Journal of Digital Imaging (2019) 32:746–760
https://doi.org/10.1007/s10278-018-00174-z

Can a Machine Learn from Radiologists' Visual Search Behaviour and Their Interpretation of Mammograms—a Deep-Learning Study

Suneeta Mall[1] · Patrick C. Brennan[1] · Claudia Mello-Thoms[1,2]

Published online: 13 August 2019

Abstract
Visual search behaviour and the interpretation of mammograms have been studied for errors in breast cancer detection. We aim to ascertain whether machine-learning models can learn about radiologists' attentional level and the interpretation of mammograms. We seek to determine whether these models are practical and feasible for use in training and teaching programmes. Eight radiologists of varying experience levels in reading mammograms reviewed 120 two-view digital mammography cases (59 cancers). Their search behaviour and decisions were captured using a head-mounted eye-tracking device and software allowing them to record their decisions. This information from radiologists was used to build an ensembled machine-learning model using top-down hierarchical deep convolution neural network. Separately, a model to determine type of missed cancer (search, perception or decision-making) was also built. Analysis and comparison of variants of these models using different convolution networks with and without transfer learning were also performed. Our ensembled deep-learning network architecture can be trained to learn about radiologists' attentional level and decisions. High accuracy (95%, p value $\cong 0$ [better than dumb/random model]) and high agreement between true and predicted values (kappa = 0.83) in such modelling can be achieved. Transfer learning techniques improve by < 10% with the performance of this model. We also show that spatial convolution neural networks are insufficient in determining the type of missed cancers. Ensembled hierarchical deep convolution machine-learning models are plausible in modelling radiologists' attentional level and their interpretation of mammograms. However, deep convolution networks fail to characterise the type of false-negative decisions.

Keywords Visual search · Breast cancer · Deep learning · Mammography · Machine learning

Introduction

Breast cancer is not only the most commonly diagnosed cancer but also the most common cause of death by cancer in women worldwide [1–3]. Mammography, the breast imaging modality of choice for cancer detection, due to its limitations [4], presents about 7–12% [5] of false positives and 4–34% [6] of false negatives. The reported sensitivity and specificity of mammography are 70–90% and 60–80%, respectively [7]. Use of adjunct imaging techniques still only improves sensitivity (85–93%) and specificity (70–85%) slightly [4, 7]. These statistics highlight the need to understand how radiologists interact with mammograms (*visual search*), how they arrive at their decisions (*cognition*), and what factors affect the errors in their interpretation of mammograms.

The radiologists' interaction with mammograms is conducted by means of the highly efficient human visual system, consisting of lens of varying resolution capabilities that is highest at the *fovea-centralis* (fovea) and decreases rapidly as one moves towards the *para-fovea* (periphery) [8]. This allows for inhomogeneous sampling of the scene, wherein different areas of the mammogram receive different levels of attention (detailed (foveal vision), less detailed (peripheral vision)). Visual search has been actively researched since 1962 [9], exploring perceptual aspects, such as search patterns [8, 10–12], search strategies [13], errors in interpretation [14–18], characteristics of regions that attracted radiologists' attention [19] and its effect on radiologists' decision [19–22] and satisfaction of search [23]. A global/local perception model [24] has been proposed describing the process of conducting the

✉ Suneeta Mall
smal5514@uni.sydney.edu.au

[1] Medical Image Optimisation and Perception Research Group (MIOPeG), Faculty of Medicine and Health, University of Sydney, 75 East Street, Lidcombe, NSW 2141, Australia

[2] Present address: Department of Radiology, University of Iowa, 200 Hawkins Drive, Iowa City, IA 52242, USA

inhomogeneous sampling of mammograms (identifying perturbations, gathering information through foveal and peripheral vision) and making a decision (reporting suspected cancer or absence of abnormalities). However, there has not been enough focus on how we can use the knowledge of attentional deployment and the interpretation of mammograms to improve the accuracy of mammographic assessments.

Direct application of this knowledge can be used in improving training and teaching platforms for radiologists. We hypothesize that the focus of training should go beyond understanding characteristics of malignancies and learning to differentiate cancer from other abnormalities (or normality). We believe that these training and teaching programmes should bring more transparency in how radiologists interact with mammograms—enabling radiologists to be more conscious about their attentional deployment and assisting them to develop an efficient visual search strategy from early on. It has been shown that the combination of visual search behaviour and characteristics of fixated mammographic regions can be used to determine the likelihood of whether a cancer will be missed and/or an erroneous decision will be made [25]. The deep convolution neural network (hereby referred as ConvNet), a specialized class of deep machine learning technique, used in Mall et al.'s [25] study is a very naïve implementation that builds a separate model for search behaviour and radiologists' decisions. It is important to evaluate whether ensembling individual models into a unified model is feasible and practical. Favorable outcome of such evaluations implies that training modules can be more customized to cater to radiologists' search behaviours and to focus not only on "where" errors were made but also on "why" errors occurred (e.g. missed during search (not fixated)). There is a possibility that the feedback of training modules can be more versatile, thus allowing for richer learning experience amidst radiologists.

False negatives (FN) are the "*white elephants*" of mammographic interpretation. It has been shown that nearly 25% of FNs are caused by radiologists not "*looking*" at the location of the cancer (i.e. missed during search, termed as "*search error*") [14]. Other 35% of FNs have been associated with radiologists "*looking*" at them but not for long enough (less than 1 s [26] termed as "*perception error*"), whereas the remaining errors of omission have been attributed to "*decision making errors*", i.e., incorrectly interpreting the finding or actively dismissing it [14]. This is another area where improvement can be imparted by building more effective training programmes by specifically focusing on areas that are likely to lead to FNs. The ability to determine if a FN will be search, perception or decision-making error will allow training programmes to specifically cater to these errors in teaching effective diagnosing techniques.

In this study, we focus on building, evaluating and comparing various machine-learning models (MLMs) that (a) simulate radiologist's attentional levels and decisions (hereby referred as iALD) and (b) determine sub-types of FN. Studies [11] have shown that radiologists visual search map (a.k.a. eye-scan path) are not reproducible, i.e., the same radiologist looking at the same case and in the same setting may follow different search maps, and despite having different search maps, radiologists can arrive at same decisions. For this reason, we are not focussing on modelling radiologists search map but only radiologist's attentional levels and decisions.

Materials and Methods

This study is a fully-crossed multi-reader, multi-case visual search study of digital mammography involving 120 two-view (craniocaudal [CC] and mediolateral oblique [MLO]) cases (59 cases depicting cancer, of which only 43 were visible in both views) obtained from a routine screening programme using a Selenia full-field digital mammography system (Hologic Inc., Marlborough, MA). Eight Mammography Quality Standards Act (MSQA)-certified radiologists of varying experience levels participated in the study.

Ground truth was established by a separate MSQA-certified breast radiologist, who did not participate as an observer in this study, using pathology reports and additional imaging. All cancer cases were biopsied, and all normal cases had a follow-up of 2 years.

Study Protocol

The radiologists wore a head-mounted eye-position tracking (ET) system (ASL Model H6, Applied Sciences Laboratory, Bedford, MA) that used an infrared beam (60 Hz temporal resolution) to calculate line of gaze by monitoring the pupil and the first corneal reflection. A magnetic head tracker was used to monitor head position, and this allowed the radiologists to freely move their heads from side to side as well as towards the displays, up to 20 cm, at which point they were outside the range of the head tracker. The ET integrates eye position and head position to calculate the intersection of the line of gaze and the display plane. The system has an accuracy (measured as the difference between true eye position and computed eye position) of less than 1° of visual angle, and it covers a visual range of 50° horizontally and 40° vertically.

The radiologists' workstation contained two calibrated 5 megapixel flat-panel portrait-mode displays (model C5i, Planar Systems Inc., Beaverton, OR), with a resolution of 2048 × 2560 pixels, typical brightness of 146 ftL and 3061 unique shades of grey. Radiologists were seated 60 cm from the workstation. Prior to the beginning of each reading session, a calibration of ET was performed wherein a 3 × 3 grid was shown on both the displays. After every five cases, the ET system was rechecked and, if necessary, it was recalibrated,

but this was only required twice at most during each reading session. After the calibration, the first (or next) case appeared on the displays wherein the left- and right-hand-side monitors would, respectively, display CC and MLO views of the case.

The eye tracker captured the visual search map (VSM), X and Y co-ordinates of fixation locations on ASL plane, dwell time, view and other details. Radiologists were advised to mark the location of perceived malignant lesions on the screen using a mouse-controlled cursor. Upon termination of search for a given case, the radiologists used a mouse-controlled cursor to click on a button in the display to select the next case of their reading sequence and were not allowed to come back to previously assessed cases.

