 DOI: 10.1784/insi.2007.49.10.603

Accuracy estimation of detection of casting defects in X-ray images using some statistical techniques

R R da Silva and D Mery

Casting is one of the most important processes in the manufacture of parts for various kinds of industries, among which the automotive industry stands out. Like every manufacturing process, there is the possibility of the occurrence of defects in the materials from which the parts are made, as well as of the appearance of faults during their operation. One of the most important tools for verifying the integrity of cast parts is radioscopy. This paper presents pattern recognition methodologies in radioscopic images of cast automotive parts for the detection of defects. Image processing techniques were applied to extract features to be used as input of the pattern classifiers developed by artificial neural networks. To estimate the accuracy of the classifiers, use was made of random selection techniques with sample reposition (Bootstrap technique) and without sample reposition. This work can be considered innovative in that field of research, and the results obtained motivate this paper.

Keywords: Casting defects, radioscopy, image processing, accuracy estimation, Bootstrap.

1. Introduction

Shrinkage, as molten metal cools during the manufacture of die castings, can cause defect regions within the workpiece. These are manifested, for example, by bubble-shaped voids, cracks, slag formation, or inclusions. Light-alloy castings for the automotive industry, such as wheel rims, steering knuckles, and steering gear boxes are considered important components for overall roadworthiness. To ensure the safety of construction, it is necessary to check every part thoroughly. Radioscopy rapidly became the accepted way for controlling the quality of die castings through computer-aided analysis of X-ray images (Mery, 2006b). The purpose of this non-destructive testing method is to identify casting defects, which may be located within the piece and thus are undetectable to the naked eye.

Two classes of regions are possible in a digital X-ray image of an aluminium casting: regions belonging to regular structures (RS) of the specimen, and those relating to defects (D). In an X-ray image we can see that the defects, such as voids, cracks and bubbles (or inclusions and slag), show up as bright (or dark) features. The reason is that X-ray attenuation in these areas is lower (or higher). Since contrast in the X-ray image between a flaw and a defect-free neighbourhood of the specimen is distinctive, the detection is usually performed by analysing this feature (see details in (Mery, 2003) and (Mery, 2006a)). In order to detect the defects automatically, a pattern recognition methodology consisting of five steps was developed (Mery, 2006b): (a) Image formation, in which an X-ray image of the casting that is being tested is taken and stored in the computer; (b) Image pre-processing, where the quality of the X-ray image is improved in order to enhance its details; (c) Image segmentation, in which each potential flaw of the X-ray image is found and isolated from the rest of the scene; (d) Feature extraction, where the potential flaws are measured and some significant features are quantified; (e) Classification, where the extracted features of each potential flaw are analysed and assigned to one of the classes (regular structure or defect).

Although several approaches have been published in this field (see for example a review in (Mery, 2006b)), the performance of the classification is usually measured without statistical validation. This paper attempts to make an estimation of the true accuracy of a classifier using the Bootstrap technique (Efron, 1993) and random selection without repositioning applied to the automated detection of casting defects. The true accuracy of a classifier is usually defined as the degree of correctness of data classification not used in its development. The great advantage of this technique is that the estimation is made by sampling the observed detection distribution, with or without repositioning, to generate sets of observations that may be used to correct for bias. The technique provides non-parametric estimates of the bias and variance of a classifier, and as a method of error rate estimation it is better than many other techniques (Webb, 2002).

The rest of the paper is organised as follows: Section 2 outlines the methodology used in the investigation; Section 3 shows the results obtained recently on real data; Finally, Section 4 provides the concluding remarks.

2. Methodologies

2.1. Processing of the casting images

The X-ray image taken with an image intensifier and a Charge Coupled Device (CCD) camera (or a flat panel detector), must be pre-processed to improve the quality of the image. In our approach, the pre-processing techniques are used to remove noise, enhance contrast, correct the shading effect, and restore blur deformation (Mery, 2006b).

The segmentation of potential flaws identifies regions in radioscopic images that may correspond to real defects. Two general features of the defects are used to identify them: (a) a flaw can be considered as a connected subset of the image, and (b) the grey level difference between a flaw and its neighbourhood is significant. According to these features, a simple automated segmentation approach was suggested in Mery and Filbert, 2002a (see Figure 1). First, a Laplacian of Gaussian (LoG) kernel and a zero crossing algorithm (Castleman, 1996) are used to detect the edges of the X-ray images. The LoG-operator involves a Gaussian lowpass filter which is a good choice for pre-smoothing our noisy images that are obtained without frame averaging. The resulting binary edge image should produce closed and connected contours at real flaws which demarcate regions. However, a flaw may not be perfectly

Romeu Ricardo da Silva and Domingo Mery are with the Departamento de Ciencia de la Computación, Pontificia Universidad Católica de Chile. E-mail: romeu@romeu.eng.br and dmery@ing.puc.cl; www.romeu.eng.br and http://dmery.ing.puc.cl

enclosed if it is located at an edge of a regular structure as shown in Figure 1(c). In order to complete the remaining edges of these flaws, a thickening of the edges of the regular structure is performed as follows: (a) the gradient of the original image is calculated (see Figure 1(d)); (b) by thresholding the gradient image at a high grey level a new binary image is obtained; and (c) the resulting image is added to the zero crossing image (see Figure 1(e)). Afterwards, each closed region is segmented as a potential flaw. For details see a description of the method in Mery and Filbert, 2002a.

