university of groningen

umcg

Radiomics and COVID-19

Predicting CO-RADS score on non-contrast chest CT using multiparametric radiomics, deep learning based noise reduction and machine learning classifiers

F.G. te Nijenhuis

First supervisor: dr. ir. P.M.A. van Ooijen
Second supervisor: Y. Nagaraj

Department of Radiology
Rijksuniversiteit Groningen
August 2020

Radiomics and COVID-19

Predicting CO-RADS score on non-contrast chest CT using multiparametric radiomics, deep learning based noise reduction and machine learning classifiers

F.G. te Nijenhuis

Abstract

The SARS-CoV-2 pandemic of 2020 brought with it new imaging challenges. To standardize the reporting of COVID-19 related findings on chest CT, the CO-RADS score is utilized. Automatic prediction of this score, using machine learning methods, would be directly beneficial to the patient by assisting the radiologist. To create an automatic CO-RADS prediction system, we obtained a dataset of non-contrast chest CT images ($n = 28$) annotated by a radiologist using the CO-RADS score. The lung fields were automatically segmented using a deep learning model. A fused model of two neural networks was used to perform noise reduction, doubling the amount of scans ($n = 56$). Histogram, shape and texture based radiomic features were extracted from normal and filtered images (1130 features per scan). Correlation based feature selection and principal component analysis were used as dimensionality reduction methods. Bayesian network, multilayer perceptron, random forest and logistic regression classifiers were trained on the original and noise reduced data. Naive Bayes was used as a baseline classifier, and classifier performance was compared using corrected paired t-tests. We found that a random forest classifier performs significantly better than the baseline classifier ($p < 0.05$) on all datasets. Best performance (90.10%($\pm$4.88) correct CO-RADS score prediction) is achieved by random forest on the deep learning based noise reduction dataset using the radiomic signature selected with correlation based feature selection, which consisted of 12 individual features. This indicates a potential role for our model, in assisting the radiologist when interpreting COVID-19 related findings on non-contrast chest CT.

Key Points

- Radiomics can be used to accurately predict CO-RADS score on chest-CT
- Deep learning based noise reduction can improve the accuracy of radiomics based methods
- Correlation based feature selection will select the most relevant radiomic signature when applied to noise reduced data, compared to other methods of dimensionality reduction
- The random forest classifier works best when predicting CO-RADS score using radiomics

Contents

1 Introduction

On 31 December 2019 the World Health Organization (WHO) was notified by the Chinese government about an outbreak of pneumonia in the city of Wuhan, situated within the Hubei province. Although the aetiology was initially unclear, a novel betacoronavirus, dubbed SARS-CoV-2, was rapidly identified as the causative agent. It is hypothesized that SARS-CoV-2 is zoonotic in origin, with the first suspected animal-to-human transmission occuring at a seafood market in Wuhan. The family of coronaviruses includes other human pathogens such as SARS-CoV, the causative agent of the Severe Acute Respiratory Syndrome (SARS) outbreak in 2002, and MERS-CoV, which caused an outbreak of Middle East Respiratory Syndrome (MERS) in 2012. Compared to SARS-CoV and MERS-CoV, SARS-CoV-2 exhibits a relatively high transmissibility and infectivity but a relatively low mortality rate. The most common mode of transmission is through droplets released by coughing and sneezing. Additionally, there is evidence of secretion of SARS-CoV-2 pathogens into bodily excretions such as saliva, urine and feces. The disease caused by SARS-CoV-2 was designated as coronavirus disease 2019 (COVID-19). Because of rapid global spread and a high contagion rate, the WHO declared COVID-19 a pandemic on the 11th of March 2020[1,2,3].

The symptomatology of COVID-19 is comparable to that of other coronaviruses. The most common symptoms at disease onset are fever and coughing. Less commonly, a sore throat, a headache, fatigue, diarrhea, and chest tightness are seen[4]. The majority of COVID-19 cases are mild, but in certain patients, a more severe form can lead to hospitalization and even admission to an Intensive Care Unit (ICU)[5]. Hospitalization occurs most frequently in the context of pneumonia, but other serious extrapulmonary manifestations of COVID-19 may occur[6]. The worst symptoms generally occur 10-14 days after initial onset. COVID-19 mainly affects the elderly, with a median age at death of 75 years[3]. As of Aug 14 2020, there were $20,900,000$ cases of COVID-19 globally, with $759,358$ confirmed deaths[7].

The main tool for the detection of COVID-19 is reverse transcription polymerase chain reaction (RT-PCR) performed on swab material obtained from oro- and nasopharynx or by broncheo-alveolar lavage[8]. This method amplifies viral RNA in the collected material to facilitate detection. In addition to the RT-PCR based methods, methods based on the detection of antibodies also exist but these antibodies appear later during the disease progression, and as such they are not suitable for initial screening. The detection of COVID-19 is complicated by the fact that no diagnostic gold-standard exists at this moment. The current RT-PCR method attains a sensitivity of 50-62%, leaving room for improvement[9].

Imaging studies can be of added value in this regard. In particular, non-contrast chest CT, while being less specific than RT-PCR methods, is highly sensitive (86-98% sensitivity[10]) to COVID-19 related changes in pulmonary parenchyma in the later stages of the disease. As such, CT imaging can be combined with RT-PCR testing to improve

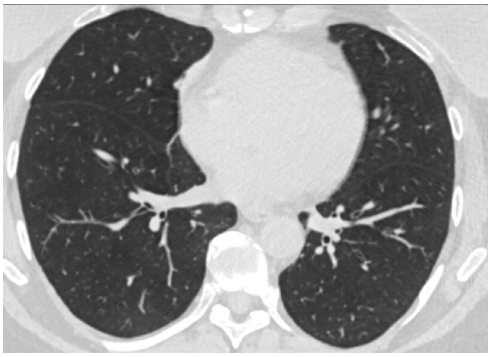
the overall diagnostic accuracy in COVID-19 related pneumonia. Additionally, it can be used to track disease severity and progression in previously confirmed cases[11].

Typically in COVID-19 pneumonia, on chest CT, multiple lobular ground glass opacities (GGOs) are seen bilaterally within the lungs, predominantly in the basal lung fields. The GGOs occur in 86% of COVID-19 patients[12] and are most prominent 0-4 days after the initial onset of symptoms[10]. Subsegmental consolidations, intralesional vascular enlargement and traction bronchiectasis are also typically seen. Crazy paving is another finding that occurs later in disease progression, because of thickening of interlobular septa superimposed on the GGO. Findings may occur in a- and presymptomatic patients, but they can also be absent in the early stages of the disease. In non-COVID-19 pneumonias, GGOs may be observed but bilateral involvement is seen less frequently. Distinguishing between COVID-19 pneumonia and other types on chest CT is a difficult task, as many of the imaging findings are nonspecific. Nevertheless, there is some evidence that COVID-19 has distinct imaging features such as the aforementioned bibasal GGOs, making screening of suspected cases and evaluation of disease progression using CT possible[4,13]. Recently, standardized CT reporting guidelines were proposed, for instance the CO-RADS system[14]. Figure 1.1 shows examples of the different CO-RADS levels. There are issues with the usefulness of these guidelines. In particular, many of the criteria for diagnosing COVID-19 on CT are nonspecific, the same radiographic abnormalities may be seen in non-COVID pneumonias. Furthermore, the guidelines give no clear recommendations regarding subsequent patient management, making them less applicable in a clinical setting[15]. Nevertheless, CO-RADS performs well in predicting COVID-19 disease progression, and it yields substantial inter-observer agreement[14]. As such, predicting a standardized score such as CO-RADS is the most straightforward way of implementing an automated COVID-19 detection method.

