Heteroscedastic Heatmap Regression for Reliable Pectoral Muscle Segmentation in Mammography

Paul Zech[1,2] P.ZECH@SIEMENS-HEALTHINEERS.COM
Christian Hümmer[1] CHRISTIAN.HUEMMER@SIEMENS-HEALTHINEERS.COM
Benjamin El-Zein[1,2] BENJAMIN.EL-ZEIN@SIEMENS-HEALTHINEERS.COM
Christopher Syben[1] CHRISTOPHER.SYBEN@SIEMENS-HEALTHINEERS.COM
Ludwig Ritschl[1] LUDWIG.RITSCHL@SIEMENS-HEALTHINEERS.COM
Steffen Kappler[1] STEFFEN.KAPPLER@SIEMENS-HEALTHINEERS.COM
Sebastian Stober[2] STOBER@OVGU.DE
[1] *Siemens Healthineers AG, Forchheim, Germany*
[2] *Otto-von-Guericke University, Magdeburg, Germany*

Editors: Under Review for MIDL 2026

Abstract

Breast cancer remains a leading cause of mortality worldwide, making accurate mammography screening essential for early detection. An important preprocessing step in mammography is the accurate segmentation of the pectoral muscle as it affects downstream tasks such as breast density estimation or automated exposure control. Existing automated segmentation methods, both traditional and deep learning-based, often lack reliable confidence measures, which becomes especially problematic in the presence of occlusions or visually confounding structures such as skin folds or other muscle fibers. To address this limitation, we propose a probabilistic framework that combines heatmap-based boundary regression with heteroscedastic uncertainty to capture input-dependent variability. Our approach not only predicts the pectoral muscle boundary but also quantifies the associated uncertainty. While mainly producing unimodal predictions, the probabilistic heatmaps reveal multimodal patterns for confounding structures, further enhancing transparency in challenging cases. We demonstrate that our method provides robust and transparent means to achieve accurate segmentation while providing meaningful uncertainty estimates.

Keywords: Pectoral muscle segmentation, heteroscedastic regression, aleatoric uncertainty

1. Introduction

Breast cancer remains one of the most prominent cancer types, especially in young and middle-aged women (Siegel et al., 2016; Ren et al., 2022). To improve early detection, the World Health Organization (WHO) recommends regular mammography screening in this population (World Health Organization, 2014). One standard projection for breast tissue characterization is mediolateral oblique (MLO) mammograms. Besides breast tissue, MLO projections also capture the pectoral muscle (PM), which typically appears as a bright and dense region in the upper corner of the image. Accurate segmentation of the PM constitutes an essential preprocessing step in the analysis of mammography images. For instance, breast positioning control systems use the PM as a key anatomical landmark (Brahim et al., 2022). Moreover, it is important in automated exposure control, where a low-dose image is acquired for breast density quantification and subsequent full-dose estimation.

In this context, accurate exclusion of the PM is essential to prevent the dose from being calibrated to the dense PM tissue, which would otherwise result in excessive radiation and overexposure. To automate the PM segmentation task, many different traditional and deep learning-based algorithms have been developed. However, the proposed solutions provide no or insufficient measures of confidence. This is specifically problematic in PM segmentation as the PM may be obscured by dense glandular tissue or confused with structurally similar skin folds and other muscle fibers. In such ambiguous cases, uncertainty modeling is essential to identify potentially unreliable local segmentation results.

To address these challenges, we propose a novel framework for PM segmentation that explicitly accounts for multiple sources of input ambiguity by modeling probabilistic heatmaps that jointly encode both the PM contour and its associated local prediction uncertainty. This method is inspired by our previous work Zech et al. (2025) and introduces substantial technical improvements, which are validated in a comprehensive experimental evaluation.

2. Related work

PM segmentation: Early developments leverage prior knowledge about the shape of the PM boundary by detecting a straight line (Karssemeijer, 1998) or fitting active contours to the PM boundary (Ferrari et al., 2004; Wang et al., 2010). Other traditional algorithms can be categorized in line detection, intensity-based, wavelet-based and statistical techniques as summarized in Ganesan et al. (2013). However, these methods rely on extensive pre- and postprocessing as well as feature engineering.

Recent advances in deep learning overcome these limitations by learning hierarchical representations directly from data using deep neural networks. Some approaches treat the task as semantic segmentation through pixel-wise classification of the PM tissue. Architectures such as U-Net (Ronneberger et al., 2015) and its variants have been widely adopted for this purpose (Ma et al., 2019; Liu et al., 2020), with extensions that incorporate adversarial training to improve robustness and anatomical plausibility (Guo et al., 2020). Other methods focus on delineating the contour of the PM, using edge detection or boundary-aware strategies, such as VGG16-based contour detection (Soleimani and Michailovich, 2020), U-Net adaptations for boundary extraction (Angelone et al., 2025), and Holistically-Nested Edge Detection (Rampun et al., 2019). In contrast, one approach formulates contour prediction as a row-wise boundary regression task (Huemmer et al., 2024), introducing an inductive bias in the network architecture that promotes continuity and structural consistency of the PM. Despite their high accuracy, the lack of a robust measure of uncertainty remains a key limitation of these solutions.

