BJR|Open https://doi.org/10.1259/bjro.20190021

Received: 01 April 2019 | Revised: 18 June 2019 | Accepted: 25 June 2019

Cite this article as:
Luo Y, Tseng H-H, Cui S, Wei L, Ten Haken RK, El Naqa I. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR Open* 2019; **1**: 20190021.

REVIEW ARTICLE

Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling

YI LUO, PhD, HUAN-HSIN TSENG, PhD, SUNAN CUI, LISE WEI, RANDALL K. TEN HAKEN, PhD and ISSAM EL NAQA, PhD

Department of Radiation Oncology, University of Michigan, 519 W William Street, Ann Arbor, MI, USA

Address correspondence to: Dr Yi Luo
E-mail: *yiyiLuo@med.umich.edu;YL1515@gmail.com*

ABSTRACT

Radiation outcomes prediction (ROP) plays an important role in personalized prescription and adaptive radiotherapy. A clinical decision may not only depend on an accurate radiation outcomes' prediction, but also needs to be made based on an informed understanding of the relationship among patients' characteristics, radiation response and treatment plans. As more patients' biophysical information become available, machine learning (ML) techniques will have a great potential for improving ROP. Creating explainable ML methods is an ultimate task for clinical practice but remains a challenging one. Towards complete explainability, the interpretability of ML approaches needs to be first explored. Hence, this review focuses on the application of ML techniques for clinical adoption in radiation oncology by balancing accuracy with interpretability of the predictive model of interest. An ML algorithm can be generally classified into an interpretable (IP) or non-interpretable (NIP) ("black box") technique. While the former may provide a clearer explanation to aid clinical decision-making, its prediction performance is generally outperformed by the latter. Therefore, great efforts and resources have been dedicated towards balancing the accuracy and the interpretability of ML approaches in ROP, but more still needs to be done. In this review, current progress to increase the accuracy for IP ML approaches is introduced, and major trends to improve the interpretability and alleviate the "black box" stigma of ML in radiation outcomes modeling are summarized. Efforts to integrate IP and NIP ML approaches to produce predictive models with higher accuracy and interpretability for ROP are also discussed.

INTRODUCTION

Radiotherapy treatment in general and personalized adaptive radiotherapy (pART) in particular have a promising prospective to improve cancer patients' therapeutic satisfaction.[1,2] However, pART success highly depends upon accurate radiation outcomes prediction. Closely related to computational statistics, machine learning (ML) explores the study design and the construction of computer algorithms to learn from data and make data-driven predictions by employing complex mathematical optimization schemes.[3,4] As more biophysical data become available before and during radiation treatment, the application of ML in radiation oncology will continue to experience tremendous growth, including treatment planning optimization,[5,6] normal tissue toxicity prediction,[7,8] tumor-response modeling,[9,10] radiation physics quality assurance.[11–13] In this paper, we focus on the application of ML approaches in radiation outcome modeling as a case study.

ML is typically classified into supervised learning, unsupervised learning and reinforcement learning methods.[14] While supervised learning handles a set of data containing both the inputs and the labeled outputs, unsupervised learning deals with a set of data with only inputs. Reinforcement learning intends to identify the best actions in a dynamic system by maximizing the cumulative reward. As multilayer neural networks to extract "complex patterns" from large scale of data sets become popular in the era of Big Data, ML approaches can also be categorized into shallow learning and deep learning (DL), where the latter combines data representation with classification/regression tasks in the same framework. As DL has garnered remarkable attention for its capacity to achieve accurate prediction in various fields, there is a growing realization that better explanation of these ML methods is equally desired. While explainability and interpretability have been used interchangeably, we would like to distinguish between them to provide more accurate definition of different ML techniques

in this review. Explainable models can be defined as those that are able to summarize the reasons for the behavior of ML algorithms, gain the trust of users, or allow the user to produce insights into the causes of the algorithm decisions.[15] Essentially, one consensus among recent studies is that explainability based on human understanding is not a monolithic concept, but rather a complex construction. According to the description of Gilpin et al and Ribeiro et al,[15,16] it can be decomposed into several human factors, such as trust, causality, transferability, and algorithm transparency. On the other hand, interpretability can be loosely defined as comprehending what a model did (or might have done) based on the inputs, with the capacity to defend its actions, provide relevant responses to questions, and be audited. Their relationship can be described as that explainable models are interpretable by default and the reverse is not always true. Although interpretability alone is insufficient for the explanation of different ML techniques, it is a necessary first step towards full explainability, and it is employed in this paper to classify existing literature in radiation oncology. In this context, ML approaches can be sorted into interpretable (IP) and non-interpretable (NIP) approaches. In addition to DL, some shallow learning approaches such as support vector machines (SVMs), random forests (RFs), and "shallow" neural network approaches belong to the NIP category. The rest of shallow learning approaches such as generalized linear models (*e.g.* linear regression, logistic regression), linear discriminant analysis, decision trees, Bayesian networks are considered IP ML approaches.

For clinical applications such as radiation outcomes prediction, the accuracy and interpretability of the ML approaches are major concerns. As accurate prediction of the treatment outcomes provides direct guidance to tailor and adapt a treatment plan in cancer therapy, and it is highly essential to use interpretable results for clinical decision-making support. If the goal is to assist physicians and patients reach the best decision, then an ML approach with a good balance between interpretability of the results and accurate predictions is needed to gain trust of the treating clinician, *i.e.* increase its credibility.[17,18] However, no single IP or NIP approach is located at a Pareto optimum, where it enjoys both the highest accuracy and the highest interpretability, but it rather exists as a comprise between them. For example, while a decision tree has more interpretable capability than the RF approach, its accuracy is generally outperformed by the latter.