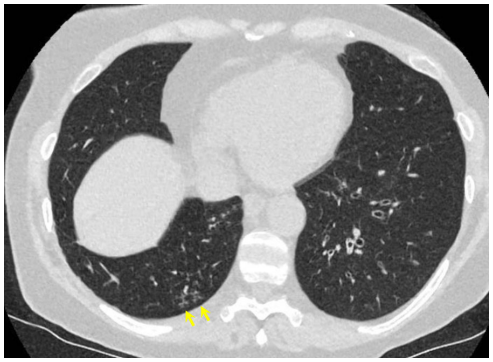
The term radiomics refers to a method of quantifying phenotypic characteristics of lesions on medical imaging using mathematical algorithms and computational tools. These so-called radiomic features can then be used to predict disease severity and progression. This quantitative process contrasts with the conventional radiological method, where the radiologist describes lesions using qualitative attributes. Radiomics originated from the field of medical oncology, in an effort to computationally quantify tumor phenotypes on medical images, and to integrate these phenotypes with information obtained at other levels of scale, such as pathomic and genomic data. In recent years, the field of radiomics has been expanding within the context of non-malignant disease[16].

Radiomic features are extracted at the sub-visual level, meaning that the computer system can detect patterns which might not be discernible by the human visual system. In this way, radiomics may provide valuable clinical information complementary to conventional radiological analysis. The extraction of radiomic features requires an annotated CT scan, where regions of interest (ROIs) have been highlighted using a digital mask. Extraction of features entails calculating quantities based on scan information in the masked areas, such as the surface area. Oftentimes, certain features are redundant or irrelevant, which means that a subset, also known as a signature, of all features predicts the disease with similar accuracy compared to using all features simultaneously. Radiomic techniques are not without issues, however. One of the most important problems is feature variability. During all steps of the image acquisition process, differences in acquisition parameters will result in different feature values. The degree to which this variation affects results is an area of active research within radiomics[17].

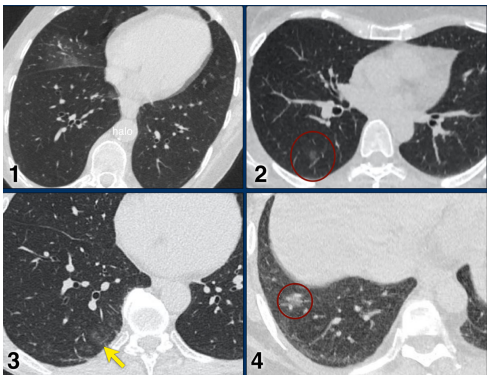
To aid the radiologist when interpreting chest imaging, machine learning methods may



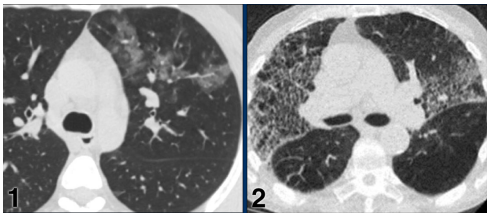
(a) CO-RADS 1



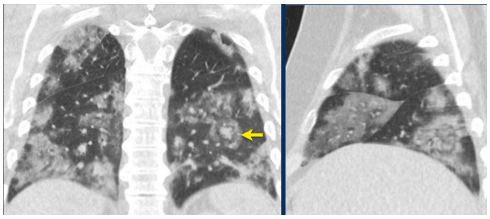
(b) CO-RADS 2



(c) CO-RADS 3



(d) CO-RADS 4



(e) CO-RADS 5

Figure 1.1: **a**: There are no significant findings, normal CT. **b**: CT shows bronchiectasis, bronchial wall thickening and tree-in-bud sign (yellow arrows), but there are no GGOs. Low suspicion of COVID-19 infection. **c**: Abnormalities indicating infection, but not certain whether COVID-19 is involved. Red circles and yellow arrow point to GGOs. **d**: High suspicion level, but findings are not extremely typical. Case 1 shows unilateral areas of GGO in left upper lobe. Case 2 shows bilateral GGO in a patient with emphysema. **e**: Typical COVID-19 findings, bilateral GGO. Images courtesy of Radiologyassistant, url: `https://radiologyassistant.nl/chest/covid-19/corads-classification`, accessed August 2020.

be applied. For example, deep learning methods have been employed with some success, however, a known issue with these methods is the lack of interpretability of the resulting neural network[18]. Radiomic features, on the other hand, have the benefit of being easy to interpret, as they are defined using simple mathematical formulae. Radiomic techniques have been successfully employed in the field of oncology, and a growing body of evidence suggests their usefulness in the context of COVID-19 detection on chest CT[19,20,21,18].

Successfully applying radiomics in this context can accelerate interpretation of the images by the radiologist, potentially reducing time until diagnosis, and therefore directly improving patient care. We hypothesize that COVID-19 pneumonia can be differentiated from non-COVID-19 pneumonia by computationally predicting the CO-RADS score on chest CT using a classification algorithm combined with an optimal radiomic signature and noise reduction techniques. Furthermore, we speculate that performing noise reduction will improve the accuracy of the classification method. To test these hypotheses, we extracted radiomic features from a publicly available dataset of chest CT-scans showing both COVID-19 and non-COVID-19 pneumonias. We compared the results from the unprocessed scans to scans subjected to noise reduction. Subsequently, feature selection was performed to obtain a relevant subset of features, and multiple machine learning classifiers were trained to predict the CO-RADS score.

2 Methods

CT scans were obtained from a single dataset and preprocessed. Subsequently, radiomic features were extracted. Finally, statistical analysis was performed. These processes are described in more detail in the following sections. Figure 2.1 portrays the workflow.

2.1 Data

We extracted radiomic features from a publicly available Italian dataset, consisting mainly of non-contrast chest CT scans, and a small number of abdominal scans showing only the basal lung fields, of 28 patients. The scans originate from multiple different scanners operated at different acquisition settings. A radiologist annotated the scans according to the CO-RADS (COVID-19 Reporting and Data System) classification scheme, as developed by the Dutch radiological society[14], see Table 2.1. For each of these patients, the percentage of lung involvement was also recorded, as well as the suspicion level and specific findings (GGO, crazy paving, pleural effusions, vascular thickening, lesion location and consolidations, see figure 1.1 for an example of the radiographic appearance). We performed noise reduction on all scans to obtain another 28 scans (total n=56), since these are essentially the same scans we copied the annotated information.

