HHS Public Access

Author manuscript

Med Phys. Author manuscript; available in PMC 2023 November 28.

Published in final edited form as:

Med Phys. 2023 August ; 50(8): 5020–5029. doi:10.1002/mp.16303.

Deep learning-based decision forest for hereditary clear cell renal cell carcinoma segmentation on MRI

Nathan Lay[1], **Pouria Yazdian Anari**[2], **Aditi Chaurasia**[2], **Fatemeh Dehghani Firouzabadi**[2], **Stephanie Harmon**[1], **Evrim Turkbey**[2], **Rabindra Gautam**[3], **Safa Samimi**[2], **Maria J. Merino**[4], **Mark W. Ball**[3], **William Marston Linehan**[3], **Baris Turkbey**[1], **Ashkan A. Malayeri**[2]

[1]Artificial Intelligence Resource, Molecular Imaging Branch, National Cancer Institute, Bethesda, Maryland, USA

[2]Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, Maryland, USA

[3]Urologic Oncology Branch, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, USA

[4]Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, USA

Abstract

Background: von Hippel–Lindau syndrome (VHL) is an autosomal dominant hereditary syndrome with an increased predisposition of developing numerous cysts and tumors, almost exclusively clear cell renal cell carcinoma (ccRCC). Considering the lifelong surveillance in such patients to monitor the disease, patients with VHL are preferentially imaged using MRI to eliminate radiation exposure.

Purpose: Segmentation of kidney and tumor structures on MRI in VHL patients is useful in lesion characterization (e.g., cyst vs. tumor), volumetric lesion analysis, and tumor growth prediction. However, automated tasks such as ccRCC segmentation on MRI is sparsely studied. We develop segmentation methodology for ccRCC on T1 weighted precontrast, corticomedullary, nephrogenic, and excretory contrast phase MRI.

Methods: We applied a new neural network approache using a novel differentiable decision forest, called hinge forest (HF), to segment kidney parenchyma, cyst, and ccRCC tumors in 117 images from 115 patients. This data set represented an unprecedented 504 ccRCCs with 1171 cystic lesions obtained at five different MRI scanners. The HF architecture was compared with U-Net on 10 randomized splits with 75% used for training and 25% used for testing. Both methods were trained with Adam using default parameters ($\alpha = 0.001,\ \beta_1 = 0.9,\ \beta_2 = 0.999$) over

Correspondence: Nathan Lay, Artificial Intelligence Resource, Molecular Imaging Branch, National Cancer Institute, Bethesda, MD, USA. nathan.lay@nih.gov.

CONFLICT OF INTEREST STATEMENT

The Artificial Intelligence Resource and Molecular Imaging Branch have a research CRADA with NVIDIA.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

1000 epochs. We further demonstrated some interpretability of our HF method by exploiting decision tree structure.

Results: The HF achieved an average kidney, cyst, and tumor Dice similarity coefficient (DSC) of 0.75 ± 0.03, 0.44 ± 0.05, 0.53 ± 0.04, respectively, while U-Net achieved an average kidney, cyst, and tumor DSC of 0.78 ± 0.02, 0.41 ± 0.04, 0.46 ± 0.05, respectively. The HF significantly outperformed U-Net on tumors while U-Net significantly outperformed HF when segmenting kidney parenchymas ($\alpha < 0.01$).

Conclusions: For the task of ccRCC segmentation, the HF can offer better segmentation performance compared to the traditional U-Net architecture. The leaf maps can glean hints about deep learning features that might prove to be useful in other automated tasks such as tumor characterization.

Keywords

decision forest; deep learning; MRI; segmentation; von Hippel–Lindau

1 | INTRODUCTION

von Hippel–Lindau syndrome (VHL) is an autosomal dominant hereditary syndrome with an increased predisposition of developing clear cell renal cell carcinoma (ccRCC). Considering the lifelong surveillance in VHL patients to monitor the disease, patients with VHL are preferentially imaged using MRI to eliminate radiation exposure. Since ccRCC is the leading cause of death in VHL patients, there is an interest towards automated analysis of renal masses in VHL patients; for example, growth prediction, which requires segmentation of the renal masses.

Use of artificial intelligence (AI) methods to further study ccRCC tumors in VHL presents some unique challenges. First, the renal imaging in VHL is predominantly using MRI while other types of RCCs are usually imaged using CT. There has been limited research in the automated renal lesion segmentation on MRI. Second, appearance of kidney parenchyma and renal lesions in VHL patients could be highly variable in appearance. Kidneys in VHL could range from being normal in size with few cysts to large kidneys with numerous renal lesions.[1] The majority of the renal lesions are cysts, although a number of other renal masses can be frequently encountered as well. Considering the challenges posed by numerous cysts of varying sizes and shapes and diversity in the appearances of kidney parenchyma, cysts, and tumors in VHL, these cases are incredibly challenging and time consuming to annotate manually for clinical and research purposes.

Deep learning, particularly U-Net,[2] has shown considerable promise in biomedical image segmentation. While U-Net is a competitive approach for segmentation in medical images, it is more difficult to interpret the model owing to deep learning methods largely being black boxes. In this work, we adapt classic decision forest to deep learning using a consistent and differentiable logic that retains all logarithmic computational guarantees of decision trees. Convolution blocks form the deep feature extractor, and the hinge forest (HF) uses the features to predict segmentation tiles. The two components are trained simultaneously

via stochastic gradient descent and backpropagation. We use this novel decision forest architecture, called HF, to segment kidney paranchyma and ccRCC masses on MRIs of VHL patients.

1.1 | Related work

This section highlights some important related works in both RCC segmentation and differentiable decision trees.