The relationship of IP and NIP ML approaches in terms of accuracy and interpretability has been studied.[19] However, the selected ML approaches refer to "off the shelf" algorithms, where they have been implemented by someone else and are available in prepackaged libraries. In other words, there will be some room to improve their accuracy or interpretability performance. In fact, researchers in the field of medical physics have been struggling to improve accuracy and interpretability of the ML approaches for radiation outcomes prediction, and their efforts were based upon both the IP and NIP ML categories. This study intends to summarize current efforts and to provide a big picture of the current trends to develop more advanced ML approach for pART. The rest of the paper is organized as follows. Section 2 introduces the strategies that have been used to increase the accuracy of IP ML approaches, Section 3 summarizes the developed strategies or tactics to improve interpretability of NIP ML approaches, discussion and conclusions are given in Section 4.

BALANCING ACCURACY AND INTERPRETABILITY OF *INTERPRETABLE* ML APPROACHES

Logistic regression

Logistic regression is most commonly used model to represent linear relationships with the assumption of uncorrelated features. For example, logistic regression has been employed to predict xerostomia after radiotherapy, and in some instances it can approximate the performance of neural networks when sigmoidal functions are used.[20] An objective statistical multivariate model was also developed to describe radiation pneumonitis risk by assessing continuous and nominal parameters to determine the optimal model order and its parameters.[21] In order to predict pneumonia risk and hospital 30-day readmission, generalized additive models (GAMs) with low-dimensional terms were developed, and pairwise interactions were added to standard GAMs resulting in GA^{2}Ms. Logistic regression with single and cross-terms not only improved accuracy compared to the GAMs, the pairwise interaction could also be visualized as a heat map[22]. It turns out adding pairwise cross-terms may improve the prediction accuracy of the logistic regression, although it might not be fully explainable.

For modern data sets with a high dimension of features, GAMs and GA^{2}Ms could be very complicated by considering all the features and their interactions. While ridge regression (logistic regression +L2 norm regularization) intends to regularize the ill-posed problems caused by high dimensional data sets,[22] least absolute shrinkage and selection operator (LASSO, logistic regression + L1 norm regularization), is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces.[23] Elastic net is another regularized regression model that linearly combines the L1 and L2 penalties of the LASSO and ridge methods to handle correlated features and high-dimensional data set, and it was used for outcome prediction in chemoradiotherapy.[24] The elastic net was reported to have similar prediction performance as RFs and yielded higher discriminative performance than decision tree, neural network, SVM and LogitBoost in chemoradiotherapy outcome and toxicity prediction, particularly, when the complexity of the input features is limited to basic clinical and dosimetric variables.[24] However, with the increment of the accuracy, elastic net trades off a little interpretability compared to logistic regression.[9]

In order to facilitate the interpretability of regression-based analyses, graphical calculating devices named "nomograms" were widely employed in clinical practice for oncology applications including radiation treatment outcomes prediction by conducting the approximate graphical computation of a regression function.[25] The group at Memorial Sloan Kettering Cancer Center has developed several nomograms for varying cancer diagnostics.[26] In addition, such nomograms have been used to predict response for treatment. For instance, a nomogram was

devised for estimating treatment failure among males with clinically localized prostate cancer treated with radical prostatectomy[27] and for predicting disease-specific survival after hepatic resection for metastatic colorectal cancer.[28] Nomograms were also employed to predict recurrence-free survival for cervical cancer based on combining individual clinical information with imaging-based fludeoxyglucose/positron emission tomography prognostic factors.[29]

Decision tree

Decision trees can model nonlinear effects and are obviously interpretable as long as the tree depth is shallow.[24] More than three decades ago, a recursive algorithm (decision tree) was applied to arbitrary dose–volume histograms to estimate the complication probability for treatment planning optimization.[30] Recently, a recursive partitioning analysis was constructed to stratify patients into risk groups for clinically significant radiation pneumonitis after chemoradiation therapy for lung cancer.[31] Additionally, decision trees were employed to predict pneumonitis in Stage I non-small cell lung cancer (NSCLC) patients after stereotactic body radiation therapy (SBRT). Ensemble techniques based on the decision tree such as boosting with RUSBoost and bagging with RFs had been used to improve its accuracy, but at the expense of losing its interpretability.[7]

In order to study weight loss in head and neck cancer patients treated with radiation therapy, a classification and regression tree (CART) prediction model was developed based on a knowledge-discovery approach. The CART not only does not require a specification of the function to model covariates, but also its prediction accuracy increases with additional treatment toxicity information.[32] It seems that tree structure has a good potential to interpret nonlinear relationships and to be integrated with other NIP ML approaches for prediction accuracy improvement. Gradient boosting machine (GBM) intends to produces a prediction model by combing weak prediction decision trees, and it has been employed to predict long-term meningioma,[9] outcomes after radiosurgery for cerebral arteriovenous malformations with a high prediction performance and a less interpretability.[33] As a tree-structured boosting, MediBoost is a new framework to construct decision trees that retain interpretability while having accuracy similar to ensemble methods.[34] While it has the same structure as CART to build a single decision tree, it has the improved accuracy by considering weighted versions of all cases at each split.[9]

