

Title

Automatic classification of benign and malignant kidney masses using radiomics. A retrospective study exploiting the KiTS19 dataset.

Authors information

Gianmarco Santini¹, Yvon Nzoughet Obame¹, Constance Fourcade¹, Noémie Moreau¹, Mathieu Rubeaux¹

¹ Keosys Medical Imaging, Nantes, France

Abstract for technical review (250 words)

Objectives: The study aims to assess the performance in differentiating benign from malignant kidney masses using a radiomics approach.

Methods: For this retrospective study we worked with the scans of 210 patients from the publicly available KiTS19 dataset. Each scan had segmentations of the healthy kidney tissue, benign lesions and malignant tumors. In Phase 1 of our study, we reduced the number of radiomic features (105) extracted from the scans by using four feature selection and ranking algorithms: recursive feature elimination (RFE), fisher score, partial least square discriminant analysis (PLS-DA) and linear support vector machine (l-SVM). The features selected by each method were then used to train a series of random forest (RF) classifiers.

In Phase 2, we trained a convolutional neural network (CNN) to automatically perform the segmentation of benign and malignant kidney masses. We then placed the best performing RF classifier from Phase 1 in series with the CNN to see if it corrected its prediction.

Results: The best classification performance was obtained when training a RF classifier with the eight features selected by the RFE method (accuracy: 0.974). This RF model applied to the segmentations derived from the neural network improved the CNN's overall results: the dice score for malignant mass went from 0.74 to 0.79 and dice score for benign mass from 0.55 to 0.80.

Conclusion: The studied radiomics approach proves to be an accurate solution to classify benign and malignant kidney masses. A deep learning algorithm has shown to also benefit from its predictive power.

Description of purpose

With an increasing use of sophisticated medical imaging techniques in the clinical and diagnostic fields incidental renal masses are regularly detected. However, differentiating benign from malignant kidney masses is not always easy, especially with cysts, as specified by the Bosniak classification [1] where for example a class III mass has a 5-10% chance of being cancerous. Nevertheless, a clear distinction between benign and malignant masses in kidneys is needed to avoid unnecessary surgical procedures.

Recently, the MICCAI KiTS19 challenge [2] was launched to find the best solution to segment kidney and malignant tumors. A series of deep learning solutions was found [3], but as required by the challenge the analysis focused only on detecting malignant masses and not on benign ones.

In this paper we want to prove the advantage of using radiomics information to improve the distinction between benign masses, like cysts, and malignant tumor in kidneys. We also verify whether the segmentation masks produced by a deep-learning-based model can benefit from the radiomics approach and doesn't affect negatively its prediction.

Method

The work is organized in two main phases as described in Figure 1.

In the Phase 1 a series of features are extracted from a training set. A feature selection operation then follows and was performed by using different feature selection and ranking algorithms. The features chosen by each algorithm were then used to train a Random Forest classifier (RF) which is used to predict on the test set whether a segmented mass is benign or malignant.

In Phase 2, we retrained a Convolutional Neural Network (CNN) used for the segmentation of kidneys and kidney tumors adding to it the segmentation of benign kidney masses. The deep learning model was then validated on the same test set as the one used in Phase 1. We corrected any misclassified masses from the CNN model with the best performing RF classifier obtained in Phase 1.

