a two-step information fusion technique allowing the combination of both dynamic and static measurements for improving performance of gas turbines faults diagnosis. Each type of available data is fed to an independent probabilistic neural network, which produces as output a health condition assessment (in the form of a probability distribution). These outputs are then entered in the first step of the fusion technique and aggregated in order to derive a probability consensus; finally in the second step this probability consensus is classified (to a particular fault) using fuzzy set theory and fuzzy logic, which constitutes final diagnostic decision.

In summary, from the above literature review it has been concluded that for the problem of gas turbines fault identification several individual classifiers have been investigated, resulting in satisfactory levels of diagnostic performance. All these classifiers have a learning phase which exploits to some extent (varying among different classifiers) and codifies pre-existing relevant knowledge. Also, to much lesser extent have been investigated some forms of combination ('fusion') of small numbers of individual classifiers, with each of them providing an independent classification (usually as a probability distribution), and the final diagnostic decision being produced through an aggregation of these individual classifications. Taking into account that each individual classifier performs some exploitation of pre-existing relevant knowledge, this combination results in a better exploitation of this knowledge. However, the combined use of larger numbers of individual classifiers (e.g. in the order of hundredths), such as Random Forests (described in more detail in the following section 3), which might result in an even better exploitation pre-existing relevant knowledge for achieving even higher levels of performance, has not been investigated. This paper contributes in filling this research gap.

3. Random Forests

Random Forests are founded on Breiman's Classification and Regression Trees and Bagging Predictors [24], [25]), which have been successful in several classification and regression tasks. Based on them there have been many research efforts for constructing various kinds of ensemble classifiers towards increasing the performance of the classification task ([26], [27], [28]). Several of them have exhibited promising results in various classification problems, such as Adaboost, Bagging and Random Forests [29]. Random Forests are a mixture of robust decision tree classifiers, such that each tree is constructed based on a set of records, each of them including values of the features and a class characterization (learning data), and has the form shown in Figure 1; also, for the construction of each tree are taken into account the values of a random vector sampled independently (e.g. see the ways of introducing randomness described below in 3.1 and 3.2) and with the same distribution for all trees in the forest. The generalization error of such a set of decision tree classifiers (known as 'forest') depends on the strength of the individual trees within the forest and their inter-correlation. The use of a random choice of features in order to split each node yields output error rates comparable to the ones of the Adaboost, which is its main ensemble rival. While previous generation tree algorithms devote significant resources for choosing how to split at a node (i.e. which feature will be used for this person), Random Forests algorithm perform this task with little computational effort. Compared with Adaboost, Random Forests have the following advantages:

- the accuracy is as good as Adaboost and sometimes better.
- they are relatively robust to outliers and noise.
- they are faster.
- they provide useful internal estimates of error, strength, correlation and variable importance.
- they are simple and easily parallelized.

In particular, a Random Forest classifier $\Theta(x)$ consists of a number of decision trees, where x is an input instance, with each tree grown using some form of randomization [30]. In particular, the process of building each decision tree consists of the following steps:

I. If the number of instances of the training set is N, sample N cases at random, but with replacement, from the original data; this sample will be the training set for building the tree.
II. If there are M input features, a number m<<M is specified, such that at each node m variables are selected at random out of the M, and the best split on these m features is used to split the node; the value of m is held constant during the forest growing.
III. Each tree is grown to the largest extent possible, so no pruning is applied.

The leaf nodes of each tree are labeled by class estimates (= the posterior distribution over the data class labels). Each internal node contains a criterion (test) that best splits the space of data to be classified. A new instance is classified by propagating it down every tree and aggregating the reached leaf distributions (e.g. through majority vore). The classification process for a new instance is shown in Figure 2.

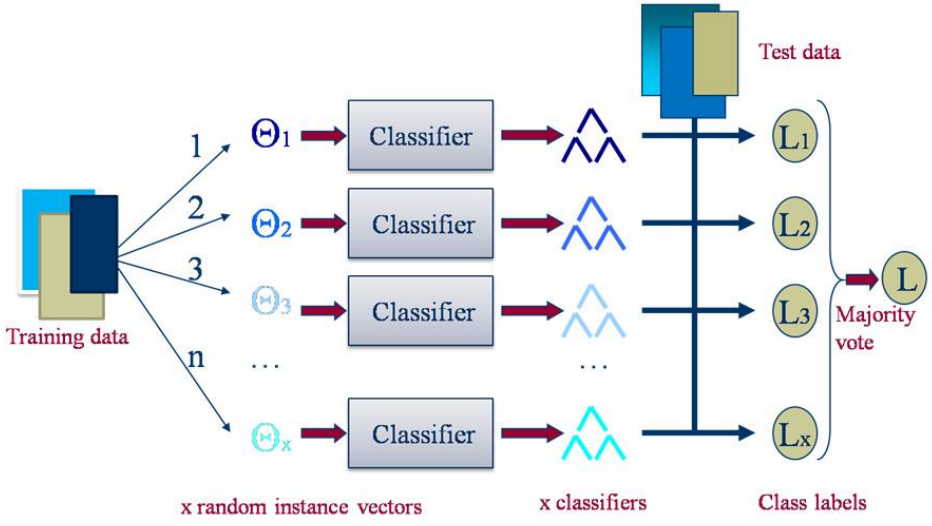

Figure 2. *Hierarchical decomposition of a Random Forests classifier on a data set*

The overall error rate of the Random Forests is influenced by two factors:

- *Robustness* (strength) of each individual tree within the forest: higher strength results in lower error rates.
- Tree *inter-correlation*: highly correlated trees result in high error rate.

