Evaluating Uncertainty-Based Deep Learning Explanations for Prostate Lesion Detection

Christopher M Trombley CHRISTOPHER.TROMBLEY@LOUISVILLE.EDU
Department of Computer Science Engineering
University of Louisville
Louisville, KY, USA

Mehmet Akif Gulum MEHMETAKIF.GULUM@LOUISVILLE.EDU
Department of Computer Science Engineering
University of Louisville
Louisville, KY, USA

Merve Ozen MERVE.OZEN@KENTUCKY.EDU
Department of Radiology
University of Kentucky
Lexington, KY, USA

Enes Esen MERVE.OZEN@KENTUCKY.EDU
Department of Radiology
Canakkale Onsekiz Mart University
Çanakkale, Turkey

Melih Aksamoglu MERVE.OZEN@KENTUCKY.EDU
Department of Radiology
Gaziantep University
Gaziantep, Turkey

Mehmed Kantardzic MEHMED.KANTARDZIC@LOUISVILLE.EDU
Department of Computer Science Engineering
University of Louisville
Louisville, KY, USA

Abstract

Deep learning has demonstrated impressive accuracy for prostate lesion identification and classification. Deep learning algorithms are considered black-box methods therefore they require explanation methods to gain insight into the model's classification. For high stakes tasks such as medical diagnosis, it is important that explanation methods are able to estimate explanation uncertainty. Recently, there have been various methods proposed for providing uncertainty-based explanations. However, the clinical effectiveness of uncertainty-based explanation methods and what radiologists deem explainable within this context is still largely unknown. To that end, this pilot study investigates the effectiveness of uncertainty-based prostate lesion detection explanations. It also attempts to gain insight into what radiologists consider explainable. An experiment was conducted with a cohort of radiologists to determine if uncertainty-based explanation methods improve prostate lesion detection. Additionally, a qualitative assessment of each method was conducted to gain insight into what characteristics make an explanation method suitable for radiology end use. It was found that uncertainty-based explanation methods increase lesion detection perfor-

mance by up to 20%. It was also found that perceived explanation quality is related to actual explanation quality. This pilot study demonstrates the potential use of explanation methods for radiology end use and gleans insight into what radiologists deem explainable.

1. Introduction

Deep learning systems for lesion detection and classification have achieved impressive accuracy as of late Alkadi et al. (2018) Schelb et al. (2019) Vente et al. (2021). This accuracy demonstrates the potential for these systems to be used as a clinical tool for radiologists. However, accuracy alone does not make a deep learning model suitable for clinical implementation. An aspect to fully realize clinical implementation is explainability of deep learning models. Deep learning algorithms are considered black-box algorithms. This means that their inner workings are not easily interpretable like a clear-box model such as a decision tree. This is an area that has attracted high interest recently and thus many methods have been proposed to interpret black-box deep learning algorithms Ribeiro et al. (2018) Lundberg and Lee (2017). However, it still unclear which explanation methods perform best in a clinical context as well as what characteristics of the explanation methods are important to clinicians.

For the sake of this manuscript, explanations are defined as methodologies that demonstrate why a model made a certain classification. Explanation methods are important for a multitude of reasons. First, a classification by itself is not necessarily insightful for a clinician. Explanation methods provide more insight into why a classification was made which can be used to improve diagnosis. Second, the explanation methods provides a sense of trust, or assurance, between the user and the deep learning system. If the user is able to see why the classification was made and affirm that the explanation is sensible then there is a higher level of assurance that the classification is correct. Third, there are legal and ethical considerations when implementing a deep learning system in a high stakes environment. An example of the legal requirements if GDRP "right to explanation" which states the explanations for deep learning systems are legally mandated Selbst and Powles (2017). The importance of this is magnified for medical tasks such as cancer diagnosis due to the complexity of the task and high risk nature.

It has been demonstrated that an explanation method by itself may not necessarily be sufficient for clinical tasks. Additional characteristics of explanation methods that have been deemed important by clinicians include feature importance, demonstrating why the model falls short, and quantifying the uncertainty Tonekaboni et al. (2019). This study focuses on explanation methods that include a uncertainty quantification in the explanations. Uncertainty quantification is important for clinical tasks due to the high-risk nature. It is important for clinicians to know what components of the explanation the model is less sure of. This can reduce diagnosis error and improve the overall diagnosis pipeline. While there are studies that evaluate explanation methods with both clinical and non-clinical populations, this is among the first that evaluate explanation methods that quantify uncertainty to the best of our knowledge.

The goal of this study was to investigate the effectiveness of uncertainty-based explanation and gain insight into what characteristics are deemed explainable by radiologists. To this end, we conduct an exploratory pilot study with a cohort of radiologists to 1) Evaluate

if uncertainty-based explanation methods improve lesion detection accuracy 2) Determine if uncertainty-based explanations are consistent with radiologist explanations and 3) Garner insight into what characteristics radiologists deem explainable.

