Multi-view Multi-task Learning for Improving Autonomous Mammogram Diagnosis

Trent Kyono TMKYONO@UCLA.EDU
Department of Computer Science
University of California, Los Angeles, United States

Fiona J. Gilbert FJG28@MEDSCHL.CAM.AC.UK
School of Clinical Medicine
University of Cambridge
Cambridge, United Kingdom

Mihaela van der Schaar MIHAELA@EE.UCLA.EDU
University of Cambridge, Cambridge, United Kingdom
Alan Turing Institute, London, United Kingdom
University of California, Los Angeles, United States

Abstract

The number of women requiring screening and diagnostic mammography is increasing. The recent promise of machine learning on medical images have led to an influx of studies using deep learning for autonomous mammogram diagnosis. We present a novel multi-view multi-task (MVMT) convolutional neural network (CNN) trained to predict the radiological assessments known to be associated with cancer, such as breast density, conspicuity, etc., in addition to cancer diagnosis. We show on full-field mammograms that multi-task learning has three advantages: 1) learning refined feature representations associated with cancer improves the classification performance of the diagnosis task, 2) issuing radiological assessments provides an additional layer of model interpretability that a radiologist can use to debug and scrutinize the diagnoses provided by the CNN, and 3) improves the radiological workflow by providing automated annotation of radiological reports. Results obtained on a private dataset of over 7,000 patients show that our MVMT network attained an AUROC and AUPRC of 0.855 ± 0.021 and 0.646 ± 0.023, respectively, and improved on the performance of other state-of-the-art multi-view CNNs.

1. Introduction

Breast cancer is the most prevalent cancer diagnosed in women, with nearly one in eight women developing breast cancer at some point in their lifetime. With the inclusion of screening mammography into breast cancer prevention and detection, randomized clinical trials have shown a 30% reduction of breast cancer mortality in asymptomatic women (Duffy et al., 2002). The success of these early breast cancer screening programs have lead to an increase in the total number of annual mammography exams conducted to nearly 40 million in the US alone (Broeders et al., 2012). Many recent research efforts have been motivated by the increasing number of mammograms requiring reading, presenting an opportunity to automate and reduce the additional workload and responsibility placed on radiologists.

We present a novel multi-view (MV) multi-task (MT) network as shown in Fig. 1. The system is comprised of two components: 1) a CNN trained using multi-task learning, which learns the radiological assessments known to be associated with cancer, such as breast density, conspicuity, suspicion (analogous to BI-RADS score), etc., that will be referred to here as MT-CNN and 2) a MVMT classifier that takes as input the predictions of the MT-CNN to concatenate image views and determine a cancer prediction. For succinct notation, we will refer to the MVMT classifier as MVMT throughout this paper.

Technical Significance MVMT provides several contributions to the machine learning for mammography literature. MT learning is used to improve the diagnostic accuracy by *predicting the radiological and patient features known to be associated with cancer*. Additionally, we concatenate and fuse mammogram views for left and right breast and corresponding mediolateral oblique (MLO) and craniocaudal (CC) views over the trained MT outputs, which provides a reduced and refined feature space to improve classification performance. Similar to our work, Geras et al. (2017) and Akselrod-Ballin et al. (2016) use full-images but report area under the receiver operating characteristic curve (AUROC) lower than ours at 0.753 and 0.78, respectively, on private datasets. Additionally, Geras et al. (2017) is the only related work to use all four of a patient's mammogram views for prediction. We report on a private dataset one of the highest AUROC (without using any ROI), and show the sources of gain attributed to MT learning, using all four mammogram views for prediction, and test-time-augmentation (TTA).

Due to the large amount of data required for training large CNNs and the limited number of available datasets, many implementations require utilization of ROIs or segmentation masks for maximizing performance. Our proposed method will allow existing mammography datasets to be leveraged without requiring a trained expert to manually annotate tumor locations or ROI. State-of-the-art implementations, utilizing region proposal networks, sliding windows, or patch classifiers for mammogram diagnosis rely on radiologist labeled ROIs and often have the highest reported diagnostic performance (Akselrod-Ballin et al., 2016, 2017; Becker et al., 2016; Carneiro et al., 2017; Jadoon et al., 2017; Jiao et al., 2016; Hepsag et al., 2017; Kooi and Karssemeijer, 2017; Mohamed et al., 2018; Platania et al., 2017; Ribli et al., 2017; Samala et al., 2017; Shen, 2017; Teare et al., 2017). However, all of these works rely on a very scarce and costly commodity, i.e., a dataset with cancer locations identified. ROI-based approaches have disadvantages other than the limitation of available location-annotated datasets. First, high-level contextual features external to the ROI are not learned (Geras et al., 2017). Secondly, in high noise scenarios where breast density may hide a visible tumor, a radiologist considers macroscopic features (not captured in ROI based methods), such as asymmetry between breasts, to assist in malignancy diagnosis (Peart et al., 2017; Scutt et al., 2006). This work does not exploit any ROI annotations and *uses only full-field mammograms*, which we refer to as image-level classification.

Clinical Relevance The recent success of convolutional neural networks (CNNs) in computer vision tasks has resulted in an influx of publications and implementations applying CNNs to mammography. There are two primary objectives or themes in the existing literature applying deep learning to mammography. The first, which occupies the majority of the research share, is to assist the radiologists in making decisions through computer-aided detection (CAD) (Abbas, 2016; Akselrod-Ballin et al., 2016; Dheeba et al., 2014; Huynh

et al., 2016; Jiao et al., 2016; Kooi et al., 2016; Qiu et al., 2016; Samala et al., 2016). The second objective, which has recently gained popularity, involves training CNNs to diagnose a patient without radiologist reading (Akselrod-Ballin et al., 2017; Becker et al., 2016; Carneiro et al., 2017; Kooi and Karssemeijer, 2017; Mohamed et al., 2018; Ribli et al., 2017; Shen, 2017). In this paper, we focus on the latter task of *autonomous diagnosis*.

In regards to clinical relevance, MVMT provides two unique benefits. First, MVMT improves radiological workflow by providing automated annotation of radiological reports through the MT predictions. Confidence estimates for each prediction are provided to a radiologist, such that only the uncertain predictions would require a radiologist's assessment reducing the overall reading and report times. Second, MVMT was especially designed to explain its predictions; it issues not only cancer predictions, but also radiological assessments, such as the conspicuity, suspicion, breast density, etc., in a similar manner as a radiologist would make an assessment. This allows our approach to provide radiologists more interpretable predictions and estimates, thereby facilitating human-machine collaboration in mammography. For example, MVMT may predict a patient have cancer, but the multi-task annotations reflect a huge discrepancy in the presentation or sign of lesion which a trained radiologist could question and reexamine the report.