All regions enclosed by edges in the binary image are considered 'hypothetical defects' (see example in Figure 1(e)). During the feature extraction process the properties of each of the segmented regions are measured. The idea is to use the measured features to decide whether the hypothetical defect corresponds to a flaw or a regular structure.

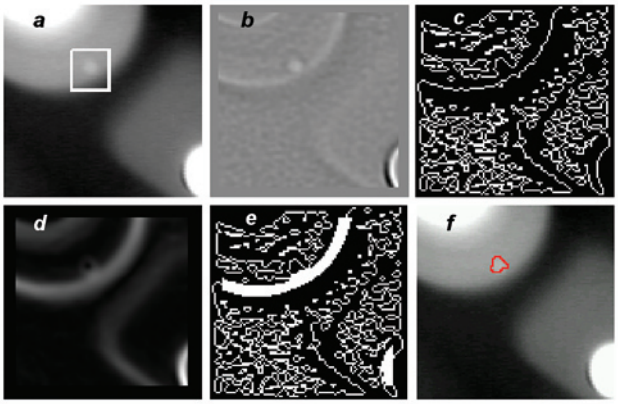

Figure 1. Detection of flaws: (a) radioscopic image with a small flaw at an edge of a regular structure, (b) Laplacian-filtered image with σ = 1.25 pixels (kernel size = 11 × 11), (c) zero crossing image, (d) gradient image, (e) edge detection after adding high gradient pixels, and (f) detected flaw using feature *F*1 extracted from a crossing line profile (Mery, 2003)

The features extracted in this investigation are described below (Table 1), and they provide information about the segmented regions and their surroundings.

The total number of features extracted is eight divided into three geometric features and five intensity features. In our work we present results obtained on 72 radioscopic images of aluminium die castings. The size of the images is 572 × 768 pixels. About 25% of the defects of the images were existing blow holes (with ∅ = 2.0 – 7.5 mm). They were initially detected by visual (human) inspection. The remaining 75% were produced by drilling small holes (with ∅ = 2.0 – 4.0 mm) in positions of the casting which were known to be difficult to detect. In these experiments, 424 potential defects were segmented, 214 of them correspond to real defects, while the others are regular structures (210).

2.2 Development of the non-linear classifiers

The non-linear classifiers were implemented using a two-layer neural network with training by error back propagation. The first step taken in the development of a non-linear classifier was to optimise the number of neurons used in the intermediate layer in order to obtain the best accuracy possible for the test sets. Some tests

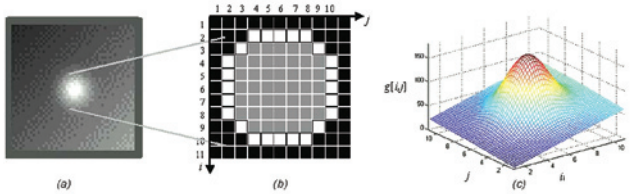

Figure 2. Example of a region. (a) X-Ray image, (b) segmented region, (c) 3D representation of the intensity (grey value) of the region and its surroundings (Mery *et al,* 2003)

Table 1. Descriptions of the features extracted

f1 and f2	Height (f1) and width (f2): height (*h*) and width (*w*) of the region (Mery and Filbert, 2002b).
f3	Area (*A*): number of pixels that belong to the region (Mery and Filbert, 2002b).
f4	Mean grey value (*G*): mean of the grey values that belong to the region (Mery and Filbert, 2002b).
f5	Mean second derivative (*D*): mean of the second derivative values of the pixels that belong to the boundary of the region (Mery and Filbert, 2002b).
f6	Crossing Line Profile (F_1): Crossing line profiles are the grey level profiles along straight lines crossing each segmented potential flaw in the middle. The profile that contains the most similar grey levels in the extremes is defined as the best crossing line profile (BCLP). Feature F_1 corresponds to the first harmonic of the fast Fourier transformation of BCLP (Mery, 2003).
f7	Contrast Kσ: standard deviation of the vertical and horizontal profiles without offset (Mery and Filbert, 2002b).
f8	High contrast pixels ratio (*r*): ratio of number of high contrast pixels to area (Mery, 2006a).

