



Published in final edited form as:

J Am Coll Radiol. 2020 December ; 17(12): 1653–1662. doi:10.1016/j.jacr.2020.05.015.

Multi-Institutional Assessment and Crowdsourcing Evaluation of Deep Learning for Automated Classification of Breast Density

Ken Chang¹, Andrew L. Beers¹, Laura Brink², Jay B. Patel¹, Praveer Singh¹, Nishanth T. Arun¹, Katharina V. Hoebel¹, Nathan Gaw¹, Meesam Shah², Etta D. Pisano^{2,3}, Mike Tilkin², Laura P. Coombs², Keith J. Dreyer^{2,4,6}, Bibb Allen^{2,5}, Sheela Agarwal^{2,6}, Jayashree Kalpathy-Cramer^{1,7}

¹Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

²American College of Radiology, Reston, VA, USA

³Beth Israel Lahey Medical Center, Department of Radiology, Boston, MA, USA

⁴Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

⁵Grandview Medical Center, Birmingham, AL, USA

⁶Lenox Hill Radiology, New York, NY, USA

⁷MGH and BWH Center for Clinical Data Science, Massachusetts General Hospital, Boston, MA, USA

Abstract

Objective: We developed deep learning algorithms to automatically assess Breast Imaging Reporting and Data System breast density.

Methods: Using a large multi-institution patient cohort of 108,230 digital screening mammograms from the Digital Mammographic Imaging Screening Trial, we investigated the effect of data, model, and training parameters on overall model performance and provided

Corresponding Author: Jayashree Kalpathy-Cramer, Athinoula A. Martinos Center for Biomedical Imaging, 149 13th Street, Charlestown, MA 02129. Phone: 617-724-4657; Fax: 617-726-7422, kalpathy@nmr.mgh.harvard.edu.

Role

Study concept and design: Chang, Beers, Brink, Shah, Pisano, Tilkin, Coombs, Dreyer, Allen, Agarwal, Kalpathy-Cramer

Acquisition, analysis, or interpretation of data: Chang, Beers, Brink, Patel, Singh, Arun, Hoebel, Gaw, Shah, Pisano

Drafting of the manuscript: Chang, Beers, Patel, Singh, Arun, Hoebel, Gaw, Coombs, Agarwal

Critical revision of the manuscript for important intellectual content: All Authors

Supervision: Pisano, Tilkin, Coombs, Dreyer, Allen, Agarwal, Kalpathy-Cramer

Statement of data access and integrity

The authors declare that they had full access to all of the data in this study and the authors take complete responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of Interest

J.K. is a consultant/advisory board member for Infotech, Soft. S.A. is an employee of Bayer HealthCare. The other authors declare no competing interests.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

crowdsourcing evaluation from the attendees of the American College of Radiology 2019 Annual Meeting.

Results: Our best performing algorithm achieved good agreement with radiologists who were qualified interpreters of mammograms, with a 4-class κ of 0.667. When training was performed with randomly sampled images from the dataset versus sampling equal number of images from each density category, the model predictions were biased away from the low-prevalence categories such as extremely dense breasts. The net result was an increase in sensitivity and a decrease in specificity for predicting dense breasts for equal class compared to random sampling. We also found that the performance of the model degrades when we evaluate on digital mammography data formats that differ from the one that we trained on, emphasizing the importance of multi-institutional training sets. Lastly, we showed that crowdsourced annotations, including those from attendees who routinely read mammograms, had higher agreement with our algorithm than with the original interpreting radiologists.

Conclusion: We demonstrated the possible parameters that can influence the performance of the model and how crowdsourcing can be used for evaluation. This study was performed in tandem with the development of the ACR AI-LAB, a platform for democratizing Artificial Intelligence.

Introduction

Breast cancer is one of the leading causes of death among women in the US, with the expected number of deaths to be over 41,000 in 2019. [1] Early mammographic screening has resulted in a decrease in breast cancer mortality. [2,3] The correct mammographic interpretation of breast density, which measures extent of fibroglandular tissue, is important in the assessment of breast cancer risk as there is increased risk with increased density. [4,5] Furthermore, the identification of dense breast may stratify patients who may have masked cancers and may benefit from additional ultrasound and/or MR imaging. As such, there is now legislation in many states that patients must be notified of their breast density after mammography. [6]

Qualitative assessment by means of the widely used Breast Imaging Reporting and Data System (BI-RADS) include four categories: a) almost entirely fatty, b) scattered fibroglandular densities, c) heterogeneously dense, or d) extremely dense. [7] These criteria are subjective, resulting in inter-rater variability among radiologists. A study by Sprague et al. showed that the likelihood of any given mammogram being rated as dense (heterogeneously dense and extremely dense) is highly dependent on the interpreting radiologist, with the percentage ranging from 6.3%–84.5%. [8] Other studies have reported intra-reader variability to be $\kappa = 0.58$ (among 34 community radiologists) and the inter-rater variability to be $\kappa = 0.643$ (between a consensus of 5 breast radiologists and the original interpreting breast radiologist). [6,9] Similarly, commercially available software shows a wide range of agreement with clinical experts and the probability of dense classification is dependent on the specific software used. [10,11] This intra- and inter-rater variability, and even inter-software variability, may confer undue patient anxiety and potential harm to the patient, i.e. possible unnecessary supplemental screening examinations.