Uncertainty quantification: Following Kendall and Gal (2017), uncertainty arises from two sources: (1) epistemic uncertainty, due to limited training data and uncertain model parameters, and (2) aleatoric uncertainty, caused by noise or ambiguity in the input. Aleatoric uncertainty can further be classified as homoscedastic (constant across inputs) or heteroscedastic (input-dependent). However, in medical imaging most methods estimate the overall predictive uncertainty rather than explicitly modeling either one (Lambert et al., 2024). Common strategies involve modeling a predictive distribution through deep ensembles such as Monte-Carlo (MC) dropout, multiple stochastic forward passes at test time (Gal and Ghahramani, 2016; Jungo et al., 2018), or ensembles of differently initialized

models (Mehrtash et al., 2020). In PM segmentation, similar strategies have been applied using MC dropout (Klanecek et al., 2023) or deep ensembles, either from model snapshots along the training trajectory (Tang et al., 2025) or from models trained on different data distributions (Huemmer et al., 2024). However, these methods do not specifically model input-dependent (aleatoric) uncertainty to capture the aforementioned input ambiguities. A straightforward approach to model aleatoric uncertainty is to use the inter-rater variability as ground-truth for the uncertainty in a supervised setting as done in Cetindag et al. (2022). A more practical approach is to implicitly learn aleatoric uncertainty from the data itself. The underlying idea is to predict the mean and variance of the predictive distribution which is trained by maximizing the log-likelihood (LL) within a heteroscedastic framework (Lambert et al., 2024). This is usually achieved by adding the variance as separate output to the mean predictions and has been successfully applied to segmentation (Graham et al., 2020) and regression tasks (Seitzer et al., 2022). Nevertheless, to the best of our knowledge, heteroscedastic uncertainty modeling has not been applied to PM segmentation.

Uncertainty-aware coordinate regression: Following Nibali et al. (2018), coordinate regression typically follows two paradigms: direct regression of numerical labels or heatmap-matching, where landmarks are represented as Gaussian blobs. Direct regression allows straightforward loss computation, while heatmap matching leverages the spatial invariance of fully-convolutional networks (FCNs) and improves generalization. The soft-argmax operator (Luvizon et al., 2017) combines these advantages by estimating coordinates as the weighted sum of softmax-activated heatmaps, and has become one standard practice in heatmap regression (Wang et al., 2026). However, in the context of aleatoric uncertainty modeling, these pseudo-probabilities are usually only used for coordinate prediction while adding a separate head for (co-)variance prediction, which decouples localization from uncertainty estimation. Only a few approaches integrate both tasks within a single probabilistic heatmap. For instance, Thaler et al. (2021) demonstrated that aleatoric uncertainty can be learned from probabilistic heatmaps by allowing the model to modulate anisotropic Gaussian blobs instead of constraining them to a fixed variance. Nonetheless, this does not directly model heteroscedasticity and requires Gaussian fitting to produce robust uncertainty predictions. In the context of facial landmark detection, Kumar et al. (2020) reported attempts to estimate heteroscedastic covariance directly from heatmap predictions, but found this unstable due to the low resolution of current heatmap-based frameworks.

3. Methods

This study extends our work (Zech et al., 2025) about heteroscedastic uncertainty modeling in PM segmentation based on probabilistic heatmaps. More specifically, we leverage anatomical prior information about the pectoral muscle in MLO projections, which allows to reformulate the PM segmentation as a row-wise column-index (CI) regression of the PM boundary, as proposed in (Huemmer et al., 2024). Accordingly, we define the uncertainty by means of a row-wise variance over the predicted boundary positions. In contrast to modeling both the boundary prediction (mean) and the uncertainty (variance) with separate prediction heads, we encode the uncertainty by directly modeling row-wise probability distributions (see Figure 1).

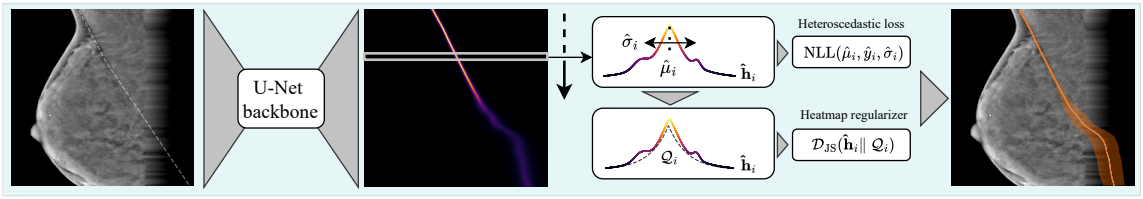


Figure 1: Overview of our method, with label-extrapolated input (left), predictive heatmap (center) and boundary prediction with uncertainty band (right).

