IMPROVED PECTORAL MUSCLE SEGMENTATION IN MAMMOGRAMS THROUGH REGRESSION-BASED DEEP LEARNING AND KNOWLEDGE DISTILLATION

Christian Huemmer[⋆], *Ramyar Biniazan*[⋆], *Manasi Datar*[⋆†], *Martin Kraus*[†], *Andreas Fieselmann*[⋆], *Steffen Kappler*[⋆]

[⋆] X-ray Products, Siemens Healthcare GmbH, Germany
[⋆†] Digital Technology & Innovation, Siemens Healthineers India LLP
[†] Digital Technology & Innovation, Siemens Healthcare GmbH, Germany

ABSTRACT

Pectoral muscle (PM) segmentation is an important step for improving the accuracy and efficiency of breast cancer screening in digital mammography. In recent years, image-to-image (I2I) deep learning (DL) methods have achieved state-of-the-art performance for automated PM segmentation by representing the PM region as a binary mask. This paper introduces a new curve regression approach by representing the PM boundary as a vector of connected points that lie on the curve separating the PM from surrounding breast tissue. This low-dimensional PM representation is used to introduce a concept of knowledge distillation (KD), which exploits an ensemble of teachers to perform loss weighting based on maximum likelihood (ML) estimation. Experiments with in-house mammography data show that DL based curve regression outperforms a reference I2I DL method (U-net) for PM segmentation. Further, application of the proposed KD concept achieves higher segmentation accuracy with only 16% of parameters and 23% of inference time compared to the U-net.

Index Terms— Knowledge distillation, pectoral muscle segmentation, teacher-student, semi-supervised learning

1. INTRODUCTION

Breast cancer is one of the most common causes of death related to cancer [1]. As a countermeasure, digital breast mammography has been established as a reliable, cost effective imaging technique with extensive use in breast cancer screening, diagnostics, and follow up. One important and challenging step in mammography image analysis is Pectoral muscle (PM) segmentation. As an example, not excluding the PM region from quantification algorithms may lead to inaccurate characterization of the breast tissue [2]. Furthermore, multi-view and longitudinal alignment of mammograms also benefit from the PM as an important anatomical landmark. Therefore, there is a strong motivation to provide technical solutions for the task of PM segmentation. However, PM segmentation in mammography images remains challenging due to the significant variability in its shape and appearance. Moreover, differences in positioning of the breast during image acquisition, blurred edges due to similarity with surrounding glandular tissue in terms of morphology and appearance, and visual obstruction due to an overlap with dense glandular tissue or skin folds constitute further challenges.

A number of different approaches have been proposed for PM segmentation in mammograms. Earlier methods relied on assumptions about the shape of the PM boundary and employed a straight line initialization like [3], followed by refinement using active contours [4] or polynomial fitting [5]. Further advancements included intensity-based filtering and region growing methods as reviewed in [6]. The above methods depend on handcrafted features, which limits their performance for the PM segmentation task. Advances in deep learning (DL), specifically deep convolutional neural networks (DCNN), have offered new possibilities to overcome these limitations. DCNN architectures specifically tailored for PM segmentation, such as modified versions of the Holistically-nested Edge Detection network [7], VGG16 [8], and U-net [2, 9] are able to learn mid- to high-level features directly from the mammogram, and further employ extensive post-processing like curve interpolation [7], scanline filling [9], or shortest path computation [8] to improve PM segmentation performance.

The methods mentioned above can be categorized to the class of supervised DL that requires large scale expert annotations. Semi-supervised DL provides a practical way of training neural networks by utilizing unlabelled data and has found various applications in medical image segmentation [10]. For instance, model compression techniques transfer information from larger to smaller models to achieve high accuracy with low footprint and fast execution time, which is especially appealing for clinical workflow integration [11]. In this context, Knowledge distillation (KD) [12] has emerged as a popular technique for transferring knowledge from a large (teacher) network to a small (student) network. While the teacher network is trained in a supervised

manner, semi-supervised techniques may be used for distillation of knowledge to the student network [13, 14]. Medical image segmentation has benefited from these techniques as well (e.g. [15, 16]), especially because of the costs associated with labeling large-scale datasets by clinical experts.

