Llobet Azpitarte, R.; Pollán, M.; Antón Guirao, J.; Miranda-García, J.; Casals El Busto, M.; Martinez Gomez, I.; Ruiz Perales, F.... (2014). Semi-automated and fully automated mammographic density measurement and breast cancer risk prediction. Computer Methods and Programs in Biomedicine. 116(2):105-115. doi:10.1016/j.cmpb.2014.01.021.

Semi-automated and fully-automated mammographic density measurement and breast cancer risk prediction

Rafael Llobet[1,*], Marina Pollán[2,3], Joaquín Antón[1], Josefa Miranda-García[4,5], María Casals[4,5], Inmaculada Martínez[4,5], Francisco Ruiz-Perales[4,5], Beatriz Pérez-Gómez[2,3], Dolores Salas-Trejo[4,5], Juan-Carlos Pérez-Cortés[1]

Instituto Tecnológico de Informática, Ciudad Politécnica de la Innovación, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia SPAIN

Abstract

The task of breast density quantification is becoming increasingly relevant due to its association with breast cancer risk. In this work, a semi-automated and a fully-automated tool to assess breast density from full-field digitized mammograms are presented. The first is based on a supervised interactive thresholding procedure for segmenting dense from fatty tissue and is used with a twofold goal: for assessing mammographic density (MD) in a more objective and accurate way than via visual-based methods and for labeling the mammograms that are later employed to train the fully-automated tool. Although most automated methods rely on supervised approaches based on a global labeling of the mammogram, the proposed method relies on pixel-level labeling, allowing

*Corresponding author. Tel: +34-96-3877069 Fax: +34-96-3877239

Email addresses: rllobet@iti.upv.es (Rafael Llobet), mpollan@isciii.es (Marina Pollán), j.anton.guirao@gmail.com (Joaquín Antón), miranda_mjo@gva.es (Josefa Miranda-García), casals_mar@gva.es (María Casals), martinez_inm@gva.es (Inmaculada Martínez), pruizp@comv.es (Francisco Ruiz-Perales), bperez@isciii.es (Beatriz Pérez-Gómez), salas_dol@gva.es (Dolores Salas-Trejo), jcperez@iti.upv.es (Juan-Carlos Pérez-Cortés)

[1]Institute of Computer Technology. Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia SPAIN.

[2]National Center for Epidemiology, Carlos III Institute of Health, Monforte de Lemos 5, Madrid 28029, Spain.

[3]Consortium for Biomedical Research in Epidemiology and Public Health (CIBER en Epidemiología y Salud Pública - CIBERESP), Carlos III Institute of Health, Monforte de Lemos 5, Madrid 28029, Spain.

[4]Valencian Breast Cancer Screening Program, General Directorate of Public Health, Valencia, Spain.

[5]Centro Superior de Investigacin en Salud Pblica CSISP, FISABIO, Valencia, Spain.

better tissue classification and density measurement on a continuous scale. The fully-automated method presented combines a classification scheme based on local features and thresholding operations that improve the performance of the classifier. A dataset of 655 mammograms was used to test the concordance of both approaches in measuring MD. Three expert radiologists measured MD in each of the mammograms using the semi-automated tool (DM-Scan). It was then measured by the fully-automated system and the correlation between both methods was computed. The relation between MD and breast cancer was then analyzed using a case-control dataset consisting of 230 mammograms. The Intraclass Correlation Coefficient (ICC) was used to compute reliability among raters and between techniques. The results obtained showed an average ICC=0.922 among raters when using the semi-automated tool, whilst the average correlation between the semi-automated and automated measures was ICC=0.838. In the case-control study, the results obtained showed Odds Ratios (OR) of 1.38 and 1.50 per 10% increase in MD when using the semi-automated and fully-automated approaches respectively. It can therefore be concluded that the automated and semi-automated MD assessment present a good correlation. Both methods also found an association between MD and breast cancer risk, which warrants the proposed tools for breast cancer risk prediction and clinical decision making. A full version of the DM-Scan is freely available.

Keywords: Mammographic density, automated density assessment, computer-aided diagnosis, computer image analysis, breast cancer risk

---

1. Introduction

Breasts are composed of fibroglandular or dense tissue (FGT) and fatty tissue. On a conventional mammography, FGT appears brighter than fatty tissue, due to the higher X-ray attenuation of the former. Mammographic density (MD) is computed as the proportion of FGT in the breast. Some examples of mammograms of different densities can be seen in Figure 1.

MD is associated with an increased risk of developing breast cancer [1, 2].

This association is more important than almost all other risk factors for the disease. Women with dense breasts are at four-to-six-fold higher risk than those with primarily fatty breasts [2, 3, 4].

In addition to the increased risk of developing breast cancer in women with high MD, high breast density impedes its diagnosis due to the fact that the high X-ray attenuation of dense tissue can obscure a tumor. This means the sensitivity of mammography for detecting breast cancer can be significantly reduced in the case of dense breasts [5]. Retrospective studies have shown that in current breast cancer screening 10% to 25% of tumors are missed by radiologists [6, 7] and some of these false negative results can be explained by higher breast density.

Since the association between mammographic density and an increased risk of breast cancer was first discovered, several metrics have been proposed to classify this parameter: Wolfe's four parenchymal patterns [1, 8], Tabar's five patterns [9] and, more recently, Boyd and BI-RADS categories. The Boyd scale [10] classifies MD in six categories: A:0%, B:$1-10$%, C:$10-25$%, D:$25-50$%, E:$50-75$% and F:$75-100$%, while BI-RADS [11] divides density into four categories: A:$0-25$%, B:$25-50$%, C:$50-75$% and D:$75-100$%. However, all these methods are based on visual analysis of the mammogram and present two major drawbacks: the subjectiveness of the categorization and the difficulty of assigning a category when the mammogram is near the boundary between two categories.

Computer-assisted measurement of breast density has been studied in the last few years in an attempt to obtain more objective risk assessments. Several semi-automated methods based on interactive thresholding techniques that compute the percentage of the dense tissue over the segmented breast area have been proposed [10, 12], as well as some fully-automated methods. Karssemeijer [13] developed an automated method in which features are calculated from gray level histograms computed in different regions of approximately equal distances from the skin line, and then classified using the k-nearest neighbor (k-NN) rule. Saha et al. [14] describe a method using a scale-based fuzzy connectivity approach.

