# Weakly Supervised Training with Explainable Artificial Intelligence to Predict Breast-Cancer Response to Neoadjuvant Chemotherapy

Minor Research Project
Medical Imaging

**Yan Novikov**

**Examination committee:**

**Supervisor: Dr. Kenneth Gilhuijs**
Associate professor, UMC Utrecht
**Second Reviewer: Dr. Alberto de Luca**
Assistant professor, UMC Utrecht

Utrecht University

The Netherlands

# Weakly Supervised Training with Explainable Artificial Intelligence to Predict Breast-Cancer Response to Neoadjuvant Chemotherapy

Yan Novikov (9391355)

## Abstract

This study investigated the feasibility of weakly-supervised deep regression for predicting patient responses to neoadjuvant chemotherapy (NAC) using Maximum Intensity Projection (MIP) images. We used radiological tumor volume ratio (RTVR) and Residual Cancer Burden (RCB) to represent radiological and pathological responses to NAC, respectively. We conducted three experiments, two with single-task regression of RTVR and RCB and one with multi-task regression. Each experiment involved training a model based on a resnet14t architecture to minimize Batch Monte-Carlo (BMC) loss designed for imbalanced regression. We evaluated the performance of each model using Spearman's correlation and Bland–Altman analysis. Spearman's correlation coefficients were calculated for the hold-out test set and were $\rho = 0.47$ for the RTVR single-task model, $\rho = 0.23$ for the RCB single-task model, and $\rho = 0.61$ and $\rho = 0.34$ for RTVR and RCB respectively, in the multi-task model. Despite the multi-task model showing a slightly better correlation, we observed a statistically significant difference neither for predicting RTVR values (P = 0.49) nor for RCB scores (P = 0.55). Deep SHapley Additive exPlanations (SHAP) provided insight into the models' decision-making processes. The results indicated that the current method could not provide clinically meaningful outputs. We discussed potential reasons for this poor performance and possible future research directions.

## Index Terms

Weakly-Supervised Learning, Breast Cancer, Residual Cancer Burden, Neoadjuvant Chemotherapy

## I. Introduction

The use of neoadjuvant chemotherapy (NAC) to treat patients with breast cancer is on the rise. This strategy makes breast-conserving surgery more feasible and allows for monitoring an individual patient's response [1]. Recent research has demonstrated that the pathological and radiological reactions to NAC can both yield important prognostic data. Achieving a pathological complete response (pCR) is associated with improved survival [2]. And while radiologic complete response (rCR) on contrast-enhanced MRI (CE-MRI) following NAC does not consistently predict pCR in early breast cancer, it strongly correlates with recurrence-free and overall survival [3] [4]. Quantitative metrics exist to measure different levels of response to NAC. The Residual Cancer Burden (RCB) score and its categorical counterpart, the RCB class, are utilized to quantify the pathological response [5]. The radiological response, on the other hand, can be gauged by the changes visible on MRI. Recent studies indicated that the change in primary tumor volume on MRI is positively associated with both recurrence-free and overall survival, as is the RCB class [6] [7].

Deep learning-based segmentation can be an effective tool for the automated assessment of breast-cancer response to NAC on breast MRI [8]. In case of incorrect segmentation, manual correction by an operator is possible. However, while the operator can remove erroneously segmented structures, the ground truth remains unknown in clinical settings. Since segmentation is not essential in predicting the response to NAC, one can skip it but still give an estimation, which a professional can check and correct.

This study aims at assessing the feasibility of predicting the radiological and pathological responses on breast CE-MRI with a weakly-supervised neural regression followed by an explainability step. The regression values are the residual tumor volume ratio (RTVR) obtained from MR scans and the RCB score determined by a pathologist. We assume that the responses are correlated and compare the performance of a multi-task model predicting both values with that of models for each score.

## II. Materials and Methods

### A. Patients

This work utilizes data from the BOGOTA study, which included $147$ female patients treated for Locally Advanced Breast Cancer (LABC) with NAC between January 1, 2011, and December 1, 2019, with the mean age of diagnosis of $50$ years (range $25 - 73$ years) [8]. Having excluded $43$ of them due to various reasons, we got left with $46$ and $55$ patients with left and right laterality of cancer, respectively, and three patients with bilateral cancer. This resulted in $104$ patients or $107$ affected breasts being available for this study.