In order to make the classification process more formal, suppose that the joint classifier $\Theta(x)$ contains x individual classifiers $\Theta_1(x), \Theta_2(x),...,\Theta_x(x)$. Let us also assume that each data instance is a pair (x,y), where x denotes the input attributes (i.e. feature values), taken from a set A_i, $i=1,...,M$, and y symbolizes the set of class labels L_j, $j=1,...,c$ (c is the number of class values). For reasons of simplicity, the correct class will be denoted as y, without any indices. Each discrete attribute A_i takes values from a set V_i, $i=1$ *to* m_i (m_i is the number of values attribute A_i can take). Finally, the probability that an attribute A_i has value v_k is denoted by $p(v_i,k)$, the probability of a class value y_j is denoted by $p(y_j)$ and the probability of an instance with attribute A_i having value v_k and class label y_j is symbolized by $p(y_j|v_i,k)$.

Each of the N instances (records) of training set can be picked at random with replacement. By this procedure, called "bootstrap replication", it is mathematically proven [11]?? that a subset of 36.6% of the training examples will not be taken into account during the tree construction phase. These out-of-bag (oob) instances allow for computing the degree of strength and correlation of the forest structure; therefore, no k-fold cross validation techniques are needed in Random Forests. Suppose that $O_k(x)$ is the set of oob instances of classifier $\Theta_k(x)$. Furthermore, let $Q(x, y_j)$ denote the subset of oob samples which were voted to have class y_j at input example x. An estimate of $p(\Theta(x) = y_j)$ is given by the following equation μήπως θα έπρεπε να είναι $p(\Theta(x) = y_j)$ και όχι $Q(x, y_j)$ το αριστερό μέλος του παρακάτω:

$$Q(x, y_j) = \frac{\sum_{k=1}^{K} I(\Theta_k(x) = y_j; (x, y) \in O_k)}{\sum_{k=1}^{K} I(\Theta_k(x); (x, y) \in O_k)}$$

where $I(\cdot)$ is the indicator function?? – ΕΞΗΓΗΣΕ ΚΑΛΥΤΕΡΑ ΤΗΝ ΠΑΡΑΠΑΝΩ ΣΧΕΣΗ

The margin function which measures the extent to which the average vote for the right class y exceeds the average vote for any other class labels is computed by:

$$\text{margin}(x, y) = P(\Theta(x) = y) - \max^{c}_{j=1, j\neq y}(P(\Theta(x) = y_j)$$

Since strength is defined as the expected margin, it is computed as the average over the training set: ΕΞΗΓΗΣΕ ΚΑΛΥΤΕΡΑ ΤΗΝ ΠΑΡΑΚΑΤΩ ΣΧΕΣΗ–ΜΗΠΙΩΣ ΘΑ ΕΠΡΕΠΕ ΝΑ ΕΧΕΙ ΠΙΘΑΝΟΤΗΤΕΣ P ΚΑΙ ΟΧΙ Q ΣΥΝΟΛΑ

$$s = \frac{1}{n}\sum_{i=1}^{n}(Q(x_i, y) - \max^{c}_{j=1, j\neq y} Q(x_i, y_j))$$

The average correlation is given by the variance of the margin over the square of the standard deviation of the forest: $\bar{p} = \frac{Var(\text{margin})}{\sigma(\Theta())^2}$, is estimated for every input example x in the training set Q(x, y_j).

In accordance to Breiman's suggestion, we examined the Random Input Forests and the Random Combination Forests randomness algorithms; for both of them the evaluation metric on which tree nodes are chosen to split is the Gini index, taken from the CART algorithm. Other similar metrics presented by researchers are Gain ratio [31], MDL [32] and Relief-F [33]. The formula for calculating the Gini index is as follows [22]:

$$Gini(A_i) = -\sum_{i=1}^{c} p(y_i)^2 - \sum_{j=1}^{m_i} p(v_{i,j}) - \sum_{j=1}^{c} p(y_i / v_{i,j})^2$$

ΕΞΗΓΗΣΕ ΚΑΛΥΤΕΡΑ ΤΗΝ ΠΑΡΑΠΑΝΩ ΣΧΕΣΗ

3.1. Random Input Forests

The simplest kind of Random Forests is formed by selecting at each node at random a small set of input variables (features) to split on. Each tree is grown using CART methodology to maximum size and is not pruned. The steps we follow are shown below in a form of pseudo-code:

For building K trees:

Build each tree by:

- Selecting, at random, at each node a small set of features (F) to split on (given M features in total). Common values of F are:
  F=1.
  F=$\log_2$(M) + 1.

- For each node split on the best of this subset (using oob instances) that maximizes the Gini index.
- Grow tree to full length.