In this paper, we present a new approach for PM segmentation by explicitly incorporating prior anatomical knowledge into the neural network design. More specifically, the boundary separating the PM from the breast tissue is represented by the column-index (CI) vector, which is defined to as a sorted vector of contiguous, row-wise column indices of the PM boundary in the mammography image. This low-dimensional representation of an image region is the basis for introducing a new KD concept, where a collection of CI vector proposals estimated by an ensemble of teachers is used to train a student based on a weighted loss derived by maximum likelihood (ML) estimation. This concept penalizes training gradient for examples with higher teacher disagreement, and is especially appealing for semi-supervised learning using a large unlabeled dataset.

2. METHODOLOGY

This paper exploits anatomical information about the continuity of the PM boundary to simplify the PM segmentation task. This is achieved by representing the PM boundary as a CI vector, which reformulates the region segmentation proposed in previous work into a new vector regression task. The CI vector definition and a corresponding supervised DL based regression method is described in Subsection 2.1. As this low-dimensional representation is well-suited for semi-supervised learning, we introduce a teacher-student concept for CI vector regression in Subsection 2.2.

2.1. Supervised CI Vector Regression

Given a $2D$ image $\mathbf{X}$ with R rows and C columns, along with the corresponding boundary mask (BM) (Figure 1), we define an ordered vector containing the CI of the PM boundary for every row in the image:

$$\mathbf{z} = [z_1, ..., z_R], \quad \text{where} \quad z_i \in [1, C], \quad \forall i \in [1, R]. \tag{1}$$

The entries z_i of the CI vector $\mathbf{z}$ are defined as:

$$z_i = \begin{cases} c & \text{if } BM_{i,c} = 1 \text{ and } \sum_{i=1}^{R} BM_{i,c} = 1 \\ C & \text{otherwise,} \end{cases} \tag{2}$$

where $BM_{i,c}$ denotes the entry at the i-th row and c-th column of the BM matrix

$$\begin{aligned} \mathbf{BM} &= \mathbf{X} - (\mathbf{X} \ominus \mathbf{SE}) \quad \text{with} \\ \mathbf{SE} &= [[0,1,0],[1,1,1],[0,1,0]]. \end{aligned} \tag{3}$$

Here $\ominus$ describes the binary erosion operator [17].

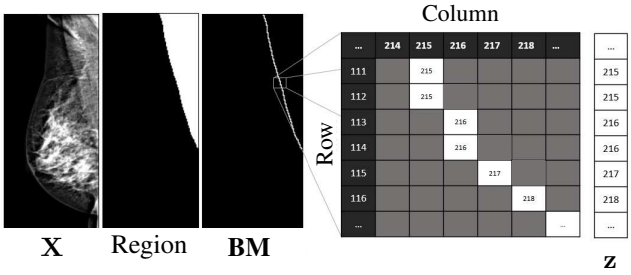

Fig. 1. Schematic overview of representing the PM obtained from expert annotations as a binary region mask, a BM obtained by performing morphological operations on the region mask, or using the proposed CI vector $\mathbf{z}$.

Representation of the PM boundary using the CI vector reformulates the PM segmentation as a low-dimensional vector regression task. Existing DL architectures can be adapted for this task by ensuring that the size of the final output layer is equal to R and predictions are in the range $[1, C]$. Further, a known regression loss function can be used to train the network. Specifics of supervised CI vector regression training employed for this work are provided in Section 3.

2.2. Semi-supervised CI Vector Regression

The CI vector provides a low-dimensional representation of the PM region and is well-suited for distilling knowledge, e.g., from an ensemble of teachers to a student network, as described in the paragraphs below. Please note that alternative KD concepts similar to those described in [18] may be adapted to work with the CI vector representation as well.

