Journal of Digital Imaging (2018) 31:929–939
https://doi.org/10.1007/s10278-018-0100-0

A Decision-Support Tool for Renal Mass Classification

Gautam Kunapuli[1] · Bino A. Varghese[2] · Priya Ganapathy[1] · Bhushan Desai[2] · Steven Cen[2] · Manju Aron[3] · Inderbir Gill[4] · Vinay Duddalwar[2]

Published online: 6 July 2018

Abstract
We investigate the viability of statistical relational machine learning algorithms for the task of identifying malignancy of renal masses using radiomics-based imaging features. Features characterizing the texture, signal intensity, and other relevant metrics of the renal mass were extracted from multiphase contrast-enhanced computed tomography images. The recently developed formalism of relational functional gradient boosting (RFGB) was used to learn human-interpretable models for classification. Experimental results demonstrate that RFGB outperforms many standard machine learning approaches as well as the current diagnostic gold standard of visual qualification by radiologists.

Keywords Renal mass · Multiphase CT · Radiomics · Statistical relational learning · Clinical decision support

Background

The National Cancer Institute projects that at least $5.1 billion will be spent on kidney cancer care in the USA by 2020 [1], assuming survival rates follow current trends. Renal cancer is, clinically, a heterogeneous disease characterized by a wide spectrum of tumor behavior. Accurate characterization of *renal masses* is necessary to determine the most beneficial course of treatment. Conventional diagnosis is based on visual qualification, where it is often difficult to discriminate between malignant tumors such as renal cell carcinomas (RCC) and benign lesions such as oncocytomas and lipid-poor angiomyolipomas [2–6]. A substantial number of renal masses also turn out to be benign (15–20%) at surgery [7]. A major contributor for false positives and overdiagnosis is that many small renal mass lesions (particularly, T1 tumors less than 7 cm in size) are difficult to characterize with conventional imaging or even after a biopsy. Factors beyond tumor size and contrast enhancement, such as tumor texture and shape, have been associated with tumor behavior but are currently not commonly used in clinical decision-making. *Radiomics*, the high-throughput extraction of such quantifiable image features, is a natural next step [8, 9].

Recently, powerful machine-learning (ML) algorithms [10] are being used to explore complex interactions in clinical data to provide diagnosis, prognosis, and treatment planning. However, applications of these algorithms come at the cost of *loss of interpretability and explainability*, particularly when incorporated into routine clinical decision-support (CDS) tools [11–16]. *Interpretability* refers to the end user understanding the rationale behind classification of a test case by a machine-learning algorithm based on extracted feature values. *Explainability* refers to the CDS providing a text-based and/or image-based information explaining the rationale behind classification decision.

Commercially available decision-support systems are either simple rule-based systems or include mathematical/physiological models to assess patient risk towards a certain disease (see Table 1). To fully exploit the heterogeneity of all available patient data (unstructured doctor's notes, multimodal imaging, epigenetic data, medical and family histories, etc.), methods that can learn complex models, identify subtle relations in data, and provide explainability must be investigated.

✉ Gautam Kunapuli
gautam@utopiacompression.com

1 UtopiaCompression Corporation, 11150 W Olympic Blvd. Suite #820, Los Angeles, CA 90064, USA

2 Department of Radiology, Keck School of Medicine, University of Southern California, 1500 San Pablo Street, 2nd Floor, Los Angeles, CA 90033, USA

3 Department of Pathology, Keck School of Medicine, University of Southern California, 2011 Zonal Avenue, Los Angeles, CA 90033, USA

4 Institute of Urology, Keck School of Medicine, University of Southern California, 1441 Eastlake Ave, Los Angeles, CA 90089, USA

Table 1 A list of commercially available CDS tools. This list is not exhaustive

CDS tool	Features	Core algorithm
IndiGO (Archimedes)	Estimates risk for diabetes, stroke, heart attacks; personalized risk management strategy	Mathematical equations; physiology models
Auminence (Autonomy Health)	Diagnosis support with symptoms entered by clinician	Rule-based/decision trees
VisualDx (VisualDx)	Differential diagnosis, curated images to support diagnosis (emergency medicine, dermatology, primary care)	Proprietary
Watson (IBM)	Research tool; mines from patient history, EHR, doctor notes, research papers; commercial use: lung cancer treatment planning	Deep learning for text mining; natural language processing
RadWise (Sage Health)	Helps in determining the appropriate imaging tests to be orders based on physician inputs	Evidence-based reasoning
LI-RADS (University of Colorado)	Diagnosis/categorization of liver carcinomas from CECT images (not commercial)	Decision flow to mimic natural thought process of radiologists

Statistical relational learning (SRL) methods provide an attractive framework for modeling CDS problems in a data-rich environment.

SRL has been applied to several *relationship discovery problems* [17] and has been successful because it combines the inferential power of probabilistic graphical models with the expressiveness of *first-order logic* (FOL). Current ML methods are based on one of two approaches: logical models, which capture relationships between objects, and statistical models, which efficiently handle noise and uncertainty. In the last decade, a great deal of progress has been made in unifying both approaches leading to the framework of SRL [18]. In this paper, we model renal cancer decision-support using relational functional gradient boosting (RFGB; [19, 20]), an SRL method . RFGB has been previously applied to several text-mining-based medical informatics tasks including predicting myocardial infarction from electronic health records (EHRs) [21], extracting adverse drug events from a drug database [22], and identifying rare diseases from patient behavioral data [23]. RFGB was also recently applied to an imaging-based medical informatics task: Alzheimer's diagnosis from MRI [24], where it significantly outperformed conventional ML approaches (AUC-ROC of 0.77 vs. 0.62; see "Results" for a discussion of AUC-ROC as an evaluation metric). The current work will be one of the earliest, if not the first attempt, to employ RFGB as a decision-support tool for cancer diagnosis using radiomics-based imaging features.

RFGB uses FOL to represent the domain and data and learns decision tree-like models that capture relationships between features and are human interpretable. In addition to inferring the most-likely diagnosis, RFGB can also provide explanations in terms of tumor shape, size, and texture metrics as well as clinical, demographic, and other factors when they are available. Explainability can greatly enhance effectiveness of decision-support and clinician confidence. Considering that FOL allows for seamless fusion of data modalities (clinical, demographic) with radiomics features, the resulting CDS tool can be thought of as an evidence-based guide for clinicians and not simply an inscrutable black box [11–16].