3.2. Random Combination Forests

This alternative approach consists of defining more features by taking random linear combinations of a number of the input variables. That is, a feature is generated by specifying L, the number of variables to be combined. At a given node, L variables are randomly selected and added together with coefficients that are uniform random numbers on [-1,1]. In this way F linear combinations are generated, and then a search is made over them for the best split that maximizes the Gini index. The steps we follow are shown below in a form of pseudo-code:

For building K trees:

Build each tree by:

- Create F random linear sums of L variables:

$$A_f = \sum_{i=1}^{L} b_{fj} x_i \text{ , where } b_{fj} = \text{uniform random number in [-1,+1]}$$

- At each node split on the best of these linear boundaries.
- Grow tree to full length

4. Application

4.1. Data

The present study is based on data acquired from dynamic measurements on an industrial power production gas turbine into which different faults were artificially generated in blades of the first stage of the compressor. Four distinct categories of dynamic measurements were performed simultaneously:

1. Unsteady internal wall pressure (using fast response pressure transducers P2 to P5 located at positions facing the first four rotors of the compressor in which the artificial blade faults were generated).
2. Casing vibration (using accelerometers A1 to A6 mounted on the outside surface of compressor casing at locations near the above pressure transducers).
3. Shaft displacement at compressor bearings (using transducer B).
4. Sound pressure levels (using double layer microphone M facing the casing at a position corresponding to the first compressor stage).

A schematic illustration of the gas turbine setting, showing the measuring instruments' arrangement, is provided in Figure 3.

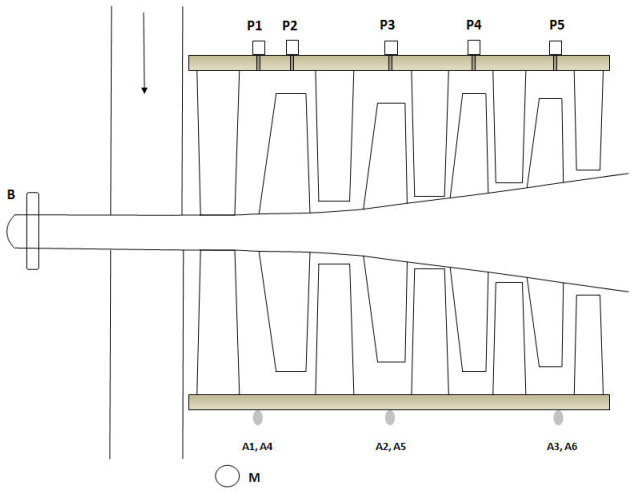

Figure 3. *Arrangement of the measuring instruments (accelerometers A1 and A4 are at the same position horizontally, with the latter being rotated by 90 degrees; similar hold for A2 and A5, and also for A3 with A6).*

Based on the above setting, five experiments were carried out, the first in a healthy condition and the other four in operation with the following four typical small and therefore not easily diagnosable blade faults (which however are rapidly growing due to the high speeds of the rotating components, so they can result in catastrophic failures) in the first stage of the compressor (which is the most vulnerable to blading faults):

- Fault 1: rotor fouling.
- Fault 2: individual rotor blade fouling.
- Fault 3: individual rotor blade twisting (by approximately 8 degrees).
- Fault 4: stator blade re-staggering (change of its angle).

Trials were performed at four different engine loads (full load, half load, quarter load and no load) both for the healthy operation and for the operation with above four faults. At each engine load four series of time domain data were acquired for each instrument (two series at a sampling frequency *l = 13 kHz* and two series at a higher sampling frequency of *m = 32 kHz*); in the healthy operation only one data series was acquired for each sampling frequency. Consequently, for every measuring instrument we have 72 different series of time domain data: 8 healthy data series (=4 loads x 2 frequencies) and 64 faulty data series (=4 faults x 4 loads x 2 frequencies x 2 series). By performing fourier analysis in each of these time series data we remarked that the main components are at the rotor's shaft rotational frequency and its harmonics, while in all the other frequencies it is noise. For this reason in all these signals we focused on the rotor's shaft rotational frequency harmonics, and in particular on the first 27 harmonics which are strong enough so that there are not buried in the noise. Therefore our data had a tabular form, consisting for each measuring instrument of 72 instances described by 27 attributes.

4.2. Calculation of Features

In order to calculate the features for each time series (signal) its 'spectral difference pattern' was calculated using the following equation:

$$P(f) = 20[\log(sp(f)) - \log(sph(f))]$$

where *P(f)* is the spectral difference pattern, which is a function of frequency *f*, *sp(f)* is the power spectrum of the signal, and *sph(f)* is the spectrum signal of the 'corresponding' healthy signal, coming from healthy operation at the same load and sampling frequency. Furthermore, since as mentioned above the most valuable diagnostic information is contained at the harmonics of the shaft rotational frequency, were filtered out the values of *P(f)* at frequencies other than the shaft rotational frequency harmonics. The resulting pattern from this filtering, *Pr(f)*, is referred to as "reduced spectral difference pattern" (and for simplicity "pattern" in the following text), and is calculated by the following equation:

$$\Pr(f) = P(f) * H(f)$$

where *H(f)=1* if f is a rotational harmonic, and *H(f)=0* for all other frequencies. An example of the patterns calculation procedure described above is shown in the following Figure 4 for the unsteady pressure transducer P2.