Generalizable Insights about Machine Learning in the Context of Healthcare

We investigate if uncertainty-based deep learning explanation improve radiologist's performance detecting prostate lesions. We also analyze which methods perform the best and why some perform better than others. In addition, one of the goals of this study was to gain insight into what a radiologist deems explainable with the context of prostate lesion detection. Our contribution can be summarized as follows:

- It is demonstrated that explanation methods in tandem with radiologists improve lesion recognition compared to various baselines.
- Each explanation method was evaluated in terms of perceived confidence, understanding, and justification. Our results suggest that perceived usefulness is related to actual usefulness.
- We provide insight into what characteristics radiologists deem to improve explainability. This insight augments our understanding of explanation methods with the potential to guide future efforts developing explanation methods

In short, we provide insight into why some explanation methods perform better than others and what characteristics radiologists deem explainable within the context of prostate lesion detection.

2. Study Methodology

We decided on the methodological approach aprior when designing the study protocol to reduce bias while conducting the study and during analysis. Our goal is to evaluate the clinical effectiveness of uncertainty-based explanation methods and determine what radiologists deem explainable. To this end, we conduct a study where we show a cohort of radiologists a series of MRI images with and without explanation methods. We measure the lesion detection performance for each case for comparison. For the MRI images provided without an explanation method, we ask the radiologist to mark the area of the image that contributes the most to their decision for later comparison. We also ask questions using a 5-point Likert scale about each explanation method to determine perceived effectiveness in terms of understanding, confidence, and justification.

2.1. Data Preprocessing

We used the PROSTATEx dataset Litjens et al. (2014) for this study. The PROSTATEx dataset consists of 330 lesions from 204 subjects. The dataset provides T2 transverse, sagittal, and coronal MRI images along with ADC, BVAL, and KTRANS. For this study, we utilize T2 images that are cropped 120x120 around the centroid of the prostate. The MRI image intensities were normalized during preprocessing and the image was cropped

120x120 pixels around the prostate. The dataset provides coordinates of the lesion centroid and a label indicating if the lesion is clinically significant or not. Recently, ground truth masks have been provided for the lesion and prostate zones Cuocolo et al. (2021). The ground truth masks were used during the learning period before the study. This dataset was split 70/30 into training and testing sets. The training set was used to train the model used in the study. The testing set was used for performance evaluation and to provide the study cohort images during the experiment.

2.2. Deep Learning Model

We learn a mapping from a T2-weighted MRI image to a binary variable representing the presence of a lesion in the image. To learn this mapping, the training set described above was used to train a convolutional neural network. We trained the model based on the VGG architecture Simonyan and Zisserman (2015) using Python 3.6 and PyTorch 1.8.0. Grid-search was utilized to select the optimal parameters of the model. Grid search is a hyperparameter selection method that iteratively tries different hyperparameter combination and selects optimal parameters based on model performance. After the parameters of models were selected using grid search, the model was tested on the testing set. The model achieves an AUC of 0.87, sensitivity of 0.85, and specificity of 0.88.

2.3. Explanation Methods

In this section, we provide an overview of the methods used during the study. Table 1 provides a brief summary of each method. Figure 1 shows an example of each uncertainty-based explanation method. The top row shows the input image, the middle row shows the explanation, and the bottom row shows the uncertainty. For BLRP, the parameters of the middle and bottom row are $\alpha = 5$ and $\alpha = 50$ respectively.

2.3.1. CXPlain

CXPlain Schwab and Karlen (2019) frames the problem of providing deep learning explanations as a causal learning task. To this end, the authors train causal explanation models to estimate the degree certain inputs cause outputs. CXPlain enables uncertainty quantification with its feature importance via bootstrap ensembling. The authors empirically demonstrate that the uncertainty estimates are strongly correlated with their ability to accurately estimate feature importance on unseen data.

2.3.2. DistDeepSHAP

Distribution DeepSHAP (DistDeepSHAP) Li et al. (2020) provides uncertainty estimations for the SHAP explanation framework. DistDeepSHAP samples the references from a distribution and calculates Shapely values for these references. To estimate the uncertainty of SHAP explanations, the authors compute a confidence interval for each assigned Shapely value using the training dataset as a reference distribution.

