PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Evaluation of different convolutional neural network encoder-decoder architectures for breast mass segmentation

Isosalo, Antti, Mustonen, Henrik, Turunen, Topi, Ipatti, Pieta, Reponen, Jarmo, et al.

Antti Isosalo, Henrik Mustonen, Topi Turunen, Pieta S. Ipatti, Jarmo Reponen, Miika T. Nieminen, Satu I. Inkinen, "Evaluation of different convolutional neural network encoder-decoder architectures for breast mass segmentation," Proc. SPIE 12037, Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications, 120370W (4 April 2022); doi: 10.1117/12.2628190

SPIE.

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Evaluation of different convolutional neural network encoder-decoder architectures for breast mass segmentation

Antti Isosalo[a,*], Henrik Mustonen[a,*], Topi Turunen[a], Pieta S. Ipatti[b], Jarmo Reponen[a,c], Miika T. Nieminen[a,b,c], and Satu I. Inkinen[a,d]

[a]Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu, Finland
[b]Department of Diagnostic Radiology, Oulu University Hospital, Oulu, Finland
[c]Medical Research Center Oulu, University of Oulu and Oulu University Hospital, Oulu, Finland
[d]Department of Radiology, HUS Diagnostic Center, Helsinki University and Helsinki University Hospital, Helsinki, Finland

ABSTRACT

In this work, we study convolutional neural network encoder-decoder architectures with pre-trained encoder weights for breast mass segmentation from digital screening mammograms. To automatically detect breast cancer, one fundamental task to achieve is the segmentation of the potential abnormal regions. Our objective was to find out whether encoder weights trained for breast cancer evaluation in comparison to those learned from natural images can yield a better model initialization, and furthermore improved segmentation results. We applied transfer learning and initialized the encoder, namely ResNet34 and ResNet22, with ImageNet weights and weights learned from breast cancer classification, respectively. A large clinically-realistic Finnish mammography screening dataset was utilized in model training and evaluation. Furthermore, an independent Portuguese INbreast dataset was utilized for further evaluation of the models. 5-fold cross-validation was applied for training. Soft Focal Tversky loss was used to calculate the model training time error. Dice score and Intersection over Union were used in quantifying the degree of similarity between the annotated and automatically produced segmentation masks. The best performing encoder-decoder with ResNet34 encoder tailed with U-Net decoder yielded Dice scores (mean±SD) of 0.7677±0.2134 for the Finnish dataset, and ResNet22 encoder tailed with U-Net decoder 0.8430±0.1091 for the INbreast dataset. No large differences in segmentation accuracy were found between the encoders initialized with weights pre-trained from breast cancer evaluation, and of those from natural image classification.

Keywords: computer aided detection, deep learning, feature pyramid network, mass detection, image segmentation, transfer learning, U-net

1. INTRODUCTION

Breast cancer is the most commonly diagnosed cancer type, with approximately 2.3 million new cases yearly.[1] Mammography is commonly used for breast cancer screening as it is a cost-effective method for early detection and does not require invasive actions. Typically, two projection images from bilateral craniocaudal (CC) and mediolateral oblique (MLO) views are taken from both breasts.

When performing automated mammography image analysis for cancer diagnosis, an initial task is to detect the potentially abnormal regions. This task is not easy, as these abnormalities can encompass only a small area in the high-resolution breast images. Moreover, abnormal breast tissue appears with varying visual characteristics in terms of shape and texture. Breast abnormalities can roughly be categorized as architectural distortions, asymmetries, calcifications and masses. Several works have been dedicated to the automated characterization of these abnormalities. Among those which focus on segmentation of architectural distortions is Ben-Ari et al.[2],

Further author information: (Send correspondence to A. I.)
A. I.: E-mail: antti.isosalo@oulu.fi, Telephone: +358 294 480 000
* These authors contributed equally to this work

Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications,
edited by Thomas M. Deserno, Brian J. Park, Proc. of SPIE Vol. 12037,
120370W · © 2022 SPIE · 1605-7422 · doi: 10.1117/12.2628190

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

segmentation of calcifications Zamir et al.[3], mass detection Ribli et al.[4] and segmentation of masses Wang et al.[5] and Lou et al.[6], to name a few.