We showcase our development of a robust ML-based radiomics pipeline for deriving diagnostically significant radiomics features and reliable tumor classification models. Our goals are twofold: (1) by exploiting the strengths of both the radiomic features and machine-learning classifiers, we aim to reduce overfitting and maximize data parsimony, that is, learn effective models with a small amount of data and (2) learn/identify potentially complex patterns from data, which can in turn be used for making actionable predictions/decisions. At this preliminary stage of our work, we focus on the *binary classification problem* of distinguishing between malignant and benign tumors. Future work will address cancer subtype classification, nuclear grade classification, and other tasks beyond diagnosis such as prognosis and treatment planning.

Methods

One hundred and fifty subjects from an IRB-approved, HIPAA-compliant study who had pre-operative, *multiphase contrast-enhanced computed tomography* (CECT) of the abdomen and pelvis and post-resected tumor pathology evaluation with expert graded histology were retrospectively selected. The post-resected tumor pathology evaluation was used as gold standard. All 150 renal masses are solid, enhancing lipid poor tumors, which is the group of masses that is responsible for the vast majority of diagnostic errors (see Table 2 for tumor distribution).

CT scans were performed on a 64-detector row helical CT scanner (Brilliance, Philips Healthcare, CT) using the following parameters: 120 kVp, variable tube current, and slice thickness of 0.5 mm with reconstruction interval of 2 mm. Images were obtained at four phases of the contrast-enhanced computed tomography. Pre-contrast CT of the abdomen was first obtained, followed by three post-contrast CT scans obtained in corticomedullary (30 s), nephrographic (90 s), and excretory (5–7 min) phases.

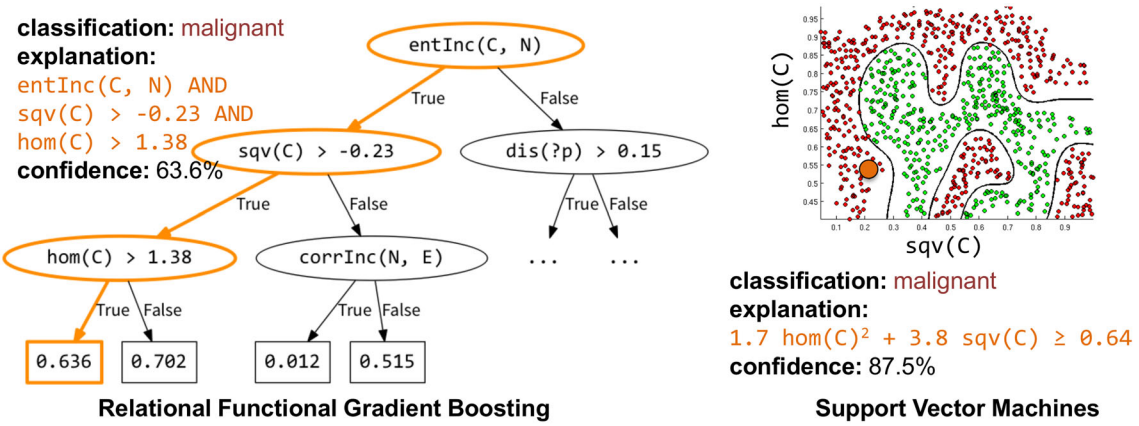

Fig. 1 RFGB learns tree models that can be easily explained (left), while SVM models (right) are far harder to interpret. RFGB's decision can be explained using intuitive comparisons and conjunctions (AND), while the SVM learns a non-linear function. Both models use radiomics features: SQV and HOM. CECT phases: corticomedullary (C) and nephrographic (N) are also captured. Illustrative example only

Tumor Segmentation

Tumors were manually segmented by an experienced radiologist in 3D using image-rendering software (Synapse 3D, Fujifilm, Stamford CT). Tumor margins were then sampled in the transverse, sagittal, and coronal dimensions. In general, the nephrographic phase provided the best delineation of the tumor; it was used as the reference template co-registering the other phases. The series of images were then co-registered with all other series using normalized mutual information cost function implemented in Statistical Parametric Mapping software package (Wellcome Trust, UK). Custom MATLAB (Mathworks, Natick MA) code was used to extract voxel data corresponding to the regions of interest (ROI). 2D CT-based texture analysis (CTTA) was performed on the largest tumor diameter within each imaging plane, and 3D CTTA was conducted on the entire tumor volume.

Radiomics-Based Feature Extraction

Texture characteristics ranging from first-order (mean, median, etc. of the intensity distribution), second-, and higher-order measures (including features such as contrast, homogeneity, entropy, etc.) were analyzed. We derived features using various techniques ranging from simple histogram analysis to more advanced techniques such as gray-level co-occurrence matrix (GLCM) and gray-level difference matrix (GLDM) [8, 9, 25, 26]. The techniques (summarized in Table 3) have been described in greater detail in previous studies [27–32].

1. *Histogram analysis*: We extract first-order statistical texture measures by assessing image intensity (gray-level distribution of an image), with no regard for the spatial location of the intensities. This includes features extracted from the histogram of the grayscale values making the tumor ROI [29] as well as histogram analysis of the 3D volume (eight features).
2. *2D and 3D GLCM and GLDM analysis*: We performed second-order statistical texture analysis, which included 2D and 3D-GLCM and GLDM analysis [28]. These analyses took into account both pixel intensities and inter-relationships, thereby providing spatial information (second order) of the intensities in various forms. For workflow implementation, the number of gray levels was reduced to 12-bit, which was determined to be sufficiently accurate for the study of texture. The co-occurrence and differences matrices were obtained in four directions (horizontal, vertical and two diagonals in the *x*-*y* plane) to capture directional texture information. Twenty different metrics were calculated: 13 based on the method by Haralick [31] and 7 additional metrics. In the 3D analysis, 5 additional directions in the *z* plane were added and the same 20 texture metrics were calculated (20 GLCM + 20 GLDM features).
3. *2D Fourier analysis*: A 512-point fast Fourier transform (FFT) was applied to all images using Matlab, after which we extracted individual frequencies, their amplitude (amount of individual frequencies), and phase (location of the frequency in the image). FFT metrics were assessed between 10 and 90% of the maximum frequency to avoid high- and low-frequency noise typical of medical images. The frequency boundary was set based on maximization of the signal-to-noise ratio [32] (three features).