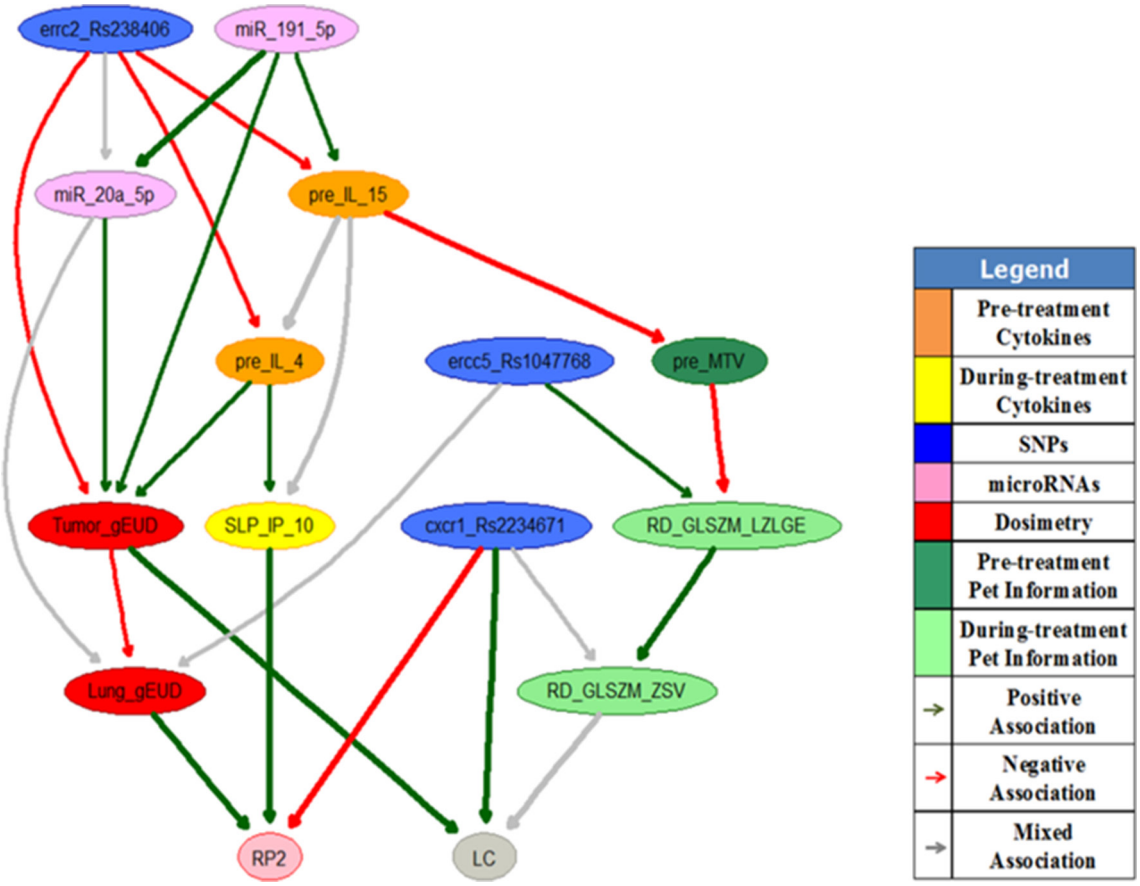
Bayesian network

Naïve Bayesian network (NBN) is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions, and it is interpretable but less accurate.[35] An advantage of the NBN is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Since independent variables are assumed, only the variances of the variables for each class need to be determined instead of the entire covariance matrix.[36] Hierarchical Bayesian networks (HBNs) are an extension of NBNs, which intends to improve inference and learning methods by using knowledge about the structure of the data. In order to predict 2-year survival in lung cancer patients treated with radiotherapy, HBN models were developed, and they were reported to outperform SVM models at handling missing data, and therefore are more suitable for the medical domain.[37] In a study of modeling local failure in lung cancer, a graphical HBN framework was generated to demonstrate that combining physical and biological factors with a suitable framework can improve the overall prediction, which highlights the potential of the integrated approach to predict post-radiotherapy local failure in NSCLC patients.[38] Additionally, a HBN was employed to estimate overall survival among colon cancer patients in a large population-based data set, resulting in a significant improvement upon existing AJCC stage-specific OS estimates.[39]

Moreover, a multiobjective HBN (MO-HBN) was developed to explore the biophysical relationships among treatment plans, patients' personal characteristics and radiation outcomes so that appropriate treatment plans before and during the course of radiotherapy can be identified.[40] Figure 1 shows an example of a during treatment MO-HBN to predict tumor local control (LC) and radiation pneumonitis toxicity Grade II or above (RP2) simultaneously in lung cancer patients. The important features for radiation outcomes prediction including tumor and lung gEUDs, three SNPs (errc2_Rs238406, ercc5_Rs1047768 and cxcr1_Rs2234671), two miRNAs (miR_20a_5p and miR_191_5p), two pre-treatment cytokines (IL_15 and IL_4), one pre-treatment radiomics feature (MTV), the relative change of one during treatment cytokine (IP_10) and the relative changes of two during treatment radiomics features (GLSZM_LZLGE, GLSZM_ZSV) were selected from a retrospective data set as denoted by the nodes in the figure. The edges of the MO-HBN, denoted by different colors, represent the biophysical relationships between the features analyzed. The study demonstrated that the MO-HBN has the potential to achieve a better performance than that of the corresponding NBN due to its hierarchical structure and additional biophysical information, and its prediction performance can be improved with patients' response during radiotherapy.[40] However, the confidence interval of the MO-HBN's prediction performance is still relatively large.

Although the BNs do not offer a significant improvement in outcome prediction over those resulting from less complex classifier algorithm as naïve Bayes, logistic regression or C4.5 decision trees, they still provide unique benefits to explore the relationship among features from large patient cohort data. The ability of carrying out causal inferences allows them to be utilized for answering complex clinical questions from unobserved evidence, and the probability distributions underlying the BN can be automatically updated with newly added patient information. Although it is hard to automatically learn a single graph that faithfully represents the casual structure of an application field, hybrid causal learning is an emerging field to show promise in obtaining causal structures with high prediction performance and causal patterns set out by domain experts (HBN with expert knowledge, HBN-EK).[41]

Figure 1. A during treatment MO-HBN for LC and RP2 prediction in lung cancer.[40] RP2, radiation pneumonitis toxicity Grade II or above; LC, local control; MO-HBN, multiobjective hierarchical Bayesian network.



