Uncertainty Measurements for the Reliable Classification of Mammograms

Mickael Tardy[1,2(✉)], Bruno Scheffer[3], and Diana Mateus[1]

[1] Ecole Centrale de Nantes, LS2N, UMR CNRS 6004, Nantes, France
{mickael.tardy,diana.mateus}@ec-nantes.fr
[2] Hera-MI, SAS, Nantes, France
[3] Institut de cancérologie de l'Ouest, Nantes, France

Abstract. We propose an efficient approach to estimate the uncertainty of deep-neural network classifiers based on the tradeoff of two measurements. The first based on subjective logic and the evidence of soft-max predictions and the second, based the Mahalanobis distance between new and training samples in the embedding space. These measurements require neither modifying, nor retraining, nor multiple testing of the models. We evaluate our methods on different classification tasks including breast cancer risk, breast density, and patch-wise tissue type and considering both an in-house database of 1600 mammographies, as well as on the public INBreast dataset. Throughout the experiments, we show the ability of our method to reject the most evident outliers, and to offer AUC gains of up to 10%, when keeping 60% of most certain samples.

Keywords: Uncertainty · Classification · Deep learning · Mammography · Breast cancer

1 Introduction

The risks of erroneous decisions are especially high when developing computer-aided systems for medical decision support. Therefore, there has been a recent interest in measuring the uncertainty of deep-learning-based predictions. In computer vision, the Out-of-distribution (OOD) detection task [5] aims at identifying whether or not a new test image belongs to the train in-distribution (ID) and can thus be classified with certainty. Current OOD benchmarks rely on public datasets that come from distinct data distributions (e.g. MNIST vs. CIFAR). In medical image analysis, the OOD detection is important but challenging because the differences between the data distributions used for training and testing are often subtle, for instance, due to variability in acquisition parameters, machine, or inclusion conditions for the patients. In addition to the in/out distribution

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-32226-7_55) contains supplementary material, which is available to authorized users.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11769, pp. 495–503, 2019.
https://doi.org/10.1007/978-3-030-32226-7_55

uncertainty, we are confronted with noisy data (*aleatoric uncertainty*), and a limited knowledge of the underlying phenomena (*model or epistemic uncertainty*), resulting from the scarcity of annotated datasets.

In this work, our goal is to provide a measure of uncertainty allowing us to identify potentially erroneous classifications, whether they come from a data uncertainty or a distribution shift. Similar to [9], the amount of tolerated uncertainty will result in a trade-off between the number of retained images and the level of accuracy. In the quest of generality, we propose combining two uncertainty measurements which neither require modifying the classification model nor re-training it with a modified loss function. The first one, based on subjective logic [6], exploits information from the predicted classification probabilities, while the second, inspired from [2], defines the region within the feature space around the known training data samples which is considered as certain.

We demonstrate the interest of our approach for different breast imaging classification tasks namely, risk assessment (high vs. low risk), breast density stratification according to BI-RADS scores, and glandular vs. conjunctive patch-tissue classification. We evaluate our method on in-house and public datasets [12] and demonstrate that our technique can effectively detect error-prone images while increasing the reliability of the retained predictions (in terms of the accuracy). For the completeness of our study, we also compare to the state-of-the-art methods [4,15]. To the best of our knowledge, we are the first to propose such uncertainty measurements for classification tasks in breast imaging analysis.

Related Work: Usually, for a given sample, a deep learning classifier yields probabilities of belonging to the given classes. Hendrycks et al. [5] established a measure of uncertainty directly from the class probabilities without any further modification or training of the model. Liang et al. [10] pushed this idea forward by proposing an additional adversarial perturbation and soft-max scaling. Similar to [5,10], our first uncertainty measurement also exploits the softmax output but interpreted through the lens of subjective logic [6].

Recently, several approaches have been proposed based on a Bayesian formulation of the uncertainty. Bayesian networks are well suited for isolating different sources of uncertainty but have an inherent high complexity. Approximations like Monte-Carlo dropout [4] have been leveraged to propose practical uncertainty estimates [7] which has been successfully used in different classification and segmentation tasks [1,3,13]. These methods, however, require the modification of the training to include dropout layers (if not present) and multiple runs during the test. Also, following a Bayesian approach, several recent works model the output of a deep network with a Dirichlet distribution [11,15]. By designing uncertainty-aware loss functions and through variational optimization, these approaches allow extracting uncertainty measurements from a unique run. However, they are still not generalizable to pre-trained models.

