Fault Classification of a Centrifugal Pump in Normal and Noisy Environment with Artificial Neural Network and Support Vector Machine Enhanced by a Genetic Algorithm

Abtin Nourmohammadzadeh[(✉)] and Sven Hartmann

Department of Informatics, Clausthal University of Technology,
Clausthal-Zellerfeld, Germany
{abtin.nourmohammadzadeh,sven.hartmann}@tu.clausthal.de
https://www.in.tu-clausthal.de/

Abstract. Fault diagnosis and detection play a crucial role in every system for its safe operation and long life. Condition monitoring is an applicable and effective method of maintenance techniques in the fault diagnosis of rotating machinery. In this paper two outstanding heuristic classification approaches, namely Artificial Neural Network (ANN) and Support Vector Machine (SVM) with four different kernel functions are applied to classify the condition of a real centrifugal pump belonging to petroleum industry into five different faults through six features which are: flow, temperature, suction pressure, discharge pressure, velocity and vibration. To increase the power of our classifiers, they are trained and tuned by Genetic Algorithm (GA) which is an effective evolutionary optimisation method. The experiments are done once with normal data and another time with noisy data in order to examine how robust the approaches are. Finally, the classification results of ANN-GA, SVM-GA, pure ANN and SVM (without GA enhancements) along with other two practical classification algorithms, namely K-Nearest Neighbours (KNN) and Decisions Tree, are compared together in terms of different aspects.

Keywords: Artificial Neural Network (ANN) · Support Vector Machine (SVM) · Genetic Algorithm (GA) · Fault diagnosis · Centrifugal pump

1 Introduction

The role of centrifugal pumps is of great importance in many industries. Hence, condition monitoring of them is absolutely necessary to prevent early failure, production line breakdown and to improve plant safety, efficiency and reliability. Furthermore, pumps, compressors and piping are causes of the major equipment failure in oil and gas plants. Centrifugal pumps are sensitive to: (1) variation in liquid condition (i.e. viscosity, specific gravity, and temperature), (2) Suction variation, such as pressure and availability of a continuous volume of fluid, and (3) variation in demand. Some of failure reasons are induced by captivation,

hydraulic instability, or other system related problems. Others are the direct result of improper maintenance, maintenance-related problems, improper lubrication, misalignment, unbalance, seal leakage, and a variety of others in which machine reliability is periodically affected.

In this research, we use the data of a real centrifugal pump used in a petroleum industry located in the south of Iran. The data consists of 7 columns, the first six are features, i.e. flow, temperature, suction pressure, discharge pressure, velocity and vibration. The last column i s the fault class related to those features ranged from 1 to 5. Table 1 shows an example of the given data and our problem of fault classification.

Table 1. A row of the given data sheet containing values of the six features and the related fault type and under that a row of features without the fault type which should be diagnosed by us.

Flow	Temperature	Suction pressure	Discharge pressure	Velocity	Vibration	Fault type
57	96	20	700	3.5	7.67	3
a	b	c	d	e	f	?

Considering the above explanations, our problem is to devise precise intelligent approaches which receive a number of data sheet's rows, learn the pattern behind the features and given fault types, and finally, are themselves able to detect faults by giving them only the features' values afterwards. Obviously, approaches with less errors or misclassification are more favourable. Due to the fact that failure diagnosis by human is time consuming and human errors may happen, using artificial intelligence and machine learning classification methods has gained popularity to develop a diagnostic scheme. Artificial Neural Networks (ANNs), which are inspired from the biological nervous systems, have been widely used by researchers in the field of classification. Support Vector Machine (SVM) presented by *Vapnik 1995* [15] is a strong classification method based on the Structural Risk Minimisation (RSM). The application of SVM in classification is called Support Vector Classification (SVC). Hence, SVM and SVC mean exactly the same in this paper.

The reminder of this paper is organised as follows: In Sect. 2 a review of the related literature and different methods used for fault classification of pumps and similar devices are presented. Our ANN and SVC approaches are described in Sect. 3. Section 4 contains the results and comparisons of the all methods applied. Finally, conclusions and a recommendation for future research are covered in Sect. 5.