In this study, each radiologist assessed all the 120 two-view cases in a different randomized order in two separate sessions that lasted about an hour each. Each radiologist completed their session (pertaining to this study) prior to their scheduled work shift. The radiologists reading environment was the same as their work reading setup—the only difference was that radiologists wore eye-tracking gear whilst participating in the study.

Definition of Attentional Levels

Based on the level of attention deployed on various mammographic regions, using radiologists' VSM, 3 types of areas, namely Foveal Areas (FA), Peripheral Areas (PA) and Never Fixated Areas (NFA), were extracted from the mammograms.

Foveal Areas Foveal areas (FAs) are breast areas measuring 2.5° radial angle (about 160 pixels × 160 pixels) consisting of at least 3 temporally sequential fixations (Fig. 1). FA extraction algorithm involves performing the following: (1) elimination of fixation points that had dwell duration less than 100 ms [8, 26, 27]; (2) fixed radius nearest neighbour algorithm using K-dimensional (KD)-tree and Bounded Deformation (BD)-tree [28] to obtain all areas containing fixation points that are within 2.5° radial angle to each other; followed by (3) removal of the redundant areas; and (4) selection of only areas that contained at least 3 temporally sequential fixation points.

Peripheral Areas Peripheral areas (PAs) are the breast areas within 2.5° radial angle from the location where a decision was made by the radiologists, consisting of less than 3 temporally sequential fixations (Fig. 2). PA defines the area where radiologists "looked" (fixated) but did not dwell long enough (< 300 ms [8, 26, 27]) within 2.5° around the centre of the mark to allow for the formation of a fixation area. To extract PA, squares of 160 pixels around the location where the radiologists made a decision but had less than 3 temporally sequential fixation points were automatically extracted from the image. These areas were retrospectively checked to ensure that they contained at least one fixation point.

Never Fixated Areas Never fixated areas (NFAs) are the breast areas that did not receive any fixation by any of the 8 radiologists (Fig. 3). NFAs were extracted by the following: (1) overlaying all 8 radiologists' VSMs on the cases; (2) identifying 2.5° radial angle areas per view per case that did not receive any fixation by any of the radiologists and extracting centre coordinates for these areas; and (3) automatically extracting these areas from the image. Only 1 such area per view per case was obtained, totaling about 240 NFAs.

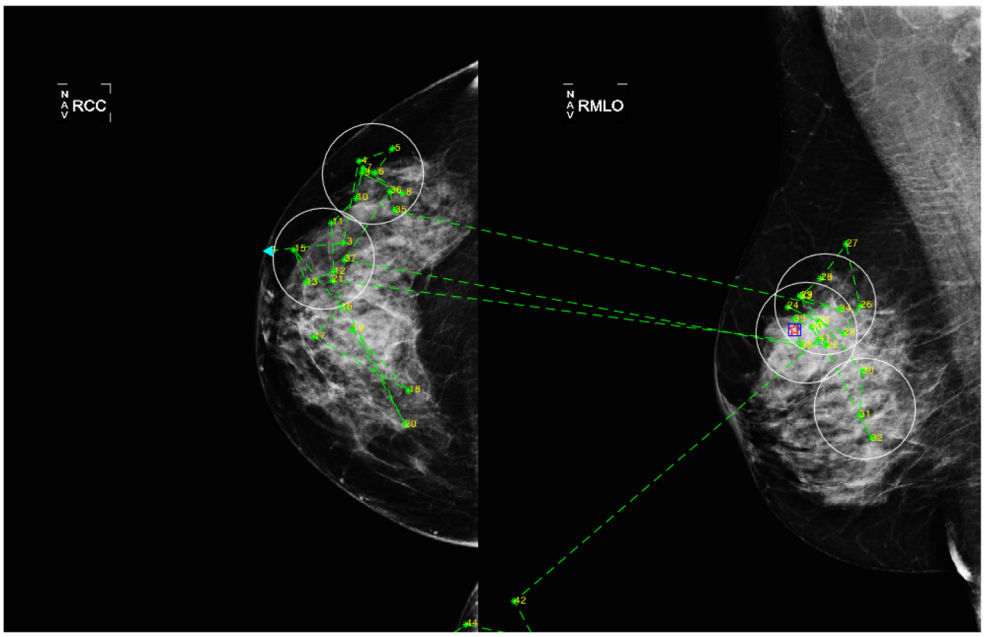

Fig. 1 Foveal areas (FAs) are the breast areas measuring 2.5° radial angle consisting of at least 3 temporally sequential fixations. These are highlighted with white circles. Red star indicates true malignancy, and blue square mark indicates location where a radiologist reported a malignant finding. Green points and dotted lines represent the temporal visual search behaviour (fixation points and the temporal sequencing amidst these points). The FAs (total 2) containing blue star in this figure on the right view have been classified as True Positive (TP) as the true cancer lies within the FA

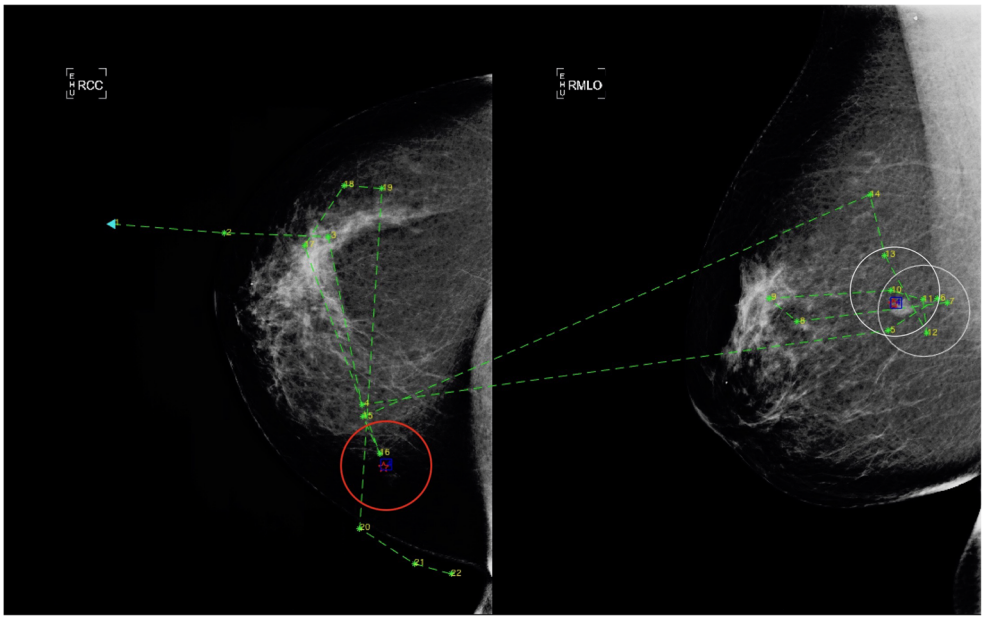

Fig. 2 Peripheral areas (PAs) are breast areas within 2.5° radial angle from a location where a decision was made by radiologists, consisting of less than 3 temporally sequential fixations. In this figure, the area shown in red circle is an example of PA. PA, in this example, is TP. For details of the figure annotations, please refer to Fig. 1 legend

Definition of Decision Outcome

Regions of mammogram were classified into the following 4 categories as follows:

1. True positive (TP): a marked region was classified as TP if it contained a true malignant lesion within 2.5° radius from the location of radiologist's mark.
2. False positive (FP): a marked region was classified as FP if it did not contain a true malignant lesion within 2.5° radius from the location of radiologist's mark.
3. True negative (TN): a fixated region was classified as TN if (a) it did not contain a true malignant lesion within 2.5° radius from the centre of fixation and (b) it was not marked as malignant by a radiologist.
4. False negative (FN): a region was classified as FN if it contained a true malignant lesion, but the radiologist failed to place a mark within 2.5° radius from the centre of the lesion.