BALANCING ACCURACY AND INTERPRETABILITY OF *NON-INTERPRETABLE* ML APPROACHES

Random forests and support vector machine

As previously stated, some shallow learning methods such as RFs and SVMs belong to the NIP ML approaches. RFs are an ensemble learning method which constructs a multitude of decision trees at training time and outputs the mean prediction of the individual trees. While variance can be controlled from the ensemble learning, the ensemble learning approach can sacrifice most of its interpretability at the same time, except that the frequency of feature appearance in the top layers of the ensemble decision tree may be used to explain their importance. However, the concept of RFs was integrated with other IP ML approaches to balance their accuracy and interpretability for radiation outcomes prediction. For example, formerly mentioned Medi-boost approach[34] attempts to emulate the performance of RFs while maintaining the intuition of classical decision trees.[42]

A SVM with a radial basis function kernel (SVM-RBF) transforms the original feature space into another space that can separate classes better. This transformation, however, can be much less intuitive than linear SVMs.[24] A non-linear SVM was developed for prediction of local tumor control after Stereotactic Body Radiation Therapy for early-stage NSCLC, and the prediction performance of the SVM model was significantly larger than that of a logistic tumor control probability model.[43] Interpreting SVM models is far from obvious, and the absence of a direct probabilistic interpretation also makes SVM inference difficult. However, work was done in providing methods to visualize SVM results as nomograms to support interpretability.[42] A nonlinear kernel, called localized radial basis function kernel (SVM-LRBF) was developed with the assumptions of intrafeature nonlinearity and interfeature independence. In addition to capturing nonlinearity of the classification function, the LRBF kernel can be visualized via nomograms. The SVM-LRBF method together with other SVM methods with linear kernel and RBF kernel had been applied for breast cancer prediction, and the study showed that while all the three kernel methods were equal in performance in terms of the area under the curve in the ROC curve, LRBF kernel was less sensitive to noise features than an RBF kernel.[44]

Deep learning

The impact of deep learning on radiation outcomes prediction

As a NIP ML approach, DL is mostly an extension of previously existing forms of artificial neural networks (ANNs) to larger number of hidden layers and artificial neurons in each layer for automatic discovery of useful features. Historically, ANNs lost popularity in favor of SVMs, RFs and gradient boosting trees due to the limited data set, computing resources and being prone to local minima. With the availability of larger data sets, graphical processing units (GPU) and stochastic gradient descent

algorithms, DL became possible to explore faster the training of larger, deeper architectures.[45] The key property of DL is that it can automatically learn useful representations of the data without conducting feature selection, which is one important component of other ML techniques. The reason why DL has relatively high prediction performance is that an architecture with sufficient depth can produce a compact representation, whereas an insufficiently deep one may require an exponentially larger architecture to represent the same function.

Now, DL based on neural networks has broad applications in radiation response, and it is poised to dominate medical image analysis for radiation outcomes prediction. Convolution neural networks (CNNs) were used to extract features from non-medical images for computer-aided diagnosis tasks,[46] extract radiomics features from the image patterns in developing a radiomics-based predictive model,[47] and extract deep features from preoperative multimodality MR images for survival prediction in Glioblastoma Multiforme.[48] Also, a CNN was developed to analyze rectum dose distributions and to predict rectal toxicity. [28] Studies show that DL has better performance than shallow learning approaches in radiation outcomes prediction. As a methodological proof-of-principle, a deep neural network (DNN) was created to predict the complete response of advanced rectal cancer after neo-adjuvant chemoradiation, which is an accurate surrogate for long-term local control. The DNN outperformed a linear regression model and a SVM model.[49] Also, CNNs and their variants were applied to the discovery of consistent patterns in three-dimensional dose plans associated with toxicities after liver stereotactic body radiotherapy. When the number of false negatives, *i.e.* missed toxicities, was minimized, DL produced almost two times fewer false-positive toxicity predictions in comparison to dose–volume histogram-based predictions.[50]

While the development of DL will transform the way we use imaging for diagnosis, treatment planning and decision making and will disrupt the way we practice medicine in a positive way, it may also affect clinical practice in negative ways.[49] Given that the field of medical physics has unique characteristics that differentiate it from those fields where these techniques have been applied successfully, the DL techniques have their limitations and nuances in radiation oncology with the limited sample sizes.[51] Another important issue with the DL is that these black-box-like networks are very difficult to debug, isolate the reason behind certain outcomes, and predict when and where failures will happen,[52] which are called interpretability challenges.[53] As the limited sample size issue could be gradually released by further collaboration of the hospitals and data centers, solving the issue of DL's interpretability becomes more important. Actually, the underlying mathematical principles of DLs are understandable. But they lack an explicit declarative knowledge representation, hence have difficulty in generating the underlying explanatory structures.[54] Although the potential of DL to improve prediction accuracy may outweigh their interpretability challenges in many industries,[32,54] professionals in the field of radiation oncology are working mostly with distributed heterogeneous and complex sources of data, and there must be a possibility to make the results re-traceable on demand.[54] There has been increasing interest in radiation oncology to make DL transparent, interpretable, and explainable, and the efforts to improve its interpretability for radiation outcomes prediction are summarized in the next section.

