UPPSALA
UNIVERSITET

Radiomics Feature Extraction and Machine Learning on MRI for predicting Treatment Response for Patients with Colorectal Cancer

Emil Åberg, Maja Arvola

Project in Computational Science: Report

Oct 2020 - Jan 2021

PROJECT REPORT

Institutionen för informationsteknologi

Abstract

Background: Today there exists no established way to predict the outcome of radiation therapy on patients with colorectal cancer. However, recent studies report that Radiomic features extracted from magnetic resonance imaging (MRI) can be used to successfully predict the treatment outcome in some situations.
Objective: This study aims to replicate the results from previous studies, that Radiomic features can predict the outcome of radiation therapy for patients with colorectal cancer, using a new data set.
Material: T2 weighted MRI and tumor segmentations created by an expert radiologist are available for 39 patients with colorectal cancer from three di erent Swedish hospitals. For every patient in the data set, the grade of tumor reduction after radiation therapy is known as one out of four categories, ranging from "No response" to "Complete response". The data is randomly divided into test data (8 patients) and training data (31 patients) with equal proportions of the outcome categories.
Method: 1578 Radiomic features are extracted for each patient using PyRadiomics. The most important features are selected using a feature selection method, either mRMR (Minimal Redundancy Maximal Relevance), LASSO (Least Absolute Shrinkage and Selection Operator) or Logistic Regression with L1-penalty. Three di erent prediction models are implemented: Random Forest regression, Random Forest classi cation, and Logistic Regression. These models are trained on the training data using 15 features selected by one of the feature selection methods.
Results: The methods are evaluated by cross-validation on the training data, where Random Forest regression together with LASSO gives the best results: R^2 score: 0.270, Accuracy: 61.3%. The same method do not perform as good on the test data: R^2 score: -0.813, Accuracy: 50.0%. In fact, the model predicts "Major response" for all patients in the test data.
Conclusions: The methods in this study are unable to predict the treatment outcome for patients not seen before. Possible explanations are that the data set is too small to successfully train a good model and class imbalance.

Keywords: Radiomic features, MRI (magnetic resonance imaging), Colorectal cancer, Machine learning

1 Introduction

Colorectal cancer is the fourth most common type of cancer in Sweden with over 7000 people diagnosed 2018 [1]. Colorectal cancer is a short notation for cancer in the colon or in the rectum. Patients with colorectal cancer are typically treated with radiation therapy prior to surgery, where malignant cancer cells are ionized with radiation. The effect of preoperative treatment varies between patients. If the outcome of radiation therapy could be predicted for individuals, it would be helpful for medical planning and it could potentially result in canceling unnecessary surgeries. However, no such method is established today.

Radiomics is an analysis method for extracting quantitative features from medical imaging. Several recent studies report successful results when predicting treatment outcome using Radiomic features [2][3][4]. The reasoning behind using Radiomics is that some of the extracted features are believed to reflect underlying physical characteristics of the tumor that corresponds to how well a patient is going to respond to treatment [5].

The purpose of this study is to investigate whether a Radiomic-based method for predicting the outcome of radiation therapy on patients with colorectal cancer can be successful. Similar to some previous studies, Radiomic features are extracted from only T2 weighted MRI [2][4][6]. In contrast to the previous studies, that considers a binary classification problem, this study investigates both classification models in the multiclass case and regression models with a numerical output.

2 Material

2.1 Study population

The study population consists of patients from three different Swedish hospitals. All patients are treated with neoadjuvant therapy for colorectal cancer between 2010 and 2017, either radiation therapy or chemoradiation therapy based on decision by oncologist, surgeon, pathologist and radiologist. After therapy, the patients undergo surgery, at which the outcome of the therapy is estimated. The outcome represents the amount of tumor regress, categorized into four categories:

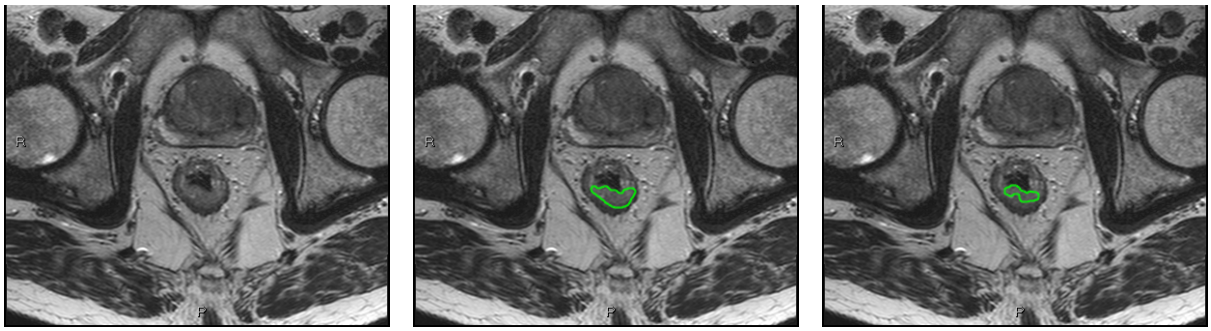
- No response (1 patient)
- Small response (10 patients)
- Major response (19 patients)
- Complete response (9 patients)

In total, 39 patients are considered in this study. This is a very small number compared with similar studies, in which 100 to 400 patients usually are considered [3][4][7].

2.2 Image data

For every patient, pre-treatment magnetic resonance images (MRI) are given. These are three dimensional, T2 weighted images, received using scanner fields strength of either 1.5 or 3.0 Tesla. This results in a variation in intensity across different images. The images also vary in voxel density. An example slice from a T2 weighted MRI is shown in Figure 1a. DICOM metadata is also available for each image, containing information such as pixel sizes, spacing between image layers, patient weight etc. [8].

Together with the image data, segmentation of the tumors are given in separate images. Figure 1b shows a segmentation conducted by looking a the T2 weighted image only, the segmentation in Figure 1c also considers diffusion weighted images, resulting in a more narrow region of interest. The segmentation is performed manually by a clinical radiologist with six years of experience.



(a) T2 weighted MRI slice.

(b) Tumor segmentation using T2 weighted image.

(c) Tumor segmentation using T2 weighted and di usion weighted images.

Figure 1: Example images from the data set.

3 Method

The workflow is divided into smaller steps shown in Figure 2. The purpose of each step is stated here, but more detailed explanations are found in the corresponding subsections. The image processing step aims to convert the input data to a format that is valid for the feature extraction step, where Radiomic features are extracted for each patient in the data set. A feature selection algorithm is then applied to remove redundant and unnecessary features. A prediction model is then trained on training data and evaluated using cross-validation on the training data. Multiple methods for feature selection and prediction are implemented, and the best combination is finally tested on the test data.

The task of predicting treatment outcome can be defined as a classification problem, which has been successful in previous studies [3][4][7]. The outcome categories, as described in Section 2.1, have a natural ordering which however motivates for treating the problem as a regression problem instead. In this study, classification models and regression models are compared. For regression models, the outcome "no, small, major and complete response" is mapped to integer values "0, 1, 2 and 3". The outputs of our regression models are floating point numbers, and to compute metrics such as accuracy and precision, the outputs are simply rounded to the closest integer.

The implementation of all steps in the workflow is done in Python 3.6. The code is publicly available on GitHub [9].

