Towards a Modular Decision Support System for Radiomics: A Case Study on Rectal Cancer

Roberto Gatta[a], Mauro Vallati[b], Nicola Dinapoli[c], Carlotta Masciocchi[a], Jacopo Lenkowicz[a], Davide Cusumano[c], Calogero Casá[a], Alessandra Farchione[d], Andrea Damiani[a], Johan van Soest[e], Andre Dekker[e], Vincenzo Valentini[c]

[a]*Istituto di Radiologia, Universitá Cattolica del Sacro Cuore. Largo F.Vito 1, 00168, Roma, Italia*
[b]*School of Computing and Engineering, University of Huddersfield, HD1 3DH Huddersfield, UK*
[c]*Polo Scienze Oncologiche ed Ematologiche, Fondazione Policlinico Universitario Agostino Gemelli. Largo A.Gemelli, 8, 00168 Roma, Italia*
[d]*Polo Scienze radiologiche e di laboratorio, Fondazione Policlinico Universitario Agostino Gemelli. Largo A.Gemelli 8, 00168, Roma, Italia*
[e]*Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Netherlands*

Abstract

Following the personalized medicine paradigm, there is a growing interest in medical agents capable of predicting the effect of therapies on patients, by exploiting the amount of data that is now available for each patient. In disciplines like oncology, where images and scans are available, the exploitation of medical images can provide an additional source of potentially useful information. The study and analysis of features extracted by medical images, exploited for predictive purposes, is termed *radiomics*. A number of tools are available for supporting some of the steps of the radiomics process, but there is a lack of approaches which are able to deal with all the steps of the process.

In this paper, we introduce a medical agent-based decision support system capable of handling the whole radiomics process. The proposed system is tested on two independent data sets of patients treated for rectal cancer. Experimental results indicate that the system is able to generate highly performant centre-specific predictive model, and show the issues related to differences in data sets collected by different centres, and how such issues can

affect the performance of the generated predictive models.

Keywords: Decision Support Systems, Radiomics, Predictive Models, Image Feature Analysis

1. Introduction

Personalized medicine is a relatively new, but already well-established, paradigm based on the principle that each individual is born with unique biological and genetic characteristics [1, 2]. The foundation of this paradigm is formed by disciplines such as Genomics –the science of studying the genes in a genome and their interactions with each other–, and proteomics –which instead focuses on proteins. Furthermore, in disciplines like oncology, where images and scans are available, the exploitation of medical images can provide an additional source of potentially useful information. Thanks also to the recent advances in computer science, it is now possible to extract a huge number of "quantitative" features from tomographic images (computed tomography [CT], magnetic resonance [MR], or positron emission tomography [PET] images), and such extracted features can then be automatically analysed in order to investigate their informativeness with regards to the evolution of the disease, or the response of the patient to a specific clinical treatment. This discipline is commonly termed *radiomics* [3], and is aimed at providing effective decision support to physicians and practitioners, and complementing the traditional "qualitative" analysis of images, commonly performed by human experts [4]. In this context, features represent a numerical synthesis of some properties of the considered image, which would not be possible to manually extract and analyse. Extracted features can then be combined with available clinical data into complex models to predict patient prognosis or benefit from a specific therapy.

Remarkably, evidence that radiomics can be helpful for predicting tumour control or clinical complications has been documented for most of the common modalities (CT, MRI, PET, etc.) and anatomical districts –such as lung, rectum, or brain– [5, 6, 7, 8].

The growing interest in radiomics lead to the development of several specifically-designed tools; examples include cGITA [9], TexRAD [10, 11], moddicom [12], Pyradiomics [13], and CERR [14]. In parallel with the grows of radiomics tools, initiative such as the Image Biomarker Standardisation

Initiative [15] and the Radiomics Ontology[1] become important to standardise the different aspects of image processing and features extraction.

However, despite the growing number of radiomics tools, to the best of our knowledge there is a lack of agents that can deal with all the steps of the radiomics process, thus providing a complete and modular environment for supporting the generation and analysis of the predictive models, and allowing the exploitation of models in everyday medical routine. Existing tools are mainly aimed at facilitating the extraction of features, and at extending the set of features that can be extracted from a medical image, only. Pivotal steps, like feature selection and generation of the actual predictive model –either via traditional statistical approaches or recent machine learning techniques– are ignored. More worryingly, existing tools do not natively provide any support for the external validation of generated models. As a result, a crucial issue of the exploitation of radiomics predictive models, is that their portability between centres or hospitals is unclear. This is also due to the fact that different machines, particularly in MRI, provide medical images with very different characteristics, particularly in terms of visual noise. Such differences can strongly affect the predictive capabilities of the generated models, and invalidate the results. A possible way for tackling this issue is to extensively exploit external independent testing sets, and providing tools that are supportive in this regards, in order to validate the generated models [16, 17, 18]. Similarly, features can be analysed in order to identify those which are more robust to common sources of image noise. However, despite the fact that empirical investigations which rely on external validation are deemed to be qualitative better than others by the TRIPOD guidelines [19], this approach can not guarantee the reproducibility of the observed results in every set [20].

The contribution of this paper is twofold. First, we introduce an approach –under the form of a medical agent-based decision support system– for supporting the whole radiomics process. In its current implementation, the agent incorporates some of the ideas and functionalities of moddicom [12]. Given a set of medical images, the proposed system is able to extract a wide range of features, to analyse and select them with regards to the outcome to predict, and to generate an optimised predictive model. When data from a new patient is provided as input, the proposed agent is able to collect features

[1] `https://bioportal.bioontology.org/ontologies/RO`

from available medical images and patient's data, and to return a prediction about the clinical outcome of a proposed treatment. In other words, the agent can be provided high level goals to achieve –such as, generate a predictive model that shows some given properties–, and it able to reason upon available knowledge in order to satisfy, whether possible, the goals. This reduces the burden on human experts, and provides a valuable decision support tool, that can also allow to investigate alternative approaches and models. The agent is centre-specific, but has been designed in order to be capable of exchanging models between agents in different centres and testing generated models on different data, thus supporting external validation. As a second contribution, we investigate the capabilities of the proposed agent in a real-world scenario. We consider two sets of medical images acquired by two different centres for treating patients affected by rectal cancer. By training the system on each set, we empirically demonstrate how the differences in data sets collected by different centres can affect the performance of the generated predictive models.

