Breast screening: understanding case difficulty and the nature of errors

Breast screening: understanding case difficulty and the nature of errors

Leng Dong*, Yan Chen, Alastair G. Gale

Applied Vision Research Centre, Loughborough University, Loughborough, UK

ABSTRACT

In the UK all screeners undertake the PERFORMS scheme where they read annual sets of challenging cases. During this assessment, they give each case a confidence rating on whether it should be recalled. If they decide to recall a case, they also indicate the center of any key mammographic features on a display of the relevant mammographic case view. Expert radiological opinion defines what the key abnormalities (targets) are in any case. Data can then be analyzed using ROC and JAFROC approaches, and particularly for the latter, assessing whether a user has correctly located a feature or not is important. Using image pixel information alone it is possible to delineate correct localization of an abnormality from an incorrect location by defining an area of interest. To explore such location information in more detail, data from the last year of the PERFORMS scheme were reanalyzed and the location responses for each of the 675 participants on 120 screening cases examined. Additionally, expert radiological opinions had been garnered for various reasons, including accurately delineating any abnormalities. An algorithmic approach is developed which assesses whether users' indications should be included as correct abnormality identification or not, based on the feedback location information of all participants' indicated locations and the relative position of an indicated location to the abnormality. This approach is proposed to be superior to simple pixel distance approaches which measure a fixed distance from the centre of a target to the user's indicated location. The approach adds to the experimenter's repertoire of tools when examining user errors and case difficulty in medical imaging research.

Keywords: performance, case difficulty, area of interest, breast screening, error margin, JAFROC

1. INTRODUCTION

Identifying early signs of breast cancer efficiently is a complex task and is affected by numerous factors. For instance, whilst the density of the screening mammogram is related to a higher probability of breast cancer, the very nature of a dense mammographic image itself can render it hard for a radiologist to perceive any early malignant signs with ease. One of our long term research projects is concerned with annually assessing the comparative performance of breast screeners in the UK and previously we have reported an investigation into what makes a particular mammographic case difficult for large numbers of radiologists and advanced practitioners which found non-simplistic findings[1]. Expert radiologists and less experienced radiologists appeared to rate case difficulty differently and various possible explanations for this were put forward. One of the factors which impacts on case difficulty is how hard it is for a radiologist to accurately identify and locate an abnormality.

All UK breast screeners undertake the PERFORMS[2,3] scheme annually where they examine recent screening cases, receiving immediate feedback. Confidence ratings on whether a suspicious case should be recalled are used to demonstrate each participant's performance. Location information of suspicious abnormal areas is also recorded. From this information ROC and JAFROC analyses can be undertaken. Data from several rounds of the scheme show that even when participants correctly recalled a case, not all of them appropriately marked the correct target abnormality location. Sometimes abnormalities were slightly missed, or clearly missed, and on other occasions other potential abnormality sites were marked. Such errors may be related to the perceived difficulty of a case by the user. To examine the underlying reasons for this behavior a detailed study was carried out on the locations identified by all UK screeners on the latest self-assessment scheme cases to try to understand in detail why some abnormal areas do not seem to be identified as would be expected.

In the latest round of the PERFORMS scheme some 675 UK screeners examined 120 challenging screening cases which had been specially selected to stretch their skills. Each case comprises two mammographic views or images, the Medio-Lateral Oblique (MLO) and the Crania-Caudal (CC) respectively of each breast. Thus each participant examined 240

*l.dong@lboro.ac.uk

Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment,
edited by Craig K. Abbey, Claudia R. Mello-Thoms, Proc. of SPIE Vol. 8673, 867316

breast images in total and for these particular screening cases 90 of these images contained one to three key mammographic features.