Klifa [15] et al. present a segmentation technique based on fuzzy clustering to quantify breast density from MRI data. Oliver et al. [16] suggest an approach based on gross segmentation and the extraction of texture features of pixels with similar tissue appearance. This work is continued in [17] in which a Fuzzy C-Means clustering approach is used for gross segmentation. Muhimmah et al. [18] use a feature extraction scheme based on a multiresolution histogram. Heine et al. [19] first perform FGT segmentation using a wavelet high pass filter. The same group later propose a new measure called *variation measure*, which is calculated as the standard deviation of the pixel values within a specific region of the breast, and find an association between this measure and the risk of breast cancer [20]. Manduca et al. [21] analyze several textural features of breast tissue and find that they predict breast cancer risk at the same magnitude as MD. Li et al. [22] propose a method based on a machine learning approach in which the MD obtained using the semi-automated tool presented in [10] is used as ground truth. A variety of measurements obtained under 15 thresholding methods are then used as features to learn the model. In general, segmentation of non-fatty tissue in mammograms appears to be more difficult than one might think, due to large differences in appearance between different parenchymal types [13]. 3D approximations to breast density assessment have also been addressed [23, 24, 25], but there is no evidence that volumetric density measurements are more strongly related to breast cancer risk than 2D measurements [26]. Moreover, very high correlation has been found between breast density measured from the cranio-caudal(CC) and mediolateral oblique (MLO) views [27], which suggests that measuring density from a mammography (a 2D projection of the breast) is an acceptable simplification.

Regardless of the method employed to classify and estimate mammographic density, this measure is of major importance as it could influence the choice of alternative screening paradigms, such as shortening the intervals between mammograms, using other methods such as magnetic resonance imaging (MRI), or to signal the need for more careful interpretation of the mammogram, such as double-reading.

This work presents a computer-assisted (semi-automated) and a fully-automated tool for the assessment of MD (DM-Scan). The aim of the semi-automated tool is twofold. Firstly, it can be used to assess MD more objectively and accurately than visually-based methods. Secondly, it is used to label each pixel of the mammogram in one of two possible classes: dense/fatty. This process, together with a feature extraction scheme, is later used to train a model to estimate MD automatically, without human intervention.

In a previous work [28], MD estimates using DM-Scan were compared with those obtained by visual inspection and by using Cumulus [10, 29]. In this work, a comparison is made between the concordance of semi-automated and fully-automated assessment tools in calculating MD, after which the correlation between MD and breast cancer risk is analyzed.

2. Materials and methods

2.1. Data sets

Two different sets of mammograms, known as DDM (655 mammograms) and Case-Control (230 mammograms), were used in this study. The DDM set was used to analyze the concordance of each of the three methods in assessing MD: visual, semi-automated and fully-automated. Unfortunately, no prospective information related to cancer development was available in this set of mammograms, so that breast cancer risk could not be analyzed in this case. For this purpose, the Case-Control set was used instead.

The DDM set consists of 655 mammograms, a subset of those used in the DDM-Spain project [30, 31]. This was a cross-sectional study which recruited 3,584 women aged 45-68 years at 7 screening centers within the Spanish breast cancer screening network. Informed consent was obtained from all participants, who agreed to their left cranio-caudal mammograms (single view) being used for study purposes. Mammograms from women participants at two screening centers equipped with full-field digital mammography machines, i.e. a total of

655 mammograms, were used in the study. The devices used in these two centers were a Hologic-Lorad M-IV and a Siemmens MAMMOMAT NovationDR.

The Case-Control set was formed from all breast cancer cases diagnosed in women attending the Burjasot screening center in Valencia, where full-field digital images had been in use for more than 4 years (Senographe 2000D Full Field Digital Mammography System). Breast cancer cases diagnosed in women screened at this center between the years 2007 and 2010 and who had attended screening in the previous round, were included in the study. For each case, a matched control was randomly chosen among women of similar age (+/- 2 years) screened the same year. Cases and controls with breast implants, surgical reduction or poor quality mammograms were excluded, so that the final sample consisted of 112 cases and 119 controls. The study was approved by the CSISP (Centro Superior de Investigacin en Salud Pblica) Ethics Committee.

2.2. MD assessment concordance

A computer-aided tool (DM-Scan) which computes MD by segmenting dense and fatty tissue in the mammogram was specifically developed for this project. The MD of each of the 655 mammograms in the DDM dataset was then assessed by three raters using this tool. In addition, a second reading of a subset of 150 randomly selected mammograms was performed two months later and both inter and intra-rater concordance was computed.

The three raters who participated in this study were experienced radiologists. Raters R1 and R2 had been reading screening mammograms from more than 10 years, with 2 years experience of full digital mammography in the former case and 6 years of indirect digital mammography in the latter. R3 had been reading mammograms for 34 years, including 2 years of indirect digital mammographs and 6 years of full digital mammograms.

The goals in this phase were 1) to provide a tool to reduce the subjectiveness of the classification process, 2) to obtain mammogram density on a continuous scale instead of in categories (as in manual classification) and 3) to obtain local (pixel-level) labeling of the mammograms. 1) and 2) were expected to improve

agreement achieved in manual classification, while 3) was to be used to train a fully-automated classifier.

In the second phase, the fully-automated classifier was implemented and the MD of each mammogram in the DDM dataset was estimated. The reproducibility of the different methods was analyzed by computing the correlation of the MDs obtained by the semi-automated and automated methods.

2.3. Breast cancer discrimination

A case-control study was conducted to analyze the ability of the proposed techniques to predict breast cancer. MD was assessed both by a rater using the semi-automated tool and by the automated system. The area under the receiver operating characteristic curve (AUC) and Odds Ratios were computed.

2.4. Semi-automated classification

Classification based on visual inspection is susceptible to human error and subjectivity. Even for experienced radiologists, it is hard to assess mammographic density objectively and accurately. Some studies show that intraobserver agreement is about 80%, whilst inter-observer agreement is below 70% when using this technique [32, 33].

As computer-assisted breast density measurement can help to reduce subjectiveness, we developed a computer-aided diagnostic tool for MD quantification (DM-Scan). This tool is freely available for noncommercial use at http://dmscan.iti.upv.es. As opposed to visual inspection methods, in which MD is usually specified as a category, it provides quantitive MD measurement on a continuous scale. Although qualitative methods have been used in other studies to measure mammographic density, these techniques involve greater subjectivity and therefore lower reproducibility. Quantitative methods can minimize such drawbacks. Figure 2 shows an example of a digital mammogram viewed in the DM-Scan screen.

The basic idea of the proposed tool is to identify pixels belonging to background, fat tissue (FT) and fibroglandular or dense tissue (FGT) by establishing

two thresholds: T1 and T2. MD is then measured as the amount of FGT in relation to breast size, i.e., MD=FGT/(FGT + FT)100.