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) variables were used (Haykin, 1994; Duda *et al*, 2001). The initialisation of the synapses and bias used the Widrow (Beale, 2001) method. All these training variations resulted in a convergence for the same range of error.

2.3 Accuracy estimation

There are various techniques to estimate the *true accuracy* of a classifier, which is usually defined as being the degree of correctness of classification of data not used in its development. The three that are most commonly used are: simple random selection of data, cross validation that really presents diverse implementations (Diamantidis, 2000), and the Bootstrap technique (Efron and Tibshirani, 1993, 1995). 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 data available and the specific classification to be made.

As described by Efron and Tibshirani (1993), two properties are important when evaluating the efficiency of an estimator $\hat{\theta}$, its bias and its variation, that are defined by the equations below:

$$Bias = E[\hat{\theta}] - \theta \quad \text{.....(1)}$$

$$Var\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] \quad \text{.....(2)}$$

where:

$E[\hat{\theta}]$: The expected value of estimator $\hat{\theta}$.

$Var\left(\hat{\theta}\right)$: The variance of estimator.

An estimator is said to be reliable if it contains low values of bias (trend) and variance. However, in practice an appropriate relation between both is desirable when looking for a more realistic objective (Efron and Tibshirani, 1993, 1995). When dealing with the accuracy of a classifier, bias and variance of the estimated accuracy are going to vary as a function of the number of data and the accuracy estimation technique used.

In this work, to calculate the classification accuracy of casting defects we first carried out the Bootstrap technique as follows:

A set of Bootstrap data (size *n*), following Efron's definition (Efron, 1993), is made up of $x_1^*, x_2^*, \cdots, x_n^*$ data, obtained in a random way and with repositioning, from an original set of data $x_1, x_2, \cdots, x_n$ (also size *n*). In this way it is possible for some data to appear 1, 2, 3 or *n* times or no times (Efron and Tibshirani, 1993). With this technique the classifier implemented using the i^{th} training set is tested with data 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. The model of Bootstrap accuracy estimation $\hat{\theta}_B$ of frequently used pattern classifiers is defined by:

$$\hat{\theta}_B = \frac{1}{b}\sum_{i=1}^{b}\left(\hat{\omega}\hat{\theta}_i + \left(1-\hat{\omega}\right)\hat{\theta}_c\right) \quad \text{.....(3)}$$

where $\hat{\theta}_c$ is the apparent accuracy (calculated with the training set data only) and the weight $\hat{\omega}$ varies between 0.632 and 1, which is normally taken as being equal to 0.632 (Efron and Tibshirani, 1993, 1995).

As a second way of estimating the accuracy of the developed classifiers, the form of random selection without data reposition was used for the formation of the training and testing sets, different from the Bootstrap technique (Silva, 2005). In addition to that, Receiver Operating Characteristic (ROC) curves were drawn to verify the reliability of the results achieved with this technique (Duda, 2001).

2.4 Features selection

In order to reduce the computation time required for classification it is necessary to select features; this way the classifier only works with non-correlated features that provide flaw detection information. There are various methods for evaluating the relevance of the extracted features. In this work the simple linear correlation coefficient was used in order to know if the features used were correlated among themselves and if they were really relevant to discriminate the two classes: regular structure (RS) and defect (D).

The correlation among the features was assessed where the linear correlation coefficient was analysed. It is well known in statistics and is calculated by the following equation (Silva *et al*, 2002):

$$\rho(x,y) = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sigma_x\sigma_y} \quad \text{.....(4)}$$

where ρ(x,y) is the linear correlation between variables *x* and *y*, $\bar{x}$ and $\bar{y}$ are the expected values of variables *x* and *y*, and the standard deviation values of variables *x* and *y* are given by σ_x and σ_y, respectively.

To check the reliability of the correlation among the extracted features, a criterion of the correlating values of the order of $2/\sqrt{N}$ was used, where *N* is the number of samples of each features/class.

3. Results

3.1 Features selection

In the first place, an evaluation was made of the relevance of the characteristics extracted from the images by calculating the linear correlation coefficients. Analysing the results of Table 2, the correlation coefficients show that the f5 and f8 features are the most relevant (largest coefficients), showing greater relevance in the discrimination between defect classes and regular structure. On the other hand, the features f1, f2, f4 and f7 also had coefficient values above the limit for a confidence interval of 95% (0.10). The darker lined cells in the last column represent the highest correlation values between the features and classes. From the results, it is clear that many features are correlated (clearer lined cells in the matrix). Although two features did not have a correlation above the limit, the first classification tests were made with the eight initial features.