In establishing each case set in the scheme the cases are selected independently by a panel of expert breast screening radiologists who examine each case on their mammographic workstations and annotate any mammographic features of interest using a large interactive graphics display. From these various annotations an agreed delineated boundary for every abnormal appearance on all images in the case set is achieved. This is then further checked independently by a consultant breast screening radiologist. Subsequently, when participants take part, then their indicated abnormality location information is compared to the expert panel's judgment of locations as part of the performance scoring process.

In the PERFORMS scheme participants examine the digital mammographic images on their clinical workstations and record their decisions using bespoke software which runs on a laptop. In recording their decisions they indicate a location on mimic images of the case being examined where they locate any abnormality by identifying the abnormality centre. Clearly the accuracy of placement of their locations is affected by the size of these images, the care taken by the participant in recording the location and other related factors. We have previously studied the accuracy with which naive participants can locate a target cross on a mammographic image transcribed to another image of differing scale and detail[4]. Increasing image scale and detail aided accurate transcription. Other approaches to examining accuracy of localization placement have employed simple pixel distance from the centre of some target.

For this investigation the location information of all the participants from the last round of the PERFORMS scheme were extracted from our database and plotted on to the relevant view of all the case images. The expert panel's delineated areas were also plotted as an Area Of Interest (AOI). As a starting point, if a participant's response location fell within the AOI then this was considered as a correct marking of that abnormality. Otherwise, the location marked by the participant was deemed to be incorrect. If a participant identified more than one feature in an image, as long as one indicated location fell within the AOI area, then this participant's response to that abnormality was counted as correct. After visualization of all the location data for all cases then summary data showing the percentage of correctly marked locations for each image were constructed. This then indicated if all participants identified the correct location for each abnormality. The results showed that among all the 90 features the overall average correct mark rate was 85.18%. This means that on average, when a feature was identified and recalled by a number of participants then 85.18% of them also marked the correct abnormality location. To illustrate this, figure 1, a right CC view of case 217, illustrates the visualized result of determining the abnormality localization for this, the most difficult image out of these particular images, as only 13.69% of participants who recalled this case also marked the correct location. In this figure all the white dots inside the AOI polygon (as determined by the experienced radiological panel) indicate those marks which are in the correct location. The other dark dots are users' location marks made in the wrong location. Clearly participants here perceived other features away from the AOI which they had considered to be important. Figure 2 shows the visualization of the left CC view of case 204 which proved to be the easiest malignant case and which scored a correct rate of 99.82%.