Firstly, a pre-process is applied to condition the image before tissue segmentation is performed. Three main operations are carried out in this phase: a) contrast and brightness normalization, b) brightness correction according to breast thickness and c) segmentation of the breast and removal of regions of no-interest.

a) **Contrast and brightness normalization:** Contrast can vary significantly from one image to another due to differences in the acquisition process (mainly acquisition device and radiation dose). To ensure that the brightness values depend as much as possible on tissue density and not on other factors related to the acquisition process, contrast and brightness normalization is desirable. Assuming that minimum and maximum tissue densities are always present in a mammography (subcutaneous fat and connective tissue respectively), minimum and maximum gray-level values should also appear in the histogram. Based on this idea, a histogram stretching operation can be set to normalize brightness and contrast. Options to manually modify brightness and contrast are also available.

b) **Brightness correction:** X-ray attenuation depends not only on the density of the irradiated tissue, but also on its thickness. The thicker the tissue irradiated, the greater the attenuation and, consequently, the brighter the image. When the mammogram is taken, the breast is compressed between two parallel flat plates, b) which causes the breast to have a uniform thickness between the plates. However, towards the edge of the breast, the thickness gradually decreases. This is a drawback when the goal is to segment dense tissue, since thicker regions may look like dense tissue and vice versa. In order to avoid this problem, a brightness correction coefficient $k_{i,j}$ was applied to each pixel $p_{i,j}$ according to a user-defined parameter $\alpha \in [0:1]$ as specified below:

$$k_{i,j} = \alpha + (1 - \alpha)d_{i,j}$$

where $d_{i,j}$ is the horizontal distance from $p_{i,j}$ to either the internal border of the image or the pectoral muscle if present divided by the total distance between this border and the breast edge at row i, i.e., $d_{i,j} = 0$ when $p_{i,j}$ coincides with the border of the image, and $d_{i,j} = 1$ when $p_{i,j}$ coincides with the edge of the breast. A value of $\alpha = 1$ leaves the image unchanged, while values of $0 \leq \alpha < 1$ attenuate brightness as we approach the internal part of the mammogram. The lower α is the greater the attenuation.

c) **Breast segmentation and removal of unwanted regions:** Mammograms usually contain other extraneous objects, such as labels and/or the pectoral muscle. Breast segmentation is semi-automatically performed by finding a threshold value T1 that discriminates between background and object pixels. The biggest object found is considered to be the breast, while the remainder are considered regions of no interest and, therefore, removed. Nevertheless, this process cannot discard objects connected to the breast. To fix this problem, the user can modify the proposed T1 threshold and also manually invalidate other regions/objects not detected in the previous process.

Once the image has been preprocessed and the breast has been segmented by means of T1, a second threshold T2 must be manually set to separate dense and fat tissue, which means the dense and non-dense or fatty tissue can be measured. Finally, MD=FGT/(FGT+FT)100 is computed. The segmentation obtained in this process is also used as a method of supervised pixel labeling, which is used to train a fully-automated classifier, as explained in the next section.

2.5. Fully-automated classification

The approach used for automated classification is based on a classical supervised machine-learning scheme. This approach needs a set of labeled samples to train a model, which will be used later in the classification stage. When manual (visual inspection) classification is performed, a global class label (semi-quantitative classification usually including between four and six categories) is assigned to each mammogram. In this case, ground truth at pixel level (local level) is not available, which prevents the implementation of a supervised

method based on local features to train the classifier. However, when addressing MD assessment tasks, methods based only on global features tend to fail due to high intra-class variability. Our approach takes advantage of the FGT segmentation performed with DM-Scan to label each breast pixel into one of the two possible classes (FGT/fatty). This makes it possible to train a model that can discriminate between both types of tissue. Figure 3 shows a diagram of the proposed scheme.

In the training phase, local features are extracted from each image in the training set. For this purpose, a local window of $N \times N$ pixels is shifted along the breast region and the gray level values are extracted, producing a local feature vector for each local window. Then, the extracted local feature vectors are projected into a lower dimensional space by using PCA analysis, and labeled with the class pertaning to the central pixel of the local window. The optimal N value for images measuring 1024×1024 was empirically found to be 31.

Once the system has been trained, the proposed classification method consists of three stages. Firstly, the breast is automatically segmented from the background of the mammogram and contrast normalization and brightness correction is performed. Secondly, each pixel of the mammogram (or a subsampling if the resolution is too high) is classified using the model learned in the previous training stage. This yields a hypothesis map (FGT/fatty) for each pixel, which could be used to compute MD. However, a third final step based on a thresholding operation is performed to improve the classification.

Breast segmentation is automated by combining automatic thresholding based on the histogram of the image, followed by a connected-component analysis algorithm that separates the breast structure from background noise. In this work we used the cranio-caudal (CC) view of the mammogram, in which the pectoral muscle is either not present or is negligible. This means the pectoral muscle is not automatically segmented. When using the mediolateral oblique (MLO) view, manual or automatic segmentation of the pectoral muscle should be carried out.

Thresholding is performed to remove the dark background by looking at the

position of the highest peak in the histogram. Connected-component labeling is then applied for blob detection. Finally, the breast region is selected as the blob with the biggest area.

Contrast normalization and brightness correction is done in the same way as in the semi-automated approach, described in Section 2.4. In this case, the α parameter is set to a fixed, empirically calculated value.

In the classification stage, the approach used to extract local feature vectors from training images is applied to the test image. Each of the feature vectors extracted are then classified by the k-NN algorithm, which produces a hypothesis label for each pixel, using the fast approximate nearest neighbor search, based on kdtrees [34]. The result of the classification stage is a binary map representing tissue types. MD could be computed from this map, but a thresholding operation is finally performed to improve the results.

Different thresholding operations are performed on the original image at all the existing gray level values t. False positives (FP_t) and false negatives (FN_t) are computed, in which FP_t are pixels classified as FGT with brightness lower or equal to t, and FN_t are pixels classified as fatty with brightness higher than t. Finally, the optimal threshold $\hat{t}$ is calculated as

$$\hat{t} = \underset{t}{\operatorname{argmin}}(FP_t + FN_t) \tag{1}$$

and every point in the image is relabeled as FGT if its gray value is higher than $\hat{t}$ orotherwise as *fatty*, leading to the final MD value as the ratio of FGT over total breast pixels.

Instead of using the hard classification scheme which labels each pixel in one of two possible classes ($FGT/fatty$), a soft or probabilistic classification can be performed. In this case, the k-NN classifier assigns to each pixel $x_{i,j}$ a probability $P(x_{i,j})$ of belonging to the FGT class (and $P(\text{fatty}|x_{i,j})$ is computed as $1 - P(\text{FGT}|x_{i,j})$). The classification stage thus produces a probability map that represents the likelihood of each pixel belonging to the FGT class. The optimal threshold $\hat{t}$ can then be computed using a probabilistic estimation of

the concepts of FP_t and FN_t as follows:

$$\hat{t} = \underset{t}{\operatorname{argmin}}(\sum_{i,j}(P(\text{fatty}|x_{i,j})\beta_{i,j,t} + P(\text{TFG}|x_{i,j})(1-\beta_{i,j,t})) \tag{2}$$

where $\beta_{i,j,t} = 1$ if the brightness of $x_{i,j}$ is greater than t and 0 otherwise. In short, the thresholding operation minimizes the sum of $P(\text{FGT}|x_{i,j})$ for those pixels with brightness lower than t plus the sum of $P(\text{fatty}|x_{i,j})$ for pixels with brightness greater than t.