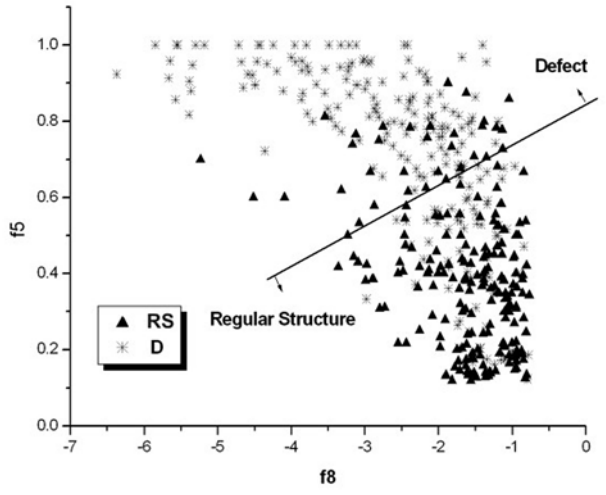

Figure 3. Graph of f8 *versus* f5 (the two most relevant features according to the correlation criterion) for the two-dimensional representation of the classes

In order to visualise the problem of separation of the RS and D classes in a two-dimensional representation space (f8xf5), since it is impossible to imagine them in a space with 8 dimensions, the graph of Figure 3 was drawn. In this graph it is seen that there is a large invasion of the domain space between both classes, and we can think of a separation of the RS and D classes through a hypothetical linear discriminator illustrated in the Figure. Even a non-linear discriminator (a curve, for example), would not result in an optimised success.

An optimised way of representing the domains of the classes of patterns of a multivariate system in two-dimensional space is by obtaining the two main discrimination components. It is known that the main linear discrimination address is called Fisher's Discriminator (Duda, 2001), and it maximises the interclass covariance matrix and minimises the intraclass covariance matrix (Duda, 2001; Silva, 2006). In this case, the first linear discrimination address of classes RS and D can be obtained going over a supervised neural network of the back propagation type with only one neuron (Haykin, 1994). Then it is possible to obtain a second main linear discrimination address, also with a neural network with only one neuron, using for the training of the network the residual information of the projection of the original information in the first discrimination address, what is called independent components (orthogonals). A detailed description of this technique is found in Silva (2004).

In this way the two main components of the linear discrimination of classes RS and D with a neural network of only one neuron, which was trained through the error back propagation algorithm using batch training (3000 periods), parameter β=0.9 and α variable, were obtained. Figure 4 shows the graph obtained with those two main linear discrimination addresses. It is evident that the separation of classes RS and D is more efficient in that representation space, because a visual analysis will make it possible to identify that there are few false positive (RS inputs in the domain space of D) and false negative (D inputs in the domain space of RS) errors. The projection of the data on the *x* axis (p1) represents what would be the best discrimination of these classes, and a projection on *y* (p2), the second best discrimination. From this graph it is concluded that the separation between RS and D can achieve good indices of success with well developed pattern classifiers.

Table 2. Matrix with the coefficients of linear correlation for the ten features

	Features ($2/\sqrt{n} = 0.10$)								*Class*
	f1	*f2*	*f3*	*f4*	*f5*	*f6*	*f7*	*f8*	
f1	1	0.68	0.85	-0.21	0.44	0.15	0.31	-0.63	-0.15
f2	0.68	1	0.87	-0.31	0.45	0.16	0.42	-0.62	-0.25
f3	0.85	0.87	1	-0.23	0.43	0.18	0.45	-0.52	-0.051
f4	-0.21	-0.31	-0.23	1	-0.62	-0.16	0.05	0.46	0.38
f5	0.44	0.45	0.43	-0.62	1	0.13	0.06	-0.65	-0.41
f6	0.15	0.16	0.18	-0.16	0.13	1	0.25	-0.05	-0.07
f7	0.31	0.42	0.45	0.05	0.05	0.25	1	-0.07	-0.20
f8	-0.63	-0.62	-0.54	0.46	-0.65	-0.05	-0.07	1	0.60

3.2 Study of neuron number in the intermediate layer

The graphs of Figures 3 and 4 showed the problem of classification of classes RS and D only from the two most relevant features, according to the linear correlation criterion or through P1 and P2. However, it is well known that the linear pattern classifiers solve well very easy class separation problems (Duda, 2001). To optimise the separation between the classes of patterns RS and D, non-linear pattern classifiers were developed through supervised neural networks with two layers of neurons and error back propagation training (Haykin, 1994).