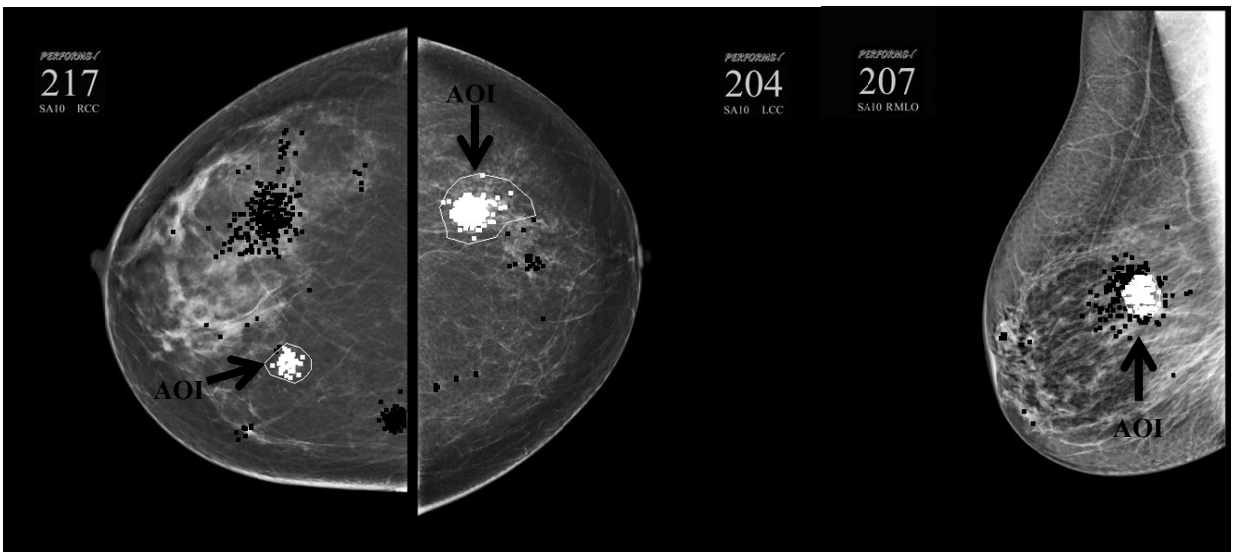

Figure 1. 217RCC Figure2. 204 LCC Figure3. 207RMO

Furthermore, figure 3 shows the localization of all marks on the right MLO view of case 207 which only had a correct rate of 55.12%. However, from the figure it can be seen that a lot of the users' marks, which could be considered as error, actually just fall outside the AOI but are still very close to the AOI boundary of the abnormality. Those marks should possibly also be included into the 'correct mark' category which would mean that the actual correct rate here should be much higher than 55.12%. The question then is how to achieve a sensible approach to account for such 'almost correct' location marks without including as correct any incorrect indicated locations?

To understand this type of situation better, an approach was proposed of using a defined Error Margin (EM) around the AOI. With the help of the margin, participants' marks that were initially considered as errors but were also actually very close to the AOI abnormality boundary, thus falling within the EM, would still be counted as correct. At the same time, a means of preventing actually wrong location marks, falling within such an EM, to be erroneously counted into the correct category should also be guaranteed. Thus the concept was of generating a family of EMs of increasing sizes and experimentally determining which is the best and most appropriate EM to use. The size of EM was also expected to differ depending upon the particular mammographic feature under investigation.

2. METHOD

Some of the mammographic images contained more than one abnormality and some of these abnormalities are very close to each other. Therefore, in order to avoid overlying by the various EMs it was decided to perform the analysis for each abnormality individually. That is, for each AOI an individual image was used to visualize all of the participants' marks. At this analysis stage those marks which are very far away from the abnormality area, which might be marks on another abnormality, or totally erroneous marks, are ignored. The marks inside the AOIs and the marks just outside the AOIs are the initial concern here.

Consequently, 106 abnormalities were individually extracted from the 90 malignant images. For each abnormality the AOI was plotted on the image from the expert panel data. Abnormality AOIs are denoted by a number of points which then comprise an irregular polygon.

The Error Margin (EM) is defined as an area with the same shape as the abnormality AOI, but of a larger size around the AOI. To achieve this, the centroid of the AOI was determined. In our system each AOI is represented as a non-self-intersecting closed polygon and each polygon is defined by n vertices (x_0, y_0), (x_1, y_1)..., (x_{n-1}, y_{n-1}). Then the centroid of the polygon (C_x, C_y) can be calculated by the following formulas[5]:

$$C_x = \frac{1}{6A}\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} + x_{i+1} y_i) \qquad \text{(Formula 1)}$$

$$C_y = \frac{1}{6A}\sum_{i=0}^{n-1}(y_i + y_{i+1})(x_i y_{i+1} + x_{i+1} y_i) \qquad \text{(Formula 2)}$$

And where A is the polygon's area,

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i y_{i+1} + x_{i+1} y_i) \qquad \text{(Formula 3)}$$

Then a distance multiplier is set which multiplies the distance from the centroid of the polygon to each polygon side, so as to obtain a family of larger polygons which mirror the AOI shape.

