CLASSIFICATION ENSEMBLES FOR SHAFT TEST DATA : EMPIRICAL EVALUATION

KYUNGMI LEE AND VLADIMIR ESTIVILL-CASTRO

School of Information and Communication Technology
Griffith University, Brisbane 4111, QLD, Australia

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.

Keywords: Pattern analysis in signals, feature extraction, ensemble of classifiers, artificial neural networks, support vector machines.

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 [Katragadda et al, 1997; Song et al, 1999], 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 misjudgment 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 that 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. Figure 1 graphically presents two different constructing methods for these two types of hybrid ultrasonic classification systems. The reason for the *ensemble of classifiers* (Type I in Figure 1) rather than *choosing the best single classifier* (Type II in Figure 1) 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. If we only rely on the system Type II, we totally dismisses the outcomes of a classifier trained by FFT feature set as the accuracy using DWT is shown to be superior.

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) [Dietterich, 1997; Valentini and Masulli, 2002]. There are two streams of model generation methods and Figure 2 presents the procedure of those various methods of model generation. 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 [Opitz, 1999] or by partitioning a data set into multiple sets [Breiman, 1996; Briem et al, 2001]. The heterogeneous method creates multiple classifiers built using different learning paradigms [Bahler and Navarro, 2000]. 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, constructed by FFT and by DWT, we will not apply such schemes as boosting or bagging [Breiman, 1996; Efron and Tibshirani, 1993] 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.

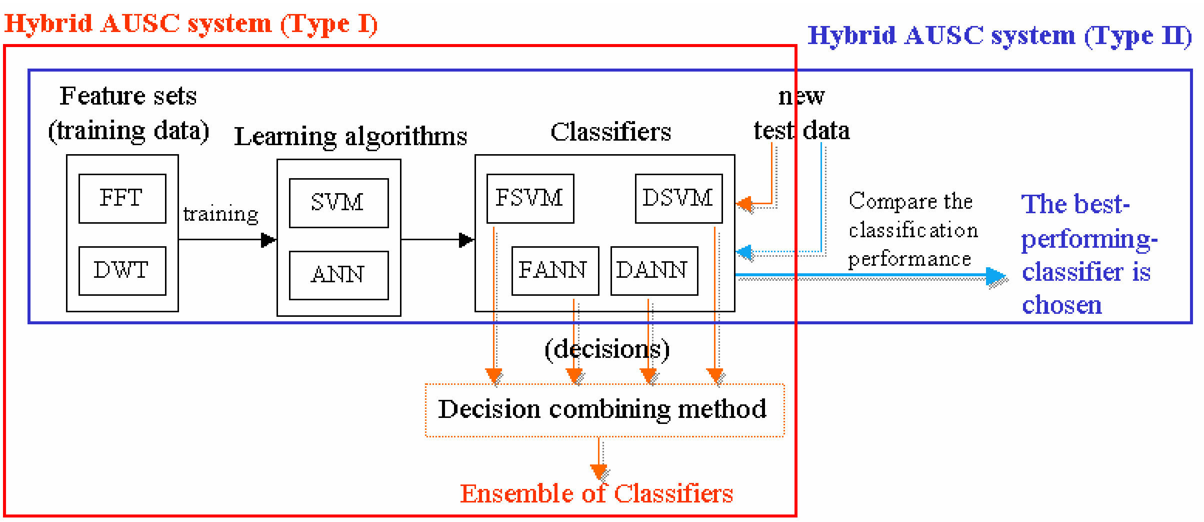

Figure 1. Constructing methods of two different AUSC systems using two feature extraction schemes (FFT and DWT) and two learning algorithms (SVM and ANN): The system Type I is constructed by combining decisions from multi-classifiers and the system Type II is by choosing the best classifiers among multi-classifiers.