3. Experiments and results

Two sets of experiments were carried out to test the performance of the proposed system. The first was aimed at testing the agreement level in the MD calculation when using the semi-automated and fully-automated systems, using the DDM dataset described in Section 2.1. In the second set, the relation between MD and the risk of developing breast cancer was analyzed by the Case-Control dataset (see Section 2.1).

3.1. Semi-automated and fully-automated correlation

In this task, a real ground truth (gold standard) as defined in other classification problems does not exist, since different experts may differ in their assessment about what is considered dense and fat tissue.

In these cases, the performance is not derived from the error rate, but from the concordance correlation among raters, which determines the degree of reproducibility of the measurement technique. With this purpose, intra and inter raters correlation using computer-aided (semi-automated) assessment, as well as correlation between fully-automated and semi-automated assessment has been computed using the Intraclass Correlation Coefficient (ICC) . The *oneway* approximation, as described in [35], has been employed.

In these cases, the performance is not derived from the error rate, but from the concordance correlation among raters, which determines the degree of reproducibility of the measurement technique. With this purpose, the intra- and

inter-rater correlation was calculated using computer-aided (semi-automated) assessment, as well as the correlation between fully-automated and semi-automated assessment, using the Intraclass Correlation Coefficient (ICC) . The *oneway* approximation, as described in [35], was employed. Preliminary experiments were carried out to test the behavior of Equations 1 and 2 in the proposed classification scheme. The best results were obtained with Equation 1 (hard classification), and the results shown below correspond to this approach.

Tables 1 and 2 show, respectively, the inter and intra-rater ICC obtained when using DM-Scan in semi-automated mode. The extent of dispersion between raters is shown in Figure 4 by means of Bland-Altman plots. Horizontal lines are plotted indicating the limits of agreement (mean and mean $\pm$ 1.96 SD).

These correlations are very good and significantly outperform the concordance obtained in visual inspection reported in previous works [31, 36, 37].

Once MD assessment by means of DM-Scan had been performed and analyzed, a fully-automated classifier was implemented as described in Section 2.5. The FGT segmentation of rater R2 was used as ground truth to train the system. The optimal values of the different parameters were empirically found to be: *local window size*=31, *number of principal components*=8, α=0.7 and K=11. A 10-fold cross-validation approach was employed to assess the MD of the 655 mammograms.

Table 3 shows the ICC comparing the fully-automated and the semi-automated (DM-Scan) methods for each rater. Figure 5 shows the Bland-Altman plot between R2 and the automated classification.

As can be seen in Tables 1 and 2, radiologists are reported to disagree on classifications, even if intra-observer agreement is analyzed (average ICC=0.92), so that a higher agreement between a human and an automated classifier cannot be expected. The reported ICC=0.838 between R2 (used to train the system) and the automatic classifier can therefore be considered a good result.

3.2. Breast cancer prediction

In the second set of experiments, MD was used as a test variable to determine the risk of developing breast cancer. Both the relation of semi-automated and fully-automated MD assessment with breast cancer risk were analyzed. The CaseControl dataset was used to test both approaches, whilst R2s DDM dataset was employed in these experiments to train the automated system. Firstly, the MD mean values for Cases and Controls were computed in both methods. Then, the area under the Receiver Operating Characteristic curve (AUC) and the Odds Ratios (OR), adjusted by age, were used to compare the semi-automated and the fully-automated approaches. OR and its 95% confidence interval were estimated using unconditional logistic models.

The mean MD obtained for Controls and Cases is shown in Table 4. As expected, MD is higher in Cases, which corroborates the existence of a relation between MD and breast cancer. However, a slight underestimation and a lower MD range is detected in the automated mode. This is due to some low contrast images present in the Case-Control dataset, which means MD is generally underrated in the automated system.

Table 5 shows the AUC obtained for semi-automated and fully-automated MD quantification. Both methods show a subtle but significant correlation with breast cancer risk, which suggests that MD can help to predict this disease. The results are very similar for both approaches, which demonstrates that the proposed automated method could be used instead of the traditional manual or semi-automated methods to estimate MD and predict breast cancer.

Table 6 shows the OR obtained at different cutpoints, with the first quartile used as reference. It can be observed that cancer risk increases with MD (average OR increase of 1.38 and 1.50 per 10% increase in MD for the semiautomated and fully-automated methods, respectively). In general, as in the AUC study, the behavior of both methods is very similar, although the automated approach gives wider confidence intervals in the last quartile.

Although the automated method presents a slightly lower predictive value than the semi-automated one at the highest cut point, the OR per 10% increase

is higher in the former, which suggests that the automated method could also be used to estimate breast cancer risk and consequently could be used in clinical decision making.

4. Discussion

A semi-automated and a fully-automated method of measuring breast density are proposed here. In the first case, MD is based on the selection of an experienced radiologists threshold, which segments dense from fatty tissue. This allows MD measurement to be made on a continuous scale, in contrast with traditional visual-based methods, which divide MD into categories. Moreover, semi-automated segmentation of dense tissue is employed to label the mammograms used later to train the automatic system. A computer tool (DM-Scan) was developed for this purpose and is now freely available for non-commercial use at http://dmscan.iti.upv.es.

Using a dataset of 655 mammograms, we found a high inter-rater correlation in semi-automated MD assessment (average ICC=0.922). Furthermore, the correlation between rater R2, using the semi-automated system, and the automated system (trained by R2) was ICC=0.838. Although this correlation is slightly lower than intra-R2 correlation (ICC=0.938), it is still high, which suggests that the proposed automated method could be a valid option for measuring MD, particularly when a large number of mammograms must be processed.

In a second set of experiments carried out on 230 mammograms (a case-control dataset), we found that both the automated and the semi-automated estimate of MD were associated to a similar extent with breast cancer (OR per 10% increase in MD: 1.38 and 1.50 for semi-automated and automated modes, respectively). These results confirm that risk assessment protocols can take advantage of the proposed MD measurement methods to improve the estimation of breast cancer risk. More accurate risk prediction would help in better clinical decision making as regards screening frequency and screening test selection.

