Classification Ensembles for Shaft Test Data: Empirical Evaluation

Kyungmi Lee and Vladimir Estivill-Castro
School of Computing and Information Technology
Griffith University, Queensland, Australia
kyungmi.lee@student.griffith.edu.au, v.estivill-castro@cit.gu.edu.au

Abstract

A-scans from ultrasonic testing of long shafts are complex signals. The discrimination of different types of echoes is of importance for non-destructive testing and equipment maintenance. Research has focused on selecting features of physical significance or exploring classifier like Artificial Neural Networks and Support Vector Machines. This paper confirms the observation that there seems to be uncorrelated errors among the variants explored in the past, and therefore an ensemble of classifiers is to achieve better discrimination accuracy. We explore the diverse possibilities of heterogeneous and homogeneous ensembles, combination techniques, feature extraction methods and classifiers types and determine guidelines for heterogeneous combinations that result in superior performance.

1. Introduction

Applications of machine learning demand exploration of feature extraction methods and classifier types in order to obtain systems with reliable highest accuracy. The industrial application discussed here is the classification of ultrasonic echoes in an A-scan. The application is particularly challenging as A-scans are taken from the end of a long large complex shaft. Although several pattern analysis and machine learning techniques have been used with success in analyzing A-scan data [11, 20], they are typically in the context of very short signals. Those cases are usually much simpler; in particular, the task reduces to detecting the existence of an echo (indicating a fault in the material). In long shafts there are many kind of echoes, and in fact there are echoes for where there is no fault. These *mode-converted* echoes are the result of reflection and other artifacts of the ultrasonic signal navigating and filling the shaft. They may cause misjudgement of the position of real faults (cracks) of shafts, thus to discriminate them from genuine echoes is important.

The relationship between ultrasonic signal characteristics and flaw classes is not straightforward. We need to extract informative set of signal features which becomes the basis of decision-making for classification. Two main issues are to identify the better set of features and to identify the more suitable learning algorithm, in order to enhance the classification performance more accurately and reliably. For ultrasonic shaft signal classification the most competitive feature extraction approaches are Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT). Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are the top two approaches to build classifiers in this field.

Previously we focused on finding the best single classifier model (between ANN and SVM) and determining the best selected feature extraction scheme (between FFT and DWT). In this paper, we learn multiple models of the shaft test data and combine their outputs for making a final decision for classification. The reason for this *ensemble of classifiers* is that FFT might reflect physical properties that are different from those DWT conveys. We suspected that including the FFT as another informant of the decision process, even if the accuracy using DWT has shown to be superior, should improve accuracy.

Constructing hybrid ensembles is not trivial. There have been various approaches for creating multiple classifiers (model generation) and for combining the outputs of multiple classifiers (model combination) [8, 23]. There are two streams of model generation methods. Homogeneous generation creates multiple models trained by multiple data sets using a single learning algorithm. Those multiple data sets are generated either by different feature extraction schemes [16] or by partitioning a data set into multiple sets [3, 4]. The heterogeneous method creates multiple classifiers built using different learning paradigms [2]. Therefore, in order to construct an effective multi-classifier system, we need to decide a scheme for model generation and also a combination method for decision making. Since we design an ensemble of models using two different feature sets (FFT and DWT), we will not apply such schemes as

IEEE COMPUTER SOCIETY

boosting or bagging [3, 9] where only one feature extraction scheme is used.

We report results of our investigation into which method for model generation offers more improvement on the accuracy achieved by a single classifier. We generate heterogeneous and homogeneous models and combine the outputs of those multi-classifiers by three widely-known combining techniques; Bayesian Combination(BC), Distribution Summation (DS) and Likelihood Combination (LC). We also explore the effect of combining multiple models not only on the overall classification performance but also on classifying each class. The analysis and investigation results obtained are the basis for the construction of an integrated multi-classifier models using both feature schemes (FFT and DWT) effectively.

Our presentation continues in Section 2 with a summary of our previous studies for improving classification system for shaft test data. It includes motivating observations for attempting ensembles of classifiers. Section 3 describes our empirical evaluation, including the description of how we generated and combined multiple models and how we evaluated the classification performance of each ensemble. Section 4 analyzes the experimental result from various combination schemes and compares and discusses their performance, followed by conclusions in Section 5.