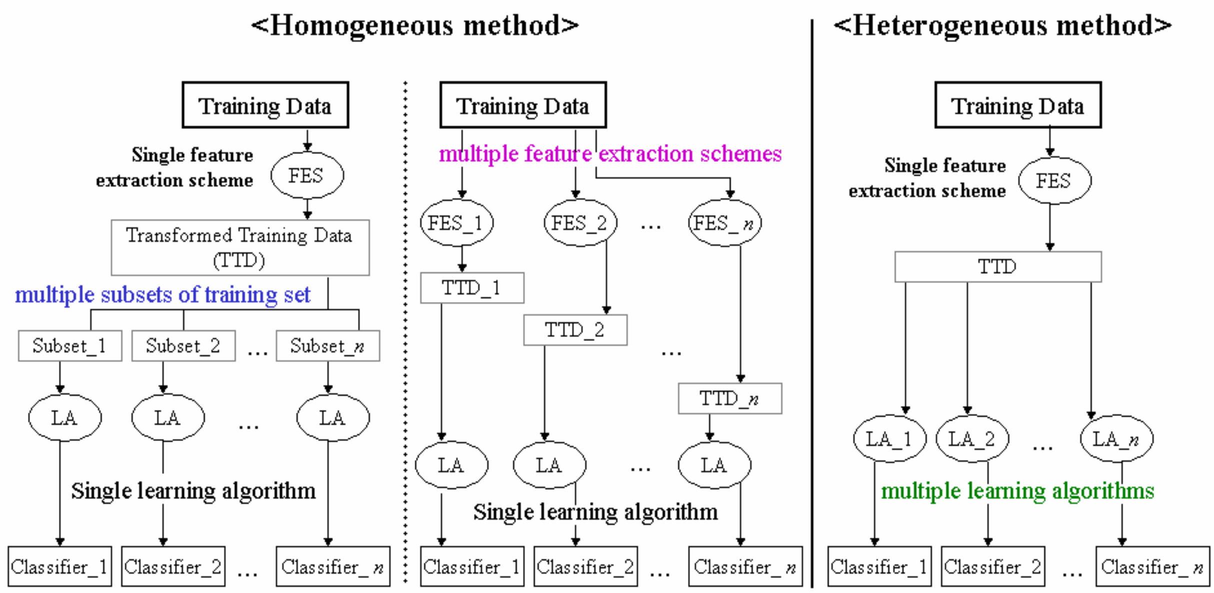

Figure 2. Methods of generating multi-classifiers for ensembles.

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 on the investigational results obtained becomes the basis for the construction of an integrated multi-classifier model using both feature extraction 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 also provides a motivation of 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 three classes; 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, fainted echoes caused by MC can confuse the signal echoes caused by CR and vice versa. Consequences of misclassification are catastrophic with enormous cost in downtime, consequential damage to associate equipment and potential injury to personnel [Cotterill and Perceval, 2001].

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 problem is that 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 [Cotterill and Perceval, 2001; Margrave et al, 1999]. This seems natural when dealing with ultrasonic signals 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, mainly because DWT provides effective signal compression and time-frequency presentation [Obaidat et al, 2001; Simone et al, 2001]. 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) [Polikar et al, 1998; Redouane et al, 2000; Spanner et al, 2000]. The first study analyzing feature extraction in more complex ultrasonic signals from shafts [Lee and Estivill-Castro, 2003] 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 [Lee and Estivill-Castro, 2004] 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 [Lee and Estivill-Castro, 2004]. 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 (homogeneous approach), or applying different learning algorithms to the same data (heterogeneous approach) creates a set of learned models. These two different approaches are summarized in Figure 2. 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