2   Related Work

In this section we aim at presenting an overview of the methods applied to classifying and clustering faults in centrifugal pumps and the likes. Researchers of

this field have widely used Artificial Intelligence (AI) due to its applicability and capability in learning complicated patterns and accurate classification. *Sun et al. 2012* [13] review Computational Intelligence (CI) approaches for oil-immersed power transformer maintenance by discussing historical developments and by presenting state-of-the-art fault diagnosis methods.

ANNs, which are of prominent approaches in AI, have been chosen as classifier in many papers. As some examples: *Unal et al. 2014* [14] propose an ANN based fault estimation algorithm verified with experimental tests and promising results. Their ANN model was modified using a genetic algorithm providing an optimal skilful fast-reacting network architecture with improved classification results. In *Azadeh et al. 2013* [2] a unique flexible algorithm is proposed for condition monitoring of a centrifugal pump into two different states based on ANN and SVM with hyper-parameters optimisation.

SVM has gained a considerable popularity among the surveys done in recent years. In *Bacha et al. 2012* [3] an intelligent fault classification with a SVM approach is applied to power transformer Dissolved Gas Analysis (DGA). An application of the SVM in multiclass gear-fault diagnosis is studied by *Bansal et al. 2013* [4]. *Bordoloi & Tiwari 2014* [5] attempt the multi-fault classification of gears by SVM learning technique using frequency domain data. *Fai & Zhang 2014* [6] applied support vector machine with genetic algorithm to fault diagnosis of a power transformer in which genetic algorithm is used to select appropriate free parameters of SVM.

An improved Ant Colony Optimisation (IACO) algorithm is proposed in *Li et al. 2013* [9] to determine the parameters of SVM and then it is applied to the rolling element bearing fault detection. *Gryllias and Antoniadis* [7] propose a hybrid two stage one-against-all SVM approach for the automated diagnosis of defective rolling element bearings. In *Muralidharan et al. 2014* [12] the application of SVM algorithm in the field of fault diagnosis and condition monitoring are discussed. *Wang et al. 2014* [16] develop a noise-based intelligent method for Engine Fault Diagnosis (EFD), so-called HHT–SVM model based on the techniques of Hilbert-Huang Transform (HHT) and Support Vector Machine (SVM). *Zhu et al. 2014* [19] train a multi-class SVM to achieve a prediction model by using Particle Swarm Optimisation (PSO) to seek the optimal parameters.

Other methods are used in this area as well. For instance: The survey of *Lei et al. 2013* [8] summarises the recent research and development of Empirical Mode Decomposition (EMD) in fault diagnosis of rotating machinery. *Azadeh et al. 2010* [1] provide a correct and timely diagnosis mechanism of pump failures by knowledge acquisition through a fuzzy rule-based inference system which could approximate human reasoning. The study of *Muralidharan & Sugumaran 2013* [11] uses vibration signals for fault diagnosis of centrifugal pumps using wavelet analysis. *Zhang and Nadi 2007* [18] propose three Genetic Programming based approaches for solving multi-class classification problems in roller bearing fault detection.

At last it is worth mentioning that although rarely but clustering approaches are used in the field of fault detection. *Zogg et al. 2006* [20] is an example

that simplifies known clustering techniques and introduces new vector clustering techniques for faults of heat pumps.

3   Description of the Applied Methods

In this section detailed explanations on the structure of our employed ANN and SVC (SVM) methods are given. Afterwards, we describe how GA is combined with these two classification approaches and illustrate the overall procedure of the devised integrated ANN-GA and SVC-GA algorithms.

3.1   The ANN-GA Framework

In machine learning, Artificial Neural Networks are a family of statistical learning algorithms inspired by biological neural networks. They are mainly used for function approximation, pattern recognition and classification. ANNs are presented as systems of interconnected "neurons" which can compute values and the combination of them leads to a network that can learn a complicated pattern between inputs and outputs. An ANN consists of nodes as neurons in different layers. Each node transmits a final value to nodes of the next layer. This value can be obtained by a function in the node called the activation function. The first layer has neurons equal to the number of inputs, whereas the last layer has neurons equal to the outputs. Between these two layers some hidden layers may exist to boost the ability of the ANN. For our classification problem we need an ANN that receives the values of six features as inputs and diagnoses the fault type based on them. Figure 1 depicts the structure of the applied ANN with nodes and the activation functions to convert a series of features to a fault type. The Network is fed by the six inputs and send them to all of the 3 neurons considered for the next layer. In each neuron i of the middle layer the weighted sum of inputs plus a constant number b_i is computed which is called T_i of the neuron, $T_i = w_{i1}x_1 + w_{i2}x_2 + ... + w_{i6}x_6 + b_i$. The three values resulted from this layer are sent to the last layer consisted of only one neuron. This last neuron acts exactly like the neurons of the previous layer and returns the weighted sum of the received values added by a constant b_4, $E = \sum_{i=1}^{3} l_i T_i + b_4$. Finally E goes through a step function, which determines the final output of the network or the fault class based on the amounts of c_i.