The remainder of this paper is organised as follows. Section 2 describes the structure of the proposed system and gives details of the considered features. In Section 3 the data sets are introduced, and empirical results are then presented. An extensive discussion is provided in Section 4. Finally, conclusions are given.

2. The Proposed Agent

The architecture of the proposed agent-based decision support system is depicted in Figure 1, and the corresponding software is available at `https://github.com/robertogattabs/RadAgent`. The exploitation of an agent-based approach has a number of advantages. An agent can cope with high level goals, such as generate models that maximise given metrics, by taking into account all the steps of the process. In facts, the agent can reason upon overall and step-specific knowledge in order to modify the behaviour of the corresponding modules, so that the overall goals are achieved. The modular architecture supports the agent by (i) allowing the development and exploitation of off-the-shelf modules that can be substituted without any modification to the rest of the architecture; (ii) providing a standardised interface between the modules; and (iii) allow to modify parameters and behaviour of each component, and isolating the effects on the overall performance.

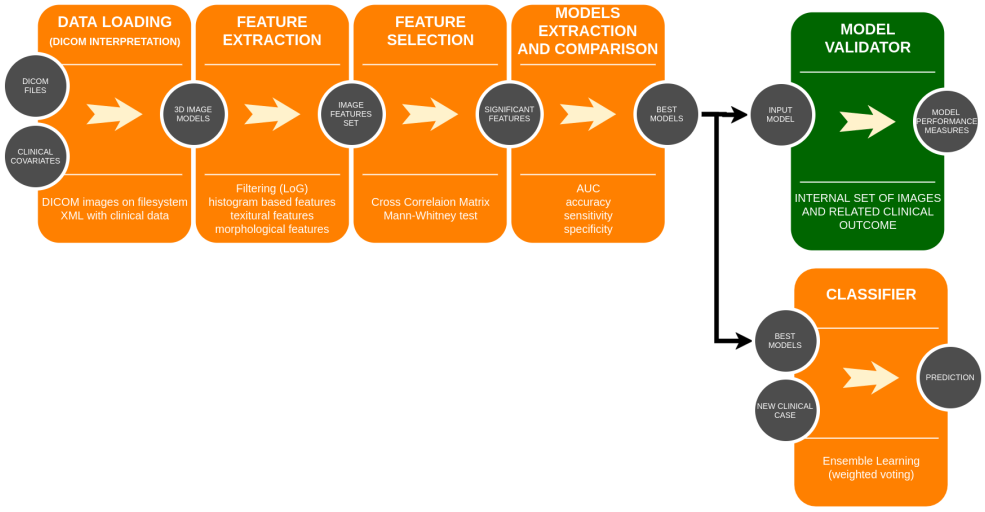

Figure 1: The overall architecture of the proposed decision support system for radiomics, in terms of modules and input/output. The modules included in the architecture correspond to the steps to be performed in the radiomics process. Generated models can be internally validated (green module) or exploited for predicting the outcome of a previously unseen clinical case, and support the physicians.

The system has been designed in order to being able to deal with all the steps of a radiomics analysis, and to provide useful information and support to the physicians. It is worth emphasising that clinicians are generally not very keen to exploit predictive models that can not be inspected and validated "clinically". For this reason, in the rest of this paper we focus on machine learning techniques, such as logic regression or decision trees, that allows to generate predictive models that can be analysed by human experts –but that yet can provide reasonably high performance.

Main functionalities of the proposed system, that can be performed automatically or required by the users, include:

- Features extraction from both original medical images, and images filtered using the well-known Laplacian over Gaussian convolution kernel (LoG) [16]. The LoG filter is commonly exploited in order to smooth the high frequency noise and enhance the variation of values among adjacent pixels in the images.

- The LoG filter can return images with very different appearance according the value of the σ parameter used. It is therefore important

to identify the σ values that lead to most significant and informative features being extracted for the considered outcome to be predicted. σ values are selected by a Mann-Whitney test with the clinical outcome (to identify the most representative σ) and exploiting a cross-correlation matrix, assessed via a p Pearson Test (to allow the use of two different σ for the same feature in case of no-correlation between the feature at the two σ values).

- Signatures of selected features, i.e. subsets of informative features, are evaluated. Signatures are generated via greedy forward selection, and are assessed according to metrics provided by the user. In our analysis we considered the AUC with regards to a logistic regression.

- Signatures are then exploited for generating predictive models, and are compared with regards to their predictive ability. The signature that leads to the best predictive model is selected in order to be used by the agent to support the decisions of the human expert, on new testing cases. Optionally, performance and characteristics of the best performing features can also be presented to the user, which can decide to exploit a different set of features than those included in the signature identified by the agent.

Noteworthy, the introduced agent has a high level of configurability, that allows it to be optimised according to the characteristics of the images and data sources of the centre. In order to maximise its compatibility with existing systems, it has been developed in R, which is one of the most used environments for statistical analysis in medicine. Results of the analysis can be exchanged between agents, in order to (externally) validate results or evolve the generated predictive models.

The architecture is agnostic with regards to the element to be predicted and to the available features. For the purposes of this work, we consider 92 types of features, that can be classified as follows:

- **BASIC**: first order image features [15] extracted by considering aspects such as Morphological (MRF), Statistical (STAT), and Intensity Histogram (HI) of the image. Features in this set also include shape properties, such as Volume, Surface, Surface to volume ratio, Compactness, Sphericity, Centre of mass shift, Mean, Variance, Skewness, Kurtosys, etc.;

- **GLCM**: Grey level co-occurrence based textural features [15]. Features in this set include Mean, Variance, Skewness, Kurtosis, 10th and 90th percentile, Robust mean absolute deviation, Energy, etc.;

- **GLRLM**: Grey level run length based textural features [15]. Examples of features in this set are Short and Long runs emphasis, Short and Long run low grey level emphasis, Grey level non-uniformity normalised, Run entropy, etc.

- **GLDZM**: Grey level size zone based textural features [15]. Features include Grey level non-uniformity, Zone size non-uniformity, Zone percentage, Zone size entropy, etc.