In the field of computer vision, pre-trained models, *e.g.*, trained on the ImageNet[7] dataset of natural images, have been shown to have a positive effect on various classification and segmentation tasks. Agarwal et al.[8] experimented with transfer learning and pre-trained weights initialization from representations learnt from natural images. They have investigated mass and non-mass areas detection using a semi-supervised image patch-based approach. Recently, Liu et al.[9] have studied very weakly-supervised segmentation from high-resolution mammograms using a network which can be trained using only image-level labels. This work is in close relation to Shen et al.,[10] which is among the first works to make trained models publicly available in the field of breast cancer detection. Their models are trained on the NYU Breast Cancer Screening Dataset which encompass 1,001,093 images from 141,472 patients.[11,12]

In this work we study two different convolutional neural network (CNN) encoder-decoder architectures with injected pre-trained encoder weights for the task of breast mass segmentation. Our objective is to find out whether pre-trained weights from breast cancer classification, namely Globally-Aware Multiple Instance Classifier (GMIC)[10] weights, can give a better standing point for the segmentation in comparison to those learned from natural images, namely ImageNet based pre-trained model weights. We use a subset of a clinically-realistic Finnish dataset in the model training and validation and further evaluate the models on a well-known Portuguese INbreast[13] dataset.

2. METHODOLOGY

2.1 Materials

2.1.1 Finnish dataset

Our Finnish dataset originates from mammography screening studies conducted over the 2011-2019 period within City of Oulu, Oulu, Finland. We later refer to this dataset as Oulu Dataset of Screening Mammography (OUDSM). A permit for a registry-based studies from the Northern Ostrobothnia Hospital District (179/2019), Finland, and the City of Oulu (35/2019), Finland, was obtained before initiating the data collection. Each study in the dataset contains digital mammograms in Digital Imaging and Communications in Medicine (DICOM) format and textual information about the study from the mammographic information system (MIS). Original size of the dataset after collection was 49,634 studies from 22,739 unique patients. Mammograms (2,934 studies) having co-reading assessment score greater than or equal to 3 in a five-level Finnish scale (1: normal, 2: benign, 3: malignancy cannot be excluded, 4: suspect for malignancy, 5: malignant), were labelled by radiology resident T. T. using a custom made MATLAB (2020a, MA, USA) based annotation tool.[14] Three groups of pixel-wise contours were drawn: malignant and benign masses, malignant and benign calcifications, and malignant and benign architectural distortions with the possibility to assign additional characterizations. Specifically, the mass masks were then utilized in this work for supervision of the training. Moreover, the labelled dataset was split into training set and holdout sets according to the Pareto Principle.

2.1.2 Portuguese dataset

Portuguese INbreast dataset originates form Centro Hospitalar de S. João, Breast Centre, Porto, Portugal.[13] The original size of the dataset is 117 studies from 108 unique patients. Each study in the dataset contains digital mammograms in DICOM format. For each study there are segmentation masks for calcifications and masses. The mass masks were utilized in this work. As the INbreast dataset files ship without parameters for windowing, all DICOM files were injected a proper Siemens value of interest lookup table (VOI LUT) to enable better contrast and compatibility for the experiments.

2.2 Post-Processing

As a post-processing step, VOI LUT mapping, *i.e.*, windowing operation described by the DICOM Standard, was performed to standardize the mammograms. All mammograms were saved as 16-bit PNG files. Moreover, mammograms were padded and resized for the experiments to 512-by-512 to retain the original aspect ratio of the imaged breasts. Segmentation masks were formatted accordingly.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