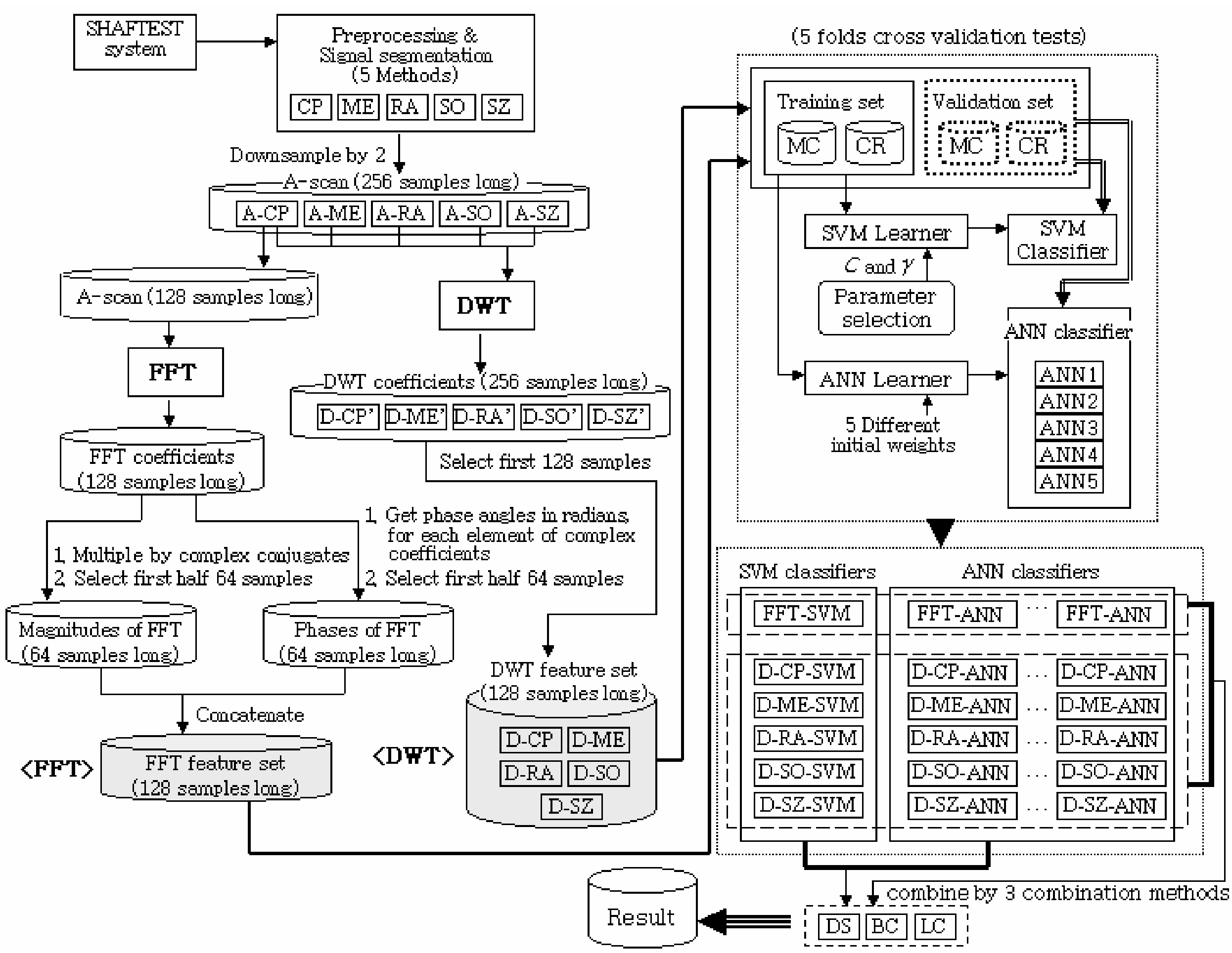

Figure 3. Overall procedure of our experiment.

The whole experimental setting consists of two steps as following:

1. We first map shaft inspection data into feature domains using two feature extraction schemes (FFT and DWT) and, using 5-folds cross-validation learning. We train them through SVM models and ANN models and 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 the classification accuracy results from the three combined models with the result using a single model

Figure 3 graphically summarizes the overall procedure of our experiment and the details are presented in following subsections.

3.1 Generation of multiple classifiers

We acquired A-scan signals from eight various shafts, ranging between 100mm to 1300mm in length using with the probe's frequency set to 2 MHz. Five of these shafts contained cracks and the rest were clean shafts. Mode-converted echoes were captured from these shafts, except for one 100mm calibration block that did not produce mode-converted echoes because of its geometry[1]. We extracted more than 1000 signal segments of interest from the whole ultrasonic A-scan signals. We also recorded whether the echoes were caused by Crack (CR), Mode-Conversion (MC) or Design Features (DF). As mentioned in Section 2.1, one of the issues of concern in ultrasonic shaft inspection is that the signal echoes caused by CR can be confused by fainted echoes caused by MC and vice versa. For our experiment, we chose each 60 signals randomly from CR data and MC data among the sample pool. Thus, in total a data set with 120 time-domain signals became the basis for further processing. We used *SHAFTEST™* [2] as a tool to capture these A-scans and build up an initial database of required echoes.