As such, there has been interest in using automated approaches to improve accuracy and consistency of breast density assessment. Commercial software utilize quantitative imaging features to assess breast density, with mixed agreement with radiologist interpretation. [11] Deep learning methods have yielded state-of-the-art results in a wide range of computer vision tasks without the need for domain-inspired hand-crafted imaging features. Moreover, recent studies have shown the potential of deep learning in medical fields such as dermatology, ophthalmology, and radiology. [12–14] A recent study from Lehman et al. demonstrates the utility of deep learning for mammographic density assessment in clinical practice at a single institution/mammography system. [6] Here, we further this work by validating the deep learning approach on a multi-institutional imaging cohort with a variety of digital-mammography systems. Furthermore, we provide an in-depth analysis of how choice of data, model, and training parameters affects algorithm performance. In addition to that, we investigate the generalizability of models across different digital-mammography data formats. Lastly, we deploy our system at the American College of Radiology (ACR) 2019 Annual Meeting for a crowdsourced evaluation.

Materials and Methods

Patient Cohort

Digital screening mammograms from 33 clinical sites were retrospectively obtained through the Digital Mammographic Imaging Screening Trial (DMIST), the details of which were previously published. [15] Each examination was interpreted by a single radiologist from a cohort of radiologists using ACR BI-RADS breast density lexicon (Category a: fatty, Category b: scattered, Category c: heterogeneously dense, Category d: extremely dense). [7] There were a total of 92 radiologists reading the exams. Readers in the United States were all qualified interpreters of mammograms under federal law. Canadian readers met equivalent standards. Each site's lead radiologist received training to read for DMIST and in turn trained the site's other readers. All images were previously de-identified before this study. The mammograms were saved in DICOM format with 4 different image data formats, corresponding to different digital-mammography systems or different versions of the same system (Table 1): 12 bit Monochrome 1 (30.3%), 12 bit Monochrome 2 (11.2%), 14 bit Monochrome 1 (58.0%), and 14-bit Monochrome 2 (0.5%). 14-bit Monochrome 2 images were excluded to ensure that each image data format included in our study had adequate representation for training of our deep learning model. Our final patient cohort consisted of 108,230 digital screening images from 21,759 patients (Table 2), which was divided into training, validation, and testing sets on the patient level. The training set was used to develop the model and the validation set was used to assess model performance during training to prevent overfitting. The test set was unseen until the model training was complete.

Experiments on Data, Model, and Training Parameters

Neural network models were implemented in DeepNeuro with Keras/TensorFlow backend. [16] We investigated the effect of data, model, and training parameters on algorithm performance. A schematic of the various experiments investigating data, model, and training parameters are summarized in Fig. 1A. To investigate the effect of training set size, we utilized various different training set sizes and assessed the resulting performance on the test

set. We tested four commonly used neural network architectures, each of which differ in number of layers and design: ResNet50, DenseNet121, InceptionV3, and VGG16. [17–21] We also investigated the benefit of pretraining by comparing ImageNet (a large computer vision dataset of natural images) pretrained versus random initialization. [22] A variety of cost-functions were also utilized (categorical cross-entropy, mean absolute error, mean squared error, and ordinal regression) in order to assess the effect of objective function (and their underlying assumptions of the nature of the labels) on performance. [23] The training set was augmented in real time by means of random horizontal/vertical flips (50% probability of each) and random rotations (0–45°). At each mini-batch, images from each breast density class were sampled with either random (weighting in the empirical density class distribution) or equal class (weighting each density class equally) probability to assess the effect of class weighting on performance. We also evaluated the effect of model ensembling by averaging the output of 2–4 trained models of the same architecture (ResNet50). Model ensembling describes the process by which several independently trained models are combined to improve performance. [24] The default model utilized 100% of the training set, ImageNet pretrained weights, ResNet50 architecture, no ensembling, categorical cross-entropy loss function, augmentation, and equal class sampling. Only one parameter was modified at a time in the experiments, keeping all other parameters the same as the default model (*ceteris paribus*).

Experiments on Image Data Formats

To visualize the differences in intensity distributions across image formats, histograms of preprocessed images from the testing set were generated. The dimensionality of histograms were then reduced to a 2-dimensional projection and plotted to inspect for similarity across image formats. [25] The effect of image format of training images on generalizability of models was investigated. We trained ResNet50 models using 12 bit Monochrome 1 images only, 12 bit Monochrome 2 images only, 14 bit Monochrome 1 images only, and all images. The performance of these models for each image format was then assessed. Projections of the intermediate output of the penultimate layer of the neural network were also plotted for images in the testing set using a model trained on all images to evaluate the learned features learned by the deep learning model. Further information about the patient cohort and experiments is available in the Supplementary Information.

Crowdsourcing Assessment

As further evaluation of our breast density algorithm, we deployed an annotation workstation at the ACR 2019 Annual Meeting. Attendees of all levels (researchers, medical students, residents, radiologists) were invited to perform annotations on a subset of images within our patient cohort. Representative images of all breast density classifications from the BI-RADS manual were provided to attendees during annotation. Attendees were able to inspect all images (all views available) from a given patient study and were asked to provide a BI-RADS breast density assessment. In total, 3,649 annotations were performed on 1083 patient studies by 17 raters (Demographics summarized in Table 3). On average, there were 3 annotations per patient study and each rater performed 215 annotations. Consensus of the crowd was determined by majority vote, with random tiebreak. In our analysis, we looked at agreement between crowd and original interpreting radiologist annotation as well

as crowd and algorithm (ResNet50). The ResNet50 model was chosen because it was the best performing architecture among those tested. Only a single model (as opposed to an ensemble model) was used to reflect the common scenario where only a single model is deployed for computation efficiency.