The third line of approaches [2,8] uses the Mahalanobis distance in the feature space (produced by the network embedding) to define a region of certainty and evaluate how far a new sample is from the known dataset. Considering the above state-of-the-art techniques and fixing as objective the practicality of implementation, we propose an efficient yet affordable method to measure the uncertainty, which combines a subjective logic interpretation of the soft-max outputs as well as the Mahalanobis Distance.

2 Methods

Let $\mathcal{X} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ be a training dataset composed of images $\mathbf{x}_i$ and class labels $\mathbf{y}_i \in \{\mathcal{C}_k\}_{k=1}^K$. Consider a classifier h that assigns to an input $\mathbf{x}_i$ a class probability vector $\hat{\mathbf{p}} = h(\mathbf{x}_i)$, where $p_{ik} \in \hat{\mathbf{p}}_i$ denotes the probability of $\mathbf{x_i}$ to belong to class $\mathcal{C}_k$. Then, suppose the classifier can be decomposed into two steps, where the first computes a feature representation $\mathbf{z}_i = g(\mathbf{x_i})$ and the second, estimates the class probabilities $\hat{\mathbf{p}} = f(\mathbf{z}_i)$. The classifier can be a deep neural network trained end-to-end, where the $g(\cdot)$ corresponds to the penultimate layer and $f(\cdot)$ stands for the soft-max. Our goal is to determine an uncertainty measurement v for each prediction of h. By defining a tolerated amount of uncertainty th_v we should be able to detect and put aside uncertain test samples while increasing the expected performance of the classifier.

In this work, we consider a combination of two uncertainty measurements $v = [u, D_{\mathrm{m}}]$. First, a prediction uncertainty $u(\mathbf{p}) : \mathbf{p} \mapsto \mathbb{R}$, based on the information contained from the probabilistic predictions. Second, a data closeness measurement $D_{\mathrm{m}}(\mathbf{z}) : \mathbf{z} \mapsto \mathbb{R}$ following a Mahalanobis approach [2] that measures the distance D_{m} of a sample to the training distribution cluster.

The **prediction uncertainty** u builds on recent works interpreting the maximum predicted probability [5], or the entropy of the probabilistic predictions [11,13] as a measure of uncertainty. However, inspired from [15] we rely on Subjective Logic [6], a formalization of Dempster-Shafer evidence theory to facilitate a direct interpretation of the uncertainty values. While Malinin et.al. [11] argue that the Dirichlet Loss function is required to induce a meaningful notion of uncertainty, we show as in [5], that the output of a classifier network trained with a soft-max layer and a cross-entropy loss still has practical value for uncertainty estimation. Formally, for K classes we have:

$$u + \sum_{k=1}^{K} b_k = 1, \quad b_k = \frac{e_k}{S}, \quad u = \frac{K}{S}, \quad S = \sum_{k=1}^{K} (e_k + 1), \tag{1}$$

where u is the sought uncertainty, b_k is the belief for the class k and e_k is the evidence provided by the network for the class k. Having $e_k + 1 = \exp^{f(x)}$ we obtain the uncertainty estimate:

$$u(\mathbf{x}) = \frac{K}{\sum_{k=1}^{K} \exp^{f(\mathbf{x})}} \tag{2}$$

The use of subjective logic requires particular attention to the logits' scale. From Eq. 1 we have $u \in [0, 1]$, with $u_{max} = 1$ corresponding to the case with no evidence. With Eq. 2 and the logits $f(\mathbf{x}) \in [-\infty, +\infty]$, we may have computational issues for large values of $f(x)$. To avoid this phenomenon, logits are rescaled, or saturated for instance to $\exp(f(x)) \in [0, 2 \cdot 10^{12}]$.

The second considered uncertainty measurement is the **Mahalanobis distance** [2] calculated from a given sample to the known distribution, as:

$$D_M(\mathbf{x}) = \sqrt{(g(\mathbf{x}) - \mu)^T \Sigma^{(-1)} (g(\mathbf{x}) - \mu)}, \tag{3}$$

where $g(\mathbf{x})$ is the output of the model's penultimate layer for a sample $\mathbf{x}$, μ and Σ are the mean and the covariance matrix of the cluster of all points in the training dataset $\mathcal{X}$, once mapped to the embedding space through function $g()$.