The row-wise mean and variance are computed as first and second order momentum directly from a single heatmap:

$$\hat{\mu}_i = \sum_j j \cdot \hat{h}_{i,j}, \qquad \hat{\sigma}_i^2 = \sum_j \hat{h}_{i,j} \cdot (j - \hat{\mu}_i)^2 \tag{1}$$

$$\hat{h}_{i,j} = \mathrm{Softmax}(h_{i,j}) = \frac{\exp(h_{i,j})}{\sum_k \exp(h_{i,k})}, \tag{2}$$

with mean $\hat{\mu}_i$ as PM boundary prediction, variance $\hat{\sigma}_i^2$ as measure for the uncertainty, and $\hat{h}_{i,j}$ as value of the softmax-activated heatmap $\hat{h}$ at row i and column j. The mean computation in (1) corresponds to the soft-argmax operation (Luvizon et al., 2017) which has been proven effective in different scenarios as it leverages the spatial generalization of FCNs (Nibali et al., 2018). Consequently, we employ a U-Net (Ronneberger et al., 2015) to learn the probabilistic heatmap h in (2).

To model heteroscedastic uncertainty in the probabilistic heatmap, we train the network by minimizing the negative log-likelihood (NLL) derived from both Laplace ($\mathcal{L}$) and Normal ($\mathcal{N}$) distributions, both well known for this purpose (Kumar et al., 2020), with

$$\mathrm{NLL}_{\mathcal{N}} := \frac{1}{N} \sum_{i=1}^{N} \left(\frac{(y_i - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2} + \frac{1}{2} \log(2\pi\hat{\sigma}_i^2) \right) \quad \text{and} \tag{3}$$

$$\mathrm{NLL}_{\mathcal{L}} := \frac{1}{N} \sum_{i=1}^{N} \left(\frac{|y_i - \hat{\mu}_i|}{\hat{b}_i} + \log(2\hat{b}_i) \right), \quad \text{where } \hat{b}_i = \sqrt{\frac{\hat{\sigma}_i^2}{2}}. \tag{4}$$

Here, y_i refers to the target at row i. In this setting, the shape of the heatmaps is only weakly supervised, since infinitely many different distributions can yield the same first- and second-order moments, which may lead to unstable training behavior. To address this, we extend the loss function with a regularization term that enforces a Gaussian or Laplacian shape on the predicted distributions, thereby stabilizing soft-argmax-based training as demonstrated by Nibali et al. (2018). Unlike static variance regularization, our approach employs a variable regularization where both the mean and variance are controlled by the heteroscedastic loss (see Figure 1). In more details, we define the regularizer as

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{D}_{\mathrm{JS}}(\hat{\mathbf{h}}_i \| \ \mathcal{Q}_i) \quad \text{with} \quad \mathcal{Q}_i = \begin{cases} \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2) & \text{(Gaussian)} \\ \mathcal{L}(\hat{\mu}_i, \hat{b}_i) & \text{(Laplace).} \end{cases} \tag{5}$$

where $\mathcal{D}_{\mathrm{JS}}(\hat{\mathbf{h}}_i \| \ \mathcal{Q}_i)$ denotes the Jensen-Shannon divergence between $\hat{\mathbf{h}}_i$ defined as the softmax-activated heatmap in row i, and a template distribution $\mathcal{Q}_i$, which is constructed from the predicted heatmaps' statistics in (1). The regularizer, scaled by a constant factor λ to control its strength, is added to the heteroscedastic loss to encourage unimodal distributions that align with the probabilistic assumptions of the loss function.

Detecting multimodality: As shown in Figure 2(b), confounding structures can induce multimodal predictive distributions, since multiple locations may represent plausible anatomical interpretations of the PM. This facilitates to identify such uncertain cases by detecting multimodality in the predicted distribution, e.g., by classifying if there exists at least one row in the predicted heatmap that contains more than one peak. In this setting, the regularizing term serves as a soft constraint which controls the trade-off between enforcing unimodal predictions for accurate training while allowing the model to express genuine multimodality in the predictions.

Image-boundary handling: In our previous study (Zech et al., 2025), the label vector was clipped to the image boundary when the PM reached the image border. However, this leads to a strong systematic overconfidence near the edges that can be addressed by edge-padding the image by one-quarter of its size with a linear gradient to zero, and linearly extrapolating the label vector into this region as illustrated in Figure 1. This preserves a smooth muscle shape beyond the visible image area and provides a more robust basis for uncertainty quantification.

4. Experiments and results

To evaluate the proposed approach, we extracted 2,847 unprocessed MLO-view mammograms from the MBTST dataset (Dahlblom et al., 2019). Segmentation labels were provided by clinical experts as binary masks and converted to contour labels by taking the first non-zero column index in each row. The dataset was split into training, validation, and test sets using a ratio of 75%/15%/10% with a uniform distribution of breast densities. This dataset split is kept consistent across all experiments. For data augmentation, we employ the image processing pipeline by (Eckert et al., 2024) which allows sampling across various processing parameters which are kept fixed for inference. Furthermore, the images were cropped to the region of interest, resized to a resolution of 256×256 and subsequently Z-normalized. As optimizer we use AdamW with weight decay 10^{-2} and an initial learning rate of 10^{-5} which was reduced by a factor of 0.1 when the loss plateaued for more than 10 epochs. For all experiments we present the average metrics across three consecutive training runs. Throughout the experiments, the model uncertainty is quantified as mean-standard-deviation (MSTD) of the predicted heatmaps distributions.