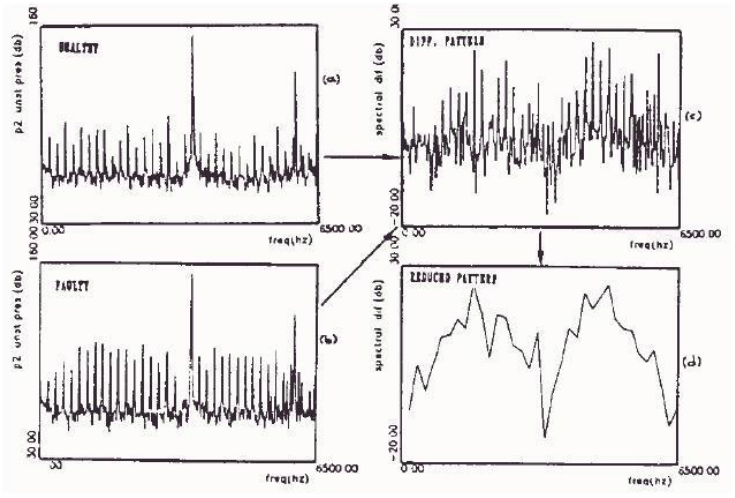

Figure 4. *Pattern calculation procedure for power spectra of unsteady pressure transducer P2.*

4.3. Preprocessing

Due to the fact that a plethora of sensors has been used, feature selection and outlier identification issues had to be

addressed, in order for the classification model to be robust and effective. In order to perform feature selection, we estimated the importance of each variable using the Information Gain criterion and arranged them in descending order. The Information Gain metric for a particular feature F is based on a statistical view of modeling uncertainty and is expressed as the difference of total Entropy (i.e. of the whole dataset) minus the sum of the Entropies of the subsets corresponding to feature's values. In particular, the Information Gain (IG) of a feature F with V different values within a dataset D of n_D instances is given by:

$$IG(F,D) = Entropy(D) - \sum_{V \in F} \frac{n_V}{n_D} Entropy(D_V)$$

The feature importance graph?? provided important information regarding the significance of each variable to the classification process; these results were verified through correlation tests.

Finally, noisy data were removed from the dataset using the CURE [29] (Clustering Using Representatives) clustering algorithm. Even though Random Forests are relative robust to outliers and noise, in order to achieve higher levels of performance we decided to perform a pre-processing noise removal process. The CURE algorithm consists of a hierarchical and a partitioning part and operates as follows: initially, a constant number of representative points *c*, is selected from each cluster (in our case as clusters we consider our 5 classes: the healthy signals, signals from fault 1, signals from fault 2, signals from fault 3, signals from fault 4). These well-partitioned data are shrunk to approach the cluster centroid, by applying a shrink factor α (when α equals 1, all the points meet to a single point, the centroid). These multiple points are better representatives of a cluster than traditional methods which make use of only a single point per cluster. Furthermore, using several points per cluster could result in clusters that are not necessarily spherical shaped offering a better representation of them. As mentioned before, CURE also used a hierarchical phase where clusters with nearest pairs of representative points are merged to form a single cluster. Figure 5 depicts the basic idea of CURE in four distinct phases. Firstly, from a given dataset (a) clusters along with representative points are being formed (b) (these points are selected in order to be far from each other and far from the middle of each cluster); then (c) two of the clusters are being merged and two new representative points are chosen; finally, in phase (d) these points shrunk towards to middle of this cluster. Note that if a single point was used for each cluster as a representative, the small cluster would have been merged with the lowest cluster rather than the upper one.

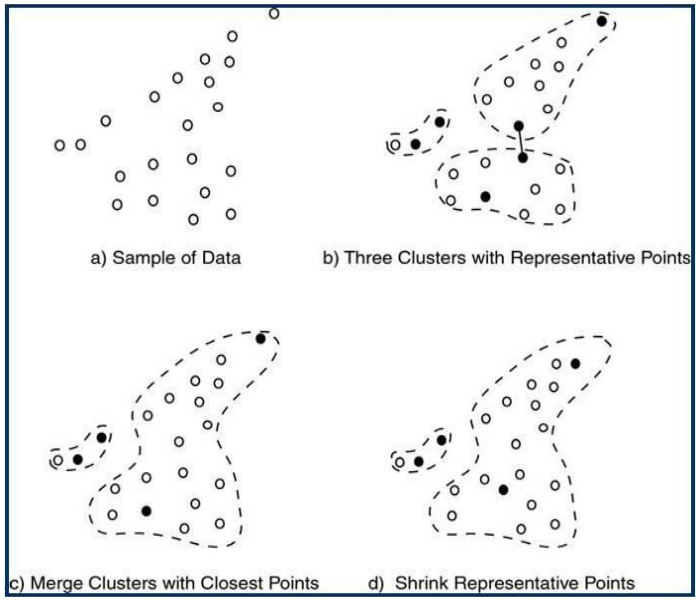

Figure 5. *The CURE algorithm for outliers removal*

CURE algorithm used memory resources efficiently since the initial cluster assignment is performed based on a random data sample. This dataset is partitioned and each part is being clustered. The resulting clusters are fully clustered in a second pass. Both sub-sampling and partitioning are performed isolated, in order to ensure that all data are fit to the memory. The time complexity of this algorithms is $O(n^2 logn)$ [30] while space complexity is only O(n), in the worst case. From a technical point of view spatial accessing methods such as R+-trees [31] or R*-trees [32]

are used. Furthermore, a stack is kept, where the number of representative points of each cluster, along with the centroid and the closest cluster are stored. CURE uses the stack to find the closest candidate clusters for the merging process [33]. The complete algorithm is shown in more detail in Table I.