2. Background and challenges

2.1. Background summary

The problem is to discriminate efficiently the different types of reflectors among the large volumes of ultrasonic shaft-test data and classify them into a) those that correspond to design features of the shaft (DF), b) those that correspond to flaws, cracks and other defects (CR) and c) the multiple reflections and mode-converted echoes (MC) of the two previous cases. Among these three causes of echoes, type DF is considered easy to distinguish compared to the other types. Also, in the field, the signal echoes caused by CR can be confused by fainted echoes caused by MC and vice versa. Consequences of misclassification are catastrophic with enormous cost in downtime, consequential damage to associate equipment and potential injury to personnel [6].

Modern signal processing techniques and artificial intelligence tools eliminate inconsistent results present even in classification by the same human expert. These approaches are integrated as automatic ultrasonic signal classification (AUSC) systems. An AUSC system, preprocesses ultrasonic flaw signals acquired in a form of digitized data and extracts informative features using digital signal-processing techniques. The main interest for the AUSC research community has been the extraction of effective sets of features from which classification might be performed more efficiently and accurately. While it is hard to determine which set of features is best, it is important to at least identify those that make the process reliable and effective in the field. It is also important to relate some features to some understanding of the phenomena (in terms of its physics). However, the physics are complex, and the relationship between signal characteristics and flaw classes is not straightforward.

The FFT is a useful scheme for extracting frequency-domain signal features [6, 14]. This seems natural when dealing with ultrasound since the traditional representation of these types of signals is by mathematical Fourier series that identify physically meaningful features, like frequency and phase. But recent studies on the ultrasonic flaw classification employ the Discrete Wavelet Transform (DWT) as part of their feature extraction scheme. DWT provides effective signal compression and time-frequency presentation [15, 19]. Many researchers have compared these two feature extraction schemes (FFT and DWT), and most comparisons showed a superiority of DWT to FFT in discriminating the type of flaw (or its non-existence) [17, 18, 21]. The first study analyzing feature extraction in more complex ultrasonic signals from shafts [13] also established experimentally that DWT was a potentially stronger feature extraction scheme for feeding ANNs. However, considering the many difficulties inherent in the ANN learning paradigm (such as generalization control, overfitting and parameter tuning) we remained more conservative about DWT's predominance. Recently, a new comparative experiment involving SVM instead of ANN models [12] confirmed the DWT as indeed the superior feature extraction scheme in the classification of echoes from ultrasonic signals in long shafts, because the statistical properties of SVM indicate robustness in its construction, especially when a limited number of training examples are available.

2.2. Open issues

We observed differences for specific classes of echoes when reflecting upon the classification result of both schemes (FFT and DWT) analyzed in [12]. A classifier constructed using one scheme of feature extraction showed more accuracy in classifying a certain type of echo than in the case of using another scheme, but the roles are reversed for other echoes. Thus, FFT, in spite of lower accuracy for overall classification, could complement the decisions based on DWT features.

Combining classifiers improves the accuracy achieved by a single classifier when different classifiers implicitly represent different useful aspects of the input data. Techniques for combining multiple classifiers must solve two issues: 1) how to generate multiple models and 2) how to combine the prediction of the multiple models to produce an overall classification. Applying a single algorithm repeatedly to different versions of the training data, or applying different learning algorithms to the same data creates a set of learned models. There are two ways of manipulating different versions of the training data; either different subset of the training data or different set of input features. The various techniques for combining the predictions obtained from the multiple classifiers are largely categorized into voting (uniform or weighted), stacking methods and cascading methods.

The diversity of techniques for generating and combining models raises the issue of which generation-combination method to choose for constructing the most effective and reliable multi-model systems for our application domain. The theory suggests generating a set of models that are diverse in the sense that they make errors in different ways. We wish to investigate the classification performance by multi-models. Is better performance obtained when participant are trained by FFT features or DWT features? We also generate multiple models using different learning algorithms (SVM and ANN), and compare which combination paradigm is more suitable. We analyze the type of errors on each combination models in order to gain insight into appropriate combining strategies.