Experiments

We conducted the study in two separate experiments—"*Modelling of Attentional Level and Decisions (iALD)*" and "*Modelling for Missed Cancers (MC)*". Both these

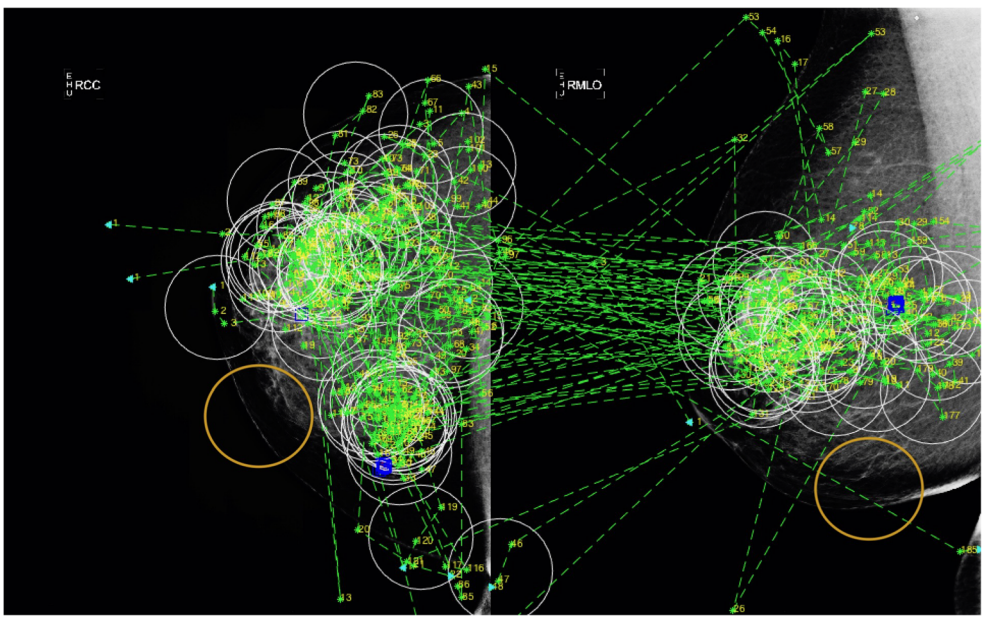

Fig. 3 Never fixated areas (NFAs) are breast areas that did not receive any fixations by any of the 8 radiologists. This figure overlays visual search behaviour of all radiologists for the case indicating areas that did not receive any attention by any of the radiologists. Example of the NFA is shown in orange circle. For details of the figure annotations, please refer to Fig. 1 legend

experiments present a specific deep network architecture that heavily utilizes ConvNet (Fig. 4). We chose ConvNet as core component of our network because it is a biologically inspired multilayer perceptron that simulates the visual cortex.

Out of several ConvNet architectures, we chose the following five contemporary networks for our analysis:

1. Residual Network (ResNet) 152 [29]
2. Inception ResNet V2 [30]
3. Inception V4 [30]
4. Neural Architecture Search Network (NASNet) [31]
5. Visual Geometry Group Network (VGGNet) 19 [32]

Our dataset is relatively smaller (details in the "Dataset collection" section) than more widely used datasets where ConvNet models have excelled (ex., ImageNet [30]). Transfer learning, a machine-learning reinforcement technique, is used to retrain a model (a.k.a. fine-tune) that was previously trained to perform a specific task T1, to perform a new task T2. It has been shown that models are able to use the knowledge they have gained to perform T1 into performing task T2 [33, 34]. This is especially true if there is an overlap in either the domain of the dataset and/or nature of task/activity (T1/T2). Negative learning, a side effect of transfer learning, wherein the retrained model does not learn enough to perform task T2 but instead gets confused and performs poorly (as compared to trained from scratch on dataset of T2) has also been reported [35].

To understand the effect of various ConvNet architectures and the effect of transfer learning on our network and its learning, we built, analysed and compared the networks in both of our experiments using each of these 5 ConvNet separately, with and without transfer learning leading us to 10 (=5 ∗ 2) training and evaluation exercises of each experiment.

Details of iALD and MC experiments and their networks are as follows:

1. iALD

The purpose of this experiment is to understand if an ensembled MLM can learn about the intricacies of visual attentional levels and if this model, with reasonably high accuracy, can predict specifics of interaction of radiologists with mammograms (i.e., the level of attention deployment on mammographic region) or accuracy of radiologists' decisions. We aim to ensemble this model in a hierarchical layered architecture (Fig. 5) and focus primarily on "local" aspect of image perception. The iALD model network is defined as top-down hierarchical classifier known as the "Local Classifier Per Node" [36] and is also a multi-label (multi-topic) classifier.

2. MC

The purpose of this experiment is to determine the nature of missed cancers and to be able to characterise "type" (search, perception, decision-making) of cancers that were missed. The network used for this model is shown in Fig. 6.

Each experiment entailed the following four steps:

Dataset collection

In total, we had 10,458 FA, 1196 PA and 240 NFA regions of mammogram which, based on the accuracy of decision outcome, were further classified into 10,106 TN, 224 TP, 147 FN and 1417 FP decision areas.

iALD All regions extracted from radiologists' VSM (10,458 FA, 1196 PA and 240 NFA) were used in modelling iALD.

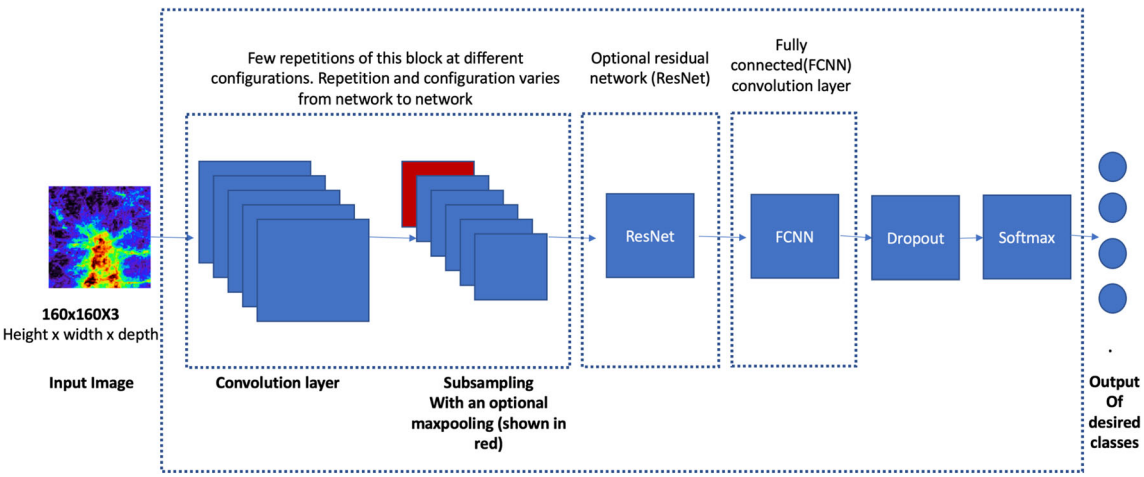

Fig. 4 Pictorial representation of high-level architecture of deep convolution neural network

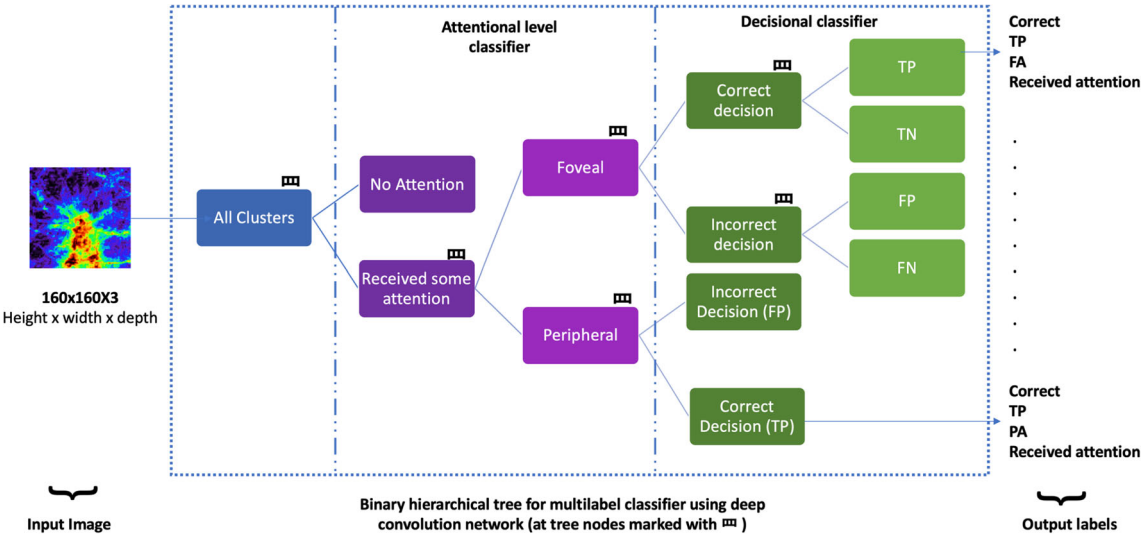

Fig. 5 Network architecture of iALD models. Nodes with symbol "" are the ConvNet nodes

We acknowledge that there is a class imbalance in our dataset. It has been shown that in convolution neural networks, oversampling does not lead to overfitting and the effect of oversampling or under-sampling is minimal, especially when the model is generalizing well to the dataset [37]. Given that our dataset is very small already, we opted to use the entire dataset instead of sub-sampling/down-sampling it.

