The 8th International Conference of the Slovenian Society for Non-Destructive Testing
»Application of Contemporary Non-Destructive Testing in Engineering«
***September 1-3, 2005, Portorož, Slovenia*, pp. 173-180**

MEASURING OF THE RELIABILITY OF NDE

F. Fücsök[1], C. Müller[2], M. Scharmach[2]

[1]Budapest Power Plant Ltd,
Budafoki ut 52, H-1117 Budapest, HUNGARY,
E-mail: ferenc.fucsok@bert.hu

[2]Federal Institute for Materials Research and Testing
Unter den Eichen 87, 12205 Berlin, GERMANY,
E-mails: Christina.Mueller@bam.de , Martina.Scharmach@bam.de

ABSTRACT

It is an important question regarding to the contemporary NDT systems how reliable is the result of the test. This question is rising connected to all diagnosis system in human medicine mechanical engineering or civil constructions. The reliability of an NDE system means the consistency of capability the system to detect, to classify and to evaluate the existing deviation within test pieces. The main elements of the reliability are: the intrinsic capability of the system, the effect of application parameters and the human factors.
At present time we cannot determine all of the effect of the elements, so we have to measure the reliability with different methods like POD (probability of determination) and ROC (Receiver Operating Characteristic) methods. According to the modular concept it is possible to determine the reliability all of the modules of NDE systems differently.
The NDT section of Scientific Society of Mechanical Engineering recognised the importance of reliability degree of NDE. So we organised a round-robin test of radiographic film evaluation as a module of radiographic test, and try to measure the effect of human factor.
The first round-robin test finished and the evaluation method is presented in details as an example of the measuring of reliability. The second round-robin test is in progress in Hungary. The Slovenian radiographers are invited to take part in this important and interesting test.

Keywords: Reliability, Radiographic testing, Round-robin test

1. Introduction

The reliability of a diagnostic system is an important question for the human doctors, for the fracture mechanic experts, or for the customers of NDT laboratories. The reliability of an NDE system means the consistency of capability the system to detect, to classify and to evaluate the existing deviation within test pieces. But measuring the reliability is a difficult problem because it depends on a lot of elements.
Many radiographic exposures and film interpretations are made daily in a typical industrial test laboratory. Yet, questions remain regarding the precise probability of detecting specific discontinuities, including the reliability of each individual inspector or laboratory. Although each

laboratory's most experienced inspectors evaluate each radiograph, the actual reliability of these inspectors remains somewhat unknown.
If you are in a fieldwork you will find that everybody (including the welders) can evaluate the radiographic films. So the most serious quarrels are about the results, with other words the reliability of the radiographic test. That was the reason why we chose the topic of an international Round robin test (RRT), the radiographic film evaluation. The NDT section of the Hungarian Scientific Society of Mechanical Engineering recognised the importance of reliability of the film evaluation, and organised a Round-robin test.
The paper is organized as follows: The next section will give a short background of the ROC method. Then the practical procedure of the RRT is described. Finally the results of the RRT of radiographic film evaluation are presented and a new RRT will be announced.

2. The ROC method

2.1 Background of the ROC method

To describe the efficiency of an NDT system it is necessary to distinguish between the devices parameters as signal to noise ratio or spatial resolution, which guarantee merely the functioning of the NDE equipment and the actual testing performance in defect detection and classification. The NDE diagnosis system acts from the interaction of rays ore waves with a defect in the material up to an indication in an inspection report with an eventful history of the useful indication signal and the noise signal. Considering the example of radiographic weld testing the physical process can be theoretically modeled or described by empirical parameters to certain accuracy up to the creation of the image on the film.
For the evaluation of the whole testing chain - especially for such NDE-systems where standards are not yet ready and the experience of the "NDE world" is poor - including the human inspector and its interaction with complex technique it is helpful to perform a statistical evaluation of defect findings. The statistics of evaluation is based on the Receiver Operating Characteristic (ROC) [1-5] which is deviated from the general theory of signal detection.
The general four possible situations in NDT diagnosis are presented in Figure 1. The idea of the ROC method is to characterize the accuracy of an inspection system by evaluating the true positive detection rate versus the false positive detection rate for a set of possible decision criteria (decision whether the signal is noise or a defect signal) which represents a varying sensitivity or recording level.

