The effect of variable labels on deep learning models trained to predict breast density

Steven Squires1, Elaine F Harkness2, D Gareth Evans2 and Susan M Astley∗

1 University of Exeter, Exeter, United Kingdom
2 University of Manchester, Manchester, United Kingdom
∗ Author to whom any correspondence should be addressed.
E-mail: sue.astley@manchester.ac.uk

Keywords: machine learning, computer vision, deep learning, mammographic density, breast cancer, artificial intelligence personalised healthcare

Abstract

Purpose. High breast density is associated with reduced efficacy of mammographic screening and increased risk of developing breast cancer. Accurate and reliable automated density estimates can be used for direct risk prediction and passing density related information to further predictive models. Expert reader assessments of density show a strong relationship to cancer risk but also inter-reader variation. The effect of label variability on model performance is important when considering how to utilise automated methods for both research and clinical purposes.

Methods. We utilise subsets of images with density labels from the same 13 readers and 12 reader pairs, and train a deep transfer learning model which is used to assess how label variability affects the mapping from representation to prediction. We then create two end-to-end models: one that is trained on averaged labels across the reader pairs and the second that is trained using individual reader scores, with a novel alteration to the objective function. The combination of these two end-to-end models allows us to investigate the effect of label variability on the model representation formed.

Results. We show that the trained mappings from representations to labels are altered considerably by the variability of reader scores. Training on labels with distribution variation removed causes the Spearman rank correlation coefficients to rise from 0.751 ± 0.002 to either 0.815 ± 0.026 when averaging across readers or 0.844 ± 0.002 when averaging across images. However, when we train different models to investigate the representation effect we see little difference, with Spearman rank correlation coefficients of 0.846 ± 0.006 and 0.850 ± 0.006 showing no statistically significant difference in the quality of the model representation with regard to density prediction.

Conclusions. We show that the mapping between representation and mammographic density prediction is significantly affected by label variability. However, the effect of the label variability on the model representation is limited.

1. Introduction

Mammographic density is an important factor when assessing women for their risk of developing breast cancer [1, 2]. There are several different methods of assessing density; one that shows a strong relationship with cancer risk is Visual Analogue Scale assessment (VAS) by radiologists and advanced practitioner radiographers [3], specifically the average score assigned by two independent readers, however the process of obtaining these is time-consuming and not scalable. Automated models that could produce similar scores would enable broader use of this measure as a risk prediction tool.

A number of automated methods have been developed to assess breast density [4–7]. Deep learning approaches, both from scratch [8], and using transfer learning [9, 10], have been shown to produce density estimates that correlate well with VAS scores.

The reason for the high performance of VAS scores for assessing risk compared with automated volumetric density measures is hypothesised to be that the domain experts are using their professional judgement rather than a more mechanical approach which determines the quantity of dense tissue. The experts are perhaps taking into account the complex interplay of areas of dense and non-dense tissue, including the
pattern and distribution, which other methods may not consider. Or perhaps they are identifying other signs such as clusters of classifications [11]. The logic of training on the assessments of these experts is to utilise their years of experience to train models to make similar predictions. Deep learning is particularly appropriate for VAS prediction because deep models automatically extracts features that might not be apparent to the conscious mind of humans. However, a significant issue with deep learning models is that they are generally considered to be ‘black boxes’ [12] where we do not know why they are making their predictions. Consequently we either need to be confident in the quality of the model predictions or provide explainable outputs [13].

An issue with the assessment of the quality of model prediction is that expert estimates show both intra- and inter-reader variability [14–16]. Training on variable data may result in models that produce inaccurate or unreliable predictions. In addition, it was shown that with variation in reader scores it becomes difficult to assess the quality of predictive models because a ‘optimal’ model would also show a relatively large error [10]. This difficulty makes it hard to decide which models to choose or to have confidence in the quality of the model performance.

If model predictions are inaccurate women may be assigned to an incorrect risk category, potentially resulting in missed opportunities to perform earlier interventions or to undertake additional investigations, or causing undue alarm. In addition, density scores can be used for other purposes such as estimating the risk of lesions being concealed by dense tissue (masking) [17, 18] or as inputs to other automated models, in which case inaccurate or unreliable models would weaken their quality. It is therefore important to investigate the effect of variability in reader labels on model predictions.

The variability of radiologist density scores has been investigated previously which showed that variation between readers could be substantial [14, 15]. There has also been work done to correct scores between different readers [16]. However, these pieces of work were interested in the variability of the scores themselves, not the effect that this variability has on deep learning-based models.

Deep learning models can be considered as a combination of a model which produces a representation in some feature space together with a mapping module which converts that feature space into a final prediction. Variability in labels may affect either, or both, of these aspects of deep models, and can make a significant difference to how models should be assessed and deployed.

In this work we first adopt a feature extraction and linear regression approach to explore the importance of label variability on the mapping from a feature space to a density score. This methodology allows us to train and test on subsets of the data which enables us to reduce the variation in the labels and assess how the performance of the model changes.

We also investigate the effect of training using variable labels on the quality of the representation created by the deep models; this requires training models end-to-end rather than using them as feature extractors. We create two fully trained models. One is trained on the averaged labels across all the readers and will include the variability due to differences in reader assessment. The second is trained with predictions made separately on the individual reader scores, by a novel alteration to the objective function. This model should be less affected by reader variability in comparison with the model trained on averaged scores.

2. Materials and methods

2.1. Reader variability

Our overall aim is to investigate how variability in reader density labels affects the representation and mapping of our models. Variability can affect the quality of the model through the impact on training and the assessment of model performance. We hypothesise that reader variability has three aspects:

1. **Fundamental differences in assessment of images:** readers have a genuinely different assessment of the density of an image. We know that individual reader assessments differ from one another [14] and also that the average of two reader assessments of density differ from the outputs of automated methods [3]. Understanding these differences may enable us to design more accurate and reliable models for density-based risk assessment by facilitating understanding of image features associated with increased risk.