4.1. Heatmap regression configuration

To identify the best configuration, we evaluate our heatmap regression framework using heteroscedastic losses ($\mathrm{NLL}_{\mathcal{L}}$ and $\mathrm{NLL}_{\mathcal{N}}$) and baseline losses—mean-absolute-error (MAE) and mean-squared-error (MSE)—along with varying regularization strengths λ. The heteroscedastic models are trained using label extrapolation, whereas the non-heteroscedastic models are trained without it to mitigate a potential performance degradation. Due to the lack of supervision from the loss function, the heatmap regularizers of the non-heteroscedastic

models were set to a fixed variance of $\sigma^2 = 10$, empirically chosen for optimal performance. The evaluation metrics MAE, root-mean-squared-error (RMSE) and LL are computed solely within the muscle region and summarized in Table 1.

Table 1: Results for different loss functions and regularization strengths. Metrics are reported as mean and (standard deviation) of 3 consecutive model runs.

Loss	λ	**MAE** $\downarrow$	**RMSE** $\downarrow$	**LL** $\uparrow$
MAE	0	1.97 (0.04)	2.46 (0.05)	-3.55 (0.19)
	10	1.94 (0.02)	2.40 (0.03)	-2.55 (0.07)
	100	1.90 (0.04)	2.35 (0.04)	-2.38 (0.03)
MSE	0	1.95 (0.07)	2.43 (0.06)	-3.85 (0.21)
	10	1.93 (0.02)	2.41 (0.04)	-3.43 (0.36)
	100	1.90 (0.01)	2.38 (0.03)	-2.83 (0.19)
$\text{NLL}_{\mathcal{L}}$	0	2.01 (0.00)	2.58 (0.02)	-2.30 (0.02)
	10	1.92 (0.04)	2.42 (0.05)	-2.28 (0.02)
	100	1.88 (0.04)	2.34 (0.05)	-2.26 (0.02)
$\text{NLL}_{\mathcal{N}}$	0	2.03 (0.03)	2.59 (0.10)	-2.31 (0.01)
	10	1.98 (0.01)	2.51 (0.01)	-2.32 (0.08)
	100	1.94 (0.09)	2.42 (0.11)	-2.30 (0.07)

The results show that the introduction of the regularization term leads to a slight performance improvement, as both MAE and RMSE decrease with increasing regularization strength across all used loss functions. A similar trend is observed for the LL values, which also increase with stronger regularization. Furthermore, both heteroscedastic loss formulations achieve MAE and RMSE values comparable to the non-heteroscedastic loss functions. At the same time, the heteroscedastic models show high LL values across all regularization strengths. Overall, the lowest errors and highest LLs are achieved by the model trained on $\text{NLL}_{\mathcal{L}}$. These findings suggest that incorporating heteroscedastic uncertainty modeling and label extrapolation does not compromise predictive performance while it allows learning stable predictive distributions across different regularization strengths. Further, it is demonstrated that the regularization term stabilizes training notably as evidenced by consistent improvement across all metrics and loss functions for higher regularization strengths. Among the evaluated configurations, the model trained with $\text{NLL}_{\mathcal{L}}$ achieves the lowest errors and highest LLs, indicating that the Laplace distribution is better suited to model the underlying heteroscedastic uncertainty. It is therefore selected as the best configuration for all subsequent experiments.

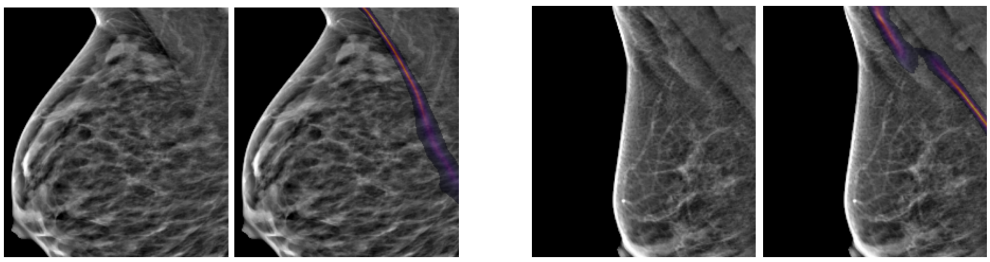
Detecting multimodality: To evaluate how the regularization term affects the number of detected multimodal distributions in the predicted heatmaps, we compare the results of the $\text{NLL}_{\mathcal{L}}$-model across different regularization strengths in Table 2.

The results show that for stronger regularization strengths, more cases are classified as unimodal. At the same time, the MAE and LL values worsen notably within the multimodal class while only marginally deteriorating for the unimodal class. This suggests that introducing the regularization term substantially stabilizes the predicted heatmaps towards

Table 2: Prediction performance for different λ of the $\text{NLL}_{\mathcal{C}}$-model according to the classification of the heatmap distribution. Metrics are reported as mean and (standard deviation). The bottom row shows the proportion of cases (%) classified as multi- vs. unimodal.