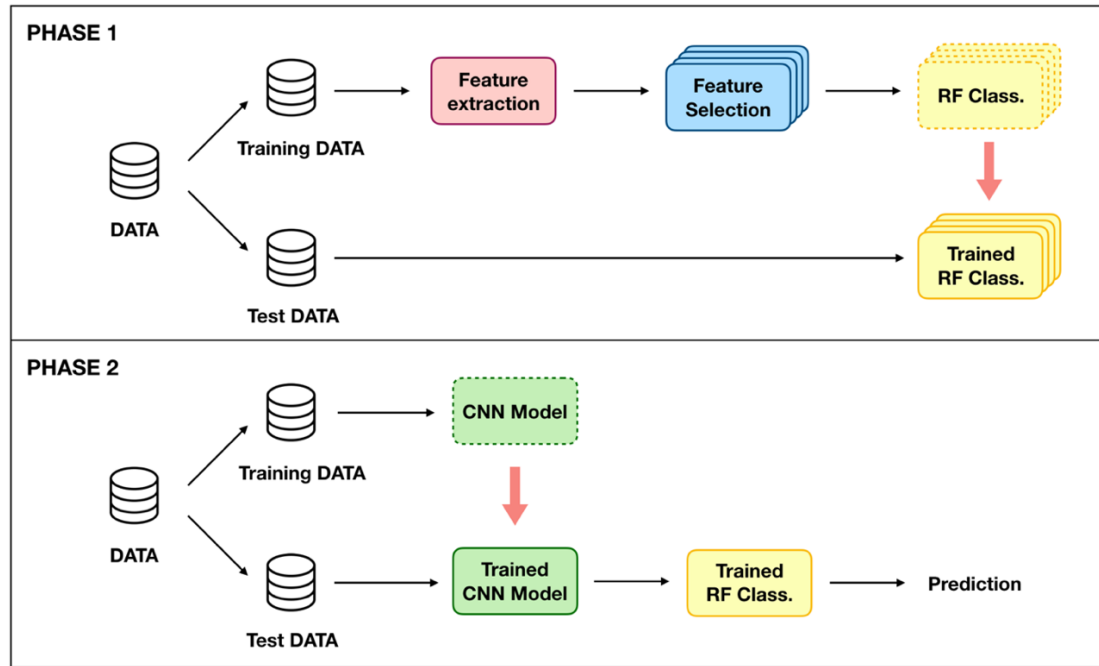


Figure 1. Two phases study. In phase 1 features extracted are reduced in number by a series of selection algorithms (represented by multiple blue boxes). The sub-set of features selected by each algorithm is used to train a separate Random Forest classifier (RF) (multiple yellow dashed boxes) and to evaluate its performance on the test set. The best trained RF classifier is used in phase 2 to correct if needed the prediction performed by the deep learning model designed to automatically segment kidneys, malignant and benign tumors.

For this study, we used the publicly available KiTS19 database [2]. It is composed of 300 cases of patients who underwent partial or radical nephrectomy collected between 2010 and 2018 by the University of Minnesota. We only worked with the 210 cases where kidney and tumor segmentation masks were available together with contrast enhanced CT images. To these cases, we manually added benign mass segmentation masks using our certified software [4]. Final segmentation masks were reviewed by an expert doctor.

Phase 1

Feature extraction: we randomly split the 210 labeled cases available into training ($n=180$) and test ($n=30$) sets. We applied a feature extraction operation on the training data by using the Pyradiomics module [5]. 105 features were extracted categorized according to their type as follows: 18 first order features, 14 shape features, 22 grey level cooccurrence matrix features (glcm), 16 gray level run length matrix features (glrlm), 16 gray level size zone matrix features (glszm), 14 gray level dependence matrix (gldm) and 5 neighbouring gray tone difference matrix features (ngtdm). We ensured that extracted features were also present in the IBSI list [6]. All the values extracted for each feature of interest were normalized to help the convergence of the subsequent algorithms.

Feature selection: to select the best set of features to train a classifier we used four different selection algorithms.

- *Recursive feature elimination (RFE)*: a wrapper-based method that recursively reduces the set of features considered. By training a designed estimator the feature importance is obtained and the less important features are pruned. The process continues until a specified minimum number of features is reached. In our case we used Logistic Regression as estimator and ten as minimum number of features to select.
- *Linear support vector classifier (l-SVM)*: it is used as embedded method to identify the most important features. With an L1-regularization the model is forced to produce a sparse solution. The non-zero coefficient assigned to the features employed are used as selection criterion to only keep the features of greater interest.
- *Partial least square discriminant analysis (PLS-DA)*: a variant of Partial Least Square regression (PLS) used in case of classification problems. Like Principal Component Analysis (PCA) it performs a dimensionality reduction and aims to find the linear boundary to separate space in two regions. In

concomitance with the Variable Importance on the Projection plot (VIP) it is possible to rank the most important features but also, by using a threshold, to select the most important features.

- *Fisher score*: a ranking feature method which assigns an importance score considering the ratio between the inter-class separation and the intra-class variance. The larger the score is, the more crucial the feature is. In this case we arbitrarily selected the first 15 top ranked features.

A 5-fold cross validation on the training set was performed on each method to select the most important features and a correlation analysis was also performed among them to further reduce their number (removal threshold > 0.90). Python scikit-learn module [7] was used to compute the feature selection algorithms. Table 1 summarizes the number of features selected and filtered according to each algorithm used.

Classification: Random forest classifier was chosen to create a machine learning model able to distinguish benign from malignant kidney masses.

For each set of selected features, we trained an independent RF model with a 5-fold cross validation.

The predictive power of each RF classifier was quantified with the test data by a series of metrics: precision, recall and F1-score were calculated for each classification label. The best RF classifier – to be used in Phase 2 - was chosen based on to the model's overall accuracy.

Phase 2

In the second part of the workflow we verified whether the best designated model from Phase 1 could also distinguish benign from malignant kidney masses if it was applied to automatic segmentations instead of manual ones as there can be slight difference between the two. Additionally, we tested the capability of the classifier to eventually correct previously misclassified masses.

For our CNN model, we used a solution we previously developed for the segmentation of kidneys and kidney tumors [8]. The same two stage model was retrained on the modified version of the dataset. i.e. segmentation masks of benign masses if present were added to the ground truth segmentation, separating them from the healthy kidney tissue. Compared to the previous implementation we modified the loss function by adding the dice loss contribution to the categorical cross entropy term as it shown to be more performant on multi-label segmentation cases with unbalanced class problem.

The training and test sets described in Phase 1 were used again here with the same sub-division to ensure that the RF classifier wasn't applied on cases previously used to train it.

We compared the segmentations produced by the deep learning model on its own to the ones produced by the same model with the added RF classifier using a dice similarity metric. We evaluated the dice score for the benign and malignant masses.