2. **Different distributions:** readers may record different absolute density values but generally (within some random error margin) agree on the order of the images from low to high density. We might expect two readers which had been exposed to different density distributions to produce different scores but still to monotonically increase together (accepting the random errors). For example, if one reader sees very few high density mammograms they might assign a higher density to images than a reader who sees a lot of high density images, due to a tendency to fill up the density space.

3. **Random errors/uncertainty:** we define random error as the difference left over between two image density assessments after accounting for the distribution and fundamental differences in density prediction. These errors should not be systematic.

2.2. Data-set and pre-processing

The characteristics of this data-set have been extensively described previously [8, 10]. There are 156,972
images from 39,357 women comprising of sets of four mammographic images for each woman corresponding to the four standard radiographic projections used in screening: RCC, LCC, RMLO, LMLO (a small number of women are missing one or more of the projections). Images are all full field digital mammograms and were produced by GE Senographe Essential mammogram systems at three different image sizes. Two independent expert readers assigned on a pragmatic basis from a pool of 19 readers assessed each image as part of the Predicting Risk Of Cancer At Screening (PROCAS) study [19] and marked an assessment of density for each radiographic projection on a 10cm Visual Analogue Scale. We refer to these as VAS density scores and they are in the range 0-100%. On a 10cm Visual Analogue Scale. We refer to these as VAS density scores and they are in the range 0-100%. The pair of VAS scores by the two readers can vary considerably from one another [10] - the purpose of this paper is to investigate how these differences alter the quality of trained models.

For our first method, using feature extraction with linear regression, we investigate how well models perform when trained on scores from individual readers and paired reader subsets of the data. To improve reliability, data from individual readers who scored the density for fewer than 4,000 images (1,000 women) were excluded, leaving data from a total of 13 readers. Data from pairs of readers who jointly scored density for fewer than 4,000 images were also excluded. The number of images read by each of these readers and reader pairs is detailed in tables 3 and 4, respectively. The choice of 4,000 image is a pragmatic one which trades of including more readers with ensuring there is enough data such that we can be confident in the conclusions.

The second method is designed to investigate the effect of label variability on the model representation in feature space. This approach uses end-to-end training: we split the whole data-set into 80% training (123169 images), 10% validation (15689 images) and 10% testing (15291 images) sets by woman so that no images from the same woman are in more than one partition.

In addition, a subset of the complete data-set contains a case control set with images from 338 women. Each of these women are matched with three women without cancer. By controlling for known risk factors we can then produce an estimate regarding the effect of the mammographic density on cancer risk [3]. None of these women were included in the training and validation partitions. We report the quality of the cancer risk predictions made by all the models produced here.

For pre-processing we scale the images down to a size of 224 × 224, flip left sided images to the right, limit the pixel value range, invert the pixels, perform histogram equalization and set the pixel range between 0 and 1. This approach has been described previously [10].

2.3. Method 1: transfer learning subset data models for the investigation of model mapping

For this method we consider the effect of training on the subsets of the data from either an individual reader or the same pair of readers. For clarity, by individual readers we mean that the images included in training are all read by the same reader and the labels used are the labels provided by that reader. By paired readers we mean the set of images read by both readers and the label is the average of the two reader scores. Due to the relatively small amount of data we need models that will produce accurate results but require only a modest number of images. In previous work it was shown that using a feature extractor with a linear regression produces reasonable density predictions. While the authors showed that a multi-layer perceptron with rectified linear unit activation functions trained on top of the deep feature extractor produced improved results over linear regression differences were modest [10]. An additional advantage of using a linear regression model trained on top of extracted features is that the results are entirely deterministic - they will be identical each time and are optimal for a linear combination of those features (under the mean squared error metric and with that level of regularisation). As we are interested in investigating the effect of variability of the reader estimates it is important to have a consistent model so that we can be confident in the conclusions we draw.

Drawing on previous work [10] we utilise the DenseNet [20] model trained on ImageNet [21] to produce feature vectors and then train a linear regression to produce density predictions. Pre-processed images (see section 2.2) are fed through a truncated DenseNet model where the final fully connected layer has been removed. The DenseNet model then outputs a 2,208 dimensional feature vector for each input image. We add a bias term (making a 2,209 feature vector) and then use linear regression to produce a density estimate using reader labels as the target. In figure 1 we show a schematic version demonstrating the approach for the inference stage.

To make our density predictions we utilise the pseudo-inverse to calculate weights:

\[
w = \left( X^T y + \lambda_2 I \right)^{-1} X^T y
\]

where \(X, y \in \mathbb{R}^{p \times (p+1)}\) is a matrix formed by stacking the \(p\)-dimensional feature vectors (with the addition of the bias term) produced by the application of the DenseNet feature vector to the training images, \(\lambda_2\) is a parameter controlling the \(L_2\) regularisation, \(I\) is the identity matrix and \(y \in \mathbb{R}^p\) is a vector containing the reader labels. To perform inference on the testing data we apply \(\hat{y} = X_w^T w^T\) where \(X_w^T\) is a matrix of the stacked feature vectors from the testing set.

As we are using deep transfer learning with a model trained from ImageNet there is no need to retrain the feature extractor model. The only part of the model we need to train at each point is the linear
regression. In terms of the computation we have a matrix $X$ containing the $n$ DenseNet derived feature vectors for all our images. Alongside that we have a $Y \in \mathbb{R}^{n \times 2}$ matrix which contains all the pairs of reader scores for all the images. Our approach requires us to sample the correct training and testing data from the matrices and find the appropriate $w$. On a standard desktop computer the training process takes seconds for the entire data-set and less for any of the subsets.

For all of this method we train our models using five-fold cross-validation to make predictions. Thus our predictions are actually made using five different models with the predictions reported as those on the held-out sets. This enables us to make predictions on all sub-sets of the data and although it does add a small amount of additional variation, due to the slightly different training sets, this is not enough to alter our conclusions.