We summarise the main parameters of this network, which have a crucial effect on its performance in matrices:

$$W = \begin{vmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ w_{2,1} & w_{2,2} & w_{2,3} \\ w_{3,1} & w_{3,2} & w_{3,3} \end{vmatrix}$$

$$B = [b_1, b_2, b_3, b_4] \ , \ L = [l_1, l_2, l_3,], \ C = [c_1, c_2, c_3, c_4].$$

$$E = \sum_{i=1}^{3} l_i N_i + b_4,$$

$$Output = \begin{cases} 1 & if\ E < c_1 \\ 2 & if\ c_1 \le E < c_2 \\ 3 & if\ c_2 \le E < c_3 \\ 4 & if\ c_3 \le E < c_4 \\ 5 & if\ E > c_4 \end{cases}$$

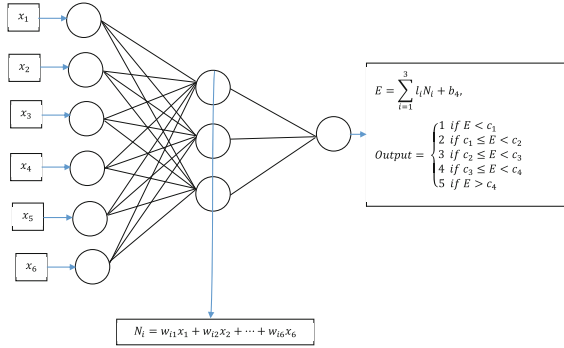$$N_i = w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{i6}x_6$$

Fig. 1. The structure of the applied ANN for fault classification

Choosing the best amounts for the above parameters can improve the classification performance of the ANN but it is a very difficult task to adjust them on the best or optimal values. Hence, besides the conventional training methods, we apply Genetic Algorithm, which is a powerful evolutionary optimisation algorithm and is able to obtain solution of good qualities in real time. For detailed explanation of GA, readers are referred to *M.D.Vose 1999* [10]. Due the fact that these parameters are continuous, we use the continuous version of GA in which the values of the parameters are considered as genes and they constitute a chromosome together. The initial population is generated by producing 200 chromosomes. The fitness of each chromosome is evaluated based on the below function:

$$Fitness\ function = 1 - percentage\ of\ correct\ predicted\ classes = 1 - \frac{N_c}{N_T}$$

Where N_c is the number of correct predicted faults and N_T is the total number of predictions. The algorithm seeks to minimise the above fitness function iteration by iteration to reach a near optimal solution in the end. The main characteristics of the applied GA are as follows: *Population size* $= 200, Crossover\ percentage = 0.7, Mutation\ percentage = 0.3, Maximum\ of\ Iterations = 100$

Figure 2 illustrates the procedure of our combined ANN-GA algorithm. According to the figure, firstly, the initial amounts of (W, B, L, C) are set. The data sheet is divided to training data set and testing data set. GA begins to optimise the amount of parameters by its selection, crossover and mutation operators. The algorithm terminates by reaching the maximum of iterations and determines the best found values for ANN parameters. Then these values are used for fault detections afterwards.