Statistical Analysis

Agreement between raters was assessed via linear κ coefficient across the four breast density categories in the testing set (4-class κ). For reference, a κ of 0.21–.40, 0.41–0.60, and 0.61–0.80 represents fair, moderate, and substantial agreement, respectively. [26]

Results

Effect of Data Parameters on Performance

The performance of training set size on testing set performance was investigated, showing that κ coefficient increases as the training set size increases. When 2% ($n = 1247$ images) of the training set was used, the mean 4-class κ was 0.563 (95% Confidence Interval, CI, 0.551–0.576). In contrast, when using 100% ($n = 62,316$ images) of the training set, the mean 4-class κ was 0.660 (95% CI 0.657–0.664) (Fig. 1B). There was a statistically significant difference between the performance of using 2%–60% and 100% of the training set (t-test $p < .05$). There was no difference in the performance of using 80% and 100% of the training set ($p = 0.291$).

Effect of Model Parameters on Performance

The mean 4-class κ of models with randomly initialized weights was 0.327 (95% CI 0.273–0.384), compared to ImageNet pretrained weights 0.660 (95% CI 0.657–0.664, $p < .001$) when using the full training set (Fig. 1C). In the experiments assessing model architecture, the mean 4-class κ of ResNet50, DenseNet121, InceptionV3, and VGG16 was 0.660 (95% CI 0.657–0.664), 0.650 (95% CI 0.640–0.659), 0.644 (95% CI 0.635–0.652), and 0.660 (95% CI 0.658–0.664), respectively. There was no statistically significant difference between the performance of the various architectures. The mean 4-class κ of no ensembling, ensembling two models, ensembling three models, and ensembling four models was 0.660 (95% CI 0.657–0.664), 0.665 (95% CI 0.664–0.666), 0.666 (95% CI 0.666–0.667), 0.667 (95% CI 0.666–0.668), respectively. The performance of ensembling four models and three models was greater than that of no ensembling ($p = 0.041$ and 0.036, respectively).

Effect of Training Parameters on Performance

For categorical cross-entropy, mean absolute error, mean squared error, and ordinal regression, the mean 4-class κ was 0.660 (95% CI 0.657–0.664), 0.649 (95% CI 0.644–0.653), 0.654 (95% CI 0.646–0.661), and 0.664 (95% CI 0.659–0.669), respectively. The performance of categorical cross-entropy and ordinal regression was significantly greater than mean absolute error ($p = 0.011$ and $p = 0.004$, respectively). The mean 4-class κ with no augmentation was 0.658 (95% CI 0.646–0.666), compared to with augmentation 0.660 (95% CI 0.657–0.664) ($p = 0.675$). The mean 4-class κ with random and equal class sampling at each mini-batch was 0.665 (95% CI 0.662–0.669) and 0.660 (95% CI 0.657–0.664), respectively ($p = 0.135$). For random class sampling, the predicted distribution of

labels on the test set was 8.1% fatty, 47.5% scattered, 40.1% heterogeneously dense, and 4.3% extremely dense. This differed from the predicted distribution of labels on the test set with equal class sampling, which was 13.5% fatty, 37.5% scattered, 36.8% heterogeneously dense, and 12.2% extremely dense ($p < 0.001$, Fig. 2B). The predicted binary distribution for random (44.4% dense) and equal sampling (49.0% dense) also differed ($p < 0.001$). For random class sampling, the mean sensitivity and specificity for classifying dense breast was 0.833 (95% CI 0.803–0.856) and 0.888 (95% CI 0.872–0.905), respectively. Comparatively, for equal class sampling, there was an increase in sensitivity (0.880, 95% CI 0.869–0.890, $p < .05$) with a decrease in specificity (0.842, 95% CI 0.828–0.857, $p < 0.001$). A display of the range of classifications for models trained with different model and training parameters for 50 patients in the testing set is shown in Fig. 2A.

Effect of Digital-Mammography Data Format on Model Generalizability

A plot of projections of intensity distributions of preprocessed images showed clustering within image format, delineating differences between image formats (Fig. 3B). Clustering by intensity distribution was preserved even after passing the images through a trained neural network, as shown by projections of the output of the penultimate layer, with the grouping by breast density occurring within the respective image format cluster (Fig. 3D-E). For all image format specific models, testing set performance was decreased for other image formats compared to the image format the model was trained on ($p < 0.001$). In contrast, a model trained on all images showed no differences in performance across image formats ($p > 0.05$, Fig. 3C).

Crowdsourcing Assessment

The 4-class κ between the crowd and algorithm (0.505, 95% CI .503–.506) was greater than agreement between crowd and original interpreting radiologist (0.463, 95% CI 0.461–0.464, $p < .001$). Agreement with the algorithm was greater than agreement with the original interpreting radiologist for both crowd participants who regularly read mammograms and those who do not (Fig. 4D). As a reference, the 4-class κ between algorithm and radiologist was 0.636 (95% CI 0.635–0.637) for the same patient studies (Fig. 4A-C).