We produce predictions on the whole data-set, by training using both the averaged reader scores as well as individual reader scores. These provide a baseline to assess the performance of models trained on the subsets of the data.

We can then examine the effect of training and testing on the individual reader scores and pairs of readers. For the former we take just those images with scores by one individual reader and produce a model which is trained using image/label pairs from that reader alone. For pairs of readers, we train models on the images and averaged scores for the pair. We present results after combining the predictions on the CC and MLO images. The differences in prediction between CC and MLO are generally small.

When we train, and test, on individual reader scores we are reducing both the variation due to the fundamental density assessment and also the distribution (points 1 and 2 in section 2.1). We consider this to be a reduction rather than a removal as individual readers may alter their judgement over time. When utilising paired reader scores we are reducing variations due to all three reasons discussed in section 2.1 although the extent of the reduction is not certain. Using different subsets of the data enables us to investigate the mapping effect and to seek insight into how much of the error is real - due to the model making predictions that differ from the true value.

To summarise the method: we take the sub-set of image/label pairs, we train five different linear regression models using five-fold cross-validation, the models are used to make predictions on the held-out set and these predictions are compared to the labels. The labels for each subset is either from the individual reader or the pair of readers and the images are those associated with the particular labels.

2.4. Method 2: end-to-end deep models to investigate the representation formed

The possible effect of variable reader scores on a representation formed by a deep model can be illustrated with an example. If we take two images which have similar true density but they have been labelled by different pairs of readers then they may have significantly different scores assigned to them. When we train on these two images the model may fluctuate in training as the model parameters are pulled in different directions trying to make the model learn the different labels. Alternatively, the model may be forced to learn noise to make a good prediction on the training set. This variability could potentially have a significant effect on the quality of the representation formed.

To create a model which should be fully affected by label variability we utilise a ResNet-18 [22] implementation pretrained on ImageNet [21]. We use ResNet-18 rather than the DenseNet model used in Method 1 because ResNet-18 is less computationally intensive to train than DenseNet. Method 1 and Method 2 are not comparable because of the different datasets used for training and testing so there is no need to make a direct comparison between them. The final fully connected layer that maps from 512 neurons to the original 1,000 output neurons (for the 1,000 classes in ImageNet) is replaced with a layer that maps to one output neuron. The model is trained using the Adam [23] optimiser with different learning rates. We augment our images by: flipping left-right with a 50% probability, random rotations and the addition of Gaussian noise. As the ResNet architecture requires

---

**Figure 1.** A schematic diagram showing the steps in the prediction procedure in the inference stage. Feature extraction involves the application of the pre-trained DenseNet model, with the fully connected part of the network removed, to a pre-processed mammogram image producing a 2,208 feature vector for that image. We add a bias term to make it a 2,209 dimensional vector. The linear regression step applies the learnt weights to the feature vector to output a final density prediction, $\rho$. 
three input channels we copy the greyscale image (after augmentation so all three images are identical) across the three channels. The training, validation and testing split was discussed in section 2.2. We train on the combination of the CC and MLO images, selecting the final model by the root mean squared error performance on the validation set. All results are reported on the testing set which consists of 15,291 images. We call this model the single-predictor.

Our second end-to-end model, the multi-predictor, is aimed at reducing the effect of the label variability on the representation formed. A schematic is shown in figure 2. The fully convolutional network is taken from the ResNet [22] model with the weights pretrained on ImageNet [21] and the final fully connected layers removed. The output of the convolution part of our model is flattened and concatenated to produce a 512-dimensional feature vector; up to this point this model is identical to the previous single-predictor model. This feature vector is then linearly mapped to \( m = 13 \) output neurons each which represent one reader score. The model is trained end-to-end with the linear mapping included in the adjusted model. By training with \( m \) output neurons, we can let the model adjust for the label variability by adjustments in the linear weights of the last layer of the model.

For an individual image we have two labels to compare with our two predicted outputs but 11 unknown labels. The entire convolutional part of the model (see figure 2) is updated for each image; but in the final layer only the linear weights of the neurons with known labels are updated. Some of the variation between the different readers will be accommodated by that linear part without affecting the rest of the model - the model should not be pulled different ways by the variability of the labels. However, we still get a signal for all of the images which can be used to adjust the fully connected part of the network.

To achieve this we utilise mean squared error as a loss function with an additional filter which prevents errors from output neurons with unknown labels from backpropagating through the network. Our prediction output is a \( m \)-dimensional vector, \( p \in \mathbb{R}^m \), where \( m \) is the total number of readers across the entire data-set (13 in our case). We then create a \( m \)-dimensional vector that is a binary filter, \( \phi \in \mathbb{R}^m \). The vector of labels, \( d \in \mathbb{R}^m \) is given by a \( m \)-dimensional vector where each element is a density score from a specific reader - all but two of which are unknown. Our model will make a prediction for all \( m \) output neurons and this filter will remove those from the objective function for which there is no label. For an individual image then we have a loss given by:

\[
\text{loss} = \sum_i^m [\phi_i(p_i - d_i)^2] \quad (2)
\]

where \( \phi_i \) is 1 if a label is known and 0 otherwise. The filter therefore prevents an error contribution from those outputs that do not have labels.

For the weights of the final fully connected part of the network only those weights that map to the output neurons with known labels will be updated, all the other neurons in the final fully connected layer will be left the same. However, as the convolutional layers all feed into the fully connected layer the entirety of the fully convolutional layers will be updated for all the images. Therefore, we update the fully convolutional layers with the entire training set whilst updating the linear layers with only the labelled data.

The training procedure we follow is identical to that of the single-predictor. This should ensure the only difference in outcomes is due to the methodological differences, not because of differences in the training process or the base-model.

The purpose of these two end-to-end models is to investigate the effect of the label variability on the quality of the representation formed. To do so we need to map the representations to the reader scores. We utilise the same approach as for Method 1 to map the representations to the final scores. We remove the final fully connected layer resulting in a 512 dimensional feature vector. Linear regression is then applied in the same manner as method 1.

