Artificial Intelligence Research and Development
A. Cortés et al. (Eds.)

doi:10.3233/FAIA220350

Analyzing the Reliability of Different Machine Radiomics Features Considering Various Segmentation Approaches in Lung Cancer CT Images

Maryam TAHMOORESI [a,1], Mohamed ABDEL-NASSER [a,b], and Domenec PUIG[a]
[a] *Department of Computer Engineering and Mathematics, Univerity Rovira i Virgili, 43007 Tarragona, Spain*
[b] *Electrical Engineering Department, Aswan University, Aswan 81528, Egypt*

Abstract. Cancer is generally defined as the uncontrollable increase of number of cells in the body. These cells might be formed anywhere in the body and spread to other parts of the body. Although the mortality rate of cancer is high, it is possible to decrease cancer cases by up to 30% to 50% through taking a healthy lifestyle and avoiding unhealthy habits. Imaging is one of the powerful technologies used for detecting and treating cancer at its early stages. Nowadays, scientists admit that medical images hold more information than their diagnosis, which is called a radiomics approach. Radiomics demonstrate that images comprise numerous quantitative features that are useful in predicting, detecting, and treating cancers in a personalized manner. While radiomics can extract numerous features, not all of them are useful. It should not be neglected that the outcome of data analysis is highly dependent on the selected features. There are different ways of finding the most reliable features. One possible way is to select all extracted features, analyze them, and find the most reproducible and reliable ones. Different statistical analysis metrics could analyze the features. To discover and introduce the most accurate metrics, in this paper, different statistical metrics used for measuring the stability and reproducibility of the features are investigated.

Keywords. CAD, reliability, radiomics, lung cancer

1. Introduction

According to WHO cancer is one of the main causes of death with around 10 million deaths in 2021. Although the mortality of cancer cases is high, cancer cases can be prevented from 30% to 50%, and the others diagnosed in the early stages and with the appropriate treatment can be controlled or be cured.

One of the powerful technologies that are used for early detection and treatment is imaging [1]. Medical images allow us to visualize the entire body that is not possible to see by the naked eye [2]. Among different image types, MRI, ultrasound, PET, and CT are used widely for early detection, finding the stage of cancer, morphology, density change, etc. [3]. In the past, medical images were used to diagnose the presence of tumors, and this diagnosis depended on the experience and knowledge of physicians, but after a

[1] Corresponding Author: maryam.tahmooresi@yahoo.com .

while, with the advancement of science and the advent of artificial intelligence, efforts to extract more accurate information images continued to reduce the dependence of results on human science, thus reducing time and error. In recent years, scientists have come to realize that medical images contain information more than they diagnose [3] and this is called a radiomics approach. Radiomics shows that images contain innumerable quantitative features which can be used for predictive, detective, prognostic, and treatment personalized of the cancers [4][5][6].

Although radiomics can extract thousands of features, it does not mean that they are all useful. In addition, it is necessary to consider that the result of data analysis methods strictly depends on the chosen features, and it can be affected by some of them, thus it might lead to achieving poor results. Therefore, finding the robustness features to make the model, is of utmost important [7].

There are different ways to find the most reliable features, one of these ways is to select all extracted features and analyze them to find the most reliable and reproducible feature [7]. There are different statistical analysis metrics to analyze the features. In this work, we investigate different statistical metrics used to measure the features' stability and reproducibility to introduce the most accurate ones.

According to the data that we have, ICC is the most useful metric, but it has some limitations mentioned in the next sections. For this reason, the main aim of this paper is to use ICC and other appropriate metrics and compare the results to find an alternative metric. Our aim is broken into three objectives that were followed. First, compare manual and semi-automatic segmentation to find out the more accurate one using ICC and Kruskal. The result of the experiments proved that semi-automatic is better than manual segmentation. Second, the comparison was repeated by Kruskal, but this time 2 trainer oncologists were eliminated to see whether it could improve the manual segmentation or not. This semi-automatic time segmentation shows better performance, too. Additionally, the feature categories were compared to determine which one had more reliable features. The results show that first order and NGTDM have the best results with 100%.