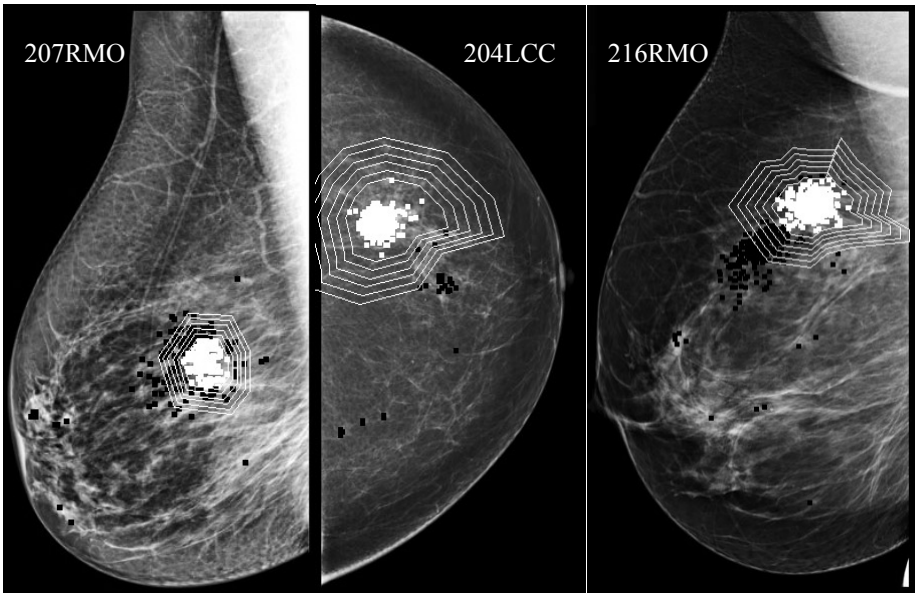

Figure 4. Images 207RMO, 204LCC and 216RMO shown with error margins

Five distance multipliers were empirically selected, which were 1.2, 1.4, 1.6, 1.8 and 2.0 respectively, so as to create five margins. Images 207RMO (figure 2) and 204LCC (Figure 3) are shown together in figure 4 as examples demonstrating what those calculated EMs look like, together with the various participants' marks on these images. For Image 207RMO, it can be clearly seen that by adding five EMs to the original AOI, more marks near the AOI are encompassed by those margins which could be counted as correct hits. Meanwhile, for image 204LC, not too many marks were found in those margins, so a larger sized EM may not be necessary for this image. After doing this for every abnormality in every image, a new problem arises which is illustrated by image 216RMO. This has an abnormal feature and many participants marked the correct location. However, a lot of participants have also marked a location at the left bottom side of the AOI where the experienced radiological panel did not think a key abnormality existed. By adding 5 EMs to this AOI, it is inevitable that these wrong marks are encompassed as if they were correct hits of the target abnormality.

The three images above demonstrate situations which should be treated differently when 5 EMs are added to each AOI. A larger sized EM is necessary for image 207RMO to include as many potential correct marks as possible. However, it is better not to add a large EM on to images 204LCC and 216RMO. For image 204LCC, few marks were found in these EMs and for image 216RMO increasing the margin size will probably cause many marks in a wrong location to be wrongly scored as hits. In order to understand the nature of these situations, the number of new marks found in each EM for these three images were counted (Table 1). We believe that the relationship between the number of marks in EM(n) and the number of marks in EM(n-1) would help us find the best fit EM to select in each case. EM 0, which has a distance multiplier of 0.8, is introduced to assist in the analysis of the number of marks in EM 1. For 207RM, the number of marks kept decreasing from EM 1 until they reached EM 4. In contrast, for 204LC no marks were found in EM 1. For 216RM the number of marks in each EM increases from EM1 to EM5. So a principle of finding the best fit EM is established based on these findings: -

By searching the number of marks in each of the 5 EMs,

1) As long as the number of marks in EM(n) is larger than that of EM(n-1), then EM(n-1) will be selected as the best fit EM.
2) If the number of marks in EM(n) is zero, then EM(n-1) will be selected as the best fit EM.

