# ESTIMATED ACCURACY OF CLASSIFICATION OF DEFECTS DETECTED IN WELDED JOINTS BY RADIOGRAPHIC TESTS

M.H.S. Siqueira[1], R. R. da Silva[1], M. P. V. de Souza[1], J. M. A. Rebello[1]
and L. P. Calôba[2], D. Mery[3]
[1]Department of Metallurgical and Materials Engineering, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro RJ Bazil, [2]Department of Electrical Engineering, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro RJ Brazil; [3]Pontificia Universidad Catolica de Chile, Escuela de Ingeniería - DCC(143),Departamento de Ciencia de la Computacion, Casilla 306, Santiago 22Chile,

**Abstract**: This work is a study to estimate the accuracy of classification of the main classes of weld defects detected by radiography test, such as: undercut, lack of penetration, porosity, slag inclusion, crack or lack of fusion. To carry out this work non-linear pattern classifiers were developed, using neural networks, and the largest number of radiographic patterns as possible was used as well as statistical inference techniques of random selection of samples with and without repositioning (*bootstrap)* in order to estimate the accuracy of the classification. The results pointed to an estimated accuracy of around 80% for the classes of defects analyzed.

**Introduction**: The non-destructive radiographic method of inspection has been widely used over the decades to evaluate the integrity of material and equipment in a wide range of industries. In the specific case of radiographs of welded materials, the research for the development of an automatic or semiautomatic system of analysis of radiographs of welded joints has grown considerably in the last years and especially in the last 10 to 15 years [1-10]. The latest publications are mainly concerned with this last stage of defect classification where the authors normally use techniques of neural networks, Fuzzy logic and hybrid systems to implement classification patterns. As in all cases the number of samples used to estimate the parameters of the classifiers (in the case of neural networks: their synapse vectors and bias) is small, making it extremely difficult to divide the training and test sets with a number of statistically significant samples to estimate the accuracy of the classification adequately with data not used in the training of the classifiers. The question arises: What is the true accuracy of weld defect classification? There are few results published relating to this matter, but notably among them is one of the last publications of Liao [7].

The aim of this present work is to present the methodologies used and the results obtained in this study to estimate the classification accuracy of the main classes of weld defects, such as: undercut (UC), lack of penetration (LP), porosity (PO), slag inclusion (SI), crack (CR) and lack of fusion (LF). The non-linear classifiers were implemented using artificial neural networks. The largest possible number of radiographic patterns and statistical interference techniques of random selection of samples with and without repositioning (*bootstrap)* was used to estimate the accuracy of the classifiers. The results are presented in tables with estimated accuracies for each classification defect studied. It should be pointed out that this work is the continuation of previous works already published, which will be commented on briefly [11-14].

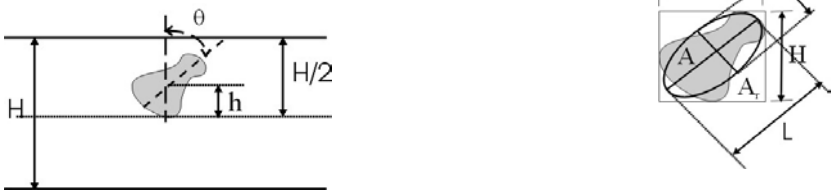
- ***Radiographic Patterns and Film Digitalization***

In previous works [11-14] only one collection of radiograph weld defect patterns from IIW (*International Institute of Welding*) containing around 86 films was used. In this work, two new collections of film patterns were used to significantly increase the amount of data for estimating the accuracy of the classifier. One of the new collections of patterns was from IIW (IIW 1290-year 95) containing 67 radiographs that were 225 mm by 50 mm in size. These radiographs were scanned in 2000 dpi with a Microtek 9800 XL scanner, which permitted an average resolution of 12.5 μm and 8 bits of grey level. The scanner had a nominal maximum density limit of 3.7 O.D., which in practice, is a low limit for the digitalization of industrial radiographs of high-density.