INTERPRETABILITY IMPROVEMENT FOR DEEP LEARNING

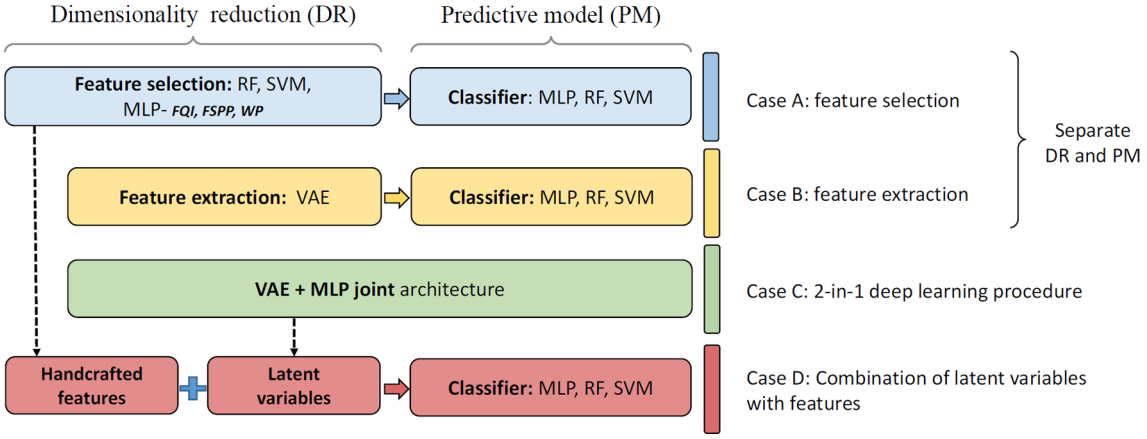
Deep learning with a combination of handcrafted features and latent variables (DL-HLV)

When a DNN is used as a feature extractor thousands of features are extracted. Unlike engineered handcrafted features, these features do not directly relate to something radiologists can easily interpret. Supplementing DL with information already known to be useful may improve the performance of these DL models and their interpretability. Previously, for survival prediction following glioblastoma multiforme, after deep features were extracted from preoperative multimodality MR images, a six-deep-feature signature was constructed by using the LASSO Cox regression model. The deep feature signature was combined with clinical risk factors to build a radiomics nomogram. The combined model not only achieved better performance for OS prediction, but also increased the interpretability of survival prediction through a nomogram construction.[48] Similarly, a methodology was developed to extract and pool low- to middle-level features using a pretrained CNN and to fuse them with handcrafted radiomic features computed using conventional CADx methods. In comparison to existing methods, the fusion-based breast CADx method demonstrated statistically significant improvements in terms of area under the curve on three different imaging modalities, and can also be used to more effectively characterize breast lesions.[55] Recently, the combination of traditional ML methods and DL variational autoencoders (VAE) techniques was developed to deal with limited datasets for radiation-induced lung damage prediction as shown in Figure 2.[56] It was demonstrated that a multilayer perceptron (MLP) method using weight pruning (WP) feature selection achieved the best performance among different hand-crafted feature selection methods, and the combination of handcrafted features and latent representation (Case D: latent Z + WP +MLP) yielded significant prediction performance improvement compared with handcrafted features only (Case A: WP + MLP), VAE-MLP disjoint (Case B) and VAE-MLP joint architectures (Case C).

Deep learning with sensitivity analysis (DL-SA)

Another method to increase the interpretability of DL is to calculate the sensitivity of the prediction with respect to changes in the input. Heat maps are visualization techniques that represent the importance of each pixel for the prediction task, which could help further optimize a CNN training approach. In a study of developing survival CNNs to predict cancer outcomes from histology and genomics, a heat map was employed to investigate the visual pattern that SCNN methods associate with poor outcomes by displaying the risks predicted by the SCNN in different regions of whole-slide images. The transparent heat map overlays in the study enable pathologists to correlate the predictions of highly accurate survival models with the underlying histology over the expanse of a whole-slide image. [57] In a DL-based radiomics model for survival prediction in glioblastoma multiforme, a

Figure 2. The evaluation of combination of handcraft features and latent variables for radiation-induced lung damage prediction,[56] where, "RF", random forest; "SVM", support vector machine; "MLP", multi-layer perceptron; "FQI", feature quality index; "FSPP", feature-based sensitivity of posterior probability; "WP", weight pruning; "VAE", variational autoencoders.



heat map was also used to show the Z-score difference of each radiomics feature between high risk and low risk group, and a consistency of radiomics feature Z-Score between the discovery data set and the validation data set.[48] In a retrospective multicohort radiomics study for lung cancer prognostication, DL was used in medical image for automated quantification of radiographic characteristics to improve patient stratification. A *gradient-weighted activation mapping* technique was employed to generate activation maps by mapping important regions in an input image with respect to predictions made, and the heat maps indicate regions in the input image having the most impact on the final prediction layer as shown in Figure 3.[52] In a previous study to develop CNNs for individualized hepatobiliary toxicity prediction after liver stereotactic body proton therapy, *saliency maps* of the CNNs were used to estimate the toxicity risks associated with irradiation of anatomical regions of specific organs at risk, and the CNN saliency maps automatically estimated the toxicity risks for portal vein regions.[50] In order to classify lung cancer using chest X-ray images, a 121-layer CNN was developed along with the transfer learning scheme. A *class-activation map* technique was employed to provide a heat map to identify the location of the lung nodule.[58]