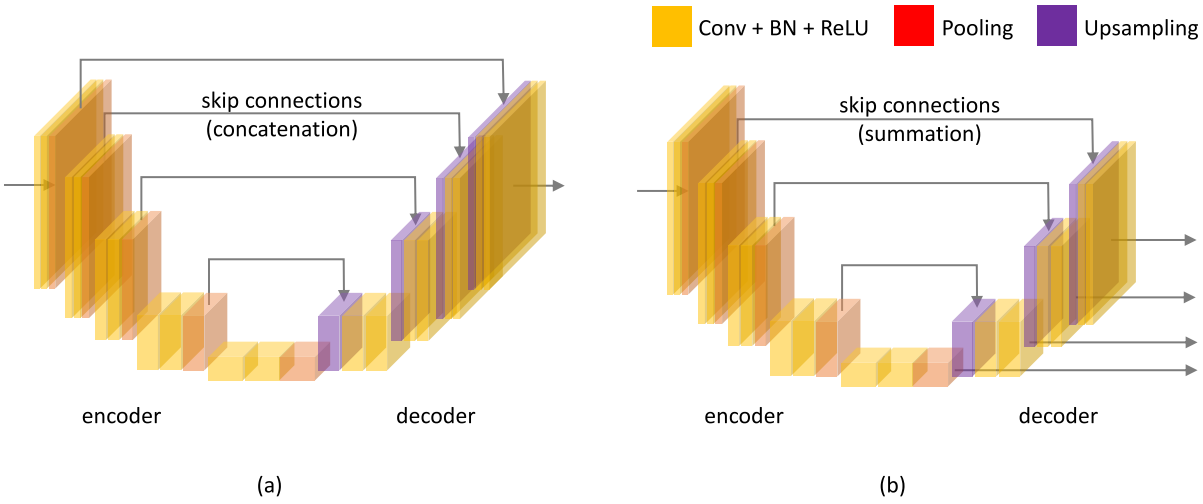

Figure 1. General principle diagrams of encoder-decoder architectures with (a) U-Net and (b) feature pyramid network (FPN) type decoders.

2.3 Network architectures

In this study, we utilize two convolutional neural network encoder-decoder architectures, namely U-Net[15] and Feature Pyramid Network[16] (FPN) (Fig. 1). Both architectures have a contracting encoder part that reduces spatial dimensions with every layer. Two different ResNet[17] encoders with injected pre-trained weights are evaluated. Pre-trained ResNet22 weights are adapted from Globally-Aware Aware Multiple Instance Classifier[10] (GMIC), trained originally with full-resolution mammograms. In addition, we use ResNet34 encoder which is initialized with ImageNet weights. The encoder is tailed with an upsampling decoder part restoring the spatial dimensions. As an output, a prediction for each pixel in the input image is produced. Skip connections between contracting and upsampling parts are for improved segmentation accuracy. FPN uses a method inspired by image pyramids to generate multiple predictions at different scales, which are then upsampled and concatenated. Spatial Dropout (PyTorch[18] Dropout2d, with rate 0.2) is used to drop entire 2D feature maps to alleviate overfitting (see also [19]). Futhermore, our models have Batch Normalization[20] (BN).

In total the ResNet34 encoder has 21,284,672 trainable parameters while the ResNet22 encoder has only 2,799,536 trainable parameters. The FPN decoder has 2,338,049 trainable parameters while the U-Net decoder has 4,773,025, therefore making the FPN decoder more memory efficient in terms of GPU memory.

2.3.1 Training details

Data augmentations: During the training randomly altered versions of data were generated on the fly to introduce regularization. The SOLT library (`https://github.com/MIPT-Oulu/solt` version 1.8.1.) was used. SOLT operates on 8-bit images. All augmentations had a 50% probability of happening (Table 1). No augmentations were used during inference (*i.e.*, when predicting segmentation masks).

For ResNet34 (with ImageNet weights) each single channel mammogram was replicated to have the input dimensions of 3-by-H-by-W, where H and W refer to height and width respectively. For ResNet22 the input dimensions were already 1-by-H-by-W, *i.e.*, single channel. Utilizing the augmentations (Table 1), mean and standard deviation were computed from the training set to normalize the input data for better convergence.

Loss functions: Two separate losses were applied to calculate the error between the predicted output and the pixel-wise reference segmentations. First, combination loss L^c of weighted sum of binary cross-entropy[22] and soft Jaccard[23] loss was applied as an initial loss, *i.e.*, for 20 epochs. Subsequently, the model was trained for 60 epochs using a soft Focal Tversky[24] loss. Best performing model based on validation loss was saved for inference. The combination loss for class c is defined as

$$L^c(w) = (1-\omega)BCE^c(w) - \omega J^c(w), \tag{1}$$

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Table 1. Augmentations and their corresponding parameter values or range used in the experiments

