Regularization Techniques in Radiomics: A Case Study on the Prediction of pCR in Breast Tumours and the Axilla

Eunice Carrasquinha[1], João Santinha[1], Alexander Mongolin[2], Maria Lisitskiya[1], Joana Ribeiro[3], Fátima Cardoso[3], Celso Matos[4], Leonardo Vanneschi[2], and Nickolas Papanikolaou[1]

[1] Center for the Unknown, Champalimaud Foundation, Computational Clinical Imaging Group, Av. Brasilia, 1400-038 Lisbon, Portugal nickolas.papanikolaou@research.fchampalimaud.org

[2] Nova Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal lvanneschi@novaims.unl.pt

[3] Champalimaud Clinical Center, Breast Unit, Av. Brasilia, 1400-038 Lisbon, Portugal joana.ribeiro@fundacao.champalimaud.org

[4] Radiology Department, Centre for the Unknown, Champalimaud Foundation, Av. Brasília, 1400-038 Lisbon, Portugal

This is the Author Peer Reviewed version of the following chapter/ conference contribution published by Springer:

Carrasquinha, E. *et al.* (2020). Regularization Techniques in Radiomics: A Case Study on the Prediction of pCR in Breast Tumours and the Axilla. In: Cazzaniga, P., Besozzi, D., Merelli, I., Manzoni, L. (eds) Computational Intelligence Methods for Bioinformatics and Biostatistics. CIBB 2019. Lecture Notes in Computer Science, vol 12313. Springer, Cham. https://doi.org/10.1007/978-3-030-63061-4_24

Regularization techniques in Radiomics: A case study on the prediction of pCR in Breast Tumours and the Axilla

Eunice Carrasquinha[*,1], João Santinha[1], Alexander Mongolin[2], Maria Lisitskiya[1], Joana Ribeiro[3], Fátima Cardoso[3], Leonardo Vanneschi[2], Nickolas Papanikolaou[1]

[1] Center for the Unknown, Champalimaud Foundation
Computational Clinical Imaging Group, Av. Brasilia, 1400-038 Lisboa,
eunice.carrasquinha@research.fchampalimaud.org * corresponding author

[2] Nova Information Management School (NOVA IMS)
Universidade Nova de Lisboa, Campus de Campolide, 1070-312, Lisboa, Portugal
{d20180761,lvanneschi}@novaims.unl.pt

[3] Champalimaud Clinical Center
Breast Unit, Av. Brasilia, 1400-038 Lisboa
joana.ribeiro@fundacao.champalimaud.org

Keywords: Radiomic features, high-dimensional data, regularization techniques, breast cancer.

Abstract. Clinicians have shown an increasing interest in quantitative imaging for precision medicine. Imaging features can extract distinct phenotypic differences of tumours, potentially they can be used as a non-invasive prognostic tool and contribute for a better prediction of pathological complete response (pCR). However, the high-dimensional nature of the data brings many constraints, for which several approaches have been considered, with regularization techniques in the cutting-edge research front. In this work, classical lasso, ridge and the recently proposed priority-lasso are applied to high-dimensional imaging data, regarding a binary outcome. A breast cancer dataset, with radiomics, clinical and pathological information as features, was used. The application of sparsity techniques to the dataset enabled the selection of relevant features extracted in MRI of breast cancer patients, in order to identify the accuracy of those features and predict the pCR in the breast and the axilla.

1 Scientific Background

High throughput experiments provide large amounts of data, holding the potential to improve our knowledge about biological processes, and eventually provide more insights regarding the treatment of many diseases, particularly cancer. In the last few years, the use of *omics* data has contributed to the identification of biomarkers, and improving diagnosis, prognostication and treatment of cancer patients. However, most of the studies are making use of molecular biomarkers including genomics and proteomics, which require biopsies or invasive surgeries to extract and analyze tumour samples.

More recently, [3] proposed the use of radiomics, which is based on the extraction of a large number of imaging features, that could contain complementary information and provide a more comprehensive view of the entire tumour. In this sense, radiomics is an emerging field that converts imaging data into a high-dimensional mineable feature space. Due to the high-dimensionality of this type of data, classical statistical approaches are no longer suitable. Several solutions to cope with this dimensionality problem can be found in the literature. The most common choice is the use of machine learning algorithms to extract knowledge from data, possibly without any assumption on the shape and characteristics of the model [1].