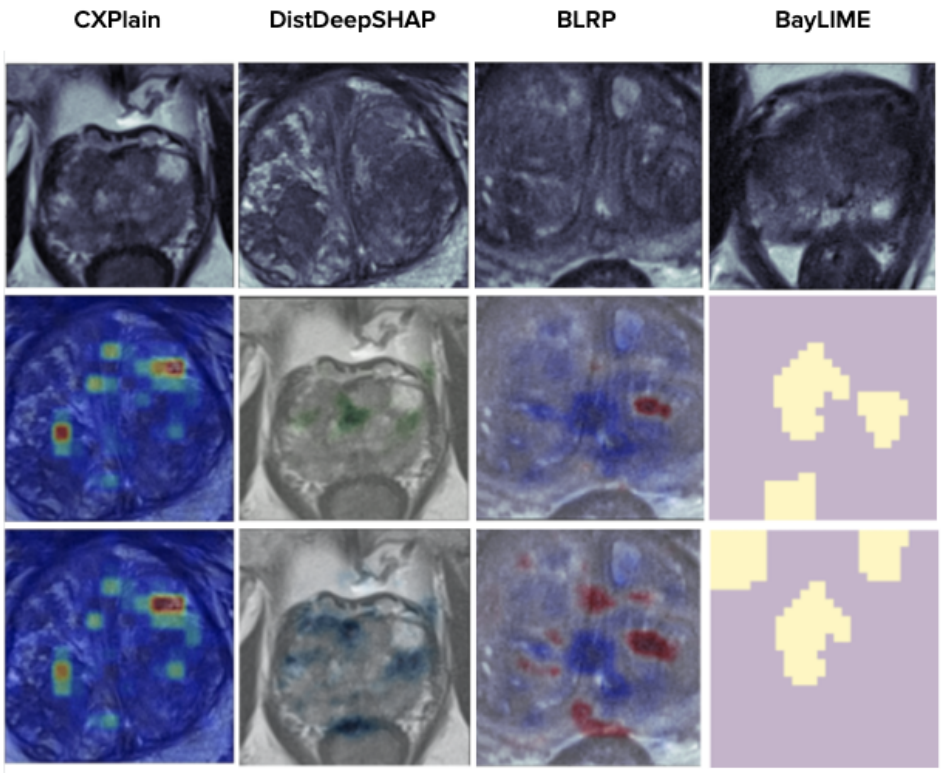

Figure 1: An overview of the uncertainty-based explanation methods used in this study.

2.3.3. BLRP

BLRP Bykov et al. (2020) is a bayesian formulation of Layerwise Relevance Propagation (LRP). It provides explanation visualization at different levels of certainty. A certainty level $\alpha = 5$ represents the most reliable relevant pixels wheresas a certainty level of $\alpha = 95$ represents the least reliable relevant pixels. In this study, we show the radiologists alpha values of 5, 25, 50, 75, and 95. An explanation of what the alpha values represent was conducted during the prestudy learning period task.

2.3.4. BayLIME

BayLIME Zhao et al. (2021) is a bayesian extension to the LIME explanation framework. BayLIME makes use of bayesian regressors as the local surrogate models as opposed to the linear models use in the original LIME framework. The authors show that the bayesian component helps improve the consistency of repeated explanations and explanation fidelity. The bayesian nature of BayesLIME enables uncertainty visualization through the variation of repeated explanations.

Table 1: An overview of the methods being evaluated in this work.

Method	Description
CXPlain	Uses causal models to provide deep learning explanations
DistDeepSHAP	Uses a reference distribution to compute Shapely values
BLRP	A bayesian extension of layer wise relevance propagation
BayLIME	Bayesian regression models are used in the LIME framework

2.4. Radiology Cohort

Ten radiologists were recruited for this study. Each radiologist holds at least a MD and is board certified in Radiology. None of the radiologists in this study had previous experience with machine learning or machine learning explanation methods. Each evaluation was performed within two weeks in a controlled, standardized online environment. Before the evaluation, an interview was held with each radiologist. During the interview, the scripted experimental procedure was described to the radiologist. This included a brief overview of each method and the instructions for the experiment. These were read directly off a pre-made script to standardize the instructions. A pre-study learning period was also administered. This learning period consisted of showing the radiologist a series of T2-weighted MRI images, corresponding explanation methods, and the ground truth label and mask to familiarize the radiologists with the explanation methods.

2.5. Study Procedure

Clinical user studies are important to evaluate the clinical effectiveness of deep learning explanation methods. However, designing such a study is a challenging task due in part to designing the experiment to answer questions such as, *Does the explanation improve clinical diagnosis? Does the explanation hinder diagnosis? Are the explanations consistent? Is the explanation clear?*

We formulate our study as a lesion detection task whereby a MRI image and corresponding explanation method is presented to a radiologist and the radiologist is asked to determine if there is a lesion present. We provide a MRI image without an explanation method as a control case for additional comparison. Note the model's predicted label was not given to the radiologist. In addition to the uncertainty-based explanation methods we include two explanations that do not quantify uncertainty. These are Grad-CAM Selvaraju et al. (2017) and Integrated Gradients Sundararajan et al. (2017). These are included to differentiate the performance between uncertainty-based explanation methods and explanation methods that do not provide uncertainty quantification.

Our underlying hypothesis is that if uncertainty-based explanation methods are beneficial for clinical diagnosis tasks then the diagnosis accuracy will improve given the presence of the uncertainty-based explanation method compared to baselines.

We provided 24 unique samples randomly selected from the PROSTATEx test set to each radiologist in the cohort. Each sample has a corresponding explanation produced by one of the explanation methods. There were four samples per explanation method shown to each radiologist. These were counterbalanced therefore the same 24 samples are provided to

each radiologist with the order randomized. In addition to this, we provide each radiologist with five additional randomly sampled MRI images separate from the ones given during the diagnosis task. Therefore, a total of 29 samples were shown to each radiologist. The radiologists were asked to highlight the region of the image that contributes the most to their diagnosis for the five additional images without a corresponding explanation. We use this data to later compare the IOU between the radiologist and uncertainty-based explanation methods. After each MRI and corresponding explanation is shown to the radiologist, we ask the radiologist to state if there is a lesion present, rate the explanation methods, and provide reasoning behind their ratings. We do this for each of the samples that have a corresponding explanation method. For the five MRI samples without a corresponding explanation, we only ask the radiologist to state if there is a lesion and mark the region of the image they contribute the most to their diagnosis.