To solve this problem, a light box was projected with an illumination potential greater than the transparency adapter of the scanner. The third collection of radiographic patterns was supplied by BAM (Federal Institute of Materials Research and Testing - Berlin) containing 67 radiographs scanned in a LS 85 SDR scanner from Lumisys/Kodak with a maximum density of 4.1 O.D. These radiographs were scanned in 12 bits of grey level and afterwards converted to 8 bits, without losing defect information. The digitalized spatial resolution of was 630 dpi (40.3μm) and the identification and classification of the defects followed the norm EN 26 520.

- ***Feature Extraction***

After digitalization, all the films were pre-processed following the same procedures already described in [11]. Using a larger quantity of radiographs, it was possible to extract data for lack of fusion (LF) and crack (CR) defects, which earlier [11-14] had not been done due to an insufficient number of radiographs containing these class defects. In this case, 7 geometric features of the defects were extracted (Figure 1a and b): Position (p=h/H): h is the distance from the centroid of the defect to the center of the weld bead that, in this case, was determined measuring a half of H. H was used to normalize the feature in relation to variations of thickness of the bead that occur in welding joints (figure 1a); Ratio of aspect (a = L/e): L is the larger axis of the ellipse of the area equivalent to the defect and e is the smaller axis (figure 1b); Ratio between the width and the Area (e/A): the ratio between the smaller axis of the ellipse and the area equivalent to the defect and the area (figure 1b); Roundness: measures the ratio $p^2/4\pi A$, where p is the perimeter and A the area of the defect; Angle (θ): the angle between the larger axis of the defect and the vertical (Figure 1a); Area/rectangle ($A/A_r$): ratio between the area of the defect and the smaller rectangle that encloses the defect (Figure 1b); Rectangle or "box" W/H: ratio between the width and the height of the smaller rectangle that encloses the defect (Figure 1b).




(a) (b)

***FIGURE 1***: illustration of the seven features extracted.

After features extraction, the set of data contained a total of 646 samples, with the distribution of classes in this way: LF (56 samples), LP(56 samples), UC (174 samples), CR ( 27 samples), SI (137 samples) and PO (196 samples) .The non-linear classifiers were implemented using a neural network of two layers with 6 or 5 neurons in the output layer in function of the number of classes. The number of neurons used in the intermediate layer was optimized in such a way as to permit improved accuracy with the test data, as was described in [14]. Some tests were carried out in terms of training parameters of the network and the best result (fastest convergence) was found when the moment (β=0.9) and α (training rate) variable were used [15]. All the training was interrupted at 3,000 epochs due to the stabilization of the error curve, as after this there was no significant decrease in the error of learning.

- ***Estimated Accuracy of Classifiers***

There are various techniques to estimate the accuracy of a classifier, but basically there are three that are the most used: simple random selection of samples, cross validation that really presents diverse implementations [16], and the *bootstrap* technique [17-18]. It is not really possible to confirm whether one method is better than the other for any specific pattern classification system. The choice of one of these techniques will depend on the quantity of samples available and the

specific classification to be made. To calculate the classification accuracy of weld defects, two techniques of random selection of samples were applied: 1) Random selection without repositioning of 80.0% of the total set of samples for the construction of the training set. The samples not chosen for training were used to set up the test set. A total of ten pairs of training and test sets were made for each defect classification. 2) Random selection with repositioning using the *bootstrap* technique. A set of *bootstrap* samples (size *n*), following Efron's own definition [17], is made up of $x_1^*, x_2^*, \Lambda\ , x_n^*$ samples, obtained in a random way and with repositioning, from an original set of data $x_1, x_2, \Lambda\ , x_n$ (also size $n$). In this manner, it is possible that some samples appear 1, 2, 3 or *n* times or no times [17]. With this technique, the classifier implemented using the i[th] set of training is tested with samples that were not used in the make up of this set, resulting in an accuracy estimator of $\hat{\theta}_i$ (for test data). This is repeated *b times*. One model of accuracy estimation $\hat{\theta}_B$ of pattern classifiers frequently used is defined by the equation:

$$\hat{\theta}_B = \frac{1}{b}\sum_{i=1}^{b}\left(0.632\hat{\theta}_i + 0.368\hat{\theta}_c\right) \quad (1),$$

where $\hat{\theta}_c$ is the apparent accuracy (calculated with the test samples) [16-17]. For this accuracy estimator, some authors [16] criticize the use of excessive weight for the training data (0.368), which could result in a very optimistic estimated accuracy (high *Bias*). In this case, an adaptation of the equation 1 is proposed in the equation 2 [17, 18.]:

$$\hat{\theta}_B{}^* = \frac{1}{b}\sum_{i=1}^{b}\left(\hat{\omega}\hat{\theta}_i + (1-\hat{\omega})\hat{\theta}_c\right) \quad (2),$$

Where the weight $\hat{\omega}$ varies between 0.632 and 1. This estimator $\hat{\theta}_B{}^*$ proportions a more adequate relation between the pessimistic estimator $\hat{\theta}_i$ and the optimistic estimator $\hat{\theta}_c$ [16-18].

In terms of weld defect classification accuracy, perhaps the only work is that of Wang [7]. Wang [7] worked with a set of 147 samples, of which he selected 27 samples for validation and 12 for test. Wang used the bootstrap technique with a selection of five samples for estimation of accuracy of classification of defects: crack, pore (gas hole), hydrogen inclusion, lack of fusion, lack of penetration and porosity.

**Results**: Tables from 1 to 7 present the results obtained in this work.

***TABLE 1***: Non-linear Classifiers Accuracy Estimators for Each Condition of Classification.

*Percentage Results obtained only with the Test Sets*

| | Sets | RESULTS (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Maximum | | Minimum | | Average | | Variation | |
| | | WR | R | WR | R | WR | R | WR | R |
| 6 Classes | Without duplication | 72.9* | 75.2* | 65.9 | 67.5 | 69.0 | 71.6 | 2.2 | 1.3 |
| | With duplication | 88.5* | 89.4* | 69.4 | 71.0 | 76.7 | 79.8 | 5.9 | 5.6 |
| 4 Classes | Without duplication | 83.0 | 83.9 | 74.1 | 75.0 | 77.8 | 79.0 | 2.7 | 2.7 |
| | With duplication | 84.0 | 85.3 | 62.8 | 63.5 | 81.4 | 82.5 | 7.3 | 7.3 |
| 5 Classes | Without duplication | 93.1 | 94.1 | 63.3 | 63.3 | 84.0 | 85.0 | 8.7 | 9.3 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **With duplication** | 97.0 | 97.5 | 77.0 | 79.1 | 85.3 | 86.4 | 6.4 | 5.9 |

***** Cells that correspond to the confusion tables presented below for the test sets.

***TABLE 2***: Confusion Table (%).

*Non-linear Classifier (without reclassification)- Without Duplication–Test Data*

| | **LF** | **LP** | **UC** | **CR** | **SI** | **PO** | ***More than one*** | ***None*** |
|---|---|---|---|---|---|---|---|---|
| **LF** | 6/46.1 | 2/15.4 | 2/15.4 | 0 | 0 | 0 | 0 | 3/23.1 |
| **LP** | 0 | 8/72.7 | 2/18.2 | 1/9.1 | 0 | 0 | 0 | 0 |
| **UC** | 1/2.7 | 1/2.7 | 33/94.3 | 0 | 0 | 0 | 0 | 0 |
| **CR** | 1/20.0 | 0 | 2/40.0 | 1/20.0 | 1/20.0 | 0 | 0 | 0 |
| **SI** | 0 | 0 | 4/17.4 | 1/4.4 | 7/30.4 | 6/26.1 | 0 | 5/21.7 |
| **PO** | 0 | 0 | 3/7.1 | 0 | 0 | 39/92.9 | 0 | 0 |