It should be noted that the value of a feature also depends on the considered σ used in the LoG filter. In this implementation of the system, for the sake of efficiency, we consider 9 possible σ values: 0.35, 0.49 , 0.54, 0.59, 0.64, 0.69, 0.74, 0.79, 0.84. Such values has been selected according to the experimental results achieved in [21]. The use of 9 possible σ values leads to a grand total of 734 features (Morphological and shape features, from the Basic set, are not affected by changes in the LoG filter) considered by the approach for generating the predictive model. The complete list of features considered in this work is provided in appendix. For a detailed description of the features, including the actual mathematical formulas, the interested reader is referred to [15].

3. Experimental Analysis

The main purpose of this experimental analysis is to assess the usefulness of the proposed radiomics agent in supporting the different steps of a radiomics investigation. It is therefore beyond the scope of this study to thoroughly compare the performance of differently generated predictive models. For a clinical evaluation of mathematical predictive models, the interested reader is referred to [16].

The experimental analysis considers two data sets of T2-weighted fast spin-echo 2D oblique images MR scans, that are used for treating patients affected by rectal cancer. The first data set includes scans of 173 patients from the Gemelli polyclinic hospital in Rome, the second set is composed by 25 clinical cases treated at the Maastro clinic of the Maastricht university medical centre. Both the data sets of images include manual contouring

of the clinical target volume (CTV) [22]. CTV includes the gross tumour volume, which is the region already affected by the tumour, as well as the regions of direct, local subclinical spread of disease that must be treated in order to stop the evolution of the tumour.

The different size of the sets provides an interesting test-bed for a radiomics decision support system. Typical medical sets can show a significant size variability, according to the typology of tumour considered and to the characteristics of the medical centre.

The scanner used at the Gemelli polyclinic hospital is a MR 1.5 T unit (Signa Excite GE Medical Systems), while the Maastro clinic is equipped with a Siemens Magnetom AVANTO machine. Acquisition parameters were homogeneous for the two data sets, and are as follows:

- repetition time, 2500–5000 msec;
- inversion time, 100–110 msec;
- pixel spacing, ca. 0.7 mm;
- echo train length, 16–24;
- section thickness, 3 mm;
- no intersection gap;

Images have been acquired in a transverse plane orthogonal to the tumour longitudinal axis. No intravenous contrast medium was administered. The subsequent manual contouring was performed by an expert radiation oncologist, using a radiotherapy delineation console (Eclipse, Varian Medical System) for the definition of lesion outline as defined in ICRU n. 83.[2]

Figure 2 shows two MR slices from the Gemelli polyclinic data set (left), and two MR slices acquired at the Maastro clinic. Noteworthy, despite the strict observance of acquisition procedures and acquisition parameters, it is easy to notice that acquired images are significantly different. Qualitatively, images acquired at the Gemelli polyclinic include more visual noise, particularly at high frequencies, than those acquired by the other centre. Moreover,

[2]https://icru.org/testing/reports/prescribing-recording-and-reporting-intensity-modulated-photon-beam-therapy-imrt-icru-report-83

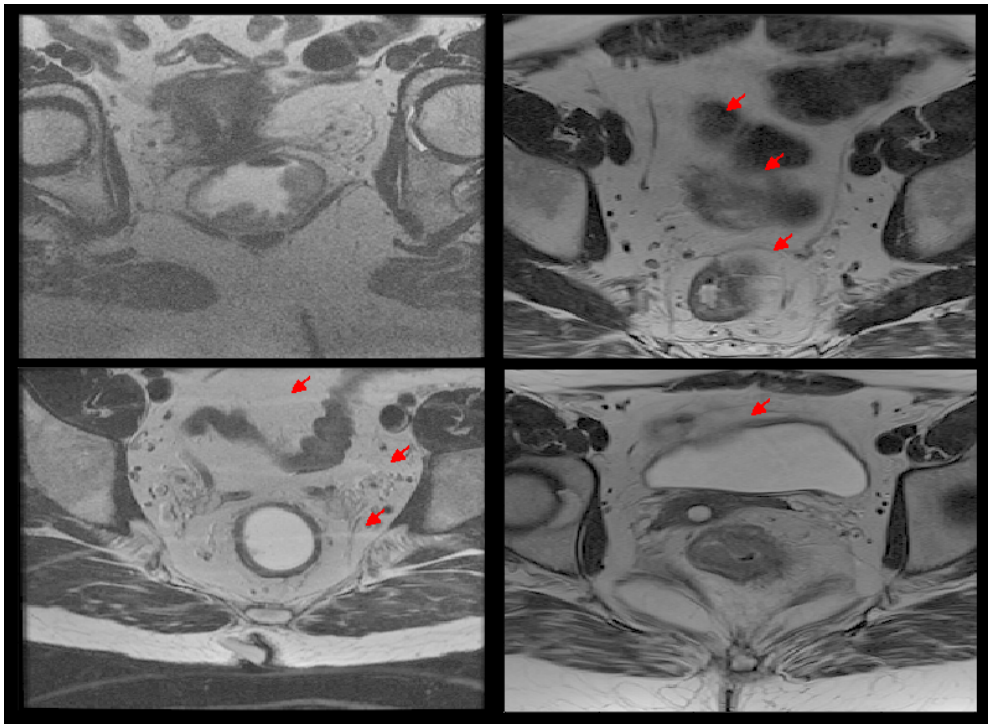

Figure 2: Two MR slices from the Gemelli polyclinic set (left) and from the Maastro clinic (right). Images show significant differences in terms of high frequency noise (upper left), horizontal lines (bottom left), and spotted blurring artefact (upper and bottom right). Such artefacts have been highlighted using red arrows, for the sake of readability.

some horizontal interferences can be spotted (and are pointed in the figure). On the other hand, images acquired at the Maastro clinic may present blurring artefacts, as highlighted in the figure.

In order to provide the appropriate input for the proposed approach, MR scans have been processed using the moddicom R library [12]. Moddicom is an open source library that allows to: (i) deal with DICOM files in order to extract images and contouring information; (ii) process and store extracted data; and (iii) analyse stored data to extract morphological and structural features. DICOM (Digital Imaging and Communications in Medicine) is a standard for storing and transmitting medical images enabling the integration of medical imaging devices such as scanners, servers, workstations, printers, network hardware, and picture archiving and communication systems (PACS) from multiple manufacturers.