Discussion

In this study, we investigated the performance of deep learning models in a large multi-institution and multi-mammography system patient cohort. Our best performing model achieved a κ of 0.667, equivalent to the agreement observed by Lehman et al., which only utilized mammograms from a single institution/mammography system. [6]

One challenge of training robust deep learning models is the availability of large annotated imaging datasets. [27] In this study, we provide empirical evidence that the size of the training set is a key determinant in the performance of neural networks, consistent with another study on abnormality classification in chest radiographs. [28] In accordance with deep learning studies in other domains, tens of thousands of annotated images are needed before model performance begins to plateau in diverse imaging cohorts, supporting the need for collaborative efforts among medical institutions. [28–30]

In our investigation of model parameters, pretraining and ensembling led to improvements in performance. Pretraining neural networks followed by fine-tuning in the target domain (also known as transfer learning) has become a well-established paradigm for medical imaging applications to achieve high performance. [12,29] In our study, we noted that pretraining on ImageNet improved performance for the breast density classification task. Further improvement in performance was seen with ensembling of independently trained models which is analogous to how a consensus of experts is more likely to be correct than any single expert. [31] Interestingly, neural network architecture did not have a significant effect on performance despite differences in model complexity and design.

One important consideration when training a model is the objective function used to optimize the algorithm, also known as a cost function. Our experiments have shown that the choice of cost function had a significant effect on model performance, mainly because each cost-function makes different assumptions about the nature of the labels. Specifically, mean absolute error, mean squared error, and ordinal regression assume that the categories are ordered while categorical cross-entropy does not. Furthermore, mean absolute error and mean squared error assume the distance between adjacent classes is equal whereas ordinal regression does not. In our application, breast density is classified on an ordered scale with undefined distances between adjacent classes (i.e. the distances between Fatty and Scattered compared to heterogeneously dense and extremely dense cannot be quantified), making ordinal regression the most appropriate cost function. This is validated in our experiments, where we find that ordinal regression exhibited the highest performance, although this was significantly different to only mean absolute error.

We also did not notice significant difference between random and equal class sampling on model performance in terms of κ coefficient. Class sampling is an important consideration in cases where there are differences in the number of patient samples from each class (i.e. when the majority class significantly outnumbers the minority class). In our study, we have more patients with scattered and heterogeneously dense breast (44.2% and 37.5% respectively) than with fatty and extremely dense breast (11.8% and 6.6%, respectively), which is the expected distribution as breast density has a normal distribution. Under random class sampling, the neural network would be exposed to more training examples of scattered and heterogeneously dense breast than of fatty and extremely dense breast. Equal class sampling can be used to mitigate this inherent class imbalance by ensuring that the neural network is adequately exposed to all classes. [32] However, it is also important to note that with equal class sampling, the distribution of predicted labels changes – specifically, minority classes are predicted with higher frequency and majority classes are predicted with lower frequency, as shown by our experimental results. The net result of this is that the sensitivity of predicting dense breast improves with equal class sampling. Moreover, equal class sampling leads to lower specificity for classification of dense breast. From a policy perspective, this can lead to more patients being notified that they have dense breast. If additional imaging is performed on these patients, this may lead to increases in the number of false positives. This is a key example of how the manner in which deep learning models are trained can have implications for clinical care.

One critical hurdle that prevents the deployment of deep learning models in the clinical work environment is their relatively poor generalizability across institutional differences, such as patient demographics, disease prevalence, scanners, and acquisition settings. In fact, other deep learning studies that have shown poor generalizability of deep learning models when applied to data from different institutions than the one they were trained on. [33,34] In our study, we found that models trained on specific digital-mammography data formats do not generalize to other data formats, and it was only after training on images from all digital-mammography data formats did our model high performance on all data formats. Indeed, several deep learning studies for mammographic breast density assessment were only validated on patient cohorts from a single institution and/or digital-mammography systems. [6,35,36] Some possible differences between different digital-mammography systems or versions of systems include the x-ray tube target, filter, digital detector technology, and control of automatic exposure. [37] Our results add to the growing body of literature that states that deep learning models do not necessarily generalize when applied to data that differs from that which the model was trained with.

Various studies have shown the utility of crowdsourcing and citizen science in biological and medical image annotation. [38–41] Crowdsourcing for annotation and evaluation is advantageous because it is scalable, high-throughput, cost-efficient, and accurate.[42–44] As such, we performed a crowdsourcing assessment of our algorithm. Notably, there was a diversity of experience of the participants in crowdsourcing, with its inclusion of students, residents, and radiologists who do not routinely read mammograms. As such, it is unsurprising that the agreement between the crowd and algorithm was lower than the agreement between the original interpreting radiologist and algorithm. Interestingly, the crowd (both participants who routinely read mammograms and those who did not) had higher agreement with the algorithm than with the original interpreting radiologist. This may reflect the consistency of the algorithm in its assessment compared to the various interpreting radiologists from different sites in the DMIST study. In other words, a single algorithm may allow for greater consistency than having different human radiologists rate each imaging study.

There are several limitations to our study. The first is that we only had one radiologist, from a cohort of radiologists, perform interpretation for each patient study. Future studies will incorporate multiple readers for each patient study. In addition, for models initialized with random weights, we did not optimize training hyperparameters such as the learning rate schedule or the duration of training. [45] It is possible that optimization would improve the performance of the randomly initialized model, but in this study, we show the performance advantage of pretrained neural networks with minimal hyperparameter tuning. Furthermore, in our investigation of augmentation, we only explored random flips and rotations, though future work will explore other augmentation techniques such as intensity scaling and elastic deformations. [46] Lastly, our algorithm was only developed to assess mammographic breast density. Future work can extend our algorithm and crowdsourcing evaluation for more complex tasks such as assigning BI-RADS categories.