1.1.1 | RCC segmentation—

Segmentation of ccRCC in MRI is sparsely explored in the literature. Two-dimensional U-Net was used for detection and segmentation of ccRCC in MRI and CT and achieved a tumor sensitivity of 0.76 and tumor DSC of 0.34 on five-fold cross validation on a cohort of 118 MR images.[3] Using attention U-Net,[4] segmentation performance (>0.4) was further improved at the cost of lower sensitivity.

Almost all recent works focus on RCC detection and segmentation in CT scans owing to the ubiquity of computed tomography (CT) imaging for renal mass assessment as well as the public availability of the KITS19 data set.[5,6] KITS19 is a CT RCC grand challenge data set with the challenge objective being to segment kidney parenchyma and RCC lesions. The vast majority (if not all) of recent renal mass segmentation methods are based on deep convolutional neural networks. Many of the state-of-the-art segmentation methods employ variations of 2D and 3D U-Net such as nnU-Net,[7] though KITS19 features three top-five submissions employing custom architectures. The top scoring method of KITS19 had a tumor Dice similarity coefficient (DSC) of 0.85.[5]

Our work differs in that we explore a novel and interpretable deep learning segmentation architecture on VHL RCC MRI cases. Additionally, existing work only considers sporadic RCC cases that are considerably less complex since these cases generally present with one tumor on imaging while VHL RCC cases often present numerous tumors and cysts. A representative example of VHL RCC can be seen in Figure 1.

1.1.2 | Differentiable decision trees—

Classic decision tree learning incrementally partitions the feature space to maximize information gain.[8] This greedy optimization can lead to suboptimal decision tree models. Differentiable decision trees were developed to find globally optimal decision tree models through gradient-based optimization. The concept is not new and has been around since at least 1999.[9] There have been many examples of these in recent years such as Refs.10–16 These can be roughly grouped into three categories: *fuzzy logic*, *routing*, and *other*. Methods such as Refs. 9–11, 15 are examples of *fuzzy logic*. These directly make Boolean logic operate with values in $[0,\ 1] \subset \mathbb{R}$ and use approximate and differentiable versions of discrete comparison operations. Methods such as Refs.12, 14 are examples of *routing* methods that discretely determine decision paths in a decision tree using a stochastic decision function. Other interpretations of decision trees include Refs., 13, 16 which do not squarely fall into either category. For example, Lee and Jaakkola[13] exploit the derivative of ReLU (which is binary thresholding) to define a differentiable decision tree.

Rather than directly *soften* operations within a decision tree to be differentiable like the *fuzzy logic* methods, we instead define a consistent and differentiable logic that operates

on the real numbers. Reinterpreting decision trees with this system of logic then directly leads to a differentiable decision tree. Unlike *fuzzy logic* values that admit fractions of membership to both `true` and `false`, our differentiable logic only admits membership to exactly one of `true`, `unknown`, and `false` by encoding the logic value in the sign of real numbers. Our system is consistent with binary logic. Because of this property, our approach shares computation efficiency advantages of *routing*, although routing is generally stochastic while our method is deterministic.

2 | METHODS

2.1 | Patient population

This single-institution study was approved by the institutional review board (protocol NCI-89-C-0086). Written informed consent was obtained from all patients. We queried our institutional records for all VHL patients who underwent surgical resection of tumors >1 cm between January 2015 and June 2021 and had a preoperative MRI. One hundred and fifteen patients with 504 clear cell carcinomas and 1171 cystic lesions were identified.

All MRI examinations were obtained from five MRI scanners (Philips Achieva 1.5T and 3T; Siemens Aera 1.5T, Verio 3T, Biograph mMR 3T). Patients underwent several routine imaging sequences including: in and out of phase T1 weighted images, precontrast T1 and post-contrast T1 (corticomedullary, nephrogenic, and delayed phases). One dose of gadobutrol (0.1 mmol per kilogram of body weight, Gadavist; Bayer, Washington, DC) was administered and followed by a 20-mL saline flush. After i.v. contrast injection, images were acquired during the corticomedullary (20 s), nephrogenic (70 s), and excretory (3 min) phases.

All images were manually segmented into kidney parenchyma, cyst, and tumors on the 3-min delay phase using ITK-SNAP.[1] All the segmentations were performed by two postdoctoral fellows each having 1 year of experience; segmentations were checked and finally revised by an experienced body radiologist with 11 years of experience. Near exact pathology-imaging correlation was achieved by denoting the tumors on presurgical MRI by lesion numbers and interpreting the pathology specimens based on the given lesion number. Since surgery often only involved one kidney, only one kidney was annotated in most cases. Solid or solid cystic tumors without pathology validation and lesions <0.5 cm were not annotated.

2.2 | Experimental approach

Images were registered to the post-contrast image using niftyreg[2] and resampled to have 1 mm × 1 mm axial slices. These MRI images were normalized using Rician normalization.[17] Examples of these T1W images and ground truth segmentation can be found in Figure 1. As many images only featured one annotated kidney, images were sagittally halved and the models trained to segment kidneys and renal masses in half-images. Images were flipped sagittally as a simple means of data augmentation (though these images are *not* symmetric). Training batches were formed using consecutive slices of a volume (modulo

the total number of slices). For further data augmentation, batches were formed relative to a randomized initial slice position.

HF was compared to U-Net on 10 randomized splits of 117 images coming from 115 patients. Here 75% was used for training and 25% was used for testing. There were no shared patients between training and test sets. Of the training set, 10% was used for validation. Validation tumor DSC was used to select optimal model states. Test DSCs for kidney, cyst, and tumor were compared, and paired *t*-test was used to establish statistical significance.[18] To help ensure that splits were representative, random splits were evenly stratified over cases with <10 and ≥10 annotated cysts. Cases with numerous cysts are generally more challenging and training or test sets with disproportionate numbers of high cyst count cases can result in deceptive performance.