Although the automated method achieves a slightly better performance in

average cases, there is a notable drop in OR in patients with high breast density (greater than 30%), due to some misclassifications in this group. Some of these errors could be due to the automated method being trained with mammograms from the DDM set, while the test images used in this experiment were from the Case-Control set. The images in the latter set were acquired by a Senographe 2000D (which was not used to acquire the DDM set) and its mammograms present bright levels very different from those in the DDM set. Although brightness and contrast normalization algorithms were applied, some images had not been correctly preprocessed, which adversely affected the classification stage.

The behavior of both approaches is in general comparable. The semi-automated method is suitable for analyzing small numbers of mammograms. However, when processing large numbers, as in screening programs, the advantages of using an automated system are obvious. Radiologists soon tire when analyzing large numbers of images and the accuracy and objectivity of the measurement process can be affected, besides which using human processors is more expensive. The evidence we found that automated MD is associated with breast cancer risk could warrant the use of this tool in a clinical setting for risk prediction and clinical decision making.

However, a number of issues should be addressed in future work in order to improve the accuracy of automated density measurement and cancer risk prediction, including brightness and contrast normalization issues by more accurate methods of finding histogram limits.

It has also been mentioned that the agreement between a human expert (rater) and the automated system is slightly lower than that among individual raters. This could be due to some manual processes (brightness correction and removal of pectoral muscle) not being addressed in the automated approach.

Brightness correction, (parameter α explained in Section 2.4) was fixed at a constant value of 0.9 in the automated method. This value is adequate in most cases, but could introduce a significant error in MD measurement. During the acquisition process, breast compression and the distance between the plates

containing the breast can be stored as metadata in the image file and could be used to estimate breast thickness [38] and, consequently, an adaptive value of parameter α.

It should also be noted that a hard classification scheme was used, assuming that each pixel can only belong to one of the two possible classes. In preliminary experiments, better concordance was found between semi-automated and fully automated assessment when using hard classification. However the relation between MD and breast cancer risk was not tested with a soft or probabilistic classification scheme, in which each pixel has an associated probability of belonging to each class. Future experiments should take these probabilities into account to test whether a probabilistic approach could improve breast cancer risk prediction.

The influence of the pectoral muscle was considered to be insignificant in the cranio-caudal (CC) view used in this study. However, here again more accurate measurements could be obtained if it were to be supressed using algorithms for automatic pectoral muscle removall [39, 40]. If mediolateral oblique (MLO) views are used instead of CC, then automatic pectoral muscle supression is mandatory. Also, improvements in contrast normalization should be addressed to avoid underestimating MD in low contrast images. Finally, other parameters could be explored that take into account not only the relative density but also the shape and the distribution of FGT in the breast [20, 22] or texture features from the mammogram [21].

In conclusion, our work shows that the semi-automated and fully-automated methods presented here showed a reasonable agreement and had substantial discriminative power to predict subsequent breast cancer development. The fully automated method will facilitate the incorporation of mammographic density assessment in clinical and screening practices. It can also be expected to facilitate the study of the evolution of breast density with time, as recent studies suggest that this evolution is even more informative as regards the risk of breast cancer [41, 42].

Acknowledgments

This work was supported by research grants from Gent per Gent Fund (EDEMAC Project); Spains Health Research Fund (Fondo de Investigacin Santiaria) (PI060386 & FIS PS09/00790); Spanish MICINN grants TIN2009-14205-C04-02 and Consolider Ingenio 2010: MIPRCV (CSD2007-00018); Spanish Federation of Breast Cancer Patients (Federacin Espaola de Cncer de Mama) (FECMA 485 EPY 117010). The English revision of this paper was funded by the Universitat Politècnica de València, Spain

References

[1] Wolfe JN. Breast pattern as an index of risk for developing breast cancer. AJR Am J Roentgenol. 1976 Jun;126(6):1130-7.

[2] McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2006 Jun;15(6):1159-69.

[3] Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, et al. Mammographic breast density as an intermediate phenotype for breast cancer. Lancet Oncol. 2005 Oct;6(10):798-808.

[4] Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast cancer risk: current understanding and future prospects. Breast Cancer Res. 2011;13(6):223.

[5] Buist DS, Porter PL, Lehman C, Taplin SH, White E. Factors contributing to mammography failure in women aged 40-49 years. J Natl Cancer Inst. 2004 Oct 6;96(19):1432-40.

[6] Domingo L, Sala M, Servitja S, Corominas JM, Ferrer F, Martínez J, et el. Phenotypic characterization and risk factors for interval breast cancers in a population-based breast cancer screening program in Barcelona, Spain. Cancer Causes Control. 2010 Aug;21(8):1155-64.

[7] te Brake GM, Karssemeijer N, Hendriks JH. Automated detection of breast carcinomas that were not detected in a screening program. Radiology. 1998 May;207(2):465-71.

[8] Wolfe JN. Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer. 1976 May;37(5):2486-92.

[9] Gram IT, Funkhouser E, Tabar L. The Tabar classification of mammographic parenchymal patterns. Eur J Radiol. 1997 Feb;24(2):131-6.

[10] Boyd NF, Byng JW, Jong RA, Fishell EK, Little LE, Miller AB, et al. Quantitative classification of mammographic densities and breast cancer risk: Results from the Canadian national breast screening study. J. Nat. Cancer Inst. 1995 May; 87:670-75

[11] American College of Radiology (ACR): Illustrated Breast Imaging Reporting and Data System (BI-RADS). 3rd edn. Reston, VA: American College of Radiology. 1998; 167-81

[12] Jamal N, Ng KH, Looi LM, McLean D, Zulfiqar A, Tan SP, et al. Quantitative assessment of breast density from digitized mammograms into Tabar's patterns. Phys Med Biol. 2006 Nov 21;51(22):5843-57.

[13] Karssemeijer N. Automated classification of parenchymal patterns in mammograms. Phys. Med. Biol. 1998; 43:365-78

[14] Saha PK, Udupa JK, Conant EF, Chakraborty DP, Sullivan D. Breast tissue density quantification via digitized mammograms. IEEE Trans Med Imaging. 2001 Aug;20(8):792-803.

[15] Klifa C, Carballido-Gamio J, Wilmes L, Laprie A, Lobo C, Demicco E, et al. Quantification of breast tissue index from MR data using fuzzy clustering. Conf Proc IEEE Eng Med Biol Soc. 2004;3:1667-70.

[16] Oliver A, Freixenet J, Bosch A, Raba D, Zwiggelaar R. Automatic classification of breast tissue. Lecture Notes in Computer Science. 2005; 3523:431-38

[17] Oliver A, Freixenet J, Mart R, Pont J, Prez E, Denton ER, Zwiggelaar R. A novel breast tissue density classification methodology. IEEE Trans Inf Technol Biomed. 2008 Jan;12(1):55-65

[18] Muhimmah I, Zwiggelaar R. Mammographic density classification using multiresolution histogram information. Proceedings of the International Special Topic Conference on Information Technology in Biomedicine. 2006 Oct.

