Using a Convolutional Neural Network to Predict Readers' Estimates of Mammographic Density for Breast Cancer Risk Assessment

Georgia V. Ionescu[a], Martin Fergie[b], Michael Berks[b], Elaine F. Harkness[b,c,f], Johan Hulleman[d], Adam R. Brentnall[e], Jack Cuzick[e], D. Gareth Evans[f,g,h], and Susan M. Astley[b,c,f]

[a]School of Computer Science, University of Manchester, Stopford Building, Oxford Road, Manchester, UK
[b]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Stopford Building, Oxford Road, Manchester, UK
[c]The University of Manchester, Manchester Academic Health Science Centre, Manchester NHS Foundation Trust, Manchester, UK
[d]School of Biological Sciences, Division of Neuroscience and Experimental Psychology, University of Manchester, Manchester, UK
[e]Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London EC1M 6BQ, UK
[f]Prevent Breast Cancer and Nightingale Breast Screening Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Southmoor Road, Wythenshawe, Manchester M23 9LT, UK
[g]9The Christie NHS Foundation Trust, Manchester Academic Health Science Centre, Withington, Manchester M20 4BX, UK
[h]Genomic Medicine, Division of Evolution and Genomic Science, Manchester Academic Health Sciences Centre, University of Manchester and Manchester University NHS Foundation Trust, Manchester M13 9WL, UK

ABSTRACT

Background: Mammographic density is an important risk factor for breast cancer. Recent research demonstrated that percentage density assessed visually using Visual Analogue Scales (VAS) showed stronger risk prediction than existing automated density measures, suggesting readers may recognise relevant image features not yet captured by automated methods.

Method: We have built convolutional neural networks (CNN) to predict VAS scores from full-field digital mammograms. The CNNs are trained using whole-image mammograms, each labelled with the average VAS score of two independent readers. They learn a mapping between mammographic appearance and VAS score so that at test time, they can predict VAS score for an unseen image. Networks were trained using 67520 mammographic images from 16968 women, and tested on a large dataset of 73128 images and case-control sets of contralateral mammograms of screen detected cancers and prior images of women with cancers detected subsequently, matched to controls on age, menopausal status, parity, HRT and BMI.

Results: Pearson's correlation coefficient between readers' and predicted VAS in the large dataset was 0.79 per mammogram and 0.83 per woman (averaging over all views). In the case-control sets, odds ratios of cancer in the highest vs lowest quintile of percentage density were 3.07 (95%CI: 1.97 - 4.77) for the screen detected cancers and 3.52 (2.22 - 5.58) for the priors, with matched concordance indices of 0.59 (0.55 - 0.64) and 0.61 (0.58 - 0.65) respectively.

Conclusion: Our fully automated method demonstrated encouraging results which compare well with existing methods, including VAS.

Keywords: breast cancer, mammographic density, deep learning, risk, VAS

1. INTRODUCTION

Mammographic density is one of the most important independent risk factors for breast cancer and is defined as the relative proportion of dense to fatty tissue in the breast as visualised in mammograms. Women with dense breasts have a 4-6 fold increased risk of breast cancer compared to women with fatty breasts.[1] Additionally, dense breasts may mask possible cancers, reducing readers' sensitivity.[2] A number of area and volumetric based methods exist to measure mammographic density (MD). These include visual area-based methods, for example BI-RADS breast composition categories,[3] Boyd categories,[4] the Visual Analogue Scale (VAS),[5] semi-automated thresholding (Cumulus).[6] The automated Densitas software[7] operates in an area-based fashion on processed "for presentation" full field digital mammograms [FFDM] whilst methods including Volpara[8] and Quantra[9] use the raw "for processing" mammogram to estimate volumes of dense fibroglandular and fatty tissue in the breast. Density measures may be expressed in absolute terms (area or volume of dense tissue) or more commonly as a percentage expressing the relative proportion of dense tissue in the breast. Recent studies have investigated density-breast cancer associations and found differences depending on the density method used.[10,11] Furthermore, it has been shown that measures of density can improve the accuracy of cancer risk prediction models.[12]