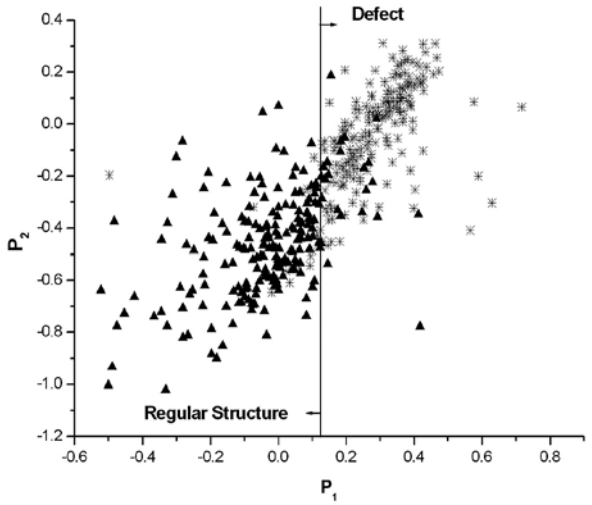

Figure 4. Graphs made with the two principal linear discrimination components

Since non-linear classifiers can have network overtraining problems, whose probability increases with increasing number of neurons in the second layer, thereby losing the capacity to generalise (Haykin, 1994). In order to decrease the probability of the existence of *overfitting* the parameters of the non-linear classifier, a study was made of the optimum number of neurons in the intermediate layer of the classifier that would make possible the best result with test sets. For that purpose, from the initial set of data with the eight features, a training set was chosen with 75% of the data chosen randomly and without reposition, and a test set with the remaining 25%, keeping the proportion between the classes. In this way the training set contained 158 samples of RS and 160 of D, and the test set had 52 of RS and 54 of D. The number of neurons in the intermediate layer of the network was varied one at a time up to 20 neurons, and the indices of success in classification and testing

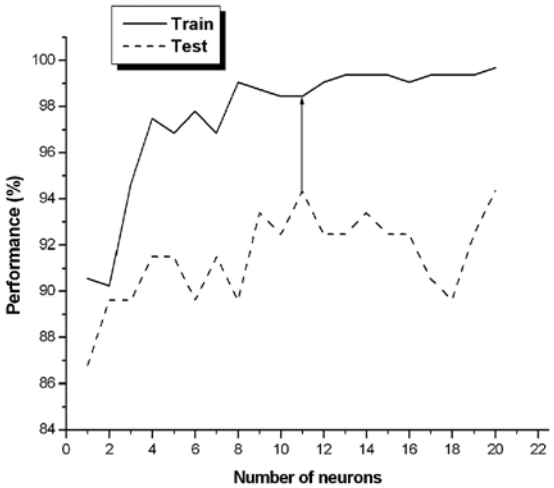

Figure 5. Graph of the indices of success obtained by the training set and the test set. The arrow indicates the point corresponding to the number of neurons chosen

Table 3. Optimisation of the number of neurons in the intermediate layer

Number of neurons	Training performance (%)	Test performance (%)
1	90.57	86.80
2	90.25	89.63
3	94.66	89.63
4	97.50	91.51
5	96.90	91.51
6	97.80	89.63
7	96.90	91.51
8	99.06	89.63
9	98.75	93.40
10	98.43	92.46
11	98.43	94.34
12	99.06	92.46
13	99.38	92.46
14	99.38	93.40
15	99.38	92.46
16	99.06	92.46
17	99.38	90.57
18	99.38	89.63
19	99.38	92.46
20	99.70	94.34

were recorded. It should be noted that, since we are dealing with only two classes of patterns, the last layer of the classifier can contain only one neuron.

The results obtained from the study of the number of neurons are shown in Table 3, and with the purpose of a graphic illustration, the graph of Figure 5 was drawn. In the table it is seen that the smallest difference between the results of the training and the tests, which theoretically can indicate a good generalisation capacity of the classifier, occurs for two neurons in the intermediate layer. However, if we analyse the increase of the performance of the classifier, which occurs significantly with the increase in the number of neurons, which is expected, a second lowest difference occurs for 11 neurons, achieving 94.34% of success with the test set. For that reason, 11 neurons were used in the intermediate layer of the neural network for the development of all the classifiers of this work having in view the estimation of the accuracy of the classification.

3.3. Accuracy estimation by the Bootstrap technique

To estimate the accuracy of the non-linear classifiers, the first technique used was a random selection with data repositioning. Once more, as an example of the operation of this technique, one can imagine a 'bag' with all the original data, then we choose data randomly from this bag to form the training sets, but every piece of data that goes to these sets returns to the 'bag' and can be chosen again several times. The data that are not chosen for training are used in the formation of the test sets.