### B. Magnetic Resonance Imaging

Throughout treatment, the patients underwent two MRI examinations, with the first before NAC and the second either halfway through the chemotherapy schedule or right before the second-to-last cycle of chemotherapy, depending on the NAC schedule [8]. The imaging was performed using $1.5$ T and $3$ T MRI units (Philips Ingenia or Achieva) with a dedicated breast coil. Dynamic contrast-enhanced T1-weighted MRI series consisted of one pre-contrast scan and at least five post-contrast scans after injection of a gadolinium-based contrast agent (Gadobutrol, $0.1$ mmol/kg). The post-contract scans were acquired at intervals between $60$ s and $90$ s. The ranges of the imaging parameters were: repetition time $3.3$ ms to $7.1$ ms, echo time $1.2$ ms to $3.4$ ms, flip angle $8°$ to $10°$, field of view $340$ mm to $426.7$ mm, voxel volumes $0.75 \times 0.75 \times 0.90$ mm$^3$ to $0.97 \times 0.97 \times 1.30$ mm$^3$. All sequences employed fat suppression.

### C. Maximum Intensity Projections

This work does not directly utilize the raw MR scans. Two-dimensional MIP images obtained by selecting the highest intensity values along the transverse axis serve as input data instead. The images are aligned so that the sternum appears at the sagittal center, and the coronal bottom point lies 5 cm below the sternum. To reduce the variation between different MR scanners, we rescaled the MIP images to the modal resolution of approximately $0.89$ mm per pixel in each direction. Then, we used cropping or zero padding to achieve the field of view to $448$ px and $224$ px in sagittal and coronal directions, respectively, while keeping the bottom center point fixed. We divided each image into two halves of size $224$ px $\times$ $224$ px and picked the one with cancer. Patients with bilateral cancer yielded two images. The resulting right breast images underwent horizontal flipping. Finally, we normalize each image to zero mean and unit variance.

### D. Residual Tumor Volume Ratios

A dedicated Biomedical Engineer analyzed the original MR scans and manually delineated the lesions. We calculated tumor volumes for each case by multiplying the number of nonzero voxels in each delineation by the size of one voxel in mm$^3$. Dividing the residual volumes by the initial ones yielded the ratios this study aims to predict. Their distribution is depicted in Fig. 1. For the deep learning experiments, we normalized the values by dividing them with rounded to first decimal their standard deviation in the training subset, which was $0.3$.

### E. Residual Cancer Burden

RCB scores were derived from the final resection specimens by a dedicated breast pathologist with 30 years of experience (PJvD) [8], following the methodology described by [5]. The distribution of the RCB discretized as described in the literature [5] is shown in Fig. 1. Similarly to the RTVR, we divided the scores by their standard deviation in the training subset rounded to the first decimal, which was $1.2$.

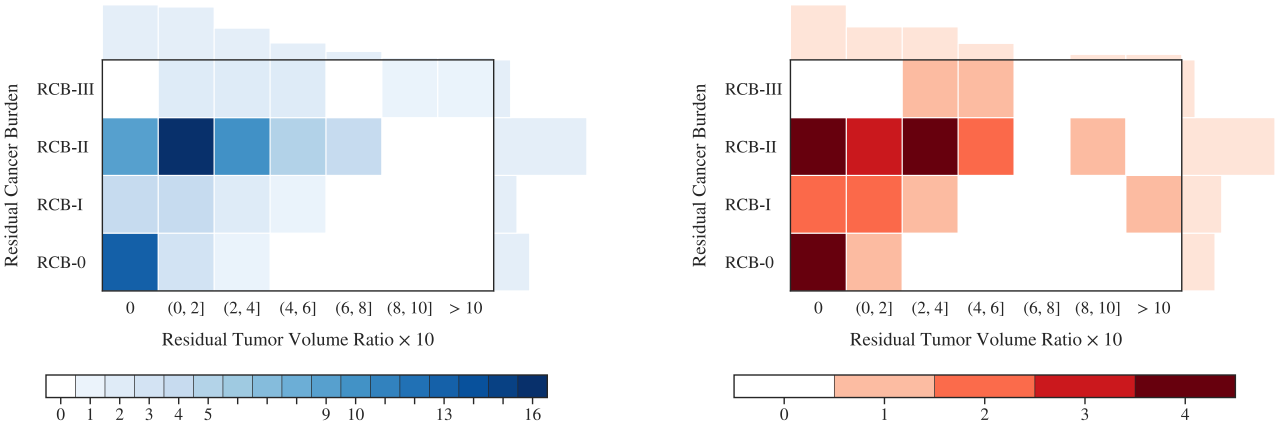
### F. Data split

Before proceeding to divide the data into training and testing subsets, we initially isolated the three bilateral cases. We then selected $80$ of the remaining $101$ patients for the training subset. To do this, we performed a stratified data split with four classes. These classes corresponded to zero/zero, zero/nonzero, nonzero/zero, and nonzero/nonzero values of residual tumor volumes and RCB scores, respectively. Finally, we included the initially set aside bilateral cases in the test subset, bringing it to a total of $24$ patients or $27$ breasts.