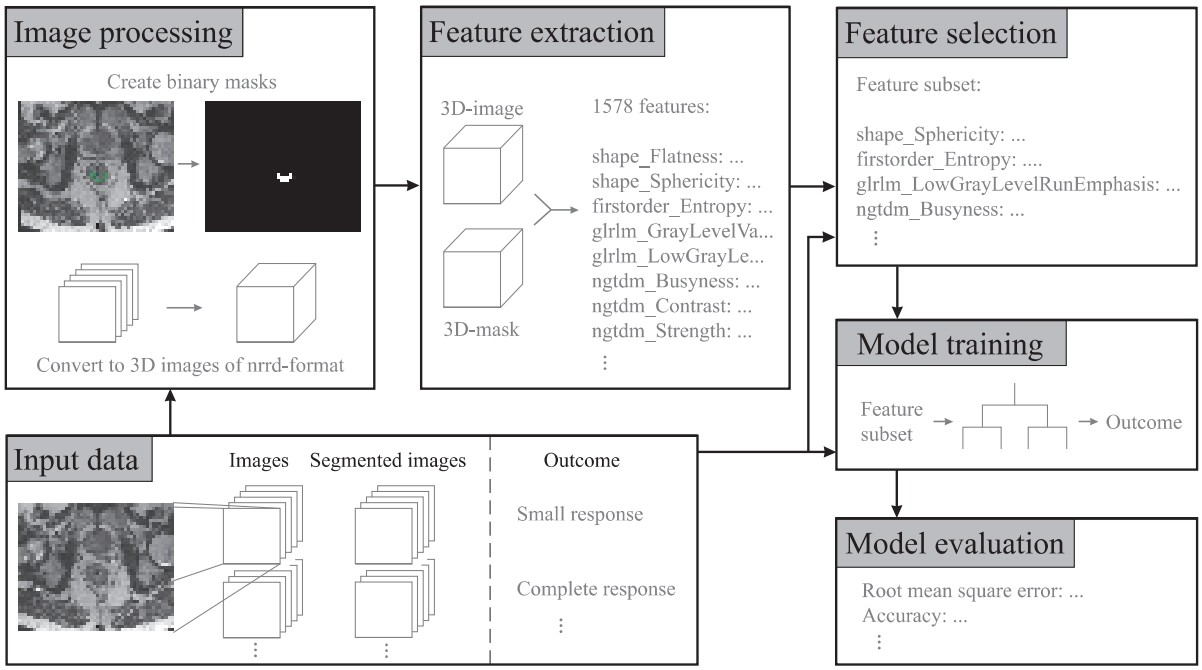


Figure 2: Schematic overview of the workflow

3.1 Image processing

The 2D image slices are converted to a 3D image of nrrd format [10] for each patient. The pixel sizes in x-, y- and z-direction varies between patients and needs to be taken into consideration when creating the 3D images. To achieve this, the pixel sizes are extracted from the DICOM metadata and thereafter included in the header of the nrrd file.

The segmented images provided by the radiologist are used to create binary masks, i.e. images where the region of interest is represented with white pixels and everything else is in black. The masks are then checked visually to ensure a correct representation of the original segmentation. When the algorithm for generating the masks fails, manual adjustment is applied. An example of a typical case where the

masking algorithm fails can be seen in Figure 3. The masks are created for the 2D slices separately, and then converted to 3D nrrd images as described in the paragraph above.

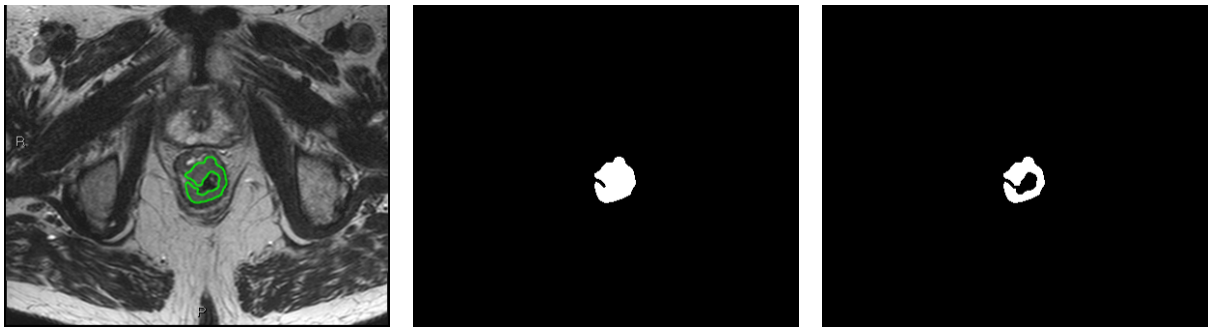


Figure 3: An example of a segmentation where the masking algorithm fails and manual adjustment is needed. To the left: images with tumor segmentation, in the middle: mask generated by masking algorithm and to the right: mask after manual adjustment.

3.2 Feature extraction

Features are extracted from the 3D images using PyRadiomics version 3.0.1, an open source Python package for extracting Radiomic features from medical images [11]. The features are extracted only from the tumor region, therefor a binary mask describing the region of interest is needed. Two different binary masks are available since the segmentation of the tumor is conducted in two different ways (see Section 2.2). The choice of mask affects the extraction and both alternatives are tested in separate runs.

Normalization and re-sampling are applied to the images before feature extraction in order to compensate for varying intensities and voxel-densities between images in the data set. A large number of Radiomic features are then extracted from every image. To extract features from one image, 16 different filters are first applied to the image (gradient filter, logarithmic filter, etc.), making a total of 17 image types, including the original image. From each image type, the features listed below are extracted, with exception for shape features that only are extracted from the original image.