Subjective assessment of percentage density recorded on Visual Analogue Scales (VAS) has a strong relationship with breast cancer risk. In a recent case-control study.[13] with three matched controls for each cancer (366 detected in the contra-lateral breast at screening on entry to the study and 338 detected subsequently), the odds ratio for screen detected cancers in the contra-lateral breast in the highest compared to the lowest quintile of percentage VAS was 4.37 (95% CI: 2.72 - 7.03) compared 2.42 (95% CI: 1.56 - 3.78) and 2.17 (95% CI: 1.413.33) for Volpara and Densitas percent density respectively. Similar results were found for subsequent cancers, with odds ratios of 4.48 (95%CI:2.79 - 7.18) for VAS, 2.87 (95%CI:1.77-4.64) for Volpara and 2.34 (95% CI: 1.50 - 3.68) for Densitas. This suggests that expert readers may recognise important features present in the mammographic images of high risk women which existing automated methods may miss. In part this may be due to their assessment of patterns of density as well as quantity of dense tissue; there is already evidence in the same case-control setting that explicit quantification of density patterns adds independent information to percent density for risk prediction.[14]

Most conventional machine learning algorithms require hand-crafted descriptive features and prior knowledge of the data. Conversely, deep learning techniques extract and learn relevant features directly from the data, without prior knowledge.[15] Convolutional neural networks (CNN) have been successfully used for a wide range of imaging tasks including image classification,[16] object detection, and semantic segmentation.[17] In mammography, deep learning has been used for breast segmentation,[18] breast lesion detection[19] and breast mass segmentation.[20] Moreover, deep learning has been employed for differentiation between benign and malignant masses[21] and to discriminate between masses and microcalcifications.[19] A method that consists of a cascade of deep learning and random forest classifiers has been used for detecting mases in mammograms.[22] Unlabelled deep learning has been used for breast density segmentation and risk scoring.[23,24] Petersen et al.[23] were amongst the first to propose autoencoders as a method for breast density estimation, using a multiscale denoising autoencoder to learn an image representation to train a machine learning model to estimate breast density. Kallenberg et al.[24] proposed a variant of the autoencoder that learns a sparse overcomplete representation of the features. A recent study employed deep learning to classify breast density BI-RADS categories and to differentiate between "scattered density" and "heterogeneously dense" breasts, showing promising results.[25]

In this paper, we trained two deep neural networks to learn features associated with breast cancer with an aim to create an automated method with the potential to match human performance on breast cancer risk assessment. Our approach is to train a CNN to predict mammographic density VAS scores with the final goal of assessing breast cancer risk.

2. DATA

We used data from the Predicting Risk Of Cancer At Screening (PROCAS) study.[26] 57,905 women were recruited to PROCAS between October 2009 and March 2015, with full-field digital mammograms available for 44,505. VAS scores were produced as described in Section 3.1. Data from women who had cancer prior to entering the

PROCAS study were excluded from this study, as were those women with additional views, to avoid ambiguity. PROCAS mammograms were in three different formats as shown in Table 1. Due to memory limitations, those with format C were excluded. The total number of exclusions for all criteria (n=21299) are shown in Table 2 leaving a total of 36606 women and 145820 mammographic images for analysis.

Table 1: Mammographic image formats in PROCAS

Format	Dimensions (pixels)	Pixel Size (µm)
A	1914x2294	94.1
B	2394x3062	94.1
C	4095x5625	54.0

Table 2: Exclusion table. [a]

Exclusion Reason	Number of Women Excluded
additional views	2384
mammographic image size	6513
prior cancer	1068
FFDM unavailable	13400

[a]Some exclusions fall into more than one category

2.1 Training data

The training set was built by randomly selecting 50% of the eligible women from PROCAS. All women that were part of the two case control test sets described in Section 2.2 were further excluded from the training set to ensure no overlap between training and test sets. The training set consisted of 67520 images from 16968 women (132 cancers and 16836 non-cancers).

2.2 Test data

We evaluated our method using three datasets: the PROCAS 50%, screen-detected cancers (SDC) and the Priors datasets. The SDC and Priors datasets are the same as the case-control studies used by Astley et al.[10]

PROCAS dataset 50%

The PROCAS 50% test set consisted of data from the remaining 50% of women (73128 images from 18360 women, 393 cancers and 17967 non-cancers) that were not used in the test set. We used all four mammographic views and analysed the data both per mammogram and per woman.

SDC dataset

The SDC dataset was a subset of PROCAS with mammographic images from 1646 women (366 cancers and 1098 non-cancers). All cancers were detected at the screen on entry to PROCAS. Mammographic density was assessed in the contralateral breast of women with cancer and in the same breast for the matched controls. Each case was matched to three controls based on age (+/-12 months), BMI category (missing, <24.9, 25.0-29.9, 30+ $[kg/m^2]$), hormone replacement therapy (HRT) use (current vs never/ever) and menopausal status (premenopausal, perimenopausal or postmenopausal). Controls had a cancer-free (normal) mammogram at entry to PROCAS, but also had a subsequent cancer-free (normal) mammogram.