In this way, 10 pairs of training and test sets were formed. It should be noted that, with this selection technique the training sets always have the same number of data as the original set (in this work, 424 data). In this way, the test sets had a number of data between 150 (≈ 35%) and 164 (≈ 38%). For the development of the classifiers, with the aim of decreasing the possibility of overtraining of its parameters (synapses and bias), use was made of a validation set formed by samples selected randomly from the Bootstrap training sets in a 10% proportion. This technique is well known as 'cross validation' (Haykin, 1994), and the end of the training was set to when the validation error increases or remains stable for 100 epochs or a maximum of 3000 epochs obviously choosing

the values of the network parameters in the situation of the least validation error. The results are presented in Table 4.

Analysing these results, it is seen that the training indices were quite high, with a mean of 98.46%, however the success indices of the test were significantly lower, with a mean of 55.61%. Calculating the accuracy estimator according to the weighting factor of 0.632 for the test set estimator and 0.368 for the training estimator (Diamantidis, 2000; Efron, 1993), the estimated accuracy result is 71.40%, which can be considered unsatisfactory for the classification of patterns for this problem of fault detection in automobile rims.

Table 4. Result of classification with the Bootstrap input sets (%)

Input sets	Training (%)	Test (%)
1	418/98.60	75/50.00
2	422/98.60	88/53.66
3	410/96.70	92/56.10
4	421/99.30	94/57.31
5	405/95.52	94/57.32
6	421/99.30	88/53.66
7	416/98.12	86/52.45
8	424/100	100/61.00
9	420/99.05	86/52.45
10	422/99.53	102/62.20
Mean	98.47	55.61
***Bootstrap* accuracy estimation** $\hat{\theta}_B = \frac{1}{b}\sum_{i=1}^{b}\left(0.632\hat{\theta}_i + 0.368\hat{\theta}_c\right)$	71.40	

The great problem for the classification of patterns, which is common to almost all work in that relation, is the lack of data to estimate with precision the true classification accuracy, so that it can be trusted that all the success indices will always be similar when the classifier is tested with a new set of data. The main objective of the use of the Bootstrap technique was to try to reproduce several sets for training and testing the classifiers as well as for estimating the accuracy expected for classes RS and D. One justification that can be thought of for the low success indices with that technique is the fact that the test sets have a large number of data in relation to the number of data used for training. Normally, in terms of pattern classification, the test or validation sets contain between 20 and 30% of data, and by the Bootstrap technique, in this paper, some test sets get to contain almost 40% of the data, and this can in fact affect the correct training of the network parameters, even using a cross validation technique to interrupt the trainings. This is even more

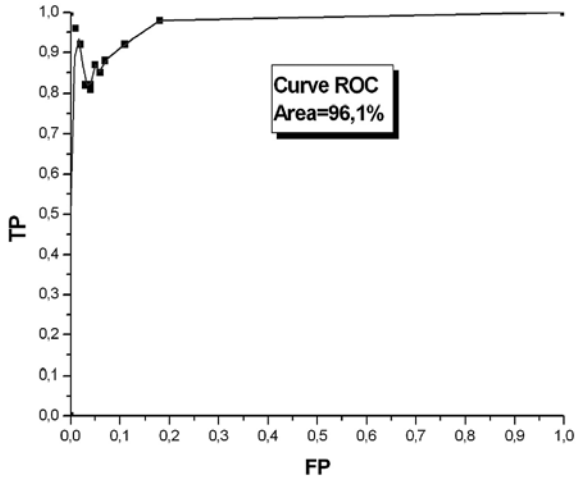

Figure 6. ROC resultant curve of the randomly selected sets without data repositioning (sixth and seventh columns of Table 5)

feasible if we think that the original data did not contain a large number of samples. To expect a success index of only about 55%, or even 71.40%, for this classification problem is too pessimistic having in mind the efficiency of the image processing techniques used and the relevance of the extracted features.

3.4 Accuracy estimation by random selection without repositioning

In the simple method of evaluation with random sampling, the original data set (with n data) is partitioned randomly into two sets: a training set containing $p \times n$ data, and a test set containing $(1 - p) \times n$ data (the values of p are chosen in a variable way case by case). This process is repeated a number of times, and the mean value is the accuracy estimator (Diamantidis, 2000). This technique was used for the first selection and formation of sets with the purpose of choosing the number of neurons of the classifier's intermediate layer. Using that simple yet very efficient technique, 10 pairs of data sets for training and testing of the classifier were chosen, and the percent proportion chosen (based on experience from other work) was 75% for training (318) and 25% for testing (106).

Table 5 contains the results achieved successfully with these sets. The fourth and fifth columns of the table refer to the number of data of each class contained in the corresponding sets. The mean was approximately 53 data of each class in each set, that is, in general there was not a significant disproportion between the number of data of each class that would affect the trainings and tests of the classifiers. The training column contains not only the percentages of success, but also the number of data classified correctly, which were as high as those obtained with the Bootstrap sets. However, it is seen that the test results were considerably higher than those achieved with the Bootstrap technique, with a mean estimated accuracy of 90.30% for the 10 test sets selected, a very satisfactory index close to the mean of 97.52% was obtained for the training sets. That small difference of about 7% is perfectly acceptable, and it shows the generalisation of the classifiers (confirmed also by the low values found for standard deviation). It should be noted that with these sets, cross validation was also used for interrupting the training in a manner similar to that used for the Bootstrap sets.