Our work used U-Net with some architectural modifications for the task of semantic image segmentation.[19] Leaky ReLU (negative slope = 0.2) was used instead of ReLU. Batch normalization[20] layers were inserted immediately before all leaky ReLU activation layers.

The HF was comprised of 100 depth 7 hinge trees and used features calculated by three consecutive VGG blocks[3,21] Each VGG block downsamples the image by a factor of 2. To produce a segmentation in the original image resolution, the HF was trained to predict 8 × 8 tiles. A graphical depiction of this architecture can be seen in Figure 2.

Both models were trained to minimize cross entropy loss using Adam[22] with learning rate $\alpha = 0.001$ and default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ over 1000 epochs using a batch size of 16. There were 54 tumors with unknown pathology that did not contribute to the loss calculation or gradient calculation. Hypointense regions inside the segmented kidney were assumed to be unannotated cysts and also did not contribute to the loss or gradient calculations. In every experiment, models were initialized with random weights.

Inference segmentations were generated by computing the argmax of the probability map that is produced by softmax. A connected component analysis was used to identify up to a single contiguous component of kidney, cyst, and tumor in each sagittally halved image. All other components were suppressed as there should be up to one kidney possibly with cysts and tumors in each half image. If there were no voxels labeled as kidney, the entire half image was suppressed of all components producing an empty segmentation for that image half.

To demonstrate quantitative interpretability, we additionally calculated leaf maps on the training set. The background was suppressed and we calculated a segmentation label count in each leaf. The leaves were ranked in ascending order by Gini impurity. We then extracted decision rules for the top leaf describing > 50% of voxels for each segmentation structure and compared with the test set. We measured test set Gini impurity and ratio of overlap of the voxels of each corresponding structure.

All experiments were conducted using PyTorch[4] 1.7.1 (HF) and 1.10.1 (U-Net) with a custom C++ PyTorch extension of HF. To expedite our experiments, we used Tesla V100 GPUs on CentOS 7.5 servers.

2.3 | Statistical analysis

Segmentation results were analyzed in two different ways: mask-based and connected component-based. The mask-based analysis treats each structure type (i.e., kidney, cyst, tumor) as a single all-inclusive binary mask independent of the number of distinct structures present. The only measurement made in the mask-based analysis is the DSC defined as

$$\mathrm{DSC}(S^*, S) = \frac{2|S^* \cap S|}{|S^*| + |S|} \tag{1}$$

where S^*, S are the ground truth and model segmentation masks, respectively, $|S|$ is the count of nonzero voxels in mask S and $|S^* \cap S|$ is the count of corresponding nonzero voxels in common between the two masks. The connected component-based analysis considers each individual connected component of each structure type in both the ground truth and segmentation masks. This kind of analysis helps facilitate both a geometric understanding and failure mode understanding. Each ground truth segmented structure S^*_{cc} is used to identify incident segmentation connected components that are combined into the mask S_{cc}. The segmentation connected component S_{cc} is considered to have *detected* ground truth connected component S^*_{cc} if the overlap ratio exceeds 0.5. Overlap ratio is defined as

$$\mathrm{OverlapRatio}(S^*, S) = \frac{|S^* \cap S|}{|S^*|} \tag{2}$$

and describes how much a given segmentation S overlaps the ground truth S^* with a value of 0 meaning no overlap and 1 being complete overlap. Then DSC, average symmetric surface distance (ASSD),[23] and Hausdorff (HD)[23] are calculated between S^*_{cc} and S_{cc}. A count of *missed* connected components S^*_{cc} is kept. Those segmentation connected components of S that do not sufficiently overlap (i.e., ≤0.5 overlap) with any ground truth connected component S^*_{cc} are counted as *extras*.

Receiver operator characteristic (ROC) curves for tumors were constructed against *detected*, *missed* (false negative) and *extra* (false positive) connected components. The score for each connected component is simply the median of the probability scores (i.e., softmax) inside of the connected component. The median was chosen since it is robust to outliers and it implies that half of the volume of the connected component exceeds that score. The median probability score was calculated against ground truth connected components S^*_{cc} independent of overlap ratio while the median probability score was calculated against individual *extra* connected components from the predicted segmentation mask S. The area under ROC curve (AUC) was calculated.

Statistics for ASSD, HD, and DSC were averaged over all structures over all splits as follows:

$$\bar{x} = \frac{1}{10}\sum_{s=1}^{10}\frac{1}{N_s}\sum_{n=1}^{N_s} x_{sn} = \frac{1}{10}\sum_{s=1}^{10}\bar{x}_s \tag{3}$$

$$\mathrm{std}(x) = \sqrt{\frac{1}{9}\sum_{s=1}^{10}(\bar{x}_s - \bar{x})^2} \tag{4}$$

where s is the index over splits and n is the index over structures (e.g., tumors) and N_s are the number of ground truth structures per split.

ROC curves were averaged by resampling all ROC curves $\{(x_{si}, y_{si})\}_{i=1}^{M_s}$ to have identical false positive rate knots $x_j,\ j = 1,\ 2, \ldots, M$, and then averaging the resampled true positive rates. The knots were chosen as the union of all false positive rates over all splits. New true positive operating point y_{sj} were linearly interpolated for each x_j for each split $s = 1,\ 2, \ldots, 10$ and the mean and standard deviation calculated $\bar{y}_j,\ \mathrm{std}(y)_j$. Because the curves were linearly interpolated, we employed composite trapezoid rule to calculate mean AUC and standard deviation as follows:

$$\overline{\mathrm{AUC}} = \int_0^1 \bar{y}(x)dx \approx \sum_{j=2}^{M} \frac{\bar{y}_{j-1} + \bar{y}_j}{2}(x_j - x_{j-1}) \tag{5}$$

$$\mathrm{std}(\mathrm{AUC}) = \sqrt{\int_0^1 \frac{1}{9}\sum_{s=1}^{10}(y_s(x) - \bar{y}(x))^2 dx} \tag{6}$$

$$\approx \sqrt{\sum_{j=2}^{M} \frac{\mathrm{std}(y)_{j-1}^2 + \mathrm{std}(y)_j^2}{2}(x_j - x_{j-1})} \tag{7}$$