	Classification	$\lambda = 0.0$	$\lambda = 10.0$	$\lambda = 100.0$
MAE ↓	Multimodal	3.49 (0.11)	3.82 (0.03)	6.08 (1.32)
	Unimodal	1.72 (0.05)	1.78 (0.05)	1.85 (0.03)
LL ↑	Multimodal	−2.77 (0.06)	−2.84 (0.03)	−3.57 (0.66)
	Unimodal	−2.21 (0.03)	−2.24 (0.03)	−2.25 (0.02)
Proportion [%]	Multimodal	16.15	7.05	0.91
	Unimodal	83.85	92.95	99.09

unimodal predictions forcing the model to ignore smaller confounding structures, leaving only large confounders which are responsible for larger prediction errors. To further support this, we analyze two qualitative examples of high uncertainty from both classes for $\lambda = 100$, depicted in Figure 2. For the unimodal case in Figure 2(a), it is observed that the muscle shows a clear edge in the upper part while being occluded by breast tissue in the lower part of the muscle to which the model reacts with increased uncertainty. For the multimodal case in Figure 2(b), there are multiple confounding anatomical structures in the image leading to multimodal predictive distributions in the heatmap. This aligns with our initial assumption that in the confounding case the underlying estimation problem becomes inherently multimodal, whereas in the occluded scenario, the potential contour location can be adequately represented by a unimodal distribution. Furthermore, these results indicate that the model can capture multimodal distributions, and that the sensitivity to such multimodality can be adjusted via the regularization term.



(a) Unimodal case: muscle occluded by breast tissue.

(b) Multimodal case: confounding structures.

Figure 2: Qualitative examples illustrating (a) unimodal and (b) multimodal predictive distributions for the $\text{NLL}_{\mathcal{C}}$ ($\lambda = 100$)-model.

4.2. Performance comparison

To verify that our approach does not compromise segmentation performance, we compare it to a classical U-Net baseline (Ronneberger et al., 2015) trained to predict binary segmentation masks. In contrast to our method, the baseline is trained without label extrapolation on the original segmentation masks. We assess the segmentation performance in terms of DICE-coefficient across different network sizes in Figure 3.

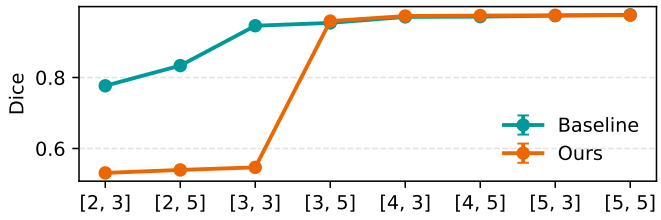


Figure 3: Segmentation performance compared to a classical U-Net baseline across different network sizes [depth, filters], where depth is the number of U-Net stages and filters is the number of filters in the first layer (doubled each stage).

For very small networks, segmentation performance of our method is noticeably lower than that of the baseline. This indicates that the model capacity is insufficient to handle the increased complexity of the task, which involves modeling both the segmentation contour and the associated uncertainty, as well as interpolating in regions where the muscle is not visible. In contrast, for medium and large network configurations, performance is on par with the baseline while the model simultaneously provides uncertainty estimates with high likelihoods as highlighted in Table 3.

Table 3: LL for different network configurations (filters × depth), reported as mean and (standard deviation).

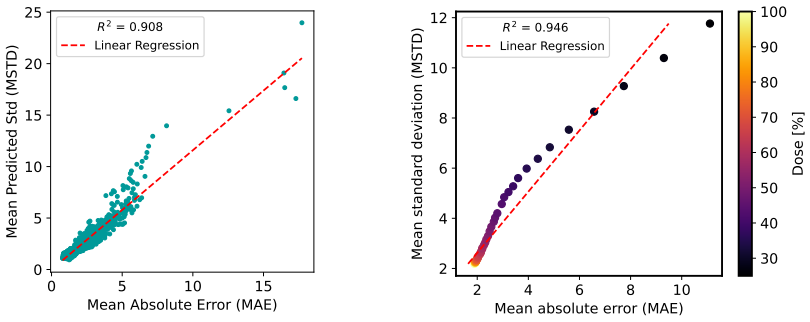
LL (filters \ depth)	**2**	**3**	**4**	**5**
3	-5.10 (0.01)	-4.95 (0.02)	-2.27 (0.01)	-2.25 (0.01)
5	-5.06 (0.00)	-2.51 (0.02)	-2.28 (0.04)	-2.26 (0.02)

4.3. Uncertainty quantification

In this section, we examine the method's ability to quantify uncertainty. We evaluate the model's response to (1) inherent dataset noise in the dataset, such as occlusions, and (2) artificially introduced unseen noise corrupting the input images. Practically, we assess how well the uncertainty predicts model error by analyzing the correlation between the mean residuals (MAE) and the predicted MSTD. The experiments are conducted on the $\text{NLL}_{\mathcal{L}}$ ($\lambda = 100$)-model (Sec. 4.1) and the results are shown in Figure 4.

Model uncertainty within the dataset: To analyze model uncertainty we employ a similar strategy to Kumar et al. (2020) by collecting all row-wise residual errors and predicted heatmap standard deviations for every image in the test set. We then sort these tuples by standard deviation and group them into equally sized bins of $N_{\text{bin}} = 50$. Within each bin, we compute the average residual error (MAE) and the MSTD; each bin therefore corresponds to a single point in Figure 4(a).