Priors dataset

The Priors dataset consisted of a case-control set of 338 cancers and 1014 controls also from the PROCAS study. All cases in this dataset were cancer-free on entry to PROCAS but diagnosed subsequently. We analysed the mammographic images of these women on entry to PROCAS. Controls had a cancer-free (normal) mammogram at entry to PROCAS, but also had a subsequent cancer-free (normal) mammogram. In this analysis we used data from all four mammographic views. Similarly to the SDC dataset, cases were matched to three controls based on age, BMI category, HRT, menopausal status and also year of mammogram.

3. MATERIALS AND METHODS

3.1 Visual assessment of density

In the Predicting Risk of Cancer At Screening (PROCAS) study, mammograms had their density assessed by two of nineteen independent readers (radiologists, advanced practitioner radiographers and breast physicians). The Visual Analogue Scale (VAS) used was a 10cm line marked at the ends with 0% and 100%. Each reader marked their assessment of breast density on one scale for each mammographic view. Readers were assigned on a pragmatic basis. The VAS score for each mammographic image was computed as the average of the two readers' scores. The VAS score per woman was averaged across all four mammographic images and across the two readers.

3.2 Deep learning model

We propose an automated method for assessing breast cancer risk based on whole-image full- field digital mammograms (FFDM) using readers' VAS scores as a measure of breast density. As a first step, we built a deep CNN that takes whole-image mammograms as input and predicts a single number between 0 and 100. This number corresponds to the VAS score (percentage density). One of the main characteristics of CNNs is that features are learned from the training data without human input and are directly optimised for the task at hand. Features (often referred to as filters) are small patches which are convolved with the input image and create activation maps that show how the input responds to the filters. The values of the features are automatically adjusted to optimise an objective function, in this case the minimisation of the squared difference between predicted and reader VAS scores. Our implementation uses the *TensorFlow* library.[27] Our network consists of 6 groups of 2 convolutional layers and a max pooling layer. Figure 1 shows a conceptual representation of the network, the complete architecture is shown in Figure 2. We use the ReLU[28] activation function and apply batch normalisation[29] before ReLU.

3.3 Pre-processing

All mammographic images had the same spatial resolution. In order to have a single mammogram size, we padded format A mammograms with zeros on the bottom and right edges to match the resolution of format B mammograms. Right breast mammograms were flipped horizontally before padding. Further, all mammograms were cropped to 2394x2995 and scaled to 512x640 pixels using bicubic interpolation. Images were downscaled due to limited memory demands. The upper bound of the pixel values was set to 75% of the pixel value range, to reduce the difference between background and breast pixel intensity. Finally, we inverted the pixel intensities and applied histogram equalisation (256 bins).[30] All pixel values were normalised in the range [0,1] before images were fed into the network.

3.4 Training the CNN

We trained two separate network architectures shown in Figure 2. The first was trained on cranio-caudal (CC) images and the second on medio-lateral oblique (MLO) images. The network takes as input pre-processed mammographic images (512x640 pixels) and outputs a single value which represents a VAS score. The CNN learns a mapping between the input mammographic image and the output VAS score. A validation set consisting of 5% of the training set was used for parameter selection and to avoid over-fitting. We used the Adam optimiser[31] with an initial learning rate of 5e-06. Training mini-batches were balanced to have an equal number of input images for each VAS score decile. The cost function was a weighted mean squared error; each weight being