2.4 | Hinge forest

Decision trees partition the feature space with respect to a loss or gain function. Each partition corresponds to a decision tree leaf and each leaf makes a constant prediction in that partition. Denote a given feature vector $\mathbf{x} \in \mathbb{R}^F$ to be an F dimensional real vector. A decision tree leaf ℓ is a path of vertices $v \in T$ in tree T. Each vertex corresponds to a decision on the feature vector $\mathbf{x}$ and the path is a conjunction over all decisions of the form

$$I_\ell(\mathbf{x}) = \bigwedge_{(d,v)\in\ell} I(d(\mathbf{x}_{f_v} > t_v)), \tag{8}$$

where $I(\cdot)$ is the indicator function, $\wedge$ is logical and, $d(x)$ represents the direction of traversal where

$$d(x) = \begin{cases} \neg x & \text{the next vertex is the left child,} \\ x & \text{otherwise,} \end{cases}$$

$\neg x$ represents logical negation of x, and $f_v,\ t_v$ are the feature index and threshold for the decision at vertex v, respectively. A simplified example of this can be seen in Figure 3.Since

partitions are disjoint, a decision tree can be interpreted as a superposition of predictions over all its leaves

$$h_T(\mathbf{x}) = \sum_{\ell \in T} w_\ell l_\ell(\mathbf{x}) . \tag{9}$$

where w_ℓ is a prediction stored in leaf ℓ.Since $I(d(\mathbf{x}_{f_v} > t_v))$ is piecewise constant, the derivative is 0 and Equation (9) is not directly useful for optimization with gradient-based methods.

Hinge trees use a consistent and differentiable system of logic on the real numbers. We define the positive reals to be *true* and the negative reals to be *false*. The special value 0 is *unknown*. Then *min*, *max*, and additive inverse are consistent with logical *and*, *or*, and *negation*. We further define the binary comparison $A > B$ in terms of a *hinge function*

$$I(A > B) = \max\{A - B, 0\} = \mathrm{ReLU}(A - B) . \tag{10}$$

Substituting this binary comparison and system of logic into Equations (8) and (9) gives

$$\hat{l}_\ell(\mathbf{x}) = \min_{(d,v) \in \ell} \{\mathrm{ReLU}(d(\mathbf{x}_{f_v} - t_v))\}, \tag{11}$$

$$h_T(\mathbf{x}) = \sum_{\ell \in T} w_\ell \hat{l}_\ell(\mathbf{x}), \tag{12}$$

$$= \sum_{\ell \in T} w_\ell \min_{(d,v) \in \ell} \{\mathrm{ReLU}(d(\mathbf{x}_{f_v} - t_v))\} . \tag{13}$$

The choice of Definition (10) ensures that at most one term of Equation (13) is nonzero.

The gradient with respect to the features $\mathbf{x}$, thresholds t, and weights w is sparse. The derivatives can be efficiently calculated for the relevant variables through classic decision tree traversal. The traversal algorithm is given in Algorithm A.1 in the appendix and the derivatives can be found in Equations (14)–(16).

$$\frac{\partial h_T}{\partial \mathbf{x}_f} = \begin{cases} w_\ell \mathrm{sgn}(\mathbf{x}_{f_{v_{\min}}} - t_{v_{\min}}) & \hat{l}_\ell(\mathbf{x}) > 0, f_{v_{\min}} = f, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

$$\frac{\partial h_T}{\partial t_v} = \begin{cases} -w_\ell \mathrm{sgn}(\mathbf{x}_{f_{v_{\min}}} - t_{v_{\min}}) & \hat{l}_\ell(\mathbf{x}) > 0, v_{\min} = v, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

$$\frac{\partial h_T}{\partial w_\ell} = \hat{l}_\ell(\mathbf{x}) . \tag{16}$$

where $v_{\min}$ corresponds to the output of Algorithm A.1 and $\hat{l}_\ell(\mathbf{x}) = |r_{\min}| = |\mathbf{x}_{f_{v_{\min}}} - t_{v_{\min}}|$. Here ℓ is also being used as an integer index to identify a leaf.

2.5 | Initialization

Classic decision trees optimize splitting dimensions and decision thresholds on an already-defined feature space. However, in a deep learning setting, feature representations are learned as the model is trained through stochastic gradient descent. This presents a conundrum: you cannot optimize the decision structure of hinge trees on feature representations that are not yet learned. The hinge trees and feature representations evolve together. This work randomly initialized the decision structure of hinge trees and relied on learnable feature representations to evolve and fit the decision structure. More investigation is needed on initialization strategies of hinge trees and is outside the scope of this work.

2.6 | Leaf maps

One way to interpret decision tree models on images is to examine how decisions behave spatially. Recall that each leaf of a decision tree uniquely represents a conjunction of decisions. The objective is then to observe how leaves behave on the image. Each tree leaf is assigned a unique integer index like the n in w_n in Figure 3. When a decision tree evaluates a voxel in the volume, only one leaf is visited and the integer index n for that specific leaf is assigned to that voxel. This results in a multilabel mask when a leaf is used for inference at that voxel. The multilabel mask is the *leaf map* and it can be used qualitatively and quantitatively to identify important decision criteria. Qualitatively, multilabel masks are often visualized as multicolor images with each label corresponding to a unique color. The idea is that decision tree leaves will identify segments of image structure (e.g., anatomical structure) and a user will be able to visualize and quickly spot this behavior. Then the features and prediction that define that particular leaf can be used for analysis or other learning tasks. An example of a leaf map can be seen in Figure 4. Quantitative analysis, for example, can be achieved by correlating tree leaves with the ground truth segmentation masks. Leaves with low impurity (e.g., entropy) for a specific segmentation label prediction correspond to discriminative features for some segmentation structure. These features can enable other kinds of analysis such as tumor characterization or growth prediction.