2. Related Works

Among different statistical metrics to evaluate the extracted features. Intra-class Correlation Coefficient is the one applied by most researchers for different reliability analysis types like test-retest, interrater, and interrater [8].

Baeßler, Weiss, and Dos Santos [9] worked on MRI to find the robustness features. Therefore, they chose different fruits/vegetables and scanned them by using FLAIR, T1W, and T2W with high and low resolutions. Later, the extracted features were used for test-retest and intraobserver and interobserver analysis., concordance correlation coefficient (CCC) was used for test-retest and, intraclass correlation coefficient (ICC) was implemented for interobserver and intraobserver. According to the achieved results, high-resolution FLAIR images showed the most reliable features, and they can be used for medical aims but in the case of T1W and T2W, it is necessary to take care to choose the features.

Lee et al. [10] focused on MRI scanning protocol parameters to find the effect of different parameters on radiomics features. They used some parameters like T1W, T2W, NEX, etc. with two scanners. ICC and COV were used to analyze the results of the test-retest scheme. The results show that scanning parameters and scanners which are used,

can affect the radiomics features and among these features, the ones with high ICC and CV can be considered reliable features to use.

Zwanenburg et al. [11] worked on robustness radiomics features for CT scan images by adding noise, translation, rotation, etc. as an alternative way for test-retest analysis. For this aim, they worked on two cancer datasets including non-small-cell lung cancer (NSCLC) and head-and-neck squamous cell carcinoma (HNSCC) to check the reproducibility of the extracted features for perturbation and test-retest and compared the results. ICC is the statistical metric that is used for the measurement and showed that this perturbation chain may use instead of test-retest.

Fiset et al. [12] worked on finding reliable radiomics features for cervical cancer and MRI is the image that they selected to work. They performed their analysis in three models including test-retest, diagnostic MRI and simulation MRI, and interobserver to find out which model can produce the most reproducible features. ICC showed that features of the test-retest chain are the most reliable and among the features' categories, shape features are the best ones.

As we mentioned before ICC is the most useful statistical metric, is used to find the robustness radiomics features, but this metric has some limitations. Here we review some papers that mentioned the ICC limitations.

Mehta et al. [13] worked on the dependency of ICC to subject distribution and sample size. They found out that convex distribution has less ICC than uniform distribution and even, less than concave distribution and this dependency is a problem to using the ICC results for reliability analysis. In the second step, they checked the effect of sample size. Thus, they used a fixed type of distribution and the findings proved that increasing the number of samples has an impact on ICC until for example n= 80 and after that, there is no effect. They believe that, although most researchers use ICC for analysis, they should be aware of its conditions, usage, and limitations.

Pleil, Wallace, Stiegel, and Funk [14] studied articles for explaining the importance of repeat measures in biomonitoring research to assess variability and eventually calculating health risk. The aim is (1) to introduce the idea of creating measurements for biomarkers, (2) to review the records of using ICC (intra-class correlation coefficients) in health-based decision making, and (3) to examine the effectiveness of various methods in ICC calculation making. According to the result of ICC, they argue that ICC estimates' precision is highly influenced by the sum of samples, the number of repeat measures, and the special sample distribution.

Chen and Barnhart [15] worked on ICC and CCC and believe these are the most common metrics which are used for analyzing reliability. Not only do they consider the effects of subject and observer with repeated measurements, but also the effects of time on data. Because practically, it is not easy to gain the true replications. these two indices of the agreement for various combinations of fixed or random effects of time and observer are compared. ultimately, 2D-echocardiogram image data is used for illustrating the suggested methodology and comparing these 2 indices. In case, of repeated measurements, one needs to choose between these two indices, using a new concordance correlation coefficient is recommended.