This study was developed in conjunction with the ACR AI-LAB, a framework for democratization of Artificial Intelligence (AI). The goal of the ACR AI-LAB is to

provide an interface for clinicians and scientists to work together to develop deep learning models. We highlight several fundamental features needed for AI democratization: First, we demonstrate the possible data, model, and training parameters that can influence the performance of the model. These parameters will be available as options in AI-LAB. We also show importance of diverse training data for model generalizability, supporting collaborative development of algorithms across institutions which the AI-LAB will facilitate. Lastly, we show how a crowdsourced annotations can be used to evaluate algorithm performance, which users will be able to do on the AI-LAB platform.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to acknowledge the computing resources provided by the MGH and BWH Center for Clinical Data Science and Amazon Web Services Machine Learning Research Awards.

Funding

Research reported in this publication was supported by a training grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under award number 5T32EB1680 to K. Chang and J. B. Patel and by the National Cancer Institute (NCI) of the National Institutes of Health under Award Number F30CA239407 to K. Chang. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This publication was supported from the Martinos Scholars fund to K. Hoebel. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Martinos Scholars fund.

This study was supported by National Institutes of Health grants U01 CA154601, U24 CA180927, and U24 CA180918 to J. Kalpathy-Cramer.

This research was carried out in whole or in part at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health.

References

- [1]. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin* 2019;69:7–34. doi:10.3322/caac.21551. [PubMed: 30620402]
- [2]. Duffy SW, Tabár L, Chen H-H, Holmqvist M, Yen M-F, Abdsalah S, et al. The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties. *Cancer* 2002;95:458–69. doi:10.1002/cncr.10765. [PubMed: 12209737]
- [3]. Tabár L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 2001;91:1724–31. [PubMed: 11335897]
- [4]. Boyd NF, Byng JW, Jong RA, Fishell EK, Little LE, Miller AB, et al. Quantitative Classification of Mammographic Densities and Breast Cancer Risk: Results From the Canadian National Breast Screening Study. *JNCI J Natl Cancer Inst* 1995;87:670–5. doi:10.1093/jnci/87.9.670. [PubMed: 7752271]
- [5]. Razzaghi H, Troester MA, Gierach GL, Olshan AF, Yankaskas BC, Millikan RC. Mammographic density and breast cancer risk in White and African American Women. *Breast Cancer Res Treat* 2012. doi:10.1007/s10549-012-2185-3.

- [6]. Lehman CD, Yala A, Schuster T, Dontchos B, Bahl M, Swanson K, et al. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology* 2019;290:52–8. doi:10.1148/radiol.2018180694. [PubMed: 30325282]
- [7]. Liberman L, Menell JH. Breast imaging reporting and data system (BI-RADS). *Radiol Clin North Am* 2002. doi:10.1016/S0033-8389(01)00017-3.
- [8]. Sprague BL, Conant EF, Onega T, Garcia MP, Beaber EF, Herschorn SD, et al. Variation in Mammographic Breast Density Assessments Among Radiologists in Clinical Practice. *Ann Intern Med* 2016;165:457. doi:10.7326/M15-2934. [PubMed: 27428568]
- [9]. Spayne MC, Gard CC, Skelly J, Miglioretti DL, Vacek PM, Geller BM. Reproducibility of BI-RADS breast density measures among community radiologists: A prospective cohort study. *Breast J* 2012;18:326–33. doi:10.1111/j.1524-4741.2012.01250.x. [PubMed: 22607064]
- [10]. Brandt KR, Scott CG, Ma L, Mahmoudzadeh AP, Jensen MR, Whaley DH, et al. Comparison of Clinical and Automated Breast Density Measurements: Implications for Risk Prediction and Supplemental Screening. *Radiology* 2016;279:710–9. doi:10.1148/radiol.2015151261. [PubMed: 26694052]
- [11]. Youk JH, Gweon HM, Son EJ, Kim J-A. Automated Volumetric Breast Density Measurements in the Era of the BI-RADS Fifth Edition: A Comparison With Visual Assessment. *Am J Roentgenol* 2016;206:1056–62. doi:10.2214/AJR.15.15472. [PubMed: 26934689]
- [12]. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8. doi:10.1038/nature21056. [PubMed: 28117445]
- [13]. Chang K, Beers AL, Bai HX, Brown JM, Ly KI, Li X, et al. Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bi-dimensional measurement. *Neuro Oncol* 2019. doi:10.1093/neuonc/noz106.
- [14]. Li MD, Chang K, Bearce B, Chang CY, Huang AJ, Campbell JP, et al. . Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *Npj Digit Med* 2020;3:48. doi:10.1038/s41746-020-0255-1.
- [15]. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *N Engl J Med* 2005;353:1773–83. doi:10.1056/NEJMoa052911. [PubMed: 16169887]
- [16]. Beers A, Brown J, Chang K, Hoebel K, Gerstner E, Rosen B, et al. DeepNeuro: an open-source deep learning toolbox for neuroimaging 2018.
- [17]. Glorot X, Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Proc Int Conf Artif Intell Stat (AISTATS'10) Soc Artif Intell Stat* 2010.
- [18]. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conf. Comput. Vis. Pattern Recognit., IEEE; 2016, p. 770–8. doi:10.1109/CVPR.2016.90.
- [19]. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely Connected Convolutional Networks 2016.
- [20]. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, 2016. doi:10.1109/CVPR.2016.308.
- [21]. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Int Conf Learn Represent* 2015:1–14. doi:10.1016/j.infsof.2008.09.005.
- [22]. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* 2015. doi:10.1007/s11263-015-0816-y.
- [23]. Cheng J, Wang Z, Pollastri G. A neural network approach to ordinal regression. *Proc. Int. Jt. Conf. Neural Networks*, 2008. doi:10.1109/IJCNN.2008.4633963.
- [24]. Dietterich TG. Ensemble methods in machine learning. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2000. doi:10.1007/3-540-45014-9_1.
- [25]. McInnes L, Healy J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction 2018.
- [26]. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 1977. doi:10.2307/2529310.