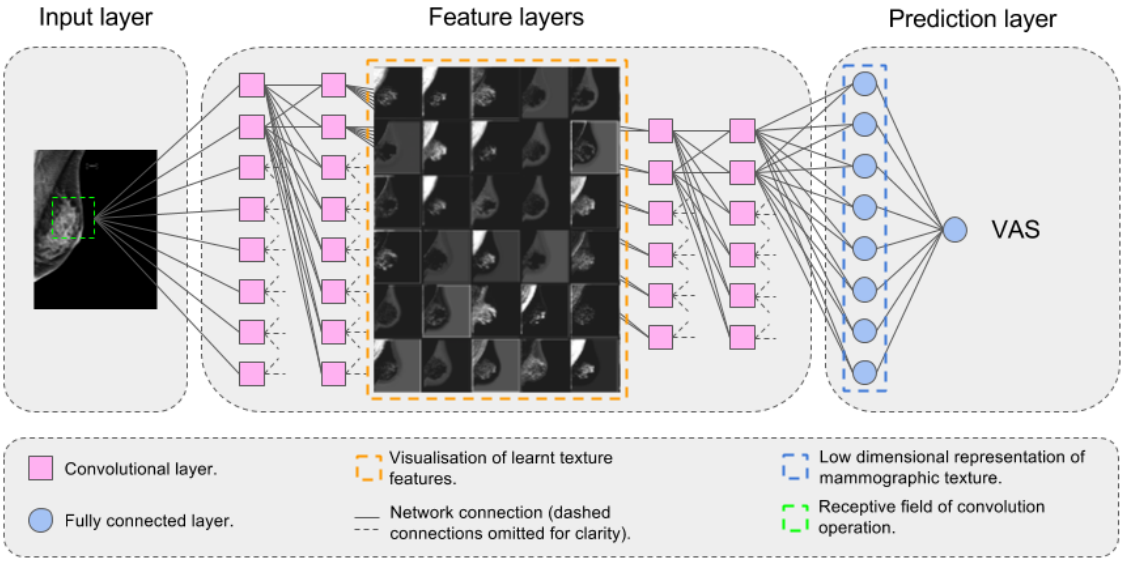


Figure 1: Conceptual diagram of our VAS-score predicting convolutional neural network.

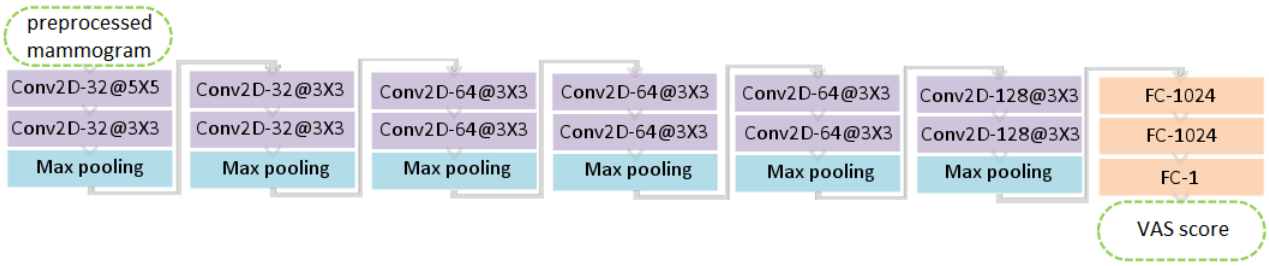


Figure 2: Network architecture and characteristics of each layer. The number of feature maps and the kernel size of each convolutional layer are shown as: *feature maps@kernel size*. The fully connected layers are marked with *FC* followed by the number of neurons in the layer.

inversely proportional with the inter-reader difference so that examples where both readers agree give a larger contribution to the loss function. We trained for 150,000 iterations and selected the model that performed best on the validation set. For the fully connected layers we used a dropout rate of 0.5 at training time.

3.5 Predicting density score

At test time, the network predicted a single VAS score for each mammogram. A small proportion of images (approximately 1%) produced a negative VAS score and were set to zero. The VAS-score for a woman was computed by averaging scores across all mammograms available (both breasts and both views).