The clinical outcome to be predicted trough the generation of radiomics-based models is the pathological complete response (pCR) after surgery, which indicates that there is no residual histological evidence of tumour after surgery. pCR is increasingly found to be a reasonable surrogate for long-term favourable outcomes [23]. In the considered datasets, 21–23% of the cases show a positive pCR. The output of the proposed approach comes in the form of probability of pCR; while the threshold can be provided as input by the user, in this case we exploited a 50%-value threshold. Remarkably, the probability value provides implicitly an estimation of the reliability of the prediction: the closer the probability is to 50%, the lower the confidence.

With the aim of limiting the possibility of overfitting, predictive models are evaluated using a 10-fold cross-validation strategy.

Performance are measured in terms of specificity and sensitivity. The former measures the so-called true negative rate, i.e., the proportion of negative cases that are correctly identified as such. In our analysis, negative cases correspond to the presence of residual histological evidence of tumour, and the absence of a complete pathological response. Sensitivity (also called the true positive rate) focuses on the proportion of correctly classified positive cases.

3.1. Results

Hereinafter we will refer to the predictive model trained, using the proposed framework, on the data set from the Gemelli polyclinic and the Maastro clinic, as respectively, *Ag.G* and *Ag.M*. On the basis of the considered training data, the optimisation procedure included in the architecture lead to the generation of differently-structured predictive models:

- The Ag.G model is based on a logistic regression built using cT (clinical T stage), the zone size entropy [15] after the application of a LoG with $\sigma = 0.35$ and the Skewness of the grey-level distribution after the application of a LoG with $\sigma = 0.59$.

- The Ag.M predictive model is based on two covariates, the Grey level co-occurrence correlation [15] obtained with a $\sigma = 0.84$ and the Grey level co-occurrence joint entropy obtained with a $\sigma = 0.54$. The agent decides automatically the number of covariates to consider according to the size of the provided training set.

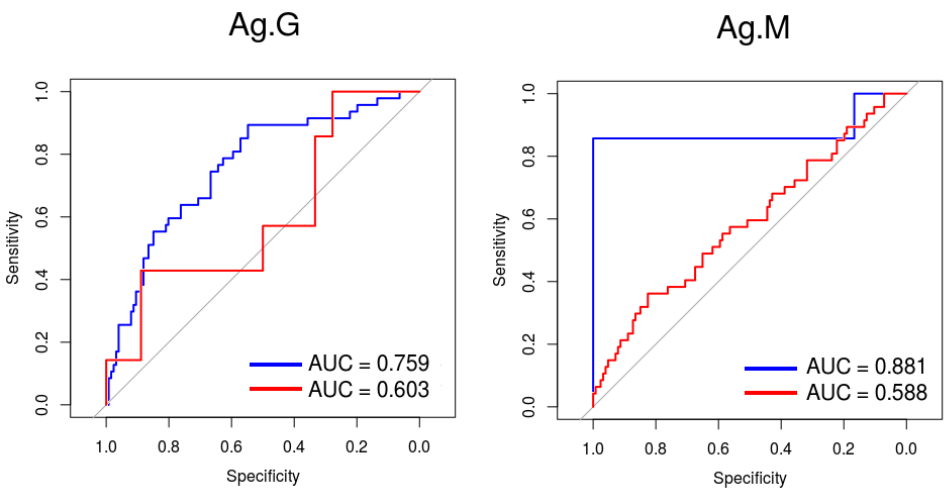

Figure 3: ROC curve of the predictive model trained on the Gemelli polyclinic's data set (left) and on the data acquired by the Maastricht clinic (right). Blue is used to indicate the ROC observed on the training set, –in cross-validation. Red indicates the ROC obtained on the (external) testing set.

It should be noted that the automated optimisation is performed greedily, following the expected AUC value.

Figure 3 shows the receiver operating characteristic curve (ROC) of Ag.G and Ag.M on both the training set (blue) and the external testing set (red). Unsurprisingly, the performance of the models tend to be better on the training set, rather than on the testing set. This is because the testing set images are affected by a different type of noise than the training set ones (examples have been discussed in Figure 2). The extremely good performance of Ag.M on the same data from the Maastricht clinic is possibly due to two main reasons: (i) the limited size of the set, which may result in some overfitting; and (ii) the fact that images acquired by the Maastricht clinic show a very limited noise, or a type of noise to which considered features are robust.

3.1.1. Features Analysis

To shed some light on the informativeness and the significance of considered features in the two data sets, we performed an univariate analysis between each feature and the pCR outcome to be predicted. The analysis was performed using the Mann-Whitney test ($p < 0.05$). Table 1 presents the results of the investigation in terms of number of features that have a

Table 1: Number of features that, at least at one σ value and according to an univariate analysis performed using the Mann-Whitney test ($p < 0.05$), are correlated with the outcome to predict. Features are grouped according to the class they belong to. Results are provided for each considered data set, and also in terms of features which are relevant for both sets.

	BASIC	**GLCM**	**GLRLM**	**GLSZM**
Policlinico Gemelli	8	9	5	4
Maastro clinic	4	9	2	3
Common features	1	4	1	1

correlation with the outcome to predict. For each feature, only the most representative σ has been considered. Features are grouped according to the class they belong to. As a first remark, we observe that out of the total set of available features, a large subset (more than 20%) has a significant correlation with the pCR outcome to be predicted. Considering that the univariate analysis can not take into account combinations of features, this result seems to suggest that considered features can be very informative, as they carry useful information for predicting the required pCR outcome.

Results presented in Table 1 also highlight the limited overlap between the features deemed to be significant between the two data sets. In total, 7 features are identified by the univariate analysis, for at least one σ value, in both the sets.

- Entropy, BASIC;
- Sum Entropy: Textural features, GLCM;
- Correlation: Textural features, GLCM;
- Sum variance: Textural features, GLCM;
- Cluster tendency: Textural features, GLCM;
- Run Entropy: Textural features, GLRLM;
- Large zone high grey level emphasis: Textural features, GLDZM.