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) [Lee and Estivill-Castro, 2003]. 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 vectors of DWT coefficients as the DWT feature set. For our experiment, we applied Daubechies wavelets [Daubechies, 1988] 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 [Lee and Estivill-Castro, 2003]. 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). The selection of these network parameters are based on the experimental result of previous study [Cotterill and Perceval, 2001] [3]. 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

[1] For this size shaft (75mm diameter × 100mm length), there is no geometric possibility for the shear waves to reflect and convert back to compression waves before they are simply absorbed, because the shear waves are generated at 33°.

[2] It is a trade mark of CCI Pope.

[3] The selection of parameters of ANNs or the selection of kernel of SVM used in this experiment is based on several previous studies [Lee and Estivill-Castro, 2004; Lee and Estivill-Castro, 2005]. These parameters or kernel were employed for comparing a SVM-based classifier with an ANN-based classifier in order to identify which single classifier system is better for ultrasonic signal classification for shaft tests. These previous experiments provide a motivation to this paper.

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[3] kernels because they provide nonlinear mapping, require comparatively small numbers of hyper parameters and offer less numerical difficulties. RBF requires a penalty parameter C and kernel parameter γ. We used a grid-searching algorithm [Hsu et al., 2003] 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 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) [Suen and Lam, 2000; Xu et al, 1992], Distribution summation (DS) [Clark and Boswell, 1991] and Likelihood Combination (LC) [Ali and Pazzani, 1996].

In a Bayesian combination method, weights are es-

Learning algorithm	SVM														
Ensemble	C(FFT,DWT-CP)			C(FFT,DWT-ME)			C(FFT,DWT-RA)			C(FFT,DWT-SO)			C(FFT,DWT-SZ)		
ϕ_e	3.75			11.33			12.58			7.5			5		
Combining method	DS	BC	LC	DS	BC	LC	DS	BC	LC	DS	BC	LC	DS	BC	LC
(1)	0.006	0.006	0.006	-0.04	-0.018	-0.018	0.010	0.01	0.033	-0.069	-0.047	-0.047	-0.04	-0.04	-0.04
(2)	-0.095	-0.175	-0.175	0.018	-0.119	-0.119	0.058	0.035	-0.051	0.007	-0.182	-0.182	-0.022	-0.13	-0.13
(3)	-0.05	-0.077	-0.077	-0.037	-0.088	-0.063	0.014	0.013	-0.013	-0.048	-0.102	-0.102	-0.048	-0.088	-0.088
(4)	-0.103	-0.103	-0.103	-0.044	-0.022	-0.022	-0.119	-0.119	-0.096	0	0.022	0.022	-0.093	-0.093	-0.093
(5)	0.206	0.126	0.126	0.075	-0.062	-0.062	-0.017	-0.04	-0.126	0.08	-0.109	-0.109	0.109	0	0
(6)	0.014	-0.013	-0.013	-0.012	-0.063	-0.038	-0.077	-0.078	-0.103	0.027	-0.027	-0.027	-0.024	-0.063	-0.063

Learning algorithm	ANN														
Ensemble	C(FFT,DWT-CP)			C(FFT,DWT-ME)			C(FFT,DWT-RA)			C(FFT,DWT-SO)			C(FFT,DWT-SZ)		
ϕ_e	3.78			3.78			12.65			7.27			6.03		
Combining method	DS	BC	LC	DS	BC	LC	DS	BC	LC	DS	BC	LC	DS	BC	LC
(1)	-0.104	-0.178	-0.163	-0.11	-0.178	-0.174	-0.093	-0.091	-0.087	-0.111	-0.128	-0.124	-0.071	-0.135	-0.125
(2)	0.045	0	0.004	0.032	-0.013	-0.01	0.128	0.107	0.107	0.109	0.023	0.031	0.125	0.024	0.016
(3)	-0.064	-0.106	-0.096	-0.072	-0.111	-0.106	-0.01	-0.013	-0.01	-0.034	-0.073	-0.067	-0.015	-0.069	-0.067
(4)	-0.021	-0.095	-0.079	-0.043	-0.111	-0.107	-0.225	-0.223	-0.219	-0.202	-0.219	-0.215	-0.082	-0.146	-0.136
(5)	0.127	0.082	0.085	0.132	0.087	0.09	-0.069	-0.09	-0.09	0.08	-0.006	0.002	0.04	-0.061	-0.069
(6)	0.016	-0.026	-0.016	0.013	-0.026	-0.021	-0.172	-0.175	-0.173	-0.075	-0.113	-0.108	-0.057	-0.112	-0.109