2. Methodology

2.1. Problem Formulation

In this section, MVMT is formalized according to the illustration in Fig. 1. The system performs two primary predictions: 1) MT-CNN feature extraction, and 2) MVMT diagnosis.

Let $\mathcal{X} = \mathcal{X}_s \times \mathcal{X}_m$, $\mathcal{X}_r$, and $\mathcal{Y}$ be three spaces, where $\mathcal{X}_s$ is the patients' non-imaging feature space (such as age), $\mathcal{X}_m$ is the patients' mammogram imaging feature space, $\mathcal{X}_r$ represents the radiologists interpreted mammogram features (such as breast density, conspicuity, etc.), and $\mathcal{Y}$ is the space of all possible diagnoses, that is $\mathcal{Y} = \{0, 1\}$, where 0 corresponds to normal and 1 corresponds to malignancy.

Given a patient, $x \in \mathcal{X}$, let a radiologist as a classifier be defined as a map, $R : \mathcal{X} \to \mathcal{X}_r \times \mathcal{Y}$, which takes as input a patient's non-imaging features, $x_s \in \mathcal{X}_s$, and mammograms, $x_m \in \mathcal{X}_m$. R provides as output the radiological annotation, $x_r \in \mathcal{X}_r$, and the patient's cancer outcome, $y_x \in \mathcal{Y}$. For patient x with mammogram views $x_m \in \mathcal{X}_m = \mathcal{X}_{m_1} \times \mathcal{X}_{m_2} \times \mathcal{X}_{m_3} \times \mathcal{X}_{m_4}$, let $\mathcal{X}_{m_i}$ represent a view from a patient's four mammogram views: mediolateral oblique (MLO) right and left, and craniocaudal (CC) right and left. Additionally, for each of x's mammogram views, $x_{m_i} \in \mathcal{X}_{m_i}$, the radiologist prediction for that i-th view is $x_{r_i} \in \mathcal{X}_{r_i}$, such that $x_r \in \mathcal{X}_r = \mathcal{X}_{r_1} \times \mathcal{X}_{r_2} \times \mathcal{X}_{r_3} \times \mathcal{X}_{r_4}$. The MT-CNN is defined by a map, $M : \mathcal{X}_{m_i} \to \mathcal{X}_{r_i} \times \mathcal{Y}_i$, where M takes as input one of a patient's mammogram views, $x_{m_i} \in \mathcal{X}_{m_i}$, and outputs the radiologist prediction for that view, $x_{r_i} \in \mathcal{X}_{r_i}$, and the patient's actual cancer outcome for that view, $y_{x_i} \in \mathcal{Y}_i$. MVMT is defined as a map, $C : \mathcal{X}_s \times \mathcal{X}_c \to \mathcal{Y}$, where C takes as input the patient's non-imaging features, $x_s \in \mathcal{X}_s$, and the MT-CNN predictions for each mammogram, $M(x_{m_1}) \times M(x_{m_2}) \times M(x_{m_3}) \times M(x_{m_4}) \in \mathcal{X}_c$. C outputs a diagnostic prediction of the actual cancer outcome, $y_x \in \mathcal{Y}$.

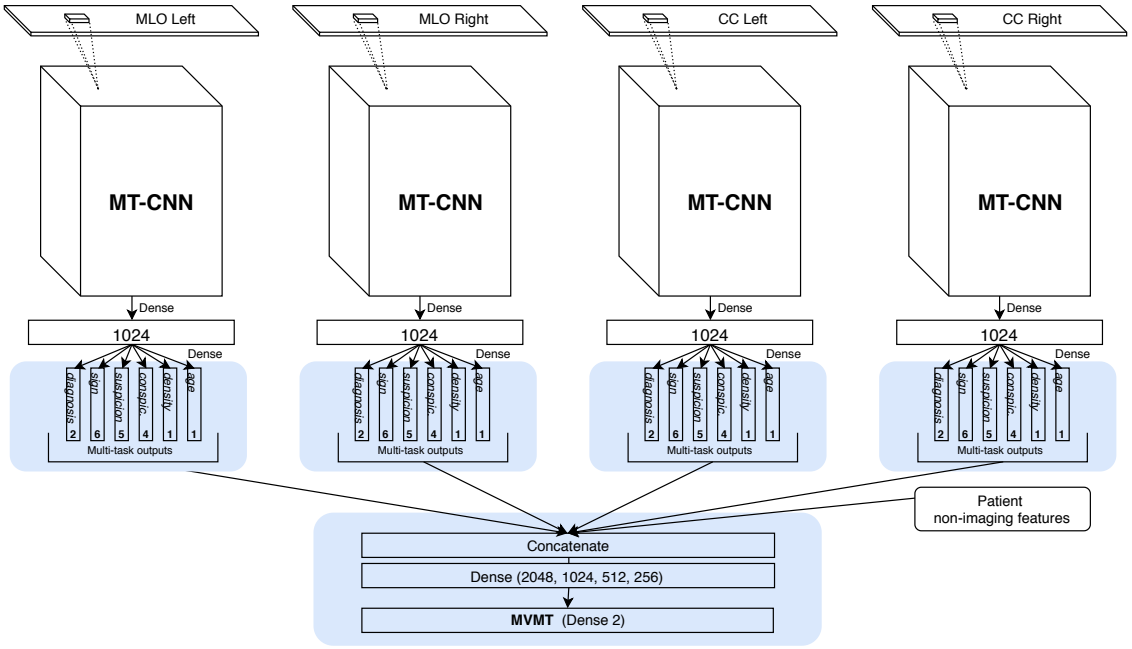

Figure 1: Full MVMT network as a stacked classifier of 4 MT-CNNs. Highlighted are the multi-task outputs of each view and how they are fused to generate the MVMT network. The mammogram conspicuity is abbreviated by conspic.

2.2. MVMT System

MVMT is a multi-view CNN designed to provide an accurate diagnosis given a patient's four mammogram views. MVMT is trained in two consecutive stages: 1) the MT-CNN over each image and 2) MVMT over a patient's four mammogram views. The neural network architecture for the the overall system is shown in Fig. 1.

The primary objective of the first training phase is to generate the MT-CNN that predicts both the diagnosis and radiological assessments on an individual mammogram view basis. The objective of our second training phase is to train a classifier that takes as input the multi-task outputs (MTO) of MT-CNN for each of a patients four mammogram views and predicts a patient-level diagnosis. In the current literature, this is done by combining multiple mammogram views at the dense layers before the final output layer (Carneiro et al., 2017; Geras et al., 2017). However, we choose to combine multiple views over the MTO for two reasons. First, the MTO are extracted imaging features that emulate radiological assessment and are what radiologists would naturally consider when reading multiple mammogram views. For example, breast density asymmetries between left and right breasts are often indicators of cancer (Scutt et al., 2006). Second, by pre-training MT-CNN in the first training phase, the MTO serves as a refined feature space for combining mammograms, requiring no retraining of layers prior to the MTO, and improves performance in scenarios where there is limited data.