3. Experiments

Fig. 1 shows the steps of the two-phase experiment.

1. We map shaft inspected data into feature domains using two feature extraction schemes (FFT and DWT) and, using 5-folds cross-validation learning, we train SVM models and ANN models. We record their performance as single models.
2. We combine single models across two dimensions: 1) combining the decisions of FFT model and DWT model trained by a single learning paradigm and 2) combining the decisions of SVM model and ANN model with same feature scheme. We apply 3 combining methods for each combination. We compared classification accuracy with the result using a single model.

3.1. Generation of multiple classifiers

We acquired A-scan signals from eight shafts, ranging between 100mm to 1300mm in length using with the probe's frequency set to 2 MHz. We extract the signal segments of interest from the whole ultrasonic A-scan signals. In order to apply a consistent way of signal segmentation which is necessary for suppressing time-variance problems with DWT, we used a systematical echo capturing method with zero-padding (SZ) [13]. Using this gating method, we capture the 768 values of long time-domain vectors, and downsampled them into 128 values for input into the FFT. We concatenate the sequences of magnitude components and phase components (FFT coefficients) into a 128 dimensional pattern vector for classification. In parallel, we compressed the 768 values representing the DWT coefficients into 128 samples by discarding the last 128 coefficients (they are supposed not to contain much information but mainly noise). We store the 128 long vector of DWT coefficients as the DWT feature set. For our experiment, we applied Daubechies wavelets [7] for filtering.

Also, we consider 4 other different schemes for the selection of the signal's region of interest; namely, Central Peak positioning (CP), Main Energy capturing (ME) RAndom positioning (RA) and Systematical echo capturing method with the preservation of Original neighboring grass (SO). The classifiers trained by DWT data using these four methods showed weaker performance compared to the DWT-based classifier using SZ [13]. Despite their comparative weakness, we included these weak classifiers as members of multi-classifiers.

We used fully connected feed-forward neural networks with 128 input nodes, two hidden layers with 64 nodes and 16 nodes and an output layer with 2 nodes for classifying the shaft signals into cracks (CR) or mode-converted echoes (MO). We trained using the back-propagation algorithm in batch mode and the topological order as the update mode of the networks. The learning rate was 0.2 and the Mean Square Error limit was 0.01 for stopping the training process. The epoch limit was 200,000 for those occasional cases where training failed to converge. The input samples were randomly divided up into 5 sets. In turn, we use 4 of these to train the network, and the remaining set to validate the network. This was repeated with all five possible combinations and furthermore, the process was repeated 5 times to get the diversity of the networks training ability by assigning 5 different initial weights to the network. As the result of this process, we produce 150 ANN models trained by six different feature sets; one FFT feature set and five DWT feature sets which were preprocessed using five different schemes (RA, CP, SO, SZ and ME).

For SVM classifiers, we employed RBF kernels because they provide nonlinear mapping, require comparatively small numbers of hyperparameters and offer less numerical difficulties. RBF requires a penalty parameter C and kernel parameter γ. We used a grid searching algorithm [10] where pairs of C and γ are tried and the one with the best 10-fold-cross-validation accuracy is picked. The result of the grid searching was the values 4, 16, 1, 2, 1 and 8 for C corresponding to the six feature sets FFT, DWT-CP, DWT-ME, DWT-RA, DWT-SO and DWT-SZ. The respective γ values are 1/32, 1/32, 1/32, 1/8, 1/8 and 1/32. Again, we used