Learning algorithm	FFT			DWT-CP			DWT-ME			DWT-RA			DWT-SZ			DWT-SO		
Ensemble	C(SVM,ANN)			C(SVM,ANN)			C(SVM,ANN)			C(SVM,ANN)			C(SVM,ANN)			C(SVM,ANN)		
ϕ_e	13.13			9.25			4.32			22.63			9.17			14		
Combining method	DS	BC	LC	DS	BC	LC	DS	BC	LC	DS	BC	LC	DS	BC	LC	DS	BC	LC
(1)	0.017	0.039	0.039	-0.064	-0.064	-0.064	-0.01	0.019	0.019	-0.029	0.009	0.009	0.009	0.013	0.013	-0.015	-0.01	-0.01
(2)	-0.056	-0.095	-0.095	0.093	0.048	0.048	-0.091	-0.118	-0.118	0.02	0.001	0	-0.014	-0.037	-0.012	0.043	0.015	0.015
(3)	-0.022	-0.025	-0.025	-0.002	-0.015	-0.015	-0.058	-0.052	-0.052	0.002	0.009	0.009	0	-0.008	0.003	0.005	-0.005	-0.005
(4)	-0.137	-0.115	-0.115	-0.025	-0.025	-0.025	-0.093	-0.065	-0.065	-0.187	-0.149	-0.149	-0.305	-0.301	-0.301	-0.128	-0.124	-0.124
(5)	0.056	0.017	0.017	-0.015	-0.059	-0.059	0.064	0.036	0.036	0.01	-0.01	-0.01	-0.004	-0.026	-0.001	-0.059	-0.088	-0.088
(6)	-0.06	-0.063	-0.063	-0.023	-0.037	-0.037	-0.038	-0.032	-0.032	-0.107	-0.099	-0.099	-0.154	-0.161	-0.151	-0.1	-0.11	-0.11

Table 1. 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:
(1) $e(C(\text{A,B}))$ - e(A) for overall (2) $e(C(\text{A,B}))$ - e(A) for classifying CR (3) $e(C(\text{A,B}))$ - e(A) for classifying MO
(4) $e(C(\text{A,B}))$ - e(B) for overall (5) $e(C(\text{A,B}))$ - e(B) for classifying CR (6) $e(C(\text{A,B}))$ - e(B) for classifying MO

tablished proportional to each individual classifier's past performance that are computed by its posterior probability using Bayes' theorem. In the Distribution summation method, distributions with each individual model are presented as a vector that 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, there 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 Table 1 and Figure 4. 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 Table 1 and Figure 4. All the combinations are carried out by three different combining methods (the 3 columns within each major column in Table 1 and Figure 4). The classification accuracy for each class is also recorded separately from the overall accuracy.

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. Table 1 shows the comparison of the decision error rate between combined models and single classifier models