- Shape (14 features)
- First Order (18 features)
- Gray Level Co-occurrence Matrix (GLCM) (23 features)
- Gray Level Run Length Matrix (GLRLM) (16 features)
- Gray Level Size Zone Matrix (GLSZM) (16 features)
- Gray Level Dependence Matrix (GLDM) (14 features)
- Neighbouring Gray Tone Difference Matrix (NGTDM) (5 features)

In total, 1578 Radiomic features are extracted from each image using the chosen mask. Age, gender, weight and type of treatment (radiation or chemoradiation) are then added to the list of features before the feature selection step. More details about the feature extraction settings, including normalization, re-sampling and image types, are found in the file `Params.yaml` [9].

3.3 Feature selection

In this section, three methods for feature selection are described. The purpose of performing feature selection is to reduce the dimensionality in the data to avoid overfitting. Redundant features and features that does not contain information related to the outcome are removed.

3.3.1 mRMR

The first feature selection model is mRMR, which stands for Minimum Redundancy Maximum Relevance. To be able to use this method, the data is first discretized and this is done by applying k-means to create four bins for each feature. This is a drawback of using mRMR, information is lost when discretizing the data. Then the python package PymRMR is used for the implementation of mRMR.

The algorithm is originally proposed by Peng et al. [12] and utilizes information theory to find a representative subset of features $S = \{x_i\}$. The aim is to select features that have maximal mutual information with the outcome target while also removing features that are highly correlated with each other and therefor can be considered redundant. The mutual information $I(x_1; x_2)$ between two random variables is defined as

$$I(x_1; x_2) = \int\int p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} dx_1 dx_2. \tag{1}$$

This expression is hard to estimate from samples of continuous variables x_1 and x_2, which is why discretization is needed. When the variables are categorical, the integrals reduces to summations, which are straightforward to implement. To select a relevant feature subset S, the relevance $D(S, y)$ is maximized,

$$\max_S D(S, y), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; y), \tag{2}$$

where y is the target variable and $|S|$ the number of selected features. At the same time, the redundancy $R(S)$ is minimized,

$$\min_S R(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j). \tag{3}$$

To combine these into one single optimization problem, consider maximizing either the difference $D - R$ or the quotient D/R. In this study, the difference is maximized.

The mRMR method is previously implemented in the context of identifying relevant features related to patient response to chemoradiation therapy [2][4][13].

3.3.2 LASSO

LASSO stands for Least Absolute Shrinkage and Selection Operator. It is a regression model that performs both feature selection and regularization. This regression method is not implemented in any previous studies found, since they treat the problem as a classification problem.

Before selecting important features with LASSO, the input feature space X is normalized by subtracting the mean and dividing with the L2-norm for all features. The objective of LASSO is then to fit a linear regression model on the normalized input X and output y by minimizing

$$\frac{1}{2N} \|y - Xw\|_2^2 + \alpha \|w\|_1, \tag{4}$$

where N is the number of samples, w are the linear coefficients, or weights, and α is a regularization parameter. By using the L1-norm of the weights as a penalty term, some weights are forced to zero, how many depends on the parameter α. This parameter is chosen to obtain around 20 features with non-zero weights. The features with largest weights (by magnitude) are then selected as important features, only features with non-zero weights are selected.

3.3.3 Logistic regression with L1-penalty

Logistic regression is a parametric model and that is widely used for classification. It is previously applied together with LASSO regularization for feature selection [3][4][7].

In the binomial case, the logistic function can be used to estimate the probability of input X belonging to class $y = 1$ according to

$$p(y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}. \tag{5}$$

The parameter β_0 is a constant value and β_1 is a vector containing weights for each input parameter. These parameters, β_0 and β_1, are commonly chosen to minimize the negative log likelihood

$$L = \log\Big(\prod_{i:y=1} p(y = 1|X_i) \prod_{j:y\neq 1} \big(1 \quad p(y = 1|X_j)\big)\Big), \tag{6}$$

meaning that the probability that samples with outcome $y = 1$ belongs to $y = 1$ is maximized while the probability that samples with $y \neq 1$ in outcome belongs to $y = 1$ is minimized.