***TABLE 3***: Confusion Table (%).

*Non-linear Classifier (with reclassification) - Without Duplication–Test Data*

| | **LF** | **LP** | **UC** | **CR** | **SI** | **PO** |
|---|---|---|---|---|---|---|
| **LF** | 7/53.8 | 2/15.4 | 2/15.4 | 0 | 2/15.4 | 0 |
| **LP** | 0 | 8/72.7 | 2/18.2 | 1/9.1 | 0 | 0 |
| **UC** | 1/2.7 | 1/2.7 | 33/94.3 | 0 | 0 | 0 |
| **CR** | 1/20.0 | 0 | 2/40.0 | 1/20.0 | 1/20.0 | 0 |
| **SI** | 1/4.4 | 0 | 4/17.4 | 2/8.7 | 9/39.1 | 7/30.4 |
| **PO** | 0 | 0 | 3/7.1 | 0 | 0 | 39/92.9 |

***TABLE 4***: Confusion Table (%).

*Non-linear Classifier (without reclassification) - With Duplication Test Data*

| | **LF** | **LP** | **UC** | **CR** | **SI** | **PO** | ***More than one*** | ***None*** |
|---|---|---|---|---|---|---|---|---|
| **LF** | 23/62.2 | 0 | 0 | 2/5.4 | 6/16.2 | 0 | 2/5.4 | 4/10.8 |
| **LP** | 0 | 43/95.6 | 0 | 2/4.4 | 0 | 0 | 0 | 0 |
| **UC** | 2/4.7 | 0 | 40/93.0 | 0 | 1/2.3 | 0 | 0 | 0 |
| **CR** | 0 | 0 | 0 | 38/100 | 0 | 0 | 0 | 0 |
| **SI** | 0 | 0 | 0 | 2/6.9 | 22/75.9 | 4/13.8 | 0 | 1/3.4 |
| **PO** | 0 | 0 | 0 | 0 | 1/2.3 | 42/97.7 | 0 | 0 |

***TABLE 5***: Confusion Table (%).

*Non-linear Classifier and Non-Hierarchy (with reclassification)- With Duplication Test Data*

| | **LF** | **LP** | **UC** | **CR** | **SI** | **PO** |
|---|---|---|---|---|---|---|
| **LF** | 24/64.9 | 0 | 2/5.4 | 4/10.8 | 7/18.9 | 0 |
| **LP** | 0 | 43/95.6 | 0 | 2/4.4 | 0 | 0 |
| **UC** | 2/4.7 | 0 | 40/93.0 | 0 | 1/2.3 | 0 |
| **CR** | 0 | 0 | 0 | 38/100 | 0 | 0 |
| **SI** | 0 | 0 | 0 | 2/6.9 | 23/79.3 | 4/13.8 |
| **PO** | 0 | 0 | 0 | 0 | 1/2.3 | 42/97.7 |

***TABLE 6***: Accuracy Estimators for *Bootstrap* for the Non-linear Classifier with 6 classes, each class with its original population of samples.

| **Sample Sets** | **Classification Criterion** | ***Bootstrap 0.632*** $\hat{\theta}_B$ | | | | ***Bootstrap 0.700*** $\hat{\theta}_B^*$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Max.** | **Min.** | **Mean** | **SD** | **Max.** | **Min.** | **Mean** | **SD** |
| **50** | **WR (%)** | 81.4 | 64.6 | 76.0 | 3.5 | 80.4 | 62.5 | 74.9 | 3.9 |
| | **R (%)** | 83.9 | 69.0 | 78.9 | 3.0 | 83.2 | 67.1 | 77.8 | 3.3 |
| **20** | **WR (%)** | 81.2 | 73.4 | 77.8 | 2.1 | 80.4 | 72.5 | 77.0 | 2.2 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **R (%)** | 83.9 | 74.8 | 80.6 | 2.1 | 83.2 | 74.0 | 79.8 | 2.2 |

***TABLE 7***: Accuracy Estimators for *Bootstrap* for the Non-linear Classifier with five classes, each class with its original population of samples.