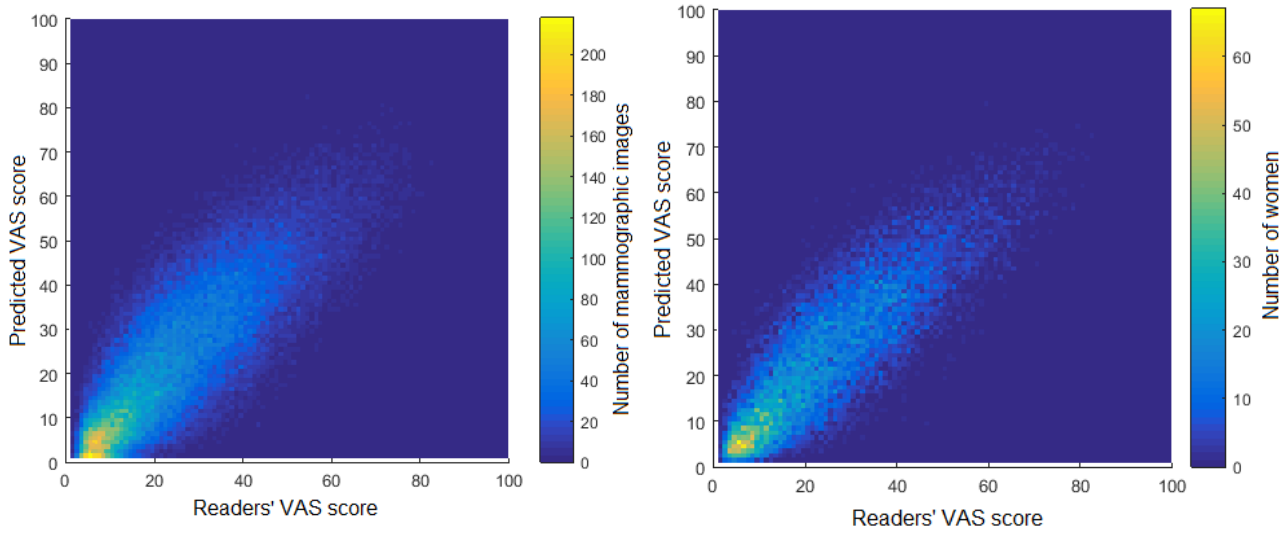
3.6 Analysis

We calculated two types of VAS scores: VAS score per image and VAS score per woman (i.e. an average of all VAS scores for all views for each woman). We evaluated our network's performance on three tasks. The first task was predicting VAS score for previously unseen mammographic images. We computed predicted VAS scores per image and per woman on the PROCAS 50%. The Pearson correlation coefficient between predicted and readers' VAS scores was calculated. Bland-Altman plots[32] were used to evaluate the agreement between readers' and predicted VAS scores and to identify any systematic bias in the predicted VAS score. We computed the reproducibility coefficient (RPC) which quantifies the agreement between readers and predicted VAS. 95% of predicted VAS scores are expected to be within one RPC from median after adjusting for the systematic bias.

Secondly, we assessed the capacity of our network to predict case-control status for screen detected cancers (SDC) and priors using the datasets as described in Section 2.2. The relationship between VAS and case-control status was analysed using conditional logistic regression with density measures modelled as quintiles based on the density distributions of controls. The difference in the likelihood-ratio chi-square between models in the subset of women who had both reader and predicted VAS scores was compared. The matched concordance (mC) index,[33] which provides a statistic similar to the area under the receiving operator characteristic curve (AUC) for matched case-control studies, with empirical bootstrap confidence intervals[33] was calculated to compare the discrimination performance of the models. All p values were two-sided.

4. RESULTS

Figure 3 shows predicted VAS plotted against readers' VAS scores for PROCAS 50%. Figure 3a shows the VAS scores for mammographic images, whilst Figure 3b shows VAS scores computed per woman. The Pearson correlation coefficient was 0.79 per mammogram and 0.83 per woman.



(a) per mammogram

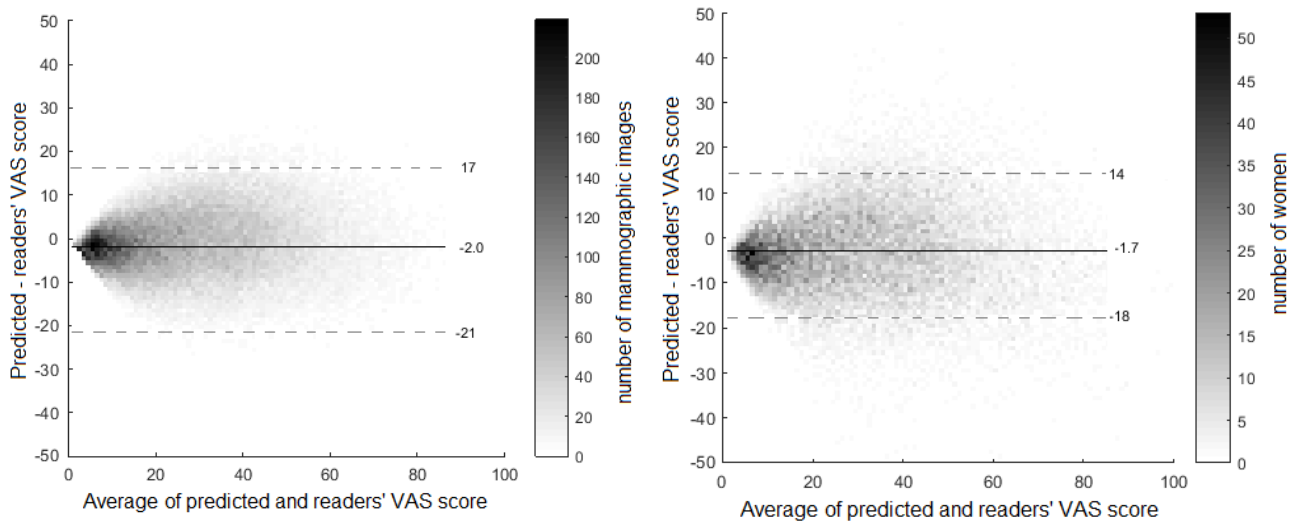
(b) per woman

Figure 3: Predicted VAS score plotted against the average VAS score of 2 readers a) per mammographic image and b) per woman

Figure 4 shows Bland-Altman plots for predicted and readers' VAS scores for PROCAS 50%. A reproducibility coefficient (RPC) of 19 was obtained for VAS scores per mammographic image with a systematic bias of -2.0. For VAS scores per woman the RPC was 16 and there was a systematic bias of -1.7 in predicted VAS.