Table 2 Tumor type distribution within the acquired CECT data set

Tumor type	No. of tumors	Class
Clear-cell renal cell carcinoma (ccRCC)	70	Malignant
Papillary renal cell carcinoma (pRCC)	20	100
Chromophobe renal cell carcinoma (chRCC)	10	
Lipid-poor angiomyolipoma (lpAML)	20	Benign
Renal oncocytoma	30	50

Table 3 The feature extraction stage of the pipeline uses the radiomics CT-based texture analysis (CTTA) panel to extract a large number of features, including first-order, second-order, and FFT features. Specifically, 51 features are extracted per phase. Recall that we have a four-phase CECT image for each subject; feature extraction produces a total of 51 × 4 = 204 features for each subject

Extracted CTTA features from multiphase CECT images of renal masses	
Gray-level histogram (3D) analysis (8 features/phase)	Mean, median, skewness, kurtosis, minimum, maximum, quartile range, standard deviation
Gray-level co-occurrence matrix (GLCM 2D/3D) analysis (20 features/phase) Gray-level difference matrix (GLDM 2D/3D) analysis (20 features/phase)	Angular second moment, contrast, correlation, dissimilarity, entropy, homogeneity, inverse difference moment mean, information measures of correlation 1 mean, information measures of correlation 2 mean, maximum correlation coefficient, root mean square, standard deviation, uniformity, variance, sum of average, sum of entropy, sum of variance, mean
Fast Fourier transform (FFT 2D) analysis (3 features/phase)	Entropy of FFT magnitude, entropy of FFT phase, complexity index

From *each phase of each subject's* multiphase CECT image, we extract 51 2D and 3D radiomics-based texture features. Recall that we obtain a four-phase CECT image from each subject, which ultimately results in 51 × 4 = 204 total features per subject.

Feature Selection

The feature extraction stage produces a large number (204) of radiomics texture features that capture various characteristics of the CECT image across four phases. While it may be beneficial to develop machine-learning models that consider all the available features, we are also interested in developing models that are *efficient and explainable* for the purposes of incorporation into a clinical decision-support system. An *efficient model* should provide a diagnosis within a reasonable amount of time. An *explainable model* should be able to explain its decision and reasoning in terms of the data features to a non-machine learning domain expert (in this case, the urologist). These two critical requirements motivate us to explore feature selection as a means of identifying the most informative features for this classification/diagnosis task.

We use a well-known technique called *Recursive Feature Elimination* (RFE; [33]), which has been extensively applied for feature selection from gene-expression DNA microarray data. Our pipeline, however, uses RFE on CECT-based radiomics features. RFE uses support vector machines (SVMs; [34]) to recursively rank and filter features with the least diagnostic value. At a high level, RFE ranks features by empirically measuring the degradation in predictive accuracy of a model if that feature was dropped. This procedure is repeated until a sufficient number of features have been recursively identified and dropped as required.

Since the data set contains 204 features over four phases, we instead consider a variant called block RFE, which drops groups of features at a time to improve efficiency of feature selection. In our case, we score the predictive accuracy of *each radiomics texture across all phases taken together* and drop textures with low diagnostic value. That is, when a texture is dropped, it is dropped across all phases. For instance, at iteration *i*, if RFE scores the texture **`DifEnt`** (GLDM entropy) the lowest, then the entire block of features **`DifEnt-PreCon`**, **`DifEnt-Cort`**, **`DifEnt-Neph`** and **`DifEnt-Excr`** (**`DifEnt`** features from all four CECT phases are dropped for all subjects). We refer to this step as recursive texture elimination; our pipeline currently uses linear SVMs for recursive texture elimination. We selected 10 radiomics texture features per phase, which produces 40 features (over the four CECT phases) per subject, that is, per training sample (see Table 4).

Machine Learning Methods

The learning task is to build a classifier to distinguish between malignant and benign tumors. We consider the following classes of methods:

1. *Conventional machine learning methods* can be categorized into two types: generative (that model the actual distribution of each class) and discriminative (that model the decision boundary between classes). The simplest form of a generative classifier is the Naïve Bayes classifier [35], Chap. 3.5, which assumes independence between features (radiomics textures) and applies Bayes theorem to predict the correct class. In the case of discriminative classifiers, such as logistic regression (LR; [36]), kernel machines [37] such as SVMs [34], and decision trees [38], single or multiple decision boundaries are learned either on the original features or by transforming the feature into higher dimensions.
2. *Conventional ensemble methods* combine the decision output of multiple weakly trained classifiers (called base classifiers) to improve generalizability and robustness in estimation. In boosting [39], several weak classifiers are learned sequentially, where each subsequent weak learner is trained explicitly on samples missed by the previous ones. In bagging (also known as bootstrap aggregating)

Table 4 The feature selection stage of the pipeline uses Recursive Feature Elimination to select 10 discriminative textures, which were used for classification of renal masses. Recall that we have a four-phase CECT image for each subject; feature selection produces a total of $10 \times 4 = 40$ features for each subject

Feature	Description
ASM	Angular second moment/sum of squared elements in the GLCM
CON	Contrast/local intensity variations between a pixel and neighbors in the GLCM
COR	Correlation of gray-level linear dependence between pixels relative to each other
DIS	Dissimilarity/difference average in the GLDM
ENT	Entropy/amount of image information needed for image compression
HOM	Homogeneity/closeness of the distribution of GLCM elements to the diagonal
IDM	Inverse difference moment/measure local homogeneity in the GLCM
SQV	Square root of variance/measure of how much elements differ from the mean
SUMAVE	Sum of averages
VAR	Variance/measure of heterogeneity

[40], base classifiers are trained independent of each other and their decisions are combined using voting or weighting schemes. Random forests [41] further extend bagging by using a random subset of features at each iteration. Ensemble methods have been highly successful over a diverse variety of tasks; as we show below, the extension of such ensemble methods with richer representations can yield powerful and explainable models.