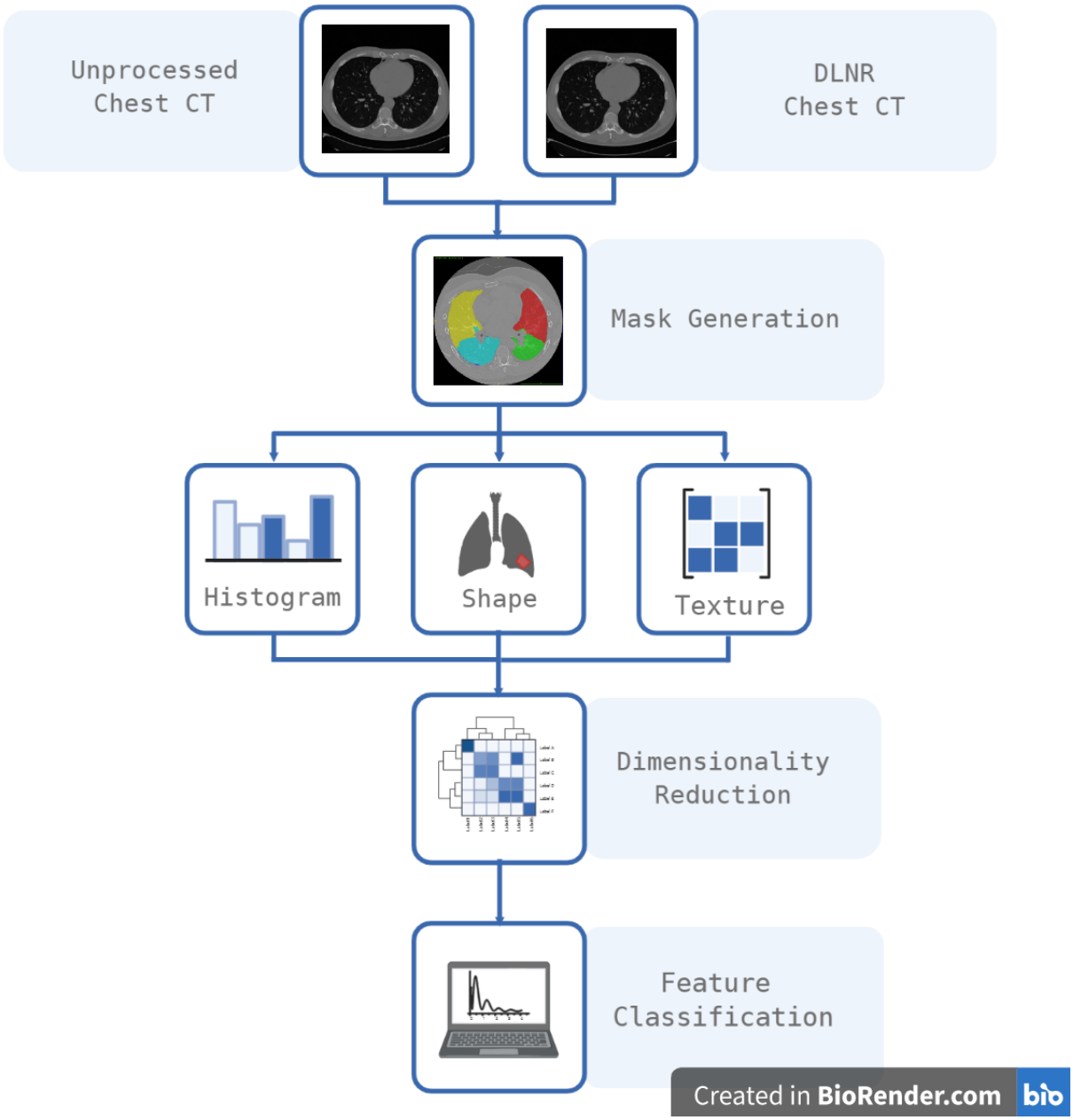


Figure 2.1: Workflow of the analysis. Images from the Italian dataset were acquired online and annotated by a radiologist. A copy of the dataset was processed using Deep learning based noise reduction (DLNR). Masks were generated for each scan. **b)** Histogram, shape and texture based radiomic features are extracted and stored. **c)** Dimensionality reduction is performed to obtain the most relevant radiomic signature. Classification is performed using a variety of algorithms. Finally, results of the classification are compared.

CO-RADS	Suspicion level	Findings
CO-RADS 0	Not interpretable	N.A.
CO-RADS 1	Very Low	Normal findings or non-infectious
CO-RADS 2	Low	Infection other than COVID
CO-RADS 3	Indeterminate	Infection, could be COVID or other
CO-RADS 4	High	Abnormalities specific for COVID
CO-RADS 5	Very high	Typical COVID findings
CO-RADS 6	Proven COVID	RT-PCR SARS-CoV-2 positive

Table 2.1: CO-RADS classification scheme

2.2 Experiments

All preprocessing and segmentation steps and subsequent radiomics experiments were performed on the University of Groningen Peregrine compute cluster, on a machine running a GNU/Linux OS with 24 cores @ 2.5 GHz (2 Intel Xeon E5 2680v3 CPUs) and 128 GB of RAM.

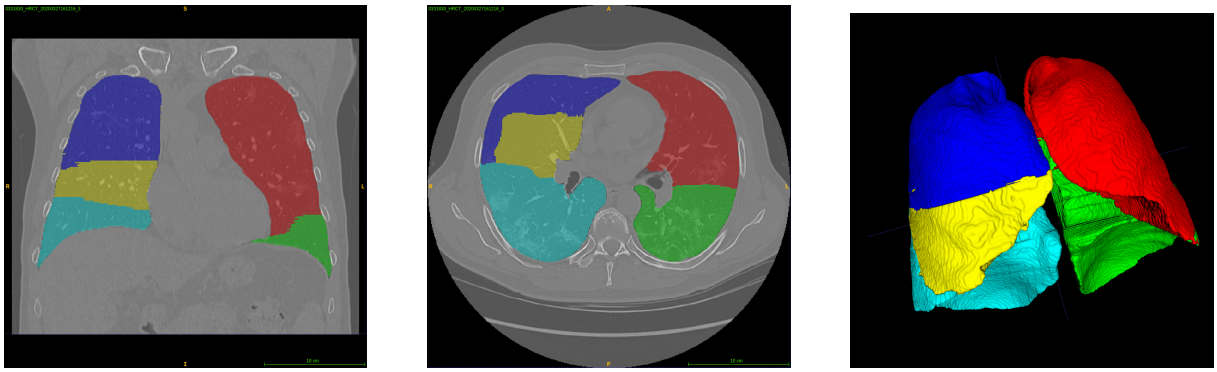
2.2.1 Preprocessing

All preprocessing as described in this section and processing in the subsequent section was performed using a custom-built command line interface (CLI) tool called `CoRa`. The CLI can perform multiple different actions, to prepare data for radiomic extraction and to extract radiomic features. `CoRa` was made specifically for this project. For more information about the specific capabilities of `CoRa`, and to download the latest version, see the software repository hosted on the GitHub website[22]. Additional instructions about using `CoRa` to reproduce the results mentioned here can be found in part 1 of the appendix.