Several machine learning approaches have been considered so far. However, due to the nature of the data, regularized optimization techniques have revealed their importance ([6] and [8]). Regularization techniques usually work by adding constraints to the cost function, with the objective of improving the generalization ability of the model. Ridge, lasso, and elastic net penalties are the most common examples of regularization techniques. The difference between them lies in the particular L_p norm used. For $p = 1$, the regularization technique is called lasso (Least Absolute Shrinkage and Selection Operator) regression, and for $p = 2$ it is called ridge regression. The result of combining the L_1 (lasso) and L_2 (ridge) norms is called elastic net [8]. In this work, the lasso and ridge regularization techniques are used.

In the context of cancer diseases, heterogeneous types of information (genomic, radiomic and clinical) are very common, and the inclusion of all information may improve the prediction accuracy of the model. Nevertheless, each type of data has generally a different structure. To overcome this problem, [2] proposed a hierarchical approach to predict a clinical outcome using different types of omics data, called priority-lasso. This method is similar to the lasso regularization technique, but takes into account groups of variables (blocks). The idea is to prioritize blocks that explain the largest possible part of the variability in the outcome.

The main goal of this work is to evaluate the predictive performance of logistic regression on a breast cancer dataset, using different regularization techniques, namely, the ridge, lasso and model extensions based on priorities.

2 Materials and Methods

2.1 Regularization Methods

In this section, the regularization techniques used for a logistic regression model are briefly presented. Logistic regression is a popular classification method that describes the relationship between one or more independent variables and a binary outcome variable Y, given by the logistic function:

$$p_i = P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta})}, \tag{1}$$

where x_i corresponds to the feature value for observation i (p is the number of features and n is the number of observations), p_i is the probability of success for observation i and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots \beta_p)$ are the unknown regression coefficients. The $\boldsymbol{\beta}$ parameters are estimated by maximizing the log likelihood function of the logistic model given by:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big\{ y_i \log p_i + (1 - y_i) \log[1 - p_i] \big\} + F(\boldsymbol{\beta}), \tag{2}$$

with $F(\boldsymbol{\beta})$ denoting the regularization term, which for the elastic net penalty takes the form:

$$F(\boldsymbol{\beta}) = \lambda \big\{ \alpha \|\boldsymbol{\beta}\|_1 + (1 - \alpha) \|\boldsymbol{\beta}\|_2^2 \big\}, \tag{3}$$

with $\lambda > 0$, which controls the penalization of the weights and $0 \leq \alpha \leq 1$ gives the balance between L_1 and L_2 norms, with the L_1 part being responsible for achieving sparsity. For $\alpha = 0$, leads to the ridge regression, for $\alpha = 1$ it corresponds to the lasso regression.

As previously mentioned, multiple types of data can improve the prediction accuracy. In order to incorporate all different types of data, different groups of features (*blocks*) should be considered. The extension of lasso, the priority lasso [2] builds a predictive model, based on different blocks of features.

Let $\boldsymbol{\pi} = (\pi_1, ..., \pi_M)$ be the permutation of $(1, ..., M)$ indicating the priority order, where M is the number of blocks. At a first step, the features from π_1 are used to fit the lasso regression model with the highest priority. Following the notation of Equation (2), the regression coefficients $\boldsymbol{\beta}$ can be estimated by minimizing:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \log p_i + (1 - y_i) \log[1 - p_i] \right\} + \lambda^{\pi_1} \|\boldsymbol{\beta}^{\pi_1}\|_1. \quad (4)$$

Secondly, lasso is applied to the block with second highest priority. In this step, a linear score obtained from the previous step is used to force the model with coefficient fixed to 1 (*offset*). Finally, in the third step, lasso is applied to the block with third highest priority, using the linear score of the previous step as an *offset*. This process is repeated until all blocks have been considered. The priority-lasso is a hierarchical approach, since blocks with higher priority tend to explain the largest part of the variability in the outcome. For more details regarding this methodology, the reader is referred to [2].