Table 5 also contains the false negative (FN) indices, real defects classified as regular structures as well as the false positive (FP) indices, regular structures classified as defects. The mean values achieved of 7.69% and 11.64%, respectively can be considered satisfactory, especially if we consider that the most critical situation is always that of false negative, and less than 8% of errors in the classification of real defects is an index that cannot be considered high for a fault detection situation in these kinds of images.

Figure 6 shows the ROC curve obtained from the interpolation of true positive (TP), 1- FN, and false positive points of Table 5.

Table 5. Results of classification with the input sets of the random selection without repositioning (%)

Input sets	Training (%)	Test (%)	RS	D	FN (%)	FP (%)
1	314/98.75	95/89.63	57	49	3.51	18.37
2	311/97.80	98/92.46	52	54	11.54	3.70
3	315/99.06	101/95.30	50	56	2.00	7.14
4	312/98.12	95/89.63	55	51	18.18	1.96
5	314/98.75	93/87.74	45	61	4.44	18.03
6	307/96.55	93/87.74	60	46	6.67	19.57
7	299/94.03	96/90.57	53	53	5.66	13.21
8	314/98.75	94/88.68	46	60	6.52	15.00
9	311/97.80	96/90.57	55	51	7.27	11.76
10	304/95.60	96/90.57	54	52	11.11	7.69
Mean (%)	97.52	90.30	≈53	≈53	7.69	11.64
Standard deviation (%)	1.28	1.61			13.03	12.53

Table 6. Optimisation of the number of neurons in the intermediate layer without features 3 and 6

Number of neurons	Training performance (%)	Test performance (%)
1	84.91	85.85
2	87.74	82.08
3	91.51	88.68
4	92.46	88.68
5	93.40	88.68
6	95.30	89.63
7	95.60	88.68
8	94.03	84.91
9	94.97	85.85
10	95.60	87.74
11	96.55	90.57
12	95.92	88.68
13	97.50	88.68
14	96.23	86.80
15	97.17	89.63
16	97.17	87.74
17	96.23	92.46
18	96.23	89.63
19	95.92	87.74
20	96.55	89.63

The area over the curve, calculated by simple integration of the interpolated curve, represents the efficiency of the system used for the detection of the real defects in the acquired images (probability of detection, PoD). In this case the value found for the area was 96.1%, which can be considered an optimum index of the efficiency and reliability of the system, higher than the 90.30% estimated accuracy value of Table 5.

3.5 Relevant features

Since the feature relevance results (section 3.1) showed a high correlation for almost all eight features extracted, except for features 3 and 6, the following step of our research was to extract those two features from the data sets and verify the possibility of not using them, provided the percentages of success did not decrease.

From the same training and test sets of Table 3, features 3 and 6 were extracted. Once again, a study was made of the number of neurons in the intermediate layer of the classifier, keeping in mind that reduction in the size of the input sets. Table 6 shows the results obtained, and again the number of 11 neurons was selected, due to the high degree of success obtained with the test set and the small difference for the training set. Higher indices occurred for 17 neurons, but since with a large number of neurons in the intermediate layer there is a higher probability of overtraining (Haykin, 1993), the value of 11 neurons was repeated, to allow also a more equitable comparison with the results of Table 5.

Table 7. Results of classification with the input sets of the random selection without repositioning (%), but without features 3 and 6

Input sets	Training (%)	Test (%)	FN (%)	FP (%)
1	302/94.97	92/86.80	14.04	14.29
2	299/94.03	94/88.68	11.54	11.11
3	302/94.97	93/87.74	8.70	15.00
4	311/97.80	92/86.80	20.00	5.88
5	309/97.17	84/79.25	31.11	13.11
6	305/95.92	89/83.97	43.33	13.04
7	291/91.51	90/84.91	11.32	18.87
8	302/94.97	93/87.74	8.70	15.00
9	292/91.83	94/88.68	10.91	13.73
10	299/94.03	92/86.80	14.04	14.29
Mean (%)	94.72	86.14	17.37	13.43
Standard deviation (%)	1.50	2.06	15.60	11.13

From the sets of Table 5, features f3 and f6 were removed for the trainings and tests. As seen in Table 7, the mean success values decreased 4% percentage wise, for both the training and the test sets. The mean of the false negative increased to 17.36%, a significantly undesirable increase. Therefore, it can be concluded that although these two features had not shown such high linear correlation values compared to the others, their use for the detection of the real defects is very important, and therefore they should not be discarded from the system.