Augmentation	Parameters
Rotation (deg)	$[-3, 3]$
Scale	$[0.7, 1.3]$
Translation (px)	50
Random Crop (px)	$[448, 448]$
Flip	Horizontal
Gamma correction	$[0.5, 1.8]$
Brightness and contrast	$[30, 100]$
Salt and pepper noise	0.1
Gaussian noise	0.5
Gaussian blur (kernel, sigma)	$[3, 7, 11]$, $[1, 5]$
Median blur (kernel, sigma)	$[3, 7, 11]$, $[1, 5]$
Cutout[21]	20%

where w are the network parameters, BCE is the binary cross-entropy loss. Furthermore, soft Jaccard loss (J) in (1) is defined as

$$J^c(w) = \frac{1}{N} \times \sum_{i=1}^{N} \frac{g_{ic} p_{i\hat{c}} + \epsilon}{g_{ic} + p_{i\hat{c}} - g_{ic} p_{i\hat{c}} + \epsilon}, \tag{2}$$

where N is the total number of pixels, g_{ic} is the binary label for pixel i and class c, and $p_{i\hat{c}}$, is the predicted probability for pixel i and class c, ϵ is for numerical stability. In the experiments, the weight ω was set 0.5. The soft Focal Tversky[24] loss is defined as

$$FT^c(w) = [1 - TI^c(w)]^{\frac{1}{\gamma}}. \tag{3}$$

TI in (3) is the Tversky index[25,26]

$$TI^c = \frac{\sum_{i=1}^{N} p_{ic} g_{ic} + \epsilon}{\sum_{i=1}^{N} p_{ic} g_{ic} + \alpha \times \sum_{i=1}^{N} p_{i\hat{c}} g_{ic} + (1 - \alpha) \times \sum_{i=1}^{N} p_{ic} g_{i\hat{c}} + \epsilon}, \tag{4}$$

where p_{ic} is the probability that pixel i is of the class c, $p_{i\hat{c}}$ is the probability that pixel i is of the class $\hat{c}$, the background class, g_{ic} is the binary label for pixel i and class c, and $g_{i\hat{c}}$ is the binary label for pixel i and class $\hat{c}$, and ϵ is for numerical stability. In the experiments, coefficient α in (4) was set 0.7 and γ in (3) was set $\frac{4}{3}$ following [24]. Typically, there exists an imbalance between healthy breast tissue and masses represented by the segmentation masks. Both Jaccard index and Tversky index alleviate this issue.

Optimizer: For the optimizer, we used the Adam[27] with multi-step learning rate scheduler[18]. Optimizer weight decay was set at 1e-4. Learning rate was set at 1e-4 in the beginning. Learning rate was decayed utilizing a multiplicative factor of 0.1 after 40, 50 and 60 epochs.

Training: We used the Finnish dataset for training the model variants. We used training batch size of 32 and validation batch size of 32 for the 512×512 resolution inputs. Number of threads was 24. Furthermore, 5-fold cross-validation was implemented for the training using a K-fold iterator with non-overlapping groups, with anonymous patient ID as group identifier. The encoder was frozen for first 2 epochs of the training.

System: The experiments were conducted using Python 3.6.13 and PyTorch 1.3.1. Furthermore, the models were trained and evaluated on a single NVIDIA Titan RTX graphics card with 24 GB of memory running on Ubuntu 18.04.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

2.4 Evaluation

For evaluation, model predictions (predicted segmentation masks) were averaged over the 5 cross-validation folds. Furthermore, 0.5 was used as a threshold value to produce the prediction maps. Dice score (DCS) and Intersection over Union (IoU), both ranging from 0 to 1, were used as evaluation metrics.

3. EXPERIMENTAL RESULTS

For the clinically-realistic Finnish mammography screening holdout dataset encoder-decoders with U-Net decoder performed the best (Table 2). In the FPN decoder case ResNet22 and ResNet34 yielded comparable performance. When predicting masses for the Portuguese INbreast dataset models with ResNet22 encoder yielded higher Dice scores (Table 2). Overall, the models performed well for INbreast. Important thing to notice is that the models in both cases were trained only on Finnish training data.

Table 2. Dice scores (DCS) and Intersection over union (IoU) metrics for different models and dataset evaluations. SD denotes standard deviation

Dataset	Model	Pre-trained weights	DCS (±SD)	IoU (±SD)
OUDSM:				
	ResNet34UNet	ImageNet	**0.7677** (±0.2134)	0.6625 (±0.2322)
	ResNet22UNet	GMIC	0.7189 (±0.2529)	0.6115 (±0.2593)
	ResNet34FPN	ImageNets	0.7328 (±0.2259)	0.6202 (±0.2380)
	ResNet22FPN	GMIC	0.6858 (±0.2676)	0.5759 (±0.2696)
INbreast:				
	ResNet34UNet	ImageNet	0.7890 (±0.1906)	0.6803 (±0.1913)
	ResNet22UNet	GMIC	**0.8430** (±0.1091)	0.7410 (±0.1358)
	ResNet34FPN	ImageNet	0.8015 (±0.1149)	0.6825 (±0.1465)
	ResNet22FPN	GMIC	0.8050 (±0.1586)	0.6974 (±0.1851)

In addition, for visual analysis, we depicted a comparison between the reference segmentation masks and the predicted ones (Fig. 2). Some of the predictions resemble the reference annotations with a high detail.