- [27]. Deng Jia, Dong Wei, Socher R Li-Jia Li, Li Kai, Fei-Fei Li. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conf. Comput. Vis. Pattern Recognit., 2009. doi:10.1109/CVPRW.2009.5206848.
- [28]. Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs. *Radiology* 2019. doi:10.1148/radiol.2018181422.
- [29]. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama* 2016;304:649–56. doi:10.1001/jama.2016.17216.
- [30]. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. *J Am Med Informatics Assoc* 2018. doi:10.1093/jamia/ocy017.
- [31]. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RVP, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *JAMA Ophthalmol* 2018;97239:1–8. doi:10.1001/jamaophthalmol.2018.1934.
- [32]. Van Hulse J, Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data, 2008. doi:10.1145/1273496.1273614.
- [33]. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med* 2018. doi:10.1371/journal.pmed.1002683.
- [34]. Albadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing: Impact. *Med Phys* 2018. doi:10.1002/mp.12752.
- [35]. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening 2019.
- [36]. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. *Med Phys* 2018. doi:10.1002/mp.12683.
- [37]. Keavey E, Phelan N, O'Connell AM, Flanagan F, O'Doherty A, Larke A, et al. Comparison of the clinical performance of three digital mammography systems in a breast cancer screening programme. *Br J Radiol* 2012. doi:10.1259/bjr/29747759.
- [38]. Shih G, Wu CC, Halabi SS, Kohli MD, Prevedello LM, Cook TS, et al. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiol Artif Intell* 2019;1:e180041. doi:10.1148/ryai.2019180041. [PubMed: 33937785]
- [39]. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology* 2019;290:498–503. doi:10.1148/radiol.2018180736. [PubMed: 30480490]
- [40]. Flanders AE, Prevedello LM, Shih G, Halabi SS, Kalpathy-Cramer J, Ball R, et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiol Artif Intell* 2020;2:e190211. doi:10.1148/ryai.2020190211. [PubMed: 33937827]
- [41]. Filice RW, Stein A, Wu CC, Arteaga VA, Borstelmann S, Gaddikeri R, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. *J Digit Imaging* 2019;1–7. doi:10.1007/s10278-019-00299-9. [PubMed: 30030764]
- [42]. Su H, Deng J, Fei-Fei L. Crowdsourcing annotations for visual object detection. *AAAI Work. - Tech. Rep*, 2012.
- [43]. Irshad H, Montaser-Kouhsari L, Waltz G, Bucur O, Nowak JA, Dong F, et al. Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd. *Pacific Symp. Biocomput.*, 2015, p. 294–305.
- [44]. Candido dos Reis FJ, Lynn S, Ali HR, Eccles D, Hanby A, Provenzano E, et al. Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer. *EBioMedicine* 2015. doi:10.1016/j.ebiom.2015.05.009.
- [45]. He K, Girshick R, Dollár P. Rethinking ImageNet Pre-training 2018.

- [46]. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. . nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. Inform. aktuell, 2019. doi:10.1007/978-3-658-25326-4_7.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

We demonstrate the effect of various parameters on the performance of the mammographic breast density classification model, the lack of generalizability across data formats, and how crowdsourcing can be used for evaluation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Take-Home Points

- The choice of data, model, and training parameters can impact deep learning model performance for evaluation of mammographic breast density. Notably, when training was performed with randomly sampled images from the dataset versus sampling equal number of images from each density category, the model predictions were biased away from the low-prevalence categories such as extremely dense breasts.
- The performance of the model degrades when evaluated on data formats that differ from the one that we trained on, emphasizing the importance of multi-institutional training sets.
- Crowdsourcing can be an effective means of evaluating model performance.
- These options for model training and evaluation will be made available in the ACR AI-LAB, a platform for democratizing Artificial Intelligence that was developed in tandem with this study.

Implications

We showcase the various data, training, and model parameters that can influence model performance, highlighting pretraining, cost-function, and sampling approach as important parameters. Furthermore, we found that model performance deteriorates when training and testing on different imaging data formats. In performing this study in tandem with the development of the ACR AI-LAB, we demonstrate important principles and pitfalls that radiologists and data scientists have to consider when training neural network models. Our hope is that radiologists who use the AI-LAB can refer to this study as an educational tool when utilizing the AI-LAB to train their own deep learning models.

