

---

# Breast Mass Classification from Mammograms using Deep Convolutional Neural Networks

---

**Daniel Lévy, Arzav Jain**

Stanford University

{danilevy, ajain}@cs.stanford.edu

## Abstract

Mammography is the most widely used method to screen breast cancer. Because of its mostly manual nature, variability in mass appearance, and low signal-to-noise ratio, a significant number of breast masses are missed or misdiagnosed. In this work, we present how Convolutional Neural Networks can be used to directly classify pre-segmented breast masses in mammograms as benign or malignant, using a combination of transfer learning, careful pre-processing and data augmentation to overcome limited training data. We achieve state-of-the-art results on the DDSM dataset, surpassing human performance, and show interpretability of our model.

## 1 Introduction

According to the International Agency for Research on Cancer, breast cancer accounts for 22.9% of invasive cancers and 13.7% of cancer-related deaths in women worldwide [5]. In the U.S., 1 in 8 women is expected to develop invasive breast cancer over the course of her lifetime [7]. Routine mammography is standard for preventive care and detection of breast cancer before biopsy. However, it is still a manual process, prone to human error due to high variability in mass appearance [2] and low signal-to-noise ratio, and thus can cause unnecessary biopsies or worse, missed malignant masses. Furthermore, efficacy is often highly correlated with radiologist expertise and workload [10].

Convolutional Neural Networks (CNN) have achieved impressive results on computer vision tasks spanning classification [16], object detection [11], and segmentation [17]. For breast mass diagnosis, deep learning techniques have been explored [1, 6, 8, 9, 24], but never in a fully end-to-end manner (directly classifying from pixel space) because of the scarcity of available data and lack of interpretability. In this work, we successfully train end-to-end CNN architectures to directly classify breast masses as benign or malignant. We obtain state-of-the-art results using a combination of transfer learning, careful pre-processing and data augmentation. We analyze the effects of these modelling choices and furthermore show how we can provide interpretability to the model’s predictions.

## 2 Related Work

While medical images differs significantly from natural images, traditional feature engineering techniques from computer vision such as scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG) have seen use and success when applied to medical images. More recently, deep learning-based approaches using CNNs have begun to achieve impressive performance on medical tasks such as chest pathology identification in X-Ray and CT [3, 23], and thoraco-abdominal lymph node detection and interstitial lung disease classification [20].

In the context of mammography, [8, 9] detect breast masses using a combination of R-CNN and random forests. Multiple works tackle the problem of breast lesion classification, but typically adopt a multi-stage approach. [24] extracts hand-engineered semantic (such as calcification) and textual features, and [6] classifies a full mammogram by extracting features from each view of the breast

and combining them to output a prediction. [1] performs extensive pre-processing using domain knowledge before training a CNN. To the best of our knowledge, ours is the first work to directly classify pre-detected breast masses using CNN architectures.

### 3 Dataset

In our experiments, we use the Digital Database for Screening Mammography (DDSM) [4], a collaboratively maintained public dataset at the University of South Florida. It comprises approximately 2500 studies each containing both mediolateral oblique (MLO) and craniocaudal (CC) views of each breast. Each image is grayscale and accompanied by a mask specifying the region of the pre-segmented mass if present. Examples of benign and malignant masses are shown in Fig. 1.

We consider only mammograms which contain masses, resulting in 1820 images from 997 patients. We split these randomly by patient into training, validation and testing sets (80%, 10% and 10% of the full dataset), constraining the validation and test sets to be balanced.

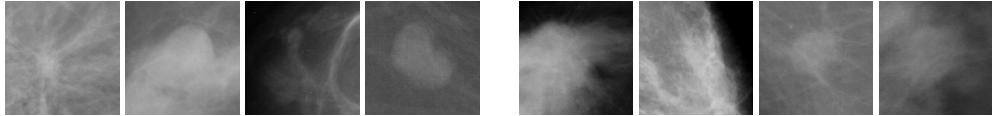


Figure 1: Benign (Left) and malignant (Right) breast masses from the dataset.

### 4 Methods

We train three different CNN architectures for breast mass classification, and analyze the effect of a number of model choices. We describe these below.

#### 4.1 CNN architectures

We evaluate three network architectures: a shallow CNN (the baseline model), an AlexNet [16] and a GoogLeNet [22]. For both the AlexNet and GoogLeNet, we use the same base architecture as the original works but replace the last fully-connected (FC) layer to output 2 classes. We also remove the two auxiliary classifiers from the GoogLeNet which we found impaired our training in practice.

The baseline model's architecture is inspired by the early layers of AlexNet [16]. We additionally use batch normalization [13]. The network takes a  $224 \times 224 \times 3$  image as input. It consists of 3 convolutional blocks composed of  $3 \times 3$  Convolutions - Batch Norm - ReLU - Max Pooling, with respectively 32, 32 and 64 filters each, followed by 3 FC layers of size 128, 64 and 2. The final layer is a soft-max layer for binary classification. We use ReLU activation functions, Xavier [12] weight initialization, and the Adam [15] update rule with a base learning rate of  $10^{-3}$  and batch size 64.