MC All 147 FN regions were only perception (81) and decision-making errors (66) (as they were extracted from FA areas). Regions where search errors occurred were extracted separately by overlaying true cancer over VSM of each radiologist. True cancers that did not intersect with any FA were extracted programmatically. In total, 445 FN due to search error were observed. All 592 (147 + 445) FNs were used to model MC.

Data Pre-Processing

Adhering to the useful field of view (2.5° radial angle), as described in [19], all regions were 160 × 160-pixel grey-scale images. These images were, then, converted to obtain coloured images using the look-up-table (LUT) approach. Prior to colour conversion, histogram normalization was applied to avoid any loss of information. The results of normalization and colour conversion are shown in Fig. 7. This step was necessary because ConvNets are designed to work with natural images that have three channels [38].

Training and validation methodology

Training A subset of image augmentation techniques, namely distortion in colour (by changing the hue, contrast, saturation) and random rotation of the images, was applied to avoid overfitting and improve model performance. Random cropping was omitted to retain information of the useful field of view, which is essentially the size of the regional image (160 × 160 pixels).

Validation We used a k-fold cross validation (wherein K = 5) approach to validate MLMs. K = 5 was chosen to match the

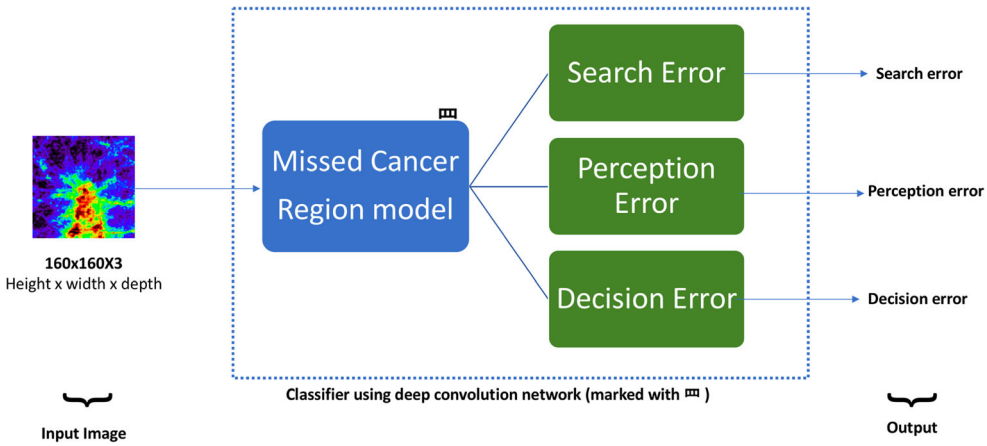

Fig. 6 Network architecture of MC models. Nodes with symbol "" are the ConvNet nodes

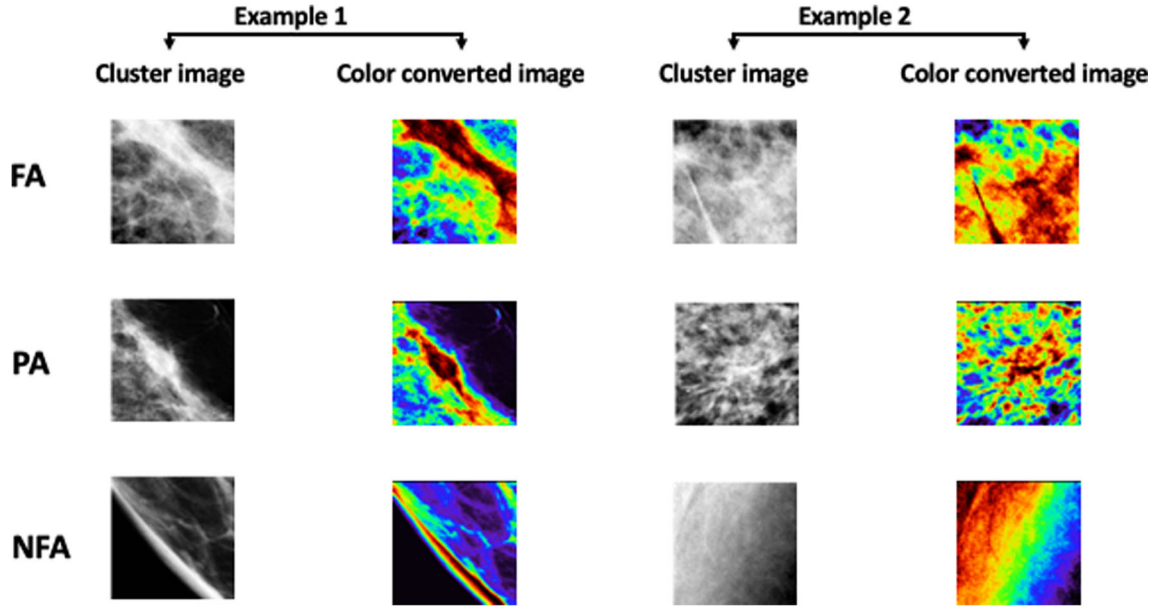

Fig. 7 Two examples of each attentional level area type (FA, PA and NFA) showing both the original area image obtained from grey-scale mammographic image and the colour-converted image (obtained by applying normalization and colour conversion using lookup table approach)

80–20% split of the training and validation samples. The final results were calculated based on the average prediction matrix of these 5-fold training/validation outcomes.

Random sampling approach was used to split the dataset into 5 (K) folds ensuring that the 80–20 split was retained across all classes. We did this to ensure that there was enough sample in the validation set for each class to conduct reasonable evaluation of the model across all categories.

Tensorflow and R-language software frameworks were used for modelling and analysis. Accelerated computing was also utilized with the aid of Graphical Processing Unit (GPU) NVIDIA GRID K520.

Evaluation of model

Each model was evaluated based on following criterion:

Model Performance Criterion The confusion matrix obtained from the model forms the basis of the evaluation. The averaged (of k-fold predictions) confusion matrix was analysed to obtain the following:

1. Per-category evaluation
   We analysed Sensitivity, Specificity, Positive (PPV) and Negative (NPV) Predictive Value, and Accuracy (i.e., (1 − *Misclassificatin Rate*)) of the model in predicting a specific category. These measures indicate how well the model learns and categorizes breast areas for a given specific category.
2. Overall evaluation
   To evaluate the overall performance of the model, we analysed accuracy (i.e., (1 − *Misclassificatin Rate*)) and confidence interval (95% CI) of accuracy. We also compared the model against null accuracy (a.k.a. no information rate, defined as the accuracy when the model is dumb and predicts only one class [arguably with highest amount of data in the dataset]) using hypothesis testing with H1: *accuracy of current model is better than "dumb model"*. To look at the agreement between truth and predicted class, we performed Cohen's kappa analysis and conducted McNemar's test.

   We also analysed micro- and macro-precision [39], recall [39] and F-score [39] for our multi-class classifier models (part of iALD and MC). Exact match ratio (a measure that allows the calculation of accuracy in correct labelling for multi-label/multi-topic classifier) [39] and Hamming loss (a measure to calculate error in labelling for multi-label/multi-topic classifier) [39] were calculated to evaluate iALD model for effectiveness in multi-labelling (a.k.a. multi-topic, for attentional level and decision outcome). Precision [39], recall [39] and F-score [39] of each sub-layer of iALD model were also evaluated to measure performance (hierarchical model evaluation).

Bias and Variance Analysis Bias and variance of iALD and MC model were also calculated using the misclassification rates, i.e., the error estimates of their respective k-iteration (of k-fold) training. The bias is defined as:

$$Bias = \frac{\sum_{i=1}^{i=k} (misclassification\ rate)_i}{k}$$

The variance in the error estimates is defined as:

$$Variance = \frac{\sum_{i=1}^{i=k} \left[(misclassification\ rate)_i - Bias \right]^2}{k}$$

Results

iALD

As shown in Fig. 5, the iALD model is logically divided into two layers—(1) attentional level and (2) decision. For simplicity of data presentation, the corresponding binary levels were merged into their respective logical layers (levels 1 and 2 into attentional level, and levels 3 and 4 into decision layer).