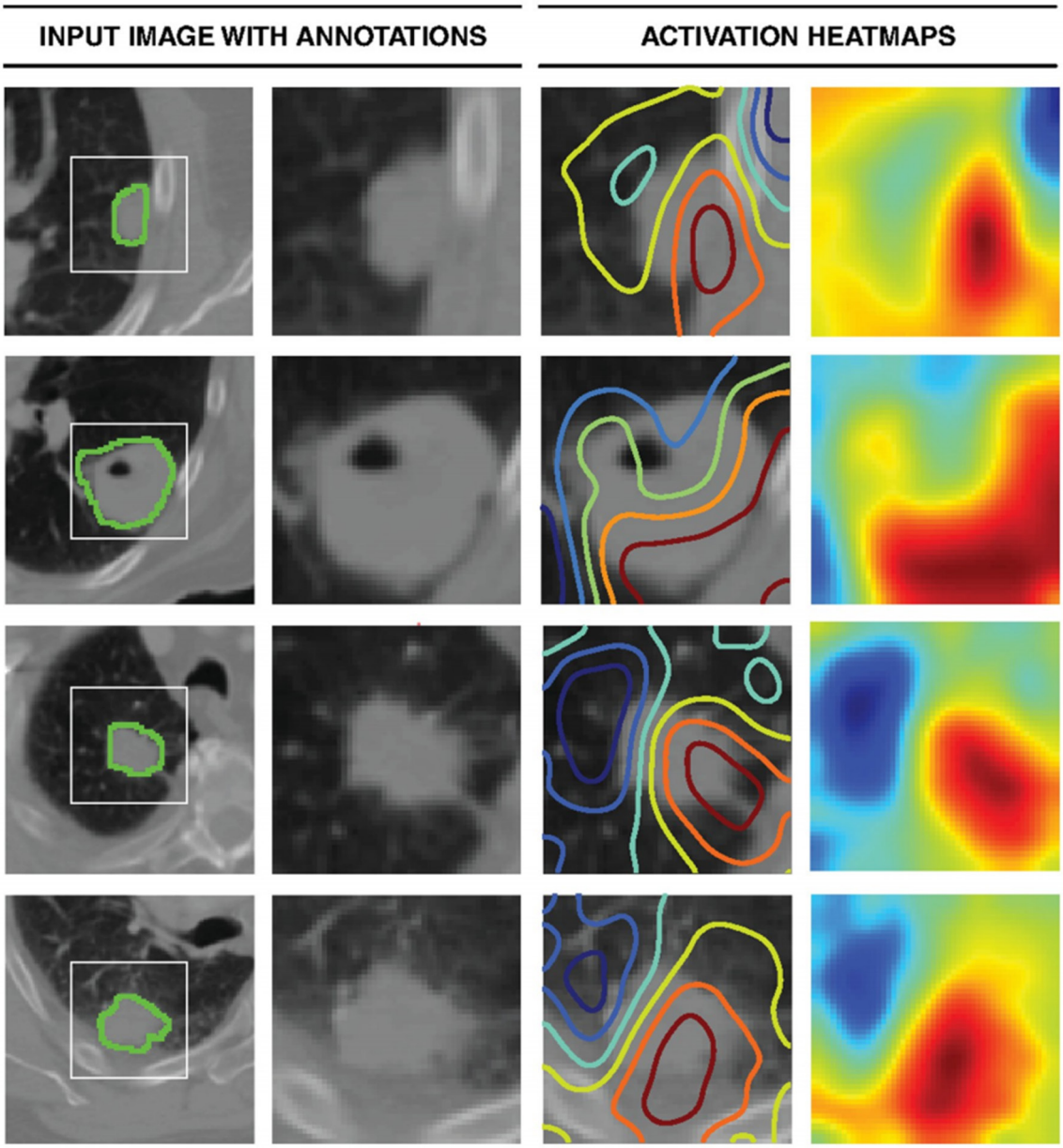
Deep learning with attention mechanisms (DL-AM)

Attention mechanisms are optional components of sequential prediction systems that allow the system to sequentially focus on different subsets of the input, and the subset selection is typically conditioned on the state of the system which is itself a function of the previously attended subsets. In addition to reducing the computational burden of processing high dimensional inputs by selecting only process subsets of the input, attention mechanisms also allow the system to focus on distinct aspects of the input and thus improve the ability to extract the most relevant information. Especially, soft attention mechanisms avoid a hard selection of which subsets of the input to attend and use a soft weighting of the different subsets for each piece of the output, thus leading to improvements in the quality of the generated outputs. The advantage brought by the soft-weighting is that it is readily amenable to efficient learning via gradient backpropagation.[59] Additionally, a gated recurrent unit (GRU)-based recurrent neural network (RNN) with *hierarchical attention* (GRNN-HA) was developed for clinical outcomes prediction through handling the high dimensionality of medical codes, modeling the temporal dependencies of healthcare events and characterizing the hierarchical structure of healthcare data. It was reported to have a better prediction accuracy and improve the interpretability of predictive models compared to baseline models by using the diagnostic codes from the medical Information Mart for Intensive Care to evaluate the model. The interpretability of the model depends on attention weights assigned to individual diagnostic codes and hospital visits, which were determined from relative importance of diagnostic codes on prediction.[60]

Deep learning with disentangled hidden layer representations (DL-DHLR)

Training DNNs with disentangled hidden layer representations is an active area of research to improve the interpretability of DL, although they have not been used for radiation outcomes prediction. The disentanglement of "the mixture of patterns" encoded in each filter of CNNs mainly disentangle complex representation in convolution-layers and transform network representations into interpretable graphs.[61] An explanatory graph represents the knowledge hierarchy hidden in convolution-layers of a CNN. While each filter in a pre-trained CNN may be activated by different object parts, part patterns can be disentangled from each filter in an unsupervised manner to clarify the knowledge representation.[62] A CNN was learned for object classification with disentangled representations in the top convolution-layer, where each filter represents a specific object part. Since the decision tree encodes various decision modes hidden inside fully connected layers of the CNN in a coarse-to-fine manner, given an input image, the decision tree infers a parse tree to quantitatively analyze rationales for the CNN prediction as shown in Figure 4.[63] The above methods focus on the understanding of a pre-trained network, but it is more challenging to learn networks with disentangled representations.[64–66]