[19] Heine JJ, Carston MJ, Scott CG, Brandt KR, Wu FF, Pankratz VS, et al. An automated approach for estimation of breast density. Cancer Epidemiol Biomarkers Prev. 2008 Nov;17(11):3090-7.

[20] Heine JJ, Scott CG, Sellers TA, et al. A novel automated mammographic density measure and breast cancer risk J Natl Cancer Inst. 2012 Jul 3;104(13):1028-37.

[21] Manduca A, Carston MJ, Heine JJ, Scott CG, Pankratz VS, Brandt KR, Sellers TA, Vachon CM, Cerhan JR. Texture features from mammographic images and risk of breast cancer. Cancer Epidemiol Biomarkers Prev. 2009 Mar;18(3):837-45

[22] Li J, Szekely L, Eriksson L, Heddson B, Sundbom A, Czene K, et al. High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. Breast Cancer Res. 2012 Jul 30;14(4):R114

[23] Gweon HM, Youk JH, Kim JA, Son EJ. Radiologist Assessment of Breast Density by BI-RADS Categories Versus Fully Automated Volumetric Assessment. AJR Am J Roentgenol. 2013 Sep;201(3):692-7.

[24] Tagliafico A, Tagliafico G, Astengo D, Cavagnetto F, Rosasco R, Rescinito G, et al. Mammographic density estimation: one-to-one comparison of digital mammography and digital breast tomosynthesis using fully automated software. Eur Radiol 2012; 22:12651270

[25] Jeffreys M, Warren R, Highnam R, Smith GD. Initial experiences of using an automated volumetric measure of breast density: the standard mammogram form. Br J Radiol. 2006 May;79(941):378-82

[26] Lokate M, Kallenberg MG, Karssemeijer N, Van den Bosch MA, Peeters PH, Van Gils CH. Volumetric breast density from full-field digital mammograms and its association with breast cancer risk factors: a comparison with a threshold method. Cancer Epidemiol Biomarkers Prev. 2010 Dec;19(12):3096-105

[27] Polln M, Ascunce N, Ederra M, Murillo A, Erdozin N, Als-Martnez JE, et al. Mammographic density and risk of breast cancer according to tumor characteristics and mode of detection: a Spanish population-based case-control study. Breast Cancer Res. 2013 Jan 29;15(1):R9.

[28] Pollán M, Llobet R, Miranda-García J, Antón J, Casals M, Martínez I, et al. Validation of DM-Scan, a computer-assisted tool to assess mammographic density in full-field digital mammograms. SpringerPlus. 2013; 2:242.

[29] Byng JW, Yaffe MJ, Jong RA, Shumak RS, Lockwood GA, Tritchler DL, Boyd NF Analysis of mammographic density and breast cancer risk from digitized mammograms. Radiographics. 1998; 18:15871598

[30] Garrido-Estepa M, Ruiz-Perales F, Miranda J, Ascunce N, González-Román I, Sánchez-Contador C, et al. Evaluation of mammographic density patterns: reproducibility and concordance among scales. BMC Cancer. 2010 Sep; 10:485.

[31] Pérez-Gómez B, Ruiz F, Martínez I, Casals M, Miranda J, Sánchez-Contador C, et al. Women's features and inter-/intra-rater agreement on mammographic density assessment in full-field digital mammograms (DDM-SPAIN). Breast Cancer Res Treat. 2012 Feb;132(1):287-95.

[32] Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F,

et al. Categorizing breast mammographic density: intra and interobserver reproducibility of birads density categories. Breast. 2005 Aug;14(4):269-75.

[33] Zhou C, Chan HP, Petrick N, Helvie MA, Goodsitt MM, Sahiner B, et al. Computerized image analysis: estimation of breast density in mammograms. Med. Phys. 2001; 28, 1056-69

[34] Arya S., Mount D.M., Netanyahu N.S., et al. An optimal algorithm for approximate nearest neighbor searching. Journal of the ACM. 1998 Nov;45(6):891-923.

[35] Shrout PE, Fleiss JL. Intraclass correlation: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420-8.

[36] Lobbes MB, Cleutjens JP, Lima Passos V, Frotscher C, Lahaye MJ, Keymeulen KB, et al. Density is in the eye of the beholder: visual versus semi-automated assessment of breast density on standard mammograms. Insights Imaging. 2012 Feb;3(1):91-9.

[37] Martin KE, Helvie MA, Zhou C, Roubidoux MA, Bailey JE, Paramagul C, et al. Mammographic density measured with quantitative computer-aided method: comparison with radiologists' estimates and BI-RADS categories. Radiology. 2006 Sep;240(3):656-65.

[38] Highnam RP, Brady JM, Shepstone BJ. Estimation of compressed breast thickness during mammography. Br J Radiol. 1998 Jun;71(846):646-53.

[39] Raba D, Oliver A, Mart J, Peracaula M, Espunya J. Breast Segmentation with Pectoral Muscle Suppression on Digital Mammograms. In Proc. of IbPRIA. 2005;471-78.

[40] Mustra M, Grgic M. Robust automatic breast and pectoral muscle segmentation from scanned mammograms. Signal Processing. In Press.

[41] van Gils CH, Hendriks JH, Holland R, Karssemeijer N, Otten JD, Straatman H, et al. Changes in mammographic breast density and concomitant changes in breast cancer risk. Eur J Cancer Prev. 1999 Dec;8(6):509-15.

[42] Kerlikowske K, Ichikawa L, Miglioretti DL, Buist DS, Vacek PM, Smith-Bindman R, et al. Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer risk. J Natl Cancer Inst. 2007 Mar 7;99(5):386-95.

Authors' contributions

JA & RLL developed DM-Scan. MP, RLL, JMG, BPG, and DST conceived the concordance study and participated in its design. JMG, RLL & JA gathered and organized the set of mammograms from the concordance study. MC, IM, & FRP were responsible for MD assessment by DM-Scan. JA, RLL & JCPC conceived and developed the algorithms for automated classification. JA & RLL carried out the experiments on automated classification. JMG & DST performed the matching process and gathered the information needed for the case-control study. RLL and MP performed the statistical analysis and the results were circulated and commented on by all the authors. RLL drafted the first version of the manuscript and it was critically reviewed by the rest of the authors. JCPC corrected the final version. All the authors have read and approved the final manuscript.

Table 1: Inter-rater ICC with their 95 % confidence intervals for semi-automated estimation.

Raters	ICC [CI 95 %]
R1 R2	0.922 [0.910 0.933]
R1 R3	0.928 [0.916 0.938]
R2 R3	0.916 [0.902 0.927]
Mean	0.922 [0.909 0.933]

Table 2: Intra-rater ICC with their 95 % confidence intervals for semi-automated estimation.