TABLE I. CURE ALGORITHM

```
Input: D=(x1,x2,…,xn); //Dataset
K; //Number of Clusters
Output : S=(<x1,Ki>,<x2,Kj>,…,<xN,Km>),
i,j,m ∈ K // Heap containing each example assigned to a cluster
MAIN BODY:

T=build_k_D_tree(D);
Q=create_heap(D;//initially one entry per item
while nodes(Q) != K
{
    u=min(Q); //finds the smallest heap item
    delete(Q,u.closest);
    w=merge(u,v);
    delete(T,v); //delete from heap
    insert(T,w); //insert from heap
for each x∈Q do
x.close=find_nearest_cluster(x);
if x is_closest_to(w) then
w.close(x);
insert(Q,w);
}
```

5. Results

The abovementioned two variations of Random Forests, Random Input (RI) Forests (described in 3.1) and Random Combination (RC) Forests (described in 3.2), were applied on the above dataset, using oob estimates. As performance evaluation metric we considered per-class Precision and Recall, which are common metrics used to validate the classification performance in the information retrieval domain. In particular, as Precision for a class is defined the percentage of correctly classified in it instances among those that the algorithm classifies in this class. As Recall for a class is then defined the fraction of correctly classified in it instances among all instances that actually belong to this class. These definitions are illustrated in the following Table II showing a "confusion matrix" which tabulates for a binary classification problem (classes **A** and **C**) the actual class distribution (vertically) against the assigned class distribution (the columns of the table). From these definitions it is clear that both these metrics assess the effectiveness of the proposed Random Forest approach in extracting, codifying and exploiting the knowledge on gas turbine blading faults identification contained in the maintenance files of the organization (in the form of digitized signals from a number of measuring instruments at various time points, together with the corresponding engine condition as diagnosed by highly knowledgeable and experienced technical personnel).

TABLE II. CONFUSION MATRIX AND RECALL AND PRECISION METRICS FOR EACH CLASS (A AND C).

		Assigned Class	
		A	**C**
Actual Class	**A**	a	b
	C	c	d

precision A= $\frac{a}{a+c}$	recall A= $\frac{a}{a+b}$	precision C= $\frac{d}{b+d}$	recall C= $\frac{d}{c+d}$

Additionally we compared the classification performance of the above two Random Forests algorithms against three other widely used classification algorithms: the Multilayer Perceptron Neural Networks, the Classification and Regression Trees (CART) and the K-Nearest Neighbor (for which cross-validation was performed).

The preprocessing phase of eliminating noisy instances using the abovementioned CURE clustering algorithm did not change the set of the initial 72 instances for all measuring instruments, as they were all found to be close to each cluster's centroid (note that each class was considered to form a separate cluster). With regard to the Random

Forests, the best results were obtained by using 450 trees and 7 input features. The results - Precision and Recall per class for each classifier - are shown in Figures 6 and 7 (F1 to F4 denote the 4 faults' categories and OK denotes the healthy one). We remark that both Random Forests variations have a good classification performance, which is at an average level of 97.5% for both Precision and Recall; this is quite satisfactory, taking into account that all four faults were small, so the proposed Random Forest approach can diagnose these faults from their very early stages. Also, they both outperform the other three examined alternative classification algorithms for all classes. We can see that RC is slightly better that RI, however the difference is very small, so we can conclude that they have similar performances. These results indicate that the proposed Random Forest approach has a good potential in extracting, codifying and exploiting the knowledge on gas turbine blading faults identification contained in the maintenance files of the organization, exceeding clearly the potential of the other three widely used alternative classifiers.

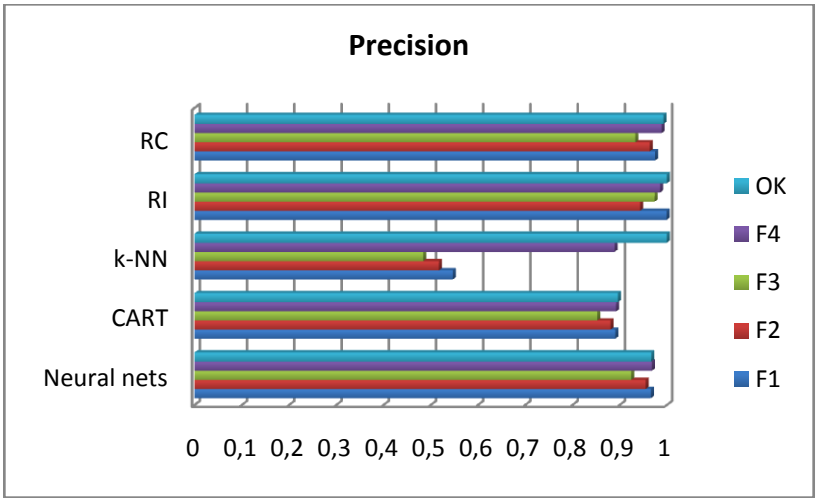

Figure 6. *Precision per class for all classifiers using data from all measuring instruments.*