3 | RESULTS

The average performance of both U-Net and HF over the 10 randomized splits is presented using DSC as a performance metric in Table 1. Some prediction examples are presented in Figure 5. More prediction examples can be found in Figure S1. Table 2 shows examples of extracted decision rules from the leaf map analysis.

4 | DISCUSSION

In this work, we introduce a novel random forest architecture adapted for deep feature-based extraction and learning, enabling interpretable model characteristics. We benchmark this methodology against U-Net for segmentation of kidney structures, including parenchyma, cysts, and tumors, in hereditary RCC.

Both AI models performed similarly in our study. In terms of mask-based DSC, the U-Net performed significantly better at kidney parenchyma segmentation but worse than the HF architecture at tumor and cyst segmentation. In terms of connected component-based

metrics, we observe the opposite: HF performs better at kidney parenchyma segmentation but worse than U-Net. The connected component overlap criteria (> 50%) biases the performance. Gross segmentation errors are suppressed and counted as either *missed* or *extra*. For structures with better performance, we tend to see higher *missed* counts. For example, HF misses a total of 90 kidney parenchyma connected components but has better DSC, ASSD, and HD scores compared to U-Net, which misses only 67 kidney parenchyma connected components. The receptive field of the VGG feature extractor is smaller than the U-Net, which could explain the worse performance on kidney parenchyma and better performance on smaller structures like renal masses that are largely uncorrelated with surrounding image context.

As RCC tumor detectors, HF bests U-Net in AUC, but both models have relatively bad AUC scores. The bad AUC scores are also reflected by mask DSC, missed and extra counts.

Both models perform less than expected at segmenting RCCs or cysts. This can be partly explained by challenges associated with the complexity of multiphase MRI data. MRI data and signal is not standardized like CT, and using multiphase scans is quite susceptible to motion artifacts. Image registration helps with motion but is not perfect. Additionally, the MRI data used in this research was obtained in five different MRI scanners and this also adds to the complexity and heterogeneity of the data. Although HF significantly outperforms U-Net at ccRCC segmentation, kidney and cyst segmentations have artificially lower DSC owing to unannotated cysts. The low DSCs can also be attributed to the challenging appearance of VHL kidneys and confusing nature of RCC lesions, which can be cystic, solid, or mixed in appearance.

Our simplistic leaf map analysis shows that there are leaves that can accurately describe the majority of kidney parenchyma and cysts based on the relatively low Gini impurity. The higher Gini impurity of RCC lesions may be due to the varying appearance of tumors and the leaf map pixelation seen in Figure 4.While the individual features of decision rules are not necessarily explained by the leaf map analysis, the collection of features that define the rules consistently map to anatomical structures. For example, the rules shown in Table 2 map to 76, 46,and 39% of all kidneys, cysts, and tumor voxels in the test set. Furthermore, examples of these rules can be quickly identified by manually examining a leaf map. The features used by rules that map to anatomical structures may then find use in other analyses or learning-based tasks like tumor characterization.

4.1 | Leaf maps compared to saliency maps

Saliency maps are heat maps that indicate likely regions of interest used by a deep neural network in its prediction. There are many ways to generate saliency maps.[24–26] One basic approach is to calculate the gradient of a model with respect to the input image. The resulting gradient image will indicate regions of importance based on pixel magnitude.[24] While qualitatively similar, leaf maps convey more information than a saliency map. Not only can this allow visualizing how decisions are attributed to different regions of an image, this can also correlate those specific decision rules across images. For example, this could be useful to identify subtypes of pathologies described by a specific rule or to identify useful deep learning features for a specific kind of anatomical structure.

4.2 | Limitations

While the HF presents promising quantitative results, predicting 8 × 8 tiles could result in a discontinuous segmentation. Surprisingly, this does not seem to happen on our data. While leaf map analysis may reveal underlying image structure and also help identify useful rules and features, this does not explain the meaning of the features.

Another limitation in this work is the use of default parameters and the lack of hyperparameter tuning. Hyperparameter tuning and learning rate schedules may vastly improve or degrade U-Net or HF model performance. We attempt to mitigate these issues through 10 randomized splits each trained over 1000 epochs to better represent general performance and give models more time to find a good local optimum.

5 | CONCLUSION

We have shown promising initial results of the novel HF in the segmentation of VHL ccRCC lesions and we have shown its performance to be comparable with U-Net. Furthermore, we have shown one way to interpret HF using leaf maps. Being logically consistent with decision trees, HF can potentially benefit from other decision tree analysis techniques.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). This research was funded by the NIH Intramural Research Program. This work was also supported by funding from the NIH Clinical Center's Research Award for Staff Clinicians (RASCL).We also thank the Office of Biomedical Translational Research Informatics (BTRIS),particularly Gloria Oshegbo, for their assistance and provision of necessary data.

