Supplementary Material

Pre-processing phase

Prior to processing each mammogram we initially checked one of its tag fields (*APs.PhotometricInterpretation*) to determine whether the image's greyscale needs to be inverted:

```
if isfield(APs, 'PixelIntensityRelationshipSign')
      if (APs.PixelIntensityRelationshipSign==1)
            Invert color;
      end
end
```

The first proper step of our algorithm is to identify the "breast-to-air" region. Image segmentation has to fulfil certain predefined criteria and is one of the most difficult tasks in image processing[29]. Similar procedures that have been used for removing labels, markers and patient's particulars in digitised film mammogram images are used here for reducing the size of FFDM images, by removing unnecessary black background. This step is carried out by identifying the vertical critical column, *j'*, in the mammogram image $\{I \in \Re^2\}$ of size $m \times r$. We use a vertical projection technique, of the intensity values, which is computationally undemanding yet very efficient in both FFDM and digitised film mammogram images. A vertical projection can be seen as a plot in which each index, in b_j, corresponds to the sum of the pixels along column *j*.

$$b_j = \int_{-\infty}^{\infty} I(i, j)dj \quad \text{(S1)}$$

This integral can be approximated by summation ($b_j = \sum_{i=1}^{m} I(i, j)$). Then, locating the global minimum in the vector *b* yields the vertical critical column: $j' = \left\{ j' / b_{j'} = \min(b) \right\}, j \in [1, 2, ..., n]$. Similarly, the projection method is adopted to locate the horizontal bottom cut in the image (i.e., where the breast ends) as shown in Figure S1 (compare (a) to (b)). The image is then down-sampled to 512 pixels width using bilinear interpolation while preserving the image aspect ratio for the height. Next, the signal-to-noise ratio is improved by means of applying a global spatial contrast adjustment to the mammograms. Limits to

contrast stretch are found by considering the non-black region after performing image intensity normalization.

Detection of breast area

To choose a threshold algorithm for detecting the breast boundary (object blob) we first examined the following 13 different threshold algorithms, in a small subset of images: *concavity method; triangle algorithm; entropy method; inter-means method; iterative inter-means; maximum likelihood (EM); mean method; median method; minimum error; iterative minimum error; moments method; p-tile method and Otsu's method*. We choose to use the *triangle algorithm* due to its invariance to intra-breast intensity variations, making it our choice for detecting breast boundaries. Our algorithm then estimates the area of the objects in the resulting binary image and retains the largest object. Whilst the latter step, removal of non-breast objects, exhibits a mild effect in FFDM images, it is of crucial importance in digitised film mammogram images where there are labels, calibration wedges and various other paraphernalia on a mammogram. The position and orientation of these labels are not standard. Belkhodja et al.[30] use the Otsu method and a Gaussian filter in their algorithm to mask out labels in digitised film images.

Segmentation and pectoral muscle detection

The next step is to invoke an image segmentation routine. Image segmentation can be defined as the partitioning of a given image into non-overlapping, constituent regions that are homogeneous with respect to some characteristic such as intensity or texture [29] Ch.10. A survey of the medical image segmentation literature is provided by [31]. In this step we use Delaunay-based image segmentation,[32] , to estimate the number of pixels with homogenous intensity. This segmentation is fully automated, fast and does not require the user to provide an initial estimate of the number of clusters. The technique resembles the dynamic thresholding method used for segmentation, but it differs in terms of divide and merge-decision making. A Delaunay triangulation (*DT*), described in detail in [33], is constructed in our case from a set of points which correspond to the probability density function (PDF) of the image intensities. The outer

boundary of a *DT* is simply the convex hull (CV) of the set of the feature points where vertices provide a direct access to intensity values to be used in image segmentation. Time complexity ($O(n\log n), n_{histogram\_bins} \le 2^8$) is considerably lower than other approaches such as region growing and the deformable contours/snakes algorithms.

Let *I* be the entire mammogram image. Image segmentation can be seen as a process that partitions *I* into *n* sub-regions ($I_1, I_2, \ldots, I_n$) based on a clear measure such as homogeneity in such a way that regions in the segmented image (*S*) comply to:

i) $\bigcup_{i=1}^{n} I_i = I$

ii) $I_i \cap I_j = \emptyset, \forall i \neq j.$

iii) I_i is a connected set, $i=1, 2 \ldots n$.