Interestingly, most of the features (4) come from the GLCM class, which includes textural features about the grey level co-occurrence. This suggests

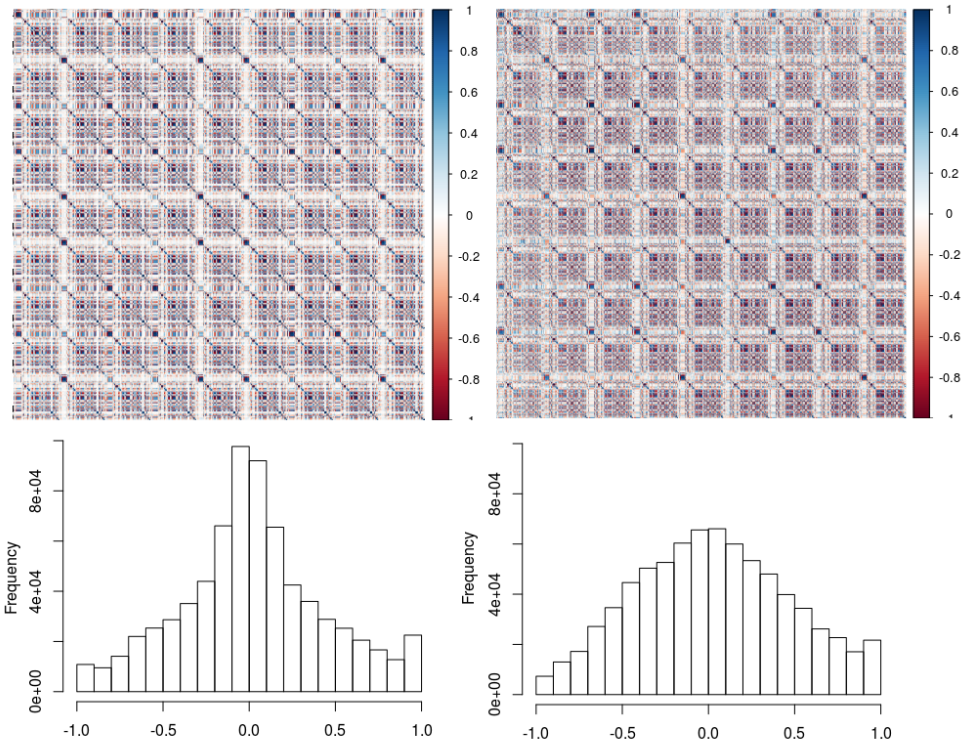

Figure 4: Cross-correlation matrices, using the Pearson correlation coefficient, obtained by analysing the data set of the Gemelli polyclinic (left) and Maastro clinic (right) are presented in the top half. Bottom half shows the distribution of the coefficients under the form of histograms. 0.0 indicates that no correlation is found, while +1 (−1) identifies cases with strong direct (inverse) correlation.

that this class is, in general, more robust with regards to the kind of noise that affects the medical images acquired by the two considered centres.

Figure 4 shows the cross-correlation matrices of the extracted features, and the bivariate correlation –measured using the Pearson correlation coefficient. For the sake of readability, features in the histograms are ordered following the order used in the matrices. Evidence seems to indicate that in the Ag.G set, features have a lower correlation: the region around 0 is very populated. This is possibly due to the noisy of the images in the set, that may reduce the informativeness of extracted information. On the contrary, features in the Ag.M model show a higher level of correlation, as correlation values are evenly distributed among the scale.

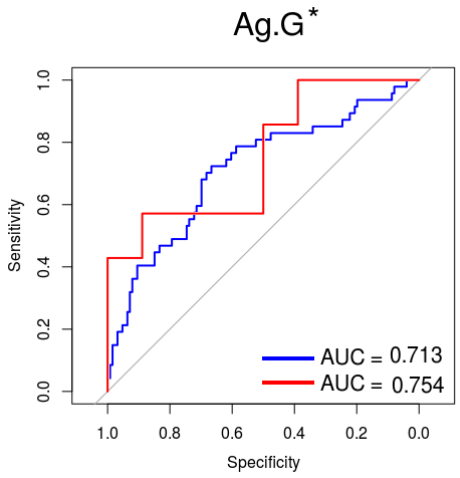

Figure 5: The ROC of a predictive model created by considering as training data images acquired by the Gemelli polyclinic. The model, called Ag.G*, shows to be more general and robust than the Ag.G and Ag.M models, but it delivers slightly worse performance. Blue is used to indicate the ROC observed on the training set, in cross-validation. Red indicates the ROC obtained on the Maastro testing set.

3.1.2. General Predictive Models

It is worth reminding that the Ag.M and Ag.G models have been optimised by the proposed system in order to maximise the performance on images from the corresponding medical centre. Results presented in Figure 3 indicate that trained model perform poorly on a different data set. Therefore, the question naturally arises: *Is it possible to generate a more general and robust predictive model?* To answer this question, we configured the proposed system in order to generate a predictive model according to the approach proposed in [21]: their work was based on a very limited set of features, and showed to be portable and robust. We refer to the resulting model as Ag.G*, because it has been trained using data from the Gemelli clinic. The Ag.G* model is based on a logistic regression built using cT (clinical T stage), the entropy of the grey-level distribution after the application of a LoG with σ=0.35 and the Skewness of the grey-level distribution after the application of a LoG with σ=0.49.

Figure 5 shows the performance of the Ag.G* predictive model. Blue is used to indicate the ROC observed on the Gemelli training set, in cross-validation. Red indicates the ROC obtained on the Maastro testing set.

The generated predictive model provides an interesting trade-off between portability and performance: while the performance on the training set are not as good as those delivered by the Ag.G or Ag.M models, the Ag.G* approach is more robust when used on data from a different centre. This seems to indicate that it is possible to generate a more general and robust model, but at the cost of reduced performance on the specific set.

4. Discussion

According to the presented results, the proposed approach is able to deal with all the steps of a radiomics analysis on data gathered by different centres. Specifically, the proposed framework showed to be capable of identifying a suitable set of informative feature to maximise the performance –measured in terms of AUC with regards to the outcome to be predicted– of a given class of predictive models. In this work, we focused on logistic regression, but the modularity of the framework allows to easily substitute logistic regression with a different class of approaches, or even to consider more approaches at once. We also highlighted how the framework can be exploited for comparing predictive models generated for different data sets, and how the corresponding features (and their characteristics) can be compared and analysed. Remarkably, this analysis can potentially lead to identify issues in the machines or in the environment, or even suggests the presence of procedural issues.