Funding information

NIH Intramural Research Program; NIH Clinical Center Research Award for Staff Clinicians (RASCL)

APPENDIX

Algorithm A.1

HingeTreeTraversal

28.8±5.7

Input: tree depth D, feature vector $\mathbf{x} \in \mathbb{R}^F$, decision thresholds $\mathbf{t} \in \mathbb{R}^{2^D-1}$, feature indices $\mathbf{f} \in \{0, 1, \ldots, F-1\}^{2^D-1}$

Output: leaf index $\ell \in \{0, 1, \ldots, 2^D-1\}$, signed margin $r_{\min} \in \mathbb{R}$, vertex index $v_{\min} \in \{0, 1, \ldots, 2^D-2\}$

Initialize $\ell = 0,\ v = 0,\ v_{\min} = 0,\ r_{\min} = \mathbf{x}_{\mathbf{f}_0} - \mathbf{t}_0$
for $i = 0$ **to** $D-1$ **do**
Set $r = \mathbf{x}_{\mathbf{f}_v} - \mathbf{t}_v$
if $|r| < |r_{\min}|$ **then**
Set $r_{\min} \leftarrow r,\ v_{\min} \leftarrow v$
end if

Set $\ell \leftarrow 2\ell + I(r > 0)$
Set $v \leftarrow 2v + I(r > 0) + 1$
end for

REFERENCES

1. Maher ER, Neumann HP, Richard S. von Hippel–Lindau disease: a clinical and scientific review. Eur J Hum Genet. 2011;19:617–623. [PubMed: 21386872]
2. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. 2015:234–241.
3. Agarwal A. Automated Detection of Renal Masses in Contrast-Enhanced MRI using Deep Learning Methods. Master's thesis. University of Guelph; 2021.
4. Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal. 2019;53:197–207. [PubMed: 30802813]
5. Heller N, Isensee F, Maier-Hein KH, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. Med Image Anal. 2020;67:101821. [PubMed: 33049579]
6. Heller N, Sathianathen NJ, Kalapara AA, et al. The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445, 2019.
7. Isensee F, Jaeger PF, Kohl SAA, Petersen JP, Maier-Hein KH. nnU-Net: self-adapting framework for u-net-based medical image segmentation. Bildverarbeitung für die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 17.bis 19.März 2019 in Lübeck.Springer Fachmedien Wiesbaden, 2019
8. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Vol 2. Springer; 2009.
9. Suárez A, Lutsko JF. Globally optimal fuzzy decision trees for classification and regression. IEEE Trans PAMI. 1999;21:1297–1311.
10. Frosst N,Hinton G.Distilling a neural network into a soft decision tree. CEUR Workshop Proceedings, 2017.
11. Hazimeh H, Ponomareva N, Mol P, Tan Z, Mazumder R. The tree ensemble layer: differentiability meets conditional computation. In: ICML. 2020:4138–4148.
12. Kontschieder P, Fiterau M, Criminisi A, Bulo SR. Deep neural decision forests. In: ICCV. IEEE; 2015:1467–1475.
13. Lee G-H, Jaakkola TS. Oblique decision trees from derivatives of ReLU networks. In: ICLR. 2019.
14. Tanno R, Arulkumaran K, Alexander D, Criminisi A, Nori A. Adaptive neural trees. In: ICML. 2019:6166–6175.
15. Yang Y, Morillo IG, Hospedales TM. Deep neural decision trees. ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden, 2018.
16. Zantedeschi V, Kusner M, Niculae V. Learning binary decision trees by argmin differentiation. In: International Conference on Machine Learning. PMLR; 2021:12298–12309.
17. Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. Magn Reson Med. 1995;34:910–914. [PubMed: 8598820]
18. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.
19. Buda M, Saha A, Mazurowski MA. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput Biol Med. 2019;109:218–225. [PubMed: 31078126]
20. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR; 2015:448–456.

21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, 2015.
22. Kingma DP, Ba J. Adam: a method for stochastic optimization. International Conference on Learning Representations, 2015.
23. Heimann T, van Ginneken B, Styner MA, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging. 2009;28:1251–1265. [PubMed: 19211338]
24. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. Workshop at International Conference on Learning Representations, 2014.
25. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. European Conference on Computer Vision. 2014:818–833.
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM:visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. 2017:618–626.

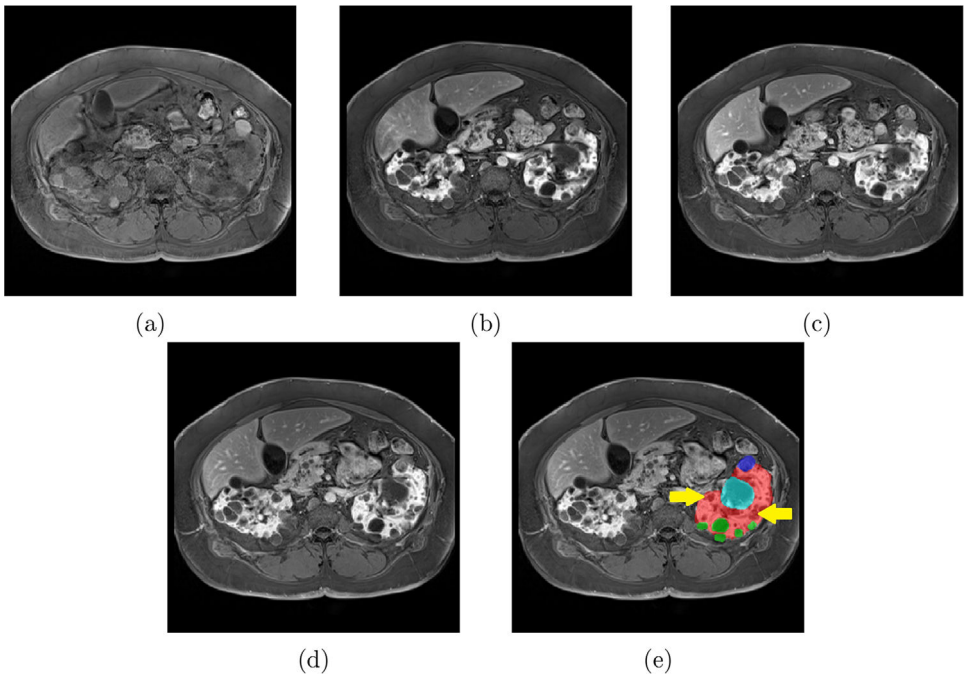


FIGURE 1.