According to the limitations of the ICC, we aim to use other metrics to find the most reliable features. ICC and other metrics are calculated for the extracted features. The achieved results will be compared to find the best metric. In the next section, we explain this process in detail.

3. Methodology

3.1. Dataset

The Cancer Imaging Archive (TCIA) including different datasets for various cancer types is used in this study. the non-small cell lung cancer (NSCLC) dataset was used which included 22 patients' CT images [16]. The dataset is comprised of manual and semi-automatic segmentation, where the manual segmentation was performed by 5 different radiation oncologists (3 experienced and 2 trainer oncologists). These 5 people also did the semi-automatic segmentation. In case of the need for any correction, they used in-house automatic segmentation tools and checking segmentation.

One patient among these 22, does not have tumor delineations and another one just has manual delineations. So, they were eliminated from our study and the rest of the 20 patients' images were used to have the same number of images for both segmentations.

3.2. Features Definitions and Extraction

By radiomics, thousands of quantitative features could be extracted, that can describe lesions, and they are divided into four main categories of shape, first-order, second-order, and higher-order features [7] [3]. We can extract these features directly or after applying any type of filter. Here each category is defined in short:

- Shape: It is defined as the main features that describe the ROI size and shape for example maximum diameters, surface area, volume, etc.
- First-order Features: The first-order features are normally based on a histogram and recount the spread and position of each voxel value without regard to kurtosis, skewness, uniformity, or other spatial relationships.
- Second-order Features: The second-order features which are generally called texture features, explain neighboring voxels' inter-relationship and it is categorized into the following subgroups: Gray-Level Size Zone Matrix (GLSZM), Gray-Level Co-occurrence Matrix (GLCM), Neighbourhood Gray-Tone Difference Matrix (NGTDM), Gray-Level Run Length Matrix (GLRLM), and Gray- Level Dependence Matrix (GLDM). Each of these subgroups contains different features.
- Higher-order Features: After the implementation of any filter or mathematical transform, higher-order features are achieved. For instance, any filter or transform such as Fractal analysis, wavelet transform, and Laplacian can be applied for bolding the details or finding the repetitive and non-repetitive patterns.

  The application used for this study was a 3D slicer to extract the features, a wavelet was used also the whole features were 851. The purpose is to select all features and analyze them by using statistical metrics to obtain the most reliable and accurate features.

3.3. Impact of Image Segmentation

Finding the incorrect region of interest (ROI) might lead to poor results, as the features are extracted from this region [18], thus, Lesion delineation is one of the most significant challenges of radiomics. Most of the tumors do not show clear borders, so it is a

challenging task and even though delegating the responsibility to the experts can be helpful, as shown in Fig 1, the result of the borders which was shown by the three expert oncologists for the same lesion, were not the same.

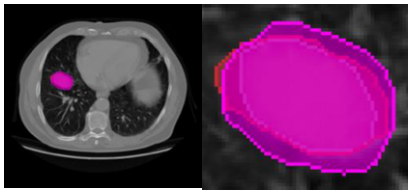

Figure 1. 3 Manual segmentation.

On the other hand, to save energy, and time and most importantly decrease the mistake level and improve the performance, automatic and semi-automatic segmentation techniques were developed by improving technology. Thus, in the present paper, one of our experiments was to compare manual and semi-automatic segmentation to see whether the semi-automatic one is more accurate or not.