#### 4.2 Aspects of model analysis

**Transfer learning** We analyze the effect of initializing networks with pre-training on the ImageNet dataset [19], and then fine-tuning on mammography images. For the AlexNet model, we initialize the convolutional layers with pre-trained weights and a smaller learning rate multiplier of 0.1, and randomly initialize the 3 FC layers. For the GoogLeNet, we use the same weight initialization scheme. We use a learning rate multiplier of 0.1 for the layers before the Inception\_5a module, 1 for the Inception\_5a and Inception\_5b modules, and 10 for the last FC layer for more aggressive learning.

We train the AlexNet with Adam, base learning rate  $10^{-3}$ , and dropout rate 0.5. We train the GoogLeNet with Vanilla SGD, base learning rate  $10^{-2}$ , and dropout rate 0.2.

**Mass context** The area surrounding a mass provides useful context for diagnosis. We explore two approaches for providing the network with context. In the first, our input to the network is the region including 50 pixels of fixed padding around the mass, providing a context size independent of mass dimensions (referred to as **Small Context**). In the second approach we use proportional padding by extracting a region two times the size of the mass bounding box (referred to as **Large Context**).

**Data Augmentation** We study the impact of data augmentation to alleviate to the relatively small size of our training set, which is characteristic of many medical image datasets. We use rotation, cropping, and mirroring transformations to increase the effective size of our dataset (referred to as **Aug**). For each training image, we perform 5 random rotations and sample 5 random crops per rotation offline, effectively increasing training set size by a factor of 25. We also perform random mirroring at train time. These augmentations are justified since masses have no inherent orientation and their diagnosis is invariant to these transformations.

## 5 Results

We first present empirical analysis of our model design using the AlexNet base architecture, then show quantitative results of our best models. Finally we use techniques for visualizing saliency maps to provide interpretability of the model. All experiments are implemented with Caffe [14], and the analysis and results are presented on the validation and test sets, respectively.

### 5.1 Empirical analysis

**Effectiveness of transfer learning** CNN-based image representations learned on large-scale annotated datasets have proven to be a useful form of pre-training that can be effectively transferred to other computer vision tasks with limited training data [18]. More recently, low-level features learned from natural images have shown to be effective for medical image classification [3, 20, 23]. In Table 1, we strongly confirm this claim by demonstrating that a fine-tuned AlexNet significantly outperforms our baseline model.

Model	Validation Accuracy
Baseline(Aug-Large Context)	0.66
AlexNet(Aug-Large Context)	<b>0.90</b>

Table 1: Effectiveness of transfer learning.

**Influence of context** To understand the influence of context around masses, we fine-tune an AlexNet on two different datasets - one with fixed padding and the other with proportional padding. The results in Table 2 show that proportional padding contains greater signal for classification, and we consequently use this for the remaining experiments.

Model	Validation Accuracy
AlexNet(No Aug-Small Context)	0.64
AlexNet(No Aug-Large Context)	<b>0.71</b>

Table 2: Influence of context around the breast mass on the model performance.

**Influence of data augmentation** Limited amounts of training data is a common bottleneck in machine learning applications to medical problems. We evaluate the utility of data augmentation schemes to increase the effective amount of training data and reduce overfitting. The training loss curves in Fig. 2 show that our data augmentation technique described in Sec. 4.2 successfully regularises the network and helps remedy the scarcity of data.

### 5.2 Performance

Our final results using the model choices described in Sec. 5.1 and all base architectures are presented in Table 3. The GoogLeNet outperforms the other models by a fair margin. It is also more suited for fine-tuning and less prone to overfitting due to its relatively small number of parameters, approximately 5 million compared to 100 million for AlexNet.

An important metric for diagnostic applications is maximizing recall, since the cost of a false negative (patient remaining undiagnosed) is much higher than a false positive (an additional biopsy). Our best model achieves 0.934 recall at 0.924 precision, outperforming human performance in a study that

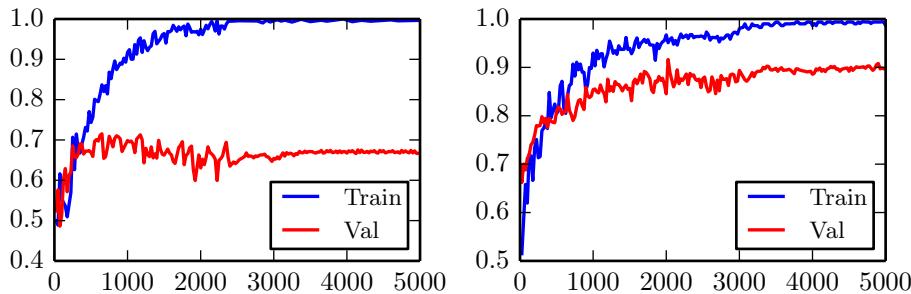


Figure 2: Accuracy curves when training on the unaugmented (Left) vs. augmented dataset (Right). x-axis is the iteration number.

shows radiologist recall between 0.745 and 0.923 [10]. This result is very promising for real-life use of such models in clinical practice.