Note that an IRB was not required for this study since the IRB recognizes that the analysis of de-identified, publicly available data does not constitute human subjects research as defined at 45 CFR 46.102 and that it does not require IRB review.

2.6. Questionnaire Development

A questionnaire was included with the experiment to measure the perceived effectiveness of each explanation method. The goal of the questionnaire was to collect qualitative data on each method and provide insight into what characteristics improve explainability according to the radiologist cohort. An existing validated survey for uncertainty based explanation methods was not found in the literature therefore we were forced to design our own questionnaire. The questionnaire was development by a multidisciplinary team including a machine learning researcher holding a PhD, a clinical radiologist holding a MD, and a radiology researcher holding a PhD and MD. This was to improve the validity of the questionnaire and to include perspectives from different fields. The questionnaire design was largely influence by Krosnick and Presser (2009). After the questionnaire was developed, it was approved by two radiologists to affirm the clarity and effectiveness of the questionnaire. We also wanted to make sure the questionnaire was able to be easily understood by the cohort. After rounds of feedback from the radiologist and the corresponding changes, the questionnaire was approved by each member of our multidisciplinary team.

The questions were selected by our team to evaluate the explanation methods using a 5-point Likert scale along three dimensions of interest influenced by Ehsan et al. (2019).

1. **Understanding** The explanation helps me understand *why* the classification was made.
2. **Confidence** The explanation gives me confidence the classification was accurate.
3. **Justification** The explanation provides sufficient justification for the classification.

The scale was from 1-5 with 1 being strongly disagree and 5 being strongly agree. This was shown after each method during the experiment. Then, in a mandatory text field each radiologist was prompted to provide reasoning behind why they rated each method the way they did. The three dimensions were select by the multidisciplinary team. They were chosen because they were considered to be important characteristics of explanation methods. The

Table 2: F1 score for radiologist, explanation, and uncertainty.

Method	F1 Score [95% CI]
Radiologist	58.82 [46.24, 64.3]
Radiologist + Explanation Technique	63.86 [53.45, 68.41]
Radiologist + Explanation Technique + Uncertainty	**66.81** [62.22, 74.36]

explanation method should provide a level understanding so the radiologist can understand why the classification was made. The explanation should also provide confidence that the model is either correct or incorrect. Lastly, the explanation method should provide a level of justification for why the classification was made.

3. Results

We report the results from 290 total responses from ten radiologists. The results section is organized in the following way. First, we investigate the clinical effectiveness of the explanation methods. Second, we evaluate explanation methods in terms of understanding, confidence, and justification. Third, we measure the consistency between the ROI highlighted by the explanation method and the radiologist. Lastly, we report our qualitative results from the questionnaire.

3.1. Do explanation methods improve diagnosis performance?

We report the lesion recognition performance with explanation methods and compare it to two baselines. The first baselines are two explanation methods that do not quantify uncertainty. This was to compare the marginal benefit of the uncertainty component. The second baseline was a MRI image without a corresponding explanation technique. This was done to determine if the presence of the explanation techniques improve lesion recognition. The results for this are shown in Figure 2.

Table 2 shows that an explanation technique in tandem with the MRI improves F1 score compared to only the MRI. It also shows, there is an additional increase in F1 score when the uncertainty component added. One radiologist noted that, in some cases, they did not see a lesion when first analyzing the MRI. However, after looking at the explanation method they were able to find a lesion in the MRI that they did not see before. Figure 2 shows that the sensitivity increase was marginal. The radiologist seem to error on the side of false positives without the explanation techniques. With the explanation technique, the true positives were better detected as well as the true negatives. It is important to note, we did not see a tradeoff here between true positive detection and true negative detection. The radiologists also noted that the uncertainty component enabled them to better determine the signal from noise if the explanation method happened to highlight multiple areas of the prostate.