In this implementation logistic regression is considered together with a penalty term, LASSO regularization. This means that the equation to minimize is

$$L + \alpha\|\beta\|_1 . \tag{7}$$

The parameter α is the same regularization parameter as described in Section 3.3.2 and it is once again chosen to obtain around 20 features with non-zero weights. The features with largest weights (by magnitude) are then selected as important features, only features with non-zero weights are selected.

This study treats the multi-class problem and our way to approach this is to use the One-vs-Rest method. This means setting up a binary problem for each class, where one class is the one of interest and the others compose the rest of the data.

3.4 Prediction models

Here we give some more details on the prediction models. These models are trained to predict the treatment outcome using the features that are selected in the previous step.

3.4.1 Random Forest

Random Forest is an ensemble method, meaning that it combines weaker models, in this case decision trees, in order to create a stronger model. It can be used for both regression and classification problems. Random Forest is previously used in Radiomic studies to predict pathological complete response in rectal cancer with varying results [4][14].

The idea behind decision trees is to divide the feature space into smaller and smaller regions using recursive binary splitting until a stopping criterion is reached [15]. Each split in the tree corresponds to a condition on one feature. To measure the quality of a split the Gini index is used for the classification problem and Mean Squared Error (MSE) for the regression problem. The algorithm is greedy in the way that it chooses the best split at a given step, without considering how it impacts the final tree.

Random Forest creates multiple decision trees using bootstrapped data. For each split, only features from a randomly selected feature subset are considered as split candidates. The purpose of this is to obtain less correlated trees than if all features are available for consideration.

The hyperparameters of the model are selected through cross-validation. The range of tested values for each parameter is specified in the list below:

- Number of decision trees (in range 5 to 75)
- Maximal tree-depth (in range 1 to 15 or no limit)
- Size of feature subset considered at each split, relative to the total number of features ($^1/_3$, $^2/_3$, 1.0 or $\sqrt{\text{number of features}}$)

3.4.2 Logistic Regression with L1-penalty

The classifier Logistic Regression (described in Section 3.3.3) is also implemented as a prediction model. The algorithm is previously implemented for classification of treatment outcome from neoadjuvant therapy with promising results [7][6]. The parameter C, corresponding to $1/\alpha$ in Equation (7), is chosen from the range 10^{-4} to 10^4 through cross-validation, as proposed in the scikit-learn documentation [16].

3.5 Train and evaluate prediction model

The data is randomly divided into training data (80%) and test data (20%), where the proportion of each outcome category is the same in both sets. This split is done at an early stage to ensure that the test data do not influence any decisions in the study. To evaluate the model performance on an independent test set gives a better generalizability than evaluating using cross-validation [17].

To optimize the parameter settings for a model, for each parameter a range of possible values is specified. The range for specific parameters are specified in the section about the corresponding prediction model (see Section 3.4.1 and 3.4.2). For all possible combinations of parameter settings a performance metric is calculated on the training data with cross-validation. The R^2 score is used for regression models and accuracy for classification models. The parameter settings with the highest score are selected as the optimal parameter settings for that particular model. All combinations of feature selection models and prediction models are evaluated with cross-validation on the training data with the optimal parameter settings for each model. The best combination of models are finally tested on the test data.

To determine whether a prediction model is useful, it is often constructive to compare it with a baseline model (i.e. a very simple model that still is reasonable). For regression problems, the simplest of baseline models might be to always predict the average outcome of the training data. When the outcome of the test data follows the same distribution as in the training data, the score $R^2 = 0$ is yielded. For classification problems, a simple baseline model is to always predict the majority class from the training data. In our data set, this corresponds to always predict "Major response". When a final regression model does not achieve a score $R^2 > 0$ or when a final classification model does not beat the accuracy of guessing majority class, the model is considered to be unsuccessful.

4 Results

The feature selection algorithms returns 15 features each that can be found in Appendix A. There are four features that are chosen by two or all three algorithms, indicating that they contain highly relevant information. These features are:

- log-sigma-5-mm-3D_firstorder_90Percentile
- log-sigma-5-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis
- logarithm_ngtdm_Contrast_T2_M
- lbp-3D-m1_firstorder_Skewness

The three prediction models are trained with the selected features as input and evaluated through cross-validation on the training data. The results from the evaluation are presented in Table 1. Since the regression model and the classification models are trained to optimize the R^2 score and accuracy respectively, these metrics are used to compare the performances between models. Thereby, the best result is obtained using the Random Forest regression model together with the selected features from LASSO as input, yielding an R^2 score of 0.270 and an accuracy of 61.3%.