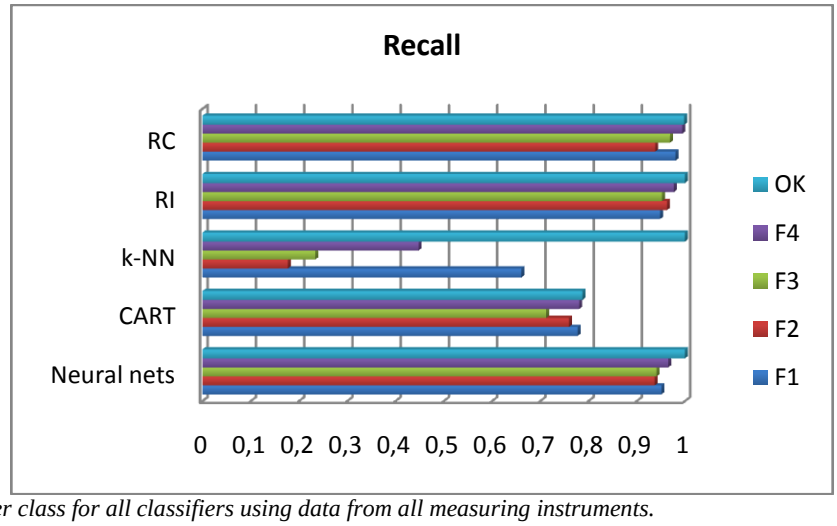

Figure 7. *Recall per class for all classifiers using data from all measuring instruments.*

Furthermore, we examined the classification performance achieved if only a subset of these measuring instruments is used. We have focused on the signals only of the 6 Accelerometers (A1-A6), which are located near the examined faults, and at the same time are much more convenient to be installed than the pressure transducers (P2-P5), which necessitate the laborious task of drilling holes in the compressor casing. The following Figures 8 and 9 show the precision and the recall achieved per class for each classifier from this reduced dataset.

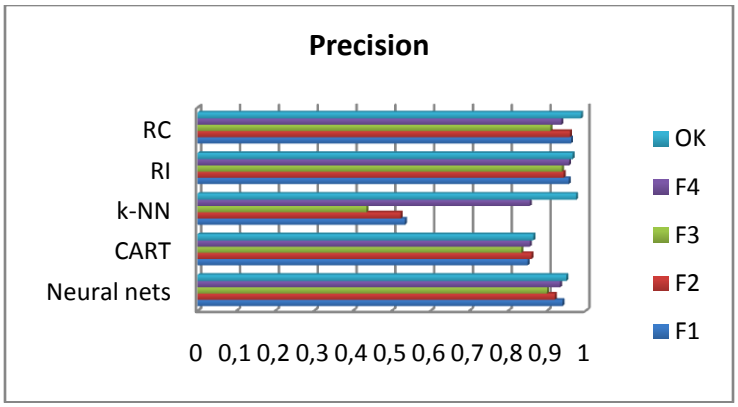

Figure 8. *Precision per class for all classifiers using data from accelerometers A1-A6.*

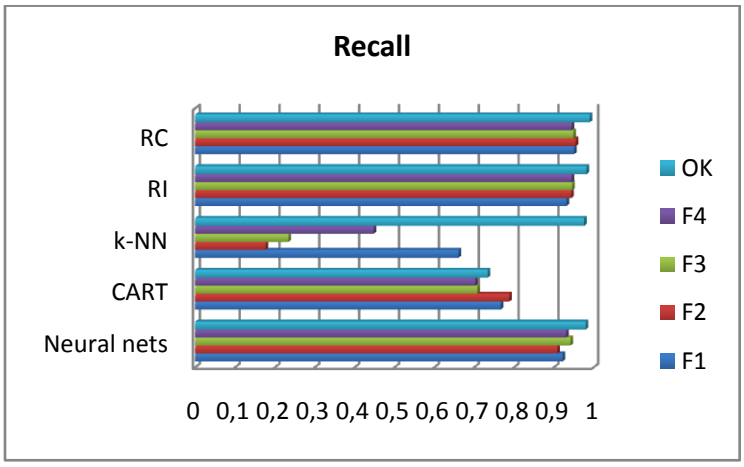

Figure 9. *Recall per class for all classifiers using data from accelerometers A1-A6.*

We can see that the results are similar to the previous ones. Again Random Forests exhibit higher precision and recall outcomes than the three examined alternative classifiers, with RC appearing to be the best performing among all classifiers. However, the use of only this type of instruments (accelerometers) results in a decrease of the performance of all classifiers in both metrics, which ranges from 2% to 3.5%, depending on the classifier. This reduction is of course expected since less information (and therefore less knowledge) is given to the inference models of each algorithm, however it is counterbalanced by the fact that accelerometers are relatively easy to be acquired and installed.