3.2   The SVM-GA Framework

In SVM (SVC), we have a set of training input $D = \{(x_1, x_2), ..., (x_i, y_i)\}$, where $x \in R^d$ and $y \in \{-1, 1\}$ is the class label, $i = 1, ..., l$. The method seeks to find
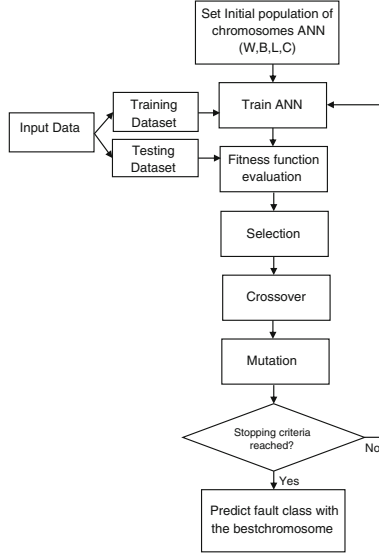
Fig. 2. The procedure of the applied ANN-GA algorithm

a separating hyper plane that maximises the distance to the nearest data points of each class. This goal is met by minimising the following objective function:

$$Max \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \varepsilon_i \tag{1}$$

$$Subject \ to \ \ y_i[W^T.\Phi(x_i)] \geq 1 - \varepsilon_i \tag{2}$$

$$\varepsilon_i \geq 0, i = 1, ..., l$$

This model is called soft margin SVM and ε_i handles misclassification, w is a weight vector, b is bias and C is the misclassification penalty to trade-off between the model complicity and training error. In equation (2), $\Phi(x_i)$ is a non-linear function and maps the input data to a high dimensional feature space where data can be separated linearly. Considering necessary condition for optimality, one can turn the above minimization problem into the following dual form:

$$Max \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j K(x_i, x_j) \tag{3}$$

$$Subject \ to \ \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{4}$$

$$0 \leq \alpha_i \leq C, i = 1, ..., l,$$

where $K(x_i, x_j)$ is a kernel function representing the inner product of $\langle \Phi(x_i), \Phi(x_j) \rangle$ and α_i is a Lagrangian multiplier. Solving the dual problem leads to the optimal separating hyper plane as following:

$$\sum_{SV} \alpha_i y_i K(x_i, x_j) + b = 0. \qquad (5)$$

The optimal classifying rule is:

$$f = sgn(b + \alpha_i [y_i K(x_i, x_j)]), \qquad (6)$$

where SV is the set of support vectors that the corresponding Lagrangian multipliers are positive for them. Figure 3 shows how a soft margin SVM with linear separating hyper plane divides the data into two classes.
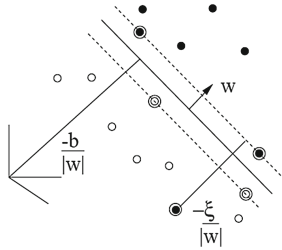

Fig. 3. Linear seperating hyper planes in soft margin SVM

We used the following kernel functions in our SVC for the fault diagnosis of the centrifugal pump:

$$Polynomial : K(x_i, x_j) = (\gamma. < x_i, x_j > +s)^d \qquad (7)$$

$$Gaussian\ basis\ function : K(x_i, x_j) = -\gamma.\|x_i - x_j\|^2 \qquad (8)$$

$$Linear : K(x_i, x_j) =< x_i, x_j > \qquad (9)$$

$$Quadratic : K(x_i, x_j) = (< x_i, x_j > +1)^2 \qquad (10)$$

As the parameters, i.e. C, γ, s, of SVM like those of ANN can strongly affect its performance, properly adjusting them can considerably improve it. Hence, other than conventional methods, these parameters are determined with the GA used for ANN. Considering the fact that SVM can only divide the data into two groups and there are 5 fault classes in our problem to be classified, we should implement SVM four times after each other. Each run of SVM classifies one fault from the remaining ones and has its own training process. Figures 4 and 5
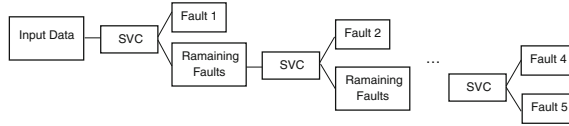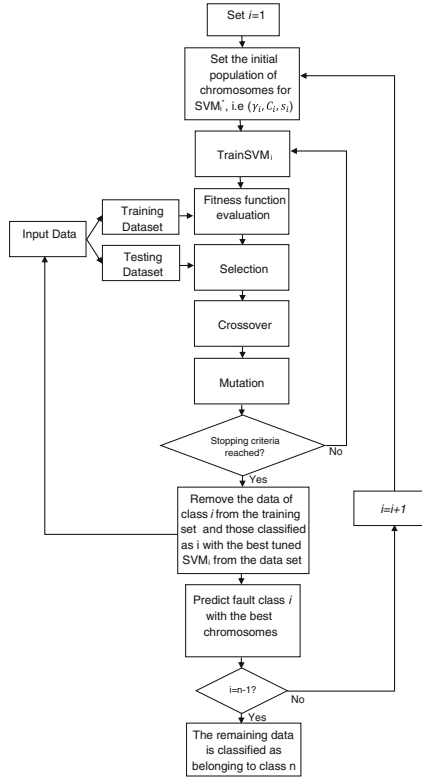
Fig. 4. SVM approach to obtain 5 classes