Note: if EM 0 is selected as the best margin, this is actually the original AOI given by the panel of experienced radiologists.

Then a computer aided method, based on the above algorithms, was developed to help identify the best EM for each image and abnormality.

	EM 0	EM 1	EM 2	EM 3	EM 4	EM 5
207RMO	38	31	29	18	11	13
204LCC	4	0	2	1	0	0
216RMO	27	18	25	32	33	34

Table 1. Number of marks found in each EM for images 207RMO, 204LCC and 216RMO

3. RESULTS

After applying the best fit margin algorithm to all the features, each feature was assigned a best margin value which indicated the margin size. For the features stated above, EM 4 was selected as the best margin for 207RMO as the number of marks kept decreasing from EM1 to EM 2 until it reached EM 5. The number of marks in EM 5 is 13 which is larger than the number of marks in EM4. So here EM 4 is chosen to be the best fit margin. For image 204LCC the best margin was found to be EM 0 (which is the original AOI) as the number of marks in EM1 is zero. For image 216RMO then EM 1 is the best margin as the number of marks in EM 2 is higher than EM 1.

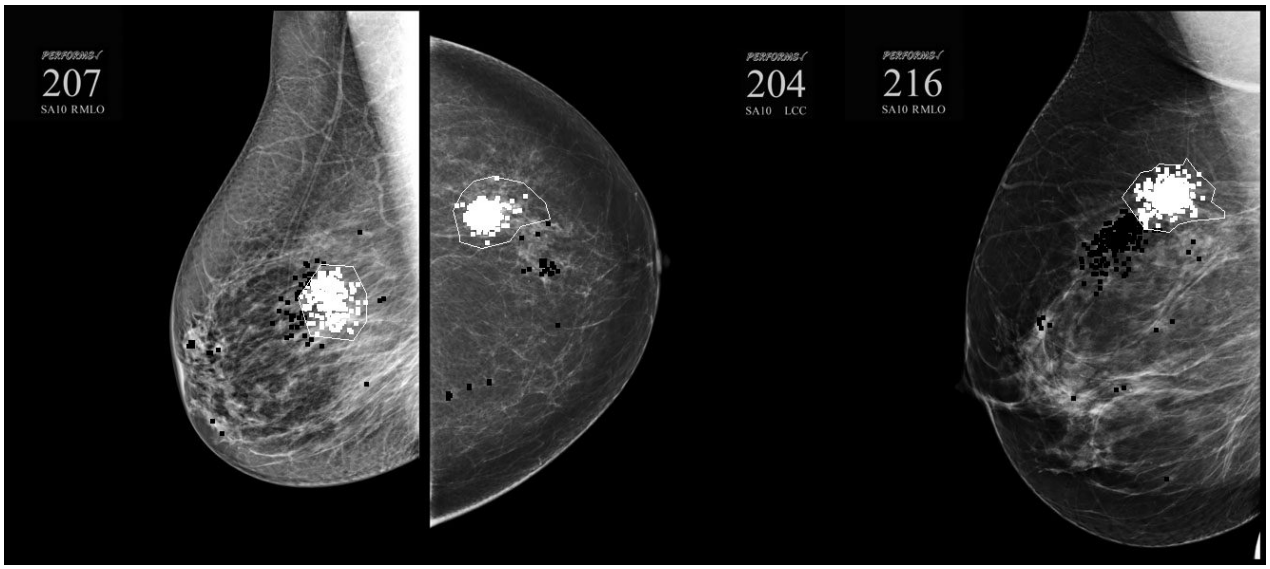

Figure 5. Best fit margins on images 207RMO, 204LCC and 216 RMO

Figure 5 shows the best fit EM for images 207RMO, 204LCC and 216RMO annotated on each image. Margin 4 with the distance multiplier of 1.8 was chosen as the best fit margin for Case 207RMO. After adding Margin 4 to the original abnormality AOI area of Case 207RMO, the rate of correct marks increased to 85.51%, which is a considerable improvement from its previous value (55.12%). For Case 204LCC, margin 0 was found to be the best fit EM, so in this case the abnormality AOI remains the same with the correct mark rate remaining the same as before (99.82%). Case 216RMO with EM 1 has its correct rate increased from 98.15% to 99.50%. To examine the overall results of this approach, the