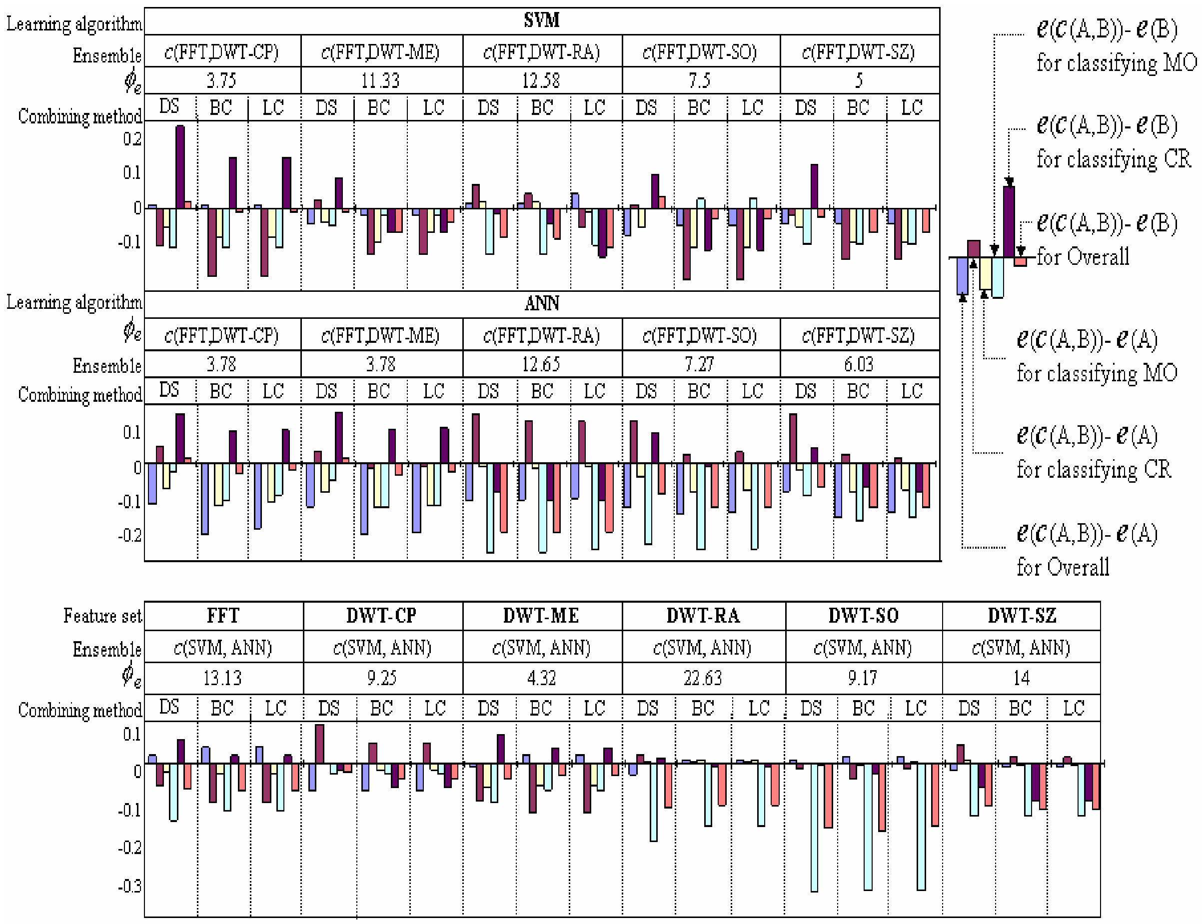

Figure 4. Bar-charts for the comparison of the combined models performance. C(A, B) indicates a combined classifier of two individual classifiers model A and model B. e(A) indicates the classification error rate of a classifier model A.

by calculating the difference of both error rates. All figures presented in Table 1 is average of result values from 5-fold cross-validation tests. Figure 4 graphically presents this comparison result, thus displays the amount of improvement in the classification performance in a form of bar charts. 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 that indicates the "fraction of correlated errors" [Ali and Pazzani, 1996] and is also listed in Table 1 and Figure 4. 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 (refer that most bars in Figure 4 point downwards).
- 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 Figure 4).
- 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 and 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 and 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 Table 1 or Figure 4, 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 methods and along the dimension of types of classifiers.

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.

The work presented in this paper can be continued in the following directions in the future.

- For the experimental comparison in this paper, we applied three generally known combining methods. We need to apply other methods such as *stacking* for combining multi-classifiers in order to extend our work.
- As mentioned before, we compared the classification performance of multi-classifiers based on two learning algorithms ANN and SVM, as the work on this paper is derived and motivated by previous studies [Lee and Estivill-Castro, 2003; Lee and Estivill-Castro, 2004; Lee and Estivill-Castro, 2005] in a single-classifier scenario with same industrial application. However, as a future work, we plan to apply various learning algorithms for generating multi-classifiers.
- We also plan to integrate other ultrasonic signal features not limited on FFT or DWT features for developing a more advanced hybrid ultrasonic signal classification system.

REFERENCES

Ali, K. & Pazzani, M. 1996, "Error reduction through Learning Multiple Descriptions". *Machine Learning* **24**(3), 173–202.

Bahler, D. & Navarro, L. 2000, "Methods for combining heterogeneous sets of classifiers". In *Pro-*

ceedings of the 17th Natl. Conf. on Artificial Intelligence, Workshop on New Research Problems for Machine Learning, AAAI Press.