After this segmentation step, a pseudo-colour pattern (Red (R), Green (G), Blue (B)) is created using the composite of the original mammogram and its segmented version (*S*). We choose (arbitrarily) to represent the original image with the *R* component, in the RGB primary colour space. While the segmented image is treated as the *G* component in the native colour space, the blue *B* (black here) component is zeroed out to eliminate third component interference. Different perturbations can obviously be obtained. However, our choice is to have dense area (pectoral muscle and fibroglandular tissue) marked with a green gamut and the radiolucent tissue marked as red-orange; see the pseudo-colour pattern in Figure S1 (b). We examined two colour spaces namely, HSV and YIQ. It is interesting to note that *I* and *Q* components in the YIQ colour space have a very appealing property, namely that they provide an excellent separation of the red-orange colour from the green gamut. The HSV, on the other hand, is non-linear, therefore it is more computationally demanding than the YIQ transformation, especially in analyses of large volumes of medical data. The breast area segmentation step can be written in a simplified form based on the following signum function.

$$sign(Q(x,y)) = \begin{cases} -: Q(x,y) \in fatty\_region\_foreground \\ 0: \quad Q(x,y) \in image\_background \\ +: Q(x,y) \in Dense\_tissue\_foreground \end{cases} \quad \text{(S2)}$$

See Figure S1 (c, e) for an example of the application of the signum function to real mammographic data. In order to remove all or most of the pectoral muscle, a mask needs to be created. Unlike existing methods where the pectoral muscle mask is constructed under the assumption that its boundary can be approximated with a single straight line on the top left corner for LMLOs, we propose the use of the convex hull allowing for fragmented straight lines. To create the mask, the smallest convex polygon (convex hull) is constructed from the binary image, $Q(x,y)<0$ (see, Figure S1 (c-d)).

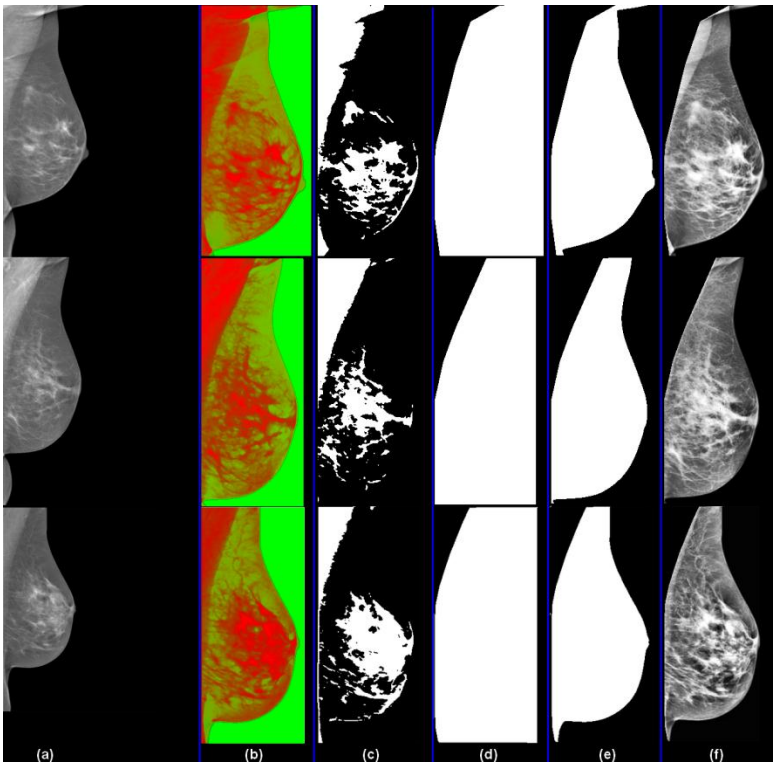

Figure.S1. Pre-processing of mammograms, (a) original mammograms, (b) pseudo-colour generation after applying the horizontal and vertical cropping, (c) the negative signal in the *Q (x,y)* colour space, detecting the non-reddish area (d) convex hull of *(c)*, (e) the final extracted breast mask, and (f) breast region after applying the contrast limited adaptive histogram equalization (CLAHE).

The accurate detection of breast and pectoral muscle boundaries is a challenging task since in some cases the latter's intensity blends together with the breast area where it becomes hard even for an observer to

identify the border of the pectoral muscle. We therefore wanted to compare our approach in this step to alternative pre-existing methods. A common approach is to use active contours (ACs), also known as Snakes or deformable contours [21] p. 299. Among those studies which used ACs in conjunction with mammography are [34,35], and the algorithm was recently evaluated in [36]. The ACs approach is based on an ingenious process that forms a set of points which aim to enclose a target shape, the shape to be extracted. The ACs approach works perfectly well in retrieving the true boundary of a fuzzy shape, thanks to its internal efficient energy minimization algorithm, making it an ideal option for semi-automatic processes. However, it fails to converge when the contour is tangent to the velocity vector and it can deform around a non-targeted object. For this reason the initialization phase can impact highly on the final result. Therefore, to enable a fair comparison with our developed algorithm, we assisted ACs to overcome these obstacles. We compared the result from our algorithm and the ACs approach in terms of the estimates of area PD to which they lead and against a third estimate (ground truth) of PD (obtained by the semi-automated approach Cumulus in a subset of 39 FFDM images. The Pearson correlation coefficients (correlation with Cumulus PD) were estimated as r=0.90 and r=0.79 for our PD proposed approach and the active contour algorithm, respectively. We measured time complexity on the larger dataset of 1011 processed FFDM images. The average time complexity was recorded as mean=0.2712 sec and mean= 8.1614 sec for our PD proposed approach and the active contour algorithm, respectively. Both procedures are fast, although our approach can lead to remarkable time gain when considering a large volume of images as will be necessary for the KARMA cohort study. For instance, to process 190,000 images (the number of images which have currently been collected within KARMA) would take in excess of two weeks using the active contour algorithm as compared to merely 15 hours using our proposed approach.