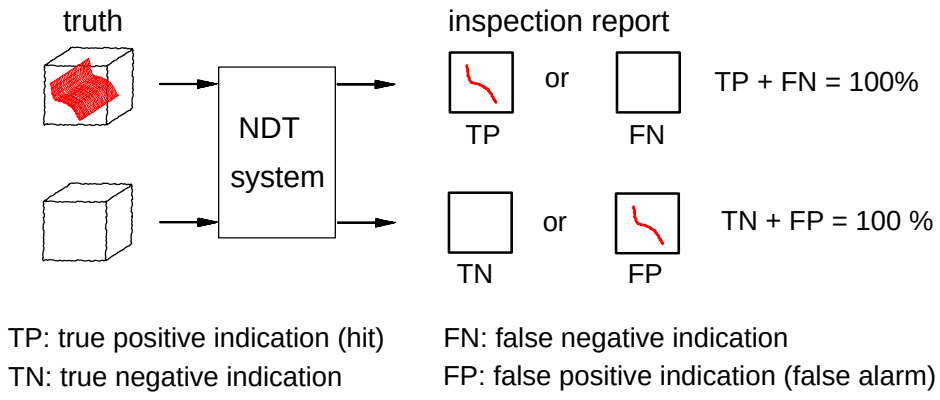

Fig. 1: The four cases of NDT-diagnosis.

The interpretations of the Figure 1 are:

TP: true positive: the defect was indicated where it was present

FN: false negative: the defect was not indicated where it was present
TN: true negative: the defect was not indicated where it was not present
FP: false positive: the defect was indicated where it was not present

The evaluation of the probabilities:

The probability of detection (POD) or other words the probability of True Positive:

$$POD = \quad P(TP) = \frac{TP}{TP + FN}$$

The probability of false alarm or other words the probability of False Positive:

$$PFA = \quad P(FP) = \frac{FP}{TN + FP}$$

2.2 An example

On the Figure 2 you can see a sketch of a welded seam, which contains 22 cells. The welding contains a 7 long cell defect, (see the thicker green line). Let suppose, the defect was detected partly at a wrong place (see the thinner red line).