To summarise this method: we train the single- and multi-predictors identically using the same training data. We remove the final layer (which maps to either 1 or 13 output neurons) and then train a linear regression on the 512-dimensional output to produce predictions of both models on the same data.
2.5 Comparisons and metrics
We utilise two metrics to compare to the labels: Root Mean Squared Error (RMSE) and Spearman rank correlation (Rank Corr.), the former to give absolute differences and the latter to give rank differences. We also use a case-control set to evaluate odds-ratios of developing cancer [3].

We can compare to either averaged pairs of reader scores or individual scores. Most of our uncertainties are found via bootstrapping and are reported at the 95% level. Where we have used other approaches they are discussed as the results are reported.

3. Results

3.1 Method 1: transfer learning subset data models for the investigation of model mapping

The prediction metrics on the whole dataset (via cross-validation), of 76,973 CC images and 77,080 MLO images are shown in table 1. The metrics are Spearman’s rank correlation coefficient (Rank Corr.) and root mean squared error (RMSE) per image. The uncertainty is at the 95% level produced by performing 1,000 sets of bootstrapping. We show the results found by training (Tr) on the averaged labels (Av) and the individual labels (Ind) along with testing on the averaged labels and the individual labels (Ts), which gives us four sets of prediction metrics in total. There are large differences in model performance between the individual and averaged test sets. However, there is a much smaller variation due to using individual or averaged labels for the training set. We note that a similar finding was made previously and discussed in some detail [10] including when they used a non-linear MLP after the feature extraction, so this is not due to the linear regression. The main purpose of these results though is to provide global metrics for comparison when we investigate the effect of training on the individual and paired values.

When we train our models on the paired and individual reader subsets we reduce the amount of data which we would expect to result in poorer predictive performance. We therefore produce a set of metrics which gives us an estimate of the predictive quality when the data-set is smaller. In table 2 we show the RMSE for different sizes of the data-set, we see a steady reduction in predictive quality with smaller numbers of images. The numbers refer to the numbers of CC or MLO images.

Tables 1 and 2 show global metrics for comparative purposes. In table 3 we show the quality of the predictions when trained and assessed on only those images labelled by the 13 individual readers (denoted as A-M) and with that individual’s density score as the label for each image. We also show the approximate (as the number of CC and MLO are not exactly equal) number of images (Number) that each reader has seen. The table also shows the average across the readers as \( Av1 = \frac{1}{n_i} \sum_{i=1}^{n} \mu_i \), where \( n_i \) is the number of readers and \( \mu_i \) the metric of the \( i \)th reader. The given uncertainty for \( Av1 \) is the standard deviation across the metrics. The metric \( Av2 \) is an average across all the images which includes repeated image-score pairs. The uncertainty for \( Av2 \) was found via bootstrapping and is not directly comparable to the uncertainty of \( Av1 \). The comparison we are interested in is with the quality of the global predictions on the individual (Ind-Ind) shown in table 1 (also shown as All in this table for ease of comparison). The predictions show most models getting significant improvements in performance. If we factor in the smaller size of the training set, by considering table 2, the gap in effective performance is even higher.

---

Table 1. Global metrics (those for the whole dataset rather than the subsets) for method 1 when training (Tr) and testing (Ts) on averaged (Av) and individual (Ind) labels. Predictions are made via five-fold cross validation with results shown for the concatenated hold out data. Uncertainty is reported at the 95% level and found via bootstrapping with 1,000 repeats. These results can be considered as the baseline we then compare the results produced on the subsets to.

| Tr       | Ts       | Rank Corr. | RMSE |
|----------|----------|------------|------|
| Av       | Av       | 0.820 ± 0.002 | 9.38 ± 0.04 |
| Ind      | Ind      | 0.751 ± 0.002 | 11.63 ± 0.05 |
| Av       | Ind      | 0.752 ± 0.002 | 11.62 ± 0.04 |
| Ind      | Av       | 0.818 ± 0.002 | 9.44 ± 0.04 |

Table 2. RMSE for different sizes of the data-set. The number of images refers to the number of CC or MLO images therefore the number that go into the model before cross-validation takes place. These show how much the quality of the metrics fall as the amount of available data decreases which will occur as we train and test our models on data subsets.

| Image number | Av-Av | Ind-Ind | Av-Ind | Ind-Ind |
|--------------|-------|---------|--------|---------|
| 20,000       | 9.54 ± 0.08 | 11.80 ± 0.11 | 11.73 ± 0.08 | 9.67 ± 0.08 |
| 10,000       | 9.73 ± 0.11 | 12.03 ± 0.16 | 11.93 ± 0.11 | 9.92 ± 0.11 |
| 7,000        | 9.79 ± 0.13 | 12.07 ± 0.19 | 11.93 ± 0.13 | 10.02 ± 0.14 |
| 4,000        | 10.05 ± 0.18 | 12.22 ± 0.24 | 12.08 ± 0.16 | 10.28 ± 0.17 |
| 2,000        | 10.33 ± 0.26 | 12.79 ± 0.36 | 12.51 ± 0.23 | 10.67 ± 0.27 |
The results in tables 1 and 2 demonstrate that assessing and, to a lesser extent, training on the averaged labels produces better model performance than using individual labels. We therefore also use the paired set of data for training and assessing and show metrics resulting from this process in table 4. The metrics associated with Av1 and Av2 are the same as in table 3. We also include an additional set of metrics called All which is the metrics when we train together on all the included paired data, i.e. it is another global metric of all the image and scores together but excluding those with fewer than 4,000 images read between them. We see the same pattern as when we trained and tested on the individual reader scores: that we find significantly improved performance compared to training on all the data together.