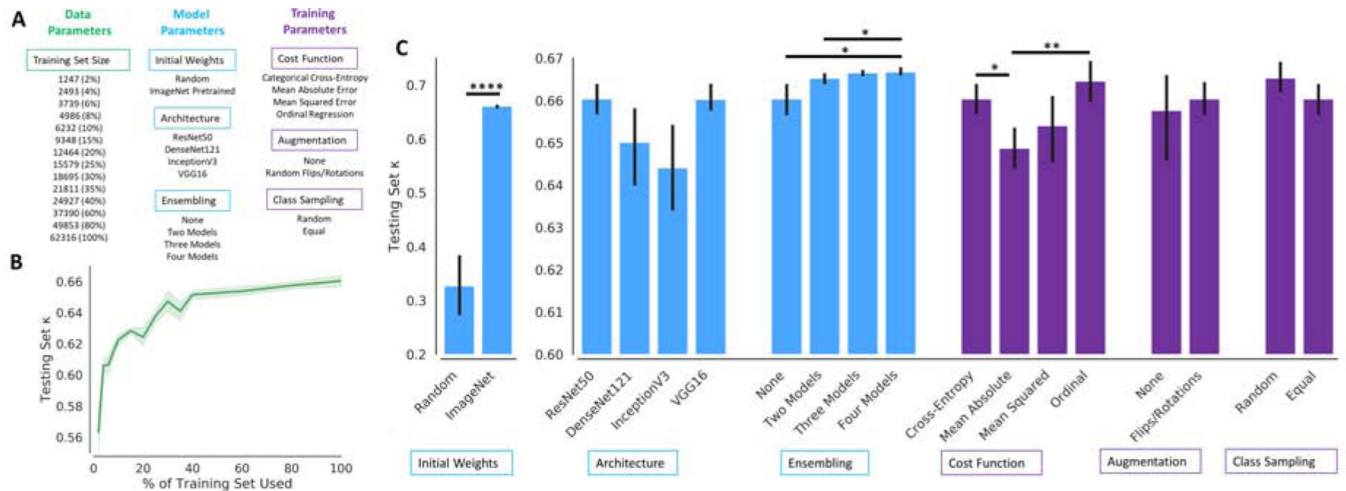


Figure 1.

(A) A summary of all the data, model, and training parameter experiments performed. (B) Performance on the testing set (measured by 4-class κ agreement with radiologist interpretation) increased as the percentage of training set used. The 95% confidence interval is plotted in light green. (C) Effect of model and training parameters on testing set 4-class κ agreement with radiologist interpretation. Black lines denote 95% confidence interval. P-values are denoted by * $p < .05$, ** $p < .01$, *** $p < .001$

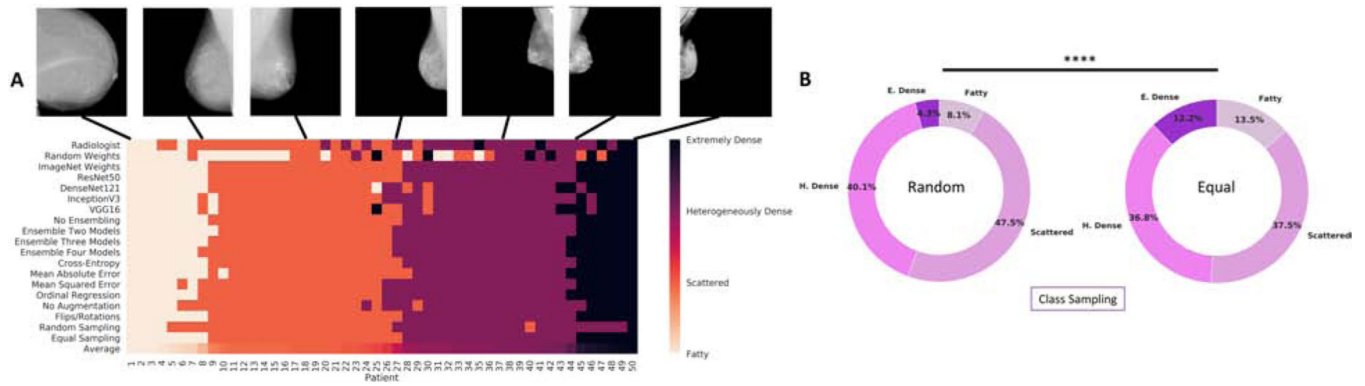


Figure 2.

(A) A visual display of the range of classifications for models trained with different model and training parameters for 50 patients in the testing set. The radiologist interpretation is displayed in the first row. The average breast density rating across all models and radiologist interpretation is displayed in the last row and was used to order the patients from least dense (left) to most dense (right). (B) The distribution of predicted breast density labels in the testing set differed for experiments with random class sampling (left) compared to equal class sampling (right) at each mini-batch. P-values are denoted by **** $p < .001$