Figure 5 illustrates the odds of developing breast cancer for women in all quintiles of predicted VAS scores compared to women in the lowest quintile for a) Screen detected dataset and b) Priors dataset.

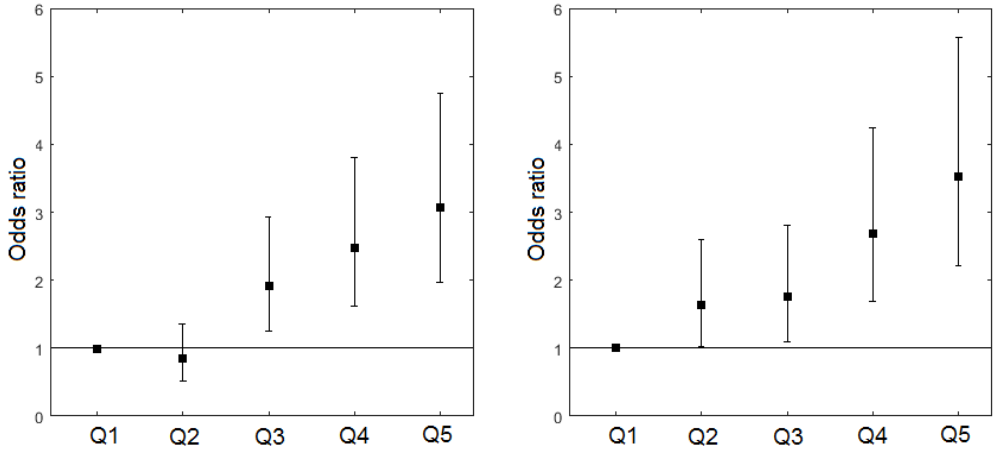
Table 3 shows the odds of developing breast cancer for women in the highest quintile of VAS scores compared to women in the lowest quintile. Predicted and readers' VAS were both significantly associated with breast cancer for SDC and priors, however the odds ratio associated with readers' VAS was higher than that for predicted VAS. For the SDC dataset, the odds ratio for women in the highest quintile compared to women in the lowest quintile of predicted VAS scores was 3.07 (95% CI: 1.97 - 4.77). In the Priors dataset the OR for predicted VAS was 3.52 (95% CI: 2.22 - 5.58). Readers' VAS was a significantly better predictor than the predicted VAS for both case-control datasets (likelihood ratio chi-square, p=0.007 for SDC and p=0.029 for the Priors dataset).



(a) per mammogram

(b) per woman

Figure 4: Bland-Altman plot of predicted and readers' VAS score. The horizontal axis shows the average between readers' and predicted VAS scores; the vertical axis shows the difference between predicted and readers' VAS scores. Solid line represents median, dashed lines show the 95% confidence limits.



(a) Screen detected cancers dataset

(b) Priors dataset

Figure 5: Risk of developing breast cancer odds ratio with 95% CI of predicted VAS for a) the screen detected cancers dataset and b) the priors dataset.

Table 3: Odds ratio (95% CI) for highest quintile compared to lowest quintile of VAS scores

Dataset	Readers' VAS	Predicted VAS
SDC test set	4.63 (2.82 - 7.60)	3.07 (1.97 - 4.77)
Prior test set	4.41 (2.76 - 7.06)	3.52 (2.22 - 5.58)

Table 4 shows the matched concordance index obtained for both case-control datasets. The matched concordance index for reader VAS was higher (0.65, 95% CI: 0.61 - 0.68 for SDC and 0.64, 95% CI: 0.60 - 0.68 for Priors) compared to predicted VAS (0.59, 95% CI: 0.55 - 0.64 for SDC and 0.61, 95% CI: 0.58 - 0.65 for Priors) showing better discrimination between cases and controls for reader VAS.