Fig. 5. The procedure of the applied SVM-GA

illustrate the framework of our SVM-GA algorithm. As it is shown by Fig. 5, the initial parameters of SVM are set at first. Then the GA searches in the space of parameter amounts for each of the 4 runs separately and as it terminates for each, it begins with the parameter setting of the next one. Finally, when the parameters have been tuned for all the runs, the procedures ends with the best values found.

Figure 6 illustrates the fitness function values of SVM-GA with Gaussian kernel function from the first up to the last iteration of GA.
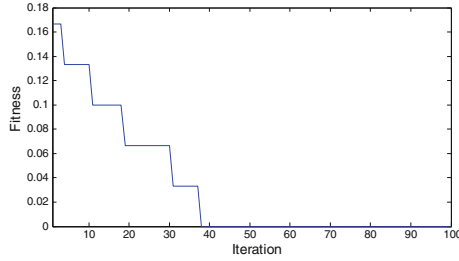
Fig. 6. Fitness function values of SVC-Gaussian-GA

4   Results and Comparisons

In this section we present a brief overview of the achieved results. We have alto-gether 100 rows of data. For feeding the algorithms, 70 % of data are randomly considered for training and 30 % as testing data. To make the data noisy for testing the robustness of the approaches, 0.1 is added to 30 % of columns 1, 3, and 6 of the data sheet. Table 2 shows percentage of correct fault diagnosis of pure GA and SVM methods without GA improvements and Fig. 7 depicts this amounts visually. SVM-Gaussian has the best performance and a good robust-ness. SVC-Linear is in the second position but it has the highest robustness among all. SVC-Quadratic, ANN and SVC-Polynomial are in the next ranks. It is worth mentioning that ANN has the worst robustness.

Table 2. Correct diagnosis proportion of the pure ANN and SVM

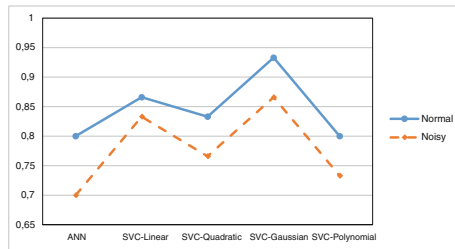|	ANN	SVM			
		Linear	Quadratic	Gaussian	Polynomial
Normal	0.8	0.866	0.833	0.933	0.8
Noisy	0.7	0.8333	0.766	0.866	0.733

Fig. 7. Correct diagnosis of the pure ANN and SVM with Normal and Noisy Data

To show the superiority of our ANN-GA and SVM-GA, the diagnosis experiments are also done with K-Nearest Neighbours (KNN) and Decision Tree, which are of high rated classification methods. Table 3 and Fig. 8 show these performance comparisons. According to the results, SVM-Gaussian has again the best performance among all and GA enhancement has enabled it to detect faults in all cases correctly both in normal and noisy environment. Therefore, it is the most robust as well, together with SVM-GA-linear. ANN-GA performs worse than SVM-GA with all the kernels in terms of correctness, and considering robustness, SVM-GA is superior except for the case of polynomial function which results almost the same as ANN-GA. Finally, the worst diagnosis performances belong to KNN and Decision Tree.