3. *Deep learning*: Artificial neural networks (ANNs) have seen renewed interest with the emergence of deep learning [42]. Deep learning has been successful for image segmentation and classification [43] and is now being applied to medical imaging [44]. Deep learning is often applied to images directly, and successive hidden layers realize various image-processing steps. For instance, the first one to three layers of a deep network behave like Gabor filters, performing edge detection and are not robust classifiers [45]. In our setting, the radiomics features, rather than raw CECT images, are used as the features. A significant drawback of deep learning is a large amount of labeled training data that is required in order to build an effective model.

4. *Statistical relational learning*: In the last decade, considerable progress has been made in SRL, which combines statistical methods (that model uncertainty) with a relational representation (to provide a richer, more natural representation of data) [17]. Numerous SRL approaches such as Markov logic networks [46] have been successful. However, many such methods require domain rules to be specified by the user or learned sequentially via *structure learning* methods before model (parameter learning). We, instead, focus on a state-of-the-art SRL approach called relational functional gradient boosting (RFGB) that can learn the structure and parameters *simultaneously*, *efficiently*, and produces *explainable models*.

Relational Functional Gradient Boosting

This approach is motivated by the intuition that finding many simple and rough rules-of-thumb to model probabilistic feature interactions *locally* can be much easier than finding a single, large, highly accurate model. Specifically, this approach turns the problem of learning SRL models into *a series of relational function approximation problems using the ensemble method of gradient-based boosting*. This is achieved by the application of Friedman's [47] gradient boosting to SRL models, where each conditional probability distribution (that models the relationship between the variables) is represented as *a weighted sum of regression models* [48]. That is, instead of representing the relationship between the various variables (or radiomics features) as a single giant relational probability tree, we use a collection of smaller relational regression trees [49].

Relational Representation Recall that a key strength of SRL methods such as RFGB is the ability to represent complex relationships between objects and attributes in a domain using first-order logic (FOL). In this domain, phases and radiomics features are the primary attributes of the various entities, that is, patients. Thus, radiomics features can be easily represented using relational features for any patient (**`PID`**) and CECT phase:

```
radiomics_feature(PID, CECTphase, value).
```

For instance, the relational feature **hom(patient38, E, 0.74)** tells us that the homogeneity in patient38's image in the excretory (**E**) phase is 0.74.

To fully utilize the representational power of FOL, we also incorporated domain knowledge by capturing the ordering of the phases, that is pre-contrast (**P**) followed by corticomedullary (**C**), nephrographic (**N**), and excretory (**E**). This is done through a predicate **follows($Phase_1$, $Phase_2$)**, which is only true when $Phase_2$ follows $Phase_1$. For instance, **follows(N, E)** will be true, while **follows(E, N)** will not. There is no way to express this information in conventional ML algorithms without considerable feature engineering as their representation is *propositional* or tabular. In contrast, as RFGB uses a logical representation, it is *relational* and can capture complex feature interactions beyond non-linear functional interactions of conventional machine learning. Intuitively, a relational representation can be thought of as a table-of-tables in a database (RDBMS).

We can also specify additional domain rules that tell us if a certain radiomics feature is increasing or decreasing between two consecutive phases. In a propositional representation, as used in classical machine learning, we would have to explicitly construct such features by considering all exhaustive combinations of features and phase-pairs. However, using first-order logic, such *relational features* can be represented compactly. To see this, consider the texture feature entropy (**ent**), from which we can identify two relational features **entInc(PID, Ph_1, Ph_2)** (the entropy increases from Ph_1 to Ph_2) and **entDec(PID, Ph_1, Ph_2)** (the entropy decreases from Ph_1 to Ph_2) through simple and natural rules. The rule(s) "entropy is increasing (decreasing) from Ph_1 to Ph_2 if its value in phase 2 is greater (lesser) than its value in phase 1" can be easily written in first-order logic as follows:

$$\mathbf{ent(PID,Ph_1,v_1),\ ent(PID,Ph_2,v_2),\ follows(Ph_1,Ph_2),\ (v_2>v_1)\ \Rightarrow\ entInc(PID,Ph_1,Ph_2).}$$

$$\mathbf{ent(PID,Ph_1,v_1),\ ent(PID,Ph_2,v_2),\ follows(Ph_1,Ph_2),\ (v_1>v_2)\ \Rightarrow\ entDec(PID,Ph_1,Ph_2).}$$

This representation is beneficial for three reasons: (1) RFGB only adds features as needed, which means that exhaustive enumeration of all combinations is avoided; (2) the rules depend on the background information **follows(Ph_1, Ph_2)**, which means that only meaningful feature combinations will be considered during learning; and finally, (3) features such as **entInc** are intuitive and interpretable to clinicians, who are not machine-learning experts. The last point is of considerable importance as we are interested in developing *explainable diagnostic models*, and the use of features such as **entInc** leads to more natural explanations.

Relational Regression Trees RFGB uses relational regression trees (RRTs) to represent local relationships and interactions between features. A regression tree [38] contains conditions on features at each node, and as we descend down the tree, we reach a leaf node that provides the regression value for that example. An RRT [49] can be viewed similarly, except the nodes define *relationships between feature combinations* rather than feature thresholds. These represent the diagnostically relevant discovered relationships in the data. The advantage of RRTs is that they incorporate domain knowledge in addition to radiomic features using FOL to represent features compactly in an interpretable tree structure. Figure 1 (left) shows an example of an RRT learned during training to classify a tumor as malignant. When a new case is presented, it is classified into a leaf node: the classification path (highlighted in orange) is the explanation and the leaf node probability is the likelihood of malignancy.

Learning of RFGB Models Recall that the learning task is to build a classifier to distinguish between malignant and benign tumors (training labels, y) using training examples, $\mathbf{x}$. RFGB tries to fit a probabilistic model $P(y \mid \mathbf{x})$; when given a patient case $\mathbf{x}$, the model returns the probability of P (y = **malignant**), that the tumor is malignant. Thus, RFGB learns a joint distribution over the labeled examples (supervised learning). The high-level idea behind RFGB is to learn multiple RRTs in a stage-wise manner by optimizing the log-likelihood over all the examples, $\mathrm{LL} = \sum_{i=1}^{N} \log P(y_i|\boldsymbol{x_i})$.

The approach learns RRTs such that the probability of observed outcomes y_i is the maximized given data $\mathbf{x}_i$. Most methods choose a family of distributions (e.g., normal distribution), and parameters of the distribution are learned from data via gradient ascent. RFGB, on the other hand, is a non-parametric approach, which does not require a family of distributions to be chosen a priori. Instead, it models the distribution using a potential function, ψ. The goal is to fit a model $P(y|\boldsymbol{x}) \propto e^{\psi(y,\boldsymbol{x})}$, where ψ is a collection of RRTs, each learned during a single iteration of RFGB.