Table 4: Matched concordance index for predicted and readers' VAS scores for both case-control datasets

Dataset	Readers' matched C-index	Predicted matched C-index
SDC dataset	0.65 (0.61 - 0.68)	0.59 (0.55 - 0.64)
Priors dataset	0.64 (0.60 - 0.68)	0.61 (0.58 - 0.65)

5. DISCUSSION

In this paper we have presented a deep learning method to predict VAS scores for breast density assessment. Subjective assessment of breast density has been shown to be a stronger predictor of breast cancer than other automated and semi-automated methods.[10] Our method is the first automated method to attempt to reproduce readers' VAS scores as an assessment of breast cancer risk; it gives promising preliminary results. We used a large dataset with 145,820 mammographic full-field digital mammograms from 36,606 women and tested our network on 3 datasets. We showed that our CNN is capable of predicting a VAS score that reflects readers' VAS which is the first step towards building a method for cancer risk prediction. Results showed a substantial agreement between readers' VAS scores and predicted VAS scores (Pearson correlation = 0.83 for VAS per woman and 0.79 for VAS per mammographic image). The mean difference (systematic bias) between the reader and predicted VAS was small, however the 95% limits of agreement showed considerable variation, which has been found to be a problem in the visual assessment of breast density both within and between readers (Sergeant et al.[34]). Despite this VAS has been found to be a significant predictor of breast cancer. Consequently we investigated our method's capacity to predict breast cancer in the datasets previously used by Astley et al.[10] Our method performed well, both in predicting breast cancer in women with SDC cancer using the contra-lateral breast, and in predicting the future development of the disease, however the ORs for predicted VAS were lower than those for readers' VAS. Despite this, our method tended to show a stronger association with breast cancer risk than percent density estimates in both SDC and priors for automated methods (Volpara, Quantra and Densitas) as reported by Astley et al.

Our method's strengths include the fact that it requires no human input and the pre-processing step is minimal. This would make it a pragmatic solution for population-based stratified screening. To further improve our method, we plan to remove the downscaling step in the pre-processing phase. We expect that inclusion of fine-scale structures would improve risk estimation, since they are visible to human readers. Moreover, we intend to test our method using processed "for presentation" mammographic images to determine whether similar results could be obtained without the need for storing raw "for processing" images. This would facilitate routine density assessment in screening programmes.

REFERENCES

[1] Huo, C., Chew, G., Britt, K., Ingman, W., Henderson, M., Hopper, J., and Thompson, E., "Mammographic densitya review on the current understanding of its association with breast cancer," *Breast cancer research and treatment* **144**(3), 479–502 (2014).

[2] Mohamed, A. A., Luo, Y., Peng, H., Jankowitz, R. C., and Wu, S., "Understanding clinical mammographic breast density assessment: a deep learning perspective," *Journal of digital imaging* , 1–6 (2017).

[3] Dorsi, C., Bassett, L., Berg, W., Feig, S., Jackson, V., Kopans, D., et al., "Breast imaging reporting and data system: Acr bi-rads-mammography," *American College of Radiology* **4** (2003).

[4] Boyd, N. F., Guo, H., Martin, L. J., Sun, L., Stone, J., Fishell, E., Jong, R. A., Hislop, G., Chiarelli, A., Minkin, S., et al., "Mammographic density and the risk and detection of breast cancer," *New England Journal of Medicine* **356**(3), 227–236 (2007).

[5] Sergeant, J. C., Warwick, J., Evans, D. G., Howell, A., Berks, M., Stavrinos, P., Sahin, S., Wilson, M., Hufton, A., Buchan, I., et al., "Volumetric and area-based breast density measurement in the predicting risk of cancer at screening (PROCAS) study," in [*International Workshop on Digital Mammography*], 228–235, Springer (2012).

[6] Byng, J. W., Boyd, N., Fishell, E., Jong, R., and Yaffe, M. J., "The quantitative analysis of mammographic densities," *Physics in medicine and biology* **39**(10), 1629 (1994).

[7] Abdolell, M., Hope, T., Zaboli, S., and Tsuruda, K., "Methods and systems for determining breast density," (July 14 2016). US Patent App. 14/912,965.

[8] Highnam, R., Brady, S. M., Yaffe, M. J., Karssemeijer, N., and Harvey, J., "Robust breast composition measurement - volpara$^{\text{TM}}$," in [*Proceedings of the 10th International Conference on Digital Mammography*], *IWDM'10*, 342–349, Springer-Verlag, Berlin, Heidelberg (2010).

[9] Pahwa, S., Hari, S., Thulkar, S., and Angraal, S., "Evaluation of breast parenchymal density with quantra software," *The Indian journal of radiology & imaging* **25**(4), 391 (2015).

[10] Astley, S. M., Harkness, E. F., Sergeant, J. C., Warwick, J., Stavrinos, P., Warren, R., Wilson, M., Beetles, U., Gadde, S., Lim, Y., et al., "A comparison of five methods of measuring mammographic density: a case-control study," *Breast Cancer Research* **20**(1), 10 (2018).