The empirical results presented in the previous section seem to confirm the importance of centre-specific radiomics-based predictive models. Figure 3 suggests that the use of "general" predictive models can lead to very poor predictive performance. However, results also clearly indicate the value of a radiomics-based decision support system, that can provide useful information to physicians and can lead to a more effective planning of the treatments for patients. A trade-off between portability and performance is presented in Figure 5: remarkably, the generated predictive model is less sensitive to the difference in the data sets, for instance in terms of image noise. On the other hand, general performance is worse than those that can be achieved by exploiting centre-specific models.

Taking a different perspective, which is necessarily more speculative than the analysis of the results presented in the previous section, we can identify a number of ways in which the presented system can be exploited with regards to radiomics:

- For the sake of the explainability of the predictive models, a number of different models can be generated for predicting the same clinical outcome. In particular, emphasis can be given to approaches that generate models easy to investigate and analyse by humans, so that an expert user can visualise the generated model, and can explore the relevance of features with regards to the considered clinical outcome. While the number of features can be extremely large, focusing on the described classes of features can highlight the importance of a set of feature, that can be used all together.

- The proposed framework can also allow users to provide as input a specific set of features to be analysed. Such features are then exploited for generating predictive models, and can be compared in terms of relation and correlation. This may allow to investigate features believed to be informative in the relevant literature, and also to assess their usefulness in the presence of images acquired by using different machines, settings, or centres.

- Different data sets can also be compared, in terms of relevant features. For instance, in the presence of large multi-centric studies, it may be useful to identify centres which acquire images with similar properties; that would reduce the noise of the analysis, and maximise the probability of generating an highly performant yet general –with regards to the considered clinics– predictive model.

The physicians involved in the experimental analysis positively evaluated the experience with the proposed agent. The agent allows the medical experts to focus on the actual goals of their investigation and analysis: optimisation and low-level details are optimised by the agent architecture without the need of human guidance. The agent, given a range of alternative modules to choose from, and the parametrisation of each module, can transparently test different alternatives in order to achieve the specified goal. In the presented experimental analysis, the goal was to generate a LR-based predictive model of the pCR of patients treated for rectal cancer. A very important aspect that the agent-based structure can support, but has not been integrated in the proposed system yet, is the ability to *explain* results, and to *motivate* the decisions. We are extremely interested in develop these aspects as part of our future work.

An important aspect to consider, particularly in the case of agent-based decision support system, is the ability to generalise on different data sets. This has been partly covered in the experimental analysis by considering images from two different centres. However, also due to the very limited amount of contoured images available in the radiomics field, it is hard to empirically demonstrate that the proposed agent will easily generalise on data sets where different type of cancer are treated. On this matter, a preliminary study performed by exploiting the proposed agent on a data set considering 15 patients affected by glioblastoma (a form of brain cancer) seems to indicate that the agent, also due to its modularity, can generalise on significantly different sets of MRI images [24].

The agent introduced in this work can play a central role in a distributed learning scenario [25], where different agents cooperate to converge to a robust and shared predictive model while preserving the privacy of patients. This can be achieved by exploiting an iterative approach, shown in Figure 6, composed by four main steps: (a) Each centre trains a local model, (b) the models are sent to a Master, (c) the Master calculates a model, considering weighting the contribute of each centre with the cardinality of the locally available sets, then calculates some new coefficients for each node, (d) the coefficients are sent to each node and the process can be repeated until in a (c) step a convergence criteria is reached.

However, we also envisage the use of the introduced agent in distributed learning scenarios where federated learning approaches are exploited [26], where there is no need for a master to coordinate learning and merge a general model.

5. Conclusion

Radiomics is a topic that is gaining a significant interest in the scientific community, as testified by the growing number of publications that can now be found on the online library of medicine-related articles Pubmed. While still in its infancy, a number of tools are now available for supporting radiomics, as well as standardisation initiatives. These initiatives, such as IBSI [15] are mainly aimed at maximising the reproducibility of results.

Despite the growing interest and the number of already available tools, there is a lack of agents that can deal with all the steps of the radiomics process. Existing tools are mainly aimed at facilitating the extraction of features, and at extending the set of features that can be extracted from a

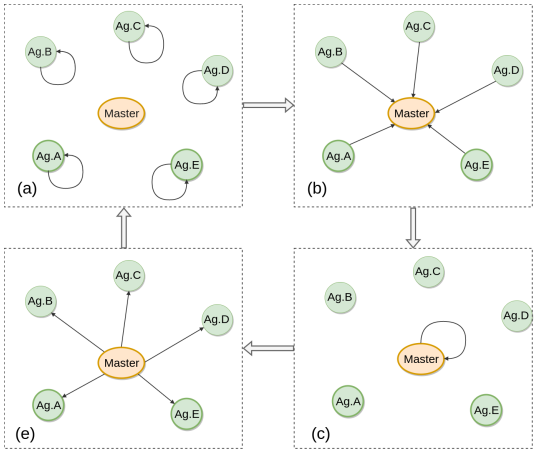

Figure 6: An example of a possible architecture of cooperative agents to converge to a robust and shared model by an iterative approach. Initially, each centre trains a local model using its local agent (a), that is then sent to a Master (b). The master agent merges the models into a general one (c), and send it back to each centre (d).

medical image, only. Crucial aspects, such as features selection, correlation between features and the outcome to predict, and the generation of the actual predictive model, are normally ignored.

In the light of the peculiarities of radiomics, such as the very different characteristics of acquired images according to the exploited machine or the location of the machine, two lines of evolution of radiomics can be envisaged:

- **General and Portable models**. By identifying features that are robust with regards to different sources of image noise that can be found in images acquired in different centres, it could be possible to generate general and portable predictive models, which would allow to exploit the availability of numerous –even though sparse– sets of images. On the other hand, the focus on portability would lead to under performing (when compared to centre-specific) models, with clear negative repercussion on the quality of the treatment delivered to patients.