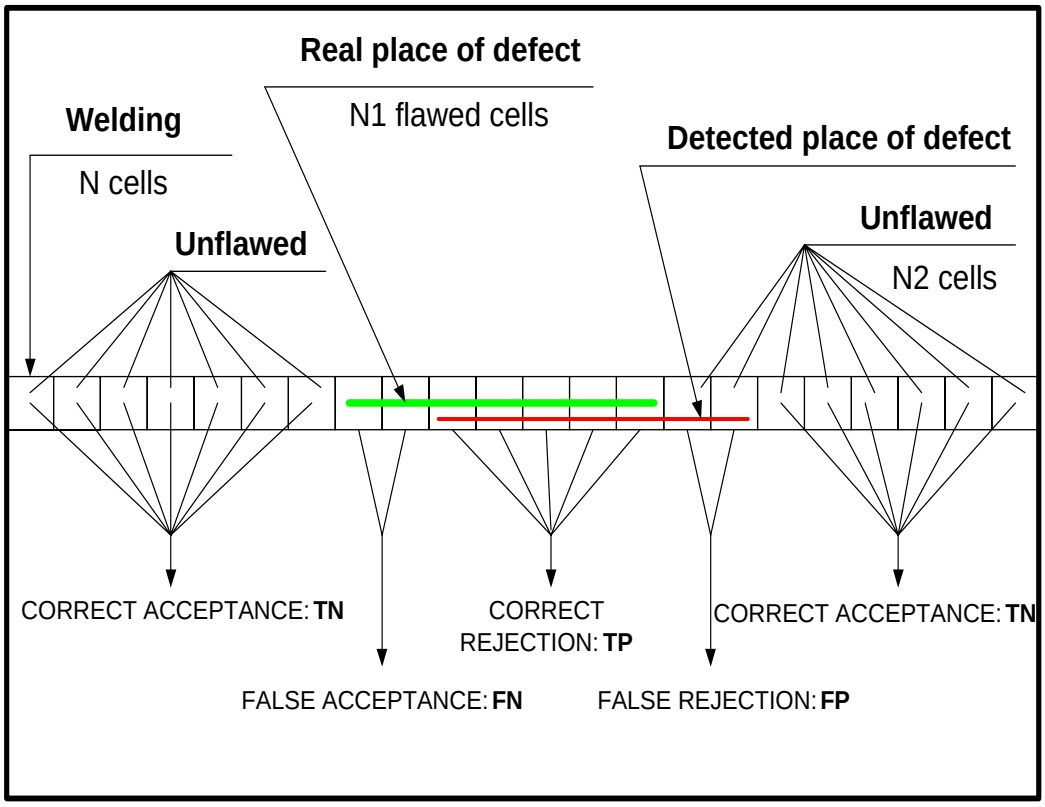

Fig. 2: A sketch of a welded seam.

Let us evaluate the probabilities.

The probability of detection:

$$POD = \frac{TP}{TP+FN} = \frac{5}{5+2} = \quad 0{,}71$$

The probability of false alarm:

$$PFA = \frac{FP}{FP+TN} = \frac{2}{2+13} = 0{,}13$$

These results were plotted on the ROC diagram at the Figure 3, where you can see the result of a perfect tester and a guessing tester, too.

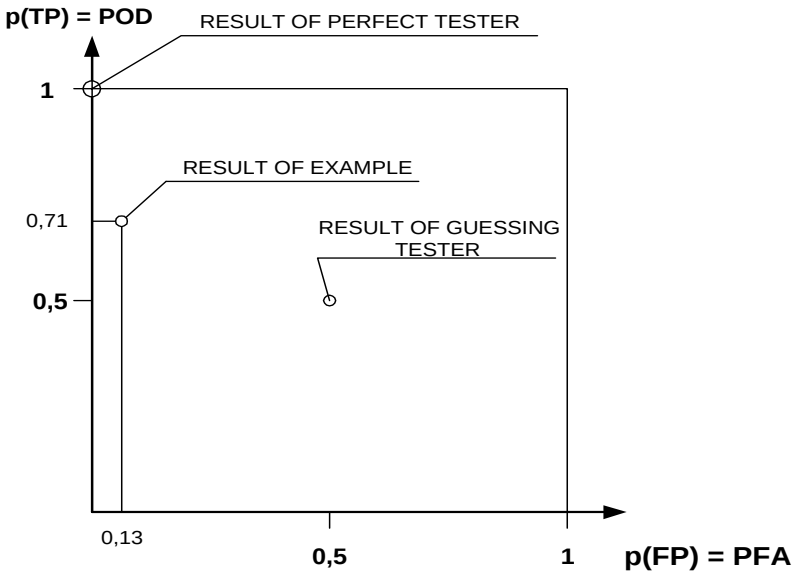

Fig. 3: The ROC diagram of the example.

2.3 Creation of ROC curves

The creation of an ROC curve is shown in Figure 4, where - following the curve from the lower left corner to the upper right - the sensitivity of the system raises. So - in the lower part of the curve the highest signals (correct indications) are included and only a small amount of noise (false calls). In the higher part more and more all of the defects are taken into account but also a greater amount of false calls has to be paid as price.