Once the image of the breast has been segmented, we enhance the local contrast to reveal features present in the breast by applying contrast limited adaptive histogram equalisation (CLAHE) with the number of tiles being equal to [8×8]. We have shown in previous studies pertaining to diabetes that CLAHE is a vital input for many imaging based medical oriented sciences; for further insights into its micro-biology

applications see[37,38]. CLAHE is described mathematically as follows. Let I be the image with the segmented breast, the CLAHE transform $\bar{h}$ is given by:

$$\bar{h}(I) = [I_{\max} - I_{\min}]\rho(f) + I_{\min} \quad \text{(S3)}$$

where $\rho(f)$ represents the CDF (cumulative distribution function) denoted by $\rho(f) = \rho(X \le x) = \int_0^x f(t)dt$.

Table S1: The derived statistical and textural features

Feature (* denotes features derived from a single region, the rest are derived from the 12 regions)	**Intensity based**	**Morphology/Shape based**
PD: Percent density (CASAM-Area)		●
f1: The DC value of the 2D discrete cosine transform	●	
f2: The coefficient of the last decomposition of single-level discrete 2D wavelet transform	●	
f3: The max coefficient of the 2D discrete Fourier transform of the local range filter	●	
f4: Kurtosis of the region of interest (ROI)intensity	●	
f5: Skewness of the ROI intensity distribution	●	
f6: Entropy of the ROI intensity	●	
f7: Mean of the entropy filtered ROI	●	
f8: Entropy of the complex imaginary part of the convolved region with log-Gabor filters	●	
f9: Entropy of the complex real part of the convolved region with log-Gabor filters	●	
f10:Entropy of the magnitude part of the convolved region with log-Gabor filters	●	
f11:The max coefficient of the 2D discrete Fourier transform applied to the ROI	●	
f12:Entropy of the normalised co-occurrence matrix of the ROI	●	
f13:The magnitude of the difference mean of the region and the mean of the complement ROI	●	
f14:The max coefficient of the 2D discrete Fourier transform of the Hessian filter applied to the ROI	●	
f15:Entropy of the entire breast area*	●	
f16:The max coefficient of the log of the magnitude part of the discrete cosine transform applied to the ROI	●	
f17:The max coefficient of the log of the magnitude part of the discrete Fourier transform applied to the ROI	●	
f18:The 4th central moment of the ROI		●
f19:Number of particles within the ROI		●
f20:Solidity of the ROI		●
f21:Eccentricity of the ROI		●
f22:Euler Number of the ROI		●
f23:Number of particles within the entire breast*		●
f24: (Skewness of the normalised singular value decomposition of the ROI) / (standard deviation of the ROI)	●	
f25:Singular value decomposition of the ROI	●	
f26:Euler number within the breast (binary image)	●	
f27:Interquartile range- the difference between the 75th and the 25th percentiles of the intensity values of the ROI	●	
f28:The 1st Fractal descriptor of the entire breast*	●	
f29:Mean intensity of the breast*	●	
f30:Skewness of the eroded ROI intensity	●	
f31:Skewness of the intensity distribution of the breast*	●	
f32:Kurtosis of the projection along the Y axis of the ROI		●
f33:Kurtosis of the projection along the X axis of the ROI		●
f34:Mean perimeter of the ROI		●
f35:Mean circularity of smallest particles 1-100 pix*		●
f36:Mean intensity of the ROI	●	
f37:Homogeneity of the breast area*	●	
f38:Median intensity of the ROI	●	