### G. Weakly Supervised Deep Regression

The model we use for all deep learning experiments consists of a convolutional backbone and a linear head. As the backbone, we utilize a resnet14t from PyTorch Image Models (**timm**) library [9]. Its architecture corresponds to a residual network with a bottleneck building block [10] modified with the ResNet-C tweak from [11]. This tweak replaces the input stem $7 \times 7$ convolution with three $3 \times 3$ convolutions, which reduces the computational cost. We start each experiment with the weights pre-trained on ImageNet-1k [12] following procedure C described in [13]. The head takes $2048$-channel input and consists of the three fully connected layers with $128$, $128$, and $1$ channel outputs. A dropout precedes each fully connected layer, and a rectified linear unit (ReLU) follows each except for the last one. In the case of multi-task learning, each task has its dedicated head.

To tackle the data imbalance problem, we use the Batch-based Monte Carlo (BMC) loss function. This loss is one of the Balanced MSE implementations proposed in [14]. The original paper claims it is the first general solution to one- or multi-dimensional imbalanced regression. The chosen implementation is the only one that requires no prior knowledge of the target distribution and better fits in our case of limited data. The loss contains a noise scale parameter $\sigma_{noise}$ jointly optimized in the training process. However, this loss is hard to interpret, so we use another balanced metric to evaluate the performance. This metric is the bin-based Euclidean distance between the output and the target values. The errors are averaged for the



(a) Training Subset

(b) Test Subset

Fig. 1: A joint histogram of the imbalanced regression targets. The RCB classes correspond to the continuous RCB scores thresholded with 0, 1.36, and 3.28, which accords to the procedure from [5]. The bins for the RTVR values are arbitrarily selected. The only exception is that zero and nonzero values are kept separate on purpose. The color represents the counts. Only the counts presented on the joint histogram have a label.

targets in each bin from Fig. 1a, and the mean of the obtained per-bin scores serve as the final performance measure. For one-dimensional regression, Euclidean distance equals L1 error.

We try to continue following procedure C from [13] used in pre-training. To optimize the model, we use SGD with Nesterov's momentum [15]. Adaptive Gradient Clipping (AGC) introduced in [16] applies to the convolutional backbone on each optimization step. Weight decay (L2 penalty) serves the required regularization. We diverge from the procedure in the batch size choice due to the small number of available training images. For the sake of ease, we also do not employ a cosine learning rate decay.

The data augmentation we apply consists of three transformations. The first is contrast adjustment via gamma transformation with gamma uniformly sampled from between $1 - \Delta\gamma_{max}$ and $1 + \Delta\gamma_{max}$. The second is adding Gaussian noise with zero mean and variance uniformly selected from zero to $\sigma_{max}$. Finally, we utilize bilinear downsampling to a random size between $s_{min} \times s_{min}$ and $224 \times 224$ followed by bilinear upsampling to the original size of $224 \times 224$. We apply these transformations in a random order, each with the probability of $p$. We normalize the intensities to zero minimum before the gamma correction. After each transformation, we again set the mean and variance of the processed images to zero and one, respectively.

In all three experiments, we optimize the hyperparameters with Optuna [17]. We utilize the k-folds technique with four folds, each consisting of 60 k$^{th}$ training/validation and 20 k$^{th}$ test samples. We stratify the splits with the four classes used for the initial data split II-F. Each hyperparameter set is assigned a score of mean test error in the five runs. We prune a trial if the k$^{th}$ test error is below 75$^{th}$ percentile of the k$^{th}$ test errors in the previous trials. Since we have no separate validation subset, we use the non-augmented training data for early stopping in case of no improvement in the validation error for 15 epochs. After 120 epochs, we stop training regardless of the validation error. We utilize TPESampler, which uses a Tree-structured Parzen Estimator algorithm, for the following hyperparameters: learning rate, momentum, weight decay, AGC value, dropout rate, initial $\sigma_{noise}$ for the BMC loss and its learning rate, the augmentation parameters $\Delta\gamma_{max}$, $\sigma_{max}$, $s_{min}$, and $p$. We try 500 hyperparameter sets per experiment. The batch size is fixed in all experiments and equals 20.

### H. Statistical Analysis

Similarly to [18], we evaluated the performance of each trained model by comparing its RTVR and RCB predictions with the ground truth values using Spearman's correlation coefficient. In addition, we estimated the bias and variance of the algorithm by implementing non-parametric Bland–Altman plots [19].