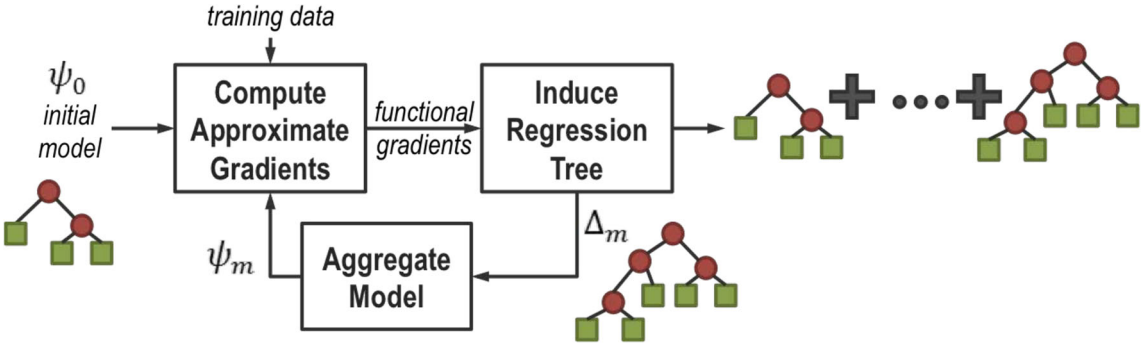

Fig. 2 RFGB visualized

At each step, RFGB learns a single RRT by computing the functional gradient, that is, by learning a tree that represents the gradient of the potential function ψ and adding features and relations only as needed. RFGB starts with an initial potential ψ_0 and iteratively adds gradients (represented as RRTs) Δ_m. After m iterations, the potential is given by $\psi_m = \psi_0 + \Delta_1 + \ldots + \Delta_m$, where Δ_m is an *RRT-based functional gradient*. This iterative learning procedure is visualized in Fig. 2. The functional gradient, while represented as tree, actually has the following form:

$$\frac{\partial \log P(y_i = \text{malignant} \mid \boldsymbol{x}_i)}{\partial \psi(y_i = \text{malignant} \mid \boldsymbol{x}_i)} = I(y_i = \text{malignant} \mid \boldsymbol{x}_i) - P(y_i = \text{malignant} \mid \boldsymbol{x}_i)$$

where I is the indicator function, that is 1, if y_i = **malignant**, and 0 if y_i = **benign**. This expression is simply *the adjustment required to match the predicted probability with the true label of the example*. If the example is malignant, and the predicted probability of malignancy is less than 1 (indicating that the model still has uncertainty about malignancy), this gradient is positive indicating that the predicted probability could be better and should move towards 1 (thus reducing uncertainty). Conversely, if the example is benign and the predicted probability is greater than 0 (indicating that the model thinks this example is more likely to be malignant), the gradient is negative, which drives its value the other way (thus reducing uncertainty).

Guarding Against Overfitting As RRTs are relational extensions of classical regression trees, they can be regularized in similar ways to guard against overfitting. Similar to decision trees, controlling the maximum number of nodes from root to leaf (depth) in the learned tree will lead to learning simpler trees. Note that RFGB does not learn large trees and then prune them but limits learning to shorter trees at the very outset in order to improve training efficiency and scalability. As RFGB is an ensemble approach, this strategy is effective since it is looking for several small, weak models rather than a giant, strong model. In addition, as each RRT also contains logical clauses in its nodes, we can also control the maximum number of clauses in the tree (paths from root to leaf), which is the maximum number of leaves. A third regularization strategy is to control the number of literals (splitting conditions) in each node. As with any other machine-learning approach, these regularization parameters and settings must be identified via cross validation for the best results.

Results

We compare the performance of several ML algorithms on the data set of radiomics features on a binary classification task of discriminating between malignant and benign tumors. All ML approaches but RFGB were evaluated using Waikato Environment for Knowledge Analysis (WEKA) software[1] (Waikato, New Zealand). Hyper-parameters (such as regularization and kernel parameters) for each method were selected using tenfold cross-validation, and all results are averaged over 10 runs, each with a different random split of training and test sets. We used the publicly available Java implementation of RFGB called RDN-Boost, which is a part of the Boost-SRL package[2]. Relational regression tree depth was limited to 3, and the node size per branch was set to 2.

We consider three metrics to compare the ML algorithms from different perspectives: (1) accuracy; (2) F-measure, also known as F_1 score ($2\frac{P \cdot R}{P+R}$, the harmonic mean of precision and recall); and (3) area under the receiver-operator characteristic (AUC-ROC). While accuracy is a standard performance metric, as the application is a clinical diagnostic system, minimizing misdiagnoses, that is, false negatives, is a priority. For this reason, we also compare the methods using F-measure and the AUC-ROC. Higher F-measure means that the classifier is better

[1] https://www.cs.waikato.ac.nz/ml/weka/
[2] https://starling.utdallas.edu/software/boostsrl/

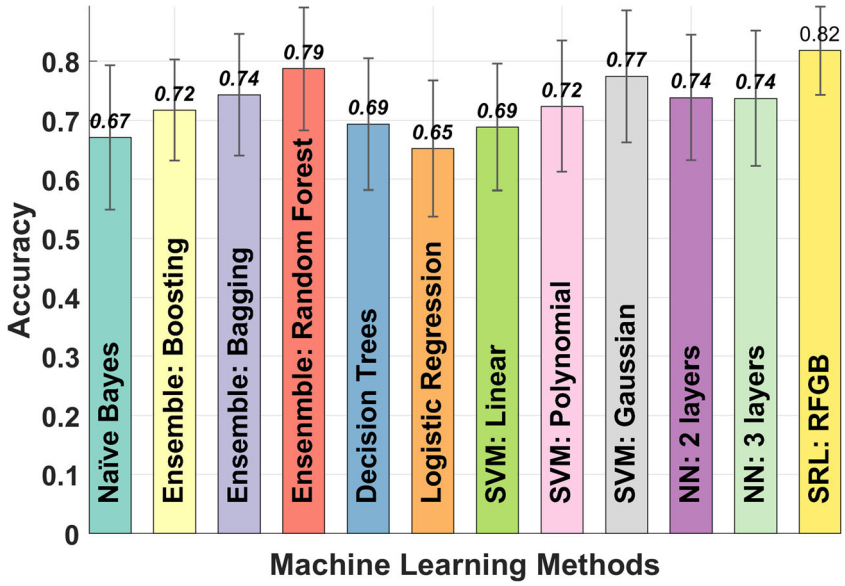

Fig. 3 Accuracy of various ML models (averaged over 10 runs) in renal mass classification compared to RFGB (bold numbers over the bars are significantly different from RFGB at $p = 0.05$)