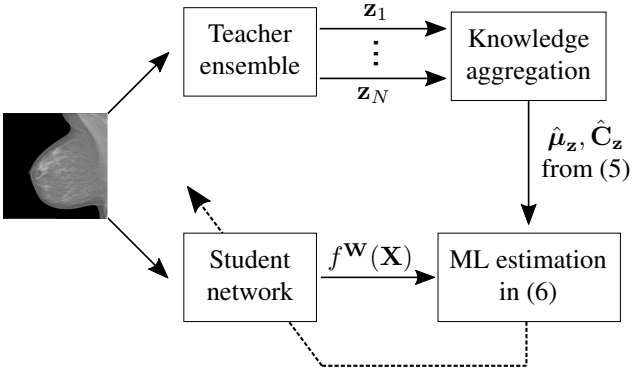

Fig. 2. Schematic overview of applying teacher-student learning to CI vector regression.

As depicted in Figure 2, we consider an ensemble of teachers to produce N proposals $\mathbf{z}_n$ $(n = 1, ..., N)$ of the latent CI vector $\mathbf{z}$. Furthermore, a student network transforms the current input image $\mathbf{X}$ to the student output vector $f^{\mathbf{W}}(\mathbf{X})$, which depends on the nonlinear function $f(\cdot)$ and the trainable parameters $\mathbf{W}$. The student output vec-

tor $f^{\mathbf{W}}(\mathbf{X})$ in Figure 2 is modeled to follow a multivariate Gaussian distribution

$$p(f^{\mathbf{W}}(\mathbf{X})) = \frac{1}{(2\pi)^{R/2}} \frac{1}{\det\{\hat{\mathbf{C}}_{\mathbf{z}}\}} \cdot \exp\left\{ -\frac{1}{2} \left(f^{\mathbf{W}}(\mathbf{X}) - \hat{\boldsymbol{\mu}}_{\mathbf{z}} \right)^{\mathrm{T}} \hat{\mathbf{C}}_{\mathbf{z}}^{-1} \left(f^{\mathbf{W}}(\mathbf{X}) - \hat{\boldsymbol{\mu}}_{\mathbf{z}} \right) \right\} \tag{4}$$

where $\det\{\cdot\}$ represents the determinant of a matrix and $(\cdot)^{\mathrm{T}}$ denotes the transpose of a vector. The mean vector $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$ and covariance matrix $\hat{\mathbf{C}}_{\mathbf{z}}$ in (4) are estimated from the teacher proposals $\mathbf{z}_n$ according to:

$$\hat{\boldsymbol{\mu}}_{\mathbf{z}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{z}_n, \;\; \hat{\mathbf{C}}_{\mathbf{z}} = \frac{1}{N-1} \sum_{n=1}^{N} (\mathbf{z}_n - \hat{\boldsymbol{\mu}}_{\mathbf{z}})(\mathbf{z}_n - \hat{\boldsymbol{\mu}}_{\mathbf{z}})^{\mathrm{T}} \tag{5}$$

Based on the probabilistic model defined by (4) and (5), ML estimation of model parameters $\mathbf{W}$ reveals:

$$-\log\left(p(f^{\mathbf{W}}(\mathbf{X})\right) \propto \frac{1}{2} \left(f^{\mathbf{W}}(\mathbf{X}) - \hat{\boldsymbol{\mu}}_{\mathbf{z}} \right)^{\mathrm{T}} \hat{\mathbf{C}}_{\mathbf{z}}^{-1} \left(f^{\mathbf{W}}(\mathbf{X}) - \hat{\boldsymbol{\mu}}_{\mathbf{z}} \right). \tag{6}$$

Here the deviation between student output vector $f^{\mathbf{W}}(\mathbf{X})$ and teacher mean vector $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$ is weighted by the inverse covariance matrix $\hat{\mathbf{C}}_{\mathbf{z}}^{-1}$. As a consequence, rows with a smaller teacher agreement are penalized and their contribution to the overall gradient is diminished during training. This is especially appealing for semi-supervised learning using a large unlabeled dataset, as shown in Section 3.