4. DISCUSSION

Overall, the results show that there are no large differences between results achieved with ResNet22 encoder and those with ResNet34. U-Net and the model with FPN decoder also give comparable performance. The lower computational cost of the FPN might make that particular architecture more appealing. The U-Net usually provides more detailed predictions than FPN, where predictions are made on different scales and then up-scaled, whereas in U-Net only the final feature map is utilized. Furthermore, it is unclear what is the role in terms of segmentation performance of the additional two channels in ImageNet trained ResNet34 for this application. Xie and Richmond[28] have used a model pre-trained on gray-scale ImageNet for disease classification from chest X-rays, with results outperforming color ImageNet initialization in terms of speed and accuracy.

We observed that the quantitative results had variabilities between the two datasets. Both dataset are imbalanced in terms of malignant and benign cases, and furthermore there is within class variation in morphology (shape, structure, texture/pattern and size). Based on empirical results,[24] the Focal Tversky loss helps alleviate the training issues typical for applications having small areas of interests (when compared to the pixel area of whole image). Furthermore, reference segmentation is sometimes not known accurately, *i.e.*, there is no pixel-level reference mask available for all masses which can be concluded by post-analysis of the raw data and annotations. Cases which have received an assessment score 'malignancy cannot be excluded' in the screening are particularly