To further compare the performance of the two unidimensional regression models with that of the multi-task model, we employed Bootstrap Resampling. This non-parametric method generated a confidence interval for the difference between Spearman's correlation coefficients calculated for the two models. After 1000 iterations, we established a distribution of these differences and calculated a 95% confidence interval [20].

### I. Shapley Additive Explanations

We used Deep SHapley Additive exPlanations (SHAP) to explain the model outputs [21]. In simple terms, SHAP measures the impact of a feature on a particular prediction compared to the average for the dataset. It relies on the concept of Shapley values from cooperative game theory [18]. It individually treats the marginal contribution of every pixel to the predicted RTVR or RCB score. These contributions can be positive and negative.

Utilizing Deep SHAP, we created a SHAP-values map for each test MIP image and each regression value. A pixel in this map represents how this particular pixel contributed to the model's output. Higher values indicate higher predictions, while lower values correspond to lower predictions [18].

We used all 80 validation (non-augmented training) images to provide the required background signal for Deep SHAP analysis [21].

## III. Results

### A. *Weakly Supervised Deep Regression*

The optimal hyperparameter values shown in Table I were comparable among the three deep-learning experiments. Only the RTVR predicting model required strong L2 regularization with optimal Weight Decay being $0.1$. Only the multi-task regression model performed better with no dropout. Optimal initial $\sigma_{noise}$ values for the BMC loss were close to zero, which makes the loss similar to the basic MSE at the beginning of the training.

| | | | | | **Experiments** | | |
|---|---|---|---|---|---|---|---|
| **Parameter** | **Scale** | **Lower** | **Upper** | **Step** | **RTVR** | **RCB** | **Multi-task** |
| Learning Rate | $\log_{10}$ | $-4.0$ | $-2.0$ | $0.5$ | $-3.5$ | $-2.5$ | $-3.5$ |
| Momentum | 1 | $0.80$ | $0.95$ | $0.05$ | $0.90$ | $0.80$ | $0.85$ |
| Weight Decay | $\log_{10}$ | $-5.0$ | $-1.0$ | $0.5$ | $-1.0$ | $-4.5$ | $-5.0$ |
| AGC Value | $\log_{10}$ | $-5.0$ | $-1.0$ | $0.5$ | $-2.5$ | $-4.5$ | $-2.5$ |
| Dropout Rate | 1 | $0.0$ | $0.2$ | $0.1$ | $0.2$ | $0.2$ | $0.0$ |
| Initial $\sigma_{noise}$ | $\log_{10}$ | $-2.0$ | $0.0$ | $0.5$ | $-2.0$ | $-2.0$ | $-1.5$ |
| Learning Rate for $\sigma_{noise}$ | $\log_{10}$ | $-6.0$ | $-4.0$ | $0.5$ | $-4.5$ | $-5.5$ | $-5.0$ |
| $\Delta\gamma_{max}$ | 1 | $0.0$ | $0.5$ | $0.1$ | $0.2$ | $0.4$ | $0.5$ |
| $\sigma_{max}$ | 1 | $0.0$ | $0.5$ | $0.1$ | $0.4$ | $0.5$ | $0.2$ |
| $s_{min}$ | 1 | 56 | 224 | 28 | 140 | 56 | 196 |
| $p$ | 1 | $0.0$ | $1.0$ | $0.1$ | $0.6$ | $0.5$ | $0.5$ |

TABLE I: Hyperparameter search settings and results for the three deep learning experiments. Logarithmic optimization is used for some parameters, as indicated.

In the hold-out test set, the Spearman's correlation between the predicted RTVR and the ground truth RTVR in single- and multi-task regression experiments were $\rho = 0.47$ (P $< 0.05$) and $\rho = 0.61$ (P $< 0.05$), respectively (Fig. 2a). Bootstrap Resampling showed no significant difference between the $\rho$ values (P $= 0.49$). The ground truth RCB scores and those predicted by single- and the multi-task models were less monotonically related, with Spearman's $\rho = 0.23$ (P $= 0.24$) and $\rho = 0.34$ (P $= 0.08$), respectively (Fig. 2b). Bootstrap Resampling also indicated no significant difference between these $\rho$ values (P $= 0.55$).