3. EXPERIMENTS AND RESULTS

This section details experiments designed to evaluate the proposed CI regression approach for PM segmentation in a supervised (Subsection 3.1) and semi-supervised (Subsection 3.2) manner. For all experiments we used PyTorch [19] as training library with a maximum number of 150 epochs.

Image preprocessing. For all experiments, image preprocessing was realized by extracting the image content from the DICOM file, performing resizing to an image dimension of 256×256 (bilinear interpolation, zero padding) and subsequently normalizing the pixel values into the range $[0, 1]$.

Labeled data. We exploit $8k$ mediolateral oblique (MLO) mammograms (both left and right laterality) acquired by Anonymous Product from 4 different clinics[1]. Ground truth segmentation was provided by clinical experts in the form of binary masks. From this labeled dataset, we extracted 16% for independent testing by keeping patient boundaries. It is important to mention that this testing data was consistently used throughout this section to report comparable performance metrics for different approaches.

[1]DICOM images were exported based on clinical collaboration and following all required local regulations and the general data protection regulation (GDPR) of the European Union. Additional ethical approval was not required.

3.1. Supervised learning

As mentioned in the description above (see *Labeled data*) we used 84% of the labeled dataset for network training. These 6720 images were split into 80% for optimizing neural network parameters and 20% for validation.

Reference Method.: U-Net [20] is a state-of-the-art neural network architecture for image-to-image (I2I) translation. In this paper, we adapted the U-Net architecture described in [2] to predict a binary mask of the PM region from a preprocessed mammography image.

CI vector regression. The widely-used Densenet-121 [21] architecture was adapted to predict the PM boundary through CI vector regression. More specifically, the linear output layer contains 256 entries to account for all 256 rows in the preprocessed mammography image.

Experimental results. A comprehensive marginal analysis was conducted to individually optimize hyper-parameters for a fair comparison. The optimizer, loss, and learning rate scheduler were found to be the most important parameters to consider.

This led to the choice of the Adam optimizer and a learning rate scheduler (learning rate=0.001, step size=10, patience=10) for both CI vector regression and the U-net. Marginal analysis further informed the choice of the loss functions as DICE loss and sum-of-absolute error for the U-net and CI vector regression, respectively.

This marginal analysis was followed by investigating the PM segmentation accuracy achieved by CI vector regression (Densenet-121) and binary mask segmentation (U-Net) for different network sizes. To this end, every training was repeated at least four times to account for slight metric fluctuations produced, e.g., by random network initialization. The results of the average DICE scores illustrated in Figure 3 reveal that the proposed CI vector regression outperforms binary mask segmentation and achieves a high PM segmentation accuracy even with a very small number of network parameters. Please note that further increasing the network sizes as displayed in Figure 3 did not remarkably increase the average DICE score values.

The superiority of CI vector regression compared to binary mask segmentation in Figure 3 is highlighted by comparing performance metrics in Table 1. To this end, we selected network architectures based on Figure 3 independently for U-Net and CI regression. The goal of this selection was to choose a good compromise between the number of parameter and the average DICE score accuracy. Inspecting the results for supervised learning (upper block of Table 1) reveals that the CI regression with a Densenet-121 of 165k parameters achieves higher PM segmentation accuracy compared to the U-Net of 483k parameters at only 36% of the inference time. This is shown by the rightmost column in Table 1 and was measured relative to the U-Net with 483k parameters on a Intel(R) Xeon(R) Silver 4116 CPU (2.10 GHz).