Table 3. Correct diagnosis proportion of ANN-GA, SVM-GA, KNN and Decision Tree with Normal and Noisy Data

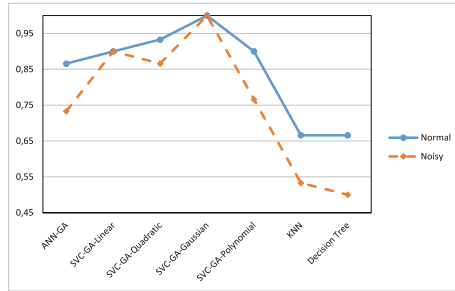|	ANN-GA	SVM-GA				KNN	Decision Tree
		Linear	Quadratic	Gaussian	Polynomial		
Normal	0.866	0.9	0.833	0.933	1	0.9	0.666
Noisy	0.733	0.9	0.866	1	0.766	0.533	0.5

Fig. 8. Correct diagnosis proportion of ANN-GA, SVM-GA, KNN and Decision Tree with Normal and Noisy Data

To show the GA effects on ANN and SVM, the performance improvements are depicted by Fig. 9. The largest improvement is for SVM-Gaussian in noisy condition and lowest for SVM-Polynomial in noisy condition.

For Comparisons of the methods in detail, McNemar's tests are executed and the results are tabulated as Table 4 to examine which model outperforms the others significantly. McNemar's is a nonparametric statistical test for two related nominal samples with a null hypothesis of marginal homogeneity in 2×2 contingency tables. A detailed explanation on McNemar's test is provided in *Webb et al. 2011* [17]. If the significance level is set to 10 %, then a p-value
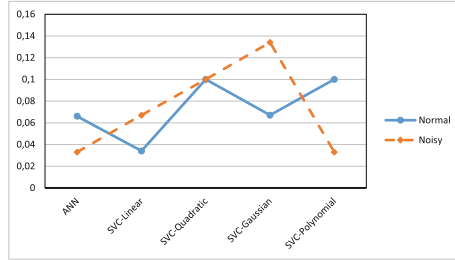
Fig. 9. The improvement of methods by GA

less than 0.1 shows that models vary significantly. The tests for SVC and SVC-GA are done with their best kernel function which is Gaussian according to the accuracy results.

Table 4. MCNemar's test results (*p*-values)

	ANN-GA	SVM	ANN	Decision Tree	KNN
Normal environment					
SVC-GA	0.1336	0.4795	0.0412	0.0044	0.0044
ANN-GA		0.6171	0.4795	0.0771	0.0412
SVC			0.1336	0.0133	0.0133
ANN				0.1138	0.0771
Decision Tree					0.7518
Noisy environment					
SVC-GA	0.1333	0.1336	0.0077	0.0003	0.0003
ANN-GA		0.2207	1	0.0455	0.0771
SVC			0.1306	0.0026	0.0044
ANN				0.771	0.1824
Decision tree					1

At the end of this section, we perform 10-fold cross-validation to evaluate the validity of models. For this sake, the data sheet is divided into 10 even subsets, then each of them is once used as the test dataset and the other 9 as training dataset. Finally, the averages of models' accuracies (proportion of correct predicted fault types) are considered for models' validity evaluation. The average of models' accuracies are presented in Table 5.

5   Conclusion

In this paper we presented fault classification algorithms by combination of two intelligent machine learning methods, namely ANN and SVM, with Genetic

Table 5. Average of accuracies in 10 fold

	Normal environment	Noisy environment
SVC-GA	0.95	0.95
SVC	0.9	0.85
ANN-GA	0.85	0.75
ANN	0.85	0.8
KNN	0.6	0.6
Decision tree	0.6	0.5

Algorithm. The results showed that GA can significantly improve the performance of the classifiers. The performances of all employed algorithms, i.e. ANN, ANN-GA, SVC, SVC-GA, KNN and Decision Tree, were compared by different tests in normal and noisy condition. The comparisons showed that SVM with Gaussian kernel function had the best accuracy in correct fault diagnosis and an excellent robustness against noise. It was also observed that SVM is superior to ANN in most of the cases. For future research in this direction, testing the ability of other optimisation algorithms to improve ANN, SVM and other classification methods is recommended.