To process the data, we first converted it from Digital Imaging and Communications in Medicine (DICOM), which is the file format scans are normally stored in on a hospitals Picture Archival and Communication System (PACS), to the Neuroimaging Informatics Technology Initiative (NIfTI) file format, which is the format our radiomics analysis software requires. This was done using the `dcm2niix` tool[23].

2.2.2 Deep Learning Based Noise Reduction

Radiomic features are very sensitive to minor variations in image acquisition and reconstruction methods[24]. To circumvent this issue, noise-reduction techniques can be applied to the images. Many different algorithms exist for this purpose. Here, we used a proprietary Deep Learning based Noise Reduction (DLNR) algorithm called Pixelshine (AlgoMedica, Inc, Sunnyvale, CA, USA), which has been shown to effectively reduce image noise and improve image quality[25]. Theoretically, this has the effect of normalizing radiomic features between scans. By using DLNR data, the radiomic signature will therefore be comparable even when different scanners and acquisition protocols are used. Both the DLNR and the unprocessed dataset were included in the analysis.



(a) Frontal view (b) Axial view (c) 3D render

Figure 2.2: 2.2a and 2.2b show the mask as generated using the `lungmask` software overlaid on a chest CT scan. 2.2c shows a 3D render of the lung mask, in a slightly rotated frontal view (note the cardiac notch on the medial aspect of the left lung). The colors represent the different lobes. Note the irregularities at the boundaries caused by uncertainty in the `lungmask` model. Original chest CT from the Italian dataset, images rendered using `ITK-SNAP` imaging software

2.2.3 Lobe segmentation

Once the files were in the correct format, we generated segmentation masks. The masks are necessary to indicate Regions Of Interest (ROIs), areas in the scan where features are to be calculated. We use the `lungmask` Python library to automatically generate masks for the different lung lobes. Specifically, the `LTRCLobes_R231` model was used, which fuses the output of two U-net convolutional neural networks to generate the segmentations. For each scan, at most 5 different ROIs are extracted, corresponding to the 5 radiological lung zones. Note that these masks were automatically generated using software, and as such are not always completely correct[26]. The mask contains the following values:

1. **Red**: Left Upper Lobe (LUL)
2. **Green**: Left Lower Lobe (LLL)
3. **Blue**: Right Upper Lobe (RLL)
4. **Yellow**: Right Middle Lobe (RML)
5. **Cyan**: Right Lower Lobe (RLL)

See Figure 2.2 for a visualization of one scan with the generated segmentations on top.

2.2.4 Radiomic extraction

To perform the radiomic analysis, we used `CoRa`, which utilizes the `PyRadiomics` library. In order to extract radiomic features, `CoRa` reads a so-called case file, which contains a list of filepaths pointing to the location of the scans and the accompanying mask files, and a parameter file which specifies the features that are to be extracted (see Listing 1 in the appendices for the full parameter file). In this way, `CoRa` acts as a wrapper which extends the functionality of the `PyRadiomics` library, which is an open source project capable of extracting radiomic features from medical imaging. Both `PyRadiomics` and `CoRa` are

Feature class	Number of features
First-Order Statistics	19
Shape-Based (2D)	10
Gray Level Co-occurrence Matrix	24
Gray Level Run Length Matrix	16
Gray Level Size Zone Matrix	16
Neighbouring Gray Tone Difference Matrix	5
Gray Level Dependence Matrix	14
Total	104

Table 2.2: Feature classes extracted using `PyRadiomics` library

written in the Python programming language. The goal of the `PyRadiomics` project is to provide a reference standard for the extraction of radiomic features, so that results can be easily compared between different radiomic analyses[27].

2.2.5 Feature classes and image types

From the segmented regions of each scan, radiomic features were derived. The extracted features can be classified as First-order (n=19), Shape-Based (n=10), Gray Level Co-occurrence Matrix (n=24), Gray Level Run Length Matrix (n=16), Gray Level Size Zone Matrix (n=16) and Gray Level Dependence Matrix (n=14) features. An overview of the feature classes is provided in Table 2.2.

All of the aforementioned feature categories except the two-dimensional Shape Based category can be derived from the original image as well as to filtered versions, further increasing the amount of extractable features. Wavelet, Laplacian of Gaussian (LoG), Square, Square Root, Logarithmic, Exponential, Gradient and Local Binary Pattern 2D (LBP2D) filters were applied to each scan. An overview of the image types is shown in Table 2.3. By applying different filters to the image and deriving features, a total of 1130 different radiomic features were extracted per scan.

The analysis proceeds on multiple cores in parallel to reduce the individual workload on the processors, exploiting the parallel properties of this problem. This processing was done on the aforementioned Peregrine cluster. The output of the radiomic extraction was stored in a `.csv` format. This output was then combined with the annotations provided by a radiologist.

2.3 Statistical analysis

The statistical analysis was performed using Waikato Environment for Knowledge Analysis (WEKA) software on a machine with an Intel Quad-Core i7-10510U CPU @ 1.80 GHz with 16 GB of RAM, running a GNU/Linux OS. The `.csv` files which resulted from the radiomic analysis were converted to `.arff` files, which are readable by WEKA.

Image type	Comment
Original	The unaltered image
Wavelet	Coiflet filtering
LoG	Laplacian of Gaussian transform
Square	Squared image
SquareRoot	Square root of image
Logarithm	Log of image
Exponential	Takes the exponential of image
Gradient	Gradient filter of image
LBP2D	Local binary pattern applied in 2D
Total	9

Table 2.3: Image types processed with `PyRadiomics` library

2.3.1 Dimensionality reduction

WEKA was used for dimensionality reduction and classification. Dimensionality reduction techniques can be categorized as feature selection (FS) or feature extraction (FE) techniques. FS is the process of selecting a subset of the original set of features. This in contrast to FE, which is the process of transforming the input features. Both processes reduce the dimension of the feature set, thereby decreasing the complexity of the classification model and reducing redundancy[28].

We compared principal component analysis (PCA, an FE method which we combined with the `Ranker` search mehod) with the correlation based subset evaluation (`CfsSubsetEval` WEKA FS method[29], together with the `BestFirst` search method) as two different methods for dimensionality reduction.

2.3.2 Classifier construction and performance evaluation

After dimensionality reduction, we used the remaining variables to build a model. The CO-RADS score was selected as the dependent variable. Multiple different models were applied and compared, a process known as model selection. We created models using the naive Bayes, Bayesian network, multilayer perceptron and random forest algorithms. Each classification algorithm was applied to the datasets generated by each dimensionality reduction method. When training every model, 10 runs of 3-fold cross-validation were performed. The performance of all classifiers was compared to that of the naive Bayes classifier, using a corrected paired t-test with Nadeau and Bengio correction. This special correction was specifically designed for use with machine learning algorithms, as it takes into account the effects of cross-validation[30]. In addition to the t-test value, true positive and false positive rate, precision, recall, F-measure, area under receiver operating curve (AUROC) and average classification accuracy was examined. These parameters can be combined to give an accurate depiction of classifier performance. Additionally, for the best performing radiomic signature, we investigated the correlation between the features.