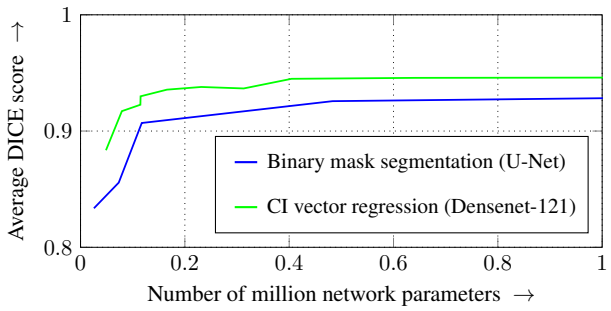

Fig. 3. Average Dice score achieved by CI vector regression (Densenet-121) and binary mask segmentation (U-Net) for varying number of network parameters.

Table 1. PM segmentation accuracy and relative inference time (RI time) for selected model architectures and number of neural network parameters (# net pars) achieved using Binary Mask (BM) segmentation (U-Net) and CI vector regression (Densenet-121).

PM segmentation method	# net pars	DICE score mean/median	RI time
U-Net for	73k	0.855/0.915	0.56
BM segmentation	117k	0.909/0.950	0.81
(supervised)	483k	0.929/0.958	1.00
Densenet-121 for	79k	0.917/0.962	0.23
CI vector regression	165k	0.935/0.970	0.36
(supervised)	400k	0.945/0.973	0.48
Densenet-121 for	79k	0.944/0.973	0.23
CI vector regression	165k	0.948/0.974	0.36
(semi supervised)	400k	0.950/0.975	0.48

These results confirm that CI vector regression provides an effective means for fast and accurate supervised PM segmentation using a labeled training dataset.

3.2. Semi-supervised learning

The labeled dataset mentioned above is complemented by $60k$ MLO unlabeled mammograms of unique patients extracted from different hospitals[2].

Teacher networks (supervised learning). We trained 6 teacher networks using CI vector regression on the labeled dataset of 6720 images as described in Subsection 3.1. This was realized based on 6-fold data splitting and independent network optimization. Please note that this approach serves as a proof of concept and that other strategies for teacher training (e.g., [18] and references therein) could be applied as well.

[2]DICOM images were exported based on clinical collaboration and following all required local regulations and the general data protection regulation (GDPR) of the European Union. Additional ethical approval was not required.

Student network (unsupervised learning). The CI vector regression estimates provided by the teacher ensemble were used to train the student network based on ML estimation as described in Subsection 2.2. To this end, we re-trained the Densetnet-121 network architectures of Table 1 as follows: The unlabeled dataset of $60k$ MLO mammograms was used to adjust network parameters, all 6720 images of the labeled dataset were used for validation. Please note that for all images an artificial ground truth was consistently provided by the teacher ensemble.

Experimental results. The lower block of Table 1 shows that re-training CI vector regression networks on the unlabeled dataset increases PM segmentation accuracy especially for small network architectures. To give an example, the Densenet-121 of $79k$ parameters outperforms the PM segmentation achieved by the U-Net of $483k$ parameters by using only 16% of parameters and 23% of the inference time. Furthermore, the accuracy drop obtained by reducing the network size is much smaller by applying knowledge distillation. Reducing the number of Densetnet-121 parameters from $400k$ to $79k$ leads to a drop of DICE scores mean values from 0.950 to 0.944 and 0.945 to 0.917 using semi-supervised and supervised learning, respectively. These results confirm that the proposed concept of CI vector regression is well-suited for creating small and efficient neural networks for the task of PM segmentation.

4. CONCLUSION

We proposed a contour-based deep learning method which offers an innovative solution to the challenging task of PM segmentation in mammography images. By training a network to learn the PM boundary by solving a vector regression instead of I2I segmentation task, we achieved improvements in accuracy, efficiency, and inference speed. Further, the proposed approach lent itself nicely to knowledge distillation, where we introduced a teacher-student concept based on ML estimation. This novel semi-supervised learning strategy resulted in a further increase in accuracy especially for small network architectures. All demonstrated benefits position our method as a promising and practical solution for accurate and quick PM segmentation. In future, the proposed strategy could also be transferred to other applications in medical imaging.

5. COMPLIANCE WITH ETHICAL STANDARDS

Data selection followed all required local regulations and the general data protection regulation (GDPR) of the European Union. Additional ethical approval was not required.