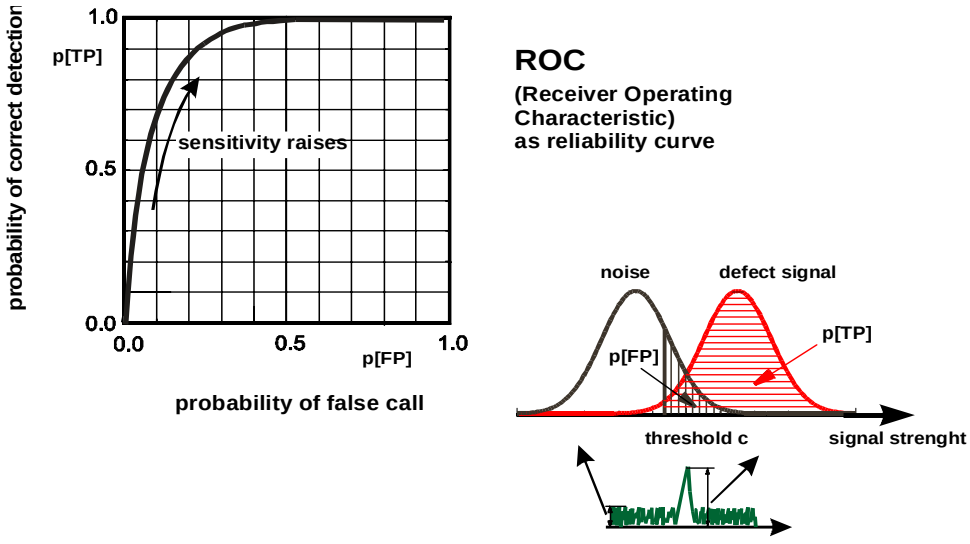

Fig. 4: Creation of an ROC curve (theory).

In practice it is not possible to apply continuously growing signal thresholds and to count correct and false call rates for each. Therefore different discrete categories of signal counting are defined to be applied by the inspectors during the non-destructive testing evaluation as indicated in Figure 5.

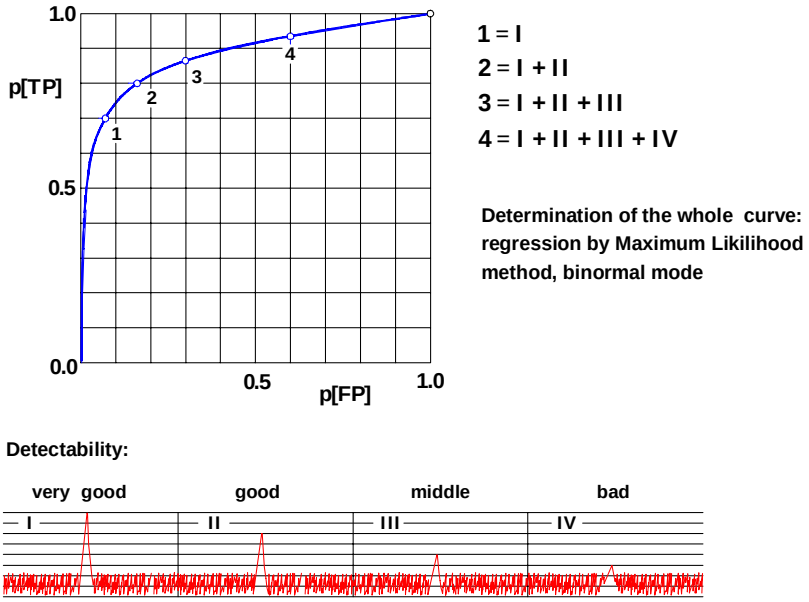

Fig. 5: Practical creation of an ROC curve.

These categories might correspond to the visibility of defects on a radiographic film or to an echo height in an ultrasonic A-scan. We call it detectability later on. So we yield five different experimental points in the ROC diagram – in the mentioned RRT investigation we reduced it to four. The maximum point represents the actual possible operating point. From the whole curve shape - which can be obtained by using a special regression method on the basis of the binormal model - the overall capability of the system is indicated. There is e.g. a forecast possible what will happen when the sensitivity of the system will be raised: Is there a gain in defect finding or is only the false alarm rate increasing?