Figure 3. Activation mapping. The first column represents the central axial slice of the network input (150 × 150 mm) with tumor annotations. In the second column, a 50 × 50 mm patch is cropped around the tumor. In the third column, activation contours are overlaid, with blue and red showing the lowest and highest contributions (gradients), respectively. Column four represents the activation heatmaps for a better visual reference.[52]

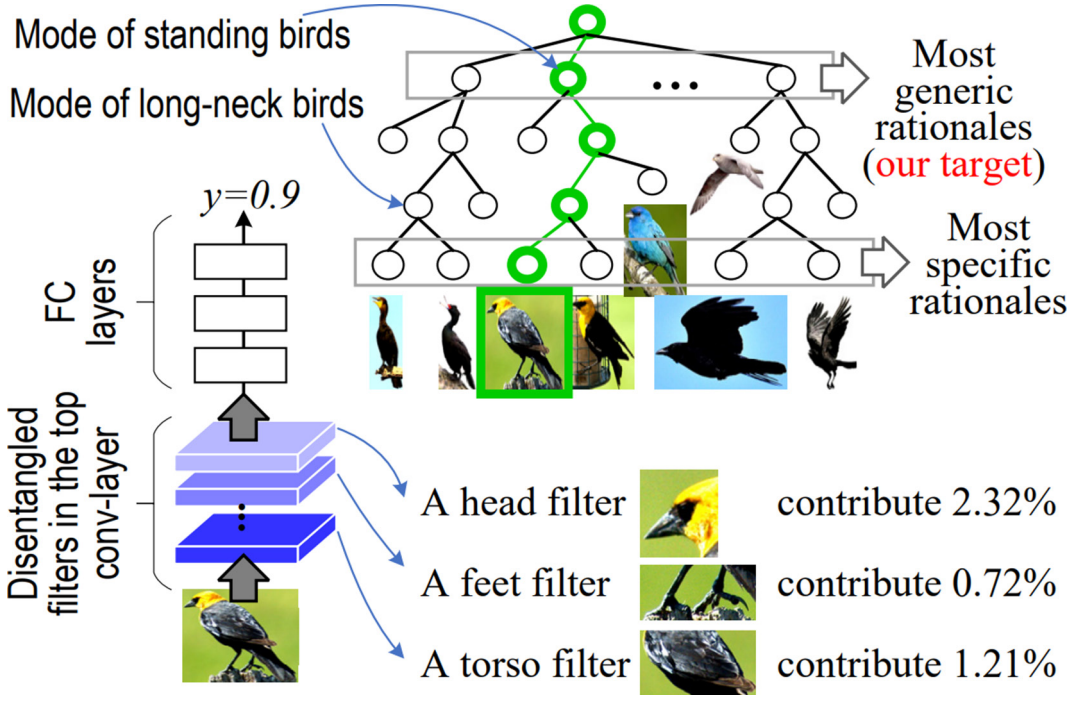


DISCUSSION AND CONCLUSIONS

Interpretability and explainability are different concepts although they have been used interchangeably. While the former is about being able to discern the predictions without necessarily knowing the underlying mechanics, the latter is being able to quite literally explain what are the mechanics that led to a particular behavior or decision by the algorithm.[67] In medicine including radiation oncology, interpretability represents physician's ability to accept, and interpret an algorithm decision in a scientifically sound manner without the need to explain algorithmic behaviors. As questions of accountability and transparency become more and more important, the interpretability of AI algorithms for radiotherapy outcomes prediction has improved in recent years, but is still far away from achieving full explainability. In this study, we focus on the trade-off of accuracy and interpretability in evaluating the prediction performance of common ML approaches, and summarize the balance of IP and NIP ML approaches for radiation outcomes prediction purposes from the current radiation oncology literature.

Table 1 shows the accuracy, interpretability and explainability levels of basic ML approaches such as logistic regression, decision trees, naïve Bayesian networks, SVM kernels, DL, and improved ML approaches associated with them. Since understanding the reasons behind prediction is quite important in assessing trust or credibility if one plans to take a clinical action

Figure 4. Decision tree that encodes all potential decision modes of the CNN in a coarse-to-fine manner. a CNN was learned for object classification with disentangled representations in the top convolution layer, where each filter represents an object part. For an input image, a parse tree (green lines) can be referred from the decision tree to semantically and quantitatively explain which object parts (or filters) are used for the prediction and how much an object part (or filter) contributes to the prediction.[63]CNN, convolution neural network.



based on a prediction, an algorithm, called local interpretable model-agnostic explanations (LIME), was developed to interpret the predictions of any classifier by approximating it locally with an interpretable model.[69] Due to its potential to enhance the interpretability of DL approaches, combining DL with LIME (DL-LIME) is considered as an improved DL approach as listed in Table 1. The number of "stars" associated to each ML approach in the table intends to describe the relative assessment of the accuracy, interpretability and explainability among these ML approaches, where the more stars represent the higher accuracy or interpretability or explainability. The evaluation of each ML approach is generated based on indicated literature next to it in the table. As can be seen from Table 1, the explainability of ML NIP methods is still a work in progress in many instances. It is interesting to evaluate the properties of these ML approaches in a general way by reviewing more literatures in different theoretic and application fields. However, it is out of the scope of this paper.