at handling the class imbalance, while higher AUC-ROC means that the classifier broadly achieves low false-negative rates across a variety of thresholds. These results are shown in Figures 3, 4, and 5. At a high-level, these results show that RFGB consistently outperforms several standard ML algorithms on all three evaluation metrics significantly at $p = 0.05$. Finally, note that visual classification by experts only achieves AUC-ROC of 0.65 using the pathological gold standard [50], which suggests that many ML approaches can already outperform human diagnostic baselines on this difficult task.

Comparison with Tree-Based and Ensemble Methods RFGB significantly outperforms decision trees [38] and ensemble methods such as AdaBoost with decision stumps [39], bagging with regression trees [40], and performs comparably to Random Forest [41]. This shows that relational features such as entInc(N, E) (which indicates increasing entropy from the nephrographic to the excretory phase; Fig. 1, left) can considerably improve classification rates. This is in stark contrast to the flat-feature representation of conventional machine learning methods (that is, a single table), which would require extensive feature engineering to achieve.

Comparison with Kernel Methods SVMs [34] have long been a de facto standard for many classification applications. One compelling reason for this is that SVMs can learn non-linear classifiers to represent complex decision boundaries through a kernel function. The kernel function $k(\mathbf{x}, \mathbf{z})$ measures similarity between two training examples $\mathbf{x}$ and $\mathbf{z}$ in a high-dimensional space without explicitly transforming $\mathbf{x}$ and $\mathbf{z}$ into this space. For example, the Gaussian kernel is an infinite-dimensional polynomial kernel [37], p. 297:

$$k(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\gamma \|\boldsymbol{x}-\boldsymbol{z}\|^2\right) = \sum_{n=0}^{\infty} \frac{\gamma^n}{n!} \left(\boldsymbol{x}^T \boldsymbol{z}\right)^n,$$

which enables us to compute the similarity of these two points in an infinite-dimensional space without explicit transformation (which is impossible anyway). This elegant property is also the SVM's most serious limitation for our purposes: we cannot visualize feature combinations expressed algebraically [37, Remark 9.9], which seriously limits explainability in terms of features (Figure 1, right). On the other hand, explanation of RFGBs result can be expressed using comparison ($\geq, \leq, \approx$) and conjunction ($\wedge$) operators (Figure 1, left). Further, as more significant features are identified by RFGB, they can be added to the model as it is learned stage-wise. Finally, RFGB comprehensively outperforms linear, polynomial, and Gaussian-kernel SVMs (Figure 1, right).

Comparison with Deep Learning/Neural Networks A deep network for an application such as this requires several layers, thousands of training examples, powerful computing, and long training times. More worryingly, as the network becomes deeper, the model identifies increasingly complex combinations of features that depend on previous layers, making the model difficult to interpret or explain. Instead of letting a deep network identify potentially unintelligible features, we use radiomic features that have proven clinical relevance [8, 9, 25, 26] as well as explainability. Using RFGB, we identify combinations of these features and discover relationships. So, given useful radiomics features, can ANNs do as well as

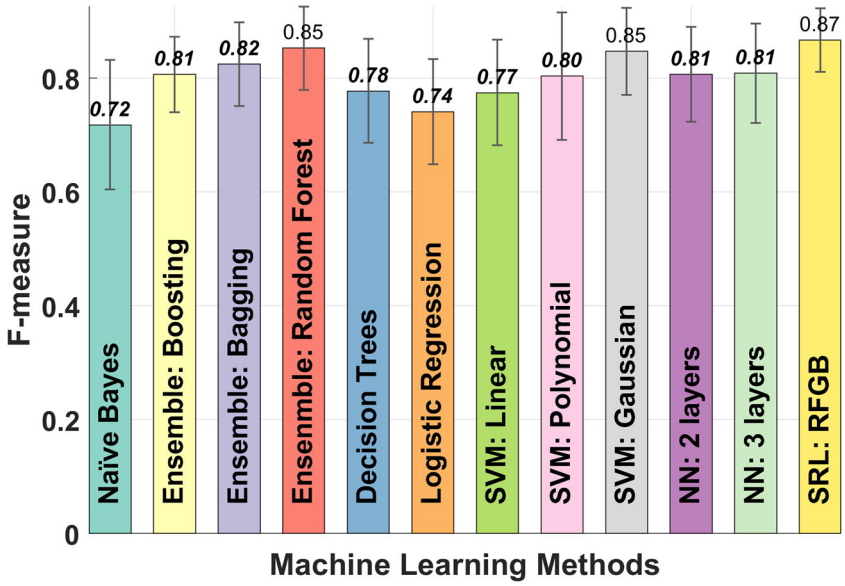

Fig. 4 F-measures of various ML models (averaged over 10 runs) in renal mass classification compared to RFGB (bold numbers over the bars are significantly different from RFGB at $p = 0.05$). The F-measure is the harmonic mean of precision and recall

RFGB? As we are limited by the data set size, we are only able to consider shallow neural networks. Figures 3, 4, and 5 show that shallow networks perform consistently worse than RFGB on all metrics. The key advantage of our pipeline is that we are able to exploit radiomics features to learn effective models with much fewer training examples.

Conclusions

We considered the task of differentiating between malignant and benign renal cell carcinomas using radiomics-based features. Our work demonstrates the usefulness of statistical relational learning, specifically, RFGB as a promising CDS tool for renal mass diagnosis as it learns models that are effective and explainable. This is beneficial from a clinical decision-support standpoint as it allows us to develop diagnostic systems that can support their decisions with explanations that are understandable to non-machine-learning domain experts such as urologists and radiologists. We are currently expanding our data set to include more cases. We are also currently expanding the radiomics platform to include shape-based and spectral/frequency-domain features [51].