| **Sample Sets** | **Classification Criterion** | ***Bootstrap 0.632*** $\hat{\theta}_B$ | | | | ***Bootstrap 0.700*** $\hat{\theta}_B^*$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Max.** | **Min.** | **Mean** | **SD** | **Max.** | **Min.** | **Mean** | **SD** |
| **20** | **SR (%)** | 93.3 | 79.8 | 86.7 | 4.1 | 92.6 | 78.7 | 85.4 | 4.4 |
| | **R (%)** | 93.6 | 80.5 | 88.0 | 3.8 | 93.0 | 79.5 | 86.9 | 4.1 |

Discussion:

- ***Non-linear Classifiers***

Different classes contained a distinct number of samples, and so in order not to favor training networks of more favorable classes, as for example porosity with 196 samples, a providential solution - the random doubling of classes with a lower number of samples - was used until equalizing with porosity, creating 196 samples for all classes.

Varying the number of neurons in the intermediary layer and checking the percentage of correctness and error after training with 3,000 epochs, the number of neurons which resulted in the best results was 16/17 neurons for the training data. However, randomly dividing the set of data into training and test pairs, and in this case without duplication so that samples used for training would not used in the test, the best test results were obtained with 8 neurons in the intermediary layer. Due to these initial results, trainings and tests made with a non-linear classifier were all carried out with 8 neurons in the intermediary layer, as a way to control the possibility of overtraining taking place.

Initially, some variations of the network parameters were tested, such as: the rate of the training variable ($\alpha$) with moment ($\beta$=0,9); rate of training fixed in 0.1 and without moment; fixed-rate in 0.1 and moment fixed in 0.9. The initialization of the synapses and *bias* used the Nguyn-Widrow [19] method. All these training variations resulted in a convergence for the same range of error. However, the parameters adjusted for variable $\alpha$ and fixed moment ($\beta$=0,9) always resulted in a faster convergence of the training error. After 3,000 epochs it was noted that there was normally error stabilization, and so this number of epochs was chosen for training with non-linear classifiers.

- ***Non-linear Classifier Accuracy Estimators***

To estimate the true accuracy of these classifiers, ten pairs of training and test sets were prepared. Table 1 shows the results obtained for the ten pairs of sets prepared for each condition evaluated: with and without duplication. The symbols WR and R represent - without reclassification and with reclassification, respectively. Summarizing, since detailed explanations may be found in the previous publication [11], the criteria of reclassification considers the output of the network indicative of the class the output that has the highest value as, even if all the outputs are negative, or if there is more than one positive output neuron. Besides studying the classification of the 6 classes of defects, a classifier for 4 classes was also tested, removing crack and lack of fusion, which had not been studied previously [11, 12, 14] in order to compare results. Table 1 also shows the results obtained for 5 classes, when the class SI was excluded from the study, since this class presented the largest number of classification problems [11,12]. In Table 1, the average indicates the estimated accuracy for each selection of 10 pairs of test sets. For 6 classes, with the data having the original quantity of samples for each class (without duplication), the estimated accuracy was 71.6% with reclassification. Although this percentage is below the indices obtained with the training sets, which oscillated in the 85.0-90.0% range, no overtraining occurred,

because the number of neurons in the intermediary layer was purposely optimized to avoid this problem. The maximum index of correctness for the test samples in this case was 72.9% without reclassification and 75.2% with reclassification. To check the indices of correctness of each class for this same set, in such a way as to show what occurred separately in each class in this classification (6 classes and set without duplication), the confusion tables of classes 2 (without reclassification) and 3 (with reclassification) were built. As the total number of samples without duplication were 646, divided in about 80.0 % for training and 20.0% for test, the training set contained 517 samples leaving 129 for the test set.

Tables 2 and 3 show the level of correctness of class UC to be 94.3%, PO 92.9% and LP 72.7%, these being the classes with the greatest accuracy. The class that resulted in the lowest index of correctness, with the test data was CR, but only 4 (four) of a total of 27 samples contained in this class, were tested, however the result was of low statistically significance. The class SI presented the greatest confusion among the other classes, confirming results shown in other works (11-12). Returning to the analysis of estimated accuracy for this classification situation, the 71.6% obtained may be considered satisfactory, bearing in mind the quantity of classes now analysed, and principally the presence of the class SI that in this new set of samples presented a large range of confusion with the other classes, justified by the wide variation of structures used this time to make up the sample set of this class (including linear and non-linear).