The best performing set-up is evaluated on the test data with following result metrics: R^2 score: 0.813, Accuracy: 50.0%, Precision: 16.7% and Recall: 33.3%. The confusion matrices for both evaluation on test data and through cross-validation on the training data are presented in Figure 4. It is clear that the model performance does not beat the established baseline models with an R^2 score below zero and an accuracy corresponding to only guessing the majority class in the prediction.

Feature selection algorithm	Prediction model	R^2 score	Accuracy	Precision	Recall
LASSO	Random Forest Regression	**0.270**	**61.3%**	50.0%	39.9%
LASSO	Random Forest Classification	0.065	53.3%	55.8%	46.5%
LASSO	Logistic Regression	-1.873	26.7%	8.9%	33.3%
Logistic Regression	Random Forest Regression	-0.316	41.9%	17.0%	23.1%
Logistic Regression	Random Forest Classification	-0.537	43.3%	41.4%	38.5%
Logistic Regression	Logistic Regression	-0.470	56.7%	54.1%	51.2%
mRMR	Random Forest Regression	-0.353	38.7%	11.1%	20.0%
mRMR	Random Forest Classification	-0.403	50.0%	35.0%	43.5%
mRMR	Logistic Regression	-1.873	26.7%	8.9%	33.3%

Table 1: Result metrics calculated with cross-validation on training data for prediction models using features from different feature selection methods. Precision and Recall score are calculated with macro average.

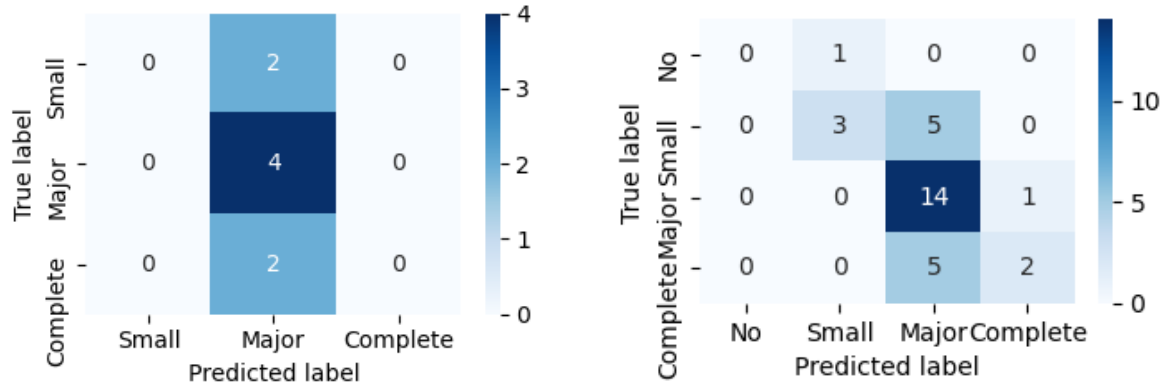


Figure 4: Confusion matrices for evaluation on test data (to the left) and through cross-validation on the training data (to the right) for Random Forest regression model. The shape difference is due to only having one sample with outcome 0, "no response", that appears in validation data but not in test data.

All results in this section are based on the segmentations that are created considering the T2 weighted images only. The same evaluations are performed on the segmentations that are created considering both T2 weighted and diffusion weighted images. This gives a slightly inferior performance when evaluating with cross-validation on the training data and is therefor excluded from the report.

5 Summary and Conclusion

The purpose of this study is to investigate if a method, based on Radiomics, can predict the outcome of neoadjuvant treatment for individuals with colorectal cancer. None of the methods implemented in this study are found to successfully predict the outcome for unseen patients. The best model slightly beats the established baseline models when cross-validating on the training data: R^2 score: 0.270, Accuracy: 61.3%. Using that model on the test data, the same outcome is predicted for all patients, a disappointing result. This yields the final result: R^2 score: 0.813, Accuracy: 50.0%.

Our approach is to extract Radiomic features from T2 weighted MRI, select the top features through feature selection and use these features to train a prediction model. Three separate algorithms for feature selection are implemented: mRMR (Minimal Redundancy Maximal Relevance), LASSO (Least Absolute Shrinkage and Selection Operator), and Logistic Regression with L1-penalty. Three different prediction models are implemented as well: Random Forest regression, Random Forest classification, and Logistic Regression. The performance of all nine possible combinations of feature selection and prediction models are evaluated.