Model uncertainty for unseen noise: Unseen noise is modeled by artificially reducing the photon counts and adding X-ray typical Poisson noise to the input images during inference using the procedure described by Eckert et al. (2019). Therefore, we linearly reduce the effective dose from 100% to 25% in 50 steps while producing 3 realizations with different seeds for each dose level. Finally, the models predict on all noised image realizations and the results are aggregated over all images. Hence, each dot in Figure 4(b) represents the whole test set at the respective noise level.



(a) Model uncertainty within the dataset.

(b) Model uncertainty for unseen noise.

Figure 4: Analysis of the correlation between MAE and MSTD for 4(a) inherent noise in the dataset and 4(b) unseen noise distorting the input for the $\text{NLL}_{\mathcal{L}}(\lambda = 100)$-model.

For the uncertainty within the test set in Figure 4(a), the results show a strong linear correlation between the MAE and MSTD, with the MSTD consistently matching or slightly exceeding the MAE. A similar behavior can be observed for the model's response to previously unseen noise in Figure 4(b). Artificially reducing the dose leads to a continuous increase in model error to which the model responds with a steady increase in the MSTD of the predicted heatmap. These results indicate that the standard deviations computed from the models' predictive distributions are highly predictive for model error for both inherent heteroscedastic noise within the dataset and previously unseen noise. It should be noted that this finding holds as an aggregated behavior when averaging over N_{bins} for Figure 4(a) or the whole test set for Figure 4(b).

To complement the quantitative analysis, we examine the model's behavior at the image level using an exemplary image under varying noise levels shown in Figure 5. Up to a 50% dose reduction, the model produces accurate predictions with low uncertainty along the entire contour. When further reducing the dose to 35%, the model's uncertainty starts to increase locally in the lower part where the muscle is originally slightly occluded, while

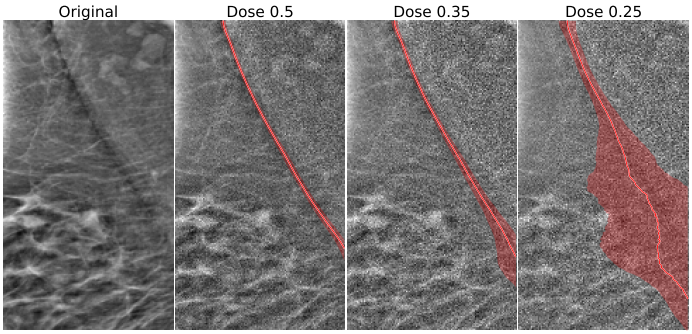


Figure 5: Predictions of the $\text{NLL}_{\mathcal{L}}$ ($\lambda = 100$)-model for an example case from the test set across different noise levels. The mean $\hat{\mu}$ is depicted as thick red line and the heatmap's standard deviation $\hat{\sigma}$ as light red area around the mean.

still producing a stable contour prediction. At 25% dose, the contour prediction degrades notably in the lower part of the muscle but at the same time the model increases uncertainty significantly as a sign for prediction failure. The observations suggest that the model is able to adjust uncertainty locally in areas of reduced visibility, while maintaining stable and confident predictions in regions where the muscle remains clearly visible.

5. Conclusion and Outlook

In this work, we present a novel framework for modeling input-dependent uncertainty in PM segmentation based on heteroscedastic regression. We show that robust mean and variance estimates can be derived from learned probabilistic heatmaps to jointly model the PM boundary and the associated predictive uncertainty. Furthermore, by representing uncertainty directly in probabilistic heatmaps, the method provides richer information than approaches that output mean and variance as isolated numerical values, as it allows to detect inherent multimodality in the predictive distributions and allows this behavior to be controlled through a dedicated regularizer. At the same time, we show that our method does not compromise segmentation performance as we achieve on par performance with a binary segmentation baseline. Last, we show that the model's uncertainty estimates correlate with model error in a global trend and demonstrate that the model reacts appropriately to previously unseen noise, increasing its predicted uncertainty when reduced visibility of the PM leads to erroneous predictions.
Although the current framework mainly models unimodal distributions, it establishes a strong foundation for future research on extending the approach to inherently capture multimodality in the predictive distributions. Moreover, evaluating the model outputs against multi-reader annotations could provide meaningful insights about the methods capability to model localized uncertainty.

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

References

Francesca Angelone, Alfonso Maria Ponsiglione, Roberto Grassi, Francesco Amato, and Mario Sansone. U-net based approach for pectoralis muscle segmentation in digital mammography. *Computer Methods and Programs in Biomedicine Update*, 8:100210, 2025. ISSN 26669900. doi: 10.1016/j.cmpbup.2025.100210.

Mouna Brahim, Kai Westerkamp, Louisa Hempel, Reiner Lehmann, Dirk Hempel, and Patrick Philipp. Automated Assessment of Breast Positioning Quality in Screening Mammography. *Cancers*, 14(19):4704, September 2022. ISSN 2072-6694. doi: 10.3390/cancers14194704. URL https://www.mdpi.com/2072-6694/14/19/4704.