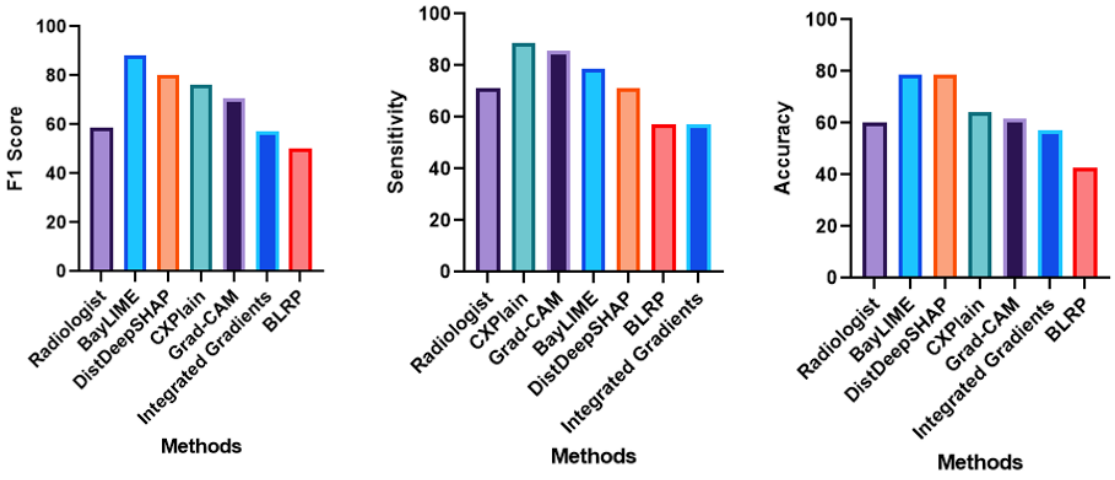

Figure 2: F1 score, sensitivity, and accuracy of each method and baseline.

3.2. Do explanations provide understanding, confidence, and justification?

We measure the perceived understanding, confidence, and justification for each explanation. We show each radiologist a MRI image and corresponding explanation technique. The radiologist is then asked from 1-5 the following questions with 5 being the positive end and 1 being the negative end. For understanding, *This visualization helped me understand why the model's classification was made.* For confidence, *This visualization makes me feel confident in the model's classification.* For justification, *This visualization adequately justifies the classification.* These results are reported in Table 3 and Figure 3. Table 3 shows the method(s) with the highest reported scores. For understanding, this was BayLIME with a score of 4.0. For confidence, these were BayLIME and BLRP with scores of 3.71. For justification, this was BayLIME with a score of 3.71. The radiologists stressed the importance of clarity for understanding. If the explanation method was noisy the radiologists rated the method lower for understanding. The radiologists noted that for confidence, it was important that the explanation method accurately segment the lesion when there was a lesion present. Methods that segmented the lesion but were not accurate were rated lower on average than methods that accurately segmented the lesion. For example, if the explanation method highlighted only half of the lesion this was considered less confident than if the method highlighted the whole lesion. Justification was similar to confidence with the major difference being that the radiologist cared more about the explanation method highlighting the lesion in general rather than an accurate segmentation. If the explanation method highlighted only half the lesion this was considered just as justified as highlighting the whole lesion. For justification, the radiologist stressed the importance of being able to look at the explanation method to confirm the lesion presence. These results along with the previous subsection suggest that lesion detection performance has a relation with perceived explanation quality in our sample. BayLIME produces the highest lesion detection accuracy and perceived quality in terms of understanding, justification, and confidence.

Table 3: Results for perceived understanding, confidence, and justification.

Method	Understanding [95% CI]	Confidence [95% CI]	Justification [95% CI]
CXPlain	3.28 [2.63, 3.93]	3.42 [2.59, 4.26]	3.14 [2.10, 4.18]
BayLIME	**4.0** [3.50, 4.49]	**3.71** [2.89, 4.52]	**3.71** [3.06, 4.36]
BLRP	3.85 [3.08, 4.26]	**3.71** [2.89, 4.52]	3.42 [2.33, 4.51]
DistDeepSHAP	3.57 [2.89, 4.24]	3.28 [2.63, 3.93]	3.14 [2.10, 4.18]

Table 4: IOU score for explanation, and uncertainty.

Method	IOU [95% CI]
Explanation Technique	61.36 [51.45, 70.41]
Explanation Technique + Uncertainty	**70.24** [61.22, 79.36]

3.3. Are explanations consistent with radiologist's explanations?

To measure the consistency of explanations, we ask each radiologist to mark the ROI in the image that contribute the most to their classification. To compare the similarity between the radiologist's explanation and the explanation method, we compute the IOU between the two. The results are reported in Table 4. This shows that uncertainty-based explanations are the most consistent with the radiologist explanations. The radiologists, on average, marked more areas than the explanation methods. Common feedback from the radiologist was that they preferred an explanation that showed a ROI that was consistent with theirs. They mentioned this made the explanations method more explainable and they were able to better interpret it.

3.4. What characteristics improve explainability for radiologists?

In this section, we investigate the qualitative responses from the radiologists to better understand what characteristics improve explainability. This insight augments our understanding of explanation methods with the potential to guide future efforts developing explanation methods. We analyze the textual responses from the questionnaire using thematic analysis J.A. (1993). The research team codified the responses from the radiologist. The codified items were then grouped together under common themes. These themes then formed the underlying characteristics for three dimensions. The research team eventually settled on the following characteristics (1) Precision (2) Clarity (3) Presentation (4) Consistency. These are described in Table 5.