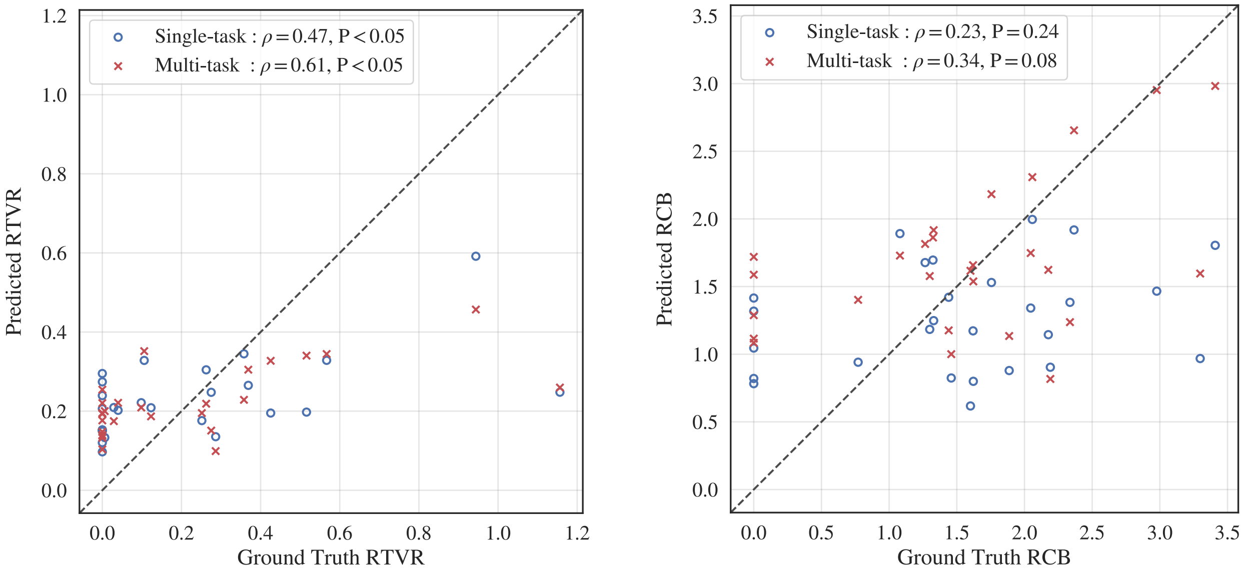
Nonparametric Bland–Altman analysis showed that RTVR predicted by single- and multi-task models had mean biases of $-0.01$ ($95\%$ limits of agreement $= -0.52$ to $= 0.49$) and $-0.01$ ($95\%$ limits of agreement $= -0.50$ to $= 0.49$), respectively, compared to the ground truth RTVR (Fig. 3a). The predicted RCB had mean biases compared to the ground truth RCB higher than did RTVR. They were $0.26$ ($95\%$ limits of agreement $= -1.59$ to $= 2.11$) and $-0.15$ ($95\%$ limits of agreement $= -1.80$ to $= 1.51$) in uni- and multi-dimensional regression, respectively (Fig. 3b).

### B. *Shapley Additive Explanations*

Explanations of the models' outputs showed that tumor and tumor-related structures, such as adjacent vessels, strongly influenced the predictions (Fig. 4). These structures in both input channels affected RTVR and RCB positively. More visible tumor tissue on the pre-treatment MIP images led to smaller outputs only in the RTVR-predicting model. The multi-task model did not focus on the volume ratios and made decisions based on other features. The explanations of the RCB-predicting model were similar to those of the multi-task model predicting both scores.

## IV. Discussion

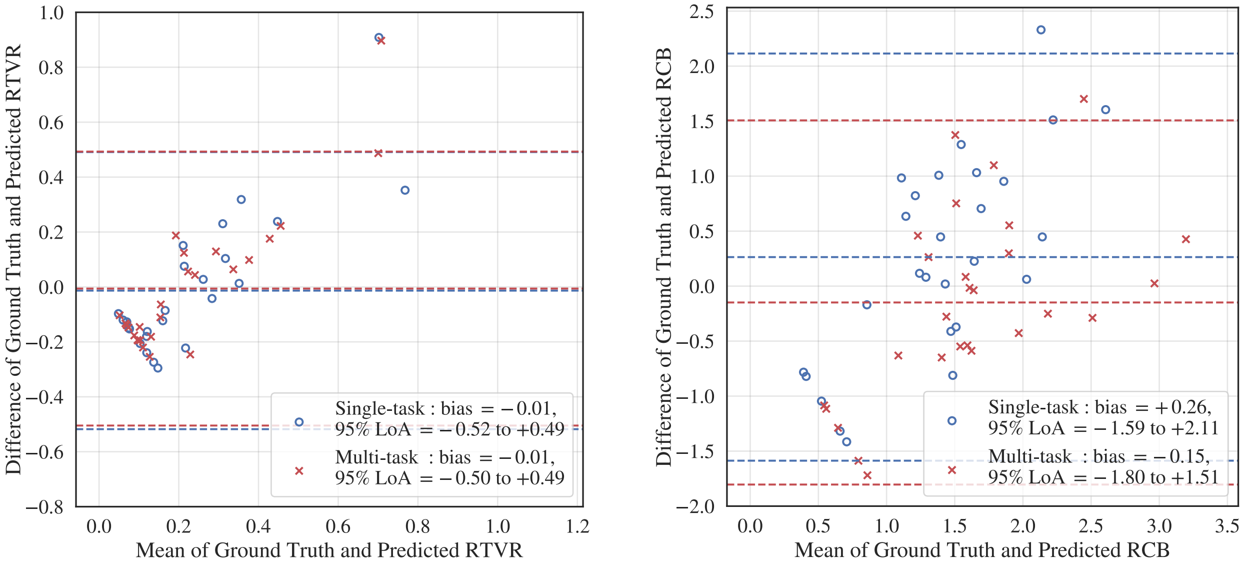
We demonstrated the performance of the resnet14t trained on limited and imbalanced data in a weakly-supervised deep regression of RTVR and RCB values. We compared the predictive ability of the models trained with each of these values as targets and that of the multi-task model. The Spearman's correlation coefficients were relatively low, showing little monotonical dependence between the predicted and ground truth values in all experiments. The multi-task model achieved a slightly better