Breiman, L. 1996, "Bagging Predictors". *Machine Learning* **24**(2), 123–140.

Briem, G., Benediktsson, J. & Sveinsson, J. 2001, "Boosting, Bagging and Consensus Based Classification of Multisource Remote Sensing Data". In *Proceedings of the 2nd Int. Workshop (MCS2001)*, J. Kitter & F. Roli, eds, LNCS 2096, Springer, 279–288.

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

Cotterill, G. & Perceval, J. 2001, "A New Approach to Ultrasonic Testing of Shafts". In *Proceedings of the 10th Asia-Pacific Conf. on Non-Destructive Testing (APCNDT).*

Daubechies, I. 1988, "Orthogonal bases of compactly supported wavelets". *Communications on Pure and Applied Mathematics* **41**, 909–996.

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

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

Hsu, C.-W., Chang, C.-C. & Lin, C.-J. 2003, "A practical guide to support vector classification". Available at http://www.csie.ntu.edu.tw/~cjlin/papers/guide.pdf.

Katragadda, G., Nair, S. & Singh, G. P. 1997, "Neuro-Fuzzy Systems in Ultrasonic Weld Evaluation". *Review of Progress in Quantitative Nondestructive Evaluation* **16**, 765–772.

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

Lee, K. & Estivill-Castro, V. 2004, "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*, CSREA Press, 848–854.

Lee, K. & Estivill-Castro, V. 2005, "Feature extraction and gating techniques for ultrasonic shaft signal classification". *Applied Soft Computing*, Elsevier (in press).

Margrave, F.W., Rigas, K., Bradley, D.A. & Barrocliffe, P. 1999, "The use of neural networks in ultrasonic flaw detection", *Measurement* **25**, 143-154.

Obaidat, M.S., Suhail, M.A. & Sadoun, B. 2001, "An intelligent simulation methodology to characterize defects in materials", *Information Sciences* **137**, 33–41.

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

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

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

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

Song, S.J., Kim, H.J. & Lee, H. 1999, "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.

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

Suen, C. & Lam, L. 2000, "Multiple classifier combination methodologies for different output levels". In *Proceedings of the 1st Workshop on Multiple Classifier Systems (MCS2000),* LNCS 1857, Springer, 52–66.

Valentini, G. & Masulli, F. 2002, "Ensembles of Learning Machines", R. Tagliaferri & M. Marinaro, eds, *Neural Nets WIRN vietri-2002*, LNCS 2486, Springer-Verlag, 3–19.

Xu, L., Krzyzak, C. & Suen, C. 1992, "Methods of combining multiple classifiers and their applications to handwritting recognition". *IEEE Transactions on Systems, Man and Cybernetics* **22**(3), 418–435.

BIOGRAPHY

KYUNGMI LEE is currently working toward a Ph.D. degree in Computer Science at Griffith University, Australia. She received a B.Sc. in Computer Science from Pusan National University in Korea. Her research interests include pattern recognition and classification, signal processing, machine learning, artificial intelligence, data mining.

VLADIMIR ESTIVILL-CASTRO obtained his PhD in 1991 from the epartment of Computer Science, University of Waterloo, in Canada. After a year as Assistant Professor at York University, Toronto, Canada, he joined consulting (in LANIA, an industry sponsored research laboratory) as a project leader and senior researcher. In 1996, he rejoined academic life at Queensland University of Technology in Australia. He has demonstrated leadership in a research environment. He developed the Machine Intelligence and Pattern Analysis Lab (Mi-PAL) at Griffith University and engaged many of the existing research staff in projects in the areas of robotics, image processing and machine learning under the research agenda of the lab. His commitment to fostering a research culture and his track record have resulted in an invitation to chair the program committee for the Australian Computer Science Conference. Estivill-Castro is currently Professor at Griffith University where he has established multi-disciplinary research collaborations. His diverse research interests include Security through Data Mining, Privacy, Machine Learning, Knowledge Discovery and Data Mining, Spatial Databases and Geographical Information Systems, Computational Geometry and Algorithms, Computer Vision and Robotics. He is internationally recognized for his work on Spatial Data Mining, Clustering and Privacy in Data Mining. He is a member of the editorial board of the Int. Journal of Data Warehousing and Mining (IJDWM).