In figure 3 we show the variability of individual reader scores for two example pairs. The top row shows the image scores for one reader against another reader, with the reader identifiers in the titles. The bottom row shows histograms of the distributions for the pair of readers. The histograms on the bottom row show the significant variation in distribution that the readers show. There is also other substantial variation but separating randomness from more fundamental differences in judgement of density is challenging.

To demonstrate the effect on the model of the label differences shown in figure 3 we train and validate purely on those images common to the pair but training separate models for each set of reader scores and then plot the predictions in figure 4. The top plots are direct comparisons of the predictions found by training and testing on the same images but with the different models trained with the individual labels from the pair of readers. The bottom row shows their prediction distributions. The model predictions and associated reader scores follow the same distribution. The pairs of models also show variation beyond just the distributional differences although substantially less than between the pairs of readers.

In previous work the relationship between the averaged VAS scores and cancer risk was shown to be high [3]. The key metric is the odds ratio between the highest and lowest density quintiles. The higher the value the stronger the relationship between the density predictions and risk of developing breast cancer. In figure 5 we show plots of these odds ratios when training on the individual and paired reader scores. Every model trained on each individual or paired set of scores show a significant relationship with cancer risk.

### Table 3. Metrics when trained and assessed on individual reader scores. We also show the approximate number of images the models are trained on. Av1 is the mean average across the reader metrics with standard deviation (in brackets) and Av2 is the average across all the images. The averages are substantially better than the overall metrics we find when we trained on the entire data-set (tables 1 and 3). All shows the global metrics trained on the entire data-set (Ind-Ind from table 1).

| Reader | Number | Rank Corr | RMSE |
|--------|--------|-----------|------|
| A      | 2263   | 0.808 ± 0.012 | 11.01 ± 0.30 |
| B      | 10821  | 0.817 ± 0.006 | 8.49 ± 0.17 |
| C      | 19262  | 0.815 ± 0.003 | 8.57 ± 0.09 |
| D      | 27829  | 0.806 ± 0.004 | 11.32 ± 0.09 |
| E      | 20022  | 0.796 ± 0.004 | 8.89 ± 0.09 |
| F      | 26239  | 0.798 ± 0.004 | 10.29 ± 0.12 |
| G      | 11469  | 0.866 ± 0.004 | 9.43 ± 0.11 |
| H      | 2606   | 0.816 ± 0.009 | 12.83 ± 0.28 |
| I      | 4467   | 0.832 ± 0.008 | 8.13 ± 0.20 |
| J      | 8915   | 0.771 ± 0.011 | 10.54 ± 0.24 |
| K      | 3249   | 0.856 ± 0.007 | 6.11 ± 0.18 |
| L      | 14137  | 0.834 ± 0.004 | 10.88 ± 0.11 |
| M      | 4230   | 0.786 ± 0.008 | 8.37 ± 0.15 |
| Av1    | 155,509| 0.815 ± 0.026 | 9.60 ± 1.69 |
| Av2    | 155,509| 0.844 ± 0.002 | 9.90 ± 0.04 |
| All    | 77,027 | 0.751 ± 0.002 | 11.63 ± 0.05 |

### Table 4. Metrics for pairs of readers when trained and assessed on the averaged scores of only those pairs. Av1 is the average across the reader metrics and Av2 is the average across all the images. All shows the metrics when training and assessing on only those images included in the paired data set. Only two of the subsets perform worse than the original model on all the data. The averages of the subset models show significant improvement over the original models.

| Number | Rank Corr | RMSE |
|--------|-----------|------|
| B_E    | 3250      | 0.818 ± 0.009 | 8.18 ± 0.22 |
| B_F    | 2655      | 0.843 ± 0.012 | 8.23 ± 0.29 |
| C_D    | 9764      | 0.86 ± 0.004  | 8.11 ± 0.10 |
| D_E    | 2581      | 0.848 ± 0.008 | 8.60 ± 0.19 |
| D_F    | 4087      | 0.85 ± 0.007  | 8.98 ± 0.18 |
| D_J    | 2379      | 0.819 ± 0.01  | 9.82 ± 0.26 |
| D_L    | 2245      | 0.863 ± 0.008 | 9.33 ± 0.22 |
| E_F    | 4088      | 0.841 ± 0.007 | 7.25 ± 0.15 |
| E_G    | 2523      | 0.864 ± 0.007 | 8.19 ± 0.20 |
| E_L    | 2202      | 0.845 ± 0.009 | 8.37 ± 0.20 |
| F_G    | 2719      | 0.861 ± 0.007 | 8.35 ± 0.22 |
| F_L    | 5254      | 0.864 ± 0.005 | 8.56 ± 0.15 |
| Av1    | 43747     | 0.848 ± 0.015 | 8.50 ± 0.63 |
| Av2    | 43747     | 0.846 ± 0.002 | 8.42 ± 0.06 |
| All    | 43749     | 0.834 ± 0.002 | 9.19 ± 0.05 |
training on the mammography data-set (labelled as Pre-trained). Uncertainties were found via bootstrapping with 1000 repeats and taking the 95% differences. The labels refer to the way we trained and tested the linear mapping: Av means we used the entire dataset with labels averaged across the pair of readers; for Ind 1 and Ind 2 we trained the linear mapping on individual reader scores and tested on the averaged scores. For Ind 1 we averaged the metrics across the readers and for Ind 2 we concatenated all the predictions together and calculated metrics from the total set.

The improvement by performing fine-tuning from the pre-trained model is substantial and statistically significant for both the multi-predictor and single-predictor models. This demonstrates that the representation formed by the pre-trained models on ImageNet can be significantly improved by adaptation to the mammography domain. However, we see little difference between training on averaged labels using the single-predictor or the individual labels using the multi-predictor. The implication of this is that the representation formed by both end-to-end models is similar, at least in their ability to produce good representations for a linear mapping to the labels.

In figure 3 we demonstrate the general improvement in performance by finetuning on the pre-trained weights. For one of the readers, we take the three sets of model predictions and show against the reader VAS scores. We see generally a stronger relationship between the reader VAS scores and the predicted scores for the trained models.