3.4.1. Understanding

Understanding aimed to gauge how much the explanation helped the radiologist understand why the classification was made. With respect to clarity, a noisy heatmap was noted to hinder understanding. However, the radiologist also noted that some noise was acceptable as long as the noise was meaningful. They mentioned that they checked each region the explanation highlighted which helped them find lesions that they may have missed at first. This was one example of the noise being helpful as long as it is meaningful. The presentation

Table 5: Major themes from the qualitative analysis.

Component	Description
Precision	Precisely localizes a lesion if one is present
Clarity	The amount of noise in the method
Presentation	The way the method is presented (e.g. color, overlay, etc)
Consistency	The explanations are consistent with the radiologists' explanations

was noted to be related to understanding. We received feedback stating that a heatmap that uses a gradient presentation such as CXPlain was preferred over a heatmap without a gradient presentation such as BayLIME. The radiologist mentioned that they used the red, yellow, green, and blue parts of the gradient to determine what areas were considered more important for the region highlighted by the explanation method. Regarding precision, the radiologist stated that if the explanation highlighted a lesion then there was almost immediate understanding why the classification was made. A common theme was the radiologist mentioning that consistency between their explanations and the explanation method explanations was important to garner increased understanding.

3.4.2. Confidence

Confidence measured how much confidence in the classification the explanation instills. Regarding precision, the radiologists mentioned that what gave them the most confidence was when the deep learning model classified the MRI as containing a lesion and the explanation method highlighted the lesion. If they were able to see a lesion and then check the explanation method also highlighted that lesion this was noted to give them a high degree of confidence that there is indeed a lesion in that area. With respect to consistency, they mentioned that if the explanation method highlighted the zone the lesion was in that this was almost as sufficient as highlighting the lesion itself. They wanted to be able to use the explanation method to verify the classification.

3.4.3. Justification

Justification evaluated the degree the explanation method justified the classification. The radiologist mentioned the presentation was important for justification. They considered it a higher level of justification if the red part of the heatmap highlighted the lesion. If the lesion was highlighted by the heatmap but not the red part of the heatmap they considered it less justified. With respect to precision, the radiologist consider the justification to be the greatest if the lesion was highlighted by the heatmap. It was noted that the explanation methods often highlighted Prostatistis instead of the lesion. This was considered to be poor justification. The radiologist wanted the justification to be presented similar to the way they justify their diagnosis. For example, highlighting the prostate zone along with the lesion was looked upon favorably.

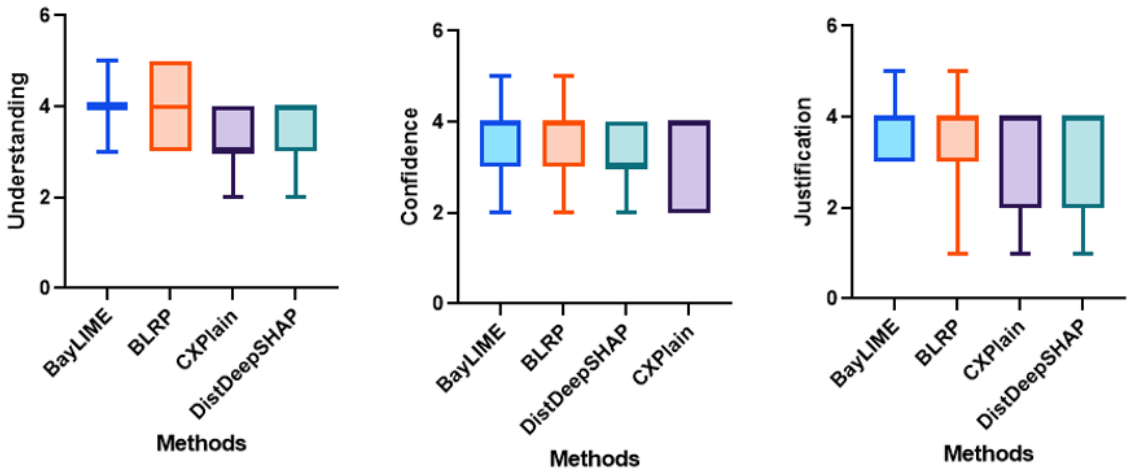

Figure 3: Graphs summarizing the perceived usefulness of each method.

4. Related Work

To the best of our knowledge, this study is the first to investigate uncertainty-based explanation methods for radiology end use. However, there are similar studies that investigate explanation methods for other clinical tasks Gulum et al. (2021). Tonekaboni et al. (2019) attempt to gain insight into what characteristics of explanation method increase clinician trust. They find clinicians deem feature importance, uncertainty, and transparency to be important components. Hegselmann et al. (2020) evaluate the interpretability of additive models. Specifically, their goal was to determine if doctors can interpret these models. They found that overall doctors were able to accurately interpret the additive models.

There have been studies that aim to translate machine learning to clinical practice. Elish (2018) examines the development of a sepsis risk machine learning model in an ER setting. Turk et al. (2016) conduct a pilot study that investigates mortality prediction using electronic medical records in a clinical setting. The authors evaluate performance and highlight some of the challenges of clinical implementation. Mazur et al. (2016) access the relationship between task demands and workload during physician-computer interaction in a simulated study. Ghassemi et al. (2018) propose a visualization system for electronic health records and evaluate accuracy and confidence in treatment decisions using a clinical cohort.