MT learning is used to fine-tune MT-CNN providing several additional benefits. The work of Argyriou et al. (2006) demonstrates both empirically and theoretically the performance advantages of learning related tasks simultaneously over each task independently. This is amplified in situations when some tasks have very few data points and would be nearly impossible to learn individually. Additionally, MT learning is leveraged to learn refined feature representations and improve classification performance of the primary task (cancer diagnosis), by obligating MT-CNN to learn the radiological assessment known to be associated with cancer, such as the breast density, conspicuity, or suspicion. Finally, concatenating and fusing mammogram views for left and right breast, including corresponding MLO and CC views for each, over the trained MTO provides a reduced (and refined) feature space that improves classification performance, particularly in data-starved scenarios (Heitz et al., 2009). Concatenation of mammogram views could be performed at a subsequent dense layer (Sesmero et al., 2007), but these layers in practice are typically larger and thus require more training data. The sources of gain attributed to MT learning are shown experimentally in the results section.

3. Experiments

3.1. Dataset

The *Tommy* dataset was originally compiled to determine the efficacy and diagnostic performance of digital breast tomosynthesis (DBT) in comparison to digital mammography. The dataset was collected through six NHS Breast Screening Program (NHSBSP) centers throughout the UK and read by expert radiologists (Gilbert et al., 2015). It is a rich and well-labeled dataset with over 7,000 patients (over 1,000 malignant) who received diagnostic mammograms, and includes radiological assessments, density estimates ($\mu = 38.2$, $\sigma = 20.7$) along a 10-cm visual analogue scale (VAS), age at examination ($\mu = 56.5$, $\sigma = 8.75$), pathology outcomes from core biopsy or surgical excision, and both mammography and DBT imaging modalities. Although not all patients in the *Tommy* dataset underwent biopsy, each patient underwent expert radiological readings of both DBT and mammography modalities that significantly reduced the likelihood of false negative readings by as much as 15 to 30 percent (Gilbert et al., 2015). The *Tommy* dataset does not contain ROI annotations, but it does contain many useful radiological assessments that we leveraged for MT learning.

The *Tommy* dataset was designed to challenge the radiologist with overlapping breast tissue cases. In this dataset, it is estimated that roughly 50% of patients have overlapping tissues that show up on standard 2D mammograms that would falsely manifest as suspicious features (Gilbert et al., 2015). The patient criteria for selection were one of the following: 1) women recalled after routine breast screening between the ages of 47 and 73, or 2) women with a family history of breast cancer attending annual screening between ages of 40 and 49.

3.2. Mammogram Preprocessing and Augmentation

Mammogram processing steps were performed in several stages. Processed mammograms were converted from DICOM (Digital Imaging and Communication in Medicine) files into uncompressed 16-bit monochrome PNG (Portable Network Graphics) files. In this step, all

mammogram views were rotated and oriented with the breast along the left margin with nipple oriented to the right. Mammograms were not cropped, and Lanczos down-scaling was used to reduce the full-field mammograms to 960 x 1264 pixels to fit within our GPU memory. This maintained and preserved the width-to-height aspect ratio of 1 : 1.3 for all mammogram fields of view.

During training, mammograms were augmented to prevent over-fitting and promote model generalizability. Image augmentation was run through the *Keras* image processing generator with random selections from the following pool of augmentations: horizontal and vertical flips, image rotations of up to 20 degrees, image shear of up to 20%, image zoom of up to 20%, and width and height shifts of up to 20%. The gray-scale augmented mammograms were then stacked into 3 channels and histogram equalized by Contrasted Limited Adaptive Histogram Equalization (CLAHE) with channel stratified clipping and grid sizes as presented in Teare et al. (2017). We used the nominal grid sizes and clip limits they presented and enhanced their approach by using it as an additional augmentation. The CLAHE grid size, g, was augmented according to the following equation:

$$a \in \mathcal{U}(-log_2(k), log_2(k)) \mid g(k) = k + a, \tag{1}$$

where k is the nominal grid size. Similarly, the CLAHE clip limit, c, was augmented as follows:

$$a \in \mathcal{U}(-log_2(l), log_2(l)) \mid c(l) = l + a, \tag{2}$$

where l is the nominal clip limit. After histogram equalization, a Gaussian noise (Neelakantan et al., 2015) was applied to each color channel with a σ of 0.01, followed by image standardization. When training $MVMT$, each of the four input mammograms were augmented with a random set of augmentations drawn from the aforementioned pool of training augmentations.

4. Results

MVMT experiments were conducted on the *Tommy* dataset. 2000 randomly selected patients were reserved for 10-fold cross-validation. The remaining patients were randomly partitioned into a MT-CNN training set and a MVMT training set of 75% and 25%, respectively. Maintaining a separate training set for MVMT provided additional samples that the MT-CNN had never seen before to promote generalizability (Heitz et al., 2009; Sesmero et al., 2007).

4.1. MT-CNN Performance

The CNN used in this work was InceptionResNetV2 (Szegedy et al., 2016). Details for it's selection are presented in Appendix A. It was instantiated with ImageNet weights and refined using MT learning with the tasks shown in Table 1. The primary output target, *diagnosis*, was one of either malignant or benign (normal) as determined by the outcome of core biopsy. Five other auxiliary output targets were trained: *sign*, *suspicion*, *conspicuity*, *density*, and *age*. The *sign*, *suspicion*, and *conspicuity* were categorical output targets representing radiologist interpretation of the observed mammogram. Both patient *age* and breast *density* were included as auxiliary tasks for improved regularization and

for their known correlation with breast cancer (Lokate et al., 2013). Breast *density* was not categorized by the traditional BI-RADS lexicon, but by a percentage density calculated from a radiologist assessment on a 10-cm VAS (visual analogue scale) as described in Gilbert et al. (2015). For this reason, breast *density* was not learned as a categorical problem but as a regression, hence the normalization. Table 1 shows the classification and regression performance of each task. The results shown are the average of 100 test-time augmentations (TTA) per sample (Wang et al., 2018). See Appendix B for in depth training details.

Table 1: Multi-task performance for MT-CNN by task. * denotes regression targets, otherwise assume categorical. The performance metric used is AUROC unless specified as MAE (mean absolute error). For multi-class (non-binary) tasks, AUROC is calculated one-vs-all. Breast density is a percentage density calculated from a radiologist assessment on a 10-cm visual analog scale (VAS). Standard deviation is provided along with metric.