[11] Eng, A., Gallant, Z., Shepherd, J., McCormack, V., Li, J., Dowsett, M., Vinnicombe, S., Allen, S., and dos Santos-Silva, I., "Digital mammographic density and breast cancer risk: a case–control study of six alternative density assessment methods," *Breast cancer research* **16**(5), 439 (2014).

[12] Brentnall, A. R., Harkness, E. F., Astley, S. M., Donnelly, L. S., Stavrinos, P., Sampson, S., Fox, L., Sergeant, J. C., Harvie, M. N., Wilson, M., et al., "Mammographic density adds accuracy to both the tyrer-cuzick and gail breast cancer risk models in a prospective uk screening cohort," *Breast Cancer Research* **17**(1), 147 (2015).

[13] Duffy, S. W. et al., "Visually assessed breast density, breast cancer risk and the importance of the cranio-caudal view," *Breast Cancer Research* **4**, 1–7 (2008).

[14] Wang, C., Brentnall, A. R., Cuzick, J., Harkness, E. F., Evans, D. G., and Astley, S., "A novel and fully automated mammographic texture analysis for risk prediction: results from two case-control studies," *Breast Cancer Research* **19**(1), 114 (2017).

[15] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I., "A survey on deep learning in medical image analysis," *Medical image analysis* **42**, 60–88 (2017).

[16] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in Neural Information Processing Systems 25*], 1097–1105, Curran Associates, Inc. (2012).

[17] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," in [*The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (June 2014).

[18] Dubrovina, A., Kisilev, P., Ginsburg, B., Hashoul, S., and Kimmel, R., "Computational mammography using deep neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* , 1–5 (2016).

[19] Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., and Li, L., "Discrimination of breast cancer with microcalcifications on mammography by deep learning," *Scientific reports* **6** (2016).

[20] Dhungel, N., Carneiro, G., and Bradley, A. P., "Deep learning and structured prediction for the segmentation of mass in mammograms," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 605–612, Springer (2015).

[21] Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., and Chen, C.-M., "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans," *Scientific reports* **6** (2016).

[22] Dhungel, N., Carneiro, G., and Bradley, A. P., "Automated mass detection in mammograms using cascaded deep learning and random forests," in [*Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*], 1–8, IEEE (2015).

[23] Petersen, K., Chernoff, K., Nielsen, M., and Ng, A. Y., "Breast density scoring with multiscale denoising autoencoders," in [*Proceedings of the STMI Workshop at the 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI12)*], (2012).

[24] Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., Igel, C., Vachon, C. M., Holland, K., Winkel, R. R., Karssemeijer, N., et al., "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring," *IEEE transactions on medical imaging* **35**(5), 1322–1331 (2016).

[25] Mohamed, A. A., Berg, W. A., Peng, H., Luo, Y., Jankowitz, R. C., and Wu, S., "A deep learning method for classifying mammographic breast density categories," *Medical physics* **45**(1), 314–321 (2018).

[26] Evans, D. G. R. et al., "Assessing individual breast cancer risk within the u.k. national health service breast screening program: A new paradigm for cancer prevention," *Cancer Prevention Research* **5**(7), 943–951 (2012).

[27] Abadi, M. et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015). Software available from tensorflow.org.

[28] Nair, V. and E. Hinton, G., "Rectified linear units improve restricted boltzmann machines," (06 2010).

[29] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*Proceedings of the 32nd International Conference on Machine Learning*], **37**, 448–456, PMLR, Lille, France (07–09 Jul 2015).

[30] Lim, J. S., "Two-dimensional signal and image processing," *Englewood Cliffs, NJ, Prentice Hall* , 710 (1990).

[31] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *CoRR* **abs/1412.6980** (2014).

[32] Altman, D. G. and Bland, J. M., "Measurement in medicine: the analysis of method comparison studies," *The statistician* , 307–317 (1983).

[33] Brentnall, A. R., Cuzick, J., Field, J., and Duffy, S. W., "A concordance index for matched case–control studies with applications in cancer risk," *Statistics in medicine* **34**(3), 396–405 (2015).

[34] Sergeant, J. C., Walshaw, L., Wilson, M., Seed, S., Barr, N., Beetles, U., Boggis, C., Bundred, S., Gadde, S., Lim, Y., et al., "Same task, same observers, different values: the problem with visual assessment of breast density," in [*Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*], **8673**, 86730T, International Society for Optics and Photonics (2013).