Model	Accuracy	Precision	Recall	# Epochs
Baseline (Aug-Large Context)	0.604	0.587	0.703	35
AlexNet (Aug - Large Context)	0.890	0.908	0.868	30
GoogLeNet (Aug - Large Context)	<b>0.929</b>	<b>0.924</b>	<b>0.934</b>	30

Table 3: Summary of performance on the test set.

### 5.3 Interpretability

Deep learning models often lack interpretability and as such are hard to adopt for practical use in medical settings. [21] describe a methodology to visualize saliency maps which show the regions of an image the network is sensitive to when making predictions. This is performed by computing the gradient of the image with respect to the unnormalized class scores. Regions with larger gradient indicate higher contribution to the prediction (brighter in Fig. 3). Both the AlexNet and GoogLeNet learn to attend to the edges of the mass, which is a high-signal criterion for diagnosis, while also paying attention to context.

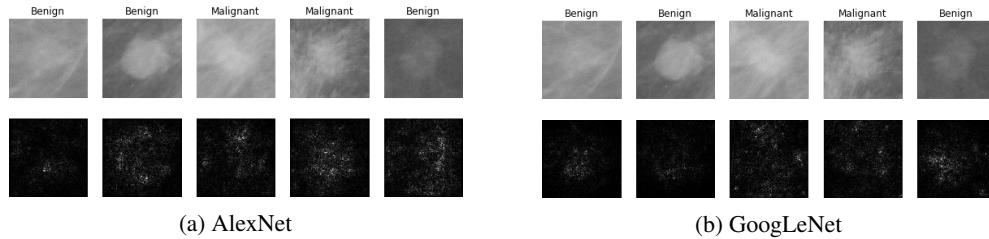


Figure 3: Saliency maps for our best AlexNet and GoogLeNet on five images from the validation set.

## 6 Conclusion

In this work, we propose an end-to-end deep learning model to classify pre-detected breast masses from mammograms. We show how careful pre-processing, data augmentation and transfer learning can overcome the data bottleneck common to medical computer vision tasks, and additionally provide a method to give more interpretability to network predictions.

Our approach obtains state-of-the-art results, outperforming trained radiologists, and the interpretability enables more comfortable adoption in real-world settings. Future work includes exploring other architectures, and integration of attention mechanisms which are more difficult to train but could provide even more concrete interpretability.

## Acknowledgements

We thank Justin Johnson for continuous feedback and guidance throughout this project as well as Serena Yeung for insightful comments on the draft. The authors also acknowledge the support of AWS Educate program for generously providing free instances with GPUs.

## References

- [1] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine*, 127:248–257, 2016.
- [2] J. E. Ball and L. M. Bruce. Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4973–4978. IEEE, 2007.
- [3] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan. Chest pathology detection using deep learning with non-medical training. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 294–297. IEEE, 2015.
- [4] K. Bowyer, D. Kopans, W. Kegelmeyer, R. Moore, M. Sallam, K. Chang, and K. Woods. The digital database for screening mammography. In *Third international workshop on digital mammography*, volume 58, page 27, 1996.
- [5] P. Boyle, B. Levin, et al. *World cancer report 2008*. IARC Press, International Agency for Research on Cancer, 2008.
- [6] G. Carneiro, J. Nascimento, and A. P. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer International Publishing, 2015.
- [7] C. DeSantis, J. Ma, L. Bryan, and A. Jemal. Breast cancer statistics, 2013. *CA: a cancer journal for clinicians*, 64(1):52–62, 2014.
- [8] N. Dhungel, G. Carneiro, and A. P. Bradley. Automated mass detection in mammograms using cascaded deep learning and random forests. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, pages 1–8. IEEE, 2015.
- [9] N. Dhungel, G. Carneiro, and A. P. Bradley. Deep learning and structured prediction for the segmentation of mass in mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–612. Springer International Publishing, 2015.
- [10] J. G. Elmore, S. L. Jackson, L. Abraham, D. L. Miglioretti, P. A. Carney, B. M. Geller, B. C. Yankaskas, K. Kerlikowske, T. Onega, R. D. Rosenberg, et al. Variability in interpretive performance at screening mammography and radiologists’ characteristics associated with accuracy 1. *Radiology*, 253(3):641–651, 2009.
- [11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [12] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.
- [13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.
- [15] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.
- [17] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
- [18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.
- [19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [20] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.
- [21] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [23] B. van Ginneken, A. A. Setio, C. Jacobs, and F. Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 286–289. IEEE, 2015.
- [24] J. Wang, X. Yang, H. Cai, W. Tan, C. Jin, and L. Li. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific reports*, 6:27327, 2016.