average value of the correct mark rate for all 90 malignant images was calculated. Overall an average of 91.98% correct mark rate was achieved after the best fit margin algorithm was adopted for each malignant case. The average Correct Rate of all cases with features increased significantly ($p<.05$) after the best margin algorithm was applied (figure 6).

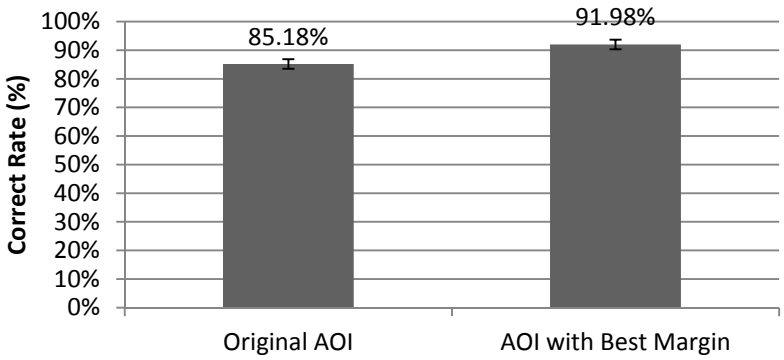

Figure 6. The average Correct Rate of all cases with features before and after the best margin algorithm was applied

The algorithm was then applied separately in this manner to all of the abnormalities in the images. Each malignant feature was assigned a margin value indicating which margin was chosen to cover as many potential correct marks as possible. To demonstrate how effective this algorithm performs in finding the best fit margins, the relationship between the size of abnormality and the best fit margin was analyzed. The size of each abnormality AOI was calculated by Formula 3 and categorized into four types: small, medium, large and extra-large. The size of the AOI relates to the type of feature, for instance a diffused mass will generally have a large extent and a spiculate mass may be quite small. It was found that these AOIs varied in pixel size (on the mimic image display) from an area denoted by pixels varying from 297 to 16,277. Here, areas less than 1,000 pixels in size are considered as "small", areas less than 2,000 and larger than 1,000 pixels are considered as "medium" and areas bigger than 2,000 pixels and less than 5,000 pixels are considered as "large", areas larger than 5,000 pixels was considered as "extra-large". Using these values then there are 25 small abnormalities, 34 medium abnormalities, 31 large abnormalities and 16 extra-large abnormalities. For each size category, the average error margin size (0-5) was calculated and is as shown in Figure 7.

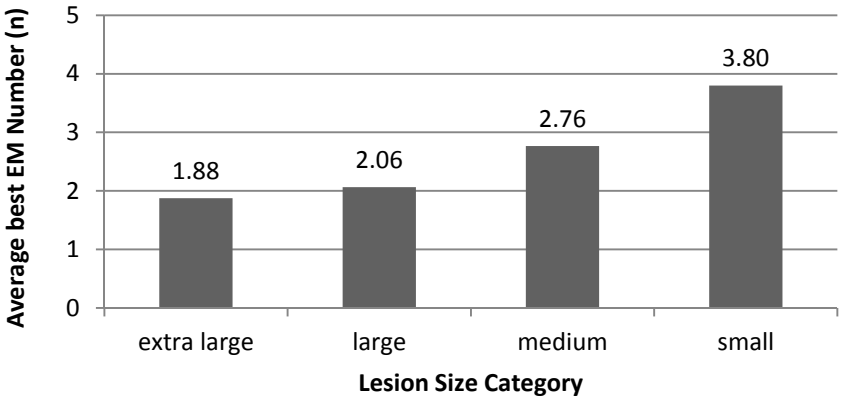

Figure 7. The average best EM number in each size category

It can be seen from Figure 7 that small abnormalities tends to have larger EMs and large abnormalities a smaller margin. This agrees with what would be expected - when the AOI is small, participants will have greater difficulty in accurately making an appropriate localization mark at/near the centre of the abnormality. In this situation, it is better to have a larger margin. On the other hand, when an abnormality is large and is easily visible then a small EM (or none at all) is needed as it is easy to correctly mark the abnormality location. A large margin here may increase the risk of counting many totally wrong location marks into a 'correct' category.

Besides the size of an abnormality, the relationship between abnormality feature types and the best fit margin is also of interest. Table 2 shows the different types of feature and how many abnormalities for each feature type there were in these images. The average size of each feature is shown in pixels in Figure 8 as measured on the displaying image. Not considering the type 'other' (which encompasses various other types of feature) then asymmetry was the largest average size and ill defined mass was the smallest average size. Figure 9 shows the relationship between feature type and the best

fit error margin number. Mainly it follows the 'smaller abnormality, larger margin' principle with only two exceptions. The 'spiculate mass' feature has an AOI average size smaller than that of 'well defined mass' but the average best fit EM number for it is still smaller than for the 'well defined mass'. The same situation occurs with the feature 'other' and 'asymmery' – this is not too surprisig as 'other' can encompass several different feature appearances. These two exceptions indicated that the best fit margin is not only affected by the size of the abnormality AOI. Other factors such as the feature type also play an important role.

Feature Type	Count of Abnormality
Architectural Distortion	9
Asymmetry	7
Calcification	37
Ill Defined Mass	10
Other	5
Spiculate Mass	26
Well Defined Mass	12