References

1. Azadeh, A., Ebrahimipour, V., Bavar, P.: A fuzzy inference system for pump failure diagnosis to improve maintenance process: the case of a petrochemical industry. Expert Syst. Appl. 37(1), 627–639 (2010). http://dx.doi.org/10.1016/j.eswa.2009.06.018
2. Azadeh, A., Saberi, M., Kazem, A., Ebrahimipour, V., Nourmohammadzadeh, A., Saberi, Z.: A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyperparameters optimization. Appl. Soft Comput. J. 13(3), 1478–1485 (2013). http://dx.doi.org/10.1016/j.asoc.2012.06.020
3. Bacha, K., Souahlia, S., Gossa, M.: Power transformer fault diagnosis based on dissolved gas analysis by support vector machine. Electric Power Syst. Res. 83(1), 73–79 (2012). http://dx.doi.org/10.1016/j.epsr.2011.09.012
4. Bansal, S., Sahoo, S., Tiwari, R., Bordoloi, D.: Multiclass fault diagnosis in gears using support vector machine algorithms based on frequency domain data. Measurement 46(9), 3469–3481 (2013). http://www.sciencedirect.com/science/article/pii/S0263224113002078
5. Bordoloi, D.J., Tiwari, R.: Optimum multi-fault classification of gears with integration of evolutionary and SVM algorithms. Mech. Mach. Theor. 73, 49–60 (2014). http://dx.doi.org/10.1016/j.mechmachtheory.2013.10.006
6. Fei, S.W., Zhang, X.B.: Fault diagnosis of power transformer based on support vector machine with genetic algorithm. Expert Syst. Appl. 36(8), 11352–11357 (2009). http://dx.doi.org/10.1016/j.eswa.2009.03.022

7. Gryllias, K.C., Antoniadis, I.A.: A support vector machine approach based on physical model training for rolling element bearing fault detection in industrial environments. Eng. Appl. Artif. Intell. 25(2), 326–344 (2012). http://dx.doi.org/10.1016/j.engappai.2011.09.010

8. Lei, Y., Lin, J., He, Z., Zuo, M.J.: A review on empirical mode decomposition in fault diagnosis of rotating machinery. Mech. Syst. Sig. Process. 35(1–2), 108–126 (2013). http://dx.doi.org/10.1016/j.ymssp.2012.09.015

9. Li, X., Zheng, A., Zhang, X., Li, C., Zhang, L.: Rolling element bearing fault detection using support vector machine with improved ant colony optimization. Meas. J. Int. Meas. Confederation **46**, 2726–2734 (2013)

10. Voser, M.D.: The Simple Genetic Algorithm: Foundation and Theory. MIT Press, Cambridge (1999)

11. Muralidharan, V., Sugumaran, V.: Rough set based rule learning and fuzzy classification of wavelet features for fault diagnosis of monoblock centrifugal pump. Measurement: Journal of the International Measurement Confederation 46(9), 3057–3063 (2013). http://dx.doi.org/10.1016/j.measurement.2013.06.002

12. Muralidharan, V., Sugumaran, V., Indira, V.: Fault diagnosis of monoblock centrifugal pump using SVM. Int. J. Eng. Sci. Technol. 17(3), 1–6 (2014). http://linkinghub.elsevier.com/retrieve/pii/S2215098614000275

13. Sun, H.C., Huang, Y.C., Huang, C.M.: Fault Diagnosis of power transformers using computational intelligence: a review. Energy Procedia 14, 1226–1231 (2012). http://dx.doi.org/10.1016/j.egypro.2011.12.1080

14. Unal, M., Onat, M., Demetgul, M., Kucuk, H.: Fault diagnosis of rolling bearings using a genetic algorithm optimized neural network. Measurement 58, 187–196 (2014). http://linkinghub.elsevier.com/retrieve/pii/S0263224114003601

15. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)

16. Wang, Y.S., Ma, Q.H., Zhu, Q., Liu, X.T., Zhao, L.H.: An intelligent approach for engine fault diagnosis based on Hilbert-Huang transform and support vector machine. Appl. Acoust. **75**, 1–9 (2014)

17. Webb, A.: Statistical Pattern Recognition. Wiley, New York (2011)

18. Zhang, L., Nandi, A.K.: Fault classification using genetic programming. Mech. Syst. Sig. Process. **21**(3), 1273–1284 (2007)

19. Zhu, K., Song, X., Xue, D.: A roller bearing fault diagnosis method based on hierarchical entropy and support vector machine with particle swarm optimization algorithm. Measurement 47, 669–675 (2014). http://www.sciencedirect.com/science/article/pii/S0263224113004569

20. Zogg, D., Shafai, E., Geering, H.P.: Fault diagnosis for heat pumps with parameter identification and clustering. Control Eng. Pract. **14**(12), 1435–1444 (2006)