2.2 Data

131 patients diagnosed with early breast cancer underwent breast Magnetic Resonance Imaging (MRI) examination at the Breast Unit of the Champalimaud Clinical Center before receiving neoadjuvant chemotherapy (NAC). After NAC, at the time of breast surgery, 58 patients presented with pCR and 73 without pCR, while for the axilla 31 patients had pCR and 36 had no pCR. Apart from the radiomic features, the dataset was composed of demographic, clinical and histopathological data. The clinical variables were nodal clinical status, tumour stage and menopausal status. For the histopathological features, four tumour variables were considered (tumour morphology, histological grade, subtypes and biological subtypes); age was the only demographic feature considered and for the radiomics features 4300 variables were used (2150 extracted from the tumour and 2150 extracted from a 5mm peritumoural zone). The Pyradiomics, was used for the extraction of the radiomics features from medical images, [7].

Arterial phase from the dynamic contrast enhanced MRI sequence was used to delineate the tumour by two radiologists (M.L. with 11 years of experience and A.U., a last year radiology resident). A second region of interest (ROI) comprising peritumoural tissue was created automatically from the tumour segmentation comprising a zone with 5mm of thickness. Both tumoural and peritumoural ROIs were used to extract radiomic features (first order, shape, GLCM, GLRLM, GLSZM,NGTDM, and GLDM of both original and filtered images - Laplacian of Gaussian, LoG, with sigma 1mm, 2mm, 3mm; wavelets - level 1 and 2) using Pyradiomics.

As classical lasso and ridge do not work with feature groups, recursive feature elimination (RFE) was used to find the most important features. The idea of RFE is based on the recursive reduction of the features. Firstly, the estimator (lasso/ridge) is trained on the full set of features and the importance of each one of them is estimated by means of coefficients. Then, the least important 50% of features are removed from the current set. The procedure is repeated on the pruned set until a target number of features is reached. The lasso and ridge regularization technique was applied only for the radiomic features, and the priority-lasso was used for all the features involved in the study. Before performing the multivariate analysis, a stability analysis for the radiomic features was conducted, using the intraclass correlation coefficient (ICC) with a threshold of 0.81, selecting 3023 out of 4300 radiomic features. Based on the correlation matrix, the absolute values of pair-wise correlations were obtained. Only features with low correlation were considered, using a threshold of 0.95. All radiomic features were normalized based on the z-score, and zero-variance and near-zero-variance features were removed from the analysis. The choice of the parameters used for regularization performed was

as follows. A cross-validation procedure was used to optimize λ, considering a training set composed of 75% of samples and 5 folds.

The open-source software R [5] was used to perform the statistical analysis. The main libraries used for the analysis were: `glmnet`, for lasso regularization for the radiomics features, and `prioritylasso`, to conduct the priority lasso for the different types of features. The open-source library `scikit-learn` [4] and `python-glmnet` were used to perform recursive feature elimination and score collection for the lasso and ridge methods.

3 Results

3.1 Breast pCR

Table 1 reports the results obtained using logistic regression for tumour pCR, without grouping features. The first line shows the results achieved with recursive feature elimination (RFE) that selected 10 most important features.To obtain these results, the following methodology was employed: 100 independent executions of each one of the algorithms were performed. For each one of these executions, 75% of the observations were randomly selected (with uniform probability) to form the training set, and the remaining 25% were used as a test set (hold out). Feature selection was executed inside of each fold, so testing/validation samples are not exposed to RFE algorithm.

In the Table 1, "CV" stands for the mean score obtained on the validation set , while "Testing" is the score obtained on the test set, both averaged over the 100 executions. From Table 1 , at least two facts can be seen: first lasso is the method that provides the

Table 1: AUC scores on Breast tumour for lasso, ridge and no regularization of logistic regression

	lasso	ridge	No regularization
With RFE	CV: 0.83(+/-0.03) Testing: 0.85(+/-0.05)	CV: 0.71(+/-0.07) Testing: 0.77(+/-0.11)	CV: 0.70(+/-0.06) Testing: 0.71(+/-0.06)
No RFE	CV: 0.85(+/-0.02) Testing: 0.82(+/-0.05)	CV: 0.58(+/-0.04) Testing: 0.58(+/-0.08)	CV: 0.55(+/-0.04) Testing: 0.57(+/-0.09)

highest performance, regardless the use or not of RFE. Secondly, in practically all the presented experiments, the results obtained on the validation set are very similar to the ones obtained on the (hold-out) test set, which indicates that the obtained models are not overfitted.