To further investigate the similarities between the representations, in table 6 we show the differences in prediction for the single- and multi-predictors. As we have utilised the same mapping function (a linear regression) between the representations in feature space of the two models the differences we are looking for will come only from the representation differences. These results imply the representation formed, with respect to the final density score, are extremely similar once we have trained the models on the mammography images.

In figure 7 we show the similarity in prediction between the three different models for one example set of reader scores. The left plot is the single predictor (S) against the pre-trained (Pre) model, the middle plot is the multi-predictor (M) against the pre-trained model and the right plot is the multi-predictor against the single-predictor. The similarity between the predictions made between the single and multi-predictor models is very high, much more so that between either of the trained models and the pre-trained model. This demonstrates that the single and multi-predictors become more similar to one another as they are trained.

In table 7 we show the dissimilarity between two runs of the multi and single predictor models.

\[ \text{Figure 3. Pairs of reader scores for two examples: D-J and B-E. The top plots show direct comparisons of the reader scores showing the large variation between the pairs of reader. The bottom plots show histograms of the pairs of reader scores which show the differences in distribution that the readers assess on the images.} \]
compared to their repeated runs - with the repeated runs having identical hyper-parameters sets. These results give empirical examples of the differences in predictions due to different training runs of the same model. The differences between different runs of the single predictor are actually larger than the differences between the single predictor and the multi-predictor (table 6) which is likely due to the high consistency we are seeing with the multi-predictor. It may be notable that the multi-predictor appears to be more consistent with repeats of the training.

In figure 8 we show the cancer odds ratios for the two models with predictions made via the individual models, labelled A–M, and the model trained on all the data (All). Similarly to the results in method 1, we see consistently good ability to make risk predictions.

Figure 4. Pairs of predicted scores from Method 1 for the same pair of examples as in figure 3. The models are trained separately on the individual scores of the pair. The top plots show direct comparisons of the predicted scores for the models trained on the different reader scores. The bottom plots show histograms of the pairs of predicted scores which show the differences in distribution due to the differences in labels from the readers.

Figure 5. Odds ratios between the highest density quintile and lowest density quintile for the models trained on individual scores (top) and paired average (bottom) as well as on the whole data-set. Higher values denote a stronger relationship between predicted density and cancer risk.
Breast density is an important risk factor for cancer risk and expert assessed scores have been shown to have a strong relationship with future breast cancer development. Accurate and consistent automated breast density prediction can be employed for masking risk estimation or cancer risk prediction models [18]. While the variability of reader labels has been investigated [14, 16], in this paper we showed the effect of this variability on deep learning based predictive models. Deep learning models can be considered to consist of two parts, one which forms the representation and a second part that performs the mapping to the final prediction. We investigated the effect of the label variability on the representation formed by the model and, separately, on the mapping from the representation to the prediction. We also considered the effect of reader density score differences on the quality of the models and on the ability to judge the quality of the models.

We showed that label variability has a considerable impact in terms of the assessment of the quality of the models. When we compared the predictions of the feature extraction linear regression models (Method 1) to averaged density scores we find better metrics than comparing to individual density scores. This demonstrates that the true quality of the models is higher than the assessment would suggest. Averaging the labels with the variability in the labels reduced, that our results improve considerably. The differences between the single and multi predictions are very small while both show much larger differences to the pretrained model. When we compare our model performance to the averaged labels we get a much better assessment of the model quality - which is true for models trained on ImageNet alone but show minor differences in performance compared to one another.

| Labels | Model    | Rank Corr | RMSE   |
|--------|----------|-----------|--------|
| Av     | Pre-trained | 0.799 ± 0.006 | 9.88 ± 0.12 |
| Av     | Single    | 0.846 ± 0.006 | 8.63 ± 0.12 |
| Av     | Multi     | 0.850 ± 0.003 | 8.50 ± 0.12 |
| Ind 1  | Pre-trained | 0.786 ± 0.022 | 9.92 ± 0.50 |
| Ind 1  | Single    | 0.850 ± 0.018 | 8.47 ± 0.48 |
| Ind 1  | Multi     | 0.854 ± 0.018 | 8.30 ± 0.48 |
| Ind 2  | Pre-trained | 0.842 ± 0.004 | 10.33 ± 0.12 |
| Ind 2  | Single    | 0.884 ± 0.004 | 8.96 ± 0.13 |
| Ind 2  | Multi     | 0.888 ± 0.003 | 8.79 ± 0.14 |

4. Discussion

Breast density is an important risk factor for cancer risk and expert assessed scores have been shown to have a strong relationship with future breast cancer development. Accurate and consistent automated breast density prediction can be employed for masking risk estimation or cancer risk prediction models [24].

The mapping of the vector representations to the final prediction is affected considerably by the variability in reader scores. All mappings show good risk prediction scores independently but combining predictions together from models trained on different datasets would need to be done with considerable care due to the different distributions of predictions produced by the models.

Table 5. Metrics for the differences of the model types (pre-trained, single and multi) for the three different label comparisons considered (Av, Ind 1, Ind 2). Both the single-predictor and multi-predictor models show statistically significant improvements from the models trained on ImageNet alone but show minor differences in performance compared to one another.