Rater	ICC [CI 95 %]
R1	0.935 [0.911 0.952]
R2	0.938 [0.915 0.955]
R3	0.900 [0.863 0.926]
Mean	0.924 [0.896 0.944]

Table 3: ICC with their 95 % confidence intervals for fully-automated estimation.

Rater	ICC [CI 95 %]
R1	0.800 [0.771 0.826]
R2	0.838 [0.814 0.860]
R3	0.785 [0.754 0.813]
Mean	0.794 [0.764 0.821]

Table 4: Mean MD for Controls and Cases in semi-automated and fully-automated methods

	Mean Controls	Mean Cases	P-value
Semi-automated	16.8 [14.4-19.2]	20.9 [18.1-23.6]	0.026
Automated	16.6 [15.05-18.1]	18.6 [17.1-20.1]	0.061

Table 5: AUC for the different markers analyzed

Marker	AUC [CI 95 %]
MD_{semi}	0.613 [0.539-0.687]
MD_{auto}	0.602 [0.530-0.672]

Table 6: Odds ratios obtained at different cutpoints

Cutpoint	Control/Case	OR	95% CI	P-value
Semi-automated				
< 10	37/21	1.00		
10-19.9	31/26	1.89	0.94-3.82	0.074
20-29.9	26/32	2.56	1.15-5.69	0.021
≥ 30	25/32	2.87	1.29-6.39	0.010
Per 10% increase		1.38	1.11-1.71	0.003
Automated				
< 10	36/22	1.00		
10-19.9	32/25	2.56	1.19-5.49	0.016
20-29.9	25/33	3.50	1.45-8.45	0.005
≥ 30	26/31	2.55	0.81-8.03	0.109
Per 10% increase		1.50	1.07-2.10	0.019

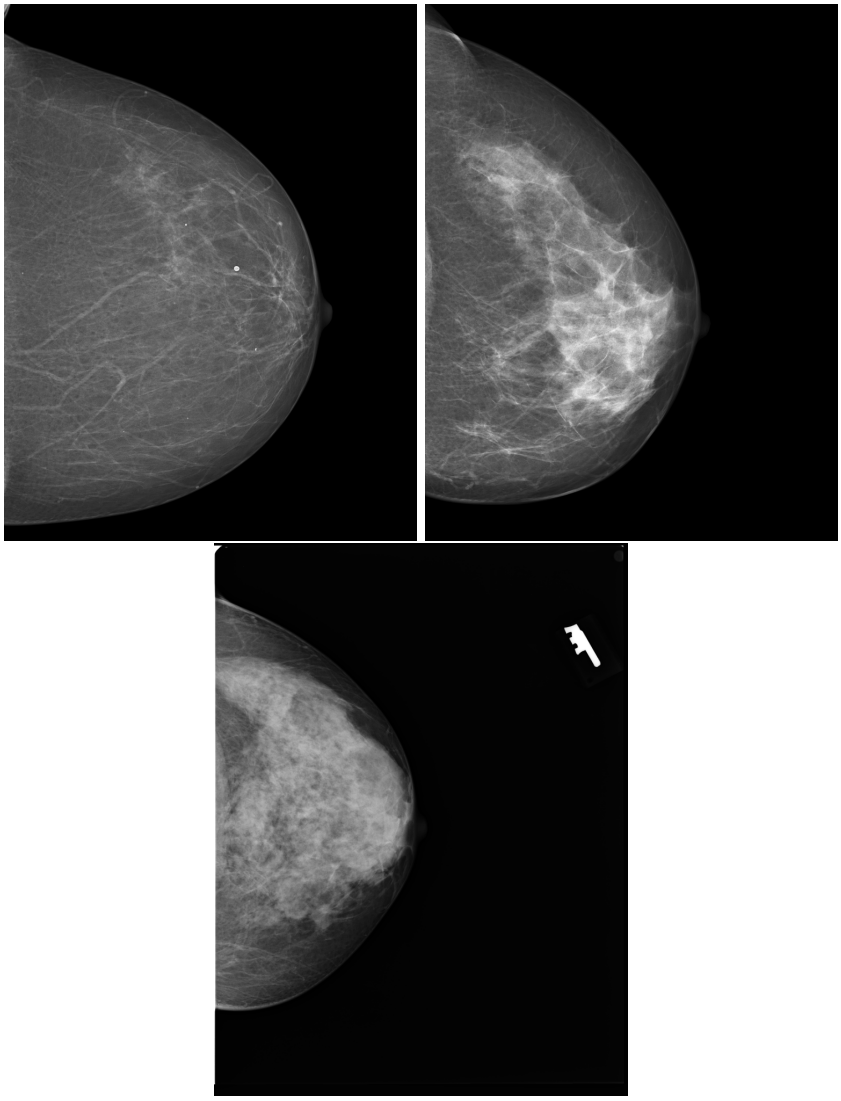

Figure 1: Breast composed of fatty tissue (top), dense and fatty tissue (middle) and mainly dense tissue (bottom)