Our next steps are to look beyond the binary classification task considered here. Consequently, we will extend RFGB to identify (1) subtype of tumors for benign (e.g., angiomyolipoma or oncocytoma) and malignant (e.g., clear cell or papillary RCC) and (2) nuclear grade: to categorize RCC according to genitourinary pathology expert classification. In addition, we are also exploring means to improve

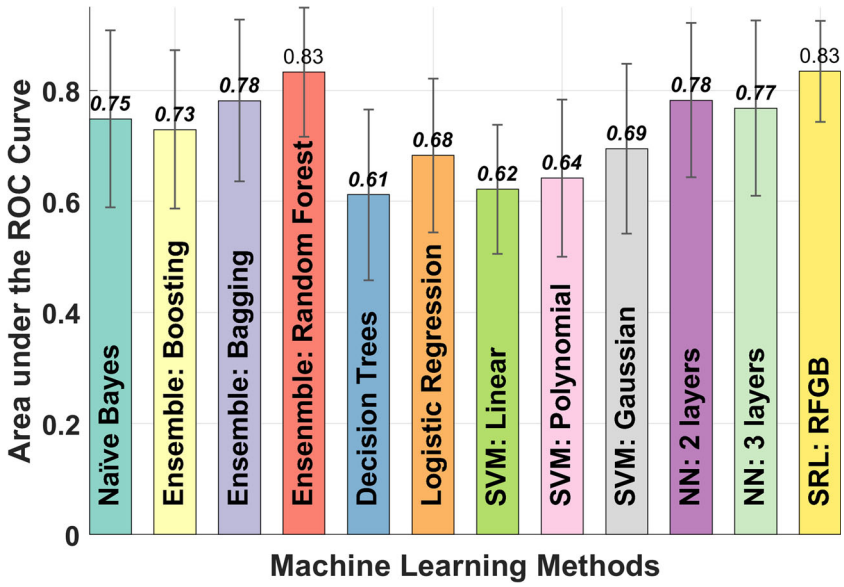

Fig. 5 Area under the receiver operator curve (AUC-ROC) of various ML models (averaged over 10 runs) in renal mass classification compared to RFGB (bold numbers over the bars are significantly different from RFGB at $p = 0.05$). The AUC-ROC is a measure of the probability that a classifier will rank a randomly chosen positive example (malignant) higher than a randomly chosen negative example (benign)

RFGB performance by including domain rules recommended by radiologists; the underlying first-order logic representation allows for the direct incorporation of such rules into the model. Also, multiple modalities beyond imaging features such as clinical and demographic information can be included as features/rules, which should also substantially improve the discriminating power of the method. Another potential pitfall we seek to address in future work is *class imbalance* (Table 2), an unequal number of malignant and benign examples. This risk can be mitigated by considering an alternative approach within the RFGB framework known as soft-RFGB [20], which implicitly weights examples to account for class imbalance.

References

1. National Cancer Institute. Cancer prevalence and cost of care projections. https:// costprojections.cancer.gov/graph.php, 2018. [Online; accessed 03-January-2018].
2. Rendon RA: Active surveillance as the preferred management option for small renal masses. Can Urol Assoc J 4:136–138, 2010
3. Adam C. Mues and Jaime Landman: Small renal masses: current concepts regarding the natural history and reflections on the American Urological Association guidelines. Curr Opin Urol 20, 2010.
4. Heuer R, Gill IS, Guazzoni G, Kirkali Z, Marberger M, Richie JP, de la Rosette JJMCH: A critical analysis of the actual role of minimally invasive surgery and active surveillance for kidney cancer. Eur Urol 57(2):223–232, 2010
5. Xipell JM: The incidence of benign renal nodules (a clinicopathologic study). J Urol 106(4):503–506, 1971
6. Gill IS, Aron M, Gervais DA, Jewett MAS: Small renal mass. N Engl J Med 362(7):624–634, 2010
7. Mindrup Steven R, Pierre Jessica S, Laila D, Konety Badrinath R: The prevalence of renal cell carcinoma diagnosed at autopsy. BJU Int 95(1):31–33, 2005
8. Duddalwar V, Zhang X, Hwang D, Cen S, Yap F, Ugwueze C, Abreu A, Aron M, Desai M, Gill I: PD14-07 Differentiation between clear cell renal cell carcinomas and oncocytomas using texture analysis of CT images. J Urol 195(4):e305, 2016
9. Bino Abel Varghese, Darryl Hwang, Steven Cen, Bhushan Desai, Felix Yap, and Vinay Duddalwar: Spectral Analysis of Renal Tumors: Evaluation of a CT Radiomic Technique. Radiol Soc N Am, 2016.
10. C. Reddy and C. Aggarwal: Healthcare Data Analytics. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2016.
11. Miller RA: Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc 1(1):8–27, 1994
12. Wyatt JC, Altman DG: Commentary: Prognostic models: clinically useful or quickly forgotten? Br Med J 311(7019):1539–1541, 1995
13. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, Spurr C, Khorasani R, Tanasijevic M, Middleton B: Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 10(6): 523–530, 2003
14. Purcell GP: What makes a good clinical decision support system: we have some answers, but implementing good decision support is still hard. Br Med J 330(7494):740–741, 2005
15. Wears RL, Berg M: Computer technology and clinical work: still waiting for Godot. J Am Med Assoc 293(10):1261–1263, 2005
16. C. Hu, R. Ju, Y. Shen, P. Zhou, and Q. Li: Clinical decision support for Alzheimer's disease based on deep learning and brain network. In 2016 IEEE International Conference on Communications (ICC), pages 1–6, 2016.
17. L. Getoor and B. Taskar. Introduction to Statistical Relational Learning. MIT Press, 2007.
18. De Raedt L, Kersting K, Natarajan S, Poole D: Statistical Relational Artificial Intelligence: Logic, Probability, and Computation, volume 32 of Synthesis Lectures on Artificial Intelligence and Machine Learning. San Rafael, CA: Morgan & Claypool, 2016
19. Sriraam Natarajan, Kristian Kersting, Tushar Khot, and Jude W. Shavlik: Boosted Statistical Relational Learners—From Benchmarks to Data-Driven Medicine. Springer Briefs in Computer Science. Springer, 2014.
20. S. Yang, T. Khot, K. Kersting, G. Kunapuli, K. Hauser, and S. Natarajan: Learning from imbalanced data in relational domains: a soft margin approach. In 2014 IEEE International Conference on Data Mining (ICDM), pages 1085–1090, 2014.
21. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D: Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. AI Mag 33(4):33, 2012
22. D. Page, S. Natarajan, V. Santos Costa, P. Peissig, A. Barnard, and M. Caldwell: Identifying adverse drug events from multi-relational healthcare data. In Proceedings of AAAI Conference on Artificial Intelligence, pages 790–793, 2012.
23. Haley MacLeod, Shuo Yang, Kim Oakes, Kay Connelly, and Sriraam Natarajan. Identifying rare diseases from behavioural data: a machine learning approach. In Proceedings of the 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016.
24. Natarajan S, Saha BN, Joshi S, Edwards A, Khot T, port EMD, Kersting K, Whitlow CT, Maldjian JA: Relational learning helps in three-way classification of Alzheimer patients from structural magnetic resonance images of the brain. Int J Mach Learn Cybern 5:659–669, 2014
25. Chen F, Huhdanpaa H, Desai B, Hwang D, Cen S, Sherrod A, Bernhard J-C, Desai M, Gill I, Duddalwar V: Whole lesion quantitative CT evaluation of renal cell carcinoma: differentiation of clear cell from papillary renal cell carcinoma. SpringerPlus 4(1):66, 2015
26. Yap F, Hwang D, Cen S, Zhang X, de Castro Abreu AL, Desai M, Aron M, Gill I, Duddalwar V: The shapely renal mass: contour evaluation of renal cell carcinoma. J Urol 195(4):e204, 2016
27. Bino Abel Varghese, Frank Chen, Darryl Hwang, Steven Cen, Inderbir Gill, and Vinay Duddalwar: Differentiation of predominantly solid, enhancing lipid-poor renal cell masses using contrast-enhanced computed tomography: evaluating the role of texture in tumor sub-typing. Am J Roentgenol, accepted (to appear), 2018.
28. C. G. Ugwueze, M. Nayyar, Darryl Hwang, Steven Cen, Felix Yap, Bhushan Desai, Inderbir S. Gill, M Desai, and Vinay Duddalwar. Texture analysis as an image-based discriminator between T1 renal cell carcinoma and pT3 renal cell carcinoma. RSNA Annual Meeting, Chicago, 2016.
29. Chen F, Gulati M, Hwang D, Cen S, Yap F, Ugwueze C, Varghese B, Desai M, Aron M, Gill I, Duddalwar V: Voxel-based whole-lesion enhancement parameters: a study of its clinical value in differentiating clear cell renal cell carcinoma from renal oncocytoma. Abdom Radiol 42(2):552–560, 2017
30. Francesco Giuseppe Mazzei, Maria Antonietta Mazzei, Nevada Cioffi Squitieri, et al. CT perfusion in the characterisation of renal lesions: an added value to multiphasic CT. BioMed Res Int, 2014, 2014. Article ID: 135013