6. DISCLAIMER

The presented methods in this paper are not commercially available and their future availability cannot be guaranteed.

7. REFERENCES

[1] R.L. Siegel, K.D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, no. 1, pp. 7–30, 2016.

[2] X. Ma, J. Wei, C. Zhou, M.A. Helvie, H. Chan, L.M. Hadjiiski, and Y. Lu, "Automated pectoral muscle identification on mlo-view mammograms: Comparison of deep neural network to conventional computer vision," *Medical physics*, vol. 46, no. 5, pp. 2103–2114, 2019.

[3] N. Karssemeijer, "Automated classification of parenchymal patterns in mammograms," *Physics in medicine & biology*, vol. 43, no. 2, pp. 365, 1998.

[4] R.J. Ferrari, A.F. Frere, R.M. Rangayyan, J. Desautels, and R.A. Borges, "Identification of the breast boundary in mammograms using active contour models," *Medical and Biological Engineering and Computing*, vol. 42, pp. 201–208, 2004.

[5] V.B. Bora, A.G. Kothari, and A.G. Keskar, "Robust automatic pectoral muscle segmentation from mammograms using texture gradient and euclidean distance regression," *Journal of digital imaging*, vol. 29, pp. 115–125, 2016.

[6] M. Moghbel, C.Y. Ooi, N. Ismail, Y.W. Hau, and N. Memari, "A review of breast boundary and pectoral muscle segmentation methods in computer-aided detection/diagnosis of breast mammography," *Artificial Intelligence Review*, vol. 53, pp. 1873–1918, 2020.

[7] A. Rampun, K. López-Linares, P.J. Morrow, B.W. Scotney, H. Wang, I.G. Ocaña, G. Maclair, R. Zwiggelaar, M.A.G. Ballester, and I. Macía, "Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network," *Medical image analysis*, vol. 57, pp. 1–17, 2019.

[8] H. Soleimani and O.V. Michailovich, "On segmentation of pectoral muscle in digital mammograms by means of deep learning," *IEEE Access*, vol. 8, pp. 204173–204182, 2020.

[9] W. Liu, C. Liu, and Y. Wei, "Utilizing deep learning technology to segment pectoral muscle in mediolateral oblique view mammograms," in *Proceedings of ICSIP*. IEEE, 2020, pp. 97–101.

[10] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," in *Proceedings of MIDL*. PMLR, 2022, pp. 820–833.

[11] J. Gou, B. Yu, S.J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[13] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proceedings of NIPS*. 2017, vol. 30, pp. 1195—-1204, Curran Associates, Inc.

[14] C. Yang, L. Xie, S. Qiao, and A.L. Yuille, "Training deep neural networks in generations: A more tolerant teacher educates better students," in *Proccedings of AAAI'19/IAAI'19/EAAI'19*. 2019, vol. 690, pp. 5628–5635, AAAI Press.

[15] Y. Zhou, H. Chen, H. Lin, and P.-A. Heng, "Deep semi-supervised knowledge distillation for overlapping cervical cell instance segmentation," in *Proceedings of MICCAI*. 2020, pp. 521–531, Springer.

[16] J.W. Choi, "Knowledge distillation from cross teaching teachers for efficient semi-supervised abdominal organ segmentation in ct," in *Proceedings of MICCAI 2022 Challenge FLARE*, pp. 101–115. Springer, 2023.

[17] P. Maragos and R. Schafer, "Morphological filters–part 1: Their set-theoretic analysis and relations to linear shift-invariant filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 8, pp. 1153–1169, 1987.

[18] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 3048–3068, 2021.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., "Pytorch: An imperative style, high performance deep learning library," in *Proceedings of NIPS*, 2019, vol. 32, pp. 8026—-8037.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of MICCAI*. Springer, 2015, pp. 234–241.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, "Densely connected convolutional networks," in *Proceeding of IEEE CPVR*, 2017, pp. 4700–4708.