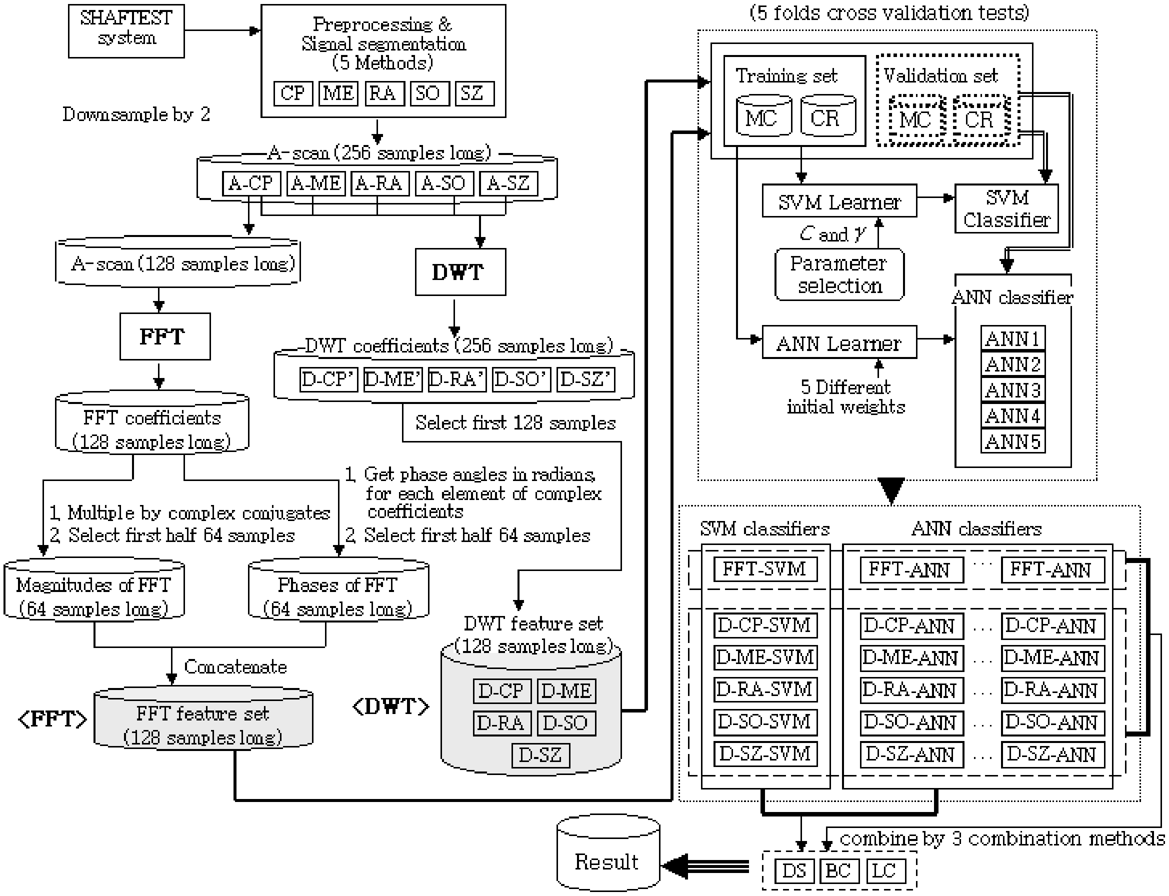
SHAFTEST system
Preprocessing & Signal segmentation (5 Methods)
CP
ME
RA
SO
SZ
Downsample by 2
A-scan (256 samples long)
A-CP
A-ME
A-RA
A-SO
A-SZ
A-scan (128 samples long)
DWT
FFT
DWT coefficients (256 samples long)
D-CP'
D-ME'
D-RA'
D-SO'
D-SZ'
FFT coefficients (128 samples long)
Select first 128 samples
1. Multiple by complex conjugates
2. Select first half 64 samples
1. Get phase angles in radians for each element of complex coefficients
2. Select first half 64 samples
Magnitudes of FFT (64 samples long)
Phases of FFT (64 samples long)
DWT feature set (128 samples long)
Concatenate
<FFT>
FFT feature set (128 samples long)
<DWT>
D-CP
D-ME
D-RA
D-SO
D-SZ
(5 folds cross validation tests)
Training set
MC
CR
Validation set
MC
CR
SVM Learner
SVM Classifier
C and γ
Parameter selection
ANN classifier
ANN Learner
ANN1
ANN2
ANN3
ANN4
ANN5
5 Different initial weights
SVM classifiers
ANN classifiers
FFT-SVM
FFT-ANN
...
FFT-ANN
D-CP-SVM
D-CP-ANN
D-CP-ANN
D-ME-SVM
D-ME-ANN
D-ME-ANN
D-RA-SVM
D-RA-ANN
D-RA-ANN
D-SO-SVM
D-SO-ANN
D-SO-ANN
D-SZ-SVM
D-SZ-ANN
D-SZ-ANN
combine by 3 combination methods
Result
DS
BC
LC

Figure 1. Overall procedure of our experiment.

