Experimental Verification of the Effectiveness of Mammography Testing Description's Standardization

Teresa Podsiadły-Marczykowska[1], Rafał Zawiślak[2]

[1] Instytut Biocybernetyki i Inżynierii Biomedycznej PAN, ul.Trojdena 4, 02-109 Warszawa, tpodsiadly@ibib.waw.pl
[2] Politechnika Łódzka, Instytut Automatyki, ul. Stefanowskiego 18/22 90-912 Łódź, Polska, rafal.zawislak@p.lodz.pl

Abstract. The article presents assumptions, and results of a test of experimental verification of a hypothesis stating that the use of the MammoEdit - a tool that uses its own ontology and the embedded knowledge of mammography – increase the diagnostic accuracy as well as the reproducibility of mammographic interpretation. The graphical user interface of the editor was similarly assessed, as well as the rules for visualization which assists the radiologist in the interpretation of the lesions' character.

Keywords: mammogram interpretation, biomedical engineering, ontology, breast scanning, cancer diagnostics

1 Introduction

Mammography is well founded, effective imaging technique that can detect breast cancer in early and even pre-invasive stage, widely used in screening programs. However, the diagnostic value of mammography is limited by significant and up to 25% high, rate of missed breast cancers. Mammography is commonly seen as the most difficult imaging modality. Computer-aided Detection (CADe) systems aiming to reduce detection errors are now commercially available and are gaining increasing practical importance, but no serious attempts have been made to apply Computer-aided diagnosis (CADx) systems for lesion diagnosis in practical clinical situations. The paper presents an assessment of diagnostic accuracy of MammoEdit - an ontology-based editor supporting mammograms description and interpretation. In our opinion MammoEdit fills the gap among the existing solutions to the problem of supporting the interpretation of mammograms.

While working on the MammoEdit project, they used the ontology of mammography, created specifically for this need. It was used as a partial set of project requirements for the user interface and database that stores patients descriptions. This role of the ontology of mammography in the MammoEdit editor project was in line with literature indications for the use of ontology in IT systems [1].

2 The Ontology of Mammography

The mammographic ontology was created using Protege-2000 editor and the OWL language. Presented version of ontology includes: comprehensive, standard description patterns of basic lesion in mammography (masses and microcalcification clusters) enhanced with subtle features of malignancy, classes presenting descriptions of real lesions carried out on the basis of completed patterns and grades modeling diagnostic categories of the BI-RADS system.

Radiologists usually agree in their assessments of the diagnostic categories of BI-RADS (*Kappa=0.73*), when they have to deal with lesions presenting features typical for radiological image of breast cancer, clusters of irregular, linear or pleomorphic microcalcifications, or with spicular massess. The agreement in assessing is significantly lower when it comes to lesions from changes of the 4 BI-RADS[1] category (*Kappa=0.28*). Why does it happen? Admittedly, the BI-RADS system controlled vocabulary contains terms important with specific diagnostic value, but it lacks complete lesions definitions and does not stress the importance of early and subtle signs of cancer. Moreover BI-RADS system recommendations for estimating lesion diagnostic category are descriptive, incomplete and imprecise, they are expressed using different generality levels. The recommendations refer to the knowledge and experience of the radiologist and to typical images of the mammographic lesions. As a consequence, there is a large margin of freedom left for individual interpretation. Those conclusions are confirmed by large study based on 36 000 mammograms [2], showing wide variability in mammograms interpretation. While creating ontology, ambiguity in BI-RADS system recommendations caused by the lack of data, necessary to express restrictions on allowed values of the ontology classes modeling diagnostic categories of the BI-RADS system. The lack of these classes in the domain model clearly prevents the use of the ontology's classification mechanism to the diagnostic categories' assessment of the real lesions. The important part of this presentation is filling of the gap in imprecisely formulated BI-RADS system recommendations. This work was necessary to define through ontology classes modeling BI-RADS categories.

3 MammoEdit – a Tool Supporting Description and Interpretation of the Changes in Mammography

Analyzing reasons for errors it can be generally concluded that potential source of mammograms' interpretation mistakes is uneven level of knowledge and diagnostic abilities of radiologists and their subjectivity. Considering the above, it should be accepted that reduction of the interpretation errors can be obtained by describing mammograms using standardized protocols which include complete, standardized definitions of the lesions taking into account initial, subtle signs of malignancy and

[1] These are suspicious changes with ambiguous configuration of feature values, without evidence of malignancy

indication of feature values qualifying the nature of the changes during the description (potentially malignant or benign).

Suggested, by the authors of the article, method of errors' reduction is to support radiologists with specialized IT tool editor – MammoEdit. In the case of the MammoEdit, created ontology of mammography was used as a set of partial project needs for the user's interface and for database storing patients descriptions. The main project assumption of the interface was that the description of the mammogram should be done in the graphical mode, using such components as menu and button, while the text of the final description should be generated automatically. The choice of the graphical mode seemed to be obvious, because radiologists are used to images. Pictographic way of recording information enables both fast entry and immediate visualization. Over 300 pictograms illustrating the most important features of the changes in a vivid and explicit way have been created for the need of the editor. The collection of graphical objects (projects of buttons, icons, pictograms etc.) was gathered after the ontological model had been created and it reflects its content. The system of mutual dependence of entering/representing data is also based on the ontology. It let to obtain legibility and explicitness of the interface and blocked entering conflict data (exclusive). Detailed description of the editor can be found in [1].

4 The evaluation of the mammogram's interpretation supporting tool

The basic aim of this testing was experimental verification of the hypothesis which states that: "***the mammogram's interpretation, using the editor that systemizes the description process and indicates the values of the features that are significant for the accurate interpretation of their nature to the radiologist, increases diagnostic efficiency of the radiologist***".

Receiver Operating Characteristic analysis (ROC) has been used to assess testing. Quantitative assessment of diagnostic test performance involves taking into account the variability of the cut-off criteria in the whole variation of the parameter. ROC curves, which are created as a result of drawing the correlation of test's sensitiveness (Y axis) in Specificity function (X axis), are served for this purpose. Sensitiveness (SE) is a test ability to detect disease among sick patients, specificity (SP) is a test ability to exclude disease for healthy patients. The whole field contained under the ROC curve (AUC – *Area Under Curve*) is interpreted as a measure of test's effectiveness [3-4]. It has been decided to use *Multireader Receiver Operating Characteristic analysis* for testing MammoEdit. The mathematic model of the method takes into account typical for radiology sources of variability in image testing assessment and possible correlations among doctor's assessments within the same technique and also possible correlations among assessed techniques. Comparing the effectiveness of different image testing interpretation techniques involves carrying out an experiment, where trial testing consists of:

- Assessing techniques – imaging, interpretation or medical image processing algorithms;
- Assessing cases (the collection must include the study of pathological states and negatives);
- Radiologists who interpret testing cases using testing techniques.

These kinds of experiments, marked briefly as „technique x radiologist x case" are based on, so-called, factorial design [5]. The choice of the scale, which was used for assessing test results reflects substantive criteria of the medical problem that is being considered. [4]. Diagnostic category of the BI-RADS system matched by radiologists was chosen to represent the assessment of test results. The test which was performed included 80 mammograms that were difficult to interpret (20 negatives and 60 pathologies) The mammograms were described by three different groups of radiologists representing different levels of competence (a trainee, a specialist, an expert). They used two interpretation methods with and without MammoEdit editor. The mammograms came from a public data base DDSM[2] (*Digital Database for Screening Mammography*). Taking under consideration different origin of testing, doctors, and widely varied, uneven level of professional competence, model: random testing/random viewer was used.

The conditions of mammograms interpretation mimicked clinical environment as closely as possible. The mammograms were presented to the radiologists on the medical screens with the MammoViewer application, which function for testing had been reduced to an advanced medical browser, in a dark room. The assessment of mammograms was held during many sessions that lasted from 1 to 3 hours. The speed and time of sessions were under doctors control, the breaks or suspensions were taken on their requests. During one session the doctor performed only one type of mammogram's assessment – with or without the assisting tool. Changes were assessed on the basis of breast images in two basic projections, without any additional diagnostic projections, clinical trials, patient's medical records or previous images to compare. Before the first mammograms reading, the meaning of scales was explained and, during testing, the scales description was available at the workplace. The minimum time interval between the interpretation of the same testing with and without MammoEdit was estimated on the basis of literature. It is assumed that it should be minimum 4-5 weeks. In the experiment it ranged from 7 to 12 weeks.

The DBM method of experiments assessment MRMC ROC [6] was used to calculate indicators of diagnostics performance, while the PropRoc method [7], which provides reliable indicators estimation for the analysis of the small trial testing and discrete assessment scale, was used to estimate ROC curves indicators.

The use of MammoEdit to support mammograms interpretation has raised the average diagnostic performance (statistically significant increase, significance level less than 0.05). Detailed results are presented in Table 1.

---

[2] The material comes from four medical centers in the USA (Massachusetts General Hospital, University of South Florida, Sandia National Laboratories and Washington University School of Medicine).

Table 1. The result of MammoEdit editor's influence on mammograms interpretation performance with 3 groups of radiologists

lesion type	interpretation methods	AUC	SL	AC%	SE%	SP%
masses	without MammoEdit	0,610	0,016	55,0	20,0	94,2
	with MammoEdit	0,853		79,3	63,3	97,1
microcalcifications clusters	without MammoEdit	0,677	0,728	71,3	36,1	96,1
	with MammoEdit	0,730		75,9	44,4	98,0

The rise of indicators mentioned above was accompanied by the reduction of the deviation/variation of the standard field under the ROC curve.

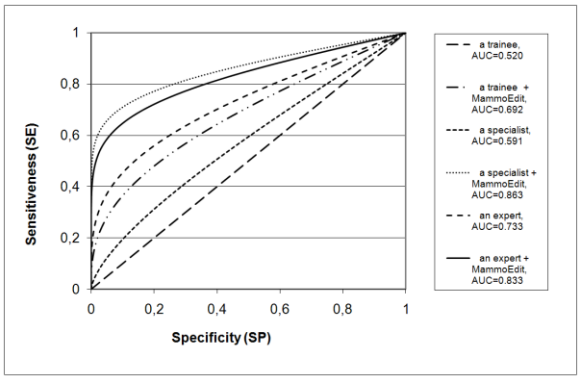

Fig. 1. MammoEdit's influence on the diagnostic effectiveness of the radiologists

The biggest increase of the diagnostic effectiveness was observed in case of a specialist radiologist, the smallest – in case of an expert radiologist. It proves bad influence of routine on the quality of mammograms assessment. The influence of MammoEdit on the variation of radiologists diagnostic opinions was assessed comparing the value of Kappa statistics achieved in the I and II testing period. The variability of opinions was assessed for the pair of radiologists and for the group of all radiologists, separately for microcalsifications, tumors and for both types of pathologies.

The use of MammoEdit for the mammograms interpretations has increased diagnostics opinions consistency for the group of radiologists for both types of changes masses and microcalcification clusters and for each of the pathologies; the biggest increase in consistency has been observed for microcalsifications clusters. MammoEdit has also influenced the increase of opinions consistency in all pairs of viewers and in all groups of changes.

5 Conclusion

The use of MammoEdit for mammograms interpretation support has raised the average diagnostic performance for the group of radiologists, other ROC analysis indi-

cators have also increased. It brings the conclusion, that initially, in the trial testing, the hypothesis of MammoEdit usefulness as a tool for mammograms interpretation support has been confirmed. The level of diagnostic knowledge representation in the ontology of mammography has also been estimated as a satisfactory one. Thorough analysis of radiologists performance, in distinction of the types changes, has proved a difference in supporting tumors and microcalsifications interpretation effectiveness (to the detriment of tumors).

The next result of the experiment is the initial assessment of the logical correctness of the class construction method. It consisted of comparing BI-RADS category of selected testing cases in DDSM data base and the assessment of the grades of the same changes performed using the MammoEdit and automatic evaluation of the changes grade on the basis of ontology classification. The experiment has proved that BI-RADS category assessment of the changes obtained with the help of the ontology's inference module is adequate to the ones in DDSM data base. What is interesting, the use of ME editor has improved asses compliance with the standard, even for the group of expert radiologists. MammoEdit supports just a part of radiologists work, we need to gain some knowledge about mathematical descriptors of all basic types for feature values of the mammograms changes, tumors and microcalsification clusters to apply it in clinical practice and integrate it with the application CAD. We also need to connect changes detection and their verified description, and finally widen the application with the cooperative function with the ontology of mammography in order to receive an automatic verification of the BI-RADS category of the change. The complement of the integrated application functionality is: administrative features (Radiology Information System), data management (data base functions) and security, finally integration with a workstation and PACS (Picture Archiving and Communication System).

References

1. Podsiadły-Marczykowska T., Zawiślak R.: The role of domain ontology in the design of editor for mammography reports, "*Dziedzinowa ontologia mammografii w projektowaniu edytora raportów*", Bazy Danych: Struktury, Algorytmy, Metody 2006, chapter 37
2. Miglioretti DL, Smith-Bindman R, Abraham L, Brenner RJ, Carney PA, et al. Radiologist Characteristics Associated With Interpretive Performance of Diagnostic Mammography. J Natl Cancer Inst 2007; 99:1854-1863
3. Swets JA.: ROC analysis applied to the evaluation of medical imaging tests. Invest Radiol 1979: tom 14: s.109-121
4. Zhou XH, Obuchowski NA, Obuchowski DM.: Statistical Methods in Diagnostic Medicine. New York: John. Wiley and Sons, 2002.
5. Obuchowski NA, Beiden SV, Berbaum KS, PhD, Stephen L. Hillis SL.: Multireader, Multicase Receiver Operating Characteristic Analysis: An Empirical Comparison of Five Methods, Acad Radiol 2004; 11:980–995
6. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA.: Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. Acad Radiol 1998; 5: s. 591–602.
7. Pesce LL, Metz CE,: Reliable and computationally efficient maximum-likelihood estimation of "proper" binormal ROC curves, Acad Radiol. 2007; 14(7): s. 814–82