The null accuracy for attentional level was 0.879, and prevalence for FA, PA and NFA was 0.8790, 0.10084 and 0.020168, respectively. Our results of evaluation of interim attentional level classifier are shown in Tables 1 and 2. The range of accuracy observed in training 10 different models of iALD was 0.88 to 0.96. Whilst training from scratch did not lead us to statistically significant results that were better than the dumb model, using transfer learning improved the efficiency of the models (of all 5 ConvNets) enough to be significantly better (p values ~ 0 to 0.03). Overall Inception ResNet and ResNet performed better than other ConvNets. ResNet with transfer learning led to very high agreement (kappa = 0.81) between the true and predicted values of attentional levels whereas Inception ResNet was at good agreement (kappa = 0.77). The second level, pertaining to decision classification, was based on the efficiency of attentional level, as this was the parent layer. This layer characterised the model network as multi-label/multi-topic. Performance of this layer is effective performance of the overall model (shown in Table 3). The iALD model results in accurately determining decision for FA and PA are shown in Tables 4 and 5, respectively.

The iALD model, similar to the attentional level model, performed statistically significantly better with transfer learning and showed no sign of negative transfer learning in any of the ConvNets. Inception ResNet and ResNet, however, were statistically significantly better than the "dumb model" both with and without transfer learning (Table 3). Performance of both Inception ResNet and ResNet, with transfer learning, was comparable (0.9459 and 0.9463, respectively). Both Inception ResNet and ResNet showed good agreement (0.65 and 0.66, respectively) without transfer learning, but agreement was improved to very good agreement (0.828 and 0.83, respectively) with transfer learning. Both these models (with transfer learning) performed well in accurately determining radiologists' attentional levels and decisions, as exact match ratios were observed to be 0.95 and 0.94, respectively, making Inception ResNet only marginally better at multi-labelling. The Hamming loss for both these models was relatively lower (0.01) than the other models (0.02) (Table 3).

We noted that Inception, NASNet and VGG without transfer learning did not learn much about the radiologists' attentional levels and radiologists' decisions, and they were mostly "dumb" for our dataset.

The maximum bias in iALD (attentional level) modelling across all 50 (=5 ConvNet ∗ 2 (with and without transfer learning ∗ 5 (k-fold)) training/validation was 0.12 (Inception, NASNet and VGG all without transfer learning), minimum was for ResNet (0.04) [with transfer learning] followed by Inception ResNet [with transfer learning] (0.05) (Table 6). The variance across k-fold training iteration of these two models was 2.80E−05 and 9.80E−06, respectively.

MC

Our results from MC modelling are shown in Tables 7 and 8. The observed null accuracy was 0.75 with prevalence for search, perception and decision-making errors set as 0.75, 0.13 and 0.11, respectively. We noted that whilst the observed accuracy of all the 10 MC models we trained ranged from 0.75

Table 1 Detailed results of iALD model evaluation in determining attentional level. Lists measures to evaluate the model performance across all classes (FA, PA, NFA). [a]Micro-precision, recall and F-score results in the same values (by equation)

	Use of transfer learning	Accuracy (95% confidence interval)	Is model better than "dumb model"? (p value)	Agreement between predicted and true values' kappa (McNemar's test [p value])	Micro-precision/ recall/F-score*	Macro-precision	Macro-recall	Macro F-score
Inception ResNet	No	0.87 (0.8598, 0.8871)	0.71	0.556 (< 2e−16)	0.87	0.57	0.68	0.62
	Yes	0.95 (0.9392, 0.9572)	≅ 0	0.77 (< 1.5e−03)	0.95	0.45	0.8	0.58
Inception	No	0.88 (0.867, 0.8938)	0.4	0.04 (< 2e−16)	0.88	0.72	0.35	0.47
	Yes	0.89 (0.8815, 0.9066)	0.01	0.23 (< 2.2e−16)	0.89	0.44	0.45	0.44
ResNet	No	0.88 (0.867, 0.8938)	0.66	0.56 (< 2e−16)	0.88	0.58	0.71	0.63
	Yes	0.96 (0.9464, 0.9632)	≅ 0	0.81 (< 2e−8)	0.96	0.44	0.88	0.58
NASNet	No	0.88 (0.8661, 0.8926)	0.47	0.03 (< 2e−16)	0.88	0.49	0.34	0.4
	Yes	0.89 (0.8797, 0.905)	0.02	0.21 (< 2e−16)	0.89	0.45	0.44	0.45
VGG	No	0.88 (0.8661, 0.8926)	0.47	0.01 (NA)	0.88	0.96	0.34	0.5
	Yes	0.89 (0.8789, 0.9042)	0.03	0.20 (< 2.2e−16)	0.89	0.47	0.43	0.45

Table 2 Detailed results of efficiency of iALD model evaluation in determining each class, i.e., FA, PA and NFA of attentional level. Lists per-class measures of the model performance

	Use of transfer learning	Type of attentional level	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Inception ResNet	No	FA	0.9	0.9	0.98	0.54
		PA	0.8	0.95	0.63	0.98
		NFA	0.33	0.93	0.1	0.99
	Yes	FA	0.97	0.88	0.98	0.81
		PA	0.81	9.97	0.77	0.98
		NFA	0.63	0.99	0.49	0.99
Inception	No	FA	1	0.02	0.88	1
		PA	0.01	1	0.6	0.9
		NFA	0.04	1	0.8	0.98
	Yes	FA	1	0.16	0.9	1
		PA	0.1	1	0.78	0.91
		NFA	0.25	1	0.8	0.98
ResNet	No	FA	0.9	0.9	0.99	0.54
		PA	0.81	0.94	0.62	0.98
		NFA	0.42	0.98	0.11	1
	Yes	FA	0.97	0.94	0.99	0.81
		PA	0.85	0.97	0.79	0.98
		NFA	0.81	0.98	0.53	1
NASNet	No	FA	1	0.02	0.88	1
		PA	0	1	0.25	0.9
		NFA	0.02	1	0.33	0.98
	Yes	FA	1	0.15	0.9	1
		PA	0.9	1	0.76	0.91
		NFA	0.02	1	0.79	0.98
VGG	No	FA	1	0	0.88	1
		PA	0	1	0.9	0.89
		NFA	0.02	1	1	1
	Yes	FA	1	0.15	0.9	1
		PA	0.9	1	0.72	0.91
		NFA	0.21	1	0.77	0.98

to 0.82, none of these models were statically significantly better than the "dumb" models (Table 7). ResNet with transfer learning was only marginally significantly (0.82, p value = 0.05) better than the "dumb" model (0.75) (Table 7). The agreement between predicted and true class of MC varied widely across models observing between poor agreement (kappa (VGG without transfer learning) = 0.13), moderate agreement (kappa (Inception ResNet with transfer learning) = 0.51, kappa (Inception with transfer learning) = 0.55, kappa (ResNet without transfer learning) = 0.40, kappa (ResNet with transfer learning) = 0.47) and kappa (NASNet with transfer learning) = 0.46) and fair agreement for other models (Table 7).

MC modelling experiment did not provide us any reasonable results. We also observed the effect of negative transfer learning in case of Inception (accuracy of 0.79 without and 0.75 with transfer learning) and Inception ResNet (accuracy of 0.813 without and 0.805 with transfer learning) models. For all other models, transfer learning improved the accuracy, but, arguably, the models were still random (p value > 0.05). The bias in the MC models across 5 (k-fold) iterations ranged from 0.18 to 0.25 (mean = 0.21, mode = 0.19) (Table 6). The variance on the other hand spanned from 0.0001 to 0.0053 with mean variance 0.00017 (Table 6).

Discussion

Efficient search of medical images is an acquired skill. Novice readers are more prone to make errors due to inefficient search strategies (e.g., search error) [14, 40] than their experienced peers, who tend to make more interpretation (perception) and

Table 3 Results of hierarchical iALD model's (both attentional and decision layers combined) performance in accurately determining radiologists' attentional levels and decisions

	Use of transfer learning	Accuracy (95% confidence interval)	Is model better than "dumb model"? (p value)	Agreement between predicted and true values' kappa (McNemar's test [p value])	Exact match ratio	Hamming loss
Inception ResNet	No	0.8648 (0.8503, 0.8784)	≅ 0	0.65 (2.2e−16)	0.86	0.02
	Yes	0.9459 (0.9361, 0.9545)	≅ 0	0.828 (NA)	0.95	0.01
Inception	No	0.8714 (0.8573, 0.8846)	0.13	0.54 (NA)	0.87	0.02
	Yes	0.884 (0.8705, 0.8966)	≅ 0	0.58 (2.2e−16)	0.88	0.02
ResNet	No	0.8665 (0.851, 0.88)	≅ 0	0.66 (2.2e−16)	0.87	0.02
	Yes	0.9463 (0.9365, 0.9549)	≅ 0	0.83 (8.7e−8)	0.94	0.01
NASNet	No	0.8702 (0.856, 0.8834)	0.18	0.53 (NA)	0.87	0.02
	Yes	0.8832 (0.8696, 0.8958)	0.002	0.57 (NA)	0.88	0.02
VGG	No	0.8698 (0.8556, 0.8831)	0.178	0.53 (NA)	0.87	0.02
	Yes	0.8824 (0.8687, 0.895)	0.002	0.58 (NA)	0.88	0.02

decision (cognition) mistakes [41]. Visual search behaviour can provide a useful source of feedback to inexperienced radiologists, allowing upskilling of their search strategies [42]. Training to build an efficient search strategy by applying vision and perception can be multifaceted. Whilst referencing others' search patterns is useful in learning [42], learning from one's own errors is instrumental as well. In other words, feedback is essential.