5 fold cross-validation test on six SVM models, which are trained by six feature sets. Thus, we manipulated 30 individual SVM classifiers.

3.2. Combination of multiple classifiers

The purpose of our experiment is to empirically investigate what combination is most fruitful. Thus, we investigate the impact on classification performance from combinations of multiple classifiers trained by different feature schemes or different learning paradigms. We explore three combining methods namely, Bayesian Combination(BC) [22, 24], Distribution summation(DS) [5] and Likelihood Combination(LC) [1].

In a Bayesian combination method, weights are established proportional to each individual classifier's past performance which are computed by its posterior probability using Bayes' theorem. In the Distribution summation method, distributions with each individual model are presented as a vector which records how many training data are correctly classified for each class. These vectors of multiple models are combined using vectorial addition for making the combined decision. The likelihood combination method is a weighted combination in which the Naive Bayes algorithm is applied to learn weights for classifiers.

We combine the outputs of two individual models under two streams; the first stream is to combine FFT models and DWT models trained by one same learning algorithm (ANN or SVM). We produced one FFT (data set) type classifier and five different DWT (data sets) type classifiers. Thus, they are five FFT-DWT combinations. Learning with SVM or with ANN from the feature-combination results in 10 ensembles. These are the first two rows in Fig. 2. The second stream is to combine one SVM model with one ANN model trained by one same feature set (FFT or DWT). Thus, we get 6 ensembles (corresponding to the third row in Fig. 2). All the combination are carried out by three different combining methods (the 3 columns within each major column in Fig. 2). The classification accuracy for each class is also recorded separately from the overall accuracy.