3 | Results

To investigate the applicability of radiomics in the context of COVID-19 prediction on chest-CT scans, radiomic features were extracted from chest-CT scans. After preprocessing, we obtained 106 instances of unprocessed data and 103 instances of noise reduced data from the original 56 scans, where an instance is defined as a combination of a scan with a mask. For each instance, a total of 1130 radiomic features were extracted and the results were stored in a `.csv` file. Features were extracted from the original image as well as from the filtered versions.

3.1 Influence of dimensionality reduction and DLNR

Using PCA, we were able to compress the 1130 features to 14, including the CO-RADS score as one of the variables, in both the unprocessed and DLNR dataset, however the compressed features were different between these datasets. The baseline classifier actually performed worse after PCA on the DLNR dataset (51.66%($\pm$7.88) average correct classification), when compared to the unprocessed PCA data (56.80%($\pm$6.39)). This difference was not significant ($p > 0.05$).

Using Correlation-based Feature Subset selection (CFS) we were able to simplify the DLNR dataset to just 13 features, including the CO-RADS score. These features are uncorrelated, which is to be expected after feature selection. Figure 3.1 shows the correlation as a heatmap. In the unprocessed dataset we ended up with 28 features including the CO-RADS score after CFS (see Table 3.1). The baseline classifier performed better after CFS on the DLNR classifier (74.10%($\pm$7.15) versus 69.75%($\pm$6.23) in the unprocessed case), but the difference again was not significant.

3.2 Predictive performance of the different classifiers

In both the unprocessed as well as the DLNR dataset, classifiers performed consistently better after CFS when compared to PCA, as indicated by the average correctness percentages in Table 3.2a.

The predictive performance of the classifiers was assessed using multiple different metrics. In particular, the Area Under Receiver Operator Curve (AUROC) value gives a clear insight into predictive capabilities of the classifiers. The random forest method performed best as a classifier, attaining an average AUROC of 0.94. Second best was the multilayer perceptron, with an average AUROC of 0.91. The third best average performance comes from the Bayes network classifier, which resulted in an AUROC of

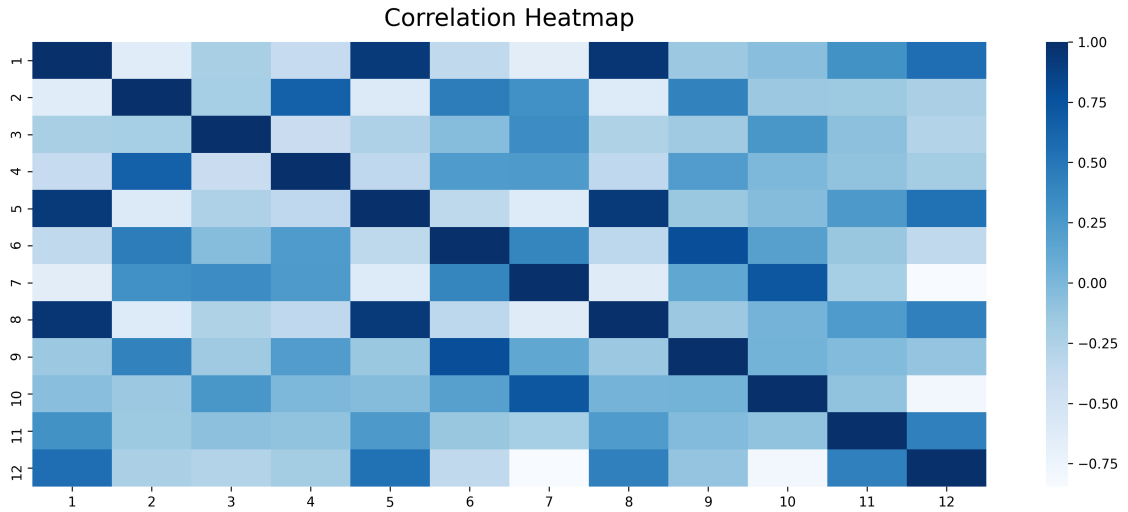


Figure 3.1: Symmetric heatmap showing the correlation between the features in the signature of the denoised data after CFS. After dimensionality reduction we expect there to be low correlation between the different features, and this is indeed the case when looking at the color values of the non-diagonal elements. Diagonal entries have a value of 1, as each feature is perfectly correlated with itself. The axis numbers correspond with the feature names provided in Table 3.1.

#	**Image Type**	**Feature Class**	**Feature Name**
1	Original	First order	10th percentile
2	Original	GLCM	Cluster prominence
3	Original	GLDM	Large dependence emphasis
4	Wavelet	First order	Variance
5	Laplacian of Gaussian	GLCM	IMC2
6	Square	First order	Range
7	Square	GLSZM	Gray level variance
8	Square	GLSZM	Small area low gray level emphasis
9	Square	NGTDM	Complexity
10	Logarithm	First order	Minimum
11	Exponential	First order	Mean
12	Exponential	First order	Minimum

Table 3.1: Radiomic signature of the DLNR data, obtained using CFS.

Dataset	DR	Classifier				
		Naive Bayes	Bayes Network	Multilayer Perceptron	Random Forest	Logistic Regression
Unprocessed	CFS	69.75% (±6.23)	75.89%(±5.59)	76.07%(±6.91)	83.44%(±5.93)	72.48%(±7.94)
	PCA	56.80%(±6.39)	50.26%(±7.36)	72.84%(±7.73)	71.63%(±7.32)	64.98%(±7.51)
DLNR	CFS	74.10%(±7.15)	85.95%(±7.34)	76.45%(±8.19)	90.10%(±4.88)	68.88%(±7.12)
	PCA	51.66%(±7.88)	58.64%(±8.06)	65.25%(±6.90)	68.46%(±6.17)	56.88%(±9.29)
Average		63.08%	67.68%	72.65%	78.41%	65.80%
t-test significant		N.A.	1/4	1/4	4/4	0/4

(a) DR: Dimensionality Reduction method, DLNR: Deep Learning based Noise Reduction, CFS: Correlation-based Feature Subset Selection, PCA: Principal Component Analysis

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	AUROC
Naive Bayes	0.67	0.01	0.72	0.63	0.65	0.85
Bayes Network	0.74	0.01	0.76	0.68	0.72	0.87
Multilayer Perceptron	0.63	0.00	0.73	0.73	0.71	0.91
Random Forest	0.75	0.00	0.81	0.78	0.79	0.94
Logistic Regression	0.75	0.01	0.69	0.66	0.65	0.85

(b) TP: True Positive, FP: False Positive, AUROC: Area Under Receiver Operator Curve

Table 3.2: Statistics describing the results. 3.2a shows correctness percentages for each classifier on the datasets. The final row counts the number of times a corrected paired t-test showed a significant difference when compared with the Naive Bayes classifier. 3.2b shows averages of descriptive statistics for each classifier.

0.87. Finally, conventional logistic regression yielded the same result on average as the baseline classifier, naive Bayes; both had an AUROC value of 0.85. The results of each classifier were compared to naive Bayes using corrected paired t-test; only the random forest performed significantly better on all four variations of the dataset. Bayes network and multilayer perceptron performed better after CFS on the DLNR dataset, and logistic regression never reached the significance threshold, defined as a p-value less than 0.05. Random forest, the best performing classifier, attains an average correctness percentage of 90.10% on the DLNR dataset, after CFS. Table 3.2 provides an overview of these statistics.