Since this study tests several methods, many of which giving successful results in similar settings, it can be argued that the reason for the poor performance lies elsewhere. Two main issues regarding the data set can be identified. The first issue is that the data set is relatively small. Other similar studies typically use a data set containing 100 to 400 patients in comparison to our 39 patients. The second issue concerns the class imbalance in the data, which is shown in Section 2.1. In our data set with 4 classes, the majority class make up 49% of all the samples while the least common class make up 2.5%. Class imbalance is a well known issue and given the prediction results on our test data (see Figure 4) it is reasonable to believe that this issue affects our predictions.

Suggestions for continuations and improvements of this project are listed below.

- Increase the data set, a good target value is at least 100 patients.
- Extract features from other types of MRI as well, for example T1 weighted and diffusion weighted images.
- Consider methods for treating the problem of class imbalance, for example over sampling the minority classes or under sampling the majority class can be investigated.
- Consider a binary problem instead of a multi class problem, for example predicting which patients are likely to achieve a complete response to treatment. This makes the study more comparable to previous studies.

Acknowledgments

We would like to thank Nafsika Korsavidou Hult, for working day and night with the data set, handcrafting all tumor segmentations and giving us the material that was needed to perform the project.

We also give our thanks to Filip Malmberg for being our supervisor, providing us with guidance and support throughout the project.

References

[1] The swedish cancer register. https://sdb.socialstyrelsen.se/if_can/val.aspx. Accessed: 2020-11-05.

[2] Joost J. M. van Griethuysen, Doenja M. J. Lambregts, Stefano Trebeschi, Max J. Lahaye, Frans C. H. Bakers, Roy F. A. Vliegen, Geerard L. Beets, Hugo J. W. L. Aerts, and Regina G. H. Beets-Tan. Radiomics performs comparable to morphologic assessment by expert radiologists for prediction of response to neoadjuvant chemoradiotherapy on baseline staging mri in rectal cancer. *Abdominal Radiology*, 45(3):632–643, March 2020.

[3] Xuezhi Zhou, Yongju Yi, Zhenyu Liu, Wuteng Cao, Bingjia Lai, Kai Sun, Longfei Li, Zhiyang Zhou, Yanqiu Feng, and Jie Tian. Radiomics-based pretherapeutic prediction of non-response to neoadjuvant therapy in locally advanced rectal cancer. *Annals of Surgical Oncology*, 26, 03 2019.

[4] Jacob T. Antunes, Asya Ofshteyn, Kaustav Bera, Erik Y. Wang, Justin T. Brady, Joseph E. Willis, Kenneth A. Friedman, Eric L. Marderstein, Matthew F. Kalady, Sharon L. Stein, Andrei S. Purysko, Rajmohan Paspulati, Jayakrishna Gollamudi, Anant Madabhushi, and Satish E. Viswanath. Radiomic features of primary rectal cancers on baseline t2-weighted mri are associated with pathologic complete response to neoadjuvant chemoradiation: A multisite study. *Journal of Magnetic Resonance Imaging*, 52(5):1531–1541, 2020.

[5] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, November 2015.

[6] Hiram Shaish, Andrew Aukerman, Rami Vanguri, Antonino Spinelli, Paul Armenta, Sachin Jambawalikar, Jasnit Makkar, Stuart Bentley-Hibbert, Armando Portillo, Ravi Kiran, Lara Monti, Christiana Bonifacio, Margarita Kirienko, Kevin Gardner, Lawrence Schwartz, and Deborah Keller. Radiomics of mri for pretreatment prediction of pathologic complete response, tumor regression

grade, and neoadjuvant rectal score in patients with locally advanced rectal cancer undergoing neoadjuvant chemoradiation: an international multicenter study. *European Radiology*, 30, 06 2020.

[7] Shi Z Yang Z Du X Zhao Z Cui Y, Yang X and Cheng X. Radiomics analysis of multiparametric mri for prediction of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. *Eur Radiol*, 29:1211–1220, 08 2018.

[8] The dicom standard specification. https://www.dicomstandard.org/. Accessed: 2020-12-18.