(a) RTVR Regression

(b) RCB Regression

Fig. 2: The correlation between values predicted and ground truth values in the hold-out test set ($N = 27$). The images also depict Spearman's $\rho$ and P in different experiments.



(a) RTVR Regression

(b) RCB Regression

Fig. 3: Bland-Altman analysis shows mean biases and 95% limits of agreement (LoA) for RTVR and RCB values predicted in uni-dimensional and multi-dimensional regressions compared to corresponding ground truth.

correlation in both regressions, but Bootstrap Resampling showed no statistically significant improvements. SHAP provided additional interpretations of the models' output.

We can infer a relatively poor performance of the proposed approach by comparing it to other weakly-supervised deep learning methods for breast MRI. For example, *van der Velden et al.* [18] studied the feasibility of automatic volumetric breast density estimation on MRI without segmentation and utilized SHAP to explain the estimated values. Despite our work having a different objective, fewer training samples, and two-dimensional MIP images instead of three-dimensional MRI, we can still compare the models' predictive abilities and SHAP explanations. The Spearman's correlation between the values predicted by the proposed CNN and the ground truth breast densities was higher than that in all our experiments. Unlike the breast density estimating model, our models did not erroneously base their predictions on the heart tissue present in the images. However,

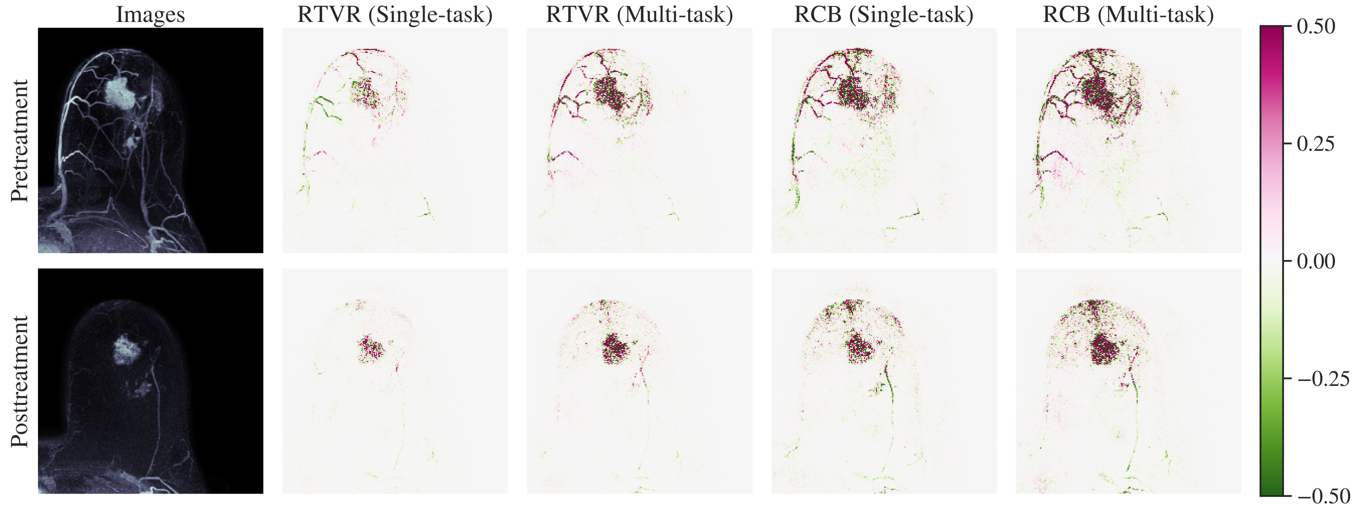


Fig. 4: Example of SHAP maps for RTVR and RCB predictions in uni-dimensional and multi-dimensional regressions. One can notice that the tumor region and adjacent vessels are responsible for both increase (pink) and decrease (green) in predictions of RTVR and RCB. The explanations in different experiments have almost no significant visual differences. Ground truth RTVR is $0.52$. RTVR predicted by single- and multi-task regression models are $0.20$ and $0.34$, respectively. Ground truth RCB is $2.98$. RCB predicted by single- and multi-task regression models are $1.46$ and $2.95$, respectively.

our models yielded SHAP maps with positive and negative values mixed inside the region of interest, which *van der Velden et al.* [18] did not observe. We can additionally compare the explanations to those presented in *Verburg et al.* [22] for the model designed to classify between breasts with and without cancer. Similarly to our work, MIP images served as input for that model. However, the training data used in that study was multi-institutional and not limited. Positive SHAP values attributed to that model were homogeneous and coincided with the location of the lesions. Despite the presence of contrast-enhanced blood vessels adjacent to the tumors, unlike in our work, only the lesions determined that model's output.

Having proven the overall poor performance of the models trained to predict RTVR and RCB value, we can compare them with each other. Interestingly, the RTVR predicting model, while having more meaningful homogeneous SHAP maps, required high regularization for optimal performance. Speculating on the reasons for this, we assume that the other two models failed to learn even the $k$ training-validation subsets, with early stopping being deployed too early. Therefore, overfitting could be the only way to train for the RCB predicting and multi-task models, requiring low regularization. Next, only the multi-task model required no dropout for optimal performance. We can speculate that the reason for this could be the same size of the network but twice more regression targets. Despite each task receiving a corresponding linear head, the convolutional body remains unchanged in this experiment.