4 | Discussion

Previous work shows that radiomics can be successfully applied in pneumonia detection, and in differentiating COVID-19 from other types of pneumonia on non-contrast chest CT[20,31]. Certain groups report diagnostic accuracy of up to 95%[32]. Pu *et al.* show that while not all COVID-19 related pneumonia might be distinguishable from other pneumonias on chest CT using radiomics, a sizable subset can be differentiated using imaging features[21]. Homayounieh *et al.* demonstrate that radiomics is superior to radiologist interpretation when predicting disease severity and outcome on non-contrast chest CT. They did not train a classifier, instead relying on multiple logistic regression analyses[19]. According to Parmar *et al.*, random forest is the best and most robust classification method in the context of lung cancer radiomics[33]. Delzell *et al.* showed, however, when investigating radiomics in lung cancer, that the random forest classifier performed poorly[34]. As such, one should still be cautious when drawing conclusions about the general performance of the classification algorithms, as great performance is not always guaranteed.

Our study demonstrates the potential applicability of radiomics when trying to predict the CO-RADS score on non-contrast chest CT, after automatic delineation of lung parenchyma, thereby predicting the extent of disease progression. By using DLNR, we were able to attain a better classification performance when compared to the unprocessed data. We observe that the random forest algorithm performed significantly better than the baseline method on all variations of the dataset. Random forest after CFS on the DLNR data attained a 90.1% average correctness and an AUROC value of 0.94.

The other classifiers we used were less successful. Interestingly, the Bayes network and multilayer perceptron classifiers performed significantly better than the baseline classifier only in the DLNR dataset after CFS. This implies that DLNR in combination with CFS is an adequate method for training a classifier, as it improves the discriminative ability of all classifiers.

All classifiers performed unsatisfactorily on the data subjected to PCA, this can be explained by the fact that PCA tries to preserve the maximum variance in the data when compressing the features. If this variance is not oriented in a way that discriminates between the underlying classes, then PCA is not an effective method of dimensionality reduction, as evidenced by the low classifier performance[35]. We would expect classifier performance to increase when using the DLNR data, but when combined with PCA all

classifiers except for the Bayes network actually perform worse than in the unprocessed case.

4.1 Limitations and future work

The field of COVID-19 research moves quickly, with new studies being published daily. Some of the material cited in this work was still in preprint version, meaning that no thorough peer-review has taken place yet. The validity of the cited preprint material does not in any way affect the results described here.

In machine learning, large datasets are often required to train reliable classifiers, whereas in our study, our dataset consisted of two variations of 28 scans (unprocessed and DLNR), for a total of 56. Additionally, the dataset we used was rather unbalanced, with only 8 COVID-19 negative scans in total. This may affect classifier performance, as the classifiers we trained here might be ill-equipped to recognize COVID-19 negative scans. Future work should entail training classifiers on larger, more balanced datasets to consolidate our findings. Using larger, more balanced datasets will also allow us to further quantify the effects DLNR has on classifier performance.

4.2 Conclusion

We have attempted to answer the question of whether we can construct a radiomic signature and classifier method that can effectively distinguish between COVID-19 and non-COVID-19 lesions on chest CT scans by estimating the CO-RADS score. Using statistical analysis, we derived a set of 12 radiomic features which can be extracted from a non-contrast chest CT scan after automatic lung segmentation. These radiomic features can be fed to a random forest algorithm to predict the CO-RADS score. The random forest classifier correctly specifies the CO-RADS score in 90.1% of cases on average, and performs significantly better than the baseline classifier on all versions of the data. By harnessing the power of radiomics, it is therefore possible to predict the CO-RADS score, and therefore COVID-19 severity, from a non-contrast chest CT with a relatively high accuracy. Performing noise reduction in the preprocessing step further improves classifier performance. Adopting the aforementioned model into clinical practice will be directly beneficial to patient care by aiding radiologists in diagnosing COVID-19. Additionally, the model can be used as a standardized tool for automatic COVID-19 scoring, to minimize inter-observer variability between different radiologists when scoring chest CTs. Lastly, this model can be used as a research tool, facilitating developments in the field of

automated COVID-19 detection, as well as other areas of clinical research.

Bibliography

[1] L. van Dorp et al. "Emergence of genomic diversity and recurrent mutations in SARS-CoV-2". In: *Infect. Genet. Evol.* 83 (May 2020), p. 104351.

[2] G. Pascarella et al. "COVID-19 diagnosis and management: a comprehensive review". In: *J. Intern. Med.* (Apr. 2020).

[3] L. Wang et al. "Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence". In: *International Journal of Antimicrobial Agents* (2020), p. 105948. ISSN: 0924-8579. DOI: https://doi.org/10.1016/j.ijantimicag.2020.105948. URL: http://www.sciencedirect.com/science/article/pii/S0924857920300984.

[4] D. Zhao et al. "A comparative study on the clinical features of COVID-19 pneumonia to other pneumonias". In: *Clin. Infect. Dis.* (Mar. 2020).

[5] J. Chen et al. "Clinical progression of patients with COVID-19 in Shanghai, China". In: *J. Infect.* 80.5 (May 2020), e1–e6.

[6] S. Behzad et al. "Extrapulmonary manifestations of COVID-19: Radiologic and clinical overview". In: *Clin Imaging* 66 (May 2020), pp. 35–41.

[7] M. Roser et al. "Coronavirus Pandemic (COVID-19)". In: *Our World in Data* (2020). URL: https://ourworldindata.org/coronavirus.

[8] *COVID-19 Richtlijn*. 2020 (accessed May 31, 2020. URL: https://lci.rivm.nl/richtlijnen/covid-19.

[9] J. He et al. "Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China". In: *Respiratory Medicine* 168 (2020), p. 105980. ISSN: 0954-6111. DOI: https://doi.org/10.1016/j.rmed.2020.105980. URL: http://www.sciencedirect.com/science/article/pii/S0954611120301207.

[10] B. Udugama et al. "Diagnosing COVID-19: The Disease and Tools for Detection". In: *ACS Nano* 14.4 (2020). PMID: 32223179, pp. 3822–3835. DOI: 10.1021/acsnano.0c02624. eprint: https://doi.org/10.1021/acsnano.0c02624. URL: https://doi.org/10.1021/acsnano.0c02624.

[11] T. Ai et al. "Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases". In: *Radiology* 0.0 (0). PMID: 32101510, p. 200642. DOI: 10.1148/radiol.2020200642. eprint: https://doi.org/10.1148/radiol.2020200642. URL: https://doi.org/10.1148/radiol.2020200642.

[12] M. Li. "Chest CT features and their role in COVID-19". In: *Radiology of Infectious Diseases* (2020). ISSN: 2352-6211. DOI: https://doi.org/10.1016/j.jrid.2020.04.001. URL: http://www.sciencedirect.com/science/article/pii/S2352621120300309.