When considering different types of features (age, clinical/pathological, radiomic), using priority-lasso, the results show that the AUC values for train and test depend on the block that is prioritized. For this analysis, two blocks were considered: one block for the age, clinical/pathological features and a second block for the radiomic features. If demographic, clinical/pathological features were prioritized, the AUC values for train and test were 89.17% and 78.57%, respectively. When priorities were switch and radiomic features block had the highest priority, the AUC values were 87.79% and 74.6%, for the training and test cohorts, respectively. Figure 1 shows the ROC curves when different priority blocks were considered.

Regarding the features selected by this approach, when the highest priority block considered was age and clinical/pathological features, four features of these block (age, tumour stage, subtypes and biological subtypes), and nine radiomic features were selected. These nine radiomic features selected: two are tumoural (LoG sigma 2mm 3D glcm Imc2 and wavelet2 LHL glszm ZoneVariance) and seven peritumoural (exponential gldm Large Dependence High Gray Level Emphasis, square gldm Small Dependence Low Gray Level Emphasis, square ngtdm Contrast, wavelet LHL glcm Imc1, wavelet LHL ngtdm Strength, wavelet2 LHH gldm Dependence Variance and wavelet2 LLH ngtdm Contrast). When the highest priority block was regarding radiomic features,

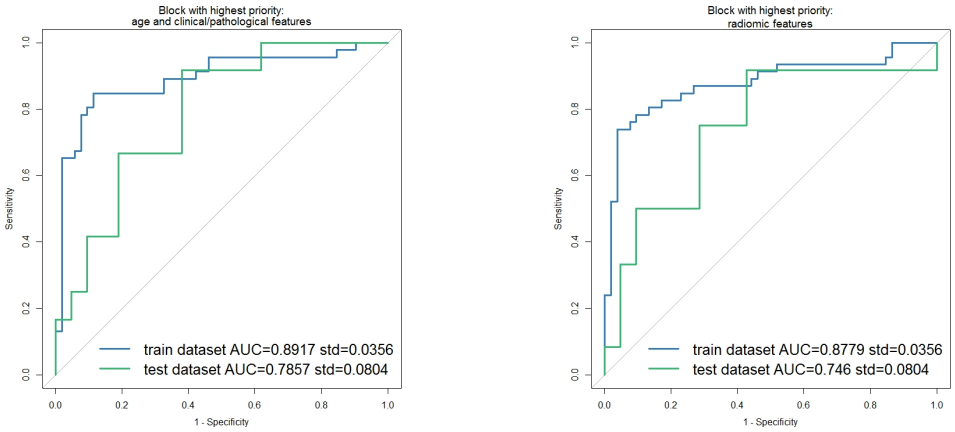

Figure 1: Roc curves for the train and test set, after priority lasso, for the Breast tumour pCR.

two features from the age, clinical/pathological block, were selected (subtypes and biological subtypes) and 11 from the radiomic features, three tumoural (LoG sigma 2 mm 3D glcm Idn, LoG sigma 2 mm 3D gldm Small Dependence Low Gray Level Emphasis and wavelet2 LHL glszm Zone Variance) and eight peritumoural (exponential gldm Large Dependence High Gray Level Emphasis, original glcm Idmn, square ngtdm Contrast, wavelet HLH glcm Imc2, wavelet LHL glcm Imc1, wavelet LHL ngtdm Strength, wavelet2 LHH gldm Dependence Variance, wavelet2 LHH glszm Zone Variance).

3.2 **Axilla pCR**

Table 2 reports the results obtained for axilla pCR. This table confirms that penalization based on L_1 is preferable for the studied data and modifications based on lasso are appropriate.

Table 2: AUC scores on axilla dataset for lasso, ridge and no regularization of logistic regression

	lasso	ridge	No regularization
With RFE	CV: 0.61(+/-0.08) Testing: 0.60(+/-0.11)	CV: 0.57(+/-0.08) Testing: 0.56(+/-0.11)	CV: 0.56(+/-0.11) Testing: 0.56(+/-0.08)
No RFE	CV: 0.59(+/-0.08) Testing: 0.58(+/-0.09)	CV: 0.55(+/-0.04) Testing: 0.54(+/-0.07)	CV: 0.52(+/-0.07) Testing: 0.53(+/-0.07)

For the priority-lasso, the blocks considered were the same as the ones mentioned in the previous section. The results show that if the block with the highest priority was regarding the demographic and clinical/pathological features, the AUC values for training and test cohorts were 87.01% and 87.5%, respectively. When priorities were switched and radiomic features block had the highest priority, AUC values were 93.83% and 86.11%, for training and test cohorts, respectively. Figure 2 shows the ROC curves when different priority blocks were considered.