In this paper, we dealt with two data-related issues. We addressed the problem of imbalanced regression by using the BMC loss and bin-based balanced metrics. As we conducted no ablation study, the effectiveness of this strategy remains unknown. Relatively low optimal initial noise scales may indicate a problem to explore. Another data-related issue was the limited number of available images. It influenced the chosen experiment design, arousing the need for cross-validation and preventing us from using a separate validation subset.

Rather than exploring the limitation of this study, future research can focus on developing a more robust approach for the same regression problem. First, one may consider the potential of designing a task-specific architecture, which remained untapped in our work. For example, a multi-branch approach would be more suitable for two unregistered input images. In contrast to the employed two-channeled one, it may prevent anatomical inconsistencies from negatively affecting the performance. Second, securing access to a larger dataset or collaborating with other institutions to merge and utilize their data can significantly increase the model's capacity to predict the RTVR and RCB values. If this is not feasible, future research may experiment with more augmentation strategies, including various spatial transforms, which remained unemployed in our study. Since MIP images contain truncated information, one can use original three-dimensional MRI if hardware allows it. In that case, generative AI can serve as an effective augmentation tool. For instance, one can consider training a VAEGAN [23] conditioned on the lesion masks using SPADE [24]. That would allow for explicitly determining the tumor volumes in the generated images and, therefore, be a suitable strategy for the RTVR regression. Finally, future research can incorporate techniques designed for multi-task learning. For example, combining the loss weighting strategy proposed by *Kendall et al.* [25] with the BMC loss would contribute to the method's robustness.

## V. Conclusion

We investigated the feasibility of predicting patients' responses to NAC directly from MIP images using a weakly-supervised deep regression model. The regression targets were RTVR and RCB values, representing radiological response on MRI and pathological response, respectively. We compared the predictive ability of single- and multi-task models utilizing Spearman's correlation coefficients. We observed no statistically significant difference between the models, although the multi-task one showed better results in estimating both regression targets.

Numerical results and SHAP explanations indicate that the current method cannot produce clinically meaningful outputs. Therefore, in its current form, the proposed method can not apply in clinical practice. Future work should consider refinement of the current approach or exploring alternative ways to achieve better performance in the regression task.

# References

[1] B. Asselain, W. Barlow, J. Bartlett, J. Bergh, E. Bergsten-Nordström, J. Bliss, F. Boccardo, C. Boddington, J. Bogaerts, G. Bonadonna, *et al.*, "Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials," *The Lancet Oncology*, vol. 19, no. 1, pp. 27–39, 2018.