Although the entropy and related measurements on the posterior probabilities are well-known to be related uncertainty, we have observed that the Mahalanobis distance brings a complementary aspect especially related to out-of-distribution cases [2]. For instance, when a classifier trained on breast images (ID) is fed with outliers from a flower dataset (OOD), we see that the rejection criterion based on the Mahalanobis distance is quite effective (See Fig. 1-left). In a situation where we artificially generate a linear transition from an ID patch to an OOD patch (Fig. 1-right) for a binary classification problem[1], we observe a similar behavior. The efficiency of the uncertainty u is obvious at the middle of the transition corresponding to a mix between an ID and the OOD patch. However, the uncertainty fails to rise after this point to indicate that the prediction of the pure OOD patch is wrong. In contrast, the Mahalanobis distance is more representative towards the OOD patch indicating an uncertain prediction.

Following the potential complementarity of the two estimates, we propose to simultaneously consider thresholds on both uncertainty measures, in order to reject uncertain predictions using $u > th_u$ as well as data points that are too far from the certain ID region $D_M > th_D$.

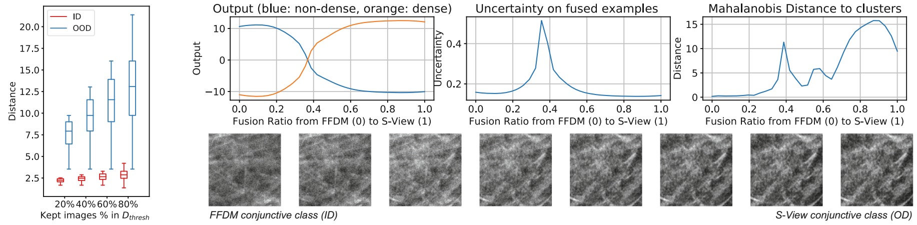

Fig. 1. Left: Toy example with OOD coming from the Flowers database **Right**: prediction probabilities (output) and variation of the uncertainty u and distance D_m measurements for the linear transition between an ID and an OOD patches.

3 Experimental Validation

To evaluate the performance of our method, we performed experiments targeting three mammography image analysis problems: risk classification, breast density

[1] Model and patches from the $TissueCLS_{raw}$ experiment described in Sect. 3.

classification, and a patch-wise tissue characterization task. For the three problems, we study the performance of the classifiers while changing uncertainty tolerance thresholds and thus the ratio of test images kept. In particular, (i) we show the precision at several cut-off values of the ratio of kept images (90%, 60%); (ii) we study the AUC and AUCPR of the predictions, and (iii) we analyze the statistics of u and D_M in the retained ID and OOD samples (see Fig. 4).

RiskCLS. We devise this experiment intending to show the generality and performance of our method on public models and databases. We focus on the image-wise risk classification according to Assessment Categories (ACR), where ACR1-2 stand for low-risk (negative) and ACR4-6 represent high risk (positive) cases. To create a basis for comparison, we rely on the VGG-based CNN model from [16], pre-trained on the DDSM (Digital Database for Screening Mammography) database. As in [16], we perform fine-tuning of the model using a second open dataset (INBreast) [12] taking 80 images for fine-tuning and keeping 305 for validation. We evaluate our method with ($RiskCLS_{tune}$) and without the fine-tuning ($RiskCLS_{init}$) step to show the behavior of the uncertainty measurements for the samples from the shifted INBreast distribution, either when it is completely or only partially unknown.

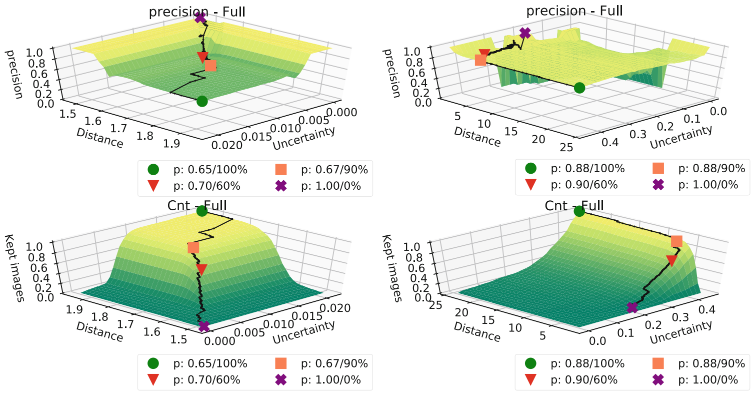

Fig. 2. Precision and ratio of kept images in the u and D_M space: without ($RiskCLS_{init}$, on the left) and with ($RiskCLS_{tune}$, on the right) fine-tuning. The legends list the precision associated to different cut-offs of the kept image ratio. (Color figure online)

In Fig. 2 we show the precision (top) and ratio of images kept (bottom) for different values of uncertainty th_u and distance thresholds th_D. We also plot in black the optimal path for the studied test dataset (thresholds that maximize the precision for a decreasing ratio of images kept) and highlight the performance at several cut-off points (colored shapes).