Task	Output	Performance metric
Diagnosis	malignant/benign	0.795 ± 0.015
Sign	none	0.720 ± 0.011
	circumscribed	0.701 ± 0.036
	spiculated	0.860 ± 0.036
	micro-calcification	0.621 ± 0.045
	distortion	0.771 ± 0.029
	asymm. density	0.641 ± 0.028
Suspicion	normal	0.672 ± 0.033
	benign	0.668 ± 0.035
	probably benign	0.657 ± 0.023
	suspicious	0.723 ± 0.018
	malignant	0.835 ± 0.014
Conspicuity	not visible	0.685 ± 0.030
	barely visible	0.748 ± 0.023
	visible, not clear	0.575 ± 0.041
	clearly visible	0.694 ± 0.017
Breast density*	0-100% VAS	13.96 ± 0.43 MAE(%)
Age*	Age 40 to 73	5.97 ± 0.17 MAE(yrs)

By providing our networks with various "perspectives" of the same mammogram, TTA mitigated the likelihood of misinterpreting a solitary sample and improved performance (Ayhan and Berens, 2018; Wang et al., 2018). Area under the receiver operating characteristic curve (AUROC) is reported for each categorical task; for regression targets mean absolute error is reported.

Table 2: Source of gain for MT-CNN and MVMT (multi-view MT-CNN) shown in terms of AUROC and AUPRC (± standard deviation) against the closest related works of Geras et al. (2017) and Zhang et al. (2018) on the *Tommy* dataset. Mammogram views (MV) is the number of input mammograms used per patient. MT is checked when multi-tasking was used, otherwise assume single-task. TA denotes if test-time augmentation was used.

Method	MV	MT	TA	AUROC	AUPRC
Zhang et al. (2018)	1			0.656 ± 0.013	0.318 ± 0.031
MT-CNN	1			0.745 ± 0.020	0.365 ± 0.025
MT-CNN	1	✓		0.752 ± 0.017	0.373 ± 0.024
MT-CNN	1		✓	0.791 ± 0.019	0.435 ± 0.030
MT-CNN	1	✓	✓	$\mathbf{0.795 \pm 0.015}$	$\mathbf{0.456 \pm 0.022}$
Geras et al. (2017)	4		✓	0.721 ± 0.024	0.425 ± 0.034
MVMT	4			0.793 ± 0.027	0.541 ± 0.029
MVMT	4	✓		0.824 ± 0.016	0.580 ± 0.028
MVMT	4		✓	0.837 ± 0.017	0.619 ± 0.021
MVMT	4	✓	✓	$\mathbf{0.855 \pm 0.021}$	$\mathbf{0.646 \pm 0.023}$

4.2. MVMT Performance

During testing the same augmentations used during training were applied and the final predictions were averaged over 100 sample iterations. Table 2 shows the sources of gain for MT-CNN and MVMT diagnostic performance. MV denotes the number of mammogram views used as input, which can be either a single view (1) or all views (4). MT is checked whenever multi-task learning was used. If MT is not checked, then the model was trained to only predict *diagnosis* with no auxiliary prediction tasks. TTA is checked whenever test-time augmentation was used (100 samples per patient). If TTA is not checked, then the AUROC and area under the precision-recall curve (AUPRC) were calculated over one sample prediction per patient. It is important to note that the reported AUROC values of 0.855 are relatively high compared to the existing state-of-the-art given the difficulty of the *Tommy* dataset. For comparison we show our proposed method in comparison to the closest image-level CNN works of Geras et al. (2017) and Zhang et al. (2018) on the *Tommy* dataset. We used the network and training methods provided in each respective publication, and conducted additional experimentation of these methods and MT-CNN on the public CBIS-DDSM dataset.

5. Discussion

Through discussion with radiologists, we have identified key requirements for MVMT acceptance and adoption in clinical practice, as well as the best options to integrate it into the radiological workflow for mammogram reading. Below, we outline two possible use cases of MVMT in practice. First, the MT output predictions of MVMT provide automated annotation of radiological reports as in Fig. 2. These reports are both time consuming and

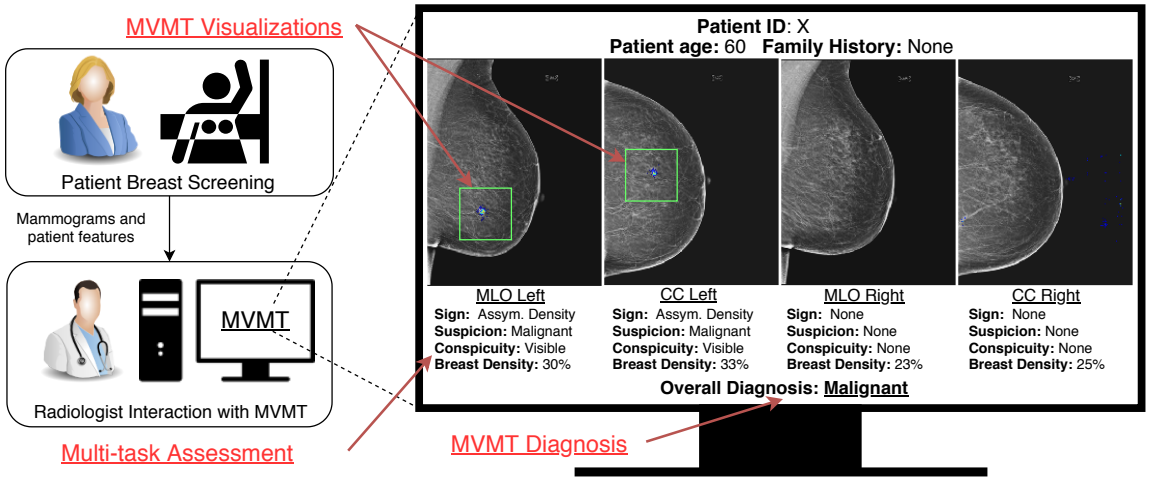

Figure 2: Example illustration of MVMT in radiological practice: 1) automated radiological reporting, and 2) MT outputs provide additional assessments for radiologists to scrutinize.

costly to generate. By providing confidence estimates for each prediction to a radiologist, the radiologist could choose to examine only the predicted radiological features that have a high degree of uncertainty reducing the overall reading and report times. Second, MVMT was designed to explain its predictions; it issues not only cancer predictions, but also radiological assessments such as the conspicuity, suspicion, breast density, etc., in a similar manner as a radiologist would make an assessment. This allows our approach to provide radiologists more interpretable predictions and estimates, thereby enabling better human-machine collaboration for mammography. MVMT provides interpretability not currently offered in the machine learning for breast cancer literature. The additional information shown in Fig. 2 can be used to debug and interpret MVMT predictions. Existing methods in machine learning for mammography provide visualizations, but do not have the ability to provide the multi-task annotations that MVMT is capable of. The multi-task outputs are extracted imaging features that emulate radiological assessment and are what a radiologist would naturally consider when examining multiple mammogram views. For example, breast density asymmetries between left and right breast are often indicators of cancer (Scutt et al., 2006). Details on our method for MVMT visualization are provided in Appendix C.

6. Related works

Networks that are trained with full images have been shown to improve diagnostic performance, but require more training data (Bojarski et al., 2017). In comparison to ROI-based methods, the training data does not require annotated locations that makes data acquisition a lot simpler, cheaper and scalable. The work of Geras et al. (2017) presented the richest mammography datasets used (with over 200,000 mammograms). Because of this, they are one of the few researchers who attempt an image-level approach utilizing all four mammogram views to predict the BI-RADS score. Our dataset is *significantly* smaller, but

we draw motivation from their work of using all four mammogram views without relying on any ROI. Results are compared to theirs for a benchmark comparison. Other related image-level and multi-view networks were presented in Akselrod-Ballin et al. (2016), Bekker et al. (2016), Carneiro et al. (2017, 2015), and Zhu and Xie (2016) and are shown in Table 3 for comparison.

MT learning has been successfully used on an ROI level in mammography (Kisilev et al., 2016; Samala et al., 2017), but our work is the first to apply MT learning to image-level mammogram classification. Bekker et al. (2016), Carneiro et al. (2015), Geras et al. (2017), and Yi et al. (2017) used multiple views for improving classification performance, however this work is the first to do so by concatenating the multi-task outputs of each mammogram view. Many early investigative works have shown the success of transfer learning using non-medical or natural images to classify mammograms (Argyriou et al., 2006). Specifically, these publications have shown performance gains from using models pre-trained with ImageNet weights, such as AlexNet, Inception or ResNet (Huynh et al., 2016; Lévy and Jain, 2016; Jiang et al., 2017; Samala et al., 2016; Yi et al., 2017).

The closest related works are presented in Table 3, and although it is difficult to draw a direct comparison to these works, we highlight the limitations of existing works in comparison to ours. MVMT has a reported AUROC of 0.855 and is higher than Geras et al. (2017) at 0.753, who predicted BI-RADS (0, 1, and 2). The works of Carneiro et al. (2017), Dhungel et al. (2014), and Zhu and Xie (2016) use the INbreast dataset to predict malignancy and have comparable AUROC ranging from 0.8 to 0.86. However, because of their small dataset size of 115 patients, the reported results could be subject to high variance. Additionally, this dataset does not have the challenging overlapping tissues present in the *Tommy* dataset.

Table 3: Comparison of related image-level and multi-view architectures. Bold represents the proposed method. * denotes μ AUROC. MGV represents the number of mammogram views used as input. ROI is Y if either ROI or segmentation masks were needed for training or inference, otherwise N. MT is Y if multitask learning used, otherwise N. MVMT is shown in bold.

Method	Dataset	Patients	MGV	ROI	MT	AUROC
Geras et al. (2017)	private	201698	4	N	N	0.753*
Akselrod-Ballin et al. (2016)	private	300	1	N	N	0.78 acc.
Carneiro et al. (2015)	DDSM	287	2	Y	N	0.91
Carneiro et al. (2017)	INbreast	115	2	Y	N	0.860
Bekker et al. (2016)	DDSM	172	2	Y	N	0.800
Dhungel et al. (2017)	INbreast	115	2	Y	N	0.800
MVMT	**Tommy**	**7060**	**4**	**N**	**Y**	**0.855**

6.1. Conclusion

We introduced a new machine learning classifier for breast cancer, called MVMT, that utilized MT learning to improve diagnostic accuracy on full image mammography in comparison to the state-of-the-art. We demonstrated on the challenging *Tommy* dataset how MT learning improved single image diagnostic accuracy in comparison to non-MT learning. Additionally, we demonstrated that concatenating images over the MT outputs led to a more refined feature representation that also resulted in increased multi-view accuracy. Lastly, we provided a brief example of how MT outputs can be leveraged for autonomous radiological report generation and improving machine learning interpretability. This method shows great promise for image-level diagnosis and with enough data may eventually surpass ROI methods.

References

Qaisar Abbas. Deepcad: A computer-aided diagnosis system for mammographic masses using deep invariant features. *Computers*, 5(4), 2016. ISSN 2073-431X. doi: 10.3390/computers5040028.

Ayelet Akselrod-Ballin, Leonid Karlinsky, Sharon Alpert, Sharbell Hasoul, Rami Ben-Ari, and Ella Barkan. A region based convolutional network for tumor detection and classification in breast mammography. In Gustavo Carneiro, Diana Mateus, Loïc Peter, Andrew Bradley, João Manuel R. S. Tavares, Vasileios Belagiannis, João Paulo Papa, Jacinto C. Nascimento, Marco Loog, Zhi Lu, Jaime S. Cardoso, and Julien Cornebise, editors, *Deep Learning and Data Labeling for Medical Applications*, pages 197–205, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46976-8.

Ayelet Akselrod-Ballin, Leonid. Karlinsky, Alon Hazan, Ran Bakalo, Ami Ben Horesh, Yoel Shoshan, and Ella Barkan. Deep learning for automatic detection of abnormal findings in breast mammography. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 321–329, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67558-9.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 41–48, Cambridge, MA, USA, 2006. MIT Press.

Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *International conference on Medical Imaging with Deep Learning*, 2018.

A. S. Becker, M Marcon, S Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss. Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investigate Radiology*, 52:434–440, July 2016. ISSN 7.

A. J. Bekker, H. Greenspan, and J. Goldberger. A multi-view deep learning architecture for classification of breast microcalcifications. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 726–730, April 2016. doi: 10.1109/ISBI.2016.7493369.

Mariusz Bojarski, Philip Yeres, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Lawrence D. Jackel, and Urs Muller. Explaining how a deep neural network trained with end-to-end learning steers a car. *CoRR*, abs/1704.07911, 2017.

Mireille Broeders, Sue Moss, Lennarth Nystrm, Sisse Njor, Hkan Jonsson, Ellen Paap, Nathalie Massat, Stephen Duffy, Elsebeth Lynge, and Eugenio Paci. The impact of mammographic screening on breast cancer mortality in europe: A review of observational studies. *Journal of Medical Screening*, 19(1_suppl):14–25, 2012.

Gustavo Carneiro, Jacinto Nascimento, and Andrew P. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 652–660, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Gustavo Carneiro, Jacinto Nascimento, and Andrew P. Bradley. Chapter 14 - deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions1. In S. Kevin Zhou, Hayit Greenspan, and Dinggang Shen, editors, *Deep Learning for Medical Image Analysis*, pages 321 – 339. Academic Press, 2017. ISBN 978-0-12-810408-8. doi: https://doi.org/10.1016/B978-0-12-810408-8.00019-5.

J. Dheeba, N. Albert Singh, and S. Tamil Selvi. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *Journal of Biomedical Informatics*, 49:45 – 52, 2014. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2014.01.010.

N. Dhungel, G. Carneiro, and A. P. Bradley. Automated mass detection in mammograms using cascaded deep learning and random forests. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, Nov 2015. doi: 10.1109/DICTA.2015.7371234.

N. Dhungel, G. Carneiro, and A. P. Bradley. Fully automated classification of mammograms using deep residual neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 310–314, April 2017. doi: 10.1109/ISBI.2017.7950526.

Neeraj Dhungel, Gustavo Carneiro, and Andrew P. Bradley. Deep structured learning for mass segmentation from mammograms. *CoRR*, abs/1410.7454, 2014.

Stephen W. Duffy, Laszlo Tabr, HsiuHsi Chen, Marit Holmqvist, MingFang Yen, Shahim Abdsalah, Birgitta Epstein, Ewa Frodis, Ljungberg Eva, Christina HedborgMelander, Ann Sundbom, Maria Tholin, Mika Wiege, Anders kerlund, HuiMin Wu, TaoShin Tung,

YuehHsia Chiu, u ChenPu Chi, ChihChung Huang, Robert A. Smith, Mns Rosn, Magnus Stenbeck, and Lars Holmberg. The impact of organized mammography service screening on breast carcinoma mortality in seven swedish counties. *Cancer*, 95(3):458–469, 2002. doi: 10.1002/cncr.10765.

Krzysztof J. Geras, Stacey Wolfson, S. Gene Kim, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *CoRR*, abs/1703.07047, 2017.

Fiona Gilbert, Lorraine Tucker, Maureen GC Gillan, Paula Willsher, Julie Cooke, Karen Duncan, Michael Michell, Hilary Dobson, Yit Yoong Lim, Hema Purushothaman, Celia Strudley, Susan M Astley, Oliver Morrish, Kenneth Young, and Stephen Duffy. The tommy trial: a comparison of tomosynthesis with mammography in the uknhs breast screening program. *Health Technology Assessment*, 19, 2015.

Nima Habibzadeh Motlagh, Mahboobeh Jannesary, HamidReza Aboulkheyr, Pegah Khosravi, Olivier Elemento, Mehdi Totonchi, and Iman Hajirasouliha. Breast cancer histopathological image classification: A deep learning approach. *bioRxiv*, 2018. doi: 10.1101/242818.

Michael Heath, Kevin Bower, Richard Moore, and W. Phillip Kegelmeyer. The digital database for screening mammography. In *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212–218. Medical Physics Publishing, 2001. ISBN 1-930524-00-5.

Geremy Heitz, Stephen Gould, Ashutosh Saxena, and Daphne Koller. Cascaded classification models: Combining models for holistic scene understanding. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, NIPS, pages 641–648, 2009.

P. U. Hepsag, S. A. Ozel, and A. Yazici. Using deep learning for mammography classification. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 418–423, Oct 2017. doi: 10.1109/UBMK.2017.8093429.

Benjamin Q. Huynh, Hui Li, and Maryellen L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3:3 – 3 – 5, 2016.

M. Jadoon, Qianni Zhang, Ihsan Ul Haq, Sharjeel Butt, and Adeel Jadoon. Three-class mammogram classification based on descriptive cnn features. *BioMed Research International*, 2017(3640901), 2017.

Fan Jiang, Hui Liu, Shaode Yu, and Yaoqin Xie. Breast mass lesion classification in mammograms by transfer learning. In *Proceedings of the 5th International Conference on Bioinformatics and Computational Biology*, ICBCB '17, pages 59–62, New York, NY, USA, january 2017. ACM. ISBN 978-1-4503-4827-0. doi: 10.1145/3035012.3035022.

Zhicheng Jiao, Xinbo Gao, Ying Wang, and Jie Li. A deep feature based framework for breast masses classification. *Neurocomputing*, 197:221 – 231, 2016. ISSN 0925-2312.

Pegah Khosravi, Ehsan Kazemi, Marcin Imielinski, Olivier Elemento, and Iman Hajirasouliha. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine*, 27:317 – 328, 2018. ISSN 2352-3964.

Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Hashoul. Medical image description using multi-task-loss cnn. In Gustavo Carneiro, Diana Mateus, Loïc Peter, Andrew Bradley, João Manuel R. S. Tavares, Vasileios Belagiannis, João Paulo Papa, Jacinto C. Nascimento, Marco Loog, Zhi Lu, Jaime S. Cardoso, and Julien Cornebise, editors, *Deep Learning and Data Labeling for Medical Applications*, pages 121–129, Cham, Sept 2016. Springer International Publishing. ISBN 978-3-319-46976-8.

Thijs Kooi and Nico Karssemeijer. Classifying symmetrical differences and temporal change in mammography using deep neural networks. *CoRR*, abs/1703.07715, 2017.

Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mrida, Clara I. Snchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal.*, 35:303–312, 2016.

Daniel Lévy and Arzav Jain. Breast mass classification from mammograms using deep convolutional neural networks. In *Neural Information Processing Systems*, volume abs/1612.00542 of *NIPS*, 2016.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

Maritte Lokate, Rebecca K. Stellato, Wouter B. Veldhuis, Petra H. M. Peeters, and Carla H. van Gils. Age-related changes in mammographic density and breast cancer risk. *American Journal of Epidemiology*, 178(1):101–109, 2013. doi: 10.1093/aje/kws446.

Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *CoRR*, abs/1512.02017, 2015. URL http://arxiv.org/abs/1512.02017.

Aly A. Mohamed, Wendie A. Berg, Hong Peng, Yahong Luo, Rachel C. Jankowitz, and Shandong Wu. A deep learning method for classifying mammographic breast density categories. *Medical Physics*, 45(1):314–321, 2018. ISSN 2473-4209. doi: 10.1002/mp.12683.

Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *CoRR*, abs/1511.06807, 2015.

Jane Peart, Glen Thomson, and Stephen Wood. Developing asymmetry in a screening mammogram: A cautionary tale of a missed cancer. *Journal of Medical Imaging and Radiation Oncology*, 62(1):77–80, 2017.

Richard Platania, Shayan Shams, Seungwon Yang, Jian Zhang, Kisung Lee, and Seung-Jong Park. Automated breast cancer diagnosis using deep learning and region of interest detection (bc-droid). In *Proceedings of the 8th ACM International Conference on Bioinformat-*

ics, Computational Biology,and Health Informatics, ACM-BCB '17, pages 536–543, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4722-8. doi: 10.1145/3107411.3107484.

Yuchen Qiu, Yunzhi Wang, Shiju Yan, Maxine Tan, Samuel Cheng, Hong Liu, and Bin Zheng. An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology. *Proc.SPIE*, 9785:9785 – 9785 – 6, 2016.

Dezso Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *CoRR*, abs/1707.08401, 2017.

Ravi Samala, Heang-Ping Chan, Lubomir M Hadjiiski, Mark A Helvie, Kenny Cha, and Caleb D Richter. Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer on mammograms. In *Physics in Medicine and Biology*, volume 62, 10 2017.

Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Jun Wei, and Kenny Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical Physics*, 43(12):6654–6666, 2016. ISSN 2473-4209.

Diane Scutt, Gillian Lancaster, and John Manning. Breast asymmetry and predisposition to breast cancer. In *Breast cancer research : BCR*, volume 8, page R14, 02 2006.

M. Paz Sesmero, Agapito I. Ledezma, and Araceli Sanchis. Generating ensembles of heterogeneous classifiers using stacked generalization. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):21–34, 2007.

Li Shen. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *CoRR*, abs/1708.09427, 2017.

Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

Philip Teare, Michael Fishman, Oshra Benzaquen, Eyal Toledano, and Eldad Elnekave. Malignancy Detection on Mammography Using Dual Deep Convolutional Neural Networks and Genetically Discovered False Color Input Enhancement. *Journal of Digital Imaging*, 2017. ISSN 1618727X. doi: 10.1007/s10278-017-9993-2.

G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *ArXiv e-prints*, July 2018.

Darvin Yi, Rebecca Lynn Sawyer, David Cohn III, Jared Dunnmon, Carson Lam, Xuerong Xiao, and Daniel L. Rubin. Optimizing and visualizing deep learning for benign/malignant classification in breast tumors. *CoRR*, abs/1705.06362, May 2017.

Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015. URL http://arxiv.org/abs/1506.06579.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. URL http://arxiv.org/abs/1311.2901.

Xiaofei Zhang, Yi Zhang, Erik Y. Han, Nathan Jacobs, Qiong Han, Xiaoqin Wang, and Jinze Liu. Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks. *IEEE Transactions on NanoBioscience*, pages 1–1, 2018.

Wentao Zhu and Xiaohui Xie. Adversarial deep structural networks for mammographic mass segmentation. *CoRR*, abs/1612.05970, 2016.

Appendix A. Candidate CNN selection

Transfer learning utilizing pretrained models on non-medical datasets has been shown to have competitive, and sometimes state-of-the-art, performance in many medical imaging and mammography tasks (Carneiro et al., 2017, 2015; Habibzadeh Motlagh et al., 2018; Huynh et al., 2016; Jiao et al., 2016; Khosravi et al., 2018; Lévy and Jain, 2016). Recent deep learning toolkits, such as *Keras*, allow practitioners to fine-tune and utilize many of the successful ImageNet models with ease. Because of this, we chose to evaluate and select the best CNN from the following ImageNet algorithms: ResNet50, VGG16, VGG19, InceptionV3, InceptionResNetV2 and Xception. We judged model performance on both ROI and full-images from the Curated Breast Imaging Subset of the Digital Database of Screening Mammography (CBIS-DDSM) (Heath et al., 2001). The DDSM is a database of 2,620 scanned film mammography studies. It contains normal, benign, and malignant cases with verified pathology information. The CBIS-DDSM collection includes a subset of the DDSM data selected and curated by a trained mammography reader. We chose this database due to the large number of related works using it, particularly with ROIs (Becker et al., 2016; Carneiro et al., 2017, 2015; Dhungel et al., 2015; Jiao et al., 2016; Lévy and Jain, 2016; Shen, 2017; Zhu and Xie, 2016).

We emulated the methodology presented in Shen (2017), a finalist in the 2016 DREAM mammography challenge, who generated a full-field mammogram classifier by first pre-training on ROIs from CBIS-DDSM. In the first step, we extracted patches from full-field mammograms without down-scaling and saved the images as 224 x 224 8-bit PNG files. Before saving patches we also standardized (0 μ, 1 σ) the entire set of patches by performing pixel-wise subtraction of the dataset mean and dividing by the dataset standard deviation. For every ROI patch saved we generated a "background" image, which was a uniformly random sampled region on the opposite (vertical and horizontal) half of the image. For training and testing we used an approximate 90-10 split, where 4000 total patches (including backgrounds) were used to train our network. We used an approximate 1:1 ratio for masses/calcification to background images. To deal with an extremely small training-set size and mitigating over-fitting, we applied random augmentation to each training image with the following specification: rotation within $\pm$25 degrees, shear up to 20 degrees counter-clockwise, horizontal flips, vertical flips, and zoom within $\pm$10%. We used a batch size of 16 and a cross entropy loss function. An iterative multi-step approach was used in training each CNN. The *Adam* optimizer with a learning rate of 10^{-3} was used for training the top layer, a learning rate of 10^{-4} for the top 50% of the network, and a learning rate of 10^{-5}

for fine tuning the rest of the network as described in Lévy and Jain (2016), Shen (2017), and Yi et al. (2017). For full-image experimentation, we used the same preprocessing, network hyperparameters and architecture used for ROIs, except we did not randomly sample background patches and also resized mammograms to 320 x 416 to preserve the aspect ratio.

Table 4 shows a comparison of candidate CNN architectures used to evaluate and test our approach. We evaluated 3 different class partitions for ROI images. In the 2-class experiment, ROI were classified as either benign or malignant. In the 3-class experiment, ROI were classified as either background, benign or malignant. And in the 5-class experiment, ROI were classified as one of background, benign calcification, benign mass, malignant calcification or malignant mass. Each of the neural networks were initialized with pre-trained ImageNet weights, and had the top-layer replaced by a global average pooling layer followed by a new fully-connected dense classifier. A single dense layer of 1024 neurons was selected to bias model fitting into the convolutional layers. Hyper-parameter tuning was forgone, since the goal of this experiment was just a ranking system for CNN selection. Inception-ResNetV2 performed the best in each classification task and metric, other than image-level AUPRC.

Table 5 shows the single mammogram classification performance of MT-CNN to the closely related works of Geras et al. (2017) and Zhang et al. (2018) on the public DDSM dataset. Each model used the same image preprocessing and augmentation presented in this section, and was trained using their published training hyperparameters and architecture. Slight modification of the network used in Geras et al. (2017) was required to accommodate a single mammogram rather than all four mammogram views. This was done by simply providing all mammograms into the first CNN they used and keeping the same subsequent layers unmodified.

Table 4: A trade-study of candidate CNN architectures on public CBIS-DDSM dataset to select the best CNN from available pre-trained ImageNet models to use as MT-CNN. For the ROI trade study, the reported values are the AUROC for 2-class, 3-class and 5-class stratifications. For full-images the AUROC and AUPRC are reported. The highest values for each experiment are in bold.

	ROI AUROC			Full-image	
Model	2-class	3-class	5-class	AUROC	AUPRC
ResNet50	0.740	0.734	0.706	0.607	0.488
VGG16	0.762	0.741	0.679	0.538	0.432
VGG19	0.783	0.739	0.665	0.542	0.402
InceptionV3	0.800	0.731	0.712	0.640	**0.541**
InceptionResNetV2	**0.842**	**0.841**	**0.844**	**0.652**	0.493
Xception	0.767	0.706	0.741	0.565	0.434

Table 5: Comparison of models on public dataset DDSM. For consistency all models are shown using TTA. Each model is trained using their published hyperparameters.

Model	AUROC	AUPRC
Geras et al. (2017)	0.490	0.408
Zhang et al. (2018)	0.531	0.423
MT-CNN	0.652	0.493

Appendix B. Training details

To bias *diagnosis* as the primary objective, loss weighting was adjusted according to the auxiliary output losses, such that the loss weight for *diagnosis* was greater than or equal to the sum of all other auxiliary output losses. Because cross-entropy loss performance deteriorates under scenarios of large class imbalance, this was mitigated by utilizing a focal loss function that is characterized by decreasing the penalty for well-classified examples (Lin et al., 2017). For a binary classification problem this is formally described as follows:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{3}$$

where $\gamma \geq 0$ is a focus tuning parameter, α_t is the inverse class frequency tuning parameter, and p_t is defined as follows:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise.} \end{cases} \tag{4}$$

In this experiment a focal loss was used for all categorical output targets with parameters α and γ initialized to 2. Focal loss was compared to cross-entropy loss as a sanity check and performance improvements were observed when using focal loss in regards to both model training time and predictive accuracy. For *age* and *density* regression targets, mean squared error (MSE) loss was used.

Because MT-CNN was initialized with ImageNet weights, lower-level CNN features were preserved by using an iterative and stratified training regime motivated by the work of Shen (2017). First, the fully-connected layer was trained for 1 epoch with the Adam optimizer and a learning rate of 10^{-3}. Then we cycled between training the top-most dense layers and the convolutional layers using a learning rate of 10^{-4} for 5 epochs each, followed by 10^{-5} for 10 epochs each. A batch size of 16 was used to train MT-CNN and was the largest that fit within GPU memory constraints. To bias MT-CNN to have the best diagnostic performance, at the end of each training epoch the model with the best AUROC for *diagnosis* was monitored and saved.

Due to the similarity in network architecture, MVMT was trained with the same training parameters, augmentation settings, preprocessing steps and loss functions mentioned when training MT-CNN. Due to the increase in network size and complexity, a batch size of 4 was required to fit within the limits of GPU memory. This required manually balancing the training classes, such that an equal number of positive and negative samples were seen

during each batch. Consider Fig. 1, only the dense layers after concatenation were trained and all other layers were preserved and not updated during back-propagation. Again, the *Adam* optimizer was used with learning rate initialized to $1e^{-4}$ for 5 epochs then $1e^{-5}$ for 15 epochs. During training all the previously mentioned augmentations were randomly applied to each input mammogram uniquely to provide the maximum amount of input variation.

All models have been generated, trained, validated and tested using Python, *Keras*, and *TensorFlow* on an Ubuntu Linux 16.04 OS and accelerated using two Nvidia GTX 1080 Ti GPUs with 11GB of memory each.

Appendix C. Visualizations

Visualizing the processing of a CNN is critical for understanding and interpreting model effectiveness and fidelity. Although several methods exist for visualization in CNNs (Mahendran and Vedaldi, 2015; Yosinski et al., 2015; Zeiler and Fergus, 2013), most require large data sets and network retraining. Instead, the method proposed in Geras et al. (2017) was used, which did not require network retraining and worked by simply examining the network's output sensitivity to perturbations in each input pixel. The premise was that a higher output variance will be observed when an "important" input pixel is perturbed. Using this method, Fig. 3 shows an example on three positive patients. Patient A was diagnosed correctly by both the radiologist and MVMT. For this patient, all of the predictions by MVMT agreed with the outcome except for the suspicion of the CC view, which MVMT deemed normal instead of suspicious. Patient B was diagnosed correctly by the radiologist but not MVMT. Patient C was diagnosed correctly by MVMT but not the radiologist. For this patient, the malignant lesion correctly identified by MVMT was also discovered by the radiologist, but was misdiagnosed as benign. Patient B has many non-breast pixels highlighted which agrees with the negative (normal) predictions of MVMT. For Patients A and C, MVMT recognized at least one "well-defined" region in either view and did not have any visible background pixels highlighted.

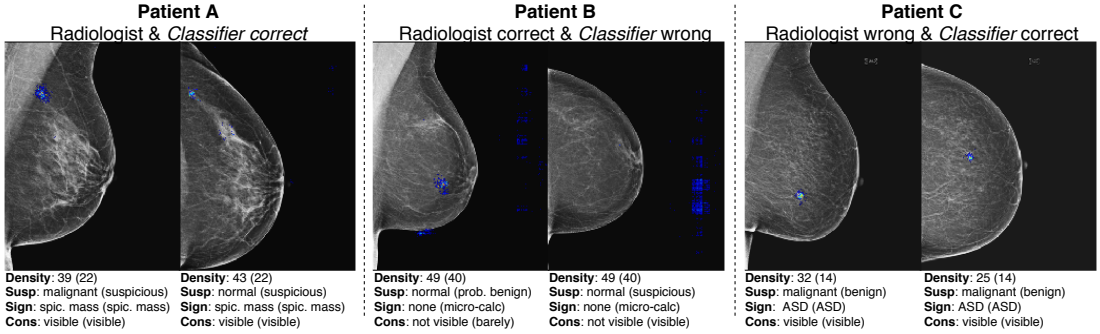

Figure 3: Example visualization of MVMT on three positive (malignant) patients. Patient A was diagnosed correctly by both the radiologist and MVMT; Patient B was diagnosed correctly by the radiologist but not MVMT; and Patient C was diagnosed correctly by MVMT, but not the radiologist. For each patient, the malignant breast is shown with MLO view on the left and CC on the right. The predicted *density*, *suspicion*, *sign* and *conspicuity* are shown, and ASD is asymmetrical density. The actual radiological annotations are in parenthesis.