Examples of four T1W image phases used as inputs. Panels (a)–(d) are pre-contrast, corticomedullary (20 s), nephrogenic (70 s), and excretory (3 min) T1W images, respectively. Panel (e) is the ground truth segmentation overlaid on the excretory image. The red, green, blue, and cyan colors correspond to kidney parenchyma, cyst, RCC lesion, and unknown pathology, respectively. Yellow arrows point out examples of unannotated cysts that are thresholded out and ignored during training. RCC, renal cell carcinoma.

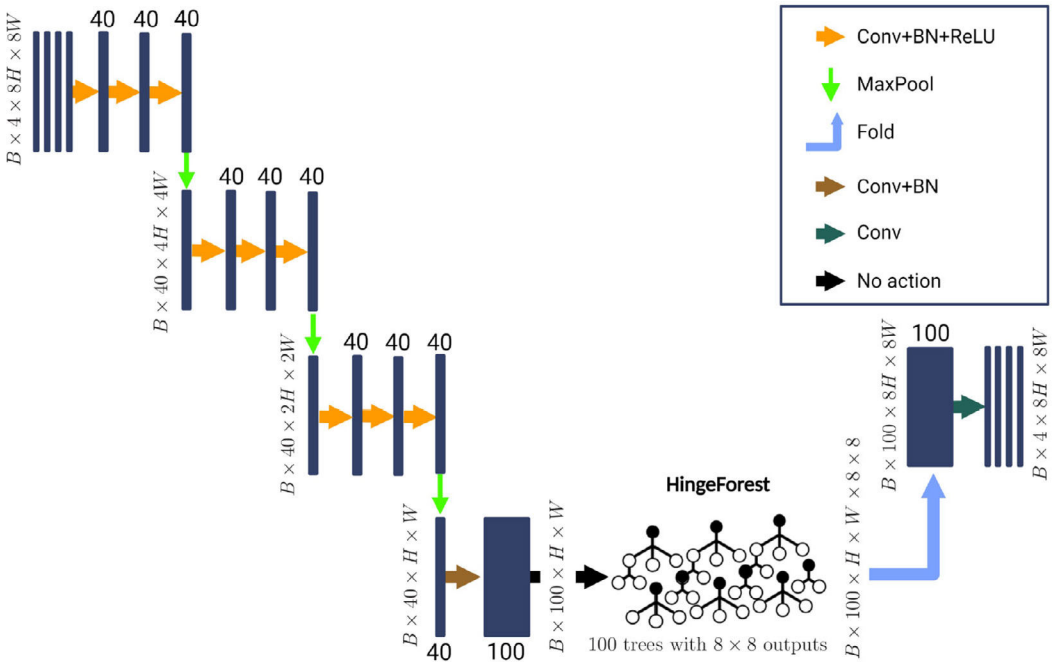


FIGURE 2.

Hinge forest segmentation architecture used in all experiments. B is the batch size and $8H \times 8W$ is the input image size. The image size must be a multiple of 8 owing to the max pool operations progressively downsampling activation maps and the subsequent upsampling by the hinge forest. Orange arrows are 3×3 convolutions with stride and padding 1. The brown and darkgreen arrows are 1×1 convolutions with stride 1 and no padding.

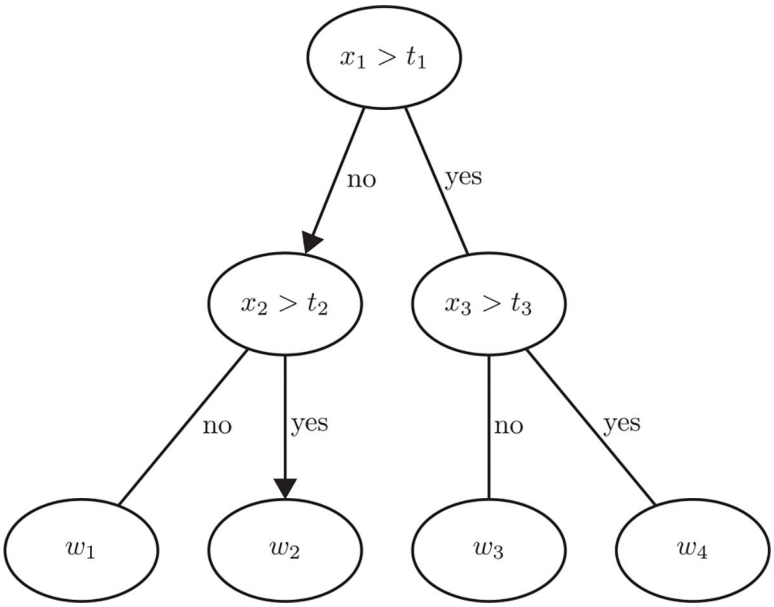
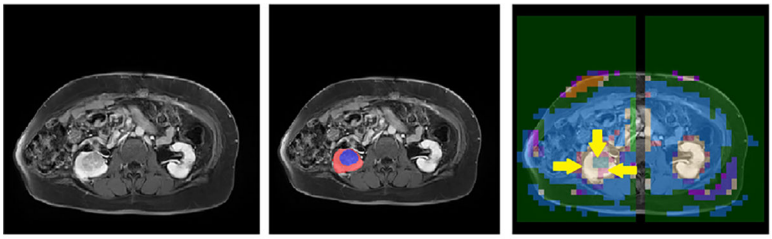


FIGURE 3.