We observe the increase of precision between $RiskCLS_{init}$ and $RiskCLS_{tune}$ models (0.65 vs. 0.88) with 100% of the data produced by the fine-tuning step.

By retaining only the 60% most certain predictions, the performance increases respectively by +5% and +2%. Note that without fine-tuning both uncertainty measurements are equally important for defining the optimal performance path. The effect of the Mahalanobis distance is reduced with fine-tuning since the shape of the distribution cluster changes and, thus, the distances of test samples towards the center of the cluster become shorter.

DensityCLS. The second experiment targets the 4-class image-wise classification of breast density based on the 4th edition of BI-RADS. The goals here are (i) to evaluate our approach when dealing with multi-class classification and (ii) to challenge it with the real-life scenario of images from a distribution shift caused by images from different manufacturers. We use the VGG-based model from [14]. The in-house training set consists of 1232 images from a Planmed Nuance Excel (PNE) mammography system. For validation, we rely on 370 PNE images as well as on 370 images from Siemens MammoNovation (INBreast [12]).

In Fig. 3, we evaluate the precision for the full test set, as well as for the ID and OOD parts separately. For the ID dataset, a significant performance improvement (+8%) is obtained retaining 60% of the data. However, for the OOD dataset, without any fine-tuning, the performance is low despite the uncertainty checks. This result shows the limits of our method for subtle distribution shifts.

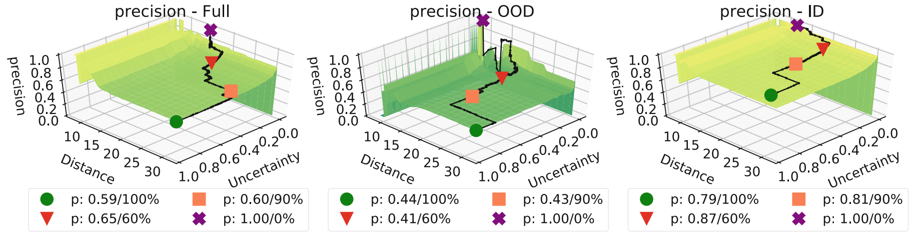

Fig. 3. Precision of kept images in the u and D_M spaces for the $DenseCLS_{raw}$. The legends list the precision associated to different cut-offs of the kept image ratio.

TissueCLS. Our final experiment is focused on the patch-wise classification of image-patches into dense and non-dense tissues. The goal of this experiment is twofold: (i) to measure the effect of a distribution shift between native 2D Full-Field Digital Mammography (FFDM) images and 2D views synthesized from 3D tomosynthesis acquisitions, which is of great clinical interest; (ii) to compare our method to state-of-the-art approaches. For the training, we used a dataset of patches from FFDM images (pixel spacing 50 µm). For validation, the ID patches came from FFDM images and OOD patches from S-View images (pixel spacing 98 µm).

Figure 4-left shows the smooth improvement of the ROC curves for a decreasing amount of kept images selected with optimal th_u and th_D. When analyzing the actual values of u and D_M on the ID and OOD samples separately (Fig. 4-right), we see that the threshold on Mahalanobis distance is more critical for the first rejected samples (from 100% up to 70%) while the effect of the uncertainty comes after (from 70%), illustrating once more their complementarity.

Finally, we compare our approach against two state-of-the-art methods using the same network architecture in all three experiments. The first consists of an MC dropout approach [4], that adds dropout layers to the existing model and keeps them active during test time to collect the variance of the predictions over different runs (here 10). The variance is then used as the uncertainty measurement. The second method results from training the same model with the Dirichlet distribution loss function from [15]. From the results reported in Table 1, we see that our approach is very competitive, while neither requiring model changes nor additional training. Gal's method [4] performs better at baseline (100%) due to the dropout training, but it is at most comparable when considering uncertainty sample pruning (90% and 60%) while requiring redesign, retraining, and multiple test runs. We also note that softmax probabilistic predictions (u_{prob}) and the entropy (u_{entr}) may be used as uncertainty alternatives with similar results. However, Subjective Logic (Eq. 1) remains competitive with the advantage of yielding directly interpretable uncertainty and belief values.