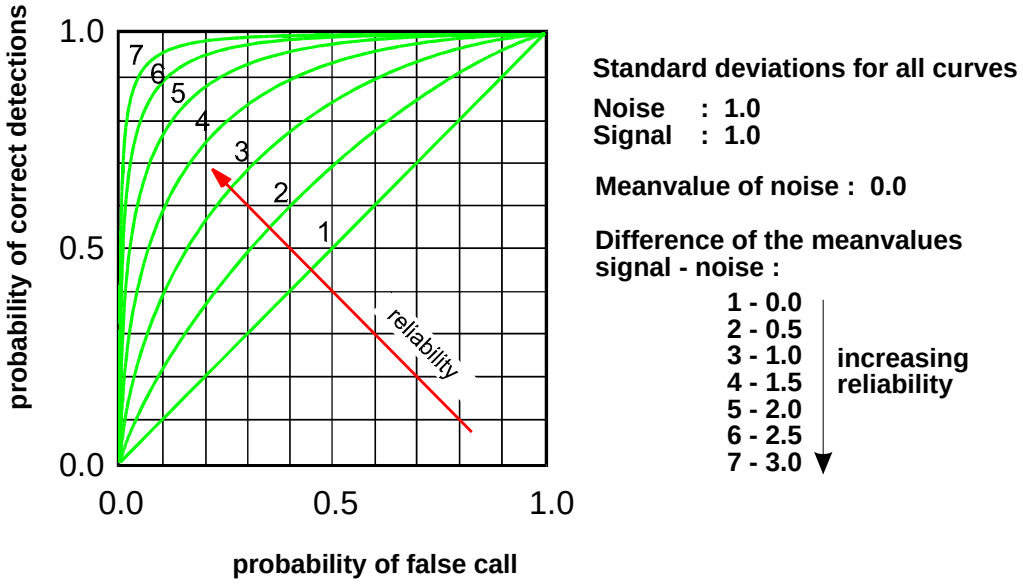

Fig. 6: Differentiating NDT-systems by ROC curves.

Considering the area under the ROC-curve (see Figure 6) it may vary from 0.5 (pure chance curve 1) up to 1.0 which corresponds to an ideal NDT system belonging to the left corner's step curve. The fictive systems, shown in Figure 6, are the performance of the system increases from curve 1 to curve 7.
With the distance from the line 1 a good summary performance value is given showing the capability of the method or the capability of human factor.

3. The RRT of film evaluation

3.1 The preparation of RRT

To try to estimate the reliability of the human factor in radiographic film evaluation an international Round-robin test was organized for Croatian, Hungarian and Polish laboratories to voluntary attendees. For the purpose of the RRT a set of films was selected in the Reliability Laboratory of BAM in Berlin, which provides the scientific and technical support. The selected 38 films contain 206 defect indications of different types and dimensions.
The selected films were scanned by a state of the art film digitizer (LS85 SDR, Lumisys). The digital images were evaluated with the help of a dedicated image processing computer program of BAM. The results of this evaluation are taken as the true values of the discontinuities. The types of the discontinuities were discussed and agreed by a small group of Hungarian experts. At the same time the X-ray films were copied with a laser printer (AGFA Scopix LR5200), for further information see [6]. Four sets of films were printed into AGFA Scopix Laser films, which were sponsored by AGFA. In this way all of the participants of the RRT evaluate exactly the same films.
The voluntary evaluators were provided with clear instructions of the procedure and specific forms for the support of the evaluation work and to aid the pre-processing of the results with the computer. The inspectors were asked to evaluate and identify each weld image cm by cm, which is a very strict prescription. Additionally they were required to fill in a form for the circumstances of the film evaluation including the length of evaluation time. The evaluation work of the 38 films took 4 to 15 hours. The longer the evaluation time the more detailed discontinuities were indicated. The evaluation results were filled in an Excel table to support the data processing of the indication results for the ROC statistics.
To take into consideration of the personal rights of the evaluators they were asked to choose a four digit personal code. So nobody knows who the owner of the best and the worst results is except for the person who had made them. The additional first digit of the personal codes is related to the country of the participants:1 = Croatia, 2 = Hungary, 3 = Poland.

3.2 The results of RRT

Till the end of the RRT a total of 60 inspectors of different laboratories were evaluated. As an example let us see the ROC curve of a Croatian participant coded 11204. On the Figure 7 you can see three curves. The higher curve is the result of the best evaluator, the lower one is the worst result. Between them you can found the result of the participant who was coded by the number in the legend. The most interesting point is the highest point of the curve, which contains the probability of all recognised discontinuities, so we called it the working point.
In the left upper corner of the diagram you can see a set of curves. They represent the requirements of the ASME code XI. Appendix VIII. The curves show the 5%, 10%....80%, 90% and 95% reliability level of probability of acceptance. The working point can be found between the 80% and 90% reliability level. It means, the evaluator will met the ASME Code requirements with more than 80% probability.
We used the ASME Code requirements arbitrary, but written requirements for probability of detection can be found only for the evaluation of ultrasonic testers.