In this study, we focussed on evaluating whether a machine can be taught about radiologists' attentional levels and their interpretation of mammogram. The reasoning behind our pursuit is to explore the intricacies of search, perception and cognition errors, and, using this acquired knowledge, build customized training programmes that can provide versatile information leading to enhanced learning experience.

MLMs have previously been used in modelling mammographic interpretations by using hand-picked features of spatial frequency analysis using artificial neural network [43]. Temporal dynamics of search behaviour have also been modelled using the support vector machine [44] indicating search behaviors are distinguisher of expertise. We have shown that our iALD model of attentional level and decisions can, with significantly better accuracy and very good agreement between truth and predicted values (Inception ResNet with transfer learning, accuracy = 95%, p value ≅ 0, k = 0.83), determine multiple facets of search behaviour and perception (especially level of attention deployment and outcome of radiologists' decision). We have also shown that these models are not just random models that perform well by chance, but they are statistically significantly better than random models. This has been missing in the present literature, as accuracy (a.k.a. correct classification rate or (1—misclassification rate)) has been used to evaluate the models. Whilst accuracy is an important measure, it can be misleading by itself (e.g., sample of 1 cancer per 100 cases would lead to 99% accuracy). The other benefit of our model is reduced selection bias due to use of deep-learning technique that learns features from the "source" itself as compared to other models [43, 44] that used a preselected set of features, forcing the model to learn from limited information that we think is relevant.

We note that accuracy of iALD models is better as compared to individual models of attentional level and decisions. The accuracy of iALD attentional level model is ≅ 96% [30] whereas independent ConvNet attentional level model is 90% [30]. Similarly, the accuracy of radiologists' iALD decisions is ≅ 95% [30] whereas independent ConvNet decision model is 92% [30] [25]. We also noted that iALD learnt better in both positive and negative prediction of decisions across all decision types (TP, FP, TN, FN) as compared to individual decision models [25] as sensitivity, specificity, PPV and NPV were all higher than ones reported by Mall et al. [25]. As for attentional level prediction of iALD, we noted that positive prediction of FA and negative prediction of both PA and NFA was higher. We also show that ensembled models for radiologists' attentional levels and decisions are feasible and in agreement with previous modelling efforts [25, 43].

Training one iALD model, on GPU, requires several days of training (< 1 week) as we trained 6 models ensembled together for 5 folds. The computation complexity in iALD model, as expected, increases as compared to individual models (as per Mall et al.'s [25] study); however, the increment is proportional. MC model training was relatively faster requiring approximately a day's worth of computation.

MC

Our MC results are indicative of the complexity in determining the type of errors for missed cancers. Type of false

Table 4 Results of hierarchical iALD model in accurately classifying radiologists' decisions on items predicted as "FA" attentional class in penultimate layer

	Use of transfer learning		Sensitivity	Specificity	Positive predictive value	Negative predictive value	Precision	Recall	F-score
Inception ResNet	No	TP	0.52	0.98	0.24	0.99	0.68	0.93	0.78
		FP	0.82	0.99	0.66	0.99	0.96	0.93	0.94
		TN	0.99	0.76	0.92	0.97	0.99	0.99	0.99
		FN	0.77	1	0.76	0.99	0.93	0.93	0.93
	Yes	TP	0.54	1	0.68	0.99	0.83	0.93	0.88
		FP	0.88	0.99	0.83	1	0.98	0.94	0.96
		TN	0.99	0.9	0.98	0.98	0.99	0.99	0.99
		FN	0.71	1	0.85	0.99	0.93	1	0.96
Inception	No	TP	0.17	1	0.93	0.97	0.71	0.93	0.81
		FP	0.37	1	0.93	0.96	0.82	0.93	0.87
		TN	0.96	0.96	0.99	0.79	0.99	0.99	0.99
		FN	0.34	1	0.87	0.97	0.89	0.86	0.88
	Yes	TP	0.19	1	0.93	0.97	0.71	0.93	0.81
		FP	0.42	1	0.93	0.97	0.82	0.93	0.87
		TN	0.96	0.96	0.99	0.8	0.99	0.99	0.99
		FN	0.36	1	0.87	0.98	0.89	0.86	0.88
ResNet	No	TP	0.56	0.98	0.24	0.99	0.68	0.93	0.78
		FP	0.86	0.99	0.66	1	0.96	0.93	0.94
		TN	0.99	0.76	0.92	0.98	0.99	0.99	0.99
		FN	0.78	1	0.76	1	0.93	0.93	0.93
	Yes	TP	0.55	1	0.68	0.99	0.68	0.93	0.78
		FP	0.92	0.99	0.76	1	0.96	0.93	0.94
		TN	1	0.9	0.98	0.99	0.99	0.99	0.99
		FN	0.77	1	0.8	1	0.93	0.93	0.93
NASNet	No	TP	0.17	1	0.93	0.97	0.71	0.93	0.81
		FP	0.36	1	0.93	0.96	0.82	0.93	0.87
		TN	0.96	0.96	0.99	0.79	0.99	0.99	0.99
		FN	0.33	1	0.87	0.98	0.89	0.86	0.88
	Yes	TP	0.18	1	0.94	0.97	0.71	0.93	0.81
		FP	0.41	1	0.93	0.96	0.82	0.93	0.87
		TN	0.96	0.96	0.99	0.8	0.99	0.99	0.99
		FN	0.36	1	0.87	0.98	0.89	0.86	0.88
VGG	No	TP	0.16	1	0.94	0.97	0.71	0.93	0.81
		FP	0.37	1	0.93	0.96	0.82	0.93	0.87
		TN	0.96	0.95	0.99	0.79	0.99	0.99	0.99
		FN	0.32	1	0.87	0.98	0.89	0.86	0.88
	Yes	TP	0.19	1	0.94	0.97	0.71	0.93	0.81
		FP	0.42	1	0.93	0.96	0.82	0.93	0.87
		TN	0.96	0.96	0.99	0.78	0.99	0.99	0.99
		FN	0.36	1	0.87	0.98	0.89	0.86	0.88

negatives (search, perception and decision making) is a more deterministic categorization on retrospection (i.e., once recognized as MC). With iALD model, we have shown that occurrences of false negative (MC) due to perception and decision-making errors can be determined using the MLM models (Inception ResNet (with transfer learning) PPV = 0.85, NPV = 0.99, ResNet is similar too) (Table 2). Two factors, namely, (1) whether the cancer was looked at and if so (2) how long it was dwelled on, are used to deterministically categorize MC into search, perception and decision-making

Table 5 Results of hierarchical iALD model in accurately characterising radiologists' decisions on items predicted as "PA" attentional class in penultimate layer

	Use of transfer learning	Type of decisions in PA	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Precision	Recall	F-score
Inception ResNet	No	TP	0.23	1	0.73	0.97	0.82	0.96	0.88
		FP	0.79	0.98	0.78	0.98	0.99	0.97	0.98
	Yes	TP	0.47	0.99	0.67	0.99	1	1	1
		FP	0.86	0.99	0.84	0.99	1	1	1
Inception	No	TP	0.67	0.95	0.02	1	1	1	1
		FP	0.5	0.94	0	1	1	1	1
	Yes	TP	0.5	0.96	0.04	10.8	0.8	0.8	0.8
		FP	0.79	0.94	0.12	1	0.95	0.95	0.95
ResNet	No	TP	0.23	1	0.71	0.97	0.77	0.96	0.85
		FP	0.79	0.98	0.79	0.98	0.99	0.95	0.97
	Yes	TP	0.41	1	0.74	0.98	0.74	0.96	0.83
		FP	0.85	0.98	0.82	0.99	0.99	0.94	0.97
NASNet	No	TP	0	0.94	0	1	NA	NA	NA
		FP	0.33	0.94	0	1	1	1	1
	Yes	TP	0.57	0.96	0.04	1	NA	NA	NA
		FP	0.82	0.94	0.12	1	1	1	1
VGG	No	TP	0	0.96	0	1	NA	NA	NA
		FP	1	0.94	0	1	1	1	1
	Yes	TP	0.57	0.96	0.04	1	NA	NA	NA
		FP	0.77	0.94	0.11	1	1	1	1