6. Conclusion

In the previous sections we have presented an Ensemble Random Forests based method for the extraction and exploitation of existing knowledge in organizations, concerning a difficult and at the same time critical problem: the identification of a very dangerous kind of faults (blading faults) in an important category of equipment (gas turbines), which are of critical importance in many industries, such as the airline and power generation industries. The proposed method uses the existing knowledge on blading faults' identification in gas turbine user organizations; this knowledge has the form of several instruments' digitized signals at many different points in time (acquired through the increasingly adopted data acquisition systems) and also the corresponding health condition of the engine (healthy or existence of a particular fault). This knowledge is extracted and codified in the form of a highly sophisticated model: a large number of decision trees (i.e. a Random Forest). Each decision tree has internal nodes corresponding to various criteria (tests) on features of signals acquired from the gas turbine (e.g. Fn > vn) and leaf nodes corresponding to classifications to particular classes (e.g. C_A) corresponding to the healthy condition or particular faults, while it can also be expressed as a set of rules. This model can be accessible through appropriate IS and networks and exploited by by the operations and maintenance personnel, who can use it for newly acquired data from the gas turbine in order to diagnose its current health condition.

Our results indicated that the proposed approach (for both examined methods of injecting randomness to the decision trees of the forest) shows a very good performance in gas turbine blade faults identification, and outperforms all the three examined widely used alternative classification approaches in terms of precision and recall. In particular, both Random Input Forests and Random Combination Forests appeared to achieve higher classification performance than Neural Networks, Classification and Regression Trees and k-Nearest Neighbor classifiers. Therefore this proposed combination of large number of decision tree classifiers, which is for the first time investigated for this critical and difficult problem (as previous research on it was focusing mainly on individual classifiers, and to a lower extent on fusion of 2-3 individual classifiers, as outlined in section 2), seems to be a highly effective mechanism of extracting, codifying and exploiting knowledge on gas turbines faults' identification, outperforming in this respect the other three widely used alternative classifiers. For this reason it can increase the effectiveness of engine condition monitoring, and through it the effectiveness of complex equipment maintenance and management: having a reliable picture of the condition of our equipment allows us to make timely and appropriate maintenance interventions, avoid catastrophic failures, replace parts and components based on their real condition (and not at predefined regular intervals, based on general manufacturer's recommendations) and make more effective maintenance plans. These can contribute to reduction of the costs of maintenance, and at the same time improvements of its effectiveness.

It should be mentioned that the price we have to pay for this higher classification performance provided by the proposed approach is the higher computational effort required for the initial training of the decision trees forest; however, taking into account that for each node it is among a limited number of input features that we search for the one to be used for the split (which reduces the required computational effort), and also that this training takes place only once in the beginning in offline mode, we do not expect that this will be a problem for the practical application of the approach. However, due to the above characteristics and requirements of the proposed approach its practical application relies critically on the use of high capabilities IS, which will provide the necessary infrastructure for performing and integrating all its basic knowledge management stages: a) acquisition and digitization of signals from various measuring instruments (which contain valuable knowledge), b) organization and storage of them, c) batch processing of them for the extraction of the knowledge they contain on gas turbine blading faults identification and construction of a set (forest) of decision trees (codification of the knowledge), and finally d) online processing of each new signal acquired from the engine and classification of it in the appropriate health condition class (exploitation of this codified knowledge).

Closing, we believe that such ensemble approaches, like the one at hand, can be successfully applied in problems of knowledge extraction, codification and exploitation of other organizational functions as well (e.g. in sales, procurement, financial management, human resources management, etc). So further research is required for investigating their performance in other types of such problems, and making the required adaptations and improvements.

References

[1] J. S. Edwards, B. Ababneh, M. Hall and D. Shaw, Knowledge management: a review of the field and of OR's contribution, Journal of the Operational Research Society, 60, 114-125, 2009.

[2] J. H. Klein, N. A. D. Connell and E. Meyer, Knowledge management: a review of the field and of OR's contribution, Journal of the Operational Research Society, 58, 1535 – 1542, 2007.

[3] J. Campos, Development in the application of ICT in condition monitoring and maintenance, Computers in Industry, 60, 1, 2009, 1-20

[4] E. Loukis, P. Wetta, K. Mathioudakis, A. Papathanasiou, K. Papailiou, Combination of Different Unsteady Quantity Measurements for Gas Turbine Blade Fault Diagnosis, 36th ASME International Gas Turbine and Aeroengine Congress, Orlando, ASME Paper 91- GT-201, 1991.

[5] E. Loukis, Contribution to Gas Turbine Fault Diagnosis Using Methods of Fast Response Measurement Analysis, Doctoral Thesis, Athens, National Technical University of Athens, 1993.

[6] G. Merrington, O. K. Kwon, G. Godwin, B. Carlsson, Fault Detection and Diagnosis in Gas Turbines, ASME Journal of Engineering for Gas Turbines and Power, 113, 11-19, 1991.