3.4. Evaluation Metrics and Statistical Analysis

As mentioned earlier, radiomics enables us to extract thousands of features but that does not guarantee the usefulness of all of them. Therefore, one feels the necessity of finding the most reproducible and reliable ones. To quantify the reproducibility of the features, Statistical metrics are implemented. Based on the type of data and the purpose, different metrics are available but in the present paper ICC and Kruskal-Wallis tests were used. Kruskal-Wallis Test was applied to see whether there is a statistically significant difference in 3 or more independent groups' medians. Kruskal-Wallis Test is a non-parametric version of the one-way ANOVA. Compared to the one-way ANOVA, in The Kruskal-Wallis test normality is not assumed in the data and it is not much sensitive to outliers, thus typically, if the normality assumption gets violated Kruskal-Wallis Test is used[2], which can be defined as follows:

$$H = (N - 1)\frac{\sum_{i=1}^{g} n_i (\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} (\bar{r}_{ij} - \bar{r})^2}$$

4. Results

4.1. Inter-rater reliability

851 features are extracted from our dataset by applying wavelet and they were divided into 9 groups (original, HHH, HHL, HLH, HLL, LHH, LHL, LLH, and LLL). Each group comprises six feature categories (first order, NGTDM, GLDM, GLCM, GLSZM, and GLRLM). In the original group, there is one more category of shape included.

[2] Zach, "Kruskal-Wallis Test: Definition, Formula, and Example," 2019. https://www.statology.org/kruskal-wallis-test/.

All features have five semi-automatic segmentation and 5 manual segmentations. To find out how many features are not affected by changing the interobserver, the first step was calculating Kruskal for both segmentations. The result is shown in Fig. 2.

Figure2. Comparison of semi-automatic and manual results.

As represented in Fig.2, the inter-rater reliability degree in semi-automatic segmentation and that of the manual segmentation is the same for some categories, but it is shown more in other categories.

The second step was comparing all categories for all groups to find out the group and category which has a more accurate interobserver degree. According to the attained results, NGTDM and first order are the ones with the most reliable features and different semi-automatic segmentation did not affect their features. Only the semi-automatic

segmentation has affected GLRLM's features in the original image type, but there is no effect after applying the filter.

As a second experiment, the ICC is calculated for the expressed features to check which of the segments has the most agreement. Table 1 shows in each group how many percent the manual segmentation showed better results than the semi-automatic segment.

Table 1. Manual segmentation's result (the percentage which manual has better results than semi-automatic)

	Original	HHH	HHL	HLH	LHH	LHL	HLL	LLH	LLL
First-order	5.5	25	28	0	22	16	0	11	18
GLDM	21	20	8	25	4	25	0	8	25
GLCM	25	57	14	21	14	21	21	14	21
GLRLM	18	18	12	25	12	0	37	18	12
GLSZM	6	25	25	12	31	25	37	31	25
NGTDM	20	40	0	40	0	40	0	20	0
SHAPE	0	-	-	-	-	-	-	-	-

Figure 3. Comparison of semi-automatic and manual segmentations' results (3 observers)

4.2. Inter-rater reliability

To calculate the Kruskal in the first experiment, we worked on the 10 segmentations (5 semi-automatic and 5 manual). In the second experiment, we plan on taking out the information of the two trainer oncologists from the calculations to see whether the results would be affected or not. As shown in Fig 3, like the first experiment, in this experiment again semi-automatic segmentation represents better results than manual segmentation. There are some cases where both have the same results but there are no features where manual segmentation shows better results.

5. Conclusion

In this paper, three objectives were followed. First, compare manual and semi-automatic segmentation to find the more accurate one. The result of the experiment proved that semi-automatic is better than manual segmentation. Second, the comparison was repeated, but this time two trainer oncologists were eliminated to see whether it could improve the manual segmentation or not. This semi-automatic time segmentation shows better performance, too. The feature categories were compared to determine which one had more reliable features. The results show that first order and NGTDM have the best result with 100%. One of the limitations we faced in this study is the implemented statistical metric. Kruskal is just applicable to accept or reject the null hypothesis, but it cannot show the degree of agreement between the observers. In future studies, we will work on this issue.

References

[1] Liu R, Elhalawani H, Radwan MA, Elgohari B, Court L, Zhu H, Fuller CD., Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer, *Clin. Transl. Radiat. Oncol.*, vol. 21, pp. 11–18, 2020, doi: 10.1016/j.ctro.2019.11.005.