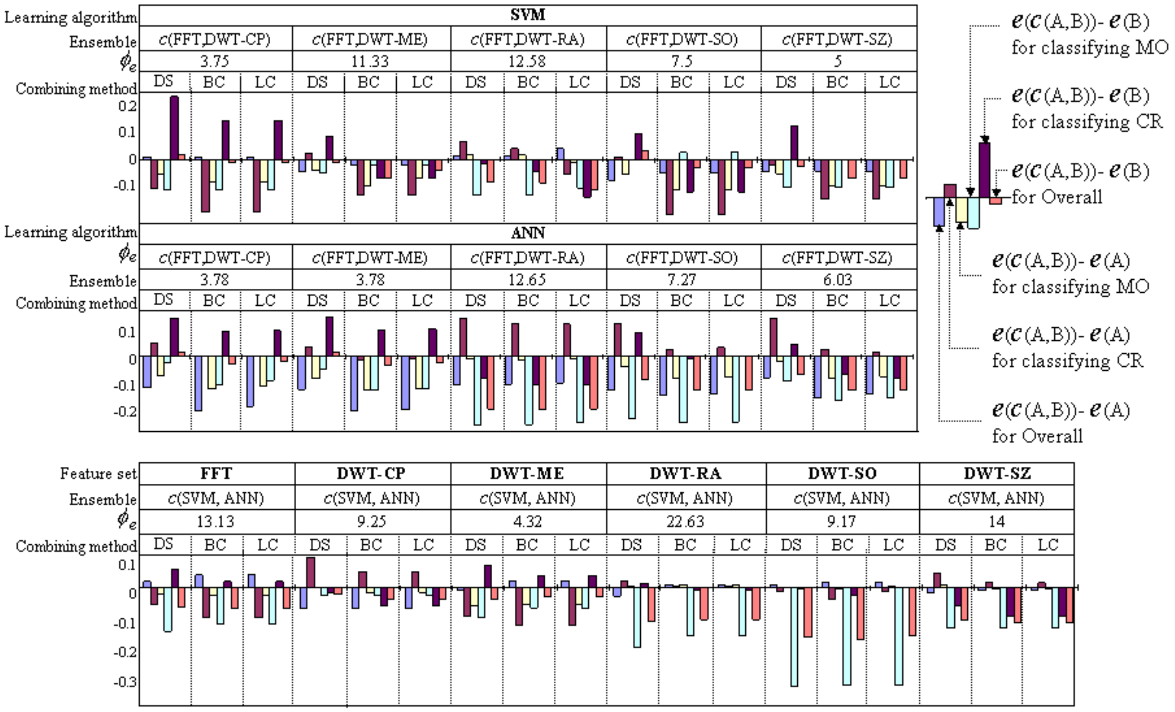

Figure 2. Comparison of the combined models performance. *c*(A, B) indicates a combined classifier of two individual classifiers model A and model B. *e*(A) indicates the classification error rate of a classifier model A.

4. Results and analysis

In order to investigate if the combined model performs more accurately in classifying input data than a single model does, we compared the decision error rate of the combined model with the decision error rate of each individual participant model. Fig. 2 displays the amount of improvement in the classification performance. This picture shows bar-charts presenting the difference between the error rate of the combined model and the error rate of the single model. Bar-charts where the combination is an improvement point downwards while if a single classifier remains better, the bar-chart points upwards. We also computed a value ϕ_e which indicates the ”fraction of correlated errors” [1] and is also listed in Fig. 2. The value of ϕ_e is generally used to measure the degree to which the errors made by models of the ensemble are correlated.

The following points are noteworthy.

- Combined models show better performance than single model in terms of the classification accuracy for the whole test data set across schemes for generating or combining multi-classifiers.
- Combining two classifiers trained by different feature sets become more advantageous when we use SVM as a learning algorithm than using ANN (refer to top 2 rows of bar-charts in Fig. 2).
- Though the overall accuracy of combined models is higher than the accuracy of single models across most types of combination, their performance in classifying each class data (MO and CR) is diverse. Especially, most FFT&DWT ensembles trained by ANN perform worse than single model in classifying CR data, whilst corresponding combined models trained by SVM perform reliably on both class except for one combination (the FFT&DWT-CP ensemble).
- Amongst the five types of DWT data combined with FFT data, DWT-SZ shows most reliability in classifying both classes regardless of learning paradigm. This implies that different echo gating preprocessing for extracting DWT features plays a role in making the DWT feature-sets. We suspect there are some implicit differences in DWT.
- The performance of the heterogeneously combined classifiers is different depending on which feature sets were used to train them.
- The value of ϕ_e is related with the amount of error reduction made by combining multi-classifiers. As

shown in Fig. 2, the value of ϕ_e seems to be much relevant to the overall error reduction rate. It seems not to have much relevance with the error reduction for each class data.

- The most suitable combination structure may depend on the interest of some particular class. For example, if accuracy for the CR class is the issue, then the SVM with DWT (single classifier) is not surpassed by the combination. Although the combination does better overall the classes.

5. Conclusion

We have explored the combining of classifiers along the dimension of feature extraction mechanism, along the dimension of combination method and along the dimension of type of classifier.

This experimental result suggests guidelines for designing an integrated multi-classifier system for shaft test data by the way of selectively employing the combining structure used in this experiment. Namely, combination in general improves the accuracy, and combining features has the potential for improvement. However, the most productive combination that offers the most improvement is usually a combination of ANN and SVM from DWT as the building feature.

References

[1] K. Ali and M. Pazzani. Error reduction through Learning Multiple Descriptions. *Machine Learning*, 24(3):173–202, 1996.

[2] D. Bahler and L. Navarro. Methods for combining heterogeneous sets of classifiers. In *Proceedings of the 17th Natl. Conf. on Artificial Intelligence (AAAI), Workshop on New Research Problems for Machine Learning*, 2000.

[3] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] G. Briem, J. Benediktsson, and J. Sveinsson. Boosting, Bagging and Consensus Based Classification of Multisource Remote Sensing Data. In J. Kitter and F. Roli, editors, *Multiple Classifier Systems. Second International Workshop, MCS 2001*, Lecture Notes in Computer Science 2096, pages 279–288. Springer-Verlag, 2001.