Results

Table 1 summarizes the results of the four RF classifiers trained in Phase 1 with different combination of the salient features.

The best performing model was the one trained with the features chosen by the RFE selection algorithm. The most representative features sorted according to their importance score in the RF model were: Root Mean Squared and 10 Percentile (first order), Run Entropy (glrlm), lcm1 and lcm2 (glcm), Max 2D diameter column (shape), Idn (glcm), Small area emphasis (glszm).

When comparing performances of the deep learning model on its own and of the model to which was added the best performing classifier, we observed a significant improvement in the distinction between benign and malignant kidney masses. The RF classifier corrected some inconsistent segmentation results as shown in Figure 2. With the RF model in series with the CNN the benign and malignant masses dice scores went from 0.55 ± 0.39 to 0.80 ± 0.13 and from 0.74 ± 0.23 to 0.79 ± 0.15 respectively.

Conclusion

With this study, we showed the advantage of using radiomics from contrast enhanced CT images to distinguish benign from malignant kidney masses. We also proved that a deep learning model can benefit from the prediction accuracy of a classification model trained with a set of key features.

Table 1. Results of the four Random Forest (RF) models trained with different combination of features selected by the four features selection algorithm. For each of them the number of features used and the overall accuracy are shown. The statistics on the single classes (malignant vs benign) are also reported.

Feature selection	N° features	RF model accuracy	Mass type	Precision	Recall	F1-score
PLS-DA	8	0.895	malignant	0.870	0.952	0.909
			benign	0.933	0.824	0.875
RFE	8	0.974	malignant	0.976	0.976	0.976
			benign	0.971	0.971	0.971
Fisher	9	0.934	malignant	0.930	0.952	0.941
			benign	0.939	0.919	0.925
l-SVM	7	0.908	malignant	0.872	0.976	0.921
			benign	0.966	0.824	0.889

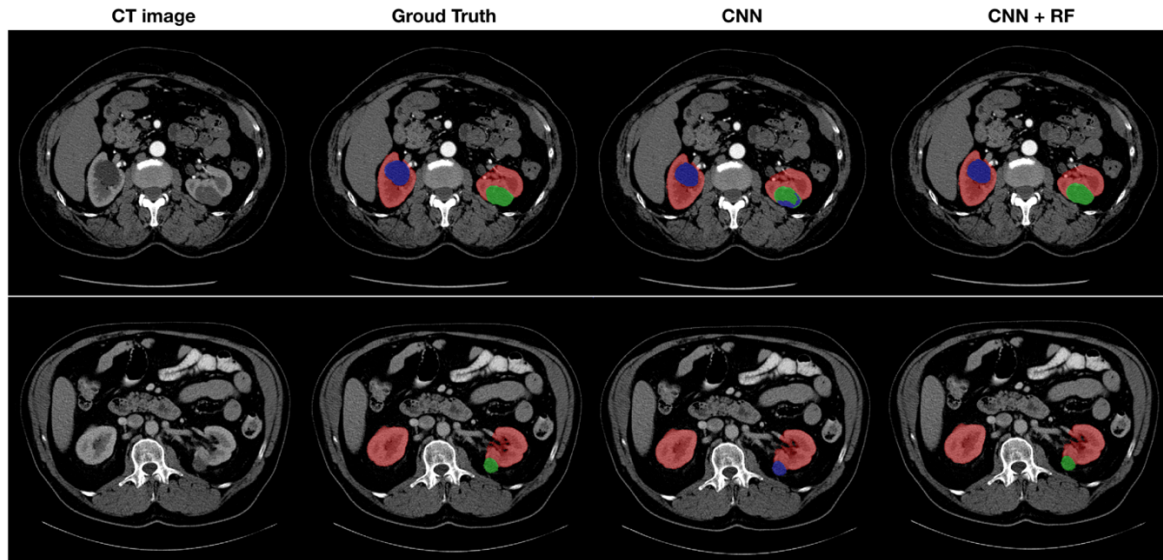


Figure 2. Examples showing the benefits of applying the Random Forest (RF) classifier after the Convolutional Neural Network (CNN). Healthy kidney tissue (red), malignant (green) and the benign masses (blue) best viewed in color.

References

1. Bosniak, M. A. (1986). The current radiological approach to renal cysts. *Radiology*, 158(1), 1-10.
2. Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., ... & Yao, G. (2019). The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 Challenge. *arXiv preprint arXiv:1912.01054*.
3. Heller, N., Sathianathan, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., ... & Dean, J. (2019). The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*.
4. <https://www.keosys.com/read-system/>
5. <https://pyradiomics.readthedocs.io/en/latest/>
6. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *arXiv preprint arXiv:1612.07003*.
7. <https://scikit-learn.org/stable/>
8. Santini, G., Moreau, N., & Rubeaux, M. (2019). Kidney tumor segmentation using an ensembling multi-stage deep learning approach. A contribution to the KiTS19 challenge. *arXiv preprint arXiv:1909.00735*.

Previous submission for publication

The present paper has not been submitted for publication or presentation before.