There have also been other human studies investigating explanation methods in a general population. Lertvittayakumjorn and Toni (2019) evaluate explanation methods for text classifications. They conduct a human study to determine which methods are most effective. They find LIME Ribeiro et al. (2016), LRP Binder et al. (2016), and DeepLIFT Shrikumar et al. (2017) perform the best overall. Jeyakumar et al. (2020) investigates explanation methods for multiple data modalities using a Mechanical Turk powered human study. The

authors find that overall prototype methods such as ExMatchina Chen et al. (2019) are the preferred explanation.

5. Discussion

This study provides insight into what uncertainty-based explanation methods are preferred by radiologists and reasoning behind why. It also investigates if uncertainty-based explanation methods improve detection of prostate lesions. It was found that BayLIME improved lesion detection the greatest up to a 20% increase in F1 score. On average, uncertainty-based explanation methods improved lesion detection by 8% and explanation methods without an uncertainty component increase lesion detection up to 5%. This demonstrates the potential of explanation methods to provide utility in diagnosis. It was noted by the radiologists that there were cases when they analyzed the MRI and did not find a lesion. However, after looking at the explanation they were to identify the lesion they previously did not see. This was mostly in cases where the lesion was small and difficult to see. The radiologists mention that the explanation methods are useful to use to flag certain regions for further inspection. The methods that performed the best qualitatively clearly localize a lesion, or the zone the lesion was located in. The methods that performed the worst were noisy and more difficult to interpret. This demonstrates that uncertainty-based deep learning explanation methods have the potential to improve overall diagnosis performance.

It was found BayLIME had the highest justification score with radiologists noting that it often showed where the lesion was located. However, it was noted that the BayLIME visualization style was the least preferred. The visualization style of CXPlain was preferred over the others. CXPlain also had the highest confidence, tied with BayLIME. BayLIME also had the highest understanding score overall. CXPlain and DistDeepSHAP had the lowest justification scores. CXPlain had the lowest understanding score. It was noted that this method marked many areas of the prostate that did not include relevant information. This was one of the common comments on why these scores were low. However, CXPlain produced the greatest increase in sensitivity. This shows that the extra noise can be useful for checking possible lesion locations but can be hard to interpret and lead to many false positives. Overall, the most common comments from the radiologist are summarized as follows. First, it is important that the explanation method highlights the lesion if there is one. The radiologists mentioned one of components they considered most useful was when the explanation method highlighted a lesion they did not see at first. They mentioned that the explanation method clarity helps understanding, and accurately localizing the lesion helps confidence and justification. The presentation of the explanation method mattered to the radiologists. They mentioned that a gradient-based presentation style such as CXPlain was preferred compared to a presentation style such as BayLIME. They mentioned the colors in the gradient were useful to determine what areas were weighted more so than others. Some explanation methods are notoriously noisy Kim et al. (2019). The radiologists mentioned that noisy methods were difficult to use to draw conclusions. However, they also mentioned that the uncertainty component helped them reason through the noise if present. Lastly, they stressed the importance of providing explanations that are consistent with the way they would provide an explanation. This makes the explanation methods easier to interpret and better able to be used as a useful tool in the diagnosis process.

These comments provide insight into what radiologists deem explainable which can guide the further development of uncertainty-based explanation methods and the evaluation of current methods. In essence, we found that radiologists want explanation methods that are precise, clear, and provide explanations in a manner similar to the way they would.

5.1. Limitations

One limitation of this study is that it only considers prostate cancer using T2 MRI images. Explanation method performance preference may change depending on the image modalities. ADC and diffusion images are also used when analyzing an MRI image. It is important to also investigate these image modalities. We did not include these to keep the study a manageable length for the radiologists and to isolate the conclusions drawn to only T2 images. Another limitation is the sample size. This study is meant to be a pilot study, however a larger sample size would be needed to draw stronger conclusion. Nonetheless, this pilot study provides insight for following studies with larger sample sizes and different image modalities.

6. Conclusion

This study investigates the clinical performance of uncertainty-based explanation methods and attempts to gain insight into what radiologists deem explainable. We perform a pilot study with a cohort of ten radiologists. We ran an experiment where we showed each radiologist an MRI image and corresponding interpretation technique. We also showed them an MRI image without a interpretation for a baseline comparison. It was found that explanation methods increase lesion detection performance up to 20% increase. The highest rated method in terms of confidence, justification, and understanding was BayLIME. However, it was also commonly noted that the BayLIME visualization was the least preferred whereas CXPlain was most preferred. The radiologists noted the most important components of the explanation methods were precision, clarity, presentation, and consistency. This study augments our understanding of what works well with current explanation and where they fall short with insight to guide future research efforts.

References

Ruba Alkadi, Dr. Fatma Taher, Ayman El-Baz, and Naoufel Werghi. A deep learning-based approach for the detection and localization of prostate cancer in t2 magnetic resonance images. *Journal of Digital Imaging*, 32, 11 2018. doi: 10.1007/s10278-018-0160-1.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *ICANN*, 2016.