There has not been much research to determine the reliability of radiographic reports using the conventional method carried out by qualified inspectors. However, recently, Fucsok [20,21] set up ROC curves [22] to estimate the probability of correct detection and interpretation of defects in radiographic films by inspectors. He arrived at a result of 68% of correct indications for all types of defects and 64% for defects that exceeded the limits of acceptability following the standard (EN 12515:1998) [20,21], which is a value below the 71.6% obtained by the accuracy estimators. However, the number of defect classes in the films used by Fucsok [20, 21] contained classes that had not been evaluated by these classifiers, such as: hydrogen inclusion, tungsten inclusion, misalignment, concavity, etc., but the classes studied in this work are the most common in terms of welding defects and those that provoked the greatest errors of interpretation.

Within the same classification problem involving 6 classes, 10 pairs of training and test sets were also made up but with duplication of samples to equalise the number of samples contained in each class making sure that each class would contain 196 samples, the same number of samples as class PO. In Table 1, it was seen that the estimated accuracy was 76.7% without reclassification, and 79.8% with reclassification, indices superior to the previous case. This increase is justified by the fact that the classes with less distribution of samples, such as CR (27), LF (56) and LP (56) used identical samples as much for training as for testing, due to duplication. However, the class PO with 196 samples not duplicated, was trained and tested with completely different sets. The class UC with 174 samples could have used some of the same samples for training and testing, although this would have been unlikely since only 22 samples were duplicated in this class. The same occurred for class SI, except in this case 59 samples, out of a total of 137, were randomly duplicated, consequently a larger proportion of identical samples would have been used for both test and training.

The confusion Tables 4 and 5 show the results obtained for the test set with the maximum accuracy (maximum of the table), proving that when there is a duplication of samples, the indices of correctness increase for practically all classes, the exception being the UC class which now had 93.0% of correctness (Table 4), compared with 94.3% previously reached, but with an insignificant difference compared to the percentage increase of the other classes, which reached a maximum index of correctness of 88.5% without reclassification (WR) and 89.4% with reclassification (R). Returning to Table 1, where there are also the accuracy estimations obtained for test sets containing only the 4 classes initially studied in the previous works [11,12]: LP, UC, SI and PO. For sets without duplication, the average results were 77.8% (WR) and 79.0% (R). However, with duplication of the samples, the results reached levels of 81.4% (WR) and 82.5%

(R). These results are extremely satisfactory, even being below the 100% that was obtained for the first set of data (estimated by the probability calculation [11,12]), because the present situation is much more representative in terms of the number of samples of radiographic images used. The largest index of correctness for a set with 4 classes was 85.3% (R) and 84.0% (WR) with the samples duplicated. For these sets, it's worth noting that there is little percentage difference of the estimators between the duplicated and non-duplicated sets. This lack of difference is because among these 4 classes, only the class LP had few samples (56) compared to the others, and so the duplication of the samples did not have much of an effect on the increase of the accuracy of the sets.

From the results presented, it can be seen that class SI presented the largest index of confusion with the other classes, principally with class PO. And so, the estimation of the accuracy of the non-linear classifier for a set not containing the class SI (only containing the five other classes) was made. In Table 1, note that the estimated accuracy is 84.0% (WR) and 85.0% (R) for the set without duplication while with duplication 85.3% (WR) and 86.4% (R) were obtained. These indices of estimated accuracy are above the indices obtained with 6 and 4 classes, when the class SI was present, a result that supports the fact that class SI contains the largest indices of classification error. Another important factor to be highlighted from this result is that there is little difference between the results of sets without duplication and those with duplication, which goes to show that the classifier is well-developed even for samples distributed unequally among the classes.