Table 6 Results of bias and variance analysis of both radiologist's attentional levels and decision (iALD) and missed cancer (MC) models across k-fold training–validation iteration

	Model type	Use of transfer learning	Range of accuracy	Bias	Variance
Inception ResNet	iALD attentional level	No	(0.85, 0.88)	0.12	1.00E−04
		Yes	(0.91, 0.96)	0.05	2.80E−05
	Missed cancer	No	(0.80, 0.82)	0.2	5.60E−05
		Yes	(0.7, 0.92)	0.19	5.30E−03
Inception	iALD attentional level	No	(0.87, 0.88)	0.12	6..4E−7
		Yes	(0.88, 0.9)	0.11	1.50E−05
	Missed cancer	No	(0.76, 0.84)	0.21	7.30E−04
		Yes	(0.64, 0.79)	0.25	2.90E−03
ResNet	iALD attentional level	No	(0.86, 0.88)	0.12	4.80E−05
		Yes	(0.95, 0.96)	0.04	9.80E−06
	Missed cancer	No	(0.79, 0.84)	0.19	3.20E−04
		Yes	(0.78, 0.88)	0.18	1.20E−03
NASNet	iALD attentional level	No	(0.87, 0.88)	0.12	2.10E−07
		Yes	(0.88, 0.89)	0.11	1.60E−05
	Missed cancer	No	(0.75, 0.83)	0.22	7.80E−04
		Yes	(0.70, 0.88)	0.19	5.10E−03
VGG	iALD attentional level	No	(0.87, 0.88)	0.12	7.60E−08
		Yes	(0.88, 0.89)	0.11	1.20E−05
	Missed cancer	No	(74, 77)	0.24	1.00E−04
		Yes	(0.79, 0.83)	0.19	1.40E−04

Table 7 Results of MC models in determining types of false negatives across all classes (search, perception and decision). [a]Micro-precision, recall and F-score results in the same values (by equation)

	Use of transfer learning	Accuracy (95% confidence interval)	Is model better than "dumb model"? (*p* value)	Agreement between predicted and true values' kappa (McNemar's test [*p* value])	Micro-precision/ recall/F-score[a]	Macro-precision	Macro-recall	Macro F-score
Inception ResNet	No	0.81 (0.7314, 0.8793)	0.08	0.38 (< 4.3e−4)	0.81	0.68	0.49	0.57
	Yes	0.81 (0.722, 0.8722)	0.12	0.51 (0.89)	0.81	0.63	0.62	0.62
Inception	No	0.79 (0.7033, 0.858)	0.23	0.22 (< 1.5e−05)	0.79	0.82	0.43	0.57
	Yes	0.75 (0.6665, 0.98288)	0.55	0.36 (< 0.16)	0.75	0.59	0.52	0.56
ResNet	No	0.81 (0.7293, 0.8782)	0.08	0.40 (< 1.3e−3)	0.81	0.72	0.51	0.6
	Yes	0.82 (0.7409, 0.8863)	0.05	0.47 (< 0.03)	0.82	0.64	0.54	0.59
NASNet	No	0.78 (0.6941, 0.8507)	0.3	0.16 (NA)	0.78	0.92	0.41	0.56
	Yes	0.81 (0.7314, 0.8793)	0.08	0.45 (0.05)	0.81	0.66	0.56	0.61
VGG	No	0.76 (0.6756, 0.8362)	0.46	0.13 (NA)	0.76	0.7	0.4	0.5
	Yes	0.79 (0.7103, 0.858)	0.23	0.20 (0.02)	0.79	0.52	0.46	0.49

Table 8 Results of MC model evaluation in determining the type of MC errors for each type of false negative

	Use of transfer learning	Error type	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Inception ResNet	No	Search	1	0.37	0.83	1
		Perception	0.31	0.98	0.71	0.9
		Decision	0.15	0.98	0.5	0.9
	Yes	Search	0.91	0.68	0.9	0.71
		Perception	0.56	0.93	0.52	0.93
		Decision	0.38	0.95	0.45	0.93
Inception	No	Search	1	0.17	0.8	1
		Perception	0.12	1	0.1	0.88
		Decision	0.15	1	0.67	0.9
	Yes	Search	0.88	0.55	0.86	0.62
		Perception	0.31	0.87	0.28	0.9
		Decision	0.38	0.97	0.62	0.93
ResNet	No	Search	0.99	0.38	0.82	0.92
		Perception	0.25	0.98	0.67	0.89
		Decision	0.3	0.98	0.67	0.92
	Yes	Search	0.97	0.51	0.86	0.88
		Perception	0.5	0.96	0.67	0.93
		Decision	0.15	0.97	0.4	0.9
NASNet	No	Search	1	0.1	0.77	1
		Perception	0.06	1	1	0.87
		Decision	0.15	1	1	0.91
	Yes	Search	0.96	0.48	0.85	0.78
		Perception	0.56	0.95	0.64	0.93
		Decision	0.15	0.98	0.5	0.9
VGG	No	Search	0.98	0.1	0.77	0.6
		Perception	0.6	0.98	0.33	0.87
		Decision	0.15	1	1	0.9
	Yes	Search	0.98	0.52	0.86	0.88
		Perception	0.25	0.93	0.36	0.89
		Decision	0.15	0.97	0.4	0.9

errors. Dwell time on fixated area is the differentiator for perception and decision-making errors. It has been shown that dwell time statistically significantly affects attentional level and decision outcome [19], but this is more an attribute of the reader/radiologist than the mammographic area. We hypothesize that perhaps not enough can be said about dwell duration simply from a region of mammography. We infer this hypothesis because sensitivity and PPV in classifying missed cancers was reasonably high, but we fail to model "why" (the type of MC) (Table 4). It is also possible that we simply do not have enough data to teach the model about MCs or even that our model is not suitable for this problem space and that we need machine learning techniques that are more spatial–temporal like structured Recurrent Neural Network [45] or other variants of ConvNets [46].

Limitations

Given the high null accuracy of the iALD model, small size and high-class imbalance in our dataset, it is hard to ascertain how generalizable the model is (and how "well" it has learned). Deep learning techniques learn to perform reasonably okay with small-medium datasets (when assisted with transfer learning, as shown in our results); however, to generalize the model and excel in its understanding, very large (of the order of millions of samples) training data is warranted.

Radiologists read CC and MLO views of each case side by side. These are the views of the same breast albeit in different projections. Radiologists' search behaviour on these views was treated independently. We acknowledge that more suitable approach would have been to transform their search map on each view onto a single plane to more accurately determine areas that received attention. This is more complex as we need to retain both spatial and temporal integrity of search behaviour during such transformation. This will be considered in future work.

Conclusions

We have shown that top down multi-label hierarchical classifier of deep convolution neural network can be trained to learn about visual radiologists' attentional levels and decisions. Highest accuracy and agreement (between true and predicted values) in modelling such behaviour are achieved by using ConvNets variant Inception ResNet (accuracy = 0.95, p value = ≅ 0, kappa = 0.828, hamming loss = 0.01) and ResNet (accuracy = 0.95, p value = ≅ 0, kappa = 0.83, hamming loss = 0.01) by reinforcing the network with transfer learning techniques. We also showed that spatial ConvNets are insufficient in modelling to determine types of missed cancers. We theorize that more spatial–temporal variants of deep network architectures might be more suitable for this task.

Acknowledgements We would like to thank the radiologists that participated in our experiment.