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

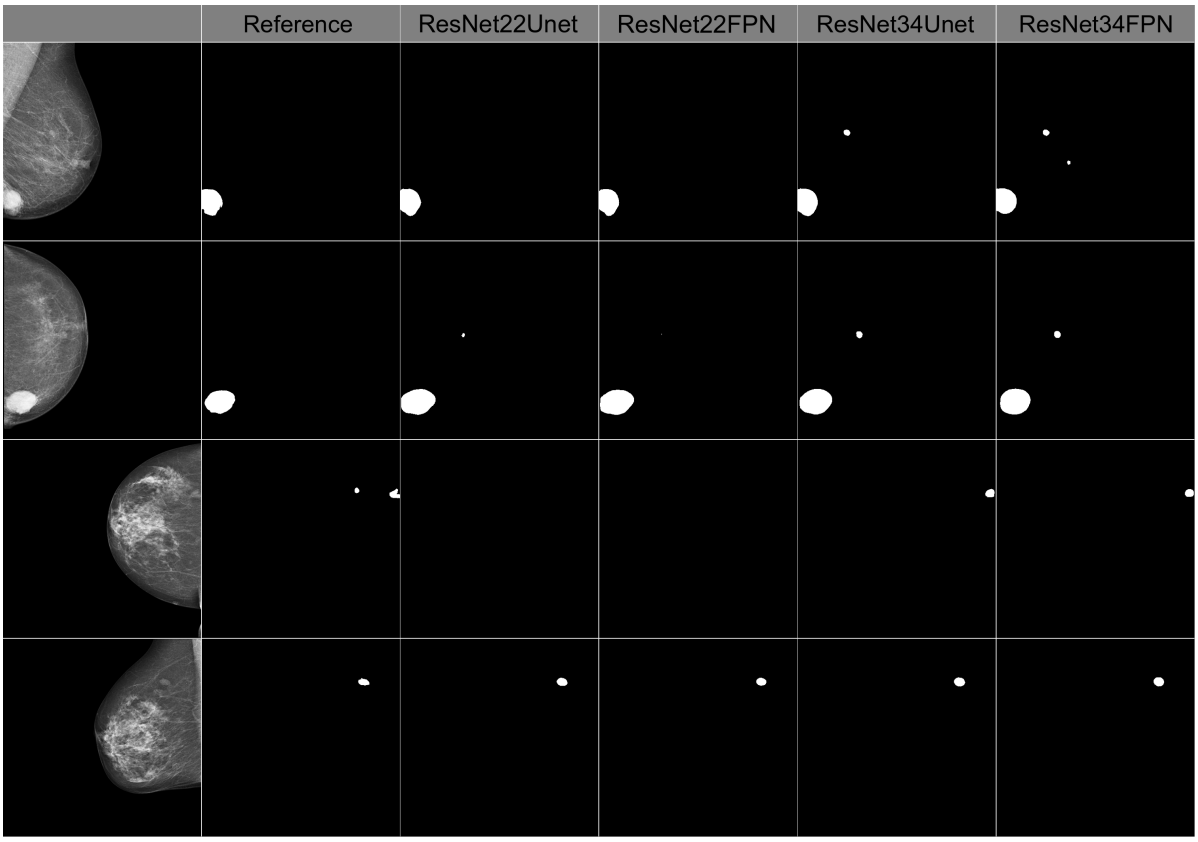

Figure 2. Qualitative segmentation results. Mammograms and reference masks courtesy of Breast Research Group, INESC Porto, Portugal.

difficult, as those may contain nonspecific texture suggesting, for example, a potential mass, but end up receiving a normal histology response.

The main benefit from pre-trained weights from breast cancer classification appeared during the model training where some of the randomly split cross-validation folds had received more challenging samples. Even though the model corresponding that particular fold did not become strong in the segmentation task, it was able to achieve better train time Dice score than the ImageNet based counterpart.

Finally, solutions which utilize a two-stage framework, *i.e.*, perform mass segmentation for masses localized first with a detection network, *e.g.*, Yan et al.[29] and Baccouche et al.,[30] are able to receive better segmentation results (*e.g.*, for INbreast DCS 0.8044 and DCS 0.9528 for Yan et al. and Baccouche et al., respectively). As lesion shape and margin are important cues when performing the classification to malignant or benign, accurate predicted segmentations have a key role. Our future studies will steer towards learning the mass characteristics for localized candidate in the original resolution.

5. CONCLUSIONS

In this paper, we have evaluated encoder-decoder methods, trained on a subset of Finnish mammography screening dataset, to assess the presence of masses in digital screening mammograms. We found no large differences between results achieved with ResNet22 encoder, pre-trained for breast cancer evaluation, and those with ResNet34, pre-trained on ImageNet. In comparison to refined datasets, clinically-realistic datasets introduce new challenges due to the natural reasons for the reference not being known with 100% accuracy, and this should be further addressed in the future.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

6. DISCLOSURES

No conflicts of interests, financial or otherwise, are declared by the authors.

7. ACKNOWLEDGEMENTS

Miika T. Nieminen received funding from the Jane and Aatos Erkko Foundation and the Technology Industries of Finland Centennial Foundation. Antti Isosalo received funding from the Jenny and Antti Wihuri Foundation (grant no. 00210099) and the Thelma Mäkikyrö Foundation. Satu I. Inkinen received funding from the Academy of Finland (project no. 316899). Inês Domingues is acknowledged for providing the Portuguese INbreast database. The authors would like to acknowledge the scientific discussions with Aleksei Tiulpin, Santeri Rytky, Egor Panfilov, Hoang Huy Nguyen and Mustafa Al-Rubaye.

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F., "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians* **71**(3), 209–249 (2021).

[2] Ben-Ari, R., Akselrod-Ballin, A., Karlinsky, L., and Hashoul, S., "Domain specific convolutional neural nets for detection of architectural distortion in mammograms," in [*IEEE 14th International Symposium on Biomedical Imaging*], 552–556, IEEE (2017).

[3] Zamir, R., Bagon, S., Samocha, D., Yagil, Y., Basri, R., Sklair-Levy, M., and Galun, M., "Segmenting microcalcifications in mammograms and its applications," in [*Medical Imaging 2021: Image Processing*], Išgum, I. and Landman, B. A., eds., **11596**, 788–795, International Society for Optics and Photonics, SPIE (2021).

[4] Ribli, D., Horváth, A., Unger, Z., Pollner, P., and Csabai, I., "Detecting and classifying lesions in mammograms with deep learning," *Scientific reports* **8**(1), 1–7 (2018).

[5] Wang, Y., Wang, S., Chen, J., and Wu, C., "Whole mammographic mass segmentation using attention mechanism and multiscale pooling adversarial network," *Journal of Medical Imaging* **7**(5), 054503 (2020).

[6] Lou, M., Qi, Y., Li, X., Xu, C., Zhao, W., Deng, X., and Ma, Y., "Aggregated pyramid attention network for mass segmentation in mammograms," *Multimedia Tools and Applications* , 1–19 (2021).