Making a general analysis of the results reached for the accuracy estimators with a random selection of 10 pairs of sets, the indices obtained can be considered satisfactory, since we know that the classifier for 6 classes of defect which are so varied in structure, especially LF and SI, and using a reduced number of features as in this work, is extremely difficult. Other authors such as Wang [7], although they obtained indices of slightly greater accuracy, normally using 10/12 features for classification of classes with structures and aspects of a more typical contrast, with hydrogen inclusion for example, and are therefore easier to classify. The accuracy study of these 6 classes (without a shadow of doubt the most important) studied with a set of 646 samples is innovating in comparison to bibliographic references cited in this area, especially in relation to works of defect classification involving geometric features of easy extraction. It should be noted also, that the greater number of features used, the greater are the problems of classifier generalisation because there will be a larger number of parameters (synapse weights and *bias*) to be estimated [15, 22]. Consequently it is preferable to use the smallest quantity possible of features, using only those that are the most relevant, which evidently permit a better generalisation of the classifier. Using the bootstrap technique to generate samples in a random way and with repositioning, 50 sets of training and 50 sets of tests were created to estimate the accuracy of the non-linear classifier. Table 6 shows the results obtained for these accuracy estimators. With 50 sets made up, the estimation accuracy, employing equation (1) is 76.0% without reclassification and 78.9% with reclassification. However, it should be noted that there were training and test pairs with extremely different results, differences above 15.0%. However, these sets were probably in a situation of overtraining despite the control of the number of neurons in the intermediary layer of the classifier, which greatly reduces this from happening [15, 22]. In this case, only the training and test pairs that had an accuracy difference of less than 15.0% (a value considered acceptable for the difference between training and tests) were considered, resulting in a total of 20 pairs. For these sets, the estimated accuracy was 77.8% (WR) and 80.6% (R) with the estimator 0.632 [17]. The estimator 0.632 is considered optimist since it weighted the training data results [16] excessively, and so the estimator using the generalised equation (2) was also calculated weighting the test data with a factor of 0.7 and training with 0.3 (Table 7), which obviously provoked a slight drop in the values of the accuracy estimator. It should be pointed out that these results are for sets with 6 classes of defects. To finalise, the same accuracy estimator was calculated for the set containing only 5 classes. In this

case, an accuracy rate of 88.0% was reached with reclassification for the estimator 0.632. It must be pointed out that all these estimators were calculated using the original set of samples without duplication, although each set of *bootstrap* could contain "duplicated" samples.

The authors of this work are certain that there is still much to be studied and researched in the sense of obtaining even more precise accuracy estimators for the classification of weld defects, principally by using larger numbers of samples for the less predominant classes such as lack of fusion and cracks, although no less important than the other classes.

**Conclusions:** This work, concludes that the classes UC, LF and PO present high accuracy of classification as much for training samples, as for sample sets only used for tests. However, the classes LF, CR and SI resulted in low accuracy indices. Within these classes, SI is the class, which presented the largest margin of confusion with the other classes, mainly PO. The crack class, the class with the least number of samples, presented a low accuracy index for the test samples. Continuation of this work should direct itself towards making a larger number of samples of these classes available, as well as the investigation and usage of other relevant features to increase the accuracy index for the class SI.

The indices of accuracy obtained in this work, for the conditions studied, certainly nears the true accuracy or the real classification of the main classes of weld defects due to the greater number of samples used when compared with previous works, also because of the quantity of sets made for the development and testing of the classifiers, making a relevant contribution to this area of research.

**Acknowledgements**: Authors wish to acknowledge CNPq (The National Council for Scientific and Technological Development, FAPERJ (Research Foundation from Rio de Janeiro) and ANP (Brazilian Agency for Petroleum) for financial support and scholarships and also to International Institute of Welding and BAM (*Bundesanstalt für Materialforschung und-prüfung* -Berlin) for permission given to publicate the present work using the radiographic patterns.

**References:**

1. Aoki K, Suga Y. Application of Artificial Neural Network to Discrimination of Defect Type Automatic Radiographic Testing of Welds. In: ISI International 1999; 39(10): 1081-1087.