Funding This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Abbreviations *CC,* Craniocaudal view; *MLO,* Mediolateral oblique view; *ET,* Eye tracking; *VSM,* Visual search map; *FA,* Foveal area; *PA,* Peripheral area; *NFA,* Never fixated area; *TP,* True positives; *FP,* False positives; *TN,* True negatives; *FN,* False negatives; *MLM,* Machine learning models; *ConvNet,* Deep convolution neural network; *ResNet,* Residual network; *NASNet,* Neural architecture search network; *VGGNet,* Visual geometry group network; *iALD,* (Eye) Attentional level and decision; *MC,* Missed cancer

References

1. AIHW: Cancer in Australia 2017," in Cancer series no. 101. Cat. No. CAN 100. Canberra: AIHW, 2017
2. (May, 2018). Australian Institute of Health and Welfare 2017. Australian Cancer Incidence and Mortality (ACIM) books: Breast Cancer. Available: https://www.aihw.gov.au/reports/cancer/acim-books
3. S. I. Ferlay J, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. (2014, 16/1/2015). GLOBOCAN 2012 v1.1, Cancer Incidence and Mortality Worldwide: IARC Cancer Base No. 11 [Internet]. Available: http://globocan.iarc.fr
4. S. Mall, S. Lewis, P. Brennan, J. Noakes, and C. Mello-Thoms, "The role of digital breast tomosynthesis in the breast assessment clinic: a review, " Journal of Medical Radiation Sciences, pp. n/a-n/a, 2017.
5. Nelson HD, Fu R, Cantor A, Pappas M, Daeges M, Humphrey L: Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. preventive services task force recommendation. Annals of Internal Medicine 164(4):244, 2016
6. Huynh PT, Jarolimek AM, Daye S: The false-negative mammogram. Radiographics 18(5):1137–1154, 1998
7. Alakhras M, Bourne R, Rickard M, Ng KH, Pietrzyk M, Brennan PC: Digital tomosynthesis: a new future for breast imaging? Clinical Radiology 68(5):e225–e236, 2013–May 2013
8. Kundel HL, Nodine CF, Toto L: Searching for lung nodules. The guidance of visual scanning," (in eng). Invest Radiol 26(9):777–781, Sep 1991
9. Tuddenham WJ: Visual search, image organization, and reader error in roentgen diagnosis. Radiology 78(5):694–704, 1962
10. Kundel HL, Lafollet PS: Visual search patterns and experience with radiological images. Radiology 103(3):523, 1972
11. Mello-Thoms C et al.: Different search patterns and similar decision outcomes: how can experts agree in the decisions they make when reading digital mammograms? In: Krupinski EA Ed.. Lecture Notes in Computer ScienceDigital Mammography, Proceedings, Vol. 5116, 2008, pp. 212–219
12. Krupinski EA: Visual scanning patterns of radiologists searching mammograms. Academic Radiology 3(2):137–144, Feb 1996
13. Kundel HL, Nodine CF, Conant EF, Weinstein SP: Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology 242(2):396–402, Feb 2007
14. Kundel HL, Nodine CF, Carmody D: Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. (in eng), Invest Radiol 13(3):175–181, May-Jun 1978
15. Nodine CF, Kundel HL: Using eye movements to study visual search and to improve tumor detection. Radiographics: a Review

Publication of the Radiological Society of North America, Inc. 7(6): 1241–1250, 1987–Nov 1987
16. Kundel HL, Nodine CF, Krupinski EA: Searching for lung nodules—visual dwell indicates locations of false-positive and false-negative decisions. Investigative Radiology 24(6):472–478, Jun 1989
17. Mello-Thoms C, Dunn S, Nodine CF, Kundel HL: Image structure and perceptual errors in mammogram reading: a pilot study. In: Krupinski EA Ed.. (Proceedings of the Society of Photo-Optical Instrumentation Engineers (Spie), no. 26)Medical Imaging 2000: Image Perception and Performance, Vol. 1, 2000, pp. 170–173
18. Nodine CF, Mello-Thoms C, Weinstein SP, Kundel HL, Toto LC: Do subtle breast cancers attract visual attention during initial impression? In: Krupinski EA Ed.. Ed. (Proceedings of the Society of Photo-Optical Instrumentation Engineers (Spie), no. 26)Medical Imaging 2000: Image Perception and Performance, Vol. 1, 2000, pp. 156–159
19. Mall S, Brennan P, Mello-Thoms C: Fixated and not fixated regions of mammograms: a higher-order statistical analysis of visual search behavior. Academic Radiology 24(4):442–455, 2017
20. Mello-Thoms C, Dunn S, Nodine CF, Kundel HL, Weinstein SP: The perception of breast cancer: what differentiates missed from reported cancers in mammography? Academic Radiology 9(9): 1004–1012, Sep 2002
21. Mello-Thoms C, Dunn SM, Nodine CF, Kundel HL: The perception of breast cancers—a spatial frequency analysis of what differentiates missed from reported cancers. Ieee Transactions on Medical Imaging 22(10):1297–1306, Oct 2003
22. Mello-Thoms C, Nodine CF, Kundel HL: Relating image based features to mammogram interpretation. In: Medical Imaging 2002 Conference, San Diego, CA, 2002, Vol. e4686, 2002, pp. 80–83
23. Berbaum KS et al.: The influence of clinical history on visual-search with single and multiple abnormalities. Investigative Radiology 28(3):191–201, Mar 1993
24. Samei E, Krupinski EA: The Handbook of Medical Image Perception and Techniques (no. Book, Whole). Cambridge: Cambridge University Press, 2010
25. Mall S, Brennan PC, Mello-Thoms C: A deep (learning) dive into visual search behaviour of breast radiologists. SPIE Medical Imaging 10577:11, 2018 SPIE
26. Hillstrom AP: Repetition effects in visual search," (in eng). Percept Psychophys 62(4):800–817, May 2000
27. Kok EM, Jarodzka H, de Bruin ABH, BinAmir HAN, Robben SGF, van Merriënboer JJG: Systematic viewing in radiology: seeing more, missing less? Advances in Health Sciences Education 21: 189–205, 07/16 2016
28. D. M Mount, S. Arya, S. E. Kemp, and G. Jefferis. (2015). Fast Nearest Neighbour Search (Wraps Arya and Mount's ANN: A Library for Approximate Nearest Neighbor Searching). Available: https://cran.r-project.org/web/packages/RANN/RANN.pdf and https://www.cs.umd.edu/~mount/ANN/
29. X. Z. Kaiming He, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," vol. arXiv:1512.03385, no. https://arxiv.org/abs/1512.03385, 2015.
30. C. Szegedy, Ioffe, S., Vanhoucke, V., "Inception-v4, Inception-resnet and the Impact of Residual Connections on Learning," vol. arXiv:1602.07261, no. https://arxiv.org/abs/1602.07261, 2016.
31. V. V. Barret Zoph, Jonathon Shlens, Quoc V. Le, "Learning Transferable Architectures for Scalable Image Recognition," vol. arXiv:1707.07012, no. https://arxiv.org/pdf/1707.07012.pdf.
32. A. Z. Karen Simonyan, "Very Deep Convolutional Networks for Large-Scale Image Recognition," vol. arXiv:1409.1556, no. https://arxiv.org/abs/1409.1556, 2014.
33. Arel I, Rose DC, Karnowski TP: Research frontier: deep machine learning—a new frontier in artificial intelligence research. Comp. Intell. Mag. 5(4):13–18, 2010
34. Greenspan H, Ginneken BV, Summers RM: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging 35(5):1153–1159, 2016
35. Pan SJ, Yang Q: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10):1345–1359, 2010
36. C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," Data Mining and Knowledge Discovery, journal article vol. 22, no. 1, pp. 31–72, January 01 2011.
37. A. M. Mateusz Buda, Maciej A. Mazurowski, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks," vol. https://arxiv.org/abs/1710.05381, no. arXiv:1710.05381, 2017.
38. Tsehay YK et al.: Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images. SPIE Medical Imaging 10134:11, 2017 SPIE
39. Sokolova M, Lapalme G: A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4):427–437, 2009/07/01/, 2009
40. Manning DJ, Ethell SC, Donovan T: Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. The British Journal of Radiology 77(915):231–235, 2004
41. Donovan T, Manning DJ: Successful reporting by non-medical practitioners such as radiographers, will always be task-specific and limited in scope. Radiography 12(1):7–12, 2006/02/01/, 2006
42. Litchfield D, Ball LJ, Donovan T, Manning DJ, Crawford T: Viewing another person's eye movements improves identification of pulmonary nodules in chest X-ray inspection. Journal of Experimental Psychology: Applied 16(3):251–262, 2010
43. Mello-Thoms C: Perception of breast cancer: eye-position analysis of mammogram interpretation. Academic Radiology 10(1):4–12, Jan 2003
44. Gandomkar Z, Tay K, Brennan PC, Mello-Thoms C: A Model Based on Temporal Dynamics of Fixations for Distinguishing Expert Radiologists' Scanpaths, Vol. 10136, 2017, pp. 1013606-1013606-9
45. A. R. Z. Ashesh Jain, Silvio Savarese, Ashutosh Saxena, "Structural-RNN: Deep Learning on Spatio-Temporal Graphs," *https://arxiv.org/abs/1511.05298,* vol. arXiv:1511.05298, 2016.
46. H. Y. Bing Yu, Zhanxing Zhu, "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting," *https://arxiv.org/abs/1709.04875,* vol. arXiv: 1709.04875, 2018.