Table 6. Metrics for the prediction differences between the pretrained (Pre), single-predictor (S) and multi-predictor (M) models. The differences between the single and multi predictions are very small while both show much larger differences to the pretrained model.

| Labels | Comparison | Rank Corr | RMSE   |
|--------|------------|-----------|--------|
| Av     | M_S        | 0.991 ± 0.001 | 1.98 ± 0.04 |
| Av     | M_Pre      | 0.931 ± 0.003 | 5.32 ± 0.10 |
| Av     | S_Pre      | 0.933 ± 0.003 | 5.17 ± 0.09 |
| Ind. 1 | M_S        | 0.984 ± 0.002 | 2.35 ± 0.13 |
| Ind. 1 | M_Pre      | 0.908 ± 0.011 | 5.66 ± 0.31 |
| Ind. 1 | S_Pre      | 0.906 ± 0.011 | 5.80 ± 0.31 |
| Ind. 2 | M_S        | 0.991 ± 0.001 | 2.31 ± 0.03 |
| Ind. 2 | M_Pre      | 0.943 ± 0.002 | 5.59 ± 0.07 |
| Ind. 2 | S_Pre      | 0.941 ± 0.002 | 5.76 ± 0.07 |
To investigate the effect on the representation we made a novel alteration to a standard deep network which enables us to train on the individual reader scores. The representation formed by this approach was compared to that of a deep network trained on the averaged labels by applying linear regression to map the representations to the labels.

Both of the models trained end-to-end on the mammograms do significantly better than the mappings applied to the representations formed by training on models only on ImageNet. The representations formed by the two different models produce similar final mapped predictions. In fact the differences between the predictions are comparable to the differences that occur due to performing multiple training runs of the same model.

The similarity in prediction between the single- and multi-predictors implies that the representations formed by the models may be either highly similar themselves or separated only by a linear mapping. This may suggest that there is a density signal in the mammograms that can be relatively easily mapped to a lower dimensional subspace even though the labels the models are trained on are highly variable.

The final linear layer of the single and multi-predictors is retrained on the same image-label pairs. This obscures exactly the similarity between the representation formed by the two models because the linear predictor is able to correct for some of the differences. But it is only a linear mapping and therefore the similarity in final predictions does imply that the representations formed by the two models are closer together than either is to the pretrained model. It could have been the case that the single and multi-predictor models would find substantially different subspaces to represent the data and that the similarity in predictions between the two models would be lower than for either to the original model trained on ImageNet.

Table 7. Metrics for the prediction differences between the repeated runs of the multi- (M) and single- (S) predictors individually. These metrics show the level of difference we might expect in the predictions due to the stochastic nature of training a deep-learning model of this type. These training differences are similar to those between the single and multi predictors.

| Labels | Comparison | Corr.  | RMSE   |
|--------|------------|--------|--------|
| Av     | S_S        | 0.989 ± 0.001 | 2.14 ± 0.04 |
| Av     | M_M        | 0.996 ± 0.000 | 1.30 ± 0.03 |
| Ind 1  | S_S        | 0.983 ± 0.002 | 2.51 ± 0.15 |
| Ind 1  | M_M        | 0.991 ± 0.001 | 1.63 ± 0.10 |
| Ind 2  | S_S        | 0.989 ± 0.001 | 2.47 ± 0.03 |
| Ind 2  | M_M        | 0.996 ± 0.001 | 1.59 ± 0.02 |

Figure 7. Plots of model predictions against other model predictions. Left) single-predictor (S) against pretrained (Pre), middle) multi-predictor (M) against pretrained, right) multi-predictor against single-predictor. These plots show how similar the single and multi predictors are to one another and the much greater difference they both have to the pretrained model.

Figure 8. Odds ratios for the two models. There are no significant differences between the models in terms of their ability to make risk predictions.

To investigate the effect on the representation we made a novel alteration to a standard deep network which enables us to train on the individual reader scores. The representation formed by this approach was compared to that of a deep network trained on the averaged labels by applying linear regression to map the representations to the labels.

Both of the models trained end-to-end on the mammograms do significantly better than the mappings applied to the representations formed by training on models only on ImageNet. The representations formed by the two different models produce similar final mapped predictions. In fact the differences between the predictions are comparable to the differences that occur due to performing multiple training runs of the same model.

The similarity in prediction between the single- and multi-predictors implies that the representations formed by the models may be either highly similar themselves or separated only by a linear mapping. This may suggest that there is a density signal in the mammograms that can be relatively easily mapped to a lower dimensional subspace even though the labels the models are trained on are highly variable.

The final linear layer of the single and multi-predictors is retrained on the same image-label pairs. This obscures exactly the similarity between the representation formed by the two models because the linear predictor is able to correct for some of the differences. But it is only a linear mapping and therefore the similarity in final predictions does imply that the two representations formed must only be different in a linear manner from one another. The rest of the differences between the two models is similar in size to the differences caused by repeated model training runs. Also the two trained models are far more similar to one another than to the pretrained model which does imply that the representations of the two models are closer together than either is to the pretrained model. It could have been the case that the single and multi-predictor models would find substantially different subspaces to represent the data and that the similarity in predictions between the two models would be lower than for either to the original model trained on ImageNet.
As our fine-tuned models both perform significantly better than ImageNet trained models, and produce very similar outputs to each other, this implies that the models must both be finding signal in the images which are not being significantly affected by label variability. This may have important implications for the deployment of these types of models.

Models trained to make predictions using different labelled data-sets may not be usable together due to the significant differences we see in the mapping of our representations to the final data. However, if the representations that are formed are consistent then it may be possible to transfer different trained models formed to produce representations. Mappings from representations do not require as much data to be effectively trained as the formation of the representation. It may be possible to train representation forming models on large data-sets and then deploy them to have their final mappings assessed on much smaller local data-sets.

In section 2.1 we detail three aspects of reader variability that all could affect the models. The first one, fundamental differences in assessment of images, clearly will have a substantial impact on the mapping from representation to density score and we would expect it to have an impact on the representation of the models as well but that is harder to quantify. The second aspect, different distributions, has a large effect on the mapping which can be seen clearly in figure 3 and 4. However, the effect on the representation formed seems to be small. The third aspect, the random errors/uncertainty likely has a larger effect on both quality of mapping and representation when there is smaller amounts of data and reduces as the data quantity rises.

5. Conclusions

In this paper we have shown that variability in reader breast density labels has a sizeable effect on the final mapping from representation space to density prediction. The consequences of this are that models trained on different data-sets cannot be meaningfully used together if the direct comparison between individuals assessed by the different models are required. However, the risk scores, when considered on their own without comparison to other model predictions, are all good.