Table 2. Number of abnormalities in each feature type

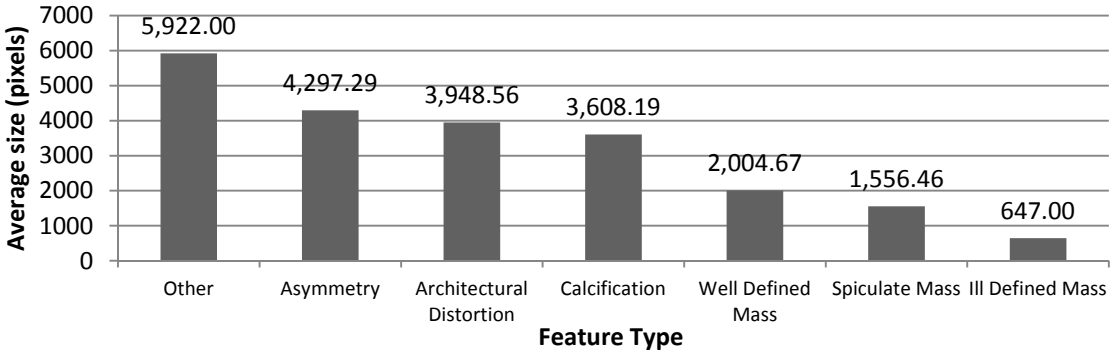

Figure 8. Average size in pixels for each feature

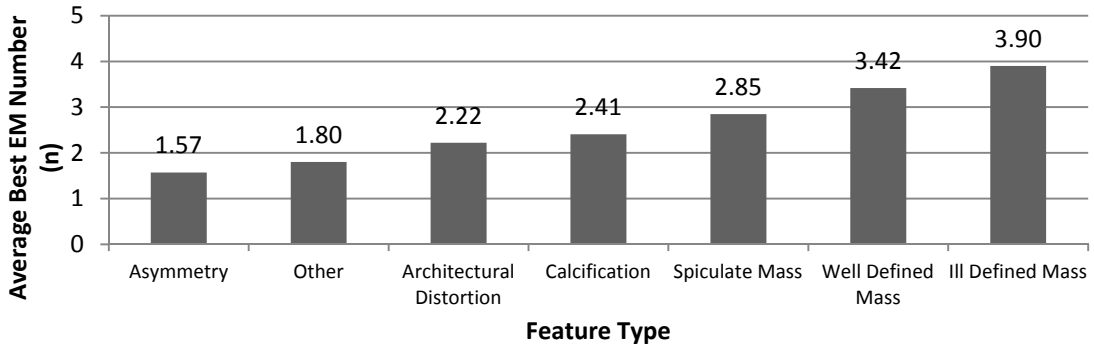

Figure 9. The average of best EM number for each feature

4. DISCUSSION & CONCLUSION

A common problem in medical imaging research is trying to interpret whether an individual has correctly identified a target abnormality or not. Typically in various studies the investigator asks the user to mark on the medical image display, or some other associated recording display, the site of the abnormality. There are various errors associated with such actions involving factors such as the size of the display being employed, the size of the abnormality being marked, user experience, etc.. Often a very simplistic approach of assessing whether the user has correctly 'hit' the target is taken

of measuring the number of pixels away from the centre of the target the indicated location is and using a fixed number of pixels to delineate an area of interest (AOI) around the target centre to assess whether a hit has been achieved.

Here, the interest was specifically in this type of task in the domain of breast screening. An approach is developed based on firstly having expert radiologists annotate accurately the outline area of interest (AOI) of the abnormality and then making judgements about users' indicated location marks as judged against this AOI. The best fit error margin (EM) algorithm was developed to build annuli polygons around the AOI and an empirical method used to judge the best fit error margin for each mammographic feature of interest. It is argued that this is a better solution than using a simple distance measure of a user's indicated location to the centre of an abnormality and helps in understanding some of the reasons for errors when difficult cases are examined.

The method, as described and implemented here has limitations. The key issue is that this algorithm is based on analyzing the data of all participants which means it can only be run after all users have taken part. This is fine for small scale experiments but in our PERFORMS situation it can take several months for all participants to complete a test scheme. This makes it impossible for us to give participants immediate feedback using the benefits of this approach on whether they had marked correct abnormality locations. To provide such feedback the abnormality AOI can be used. Despite this drawback, the approach seems very appropriate for analyzing free-response receiver operating characteristic (FROC) observer performance data. Ongoing research is examining whether a suitable solution can be found as to whether to include marked abnormality locations falling outside an abnormality AOI as a correct hit of that abnormality without requiring all participants take part.

5. ACKNOWLEDGEMENTS

This work is partly supported by the UK National Health Service Breast Screening Programme. PERFORMS is a registered trade mark.

REFERENCES

[1] Gale A.G., "PERFORMS – a self-assessment scheme for radiologists in breast screening," Seminars in Breast Disease: Improving and monitoring mammographic interpretative skills, 6(3), 148-152, (2003)

[2] Gale A.G., "Maintaining quality in the UK breast screening program", In D.J. Manning & C. Abbey (Eds.) Proc. SPIE Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment. 7627, 1-11 (2010).

[3] Scott H.J., and Gale A.G. "Breast screening: when is a difficult case truly difficult and for whom?" In Image Perception, Observer Performance, and Technology Assessment. MP. Eckstein and Y Jiang (eds.) Proceedings of SPIE, Vol. 5749, 2005

[4] Hatton J., Wooding, D.S. & Gale A.G., "Accuracy of transcribing locations on mammograms: implications for the user interface for recording and assessing breast screening decisions." In Medical Imaging 2003: image perception and performance. In Krupinski E. (ed.) Proceedings of SPIE Medical Imaging conference. Vol. 5034, 2003.

[5] Bourke, P., "Calculating the area and centroid of a polygon." http://local.wasp.uwa.edu.au/-pbourke/geometry/polyarea (1988).