Sabri Can Cetindag, Mert Yergin, Deniz Alis, and Ilkay Oksuz. Meta-learning for Medical Image Segmentation Uncertainty Quantification. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, volume 12963, pages 578–584. Springer International Publishing, Cham, 2022. ISBN 978-3-031-09001-1 978-3-031-09002-8. doi: 10.1007/978-3-031-09002-8_51. URL https://link.springer.com/10.1007/978-3-031-09002-8_51. Series Title: Lecture Notes in Computer Science.

Victor Dahlblom, Magnus Dustler, and Sophia Zackrisson. Mammograms from cancer screening. Available: AIDA Data Hub, 2019. URL https://datahub.aida.scilifelab.se/10.23698/aida/mbtst-dm.

Dominik Eckert, Sulaiman Vesal, Ludwig Ritschl, Steffen Kappler, and Andreas Maier. Deep Learning-based Denoising of Mammographic Images using Physics-driven Data Augmentation, 2019. URL https://arxiv.org/abs/1912.05240. Version Number: 1.

Dominik Eckert, Christopher Syben, Christian Hümmer, Ludwig Ritschl, Steffen Kappler, and Sebastian Stober. Stylex: A trainable metric for x-ray style distances. *arXiv*, 2024. doi: 10.48550/ARXIV.2405.14718.

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*, 42:201–208, 2004.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd international conference on machine learning*, volume 48 of *Proceedings of machine learning research*, pages 1050–1059, New York, New York, USA, June 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Karthikeyan Ganesan, U. Rajendra Acharya, Kuang Chua Chua, Lim Choo Min, and K. Thomas Abraham. Pectoral muscle segmentation: A review. *Computer Methods and Programs in Biomedicine*, 110(1):48–57, April 2013. ISSN 01692607. doi: 10.1016/j.cmpb.2012.10.020.

Mark S. Graham, Carole H. Sudre, Thomas Varsavsky, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Hierarchical Brain Parcellation with

Uncertainty. In Carole H. Sudre, Hamid Fehri, Tal Arbel, Christian F. Baumgartner, Adrian Dalca, Ryutaro Tanno, Koen Van Leemput, William M. Wells, Aristeidis Sotiras, Bartlomiej Papiez, Enzo Ferrante, and Sarah Parisot, editors, *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, volume 12443, pages 23–31. Springer International Publishing, Cham, 2020. ISBN 978-3-030-60364-9 978-3-030-60365-6. doi: 10.1007/978-3-030-60365-6_3. URL https://link.springer.com/10.1007/978-3-030-60365-6_3. Series Title: Lecture Notes in Computer Science.

Yongze Guo, Wenhui Zhao, Songfeng Li, Yaqin Zhang, and Yao Lu. Automatic segmentation of the pectoral muscle based on boundary identification and shape prediction. *Physics in Medicine & Biology*, 65(4):045016, February 2020. ISSN 1361-6560. doi: 10.1088/1361-6560/ab652b. URL https://iopscience.iop.org/article/10.1088/1361-6560/ab652b.

Christian Huemmer, Ramyar Biniazan, Manasi Datar, Martin Kraus, Andreas Fieselmann, and Steffen Kappler. Improved Pectoral Muscle Segmentation in Mammograms through Regression-Based Deep Learning and Knowledge Distillation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, Athens, Greece, May 2024. IEEE. ISBN 9798350313338. doi: 10.1109/ISBI56570.2024.10635448.

Alain Jungo, Richard McKinley, Raphael Meier, Urspeter Knecht, Luis Vera, Julián Pérez-Beteta, David Molina-García, Víctor M. Pérez-García, Roland Wiest, and Mauricio Reyes. Towards Uncertainty-Assisted Brain Tumor Segmentation and Survival Prediction. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, volume 10670, pages 474–485. Springer International Publishing, Cham, 2018. ISBN 978-3-319-75237-2 978-3-319-75238-9. doi: 10.1007/978-3-319-75238-9_40. URL http://link.springer.com/10.1007/978-3-319-75238-9_40. Series Title: Lecture Notes in Computer Science.

N Karssemeijer. Automated classification of parenchymal patterns in mammograms. *Physics in Medicine and Biology*, 43(2):365–378, February 1998. ISSN 0031-9155, 1361-6560. doi: 10.1088/0031-9155/43/2/011.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, October 2017. URL http://arxiv.org/abs/1703.04977. arXiv:1703.04977 [cs].

Zan Klanecek, Tobias Wagner, Yao-Kuan Wang, Lesley Cockmartin, Nicholas Marshall, Brayden Schott, Ali Deatsch, Andrej Studen, Kristijana Hertl, Katja Jarm, Mateja Krajc, Miloš Vrhovec, Hilde Bosmans, and Robert Jeraj. Uncertainty estimation for deep learning-based pectoral muscle segmentation via Monte Carlo dropout. *Physics in Medicine & Biology*, 68(11):115007, June 2023. ISSN 0031-9155, 1361-6560. doi: 10.1088/1361-6560/acd221. URL https://iopscience.iop.org/article/10.1088/1361-6560/acd221.

Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8233–8243, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00826. URL https://ieeexplore.ieee.org/document/9157108/.

Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. *Artificial Intelligence in Medicine*, 150: 102830, April 2024. ISSN 09333657. doi: 10.1016/j.artmed.2024.102830. URL https://linkinghub.elsevier.com/retrieve/pii/S0933365724000721.

Wei Liu, Chengqian Liu, and Yiran Wei. Utilizing Deep Learning Technology to Segment Pectoral Muscle in Mediolateral Oblique View Mammograms. In *2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP)*, pages 97–101, Nanjing, China, October 2020. IEEE. ISBN 978-1-7281-6896-8. doi: 10.1109/ICSIP49896.2020.9339411. URL https://ieeexplore.ieee.org/document/9339411/.

Diogo C. Luvizon, Hedi Tabia, and David Picard. Human Pose Regression by Combining Indirect Part Detection and Contextual Information, October 2017. URL http://arxiv.org/abs/1710.02322. arXiv:1710.02322 [cs].

Xiangyuan Ma, Jun Wei, Chuan Zhou, Mark A. Helvie, Heang-Ping Chan, Lubomir M. Hadjiiski, and Yao Lu. Automated pectoral muscle identification onmlo-view mammograms: Comparison of deep neural network to conventional computer vision. *Medical Physics*, 46 (5):2103–2114, May 2019. ISSN 0094-2405, 2473-4209. doi: 10.1002/mp.13451.

Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, December 2020. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2020.3006437. URL https://ieeexplore.ieee.org/document/9130729/.

Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical Coordinate Regression with Convolutional Neural Networks, May 2018. URL http://arxiv.org/abs/1801.07372. arXiv:1801.07372 [cs].

Andrik Rampun, Karen López-Linares, Philip J. Morrow, Bryan W. Scotney, Hui Wang, Inmaculada Garcia Ocaña, Grégory Maclair, Reyer Zwiggelaar, Miguel A. González Ballester, and Iván Macía. Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network. *Medical Image Analysis*, 57:1–17, October 2019. ISSN 13618415. doi: 10.1016/j.media.2019.06.007.

Wenhui Ren, Mingyang Chen, Youlin Qiao, and Fanghui Zhao. Global guidelines for breast cancer screening: A systematic review. *The Breast*, 64:85–99, August 2022. ISSN 09609776. doi: 10.1016/j.breast.2022.04.003.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pages 234–241. Springer International Publishing, Cham, 2015. ISBN 978-3-319-24573-7. doi: 10.1007/978-3-319-24574-4_28. URL http://link.springer.com/10.1007/978-3-319-24574-4_28. Series Title: Lecture Notes in Computer Science.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks, April 2022. URL http://arxiv.org/abs/2203.09168. arXiv:2203.09168 [cs].

R.L. Siegel, K.D. Miller, and A. Jemal. Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66(1):7–30, 2016.

Hossein Soleimani and Oleg V. Michailovich. On Segmentation of Pectoral Muscle in Digital Mammograms by Means of Deep Learning. *IEEE Access*, 8:204173–204182, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3036662. URL https://ieeexplore.ieee.org/document/9252130/.

Yutao Tang, Yongze Guo, Huayu Wang, Ting Song, and Yao Lu. Uncertainty-Aware Semi-Supervised Method for Pectoral Muscle Segmentation. *Bioengineering*, 12(1):36, January 2025. ISSN 2306-5354. doi: 10.3390/bioengineering12010036. URL https://www.mdpi.com/2306-5354/12/1/36.

Franz Thaler, Christian Payer, Martin Urschler, and Darko Štern. Modeling Annotation Uncertainty with Gaussian Heatmaps in Landmark Localization. *Machine Learning for Biomedical Imaging*, 1(UNSURE2020):1–27, September 2021. ISSN 2766-905X. doi: 10.59275/j.melba.2021-77a7. URL https://melba-journal.org/2021:014.

Anbang Wang, Marawan Elbatel, Keyuan Liu, Lizhuo Lin, Meng Lan, Yanqi Yang, and Xiaomeng Li. Geometric-Guided Few-Shot Dental Landmark Detection with Human-Centric Foundation Model. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964, pages 197–207. Springer Nature Switzerland, Cham, 2026. ISBN 978-3-032-04970-4 978-3-032-04971-1. doi: 10.1007/978-3-032-04971-1_19. URL https://link.springer.com/10.1007/978-3-032-04971-1_19. Series Title: Lecture Notes in Computer Science.

Lei Wang, Miao-liang Zhu, Li-ping Deng, and Xin Yuan. Automatic pectoral muscle boundary detection in mammograms based on markov chain and active contour model. *Journal of Zhejiang University SCIENCE C*, 11(2):111–118, February 2010. ISSN 1869-1951, 1869-196X. doi: 10.1631/jzus.C0910025.

World Health Organization. *WHO position paper on mammography screening*. World Health Organization, Geneva, 2014. ISBN 978-92-4-150793-6.

Paul Zech, Christian Huemmer, Christopher Syben, Ludwig Ritschl, and Sebastian Stober. Uncertainty-aware pectoral muscle segmentation based on heteroscedastic regression. In *Medical Imaging with Deep Learning - Short Papers*, 2025. URL https://openreview.net/forum?id=dlnlmPxgvm.