f39:Skewness of the gradient of the intensity within the ROI	●	
f40:Energy property of the co-occurrence matrix of the selected area - horizontal shift (2 pixels)	●	
f41:Energy property of the co-occurrence matrix of the selected area - diagonal shift (8 pixels)	●	
f42:Variance of the intensity within the breast area*	●	
f43:Mean of the local binary pattern (LBP) of the whole breast region*		●
f44:Kurtosis of the local LBP histogram of the whole breast region*		●
f45:Mean intensity of the pectoral muscle*	●	
f46:RunLengthCode: Short Run Emphasis (SRE), Long Run Emphasis(LRE), Gray Level Non-Uniformity (GLN), Run Percentage (RP), Run Length Non-Uniformity (RLN), Low Gray Level Run Emphasis (LGRE), High Gray Level Run Emphasis (HGRE)*	●	
f47:Entropy of the sum of the average values of co-occurrence matrix of the breast along the horizontal and diagonal directions *	●	
f48:Standard deviation of FD (fractal descriptor) for the selected area *	●	
F49: Lacunarity of FD for the selected area *	●	
f50: Laws Texture Energy Measures: Entropy Level, Entropy Edge, Entropy Spot, Entropy Ripple *	●	
f51: Microcalcifications *		
f52: Intensity ratio(dense/breast) *	●	
f53: Dense area size *		●
f54: Breast area *		●
f55: Percentage Fatty *		●

Table S2. The table depicts the twelve different regions used in our approach from which features in Table S1 are derived.

Region Label	Description
Region 1	Finds the *dense* region in mammographic images using the *p-tile method*
Region 2	Finds the *dense* region in mammographic images using the *entropy of the histogram*.
Region 3 (*)	Finds the *dense* region in mammographic images usingthe *moment preserving thresholding method*
Region 4	Finds the *dense* region in mammographic images using the *Otsu's method*.
Region 5	Finds the *dense* region in mammographic images using the *Iterative Otsu's method*.
Region 6	Finds the *dense* region in mammographic images using the *arithmetic mean of pixels.*
Region 7 (~)	Finds the *fatty* region in mammographic images using our customized threshold (see Fig.S1(c) & Eq.S2)
Region 8 (~)	Finds the *fatty* region in mammographic images using the *entropy of the histogram*.
Region 9 (~)	Finds the *fatty* region in mammographic images using the *p-tile method*
Region 10 (~)	Finds the *fatty* region in mammographic images using the *Otsu's method*.
Region 11 (~)	Finds the *fatty* region in mammographic images usingthe *moment preserving thresholding method*
Region 12	Finds the *dense* region in mammographic images using our customized threshold (see Fig.S1(c) & Eq.S2)

(~) Denotes the logical "NOT" of a binary matrix. (*) CASAM-Area is calculated based on Region 6 in Eq.1 (in the main manuscript).

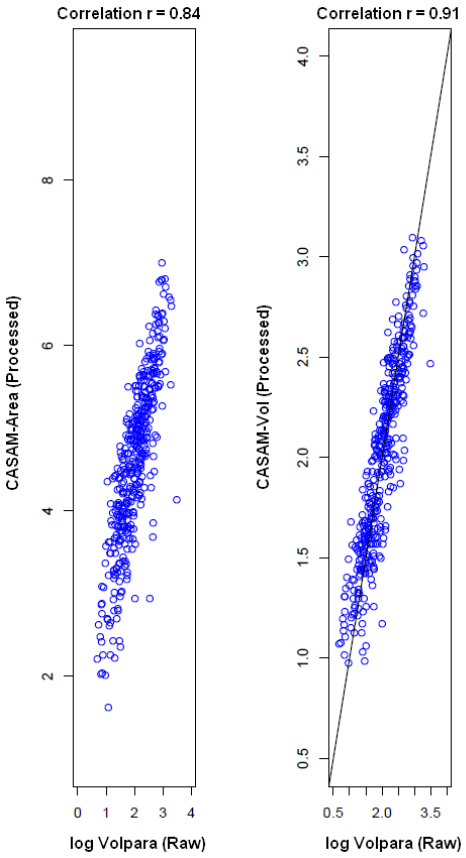

Figure.S2. Scatter plot showing the correlation between log (Volpara^{TM}) on raw images and CASAM-Area and the predicted CASAM-Vol on the processed images for the training data set (403).

Table S3: p-values assessing the (residual) association between automated measures of mammographic density and case-control status (after adjustment for one other density measure). (a) with partial adjustment (age and BMI), (b) with full adjustment (age, BMI, menopausal status, HRT use, parity and age at first birth).

(a)			**Variables adjusted for**	
Covariates (n=1059)	Standard	Standard+ Volpara^{TM}	Standard+ CASAM-Area (Processed)	Standard+ CASAM-Vol
Volpara^{TM}	0.005	-	0.479	**0.823**
CASAM-Area (Processed)	0.012	0.163	-	**1.000**
CASAM-Vol (Processed)	0.023	**0.088**	**0.235**	-
(b)				
Volpara^{TM}	0.010	-	0.527	**0.560**
CASAM-Area (Processed)	0.023	0.173	-	**0.777**
CASAM-Vol (Processed)	0.065	**0.054**	**0.158**	-