4. Conclusions

From the results obtained, it can be concluded that the extracted features, after applying the image processing techniques, are highly relevant for the detection system, particularly features f4, f5 and f8.

As to the Bootstrap technique, the accuracy results were well on this side of acceptable, and that can be explained by the small amount of data available in the training sets.

The estimation of the accuracy of classification with the random selection technique without data repositioning, with fixed values of 25% of data for the test sets, had high indices of correctness, showing the efficiency of the system developed for the detection of defects, which was also evident from the drawing of the ROC curve for the system.

It must be pointed out that this work does not exhaust the research in this field, and that much can still be done to increase the reliability of the results obtained as well as to increase the number of features to be extracted to increase the degree of success in the detection of faults. However, this paper can be considered pioneering dealing with defects in automobile wheels, and there are no results on estimated accuracy in other papers that could be used for comparison with these results.

Acknowledgement

This work was supported in part by FONDECYT – Chile (International Cooperation), under grant No. 7060170. This work has been partially supported by a grant from the School of Engineering at Pontificia Universidad Católica de Chile.

References

1. M Beale, Neural Network Toolbox for Use with Matlab User's Guide Version 4, USA. The MathWorks, 2001.
2. K Castleman, Digital Image Processing, Prentice-Hall, Englewood Cliffs, New Jersey, 1996.
3. N A Diamantidis, D Karlis and E A Giakoumakis, 'Unsupervised stratification of cross-validation for accuracy estimation', Artificial Intelligence, 116: 1-16, 2000.
4. R O Duda, P E Hart and D G Stork, Pattern Classification, 2nd edition , USA, John Wiley & Sons, 2001.
5. B Efron and R J Tibshirani, 'Cross-validation and the Bootstrap: estimating the error rate of the prediction rule', Technical Report 477, Stanford University, http://utstat.toronto.edu/tibs/research.html, 1995.
6. B Efron and R J Tibshirani, An Introduction to the Bootstrap, New York, Chapman & Hall/CRC, 1993.
7. S Haykin, Neural Networks – A Comprehensive Foundation, USA. Macmillan College Publishing Inc, 1994.
8. D Mery and D Filbert, 'Automated flaw detection in aluminium castings based on the tracking of potential defects in a radioscopic image sequence', IEEE Trans. Robotics and Automation 18, (2002a), 890-901, 2002.

9. D Mery and D Filbert, 'Classification of potential defects in automated inspection of aluminium castings using statistical pattern recognition', (2002b), In Proceedings of 8th European Conference on Non-Destructive Testing (ECNDT 2002), Barcelona, 17-21 June 2002.
10. D Mery, 'Crossing line profile: a new approach to detecting defects in aluminium castings', Lecture Notes in Computer Science, LNCS 2749: 725-732, 2003.
11. D Mery, R R Silva, L P Caloba and J M A Rebello, 'Pattern recognition in the automatic inspection of aluminium castings', Insight, 45(7): 431-439, 2003.
12. D Mery, 'High contrast pixels: a new feature for defect detection in X-ray testing', Insight, (2006a), (46)12: 751-753, 2006.
13. D Mery, Automated radioscopic testing of aluminium die castings', Materials Evaluation, (2006b), 764(2): 135-143, 2006.
14. H Schulenburg and M Purschke, 'Advances in the automatic evaluation of radioscopic images', In: International Conference on Computerized Tomography for Industrial Applications and Image Processing in Radiology, Berlin, 241-243, 15-17 March 1999.
15. R R Silva, L P Calôba, M H S Siqueira and J M A Rebello, 'Pattern recognition of weld defects detected by radiographic test', NDT&E Int, 37(6): 461-70, 2004.
16. R R Silva, L P Calôba, M H S Siqueira, L V S Sagrilo and J M A Rebello, 'Evaluation of the relevant characteristic parameters of welding defects and probability of correct classification using linear classifiers. Insight, Vol 44, No 10, pp 616-622, 2002.
17. R R Silva, M H S Siqueira, H S Marcio, M P V Souza, J M A Rebello and L P Calôba, 'Estimated accuracy of classification of defects detected in welded joints by radiographic tests', NDT&E International, UK, Vol 38, pp 335-343, 2005.
18. R R Silva, S D Soares, L P Calôba, M H S Siqueira and J M A Rebello, 'Detection of the propagation of defects in pressurized pipes by means of the acoustic emission technique using artificial neural networks', 48(1): 45-51, 2006.
19. A Webb, Statistical Pattern Recognition, John Wiley & Sons Inc., 2 Edition, Hoboken, USA, 2002.