- **Centre-specific models**. By dropping any requirement related to the portability of models, a significant performance boost can be obtained by highly optimised centre-specific predictive models. This would allow every centre to train a model that is specific for the characteristics of the machines, and for the typology of noise which is included in the

acquired images. A significant drawback would then be that any change in the environment, e.g. a new machine is bought to substitute and obsolete one, may dramatically reduce the reliability of the generated model. Furthermore, this approach does not allow to investigate, in a general sense, the importance and robustness of features.

In this paper, we introduced a medical agent-based decision support system which is capable of supporting the whole radiomics process[3]. The agent can be given a high level goal, and is then able to reason in order to achieve it. Given a set of medical images, the proposed system is able to extract a wide range of features, to analyse and select them with regards to the outcome to predict, and to generate an optimised predictive models. When data from a new patient is provided as input, the proposed agent is able to collect features from available medical images and patient's data, and to return a prediction about the clinical outcome of a proposed treatment. Our experimental analysis demonstrated the ability of the system, and highlighted that the proposed architecture is capable of supporting both the lines of research mentioned above: predictive models can be optimised for a specific centre, and then exchanged in order to analyse the differences. Furthermore, data sets can be merged in order to generated general predictive models, or more general approaches can be used for the creation of predictive models.

We see several avenues for future work. We are actively working on four aspects.

1. The exploitation of additional data sets for testing the capability of the proposed medical agent-based decision support system to generalise on different types of images and contouring.
2. A graphical user interface, that would create a more comfortable environment for researchers.
3. The development of additional modules for performing different kind of features selection algorithms, and extend the set of techniques that can be used for generating the actual predictive model. Specifically, we are looking into neural networks [27], SVM [28], and decision trees [29]. Neural networks will need only a subset of the currently developed modules of the proposed decision support agent, but this aspect is already supported by the modularity of the system.

[3]The agent-based decision support system can be downloaded from `https://github.com/robertogattabs/RadAgent`

4. An approach for extracting information about the spectral components (and other measurable aspects) of image noise of images included in the considered data set. Such analysis will allow to assess the impact of different sort of noise on the predictive capabilities of (some set of) considered features, and to better counter-balance it. As a result, it would be possible to generate more robust predictive models.
5. Improving the capabilities of the agent, so that it can explain the obtained results and motivate the decisions taken.
6. An architecture to support multi-centric investigation based on the distributed learning principles.

Acknowledgement

We want to thank Silvia Chiesa for providing us insights and data of MRI images of brain cancer patients.

[1] F. Collins, The language of life: DNA and the revolution in personalised medicine, Profile Books, 2010.

[2] L. Wen-Ling, T. Fuu-Jen, Personalized medicine: A paradigm shift in healthcare, BioMedicine 3 (2013) 66 – 72.

[3] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, et al., Radiomics: extracting more information from medical images using advanced feature analysis, European Journal of Cancer 48 (2012) 441–446.

[4] R. J. Gillies, P. E. Kinahan, H. Hricak, Radiomics: Images are more than pictures, they are data, Radiology 278 (2016) 563–577.

[5] E. Sala, E. Mema, Y. Himoto, H. Veeraraghavan, B. JD., et al., Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging, Clinical Radiology 72 (2017) 3–10.

[6] M. Hatt, F. Tixier, L. Pierce, P. Kinahan, C. Le Rest, D. Visvikis, Characterization of pet/ct images using texture analysis: the past, the present any future?, European Journal of Nuclear Medicine and Molecular Imaging 44 (2017) 151–165.

[7] G. Lee, H. Lee, H. Park, M. Schiebler, E. van Beek, Y. Ohno, J. Seo, A. Leung, Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art, European Journal of Radiology (2016).

[8] S. Alobaidli, S. McQuaid, C. South, V. Prakash, P. Evans, N. A, The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning, British Journal of Radiology (2014).

[9] Y.-H. D. Fang, C.-Y. Lin, M.-J. Shih, et al., Development and evaluation of an open-source software package cgita for quantifying tumor heterogeneity with molecular images, BioMed Research International (2014).

[10] C. Chatwin, R. Young, B. Ganeshan, Texrad-feedback plc - cancer management imaging software. project report (2015).

[11] M. Strzelecki, P. Szczypinski, A. Materka, A. Klepaczko, A software tool for automatic classification and segmentation of 2d/3d medical images, Nuclear Instruments and Methods In Physics Research 702 (2013) 137–140.

[12] N. Dinapoli, A. Alitto, M. Vallati, R. Gatta, R. Autorino, L. Boldrini, A. Damiani, V. Valentini, Moddicom: a complete and easily accessible library for prognostic evaluations relying on image features, Conference Proceeding IEEE Engineering in Medicine and Biology Society (2015) 771–774.

[13] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer research 77 (2017) e104–e107.

[14] L. Zhang, D. Fried, X. Fave, L. Hunter, J. Yang, C. L., ibex: An open infrastructure software platform to facilitate collaborative work in radiomics, Medical Physics 42 (2015) 13411353.

[15] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, Image biomarker standardisation initiative - feature definitions, CoRR abs/1612.07003 (2016).

[16] N. Dinapoli, C. Cas, B. Barbaro, G. Chiloiro, A. Damiani, M. Di Matteo, A. Farchione, M. Gambacorta, R. Gatta, V. Lanzotti, C. Masciocchi, V. Valentini, Radiomics for rectal cancer, Translational Cancer Research 5 (2016).

[17] B. Altazi, G. Zhang, D. Fernandez, M. Montejo, D. Hunt, J. Werner, M. Biagioli, E. Moros, Reproducibility of f18-fdg pet radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms, Journal of Applied Clinical and Medical Physics (2017).

[18] A. J. Wong, A. Kanwar, A. S. Mohamed, C. D. Fuller, Radiomics in head and neck cancer: from exploration to application, Translational Cancer Research 5 (2016).

[19] G. Collins, J. Reitsma, D. Altman, K. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement, Annals of Internal Medicine 6 (2015) 55–63.

[20] S. Powers, V. McGuire, L. Bernstein, A. Canchola, A. Whittemore, Evaluating disease prediction models using a cohort whose covariate distribution differs from that of the target population, Stat Methods Med Res (2017).