Regarding the features selected, if the demographic and clinical/pathological block was considered with the highest priority, age, tumour stage, tumour morphology, histologic grade, subtypes and biological subtypes were selected. When the radiomic block was considered first, five features (tumour stage, tumour morphology, histologic grade, subtypes and biological subtypes) were selected from the demographic and clinical/pathological block and four features were selected from radiomic block: three tumoural (exponential first order Median, LoG sigma 2mm 3D first order Skewness and wavelet LHL gldm Low Gray Level Emphasis) and one peritumoural (original glcm Idn).

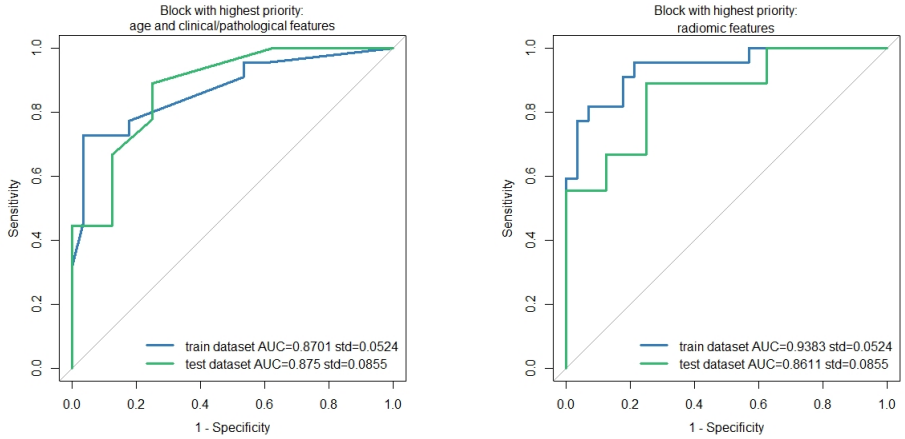

Figure 2: Roc curves for the train and test set, after priority lasso, for the axilla pCR.

4 Conclusions

In this work, we evaluated the predictive power of logistic regression using the ridge, lasso and priority-lasso regularization techniques. The presented results confirmed that a better prediction accuracy can be obtained using the lasso and priority-lasso regularization techniques regarding the achievement of pCR after NAC for breast cancer. Lasso regularization techniques showed a better performance, irrespectively of whether RFE was used or not. For the axilla pCR, the results showed that the best performance was obtained using priority-lasso. In this sense, the presented results showed that data of different nature should be considered into different groups, but used by one method that supports prioritization. This work may pave the way for further research aimed at defining other regularization methods to support different blocks of features. Also, considering the high scores achieved in this work, similar data collection and classification methods could be used by clinicians in the future, to support medical decisions.

Acknowledgments

This work was partially supported by national Portuguese funds through FCT (Fundação para a Ciência e a Tecnologia) under project PTDC/CCI-INF/29168/2017 (BINDER).

References

[1] A.-L. Boulesteix, M. N. Wright, S. Hoffmann and I.R. König, "Statistical learning approaches in the genetic epidemiology of complex diseases". *Human Genetics*, pp. 1-12, 2019.

[2] S. Klau, V. Jurinovic, R. Hornung, T. Herold and A.-L. Boulesteix. "Priority-Lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data". *BMC Bioinformatics*, vol.19, no.1, pp. 1-14, 2018.

[3] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G. Stiphout, P. Granton, C.M. Zegers, R. Gillies, R. Boellard, A. Dekker and H.J. Aerts. "Radiomics: Extracting more information from medical images using advanced feature analysis". *European Journal of Cancer*, vol.48, no.4, pp. 441-446, 2012.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research*, vol.12, pp. 2825-2830, 2011.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[6] R. Tibshirani. "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society. Series B (Methodological)*, vol.58, no.1, pp. 267-288, 1996.

[7] J.J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, S. and H.J. Aerts . "Computational Radiomics System to Decode the Radiographic Phenotype". *Cancer Research*, vol.77, no.21, pp. e104-e107, 2017.

[8] H. Zou and T. Hastie. "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol.67, no.2, pp. 301-320, 2005.