2. Kato Y, Okumura T, Matsui S, et al. Development of an Automatic Weld Defect Identification System for Radiographic Testing. Welding in the Word 1992; 30 (7/8):182-188.

3. Liao T W, Ni J. An Automated Radiographic NDT System for Weld Inspection: Part I – Weld Extraction. NDT&E International 1996; 29(3):157-162.

4. Liao T W, Li Y. An Automated Radiographic NDT System for Weld Inspection: Part II – Flaw Detection. NDT&E International 1998. 31(3):183-192.

5. Liao T W, Li D, Li Y. Detection of Welding Flaws from Radiographic Images with Fuzzy Clustering Methods. Fuzzy sets and Systems 1999; 108:145-158.

6. Sankaran V, Chartrand B, Millard D, et al. Automated Inspection of Solder Joints-A neural Network Approach. European Journal Mechanical Engineering; 43(3):129-153.

7. Wang, G, Liao, T. W. Automatic Identification of Different Types of Welding Defects in Radiographic Images. NDT&E International 2002; 35: 519-528.

8. Lashkia V. Defect detection in X-ray images using fuzzy reasoning. Image and Vision Computing 2001; 19: 261-269.

9. Shafeek, H.I, Gadelmawla, E., Abdel-Shafy, A.A, Elewa, I.M. Automatic Inspection of Gas Pipeline Welding Defects Using an Expert System. NDT&E International, 2004.

10. Shafeek, H.I, Gadelmawla, E., Abdel-Shafy, A.A, Elewa, I.M. Assessment of Welding Defects for Gas Pipeline Radiographs Using Computer Vision. NDT&E International, 2004.

11. Silva R R, Siqueira M H. S, Calôba L P, et al. Radiographics Pattern Recognition of Welding Defects using Linear Classifiers. Insight 2001; 43(10): 669-674.

12. Silva R R., Siqueira M H. S, Calôba L P, et al. Contribution to the Development of a Radiographic Inspection Automated System. In: 8th European Conference on Non Destructive Testing. June 17-21, 2002.
13. Silva R R, Siqueira M H. S, Calôba L P, Rebello, J.M.A. Evaluation of the Revelant Characteristic Parameters of Welding Defects and Probability of Correct Classification using Linear Classifiers. Insight 2002; 44(10): 616-622.
14. Silva R R, Calôba L P, Siqueira M H.S, Rebello, J.M.A. Pattern Recognition of Weld Defects Detected by Radiographic Test. NDT&E International, 2004.
15. Haykin S. Neural Networks – A Comprehensive Foundation. Macmillian College Publishing. Inc. 1994.
16. Diamantidis, N.A., Karlis, D., Giakoumakis, E.A. Unsupervised Stratification of Cross-Validation for Accuracy Estimation. Artificial Intelligence 2000. 116: 1-16.
17. Efron, B., Tibshirani, R. J. An Introduction to the Bootstrap. New York, Chapman & Hall/CRC, 1993.
18. Efron, B., Tibshirani. Cross-Validation and the Bootstrap: Estimating the Error Rate of the Prediction Rule. Technical Report 477, Stanford University 1995. http://utstat.toronto.edu/tibs/research.html.
19. Beale, M. Neural Network Toolbox for Use with Matlab User's Guide Version 4. The MathWorks, 2001.
20. Fucsok, F., Scharmak, M. Human Factors: The NDE Realibility of Routine Radiographic Film Evaluation. In: 15th World Conference on Non-Destructive Testing. Rome, October 15-21 2000.
21. Fucsok, F., Muller, C., Scharmak, M. Reliability of Routine Radiographic Film Evaluation – An Extended ROC Study of the Human Factor. In: 8th European Conference on Non Destructive Testing. Barcelona, June 17-21 2002.
22. Duda, R.O., Hart, P.E., Stork, D.G. Pattern Classification. 2nd edition, U.S.A., John Wiley & Sons 2001.