[2] P. Cortazar, L. Zhang, M. Untch, K. Mehta, J. P. Costantino, N. Wolmark, H. Bonnefoi, D. Cameron, L. Gianni, P. Valagussa, *et al.*, "Pathological complete response and long-term clinical benefit in breast cancer: the ctneobc pooled analysis," *The Lancet*, vol. 384, no. 9938, pp. 164–172, 2014.

[3] S. P. Gampenrieder, A. Peer, C. Weismann, M. Meissnitzer, G. Rinnerthaler, J. Webhofer, T. Westphal, M. Riedmann, T. Meissnitzer, H. Egger, *et al.*, "Radiologic complete response (rcr) in contrast-enhanced magnetic resonance imaging (ce-mri) after neoadjuvant chemotherapy for early breast cancer predicts recurrence-free survival but not pathologic complete response (pcr)," *Breast Cancer Research*, vol. 21, pp. 1–11, 2019.

[4] C. E. Loo, L. S. Rigter, K. E. Pengel, J. Wesseling, S. Rodenhuis, M.-J. T. V. Peeters, K. Sikorska, and K. G. Gilhuijs, "Survival is associated with complete response on mri after neoadjuvant chemotherapy in er-positive her2-negative breast cancer," *Breast Cancer Research*, vol. 18, pp. 1–12, 2016.

[5] W. F. Symmans, F. Peintinger, C. Hatzis, R. Rajan, H. Kuerer, V. Valero, L. Assad, A. Poniecka, B. Hennessy, M. Green, *et al.*, "Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy," *Journal of Clinical Oncology*, vol. 25, no. 28, pp. 4414–4422, 2007.

[6] N. M. Hylton, J. D. Blume, W. K. Bernreuter, E. D. Pisano, M. A. Rosen, E. A. Morris, P. T. Weatherall, C. D. Lehman, G. M. Newstead, S. Polin, *et al.*, "Locally advanced breast cancer: Mr imaging for prediction of response to neoadjuvant chemotherapy—results from acrin 6657/i-spy trial," *Radiology*, vol. 263, no. 3, pp. 663–672, 2012.

[7] N. M. Hylton, C. A. Gatsonis, M. A. Rosen, C. D. Lehman, D. C. Newitt, S. C. Partridge, W. K. Bernreuter, E. D. Pisano, E. A. Morris, P. T. Weatherall, *et al.*, "Neoadjuvant chemotherapy for breast cancer: functional tumor volume by mr imaging predicts recurrence-free survival—results from the acrin 6657/calgb 150007 i-spy 1 trial," *Radiology*, vol. 279, no. 1, pp. 44–55, 2016.

[8] M. H. Janse, L. M. Janssen, B. H. van Der Velden, M. R. Moman, E. J. Wolters-van der Ben, M. C. Kock, M. A. Viergever, P. J. van Diest, and K. G. Gilhuijs, "Deep learning-based segmentation of locally advanced breast cancer on mri in relation to residual cancer burden: A multi-institutional cohort study," *Journal of Magnetic Resonance Imaging*, 2023.

[9] R. Wightman, "Pytorch image models." https://github.com/huggingface/pytorch-image-models, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[11] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 558–567, 2018.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[13] R. Wightman, H. Touvron, and H. Jegou, "Resnet strikes back: An improved training procedure in timm," in *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*.

[14] J. Ren, M. Zhang, C. Yu, and Z. Liu, "Balanced mse for imbalanced visual regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7926–7935, 2022.

[15] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, pp. 1139–1147, PMLR, 2013.

[16] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *International Conference on Machine Learning*, pp. 1059–1071, PMLR, 2021.

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[18] B. H. van der Velden, M. H. Janse, M. A. Ragusi, C. E. Loo, and K. G. Gilhuijs, "Volumetric breast density estimation on mri using explainable deep learning regression," *Scientific Reports*, vol. 10, no. 1, p. 18095, 2020.

[19] J. M. Bland and D. G. Altman, "Measuring agreement in method comparison studies," *Statistical methods in medical research*, vol. 8, no. 2, pp. 135–160, 1999.

[20] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[22] E. Verburg, C. H. Van Gils, B. H. Van Der Velden, M. F. Bakker, R. M. Pijnappel, W. B. Veldhuis, and K. G. Gilhuijs, "Deep learning for automated triaging of 4581 breast mri examinations from the dense trial," *Radiology*, vol. 302, no. 1, pp. 29–36, 2022.

[23] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*, pp. 1558–1566, PMLR, 2016.

[24] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2337–2346, 2019.

[25] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.