[7] E. Loukis, K. Mathioudakis, Papailiou, K. D., Optimizing Automated Gas Turbine Fault Detection Using Statistical Pattern Recognition, Journal of Engineering for Gas Turbine and Power - ASME, 116(1), 165-171, 1994.

[8] N. Aretakis, K. Mathioudakis, Classification of Radial Compressor Faults Using Pattern-Recognition Techniques, Control Engineering Practice, 6, 1217-1223, 1998.

[9] N. Aretakis, K. Mathioudakis, A. Stamatis, Identification of sensor faults on turbofan engines using pattern recognition techniques, Control Engineering Practice, 12, 827-836, 2004.

[10] J.S. Breese, E.J. Horvitz, M.A. Peot, R. Gay, G. H. Quentin, Automated Decision-Analytic Diagnosis of Thermal Performance in Gas Turbines, ASME paper No. 92-GT-399, 1992.

[11] H. DePold, D. Gass, The Application of Expert Systems and Neural Networks to Gas Turbine Prognostics and Diagnostics, Journal of Engineering for Gas Turbines and Power, ASME, 121, pp. 607-612, 1999.

[12] C. Siu, Q. Shen, R. Milne, TMDOCTOR: A Fuzzy Rule- and Case- Based Expert System for Turbomachinery Diagnosis, Proceedings of IFAC Symposium - SAFEPROCESS '97, Vol. 1(1), 556-563, 1997.

[13] R. Ganguli, R. Verma, N. Roy, Soft Computing Application for Gas Path Fault Classification, ASME Paper No. GT-2004-53209, 2004.

[14] R. Verma, N. Roy, R. Ganguli, Gas turbine diagnostics using a soft computing approach, Applied Mathematics and Computation, 172, 1342-1363, 2006.

[15] S.O.T. Ogaji, L. Marinai, S. Sampath, R. Singh, S.D. Prober, Gas-turbine fault diagnostics: a fuzzy-logic approach, Applied Energy, 82, 81-89, 2005.

[16] C. Romessis, K. Mathioudakis, Bayesian Network Approach for Gas Path Fault Diagnosis, Journal of Engineering for Gas Turbines and Power, ASME, 128(1), 64-72, 2006.

[17] C. Angelakis, E. Loukis, A. Pouliezos, G.Stavrakakis, A Neural Network based Method for Gas Turbine Blading Fault Diagnosis, International Journal of Modelling and Simulation, 21(1), 51-60, 2001.

[18] C. Romessis, A. Stamatis, K. Mathioudakis, A Parametric Investigation of the Diagnostic Ability of Probabilistic Neural Networks on Turbofan Engines, ASME Paper No. 2001-GT-0011, 2001.

[19] R. B. Joly, S. O. T. Ogaji, R. Singh, S. D. Probert, Gas-turbine diagnostics using artificial neural-networks for a high bypass ratio military turbofan engine, Applied Energy, 78, 397-418, 2004.

[20] T. Palme, M. Fast, M. Thern, Gas turbine sensor validation through classification with artificial neural networks, Applied Energy, 2011 (article in press).

[21] A. Volponi, T. Brotherton, R. Luppold, D. L. Simon, Development of an Information Fusion System for Engine Diagnostics and Health Management, NASA TM-2004-212924, Febr. 2004.

[22] P. Dewallef, C. Romessis, O. Léonard, K. Mathioudakis, Combining Classification Techniques with Kalman Filters for Aircraft Engine Diagnostics, Journal of Engineering for Gas Turbines and Power, 128(2), 281 – 288, 2006.

[23] A. Kyriazis, K. Mathioudakis, Gas Turbine Fault Diagnosis Using Fuzzy-Based Decision Fusion, Journal of Propulsion and Power, 25(2), 335-343, 2009.

[24] L. Breiman, J. H. Friedman, R. A. Olshen, and Ch. J. Stone, Classification and regression trees. Wadsworth Inc., Belmont, California, 1984.

[25] L. Breiman. Bagging predictors. Machine Learning Journal, 26(2), 123-140, 1996.

[26] L. Breiman. Random forests, Machine Learning Journal, 45(1), 5-32, 2001.

[27] Y. Freund and R. E. Shapire, Experiments with a new boosting algorithm, in Lorenza Saitta, editor, Machine Learning: Proceedings of the Thirteenth International Conference (ICML96). Morgan Kaufmann, 1996.

[28] Y. Amit and D. Geman, Shape quantization and recognition with randomized trees, Neural Computation, 9, 1545-1588, 1997.

[29] T. K. Ho, The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), pp. 832-844. 1998.

[30] I. Kononenko, Estimating attributes: analysis and extensions of Relief, in Luc De Raedt and Francesco Bergadano (Editors), Machine Learning: ECML-94, 171-182, Springer Verlag, Berlin, 1994.

[31] J. Ross Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, 1993.

[32] I. Kononenko, On biases in estimating multi-valued attributes, in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI95), 1034-1040. Morgan Kaufmann, 1995.

[33] L.Breiman, Looking Inside the Black Box, Wald Lecture II, Department of Statistics, California University, 2002.