[13] W. Zhao et al. "Relation Between Chest CT Findings and Clinical Conditions of Coronavirus Disease (COVID-19) Pneumonia: A Multicenter Study". In: *AJR Am J Roentgenol* 214.5 (May 2020), pp. 1072–1077.

[14] M. Prokop et al. "CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19Definition and Evaluation". In: *Radiology* 296.2 (2020). PMID: 32339082, E97–E104. DOI: 10.1148/radiol.2020201473. eprint: https://doi.org/10.1148/radiol.2020201473. URL: https://doi.org/10.1148/radiol.2020201473.

[15] M. M. Hammer et al. "Challenges in the interpretation and application of typical imaging features of COVID-19". In: *Lancet Respir Med* 8.6 (June 2020), pp. 534–536.

[16] S. S. Yip and H. J. Aerts. "Applications and limitations of radiomics". In: *Phys Med Biol* 61.13 (July 2016), R150–166.

[17] C. Haarburger et al. "Radiomics feature reproducibility under inter-rater variability in segmentations of CT images". In: *Sci Rep* 10.1 (July 2020), p. 12688.

[18] L. Li et al. "Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT". In: *Radiology* 0.0 (0). PMID: 32191588, p. 200905. DOI: 10.1148/radiol.2020200905. eprint: https://doi.org/10.1148/radiol.2020200905. URL: https://doi.org/10.1148/radiol.2020200905.

[19] F. Homayounieh et al. "CT Radiomics, Radiologists and Clinical Information in Predicting Outcome of Patients with COVID-19 Pneumonia". In: *Radiology: Cardiothoracic Imaging* 2.4 (2020), e200322. DOI: 10.1148/ryct.2020200322. eprint: https://doi.org/10.1148/ryct.2020200322. URL: https://doi.org/10.1148/ryct.2020200322.

[20] M. Fang et al. "CT radiomics can help screen the Coronavirus disease 2019 (COVID-19): a preliminary study". In: *Science China Information Sciences* 63 (July 2020). DOI: 10.1007/s11432-020-2849-3.

[21] J. Pu et al. "Any unique image biomarkers associated with COVID-19?" In: *Eur Radiol* (May 2020).

[22] F.G. te Nijenhuis. *CoRa (COVID Radiomics) CLI*. https://github.com/FrankTN/SW_COVID. 2020.

[23] X. Li et al. "The first step for neuroimaging data analysis: DICOM to NIfTI conversion". In: *J Neurosci Methods* 264 (2016), pp. 47–56. DOI: 10.1016/j.jneumeth.2016.03.001.

[24] A. Ibrahim et al. "Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework". In: *Methods* (June 2020). DOI: 10.1016/j.ymeth.2020.05.022.

[25] S. Tian et al. "Potential value of the PixelShine deep learning algorithm for increasing quality of 70 kVp+ASiR-V reconstruction pelvic arterial phase CT images". In: *Japanese Journal of Radiology* 37 (Dec. 2018). DOI: 10.1007/s11604-018-0798-0.

[26] J. Hofmanninger et al. *Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem*. 2020. arXiv: 2001.11767 [eess.IV].

[27] J.M. van Griethuysen et al. "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Research* 77.21 (2017), e104–e107. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-17-0339. eprint: https://cancerres.aacrjournals.org/content/77/21/e104.full.pdf. URL: https://cancerres.aacrjournals.org/content/77/21/e104.

[28] S. Khalid, T. Khalil, and S. Nasreen. "A survey of feature selection and feature extraction techniques in machine learning". In: *2014 Science and Information Conference*. 2014, pp. 372–378.

[29] M. A. Hall. "Correlation-based Feature Subset Selection for Machine Learning". PhD thesis. Hamilton, New Zealand: University of Waikato, 1998.

[30] C. Nadeau and Y. Bengio. "Inference for the Generalization Error". In: *Mach. Learn.* 52.3 (Sept. 2003), pp. 239–281. ISSN: 0885-6125. DOI: 10.1023/A:1024068626366. URL: https://doi-org.proxy-ub.rug.nl/10.1023/A:1024068626366.

[31] L. Li et al. "Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy". In: *Radiology* 296.2 (Aug. 2020), E65–E71.

[32] S. A. Harmon et al. "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets". In: *Nature Communications* 11.1 (2020). DOI: 10.1038/s41467-020-17971-2.

[33] C. Parmar et al. "Machine Learning methods for Quantitative Radiomic Biomarkers". In: *Sci Rep* 5 (Aug. 2015), p. 13087.

[34] D. A. P. Delzell et al. "Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data". In: *Frontiers in Oncology* 9 (2019), p. 1393. ISSN: 2234-943X. DOI: 10.3389/fonc.2019.01393. URL: https://www.frontiersin.org/article/10.3389/fonc.2019.01393.

[35] A. Cheriyadat and L. M. Bruce. "Why principal component analysis is not an appropriate feature extraction method for hyperspectral data". In: *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*. Vol. 6. 2003, 3420–3422 vol.6.

Appendices

These appendices contain additional technical information about the software that was used in this project

---

CoRa

`CoRa` is a command line interface program created to simplify and standardize the process of extracting radiomic features from chest CT scans. Here we describe the steps we took to obtain the results mentioned in the report. For installation instructions and the full `README`, see[22]. We begin by generating the segmentations of the lungs, by using the `CoRa` command:

```
$: cora masks
```

To facilitate preparation of the case file, the `cases` command was added to `CoRa`.

```
$: cora cases -o {CASES.csv} -c {CASE_TYPE}
```

By varying the case type, different case files are generated which correspond to different datasets. This input file will then point to the correct scans for the subsequent feature extraction, provided that the data is set up correctly. Currently, only the `UMCG_R` and `UMCG_D` types are supported. These will point to the cases for the unprocessed and the noise reduced scans of the dataset we used, respectively.

Extraction of radiomic features was performed with the `run` command. Inputs to this command are the case files which were made using the `cases` command and the `params.yaml` file. Note that the input file used in this command should be the same as the output file generated by the `cases` command. Alternatively, the user can manually edit the casefile to point to other scans.

```
$: cora run -p -i {CASES.csv} -o {OUTPUT.csv}
```

Parameter file

```yaml
# Using all image types
imageType:
    Original: {}
    Wavelet:
        binWidth: 10
    LoG:
     sigma: [1.0, 1.0]
    Square: {}
    SquareRoot: {}
    Logarithm: {}
    Exponential: {}
    Gradient: {}
    LBP2D: {}

# Currently enabling all features
featureClass:
    firstorder:
    shape:
    glcm:
    glrlm:
    glszm:
    gldm:
    ngtdm:

# Default settings, force 2D
setting:
    binWidth: 25
    force2D: True
    resampledPixelSpacing:
    interpolator: sitkBSpline
    normalize: False
    removeOutliers:
    geometryTolerance: 0.001 # Very generous tolerance
    correctMask: True
```

Listing 1: Parameter file used in the computations (`params.yaml`).