[7] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (2009).

[8] Agarwal, R., Diaz, O., Lladó, X., Yap, M. H., and Martí, R., "Automatic mass detection in mammograms using deep convolutional neural networks," *Journal of Medical Imaging* **6**(3), 1–9 (2019).

[9] Liu, K., Shen, Y., Wu, N., Chłędowski, J. P., Fernandez-Granda, C., and Geras, K. J., "Weakly-supervised high-resolution segmentation of mammography images for breast cancer diagnosis," *Medical Imaging with Deep Learning* **4** (2021).

[10] Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S. G., Moy, L., Cho, K., and Geras, K. J., "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Medical Image Analysis* **68**, 101908 (2020).

[11] Wu, N., Phang, J., Park, J., Shen, Y., Kim, S. G., Heacock, L., Moy, L., Cho, K., and Geras, K. J., "The NYU breast cancer screening dataset v1.0." New York University, New York, NY, USA, Tech. Rep. (2019). (Accessed: 28 Oct 2021).

[12] Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L. L. Y., Ho, K., Weinstein, J. D., Reig, B., Gao, Y., Toth, H., Pysarenko, K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S. G., Heacock, L., Moy, L., Cho, K., and Geras, K. J., "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Transactions on Medical Imaging* **39**(4), 1184–1194 (2020).

[13] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., and Cardoso, J. S., "Inbreast: Toward a full-field digital mammographic database," *Academic Radiology* **19**(2), 236–248 (2012).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

[14] Isosalo, A., Heino, H., Inkinen, S. I., and Nieminen, M. T., "Mammogram annotation tool." GitHub, 14 Sep 2021, https://github.com/MIPT-Oulu/MammogramAnnotationTool_public (2021). (Accessed: 29 Oct 2021).

[15] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical image computing and computer-assisted intervention, Lecture Notes in Computer Science*], **9351**, 234–241, Springer (2015).

[16] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S., "Feature pyramid networks for object detection," in [*IEEE conference on computer vision and pattern recognition*], 2117–2125 (2017).

[17] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 770–778, IEEE (June 2016).

[18] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., "Pytorch: An imperative style, high-performance deep learning library," in [*Advances in Neural Information Processing Systems*], Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., eds., **32**, 8024–8035, Curran Associates, Inc. (2019).

[19] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C., "Efficient object localization using convolutional networks," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 648–656 (June 2015).

[20] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*International conference on machine learning*], 448–456, PMLR (2015).

[21] DeVries, T. and Taylor, G. W., "Improved regularization of convolutional neural networks with cutout," *arXiv:1708.04552v2* (2017).

[22] Cox, D. R., "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)* **20**(2), 215–232 (1958).

[23] Rahman, M. A. and Wang, Y., "Optimizing intersection-over-union in deep neural networks for image segmentation," in [*International symposium on visual computing*], 234–244, Springer (2016).

[24] Abraham, N. and Khan, N. M., "A novel focal tversky loss function with improved attention U-Net for lesion segmentation," in [*IEEE 16th International Symposium on Biomedical Imaging*], 683–687, IEEE (2019).

[25] Tversky, A., "Features of similarity," *Psychological review* **84**(4), 327–352 (1977).

[26] Salehi, S. S. M., Erdogmus, D., and Gholipour, A., "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in [*Machine Learning in Medical Imaging, Lecture Notes in Computer Science*], Wang, Q., Shi, Y., Suk, H. I., and Suzuki, K., eds., **10541**, 379–387 (2017).

[27] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv:1412.6980v9* (2014).

[28] Xie, Y. and Richmond, D., "Pre-training on grayscale imagenet improves medical image classification," in [*Proceedings of the European Conference on Computer Vision (ECCV) Workshops*], 476–484, Springer (September 2019).

[29] Yan, Y., Conze, P.-H., Quellec, G., Lamard, M., Cochener, B., and Coatrieux, G., "Two-stage multi-scale breast mass segmentation for full mammogram analysis without user intervention," *Biocybernetics and Biomedical Engineering* **41**(2), 746–757 (2021).

[30] Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C., and Elmaghraby, A. S., "Connected-unets: a deep learning architecture for breast mass segmentation," *NPJ Breast Cancer* **7**(1), 1–12 (2021).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 08 Apr 2022
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use