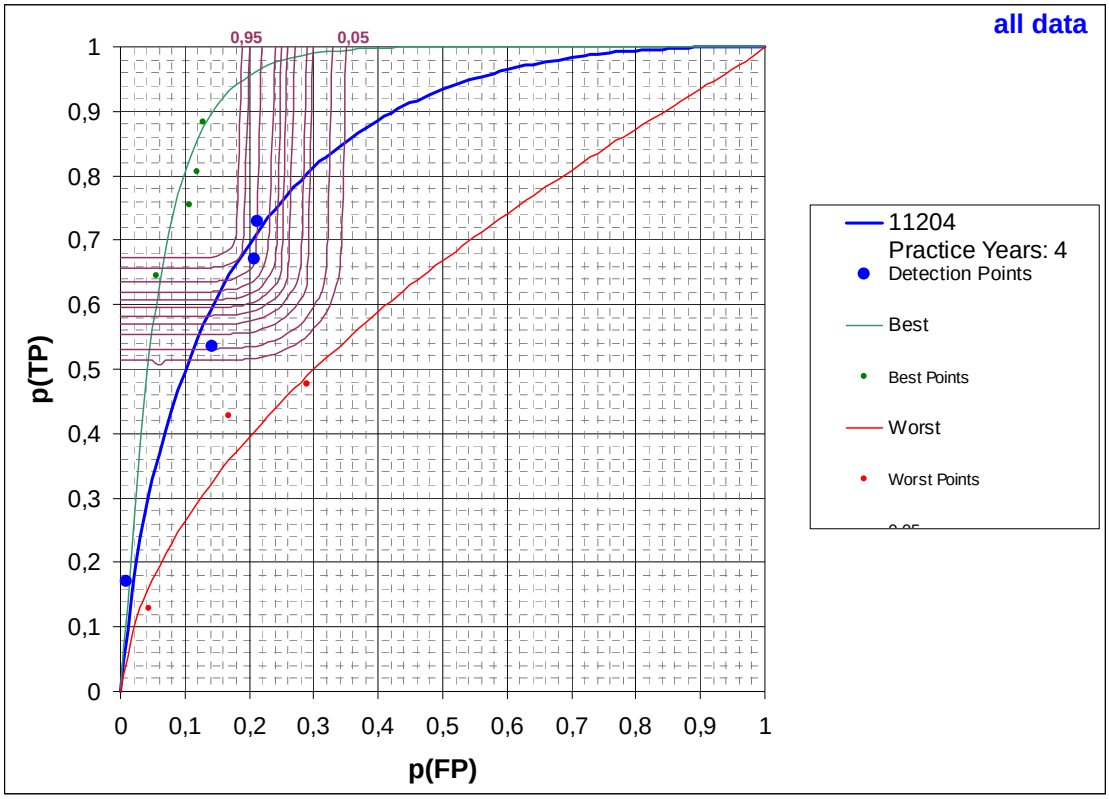

Fig. 7: The ROC curve of Croatian participant coded 11204.