[21] N. Dinapoli, J. van Soest, C. Masciocchi, C. Cas, V. Lanni, A. Damiani, R. Gatta, B. Barbaro, M. Di Matteo, F. Cellini, M. Gambacorta, A. Dekker, P. Lambin, V. V, Radiomics in magnetic resonance imaging for prognosis in patients with rectal cancer: An independent external validation, Radiation Oncology 96 (2016) E180–E181.

[22] N. G. Burnet, S. J. Thomas, K. E. Burton, S. J. Jefferies, Defining the tumour and target volumes for radiotherapy.cancer imaging, Cancer Imaging (2004).

[23] L. Ferrari, A. Fichera, Neoadjuvant chemoradiation therapy and pathological complete response in rectal cancer, Gastroenterology Report 3 (2015) 277–288.

[24] S. Chiesa, M. Lupattelli, R. Gatta, I. Palumbo, M. Balducci, R. Tarducci, R. Cusumano, C. Masciocchi, J. Lenkowicz, M. Martucci,

P. Floridi, N. Dinapoli, F. Beghella Bartoli, V. Valentini, C. Aristei, C035 delta radiomica delle caratteristiche delle immagini per predire gli outcomes nei pazienti con glioblastoma multiforme: studio prospettico multicentrico- gli.f.a. project (english), in: Associazione Italiana Radioterapia Oncologica (AIRO).

[25] A. Damiani, M. Vallati, R. Gatta, N. Dinapoli, A. Jochems, T. Deist, J. van Soest, A. Dekker, V. Valentini, Distributed learning to protect privacy in multi-centric clinical studies, in: Artificial Intelligence in Medicine, AIME, pp. 66–75.

[26] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, W. Shi, Federated learning of predictive models from federated electronic health records, International Journal of Medical Informatics 112 (2018) 59 – 67.

[27] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, Journal of biomedical informatics 35 (2002) 352–359.

[28] C. Cortes, V. Vapnik, Support vector machine, Machine learning 20 (1995) 273–297.

[29] W.-Y. Loh, Classification and regression trees, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (2011) 14–23.

Appendix A. Detailed List of Features

Here we provide the list of 92 features exploited in this work. They are described in the Image biomarker standardisation initiative Reference manual [15]. In order to make it easier, for the interested reader, to identify the features in the reference manual, the same id is used in the following list.

- **BASIC**
  - 4.1.1 Volume
  - 4.1.3 Surface area
  - 4.1.4 Surface to volume ratio

- 4.1.5 Compactness 1
- 4.1.6 Compactness 2
- 4.1.7 Spherical disproportion
- 4.1.8 Sphericity
- 4.1.9 Asphericity
- 4.1.10 Centre of mass shift
- 4.1.11 Maximum 3D diameter
- 4.1.12 Major axis length
- 4.1.13 Minor axis length
- 4.1.14 Least axis length
- 4.1.15 Elongation
- 4.1.16 Flatness
- 4.3.1 Mean
- 4.3.2 Variance
- 4.3.3 Skewness
- 4.3.4 Kurtosis
- 4.3.5 Median
- 4.3.6 Minimum grey level
- 4.3.7 10th percentile
- 4.3.8 90th percentile
- 4.3.9 Maximum grey level
- 4.3.10 Interquartile range
- 4.3.11 Range
- 4.3.12 Mean absolute deviation
- 4.3.13 Robust mean absolute deviation
- 4.3.17 Energy
- 4.3.18 Root mean square
- 4.4.18 Entropy
- 4.4.19 Uniformity

- **Grey level co-occurrence based features – Texture features (GLCM)**
  - 4.6.1 Joint maximum
  - 4.6.2 Joint average
  - 4.6.3 Joint variance
  - 4.6.4 Joint entropy
  - 4.6.5 Difference average
  - 4.6.6 Difference variance
  - 4.6.7 Difference entropy
  - 4.6.8 Sum average
  - 4.6.9 Sum variance
  - 4.6.10 Sum entropy
  - 4.6.11 Angular second moment
  - 4.6.12 Contrast
  - 4.6.13 Dissimilarity
  - 4.6.14 Inverse difference
  - 4.6.15 Inverse difference normalised
  - 4.6.16 Inverse difference moment
  - 4.6.17 Inverse difference moment normalised
  - 4.6.18 Inverse variance
  - 4.6.19 Correlation
  - 4.6.20 Autocorrelation
  - 4.6.21 Cluster tendency
  - 4.6.22 Cluster shade
  - 4.6.23 Cluster prominence
  - 4.6.24 First measure of information correlation
  - 4.6.25 Second measure of information correlation
- **Grey level run length based features – Texture features (GLRLM)**

673 – 4.7.1 Short runs emphasis
674 – 4.7.2 Long runs emphasis
675 – 4.7.3 Low grey level run emphasis
676 – 4.7.4 High grey level run emphasis
677 – 4.7.5 Short run low grey level emphasis
678 – 4.7.6 Short run high grey level emphasis
679 – 4.7.7 Long run low grey level emphasis
680 – 4.7.8 Long run high grey level emphasis
681 – 4.7.9 Grey level non-uniformity
682 – 4.7.10 Grey level non-uniformity normalised
683 – 4.7.11 Run length non-uniformity
684 – 4.7.12 Run length non-uniformity normalised
685 – 4.7.13 Run percentage
686 – 4.7.14 Grey level variance
687 – 4.7.15 Run length variance
688 – 4.7.16 Run entropy

689 • **Grey level size zone based features – Texture features (GLDZM)**

690 – 4.8.1 Small zone emphasis
691 – 4.8.2 Large zone emphasis
692 – 4.8.3 Low grey level zone emphasis
693 – 4.8.4 High grey level zone emphasis
694 – 4.8.5 Small zone low grey level emphasis
695 – 4.8.6 Small zone high grey level emphasis
696 – 4.8.7 Large zone low grey level emphasis
697 – 4.8.8 Large zone high grey level emphasis
698 – 4.8.9 Grey level non-uniformity
699 – 4.8.10 Grey level non-uniformity normalised
700 – 4.8.11 Zone size non-uniformity

- 4.8.12 Zone size non-uniformity normalised
- 4.8.13 Zone percentage
- 4.8.14 Grey level variance
- 4.8.15 Zone size variance
- 4.8.16 Zone size entropy