[5] P. Clark and R. Boswell. Rule Induction with CN2: Some Recent Improvements. In *Proceedings of the European Working Session on Learning*. Pitman, 1991.

[6] G. Cotterill and J. Perceval. A New Approach to Ultrasonic Testing of Shafts. In *Proceedings of the 10th Asia-Pacific Conference on Non-Destructive Testing (APCNDT)*, 2001.

[7] I. Daubechies. Orthogonal bases of compactly supported wavelets. *Commun. Pure Appl. Math.*, 41:909–996, 1988.

[8] T. G. Dietterich. Machine-learning research: Four current directions. *AI Magazine*, 18(4):97–136, 1997.

[9] B. Efron and R. Tibshirani. *An introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[10] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification.

[11] G. Katragadda, S. Nair, and G. P. Singh. Neuro-Fuzzy Systems in Ultrasonic Weld Evaluation. *Review of Progress in Quantitative Nondestructive Evaluation*, 16:765–772, 1997.

[12] K. Lee and V. Estivill-Castro. Support Vector Machine Classification of Ultrasonic shaft Inspection Data Using Discrete Wavelet Transform. In *Proceedings of the 2004 International Conference on Machine Learning; Models, Technologies and Applications*, pages 848–854.

[13] K. Lee and V. Estivill-Castro. Classification of Ultrasonic Shaft Inspection Data Using Discrete Wavelet Transform. In *Proceedings of the IASTED international conferences on Artificial Intelligence and appliction*, pages 673–678. ACTA Press, 2003.

[14] F. W. Margrave, K. Rigas, D. A. Bradley, and P. Barrocliffe. The use of neural networks in ultrasonic flaw detection. *Measurement*, 25:143–154, 1999.

[15] M. S. Obaidat, M. A. Suhail, and B. Sadoun. An intelligent simulation methdology to characterize defects in materials. *Information Sciences*, 137:33–41, 2001.

[16] D. Opitz. Feature selection for ensembles. In *Proceedings of the 16th National Conference on Artificial Intelligence*, pages 379–384. AAAI Press, 1999.

[17] R. Polikar, L. Udpa, S. S. Udpa, and T. Taylor. Frequency Invariant Classification of Ultrasonic Weld Inspection Signals. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 45(3):614–625, May 1998.

[18] D. Redouane, K. Mohamed, and B. Amar. Flaw Detection in Ultrasonics Using Wavelets Transform and Split Spectrum. In *Proceedings of the 15th World Conference on Non-Destructive Testing*, 2000.

[19] G. Simone, F. C. Morabito, R. Polikar, P. Ramuhalli, L. Udpa, and S. Udpa. Feature extraction techniques for ultrasonic signal classification. In *Proceedings of the 10th Int. Symposium on Applied Electromagnetics and Mechanics (ISEM 2001)*, 2001.

[20] S. J. Song, H. J. Kim, and H. Lee. A systematic approach to ultrasonic pattern recognition for real-time intelligent flaw classification in weldments. *Review of Progress in Quantitative Nondestructive Evaluation*, 18:865–872, 1999.

[21] J. Spanner, L. Udpa, R. Polikar, and P. Ramuhalli. Neural networks for ultrasonic detection of intergranular stress corrosion cracking. *The e-Journal of Nondestructive Testing And Ultrasonics*, 5(7), July 2000.

[22] C. Suen and L. Lam. Multiple classifier combination methodologies for different output levels. In *Multiple Classifier Systems. First International Workshop, MCS 2000*, Lecture Notes in Computer Science 1857, pages 52–66. Springer-Verlag, 2000.

[23] G. Valentini and F. Masulli. Ensembles of Learning Machines. In R. Tagliaferri and M. Marinaro, editors, *Neural Nets WIRN vietri-2002*, Lecture Notes in Computer Science 2486, pages 3–19. Springer-Verlag, 2002.

[24] L. Xu, C. Krzyzak, and C. Suen. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3).