[2] Mondal SB, O'Brien CM, Bishop K, Fields RC, Margenthaler JA, Achilefu S. Repurposing molecular imaging and sensing for cancer image-guided surgery, *J. Nucl. Med.*, vol. 61, no. 8, pp. 1113–1122, 2020, doi: 10.2967/jnumed.118.220426.

[3] Jie T, Di D, Zhenyu L, Jingwei W. Radiomics and Its Clinical Application: Artificial Intelligence and Medical Big Data. Elsevier Science, 2021.

[4] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data, *Radiology*, vol. 278, no. 2, pp. 563–577, 2016, doi: 10.1148/radiol.2015151169.

[5] Yang F, Dogan N, Stoyanova R, Ford JC. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: A simulation study utilizing ground truth, *Phys. Medica*, vol. 50, no. December 2017, pp. 26–36, 2018, doi: 10.1016/j.ejmp.2018.05.017.

[6] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, Sanduleanu S, Larue RTHM, Even AJG, Jochems A, Van Wijk Y, Woodruff H, Van Soest J, Lustberg T, Roelofs E, Van Elmpt W, Dekker A, Mottaghy FM, Wildberger JE, Walsh S. Radiomics: The bridge between medical imaging and personalized medicine, *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, 2017, doi: 10.1038/nrclinonc.2017.141.

[7] Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, Bellomi M .Radiomics: the facts and the challenges of image analysis, Eur. Radiol. Exp., vol. 2, no. 1, 2018, doi: 10.1186/s41747-018-0068-z.

[8] Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research, J. Chiropr. Med., vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.

[9] Baeßler B, Weiss K, Santos DPD. "Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study," Invest. Radiol., vol. 54, no. 4, pp. 221–228, 2019, doi: 10.1097/RLI.0000000000000530.

[10] Lee, J, Steinmann A, Ding Y, Lee H, Owens C, Wang J, Yang J, Followill D, Ger R, MacKin D, Court LE. Radiomics feature robustness as measured using an MRI phantom, Sci. Rep., vol. 11, no. 1, pp. 1–15, 2021, doi: 10.1038/s41598-021-83593-3.

[11] Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, Löck S. Assessing robustness of radiomic features by image perturbation, Sci. Rep., vol. 9, no. 1, pp. 1–11, 2019, doi: 10.1038/s41598-018-36938-4.

[12] Fiset S,Welch ML,Weiss J,Pintilie M, Conway JL., Milosevic M, Fyles A, Traverso A, Jaffray D, Metser U, Xie J, Han K. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer, Radiother. Oncol., vol. 135, pp. 107–114, 2019, doi: 10.1016/j.radonc.2019.03.001.

[13] Mehta S, Bastero-Caballero RF, Sun Y, Zhu R, Murphy DK, Hardas B, Koch G. Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies, Stat. Med., vol. 37, no. 18, pp. 2734–2752, 2018, doi: 10.1002/sim.7679.

[14] Pleil JD, Wallace MAG, Stiegel MA, Funk WE. Human biomarker interpretation: the importance of intra-class correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates, *J. Toxicol. Environ. Heal. - Part B Crit. Rev.*, vol. 21, no. 3, pp. 161–180, 2018, doi: 10.1080/10937404.2018.1490128.

[15] Chen CC, Barnhart HX. Assessing agreement with intraclass correlation coefficient and concordance correlation coefficient for data with repeated measures, *Comput. Stat. Data Anal.*, vol. 60, no. 1, pp. 132–145, 2013, doi: 10.1016/j.csda.2012.11.004.

[16] Aerts HJWL, Wee L, Rios Velazquez E, Leijenaar RTH, Parmar C, Grossmann P, Lambin P. NSCLC-Radiomics[Dataset], 2019. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI (accessed Dec. 16, 2021).

[17] Li R, Xing L, Napel S, Rubin DL, Radiomics and radiogenomics: technical basis and clinical applications. 2019.