The quality of the representation formed when deep learning models are trained on the variable data appear to be similar and not significantly affected by the variability in the labels. If the variability in the labels had a significant effect on the model representation we would expect to see significant differences in the outputs of the two models when we applied linear regression to them. However, we see very similar results, implying that the representations formed by both models are, in some sense, extremely similar. We may see less variability in repeated training runs of the multi-predictor approach which may imply it is more stable in training. The similarity in quality of the representation of the two trained deep models provides evidence that there is a strong density-related signal in the images which the models are easily able to learn.

We conclude that the deep learning representations formed by training on these types of labels are similar regardless of the method used. However, the final mapping from representation space has to be considered very carefully to ensure that false conclusions are not being drawn.

Acknowledgments

D. Gareth Evans, Elaine Harkness and Susan M Astley are supported by the National Institute for Health Research (NIHR) Manchester Biomedical Research Centre (IS-BRC-1215-20007). Steven Squires was supported by CRUK AI-informed screening (A29024). Steven Squires is now affiliated with the University of Exeter but all work for this paper was performed whilst he was working at the University of Manchester.

Data availability statement

No new data were created or analysed in this study.

Disclosures

The authors declare no conflicts of interest.

Ethical statement

Ethics approval for the PROCAS study was through the North Manchester Research Ethics Committee (09/H1008/81). Informed consent was obtained from all participants on entry to the PROCAS study.

ORCID iDs

Steven Squires @ https://orcid.org/0000-0003-3681-1342

References

[1] Boyd N F, Martin L J, Bronskill M, Yaffe M J, Duric N and Minkin S 2010 Breast tissue composition and susceptibility to breast cancer Journal of the National Cancer Institute 102 1224–37

[2] Hua C, Chew G, Britt K, Ingman W, Henderson M, Hopper J and Thompson E 2014 Mammographic density review on the current understanding of its association with breast cancer Breast Cancer Research and Treatment 144 479–502

[3] Astley S M et al 2018 A comparison of five methods of measuring mammographic density: a case–control study Breast Cancer Research 20 10

[4] Fonseca P, Mendoza J, Wainer J, Ferrer J, Pinto J, Guerrero J and Castaneda B 2015 Automatic Breast Density Classification Using a Convolutional Neural Network Architecture Search Procedure Medical Imaging 2015: Computer-Aided Diagnosis vol 9414 (International Society for Optics and Photonics) 941428
[5] Kallenberg M et al 2016 Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring IEEE Trans. Med. Imaging 35 1322–31
[6] Highnam R, Brady S M, Yaffe M J and Harvey J 2010 Robust breast composition measurement volpara tm International Workshop on Digital Mammography (Springer) pp 342–9
[7] Keller B M, Nathan D L, Wang Y, Zheng Y, Gee J C, Conant E F and Kontos D 2012 Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation Med. Phys. 39 4903–17
[8] Ionescu G V, Fergie M, Berks M, Harkness E F, Hulleman J, Brentnall A R, Cuzick J, Evans D G and Astley S M 2019 Prediction of reader estimates of mammographic density using convolutional neural networks Journal of Medical Imaging 6 031405
[9] Squires S, Ionescu G, Harkness E F, Mackenzie A, Evans D G, Maxwell A, Howell S and Astley S M 2020 Automatic Density Prediction in Low Dose Mammography 15th International Workshop on Breast Imaging (IWBI2020) vol 11513 (International Society for Optics and Photonics) 115131D
[10] Squires S, Harkness E, Evans D G and Astley S M 2022 Automatic assessment of mammographic density using a deep transfer learning method Journal of Medical Imaging 10 024502
[11] Alsheh Ali M, Czene K, Hall P and Humphreys K 2019 Association of microcalcification clusters with short-term invasive breast cancer risk and breast cancer risk factors Sci. Rep. 9 1–8
[12] Castelvecchi D 2016 Can we open the black box of ai? Nature News 538 20
[13] Amann J, Blasimme A, Vayena E, Frey D and Madai V I 2020 Explainability for artificial intelligence in healthcare: a multidisciplinary perspective BMC Med. Inf. Decis. Making 20 1–9
[14] Sergeant J C et al 2013 Same Task, Same Observers, Different Values: the Problem with Visual Assessment of Breast Density Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment vol 8673 (SPIE) pp 197–204
[15] Sprague B L et al 2016 Variation in mammographic breast density assessments among radiologists in clinical practice: a multicenter observational study Annals of internal medicine 165 457–64
[16] Sperrin M, Bardwell L, Sergeant J C, Astley S and Buchan I 2013 Correcting for rater bias in scores on a continuous scale, with application to breast density Statistics in Medicine 32 4666–78
[17] Mainprize J G, Alonzo-Proulx O, Jong R A and Yaffe M J 2016 Quantifying masking in clinical mammograms via local detectability of simulated lesions Med. Phys. 43 1249–58
[18] Mainprize J G, Alonzo-Proulx O, Alshafeiy T I, Patrie J T, Harvey J A and Yaffe M J 2019 Prediction of cancer masking in screening mammography using density and textural features Academic Radiology 26 608–19
[19] Evans D G R et al 2012 Assessing individual breast cancer risk within the uk national health service breast screening program: a new paradigm for cancer prevention Cancer Prevention Research 5 943–51
[20] Huang G, Liu Z, Van Der Maaten L and Weinberger K Q 2017 Densely connected convolutional networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 4700–8
[21] Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L 2009 Imagenet: A large-scale hierarchical image database 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE pp 248–53
[22] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 770–8
[23] Kingma D P and Ba J 2014 Adam: A method for stochastic optimization arXiv:1412.6980
[24] Brentnall A R et al 2015 Mammographic density adds accuracy to both the tyrer-cuzick and gail breast cancer risk models in a prospective uk screening cohort Breast Cancer Research 17 147