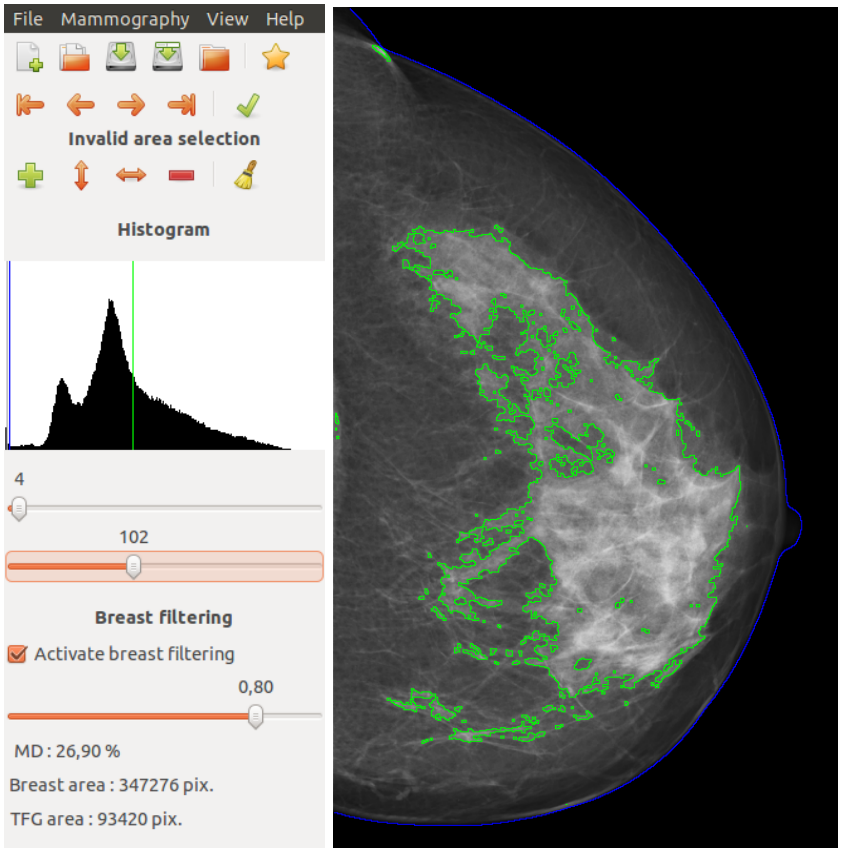

Figure 2: Example of FGT segmentation in DM-Scan

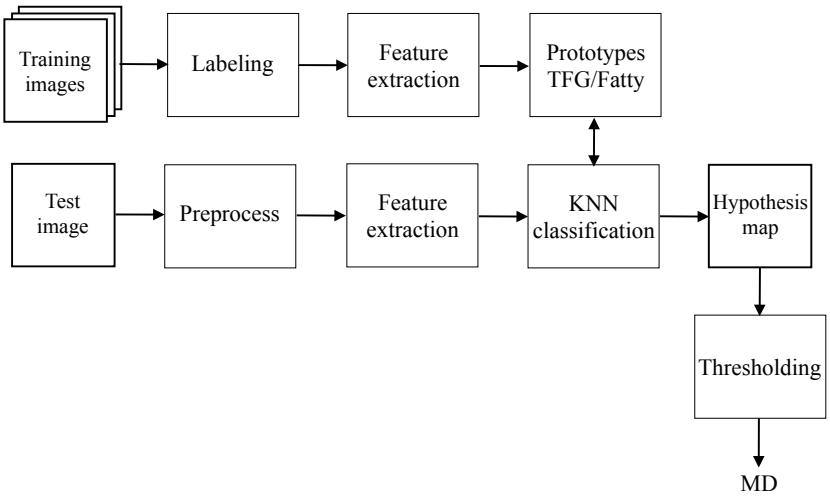

Figure 3: Scheme of the automated process.

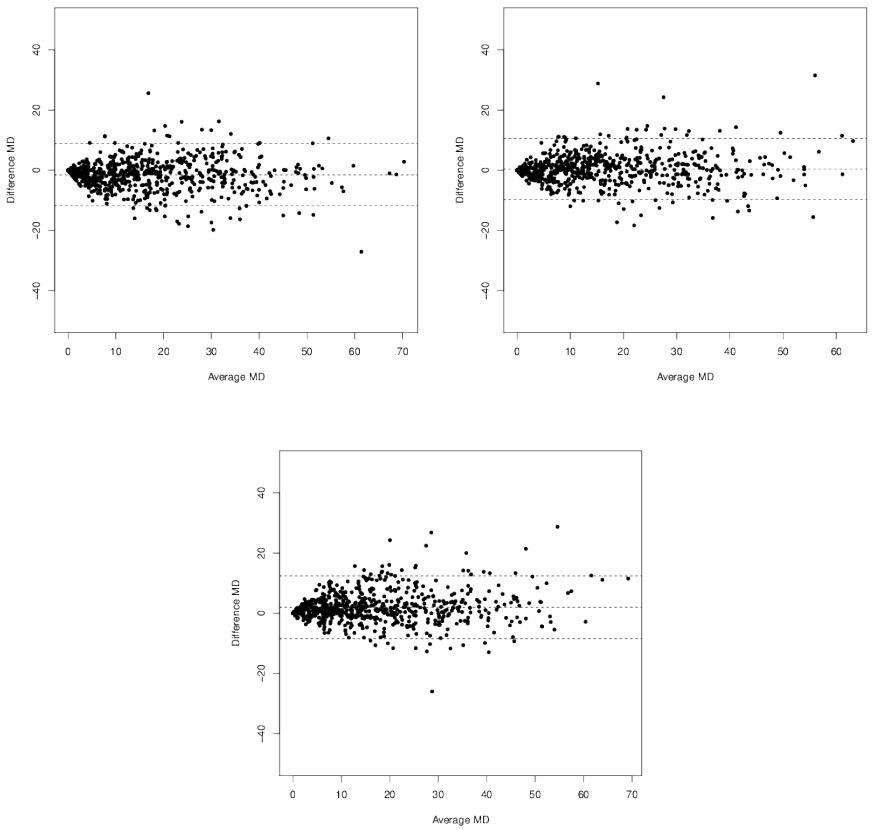

Figure 4: Bland-Altman plot for raters R1-R2 (top), R1-R3 (medium) and R2-R3 (bottom) using DM-Scan

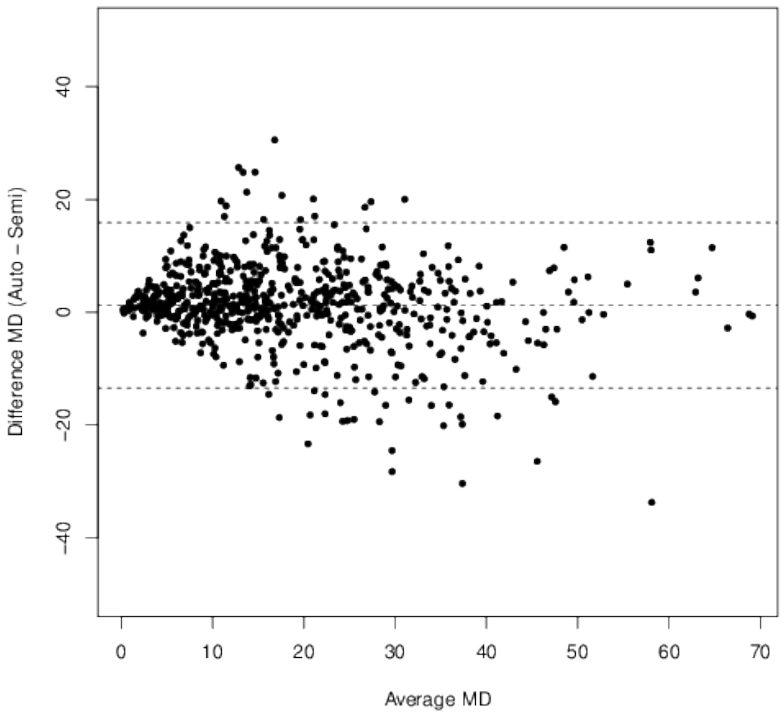

Figure 5: Bland-Altman plot for automated classification and semi-automated classification (rater R1) with Case-Control dataset