Feature classes and Image types

Here we describe the feature classes and image types used in the radiomic analysis in more detail. First-Order Statistics quantify the lesions inside the mask using basic metrics,

such as entropy, which uses a histogram of the voxel intensities in the ROI to quantify the amount of probabilistic uncertainty in the image values, and energy, which is simply the sum of squares of all the voxel intensities within the mask. Other measures in this category expand on the basic idea of earlier features. An example of this would be total energy, which takes the aforementioned energy value and scales it based on the volume of the lesion. Note that both energy and total energy are volume-confounded, meaning that lesions of different volume cannot be compared. Other, less complex first order statistics, such as the minimum value (in Hounsfield Units, HU) of voxel intensities in the image, the mean, median and range of voxel intensities are also included in this category. There are 19 features in total in the First Order Statistics category.

The next category of features are based on shape of the ROI. Shape Based features can be extracted both in 2D and 3D, but since one of our datasets is 2-dimensional, consisting of single slices from different scans, we can only analyze data using 2D settings. These features are independent of gray-level intensity distribution and therefore they can only be computed using the mask and the original image. To derive 2D shape features, a circumference mesh is first constructed, which defines the approximate shape of the ROI. An adapted version of the famous marching cubes algorithm is used to generate this mesh. Values include mesh surface area, which measures the size of the triangulated area, approximately corresponding to the area of the ROI, and the perimeter of the mesh. These features can be combined to obtain more complex features, such as the perimeter to surface ratio, which, as the name suggests, divides the perimeter by the surface area. A low value indicates a compact shape like a circle, while a higher value indicates a very irregular ROI. This value depends on the total surface area, thus ROIs of different sizes cannot be readily compared. Another feature in this category is the major axis length, which measures the length of the largest principal component of an enclosing ellipsoid. There are 10 2D shape based features.

The following feature categories are based on matrix algebra. A matrix is a two-dimensional array of numbers. Mathematically speaking, multiplying anything with a matrix constitutes a linear transformation. In the context of radiomic image analysis, these matrices are mainly used to store values, so that feature calculations can be performed using their contents. Note that in the sense of this definition, the 2D image slices in our datasets can also be considered as matrices, where the entries correspond to voxel intensity values.

The Gray Level Co-occurence Matrix (GLCM) category contains features based on the GLCM. The GLCM is constructed by considering the number of times two pixels occur together within a ROI, at a certain distance from eachother. In this construction, the (i, j)th entry in the matrix represents the number of times levels i and j occur together, separated by a distance δ at an angle θ. By default, *PyRadiomics* uses a distance $\delta = 1$. In *PyRadiomics*, the GLCM is symmetric by default, meaning that entry (i, j) is always the same as (j, i). It will have a size equal to the number of discrete intensity values in the image. By default the values are calculated at all possible values of θ and then averaged. This has the effect of making the features rotationally invariant, a desirable property in image analysis, because the average value from all angles is taken. Example features in this category include autocorrelation, which measures the coarseness of the texture of gray-levels within the ROI, and joint energy, which measures the amount of pattern homogeneity in the image by summing the squared values of the normalized co-occurence matrix. Cluster tendency is mathematically more complex, it measures the amount of groupings of voxels with similar gray level values. The GLCM features are useful when

quantifying textural information in the image. 24 GLCM features can be extracted by *Pyradiomics*.

The Gray Level Size Zone Matrix (GLSZM) category is based on another matrix computed from the ROI in the base image. The (i, j)th entry in this matrix represents the number of connected gray level zones of intensity i and size j, where two voxels are connected if there is a distance of 1 between them. In 2D, this means that a single voxel can potentially be connected to its 8 direct neighbours. Features in this category include Small Area Emphasis (SAE), which is an indication of the amount of small size gray level zones. A large value signifies a multitude of small areas, meaning a finely textured ROI. Another feature in this category is Gray Level Non Uniformity (GLN), which measures the variation in gray level intensity values in the image, with low values indicating homogeneity. There are a total of 16 features in this category.

A Gray Level Run Length Matrix (GLRLM) counts the length of the same consecutive voxel intensity along a certain angle. The (i, j)th entry in this matrix represents the number of runs of gray level i of length j along angle θ. Like with the GLCM mentioned previously, the values of features in the GLRLM are calculated at every angle θ, after which the mean value is calculated. Example features include the Short Run Emphasis (SRE) value, which is increased when there are many short runs, indicating fine textural structure within the ROI and the Run Percentage, which is simply the amount of runs divided by the amount of pixels in the ROI. Values close to 1 indicate that there are many runs, meaning a fine texture in the image. This feature therefore quantifies the coarseness of a texture. The GLRLM category contains 16 features.

One of the more complicated matrix categories is based on the concept of a Neighboring Gray Tone Difference Matrix (NGTDM). This matrix contains information about the sum of absolute differences for each gray level between a gray level at a specific voxel and the average value of the neighborhood around this point. An example feature is the coarseness, which measures the average distance between the center voxel and its neighbors, thus giving an idea of the local spatial rate of change. Another feature within this category is the complexity. This value is large when the image is non-uniform, with rapid changes in gray level intensity, which constitutes a complex image. 5 features can be derived from an image using NGTDM analysis.

Lastly, there is the Gray Level Dependence Matrix (GLDM). A voxel is dependent on another voxel if the difference between their intensities is less than or equal to some constant α. In the GLDM, the (i, j)th element signifies the amount of occurrences of a voxel with intensity i with j dependent voxels in its neighborhood. Example features include Small Dependence Emphasis (SDE), where a greater value is indicative of less homogeneous textures, and Gray Level Non-Uniformity (GLN), which measures the similarity of different gray levels in the image using the GLDM. There are 14 GLDM features available for use in radiomic analysis.

All of the aforementioned feature categories except the Shape Based category can be applied to the original image as well as to filtered versions, further increasing the amount of extractable features.

The wavelet image is created by filtering the image using a high or lowpass filter. The Laplacian of Gaussian (LoG) image is created by taking the convolution of the image with the second derivative, also known as the laplacian, of the gaussian filter. This can be seen as an edge detection filter, the value in the constructed image will be high when there is a sudden gray level change in the original image. The σ setting, which defines the width of the filter, specifies which type of textures are emphasized by the filter. A low σ value

emphasizes fine textures, whereas a higher σ can be used to highlight coarse textures.

The square image is obtained by taking the square of the voxel intensities. Note that negative values in the original will be made negative again after taking their square. Similarly, the square root image is obtained by taking the square root of the original intensities. Again, negative originals will remain negative after the application of this filter. The logarithm image returns takes the logarithm of the original image and returns it as a new image, and the exponential image is created by replacing the original image intensity I with e^{cI}, where c is a constant used to rescale the value back to the original range. Again, both the logarithmic and the exponential filter maintain the original sign of the image. The gradient image is computed by convolving the original image with a gradient kernel. The gradient image is well suited for edge detection. The last filtered image we use is the Local Binary Pattern (LBP) image. This image is created by taking the local binary pattern around a voxel for each voxel in the image.

DLNR comparison

(a) Unprocessed

(b) DLNR

Figure 4.1: Here we compare the unprocessed scan to the same slice after DLNR. Even though the difference might seem subtle, it has significant effects on the effectiveness of the algorithm.