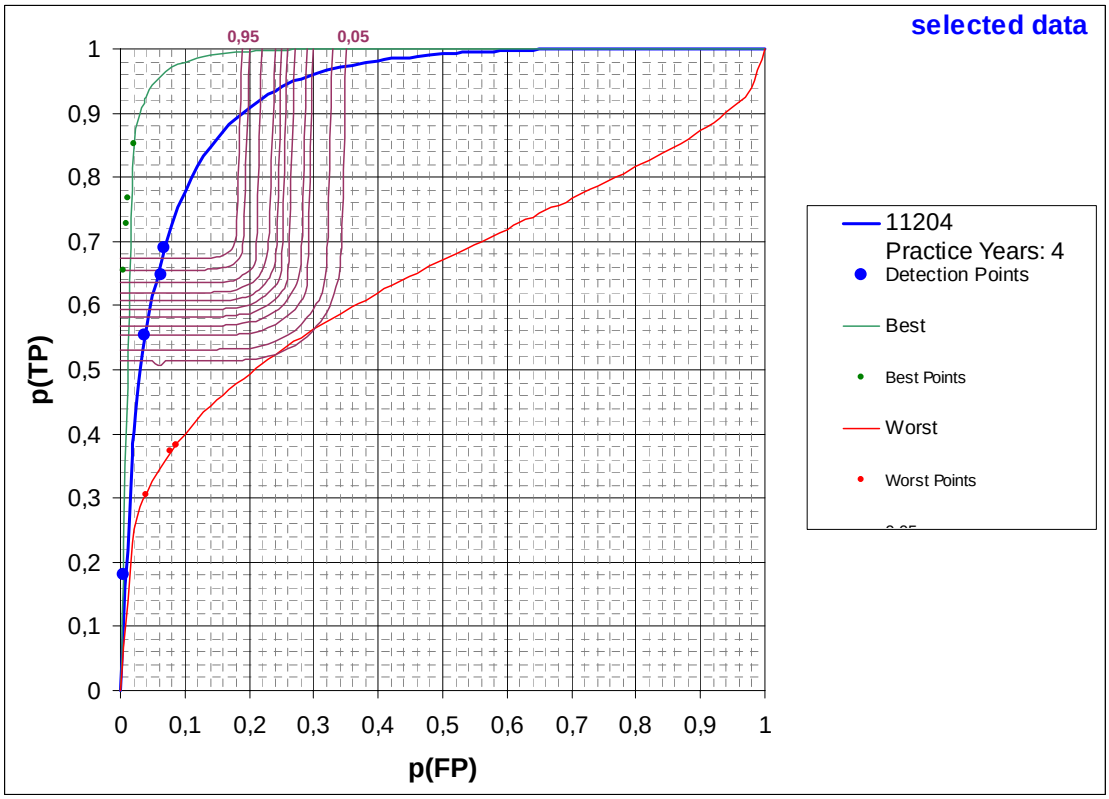

Fig. 8: The ROC curve of Croatian participant coded 11204 according to the New evaluation.

Discussing the circumstances of the RRT with the participants, many of them expressed, they could not exactly follow the prescription of the test. They have known that the small discontinuities could be omitted because of the strictest acceptance level, so they omit them. So the most practiced evaluators, supporting their long term practice, did not write the small gases and slags into the list, and they have got wrong reliability of detection. So we have to do two types of evaluations: taking into consideration of all discontinuity (called Old evaluation) and only the bigger ones which are over the acceptance level 1 of the EN 12517:1998 (called New evaluation).
On the Figure 8 can be seen the results of the evaluation of the Croatian participant coded 11204 only the flaws over the acceptance level 1. The working point can be found over the 95% reliability level. It means, the evaluator will met the ASME Code requirements with more than 95% probability.

4. Conclusions

From the evaluation of the results of the RRT we learned, the reliability of the work of the key persons of laboratories can be increased. The best results of probability of detection is 88,3 % with 12,9% false alarm in the old evaluation, and 85,2 % POD with 2,1 % PFA in the new evaluation. These results are reasonable, but the overall results are worse with high false alarm rate.
It is clear, the reliability of the film evaluation have to increase. Refresh trainings and Round robin tests are necessary. The Reliability Laboratory in BAM has a lot experiences in the field of RRT. For this reason, we continue this RRT with a new set of film.
We collected a well-selected set of radiographic shots which fulfil the following demands:

- technically good,
- have enough planar and volumetric failures,
- the true values have to be as correct as possible,
- contain different grey scale indications.

The new Round robin test with the new set of films is in progress in Hungary. We invite the Slovenian radiographers to take part in this important and interesting test.

5. References

[1] Metz, C.E. 'Basic Principles of ROC analysis' Seminars in Nuclear Medicine 8 4 (1978).
[2] Metz, C.E. 'Some practical issues of experimental design and data analysis in radiological ROC studies' Invest Radiol. 24 (1989), pp 234-245.
[3] Swets, J.A. 'Assessment of NDT systems-Part I' Mater. Eval. 41 11 (1983), pp 1294-1298.
[4] Swets, J.A. 'Assessment of NDT systems-Part II' Mater. Eval. 41 11 (1983), pp 1299-1303.
[5] Nockemann, C., Heidt, H., and Thomsen, N. 'Reliability in NDT: ROC study of radiographic weld inspections' NDT&E International 24 5 (1991), pp 235-245.
[6] Zscherpel, U., 'Film Digitisation Systems for DIR: Standards, Requirements, Archiving and Printing', the e-Journal of NDT & Ultrasonics, Vol.5 (2000) No.5 www.ndt.net/v05n05.htm.