Kirill Bykov, Marina M-C Höhne, Klaus-Robert Müller, Shinichi Nakajima, and Marius Kloft. How much can i trust you?–quantifying uncertainties in explaining neural networks. *arXiv preprint arXiv:2006.09000*, 2020.

Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. In *NeurIPS*, 2019.

Renato Cuocolo, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, and Massimo Imbriaco. Quality control and whole-gland, zonal and lesion annotations for the prostatex challenge public dataset. *European Journal of Radiology*, 138:109647, 2021. ISSN 0720-048X. doi: https://doi.org/10.1016/j.ejrad.2021.109647. URL https://www.sciencedirect.com/science/article/pii/S0720048X21001273.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. Automated rationale generation: A technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 263–274, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302316. URL https://doi.org/10.1145/3301275.3302316.

Madeleine Clare Elish. The stakes of uncertainty: Developing and integrating machine learning in clinical care. *Ethnographic Praxis in Industry Conference Proceedings*, 2018.

Marzyeh Ghassemi, Mahima Pushkarna, James Wexler, Jesse Johnson, and Paul Varghese. Clinicalvis: Supporting clinical task-focused design evaluation, 10 2018.

Mehmet A. Gulum, Christopher M. Trombley, and Mehmed Kantardzic. A review of explainable deep learning cancer detection models in medical imaging. *Applied Sciences*, 11 (10), 2021. ISSN 2076-3417. doi: 10.3390/app11104573.

Stefan Hegselmann, Thomas Volkert, Hendrik Ohlenburg, Antje Gottschalk, Martin Dugas, and Christian Ertmer. An evaluation of the doctor-interpretability of generalized additive models with interactions. In *Machine Learning for Healthcare (MLHC)*, 2020.

Aronson J.A. A pragmatic view of thematic analysis. *Qual Rep*, 2, 11 1993. doi: 10.46743/2160-3715/1995.2069.

Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation

methods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4211–4222. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/2c29d89cc56cdb191c60db2f0bae796b-Paper.pdf.

Beomsu Kim, Junghoon Seo, Seunghyun Jeon, Jamyoung Koo, Jeongyeol Choe, and Taegyun Jeon. Why are saliency maps noisy? cause of and solution to noisy saliency maps. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4149–4157, 2019.

J. A. Krosnick and S Presser. Question and questionnaire design'. handbook of survey research (2nd edition), 2009.

Piyawat Lertvittayakumjorn and Francesca Toni. Human-grounded evaluations of explanation methods for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5198–5208, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1523. URL https://www.aclweb.org/anthology/D19-1523.

Xiaoxiao Li, Yuan Zhou, Nicha C. Dvornek, Yufeng Gu, Pamela Ventola, and James S. Duncan. Efficient shapley explanation for features importance estimation under uncertainty. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 792–801, Cham, 2020. Springer International Publishing.

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in mri. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014. doi: 10.1109/TMI.2014.2303821.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Lukasz Mazur, Prithima Mosaly, Carlton Moore, Elizabeth Comitz, Fei Yu, Aaron Falchook, Michael Eblan, Lesley Hoyle, Gregg Tracton, Bhishamjit Chera, and Lawrence Marks. Toward a better understanding of task demands, workload, and performance during physician-computer interactions. *Journal of the American Medical Informatics Association*, 23:ocw016, 03 2016. doi: 10.1093/jamia/ocw016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), Apr. 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Patrick Schelb, Simon Kohl, Jan Radtke, Manuel Wiesenfarth, Philipp Kickingereder, Sebastian Bickelhaupt, Tristan Kuder, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, Klaus Maier-Hein, and David Bonekamp. Classification of cancer at prostate mri: Deep learning versus clinical pi-rads assessment. *Radiology*, 293:190938, 10 2019. doi: 10.1148/radiol.2019190938.

Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3ab6be46e1d6b21d59a3c3a0b9d0f6ef-Paper.pdf.

Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 12 2017. ISSN 2044-3994. doi: 10.1093/idpl/ipx022. URL https://doi.org/10.1093/idpl/ipx022.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

Sana Tonekaboni, Shalmali Joshi, Melissa McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare (MLHC)*, 2019.

Benjamin Turk, Arona Ragins, Jason Ha, Brian Hoberman, Steven LeVine, Mamuel Ballesca, Vincent Liu, and Patricia Kipnis. Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals. *Journal of Hospital Medicine*, 11:S18–S24, 11 2016. doi: 10.1002/jhm.2652.

Coen de Vente, Pieter Vos, Matin Hosseinzadeh, Josien Pluim, and Mitko Veta. Deep learning regression for prostate cancer detection and grading in bi-parametric mri. *IEEE Transactions on Biomedical Engineering*, 68(2):374–383, 2021. doi: 10.1109/TBME.2020.2993528.

Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. Baylime: Bayesian local interpretable model-agnostic explanations. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 887–896. PMLR, 27–30 Jul 2021.