31. Haralick RM, Shanmugam K, Dinstein I: Textural features for image classification. IEEE Trans Syst Man Cybern SMC-3(6):610–621, 1973
32. Bino A. Varghese, Darryl H. Hwang, Steven Y. Cen, Bhushan B. Desai, Felix Yap, Inderbir Gill, Mihir Desai, Manju Aron, Gangning Liang, Michael Chang, Christopher Deng, Mike Kwon, Chidubem Ugweze, Frank Chen, and Vinay A. Duddalwar. Fast Fourier transform-based analysis of renal masses on contrast-enhanced computed tomography images for grading of tumor. In Proceedings Volume 10160, 12th International Symposium on Medical Information Processing and Analysis, 2017.
33. Guyon I, Weston J, Barnhill S, Vapnik V: Gene selection for cancer classification using support vector machines. Mach Learn 46(1): 389–422, 2002
34. Cortes C, Vapnik V: Support-vector networks. Mach Learn 20(3): 273–297, 1995
35. Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
36. P. McCullagh and John A. Nelder. Generalized Linear Models, 2nd edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, 1989.
37. John Shawe-Taylor and Nello Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
38. Leo Breiman, Jerome H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, 1984.
39. Robert E. Schapire. A brief introduction to boosting. In Proceedings of the 16th International Joint Conference on Artificial Intelligence—Volume 2, pages 1401–1406, 1999.
40. Breiman L: Bagging predictors. Mach Learn 24(2):123–140, 1996
41. Breiman L: Random forests. Mach Learn 45(1):5–32, 2001
42. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.
43. Druzhkov PN, Kustikova VD: A survey of deep learning methods and software tools for image classification and object detection. Pattern Recognit Image Anal 26(1):9–15, 2016
44. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Snchez CI: A survey on deep learning in medical image analysis. Med Image Anal 42:60–88, 2017
45. Andrearczyk V, Whelan PF: Using filter banks in convolutional neural networks for texture classification. Pattern Recogn Lett 84: 63–69, 2016
46. P. Domingos and D. Lowd. Markov Logic: An Interface Layer for Artificial Intelligence. Morgan & Claypool Publishers, 2009.
47. J.H. Friedman: Greedy function approximation: a gradient boosting machine. Ann Stat 29, 2001.
48. Natarajan S, Khot T, Kersting K, Gutmann B, Shavlik J: Gradient-based boosting for statistical relational learning: the relational dependency network case. Mach Learn 86(1):25–56, 2012
49. Blockeel H, De Raedt L: Top-down induction of first-order logical decision trees. Artif Intell 101:285–297, 1998
50. Shin T, Duddalwar VA, Ukimura O, Matsugasumi T, Chen F, Ahmadi N, de Castro Abreu AL, Mimata H, Gill IS: Does computed tomography still have limitations to distinguish benign from malignant renal tumors for radiologists? Urol Int 99(2):229–236, 2017
51. Yap FY, Hwang DH, Cen SY, Varghese BA, Desai B, Quinn BD, Gupta MN, Rajarubendra N, Desai MM, Aron M, Liang G, Aron M, Gill IS, Duddalwar VA: Quantitative contour analysis as an image-based discriminator between benign and malignant renal tumors. Urology 114:121–127, 2018