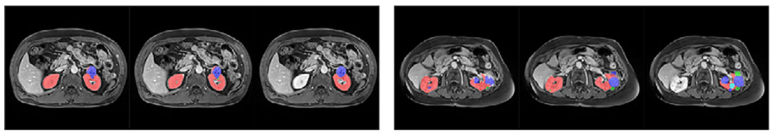
Simplified example of a depth 2 decision tree. The internal vertices have threshold conditions for a given feature vector x that result in yes/no answers. Following these conditions lead to leaf vertices that predict some arbitrary value like ω_1. The arrows from the root vertex to leaf vertex ω_2 correspond to the condition $x_1 \leq t_1 \wedge x_2 > t_2$ and results in a prediction of $\omega_2 I(x_1 \leq t_1)/(x_2 > t_2) = \omega_2$. When this condition is true, all other leaves predict 0 since the indicator function of other conditions is 0. The corresponding hinge tree condition is identical to $\min\{\text{ReLU}(t_1 - x_1), \text{ReLU}(x_2 - t_2)\}$ and results in a decision margin-weighted prediction of $\omega_2 \min\{\text{ReLU}(t_1 - x_1), \text{ReLU}(x_2 - t_2)\}$. When this condition is nonzero, the other leaves predict 0 since the ReLU of other conditions is 0.



(a) (b) (c)

FIGURE 4.

An example of a leaf map from one tree on a test case. Panel (a) is the excretory contrast image and panel (b) is the ground truth segmentation (red is kidney parenchyma and blue is RCC lesion). Panel (c) is a leaf map and the colors represent tree leaves. The white region consistently coincides with kidney parenchyma across images while green and red regions consistently coincide with RCC lesion across images. The pixelation is due to the hinge forest operating on downsampled activation maps. RCC, renal cell carcinoma.



(a) (b)

FIGURE 5.

Examples of the best and average hinge forest, U-Net test set segmentations, and ground truths, respectively, on the first random split overlaid on the corresponding excretory T1w contrast image. The red, green, blue, and cyan colors correspond to kidney parenchyma, cyst, RCC lesion, and unknown pathology, respectively. The mask-based tumor DSCs for hinge forest and U-Net, respectively, are (a) 0.86, 0.74 (b) 0.56, 0.36. DSC, Dice similarity coecient; RCC, renal cell carcinoma.

Med Phys. Author manuscript; available in PMC 2023 November 28.

TABLE 1

Average test set mask dice similarity coefficient (DSC), tumor DSC, average symmetric surface distance (ASSD), Hausdorff distance (HD), and area under ROC curve (AUC), respectively, over all scans and over all 10 random splits.

Structure	Mask DSC	DSC	ASSD (mm)	HD (mm)	AUC	Total #	Missed #	Extra #
Hinge forest			0.97M # Params			≈ 65-h training time		
Kidney	0.75 (0.03)	**0.82** (0.01)	**1.59** (0.12)	**16.10** (1.41)	-	334	90	292
Cyst	**0.44** (0.05) •	0.61 (0.05)	4.03 (1.66)	13.15(3.77)	-	2823	1738	**2700**
Tumor	**0.53** (0.04) •	0.56 (0.07)	8.59 (3.38)	29.06 (7.28)	**0.68** (0.07)	773	**397**	**2062**
U-Net			7.76M # Params			≈ 66-h training time		
Kidney	**0.78** (0.02)	0.81 (0.02)	1.98 (0.64)	17.84 (1.73)	-	334	**67**	**238**
Cyst	0.41 (0.04)	**0.63** (0.05)	**4.01** (1.87)	**12.81** (4.27)	-	2823	**1636**	4109
Tumor	0.46 (0.05)	**0.61** (0.07)	**5.35** (2.90)	**20.88** (7.76)	0.57 (0.09)	773	552	2128

Standard deviations are in parentheses. Total, missed, and extra fields are a count of total ground truth connected components, missed ground truths, and false positive connected components, respectively, over all scans over all 10 random splits. Ground truth connected components were considered *missed* if the incident segmentation connected components overlapped the ground truth by less than 50%.Bold text indicates best score while

•indicates paired *t*-test statistical significance ($\alpha < 0.01$).

TABLE 2

Top decision rules that describe > 50% of voxels occurring in each structure on the training set.

Structure	Gini	Ratio explained	Leaf rule
Kidney	0.24	0.76	$/(\mathbf{x}_{54} \leq -0.35) \wedge /(\mathbf{x}_{10} \leq -1.22) \wedge /(\mathbf{x}_{21} > -0.65) \wedge /(\mathbf{x}_{51} \leq 2.01) \wedge /(\mathbf{x}_{69} \leq 2.40) \wedge /(\mathbf{x}_{27} \leq 1.45) \wedge /(\mathbf{x}_{42} \leq 2.66)$
Cyst	0.28	0.46	$/(\mathbf{x}_{73} > -3.65) \wedge /(\mathbf{x}_{55} > -0.52) \wedge /(\mathbf{x}_{62} \leq -1.10) \wedge /(\mathbf{x}_{25} \leq 1.77) \wedge /(\mathbf{x}_{63} \leq 0.08) \wedge /(\mathbf{x}_{82} \leq -0.49) \wedge /(\mathbf{x}_{23} > -2.18)$
Tumor	0.83	0.39	$/(\mathbf{x}_{0} > 1.49) \wedge /(\mathbf{x}_{64} > -0.57) \wedge /(\mathbf{x}_{54} \leq 1.24) \wedge /(\mathbf{x}_{48} > -0.12) \wedge /(\mathbf{x}_{94} > -1.32) \wedge /(\mathbf{x}_{32} \leq -0.82) \wedge /(\mathbf{x}_{90} \leq -0.14)$

The Gini impurity and ratio explained reflect label impurity and proportion of voxels described for each structure on the test set.