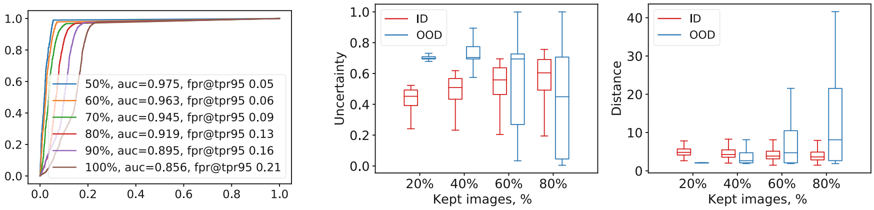

Fig. 4. TissueCLS experiment. **Left**: ROC curves with kept images ratio, AUC and FPR@TPR95, **Right**: statistics of u and D_M among the retained samples, for an increasing amount of kept images.

Table 1. Precision, AUC and AUCPR of different models on the thresholded datasets. Cut-offs of 100%, 90%, 60% images are reported.

Model	Precision			AUC			AUCPR		
	100%	90%	60%	100%	90%	60%	100%	90%	60%
Gal [4]	0.74	0.75	0.81	**0.89**	0.89	0.87	**0.87**	**0.84**	0.72
Sensoy [15]	0.87	0.89	0.93	0.87	0.90	0.93	0.81	0.82	0.83
Ours $u_{prob} + D_M$	**0.89**	**0.90**	**0.94**	0.88	**0.90**	**0.96**	0.81	**0.84**	**0.90**
Ours $u_{entr} + D_M$	**0.89**	**0.90**	**0.95**	0.88	**0.91**	**0.96**	0.81	**0.84**	**0.91**
Ours $u_{SL} + D_M$	**0.89**	**0.90**	**0.95**	0.86	**0.90**	**0.96**	0.75	0.81	**0.91**

4 Discussion and Conclusion

In the context of mammography image classification problems, we have studied the problem of uncertainty measurement, aiming to define a method capable of differentiating certain from uncertain predictions, and thus increasing the safety of CAD system suggestions. Uncertainty measurements based on the probability predictions and the Mahalanobis distance have been shown to be effective tools towards this end.

With the proposed combination of the two measurements we have demonstrated that it is possible to detect evident out-of-distribution samples (as the flowers) while achieving more moderate improvements of performance for subtle forms of distribution shift (e.g. scanned films vs FFDM or FFDM of different manufacturers). In these cases, our method deployed on a validation dataset may be useful to detect the effectiveness of augmentation and fine-tuning strategies when dealing with small datasets.

With respect to the uncertainty measure based on the probabilistic predictions, the scale of the logits used for the estimate u is worthy of attention: when using subjective logic a rescaling may be needed. However, we showed, that entropy or probability may yield similar results (see Table 1). A limitation of Mahalanobis distance is that it requires having access to the training dataset in order to compute the covariance matrix, which may not always be possible. Also, despite the effectiveness of the combination further research is required on automatic ways to find the optimal thresholds.

Finally, the effectiveness of our method has been shown in several mammography classification tasks. Given that no changes in the model nor retraining are required, our findings can be easily generalized to other medical image analysis problems confronted to uncertainties coming from the data but also from the distribution shifts.

References

1. Bragman, F.J.S., et al.: Uncertainty in multitask learning: joint representations for probabilistic MR-only radiotherapy planning. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_1
2. Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., Vernekar, S.: Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. CoRR abs/1812.02765 (2018)
3. Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J.: Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 691–699. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_78
4. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)

5. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. CoRR abs/1610.02136 (2016)
6. Jøsang, A.: Subjective Logic: A Formalism for Reasoning Under Uncertainty. Springer (2018, Incorporated)
7. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30, pp. 5574–5584. Curran Associates, Inc. (2017)
8. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 7167–7177. Curran Associates, Inc. (2018)
9. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. **7**(1), 17816 (2017)
10. Liang, S., Li, Y., Srikant, R.: Principled detection of out-of-distribution examples in neural networks. CoRR abs/1706.02690 (2017)
11. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 7047–7058. Curran Associates, Inc. (2018)
12. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: INbreast: toward a full-field digital mammographic database. Acad. Radiol. **19**(2), 236–248 (2012)
13. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2018, pp. 655–663. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_74
14. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. **8**(1), 4165 (2018)
15. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 3183–3193. Curran Associates, Inc. (2018)
16. Shen, L.: End-to-end training for whole image breast cancer diagnosis using an all convolutional design. arXiv preprint arXiv:1708.09427, November 2017