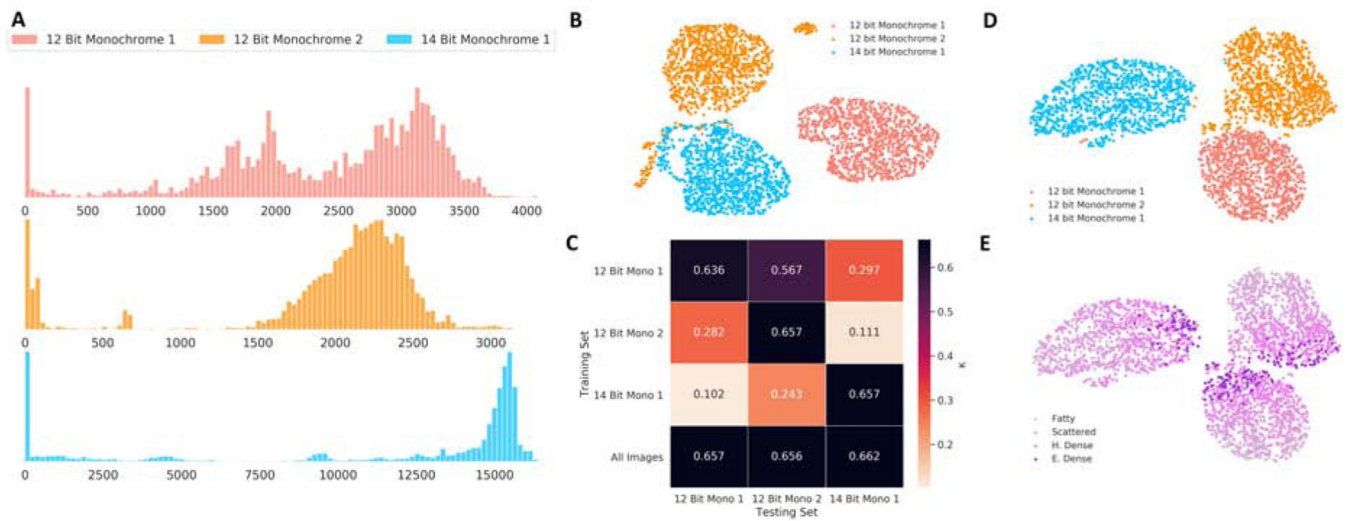


Figure 3.

(A) Intensity distribution histogram (Frequency vs Intensity Value) of 100 randomly selected images of each pixel format. (B) Visualization of the histogram of intensities of 3000 preprocessed images from the testing set demonstrating clustering of images by image format. (C) Performance of models trained on specific image formats as well as all images, showing that for image format specific models, testing set performance was decreased for other image formats compared to the image format the model was trained on. (D-E) Visualization of an intermediate layer of the trained neural network for 3000 images in the testing set, color-coded by image format and radiologist interpretation of breast density.

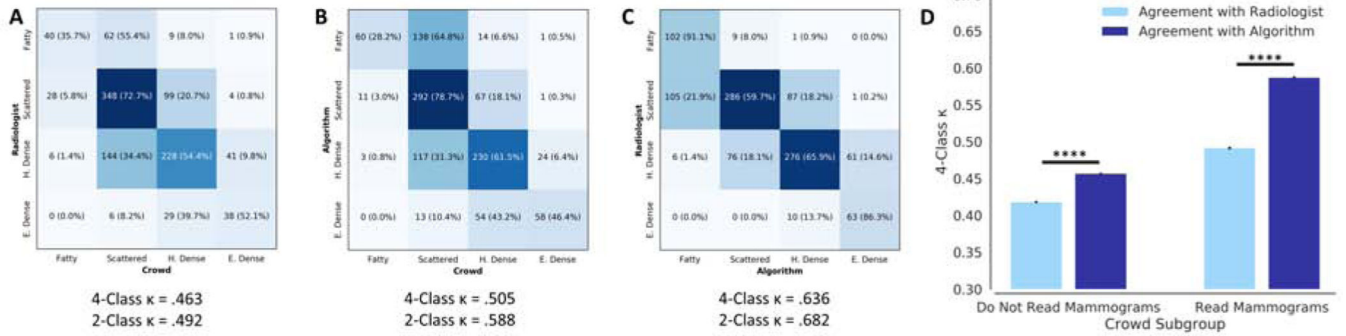


Figure 4. Confusion matrices showing the agreement between original interpreting radiologist, algorithm, and crowd. The agreement between the algorithm and crowd (B) was greater than the agreement between crowd and original interpreting radiologist (A). The agreement between algorithm and original interpreting radiologist for the same patient studies (C) shown for reference. (D) There was higher agreement, in terms of 4-class κ , with the algorithm than with the original interpreting radiologist from the DMIST trial for both crowdsourcing participants who read mammograms and those who do not. P-values are denoted by * $p < .001$

Table 1.

Breakdown of data format by digital mammograph system.

12 Bit Monochrome 1	12 Bit Monochrome 2	14 Bit Monochrome 1
Senoscan (99.9%) Kodak Lumiscan 75 (.1%)	Senograph (93.8%) Other (6.1%) Mammo-Clinical (.1%)	Senograph (94.1%) Mammo-Clinical (5.9%)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Summary of demographics in the patient cohort with regard to age, sex, race, and breast density.

	Training (n = 62316)	Validation (n = 6978)	Testing (n = 38936)
Age (median years, IQR)	46 (53–61)	46 (53–61)	47 (53–61)
Female (%)	100	100	100
Race			
White	50414	5622	30845
Black or African American	8389	925	5733
Hispanic or Latino	2273	289	1416
Asian	819	62	633
American Indian or Alaska	63	8	19
Other or Unknown	358	72	290
Radiologist-assessed breast density			
Fatty	6980 (11.2%)	873 (12.5%)	4575 (11.8%)
Scattered	27733 (44.5%)	2985 (42.8%)	17191 (44.2%)
Heterogeneously dense	23987 (38.5%)	2753 (39.5%)	14585 (37.5%)
Extremely dense	3616 (5.8%)	367 (5.3%)	2585 (6.6%)

Table 3.

Demographics of participants of the crowdsourcing assessment.

	N
Experience	
Radiologist (Breast)	3
Radiologist (Other)	10
Resident	2
Student	2
Read Mammograms	
No	10
Yes	7

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript