BHPA-symposium 2022

Scientific Committee

Milan Tomsej (chair), CHU Charleroi

Alain Seret, Université de Liège

Caro Franck, Universitair Ziekenhuis Antwerpen

Damien Dumont, UCLouvain

Kristof Baete, KU Leuven

Lara Struelens, SCK-CEN

Federica Zanca, Palindromo

Koen Tournel, Jessa Ziekenhuis

Nico Buls, UZ Brussel

Gert Van Gompel, UZ Brussel

Thierry Gevaert, UZ Brussel

Scientific Session

Topic: AI and Image processing

Chair: Thierry Gevaert (UZ Brussel)

Friday 29/04/2022 11h00-12h15

Auditorium 2000

Deep learning dose prediction as a tool to evaluate treatment complications based on NTCP models

Camille Draguet[1,2], Ana M. Barragán-Montero[1], Pieter Populaire[2,3], Gilles Defraene[2], Melissa Thomas[2,3], Karin Haustermans[2,3], John A. Lee[1], Edmond Sterpin[1,2]

[1] UCLouvain – Institut de Recherche Expérimentale et Clinique - Molecular Imaging Radiotherapy and Oncology (MIRO), Brussels, Belgium

[2] KU Leuven – Department of Oncology – Laboratory of Experimental Radiotherapy, Leuven, Belgium

[3] University Hospitals Leuven, Department of Radiation Oncology, 3000 Leuven, Belgium

ABSTRACT - In this study, we demonstrated the ability of our dose prediction models for IMRT and PBS to be used in combination with a pulmonary NTCP model in order to redirect patients towards PBS when the risk of developing a pulmonary complication in a population of esophageal cancer patients is above 10%.

KEY WORDS – Esophageal cancer, AI, clinical decision tool, model-based approach, NTCP model

Introduction

Dose prediction models based on deep learning (DL) algorithms become a common tool for automatic radiotherapy treatment planning. The combination of these algorithms with normal tissue complication probability (NTCP) models has the potential to automatically detect patients at high risk of toxicity, and thus, to trigger the evaluation of alternative treatments to reduce such complications, without the need of spending countless hours planning treatments manually. This work aims to evaluate the accuracy of our DL dose prediction models for intensity-modulated radiation therapy (IMRT) and for pencil beam scanning (PBS) treatments in combination with NTCP models to detect esophageal cancer patients at high risk of toxicity.

Materials and methods

Our DL models, two identical UNet architectures with dense connections, were trained with a database of 40 patients. For each modality, the DL model was trained and tested 4 times, by using a circulating test set of 10 patients out of 4 folds, in order to obtain a prediction of the dose distribution for each patient. The predicted and ground truth 3D dose distributions were used to extract relevant dose-volume metrics that were the input for a NTCP model. The NTCP model was used to evaluate postoperative pulmonary complications such as pneumonia, respiratory failure and respiratory distress syndrome [1]. The predicting variables in the NTCP model were the mean lung dose, age, histology type and body mass index.

The absolute errors for the NTCP model for the predicted and ground truth doses were compared for both modalities (IMRT and PBS). The accuracy of our DL models for patient referral is then evaluated based on ΔNTCP thresholds between IMRT and PBS plans. According to the Dutch Society for Radiotherapy and Oncology, a ΔNTCP ≥ 10 % for grade 2 complications is a necessary condition for redirecting a patient to proton therapy.

Results

A clinical decision for redirecting patients towards proton therapy should be made based on the results of the complication probability model. A comparison between the ΔNTCPs computed from the predicted plans and from the ground truth plans is shown in Figure 1. Our models succeed in predicting dose distributions that are close to the ground truth dose distributions. This can be observed in Figure 1. All patients are close to the ideal scenario (ΔNTCP ground truth = ΔNTCP predicted), highlighted with the gray line. As a result of this success, all patients are selected for the correct modality based on the predicted plans.

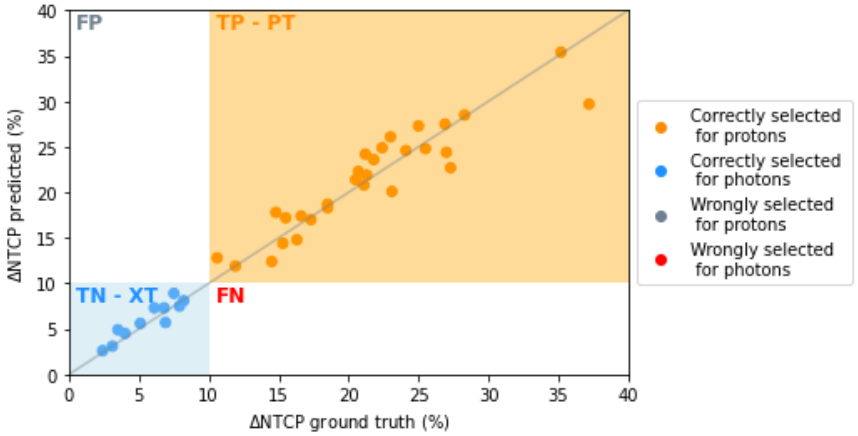


Figure 1: Comparison of the clinical decisions for each patient, based on the ground truth ΔNTCP and predicted ΔNTCP. The gray line represents the perfect scenario where the ground truth ΔNTCP and the predicted ΔNTCP are always equal. When a patient is in the orange zone, the criteria for PBS is met both with the ground truth plans and the predicted plans. When a patient is in the blue zone, the criteria for PBS is not met both with the ground truth plans and the predicted plans and the patient should go to IMRT.

Conclusion

This study evaluates our DL dose prediction models in a broader patient referral context and demonstrates their ability to be used for clinical decisions. Future research should focus on the detection of outliers to reduce the impact of dose prediction error on the clinical decision for those patients.

References

[1] Thomas, M., Defraene, et al. (2019). NTCP model for postoperative complications and one-year mortality after trimodality treatment in oesophageal cancer. Radiotherapy and Oncology, 141, 33-40. https://doi.org/10.1016/j.radonc.2019.09.015

Treatment plan prediction using convolution neural networks for breast VMAT

L. Vandewinckele[1], T. Reynders[1], S. Petillion[1], C. Weltens[1,2], F. Maes[3,4], W. Crijns[1, 2]
[1] *Department of Oncology, Laboratory of Experimental Radiotherapy, KU Leuven*
[2] *Department of Radiation Oncology, UZ Leuven*
[3] *Department ESAT/PSI, KU Leuven*
[4] *Medical Imaging Research Center, UZ Leuven*

ABSTRACT – This work investigates two deep learning networks that predict the VMAT apertures and MU directly for a three arc VMAT breast plan. The two neural networks have a Unet architecture, but one of them is trained using a GAN structure while the other has a normal loss function. The GAN is found to predict the VMAT plan significantly better than the other network.

KEY WORDS – Radiotherapy, treatment planning, deep learning

Introduction

Adaptive radiotherapy requires the instant generation of a treatment plan adjusted to the anatomy at that moment. Current research in adaptive radiotherapy treatment planning is directed towards 3D dose predictions from patient anatomy followed by an optimization process that translates the 3D dose predictions into a treatment plan. This optimization procedure is time consuming and diminishes the 3D information to 1D DVHs. This work investigates two deep learning networks that predict the VMAT apertures and MU directly for a three arc VMAT breast plan.

Materials and methods

To obtain a homogeneous data set, a dataset consisting of 135 patients (110 for training, 25 for testing) treated in our institution for right breast cancer was replanned for VMAT using our clinical RapidPlan model. A convolutional neural network, type Unet, predicts from CT and contour information, an aperture and MU per control point. The input consist of 2D projections of the CT and contours along the beams eye view. The network predicts for all control points an image of which the shape equals the MLC beam aperture and the intensity equals the MU-weight.

Two types of neural networks were trained. In the first, the Unet has as loss function the 'mean absolute error' between the true and the predicted outputs. In the second, the Unet is combined with a discriminator network that needs to learn to distinguish true and predicted outputs[1]. When the discriminator is able to distinguish them, the true and predicted outputs do not resemble each other enough and this pushes the Unet to learn. The Unet is here called the generator network and the combination is called a GAN (Generative Adversarial Neural Network). The output of the two networks is compared by calculating the mean absolute error (MAE) and gamma pass rate (GPR) between the predicted and true output fluence maps, obtained by summing the VMAT apertures.

Results

In Figure 1, an example of the ground truth and predicted output for one control point is shown. In Figure 2, an example is shown of the ground truth vs predicted fluence map for the two different networks. In Figure 3, boxplots are presented showing the difference of the metrics (MAE and GPR) between the GAN trained Unet and the no GAN trained Unet. The boxplots for the GAN lie better for the different metrics and the differences are statistically significant based on a paired Wilcoxon rang sum test and a significance level of 0.05 (p-values: MAE: 2.38×10^{-7}, GPR: 2.38×10^{-7}).

Figure 1: An example of the ground truth and predicted output of the no GAN and GAN network respectively for one of the control points (MLC aperture + MU) of a random patient.

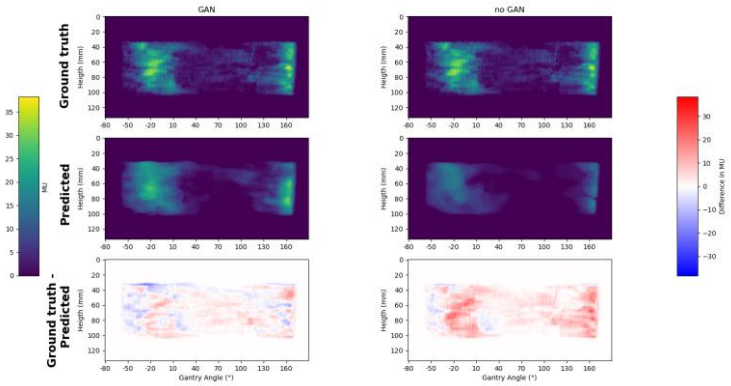


Figure 2: An example of the fluence maps of a random patient obtained by summing the ground truth and predicted MLC apertures and MUs.

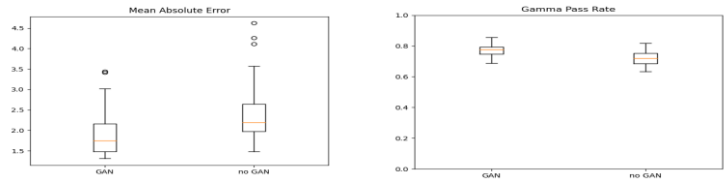


Figure 3: Boxplots of the mean absolute error, structural similarity index and the gamma pass rate over the patients in the test set between the fluence maps obtained by summing the ground truth and predicted MLC apertures and MUs.

Conclusion

The Unet with the GAN structure is statistically better in predicting the breast VMAT plan than the Unet trained with the 'mean absolute error' as loss function.

References

[1] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. Proc - 30th IEEE Conf Comput Vis Pattern Recognition.

Comparison of radiomic feature extraction results between the 2D lattice-based approach and the 1D approach depending on the BI-RADS density classification

Wagner T[1], Cockmartin L, Marshall NW[1,2], and Bosmans H[1,2]
[1]KU Leuven, Department of Imaging and Pathology, Division of Medical Physics & Quality Assessment, Herestraat 49, 3000 Leuven, Belgium
[2]UZ Leuven, Department of Radiology, Herestraat 49, 3000 Leuven, Belgium

ABSTRACT - Maximum 80 words

In the analysis of mammographic images, radiomic features are often used to quantitatively describe the parenchymal pattern of a breast. The extraction can be done in a 1D approach, yielding one feature value for the entire breast, or a 2D approach, yielding multiple feature values across the breast. This work focuses on comparing the results obtained from both approaches and analyzes different ways to interpret the result of the 2D approach for different BI-RADS density classifications.

KEY WORDS – Radiomics, BI-RADS density, processed images, mammography

Introduction

To objectively quantify mammographic images, radiomic features are used, giving properties of the breast like parenchymal pattern and the intensity distribution of the mammographic image a numerical value. For the extraction, a distinction is made between the 1D case, where the radiomic features are computed on the whole breast, and the 2D case, where the extraction is made on several smaller cells, yielding a lattice of radiomic results over the breast. In this work, several statistical evaluations of the 2D grid results are evaluated and compared to the 1D result for different BI-RADS density classifications.

Materials and methods

In this study, a dataset of 2109 patients from the UZ Leuven in Belgium are used. Due to difficulties of segmenting the pectoral muscle from the breast, only CC view images are taken into account.

The radiomic features will be extracted using the PyRadiomics[1] library, which offers both 1D and 2D extraction.

The lattice-based extraction will be conducted in a manner similar to the one mentioned by Zheng et al[2]. The breast is divided up into grid cells,

where in each of them the radiomic features are calculated, resulting in a spatial resolution of the parenchymal patterns of the breast.

Results

Preliminary results show indeed a strong local dependence of radiomic features across the breast. Depending on the given feature the variance in feature values can be significant, especially for heterogeneous breasts. The dependence of this variance on the BI-RADS density classification will be analyzed in the further process. Furthermore, different statictical approaches will be tested to compare the 2D results to the 1D results.

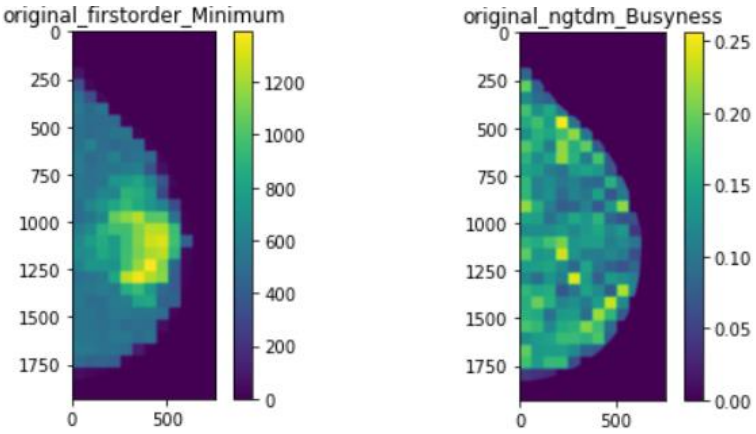


Figure 1: 2D map of extracted radiomic features for the original_firstorder_Minimum feature(left) and the original_ngtdm_Busyness feature(right).

Conclusion

Extracting radiomic features with a 2D lattice-based approach allows for a spatial resolution of radiomic feature values across the breast. Depending on the heterogeneity of the breast, the additional information gained by the 2D approach might be helpful in determining risk factors in breast cancer research.

References

[1] van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G.,
Fillion-Robin, J.-C., Pieper, S., and Aerts, H. J., “Computational radiomics system to decode the radiographic phenotype,” Cancer Research 77, e104–e107 (Nov 2017).
[2] Zheng, Y., Keller, B. M., Ray, S., Wang, Y., Conant, E. F., Gee, J. C., and Kontos, D., “Parenchymal
texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment:
Parenchymal texture analysis for breast cancer risk assessment,” Medical Physics 42, 4149–4160 (Jun 2015).

The evaluation of radiomic authenticity in mammographic background parenchymal patterns generated by VICTRE digital phantom

Wang YK [1], Cockmartin L[2], Marshall NW[1,2], Bosmans H[1,2]
[1]KU Leuven, Department of Imaging and Pathology, Division of Medical Physics & Quality Assessment, Herestraat 49, 3000 Leuven, Belgium
[2]UZ Leuven, Department of Radiology, Herestraat 49, 3000 Leuven, Belgium

ABSTRACT – Mammographic background parenchymal patterns are crucial to the breast cancer risk. To test the authenticity of the simulated mammograms generated by the VICTRE digital phantom, phantom- and patient-based mammograms are compared by the projection-invariant radiomic features.

KEY WORDS – breast cancer risk, parenchymal patterns, radiomics, digital phantom, virtual clinical trial

Introduction

Breast cancer risk prediction is a crucial step to achieve accurate patient stratification in the breast cancer screening. The focus of recent research has been on background parenchymal patterns in mammography. The quantitative imaging biomarkers extracted by radiomics have shown the ability to be a better risk factor than the previous established percentage breast density.

Under the restriction of ethical constraints, the virtual clinical trial with digital breast phantoms is a good compromise to test different clinical settings. However, the authenticity of the phantoms has to be investigated according to different clinical tasks.

In this study, the Graff digital breast phantoms from the VICTRE simulation platform was adopted to simulate lesion-free mammograms in order to focus on the background parenchymal patterns. The generated mammograms will be compared to a real patient population (lesion free, with BIRADS assessment 1) with radiomic features. Finally, the difference of mammographic background parenchymal patterns between phantom- and patient-based will be characterized by radiomics.

Materials and methods

The virtual imaging clinical trial for regulatory evaluation (VICTRE) dataset [1] archived on the TCIA database, was adopted in this study. This dataset contains realizations of the digital VICTRE phantom, generated for a DBT system that acquires 25 PVs over a total angular range of 50 degree. The simulated DBT system provides projection

views, reconstructed DBT, and the mammogram. In this study, only lesion free mammograms were selected in the archived dataset. A real patient cohort will also be selected with BIRADS assessment 1 to make sure the absence of lesions.

Projection-invariant radiomic features were selected by finding the reproducible features with intra-class correlation coefficient (ICC) across projection views with small projection angle perturbations. The selected robust features will be compared across real and simulated images.

Results

In Fig1, projection invariant features selected by projection views. In the final submission, those features will be compared across real and simulated datasets.

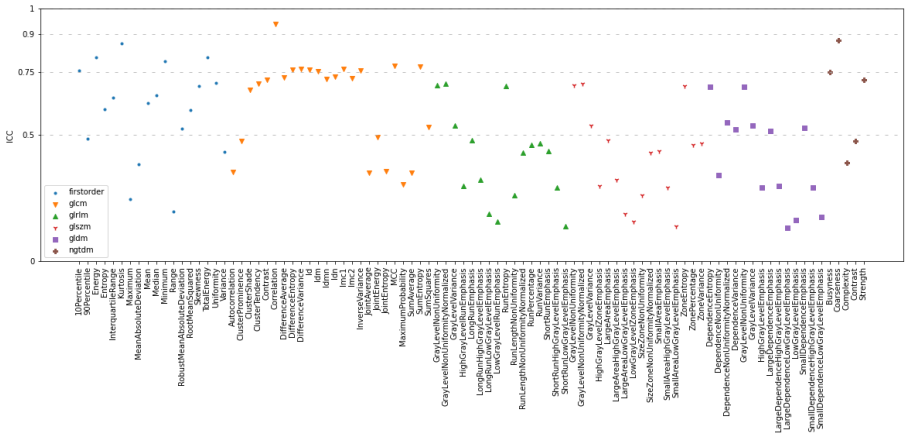


Figure 1: Reliability test presented by ICC across projection views with projection angle variation of 41.6 degree for 93 features. ICC interval from ICC 0 to 1 separated by grid lines indicate 4 levels of reliability: poor, moderate, good, and excellent.

Conclusion

The authenticity of the VICTRE digital phantom will be tested according to the proposed projection-invariant radiomic features with the focus on the mammographic background parenchymal patterns.

References

[1] Badano, A.; Graff, C. G.; Badal, A.; Sharma, D.; Zeng, R.; Samuelson, F. W.; Glick, S. J.; Myers, K. J. Evaluation of Digital Breast Tomosynthesis as Replacement of Full-Field Digital Mammography Using an In Silico Imaging Trial. JAMA Netw Open 2018, 1 (7), e185474. https://doi.org/10.1001/jamanetworkopen.2018.5474.

An in-depth comparison of IBSI-compliant radiomics software using digital phantom and patient data

Z. Paquier[1], S. Chao[1], A. Acquisto[2], C. Fenton[3], T. Guiot[1], J. Dhont[1], H. Levillain[1], A. Gulyban[1], M. Bali[2] , N. Reynaert[1]

[1] Medical Physics Dpt., Institut Jules Bordet – Université Libre de Bruxelles, Brussels, Belgium
[2] Radiology Dpt., Institut Jules Bordet – Université Libre de Bruxelles, Brussels, Belgium
[3] Radiology Dpt., Erasme – Université Libre de Bruxelles, Brussels, Belgium

ABSTRACT – Three IBSI-compliant radiomics software platforms were evaluated by comparing their respective feature values against the phantom-based IBSI reference, and among the platforms using two clinical image datasets (CT, MRI). We observed varying naming conventions between platforms, making direct comparison difficult. For common features, feature values showed disagreements depending on the imaging modality. These results indicate that reporting radiomics features could be different using various software and requires further harmonization.

KEY WORDS – Radiomics, Software, Comparative study, IBSI

Introduction

Radiomics has emerged as a promising clinical tool in medical imaging. However, the reproducibility of radiomics studies is cumbersome, up to the level of implementation in (commercially available) radiomics software. The lack of standardization of radiomics feature computation was aimed to be addressed by the Image Biomarker Standardisation Initiative (IBSI). This study evaluates IBSI-compliant radiomics software platforms, with a focus on both feature names and formulas, and output feature values.

Materials and methods

The three radiomics platforms evaluated in this comparison were RadiomiX Research Toolbox, LIFEx v7.0.0, and syngo.via Frontier Radiomics v1.2.5, which is based on PyRadiomics v2.1. Three different image data sets were used: a digital phantom provided by IBSI for benchmark, 27 contrast-enhanced computed tomography (CECT) of colorectal liver metastases, and 39 magnetic resonance imaging (MRI) of breast cancer, comprising two maps (D and f) from the intravoxel incoherent motion (IVIM) model and three dynamic contrast-enhanced MRI maps (K^{trans}, k_{ep}, IAUG). The comparison was divided into four phases: (1) matching the feature names based on the formulas; (2) evaluation based on IBSI reference values using the digital phantom; (3) software version and (4) inter-software comparisons using clinical data. The coefficient of variation was used to evaluate the agreement with IBSI reference, applying a classification of excellent ($CV \leq 1\%$), good ($1\% < CV \leq 5\%$), moderate ($5\% < CV \leq 10\%$) or poor ($CV > 10\%$). For version and inter-software comparisons, the lower boundary of the 95% confidence interval of the intraclass correlation coefficient (ICC) was used, with a stratification of poor (<0.5), moderate (0.5-0.75), good (0.75-0.9) and excellent (>0.9).

Results

The three radiomics platforms have 41 features in common (3 shape, 8 intensity, 30 texture) out of a total of 172 features for RadiomiX, 84 features for LIFEx and 110 features for syngo.via. The three platforms have also the option to perform wavelet filtering. The naming convention is, however, different between them. For example, one uses acronyms, while others provide full description for the same feature, while for others the formula is identical under completely different names.

Syngo.via achieved excellent agreement with the IBSI benchmark, while LIFEx and RadiomiX showed slightly worse agreement (Figure 1).

Software platform version affected feature reliability, heavily depending on the imaging type in the case of PyRadiomics (v.3.0.1 and v2.1.0, worse for MRI compared to CECT).

Excellent reproducibility between the three platforms was achieved for shape features only. Intensity and texture feature results varied considerably across the image types (Figure 2). Wavelet features produced the largest discrepancies between the software platforms, with the percentage of features with poor agreement at 69% for Kep, 77% for Ktrans, 78% for CECT, 84% for IAUG, 93% for IVIM-f and 94% for IVIM-D.

Figure 1: Number of features with excellent (green), good (yellow), moderate (orange) or poor (red) agreement with IBSI reference values for each radiomics software platform. Abbreviation: CV = Coefficient of variation.

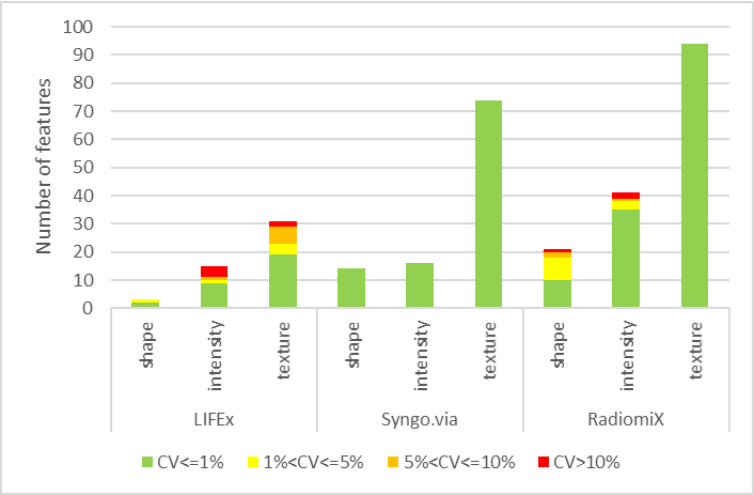
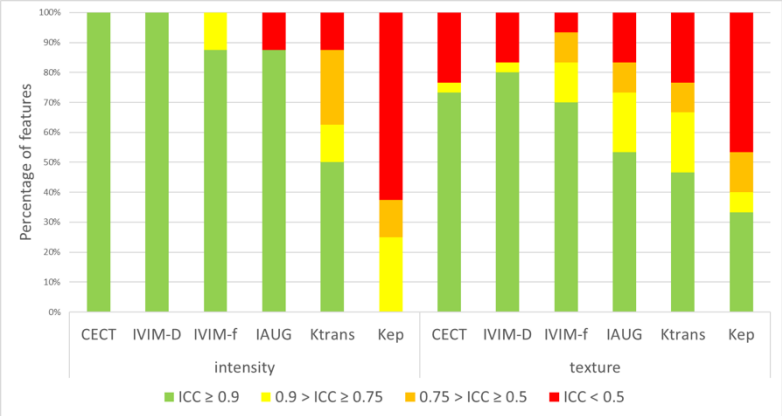


Figure 2: Number of intensity and texture features of excellent (green), good (yellow), moderate (orange) and poor (red) agreement between RadiomiX, LIFEx and syngo.via. Abbreviation: ICC = Intraclass correlation coefficient.



Conclusion

The results in this study show that even with IBSI-compliant software, the reproducibility of features between radiomics platforms is not guaranteed and the translation between platforms of models directly using these features should be performed with care. Furthermore, the results show commissioning of radiomics software should be repeated when updating to a new version or when adding a new imaging modality.

Automatic Segmentation of Kidneys in Computed Tomography for Total Kidney Volume Measurements

Koukoutegos K.[1,2,3], De Buck S.[2,3], De Keyzer F.[1,2,3], Kozari N. T.[3] , Maes F.[3], Bosmans H.[1,2,3]

[1]KU Leuven, Department of Imaging and Pathology, Division of Medical Physics & Quality Assessment, Herestraat 49, 3000 Leuven, Belgium
[2]UZ Leuven, Department of Radiology, Herestraat 49, 3000 Leuven, Belgium
[3]UZ Leuven, Medical Imaging Research Center, Herestraat 49, 3000 Leuven, Belgium

ABSTRACT

Kidney segmentation is an essential and time consuming task performed by clinical experts in order to evaluate renal functionality, diagnose the risk of renal diseases and improve treatment planning. Measuring the total kidney volume is a key part of this process as it provides a powerful prognostic biomarker that predicts adequately the risk of kidney insufficiency.

KEY WORDS

Kidney segmentation, Kidney volume, Deep Learning, Convolutional Neural Networks, Computer Vision

Introduction

Renal volumetry is the first task performed by radiologists in order to assess the risk of reduction in kidney functionality. Up to date, manual segmentations and the ellipsoid method are considered the two most widely used approaches to this problem [1]. The former one, though accurate, requires clinical expertise, it is subject to intra-observer variability and takes a lot of time as it consists of segmenting the kidneys slice-wise in a whole 3D CT imaging. The ellipsoid method on the other hand depends on the calculation of 3 orthogonal axes of each kidney, thus making it faster but not quite accurate. Deep Learning has proven, over the

last decade, that given a sufficient amount of data and big enough parameter space, it can find meaningful patterns and provide a standardized way of separating the data. To this end, we propose a segmentation framework which depends on Convolutional Neural Networks, the dominant architecture for Computer Vision tasks. The network segments kidneys of CT images in a fully automated manner, achieving near-expert accuracy and requiring only a limited amount of time and resources.

Materials and methods

Dataset

Our database consists of 89 pre-transplant cases from UZ Leuven with a big range of different kidney volumes. Computed Tomography images were chosen as those are more common for volumetry purposes. Using 3D Slicer [2], we manually segmented all cases following the guidelines of clinical experts. From the total of 89 CTs we use 62 for training, 14 for validation and the remaining 13 for testing our model.

Model training

The UNet architecture [3], is the most widely adopted one during the last years for medical image segmentation tasks. Our model is based on UNet but includes some minor modifications as described by [4], which make it more accurate and easier to train. Within our set of experiments we have used Adam [5] and AdamW [6] optimization methods which both converge in quite similar results, though in different number of epochs. Our learning rate follows the Cosine Annealing with warm Restarts scheduling method [7], which helps to explore better the loss curve and avoids getting stuck in sup-optimal regions.

Results

We left our network training for an undefined number of iterations, in order to give it the sufficient amount of time it needs to find an optimal solution to the segmentation problem. The network reached an average Dice of 0.8903 on the validation set by epoch 352 but was able to improve this even further to 0.8953 at step 412. This validates that our learning rate scheduling does actually help in exploring the loss curve more accurately. We also make use of data augmentation techniques such as random flipping of axes, shifting pixel intensity values and adding Gaussian noise. Using our model's output, we can rephrase the problem of measuring the kidney volume as counting the number of predicted voxels. In Figure 1, you can see 2 coronal slices with the red contour showing our manual segmentations and the black one showing the model prediction.