[9] Åberg E. and Arvola M. Radiomics on mri for predicting colorectal cancer treatment response. https://github.com/majaarvola/Radiomics_on_MRI_for_predicting_colorectal_cancer_treatment_response, 2021.

[10] Definition of nrrd file format. http://teem.sourceforge.net/nrrd/format.html. Accessed: 2020-12-18.

[11] Pyradiomics documentation. https://pyradiomics.readthedocs.io/en/latest/index.html. Accessed: 2020-12-01.

[12] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, 27:1226–38, 09 2005.

[13] Talasila N Bera K Brady JT Gollamudi J Marderstein E Kalady MF Purysko A Willis JE Stein S Friedman K Paspulati R Delaney CP Romero E Madabhushi A Viswanath SE.c Alvarez-Jimenez C, Antunes JT. Radiomic texture and shape descriptors of the rectal environment on post-chemoradiation t2-weighted mri are associated with pathologic tumor stage regression in rectal cancers: A retrospective, multi-institution study. *Cancers (Basel)*, 12(8), 07 2020.

[14] Meyer H. J. Hambsch P. Wolf U. Kuhnt T. Hoffmann K. T. Surov A. Hamerla, G. Radiomics model based on non-contrast ct shows no predictive power for complete pathological response in locally advanced rectal cancer. *Cancers (Basel)*, 11(11), 10 2019.

[15] T. Hastie R. Tibshirani G. James, D. Witten. *An Introduction to Statistical Learning*. Springer, New York, NY, 7 edition, 2013.

[16] 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*, 12:2825–2830, 2011.

[17] Bino A. Varghese, Steven Y. Cen, Darryl H. Hwang, and Vinay A. Duddalwar. Texture analysis of imaging: What radiologists need to know. *American Journal of Roentgenology*, 212(3):520–528, March 2019.

Appendix A Selected Features

Logistic Regression
lbp-3D-m1_firstorder_Range
lbp-2D_glcm_JointEntropy
log-sigma-5-mm-3D_firstorder_90Percentile
square_glcm_Imc1
wavelet-LH_gldm_DependenceVariance
exponential_glcm_MCC
lbp-3D-m2_firstorder_Median
wavelet-HL_glszm_ZonePercentage
lbp-3D-m1_firstorder_Skewness
log-sigma-5-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis
log-sigma-1-mm-3D_glrlm_LongRunHighGrayLevelEmphasis
log-sigma-1-mm-3D_glszm_SmallAreaEmphasis
logarithm_ngtdm_Contrast
wavelet-HH_glcm_SumSquares
log-sigma-3-mm-3D_firstorder_Skewness

Table 2: Top 15 features selected by Logistic Regression algorithm

LASSO
square_glcm_Idmn
log-sigma-5-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis
wavelet-LH_glrlm_ShortRunLowGrayLevelEmphasis
lbp-2D_glrlm_ShortRunEmphasis
logarithm_ngtdm_Contrast
wavelet-LL_glcm_InverseVariance
logarithm_gldm_DependenceNonUniformityNormalized,
log-sigma-5-mm-3D_ngtdm_Contrast
wavelet-HH_firstorder_Skewness
log-sigma-3-mm-3D_glszm_GrayLevelVariance
wavelet-HL_firstorder_Skewness
log-sigma-5-mm-3D_firstorder_90Percentile
lbp-3D-m1_glrlm_ShortRunEmphasis
log-sigma-3-mm-3D_firstorder_Median
squareroot_glrlm_LongRunHighGrayLevelEmphasis

Table 3: Top 15 features selected by LASSO algorithm

MRMR
wavelet-LL_glcm_Idn
log-sigma-5-mm-3D_firstorder_90Percentile
square_firstorder_InterquartileRange
wavelet-HL_glrlm_LongRunLowGrayLevelEmphasis
wavelet-HH_glcm_ClusterShade
log-sigma-1-mm-3D_firstorder_Kurtosis
wavelet-LH_glszm_SmallAreaEmphasis
square_glcm_Correlation
lbp-3D-m1_firstorder_MeanAbsoluteDeviation
wavelet-HL_glcm_DifferenceVariance
wavelet-HH_glszm_ZonePercentage
lbp-3D-m1_firstorder_Skewness
logarithm_ngtdm_Strength
wavelet-LH_glcm_SumSquares
lbp-3D-k_glrlm_ShortRunLowGrayLevelEmphasis

Table 4: Top 15 features selected by MRMR algorithm