Due to the limited data sizes in radiation treatment and the requirement of clinical decision-making for pART, developing unique ML approaches to achieve the Pareto optimum of accuracy and explainability is necessary and challenging at the same time. Explainable AI was proposed based on the trade-off between prediction accuracy and explainability by producing more explainable models and maintaining a high level of learning performance.[70] However, as previously stated, we only focus on accuracy and interpretability, an initial stage towards full explainability, in this paper. The efforts to balance them not only came from IP ML approaches but also from NIP ML aspects as illustrated in Figure 5. The *y*- and *x*-axes of the figure represent the accuracy and interpretability of IP and NIP ML approaches for radiation outcomes prediction, respectively. Then the locations of these ML approaches were determined based on radiation oncology literatures as shown in Table 1. For the sake of clear description, blue or green color was used to represent NIP or IP ML approaches in terms of accuracy or interpretability. While the deeper the color indicates the higher accuracy or interpretability, ideal approaches to balance them are denoted as cyan color, which is the mixed color of blue and green. In addition to giving a general concept of the current status of ML approaches for radiation outcomes prediction, the figure also shows potentially possible trends to develop more balanced ML approach for pART.

The rising of DL approaches is attributed to their potential for high accuracy performance when sufficient data and computational support are available. These black box models create nonlinear predictors and automatically take into consideration a large number of implicit variable interactions. However, what makes them accurate is what makes their predictions difficult to understand; they are too complex. The exact DL architecture does not seem to be the most important determinant in getting a good solution for both accuracy and interpretability. A key aspect that is often overlooked is that expert knowledge about the task to be solved can provide advantages that can go beyond adding more layers to a CNN.[59]

Table 1. The evaluation of the accuracy (**A**), interpretability (**I**) and explainability (**E**) of ML approaches in radiation outcomes prediction

Basic ML	Type	A	I	E	Improved ML	Type	A	I	E
Logistic regression[20,21]	IP	*	****	***	GA^2M[68]	IP	**	***	**
					Ridge Regression[22]	IP	**	**	*
					LASSO[23]	IP	**	***	**
					Elastic Net[9,24]	IP	***	**	*
Decision tree [24,30,31]	IP	**	*****	*****	CART[32]	IP	***	****	*****
					Random Forests[7]	NIP	****	*	NA
					GBM[9,33]	NIP	****	*	NA
					MediBoost[9,34]	IP	****	**	*
Naïve BN [35,37]	IP	*	****	****	HBN[38,40]	IP	**	***	**
					HBN-EK[41]	IP	**	****	***
Linear SVM [24]	NIP	**	**	*	SVM-RBF[43]	NIP	***	*	NA
					SVM-LRBF[44]	NIP	***	**	*
Deep learning[49,50]	NIP	****	*	NA	DL-HLV [48,55,56]	NIP	*****	**	NA
					DL-SA[52,57] /AM[59,60]	NIP	*****	**	NA
					DL-DHLR[61–63]	NIP	*****	***	NA
					DL-LIME[69]	NIP	*****	***	NA

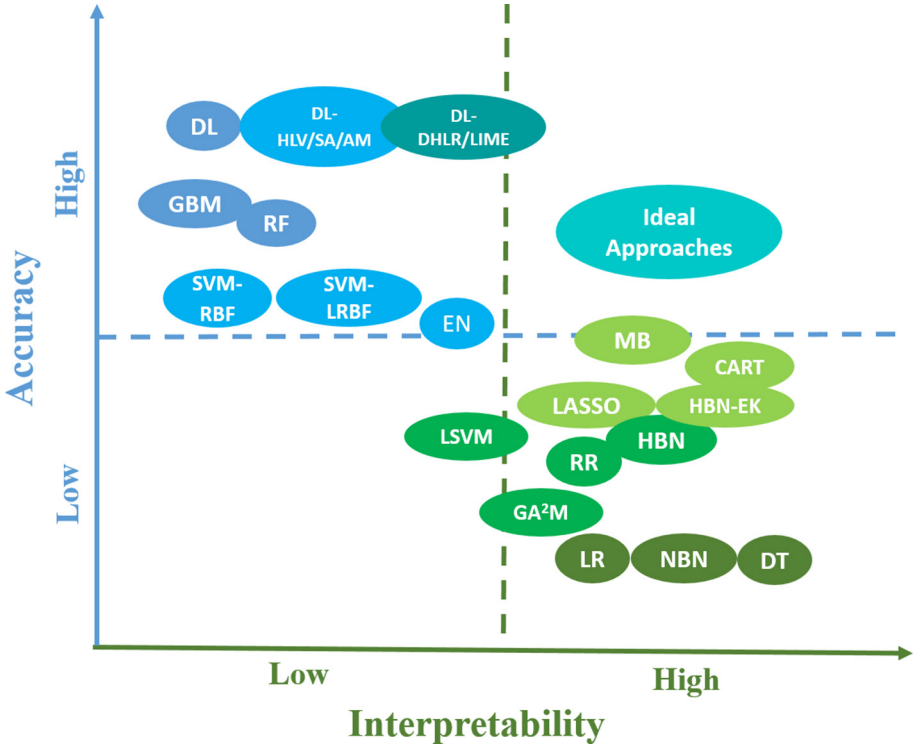
BN, Bayesian network; CART, classification and regression tree; DHLR, disentangled hidden layer representation; DL-AM, deep learning withattention mechanisms; DL-HLV, deep learning withcombination of handcrafted features and latent variables; GBM, gradient boosting machine; HBN, hierarchical Bayesian network; HBN-EK, hierarchical Bayesiannetwork with expert knowledge; HLV, handcrafted features and latent variables; IP, interpretable; LASSO, least absolute shrinkage and selection operator; LIME, local interpretable model-agnostic explanation; ML, machine learning; NIP, non-interpretable; SVM, support vector machine.

On the other hand, the main purpose of a predictive model's interpretability is to conduct statistical inferences, which intends to use the model to learn about the data generation process. However, none of NIP ML methods are able to conduct inference. In contrast, linear regression models, which assume that the data follow a Gaussian distribution, determine the standard error of the coefficient estimates and output confidence intervals. Since they allow us to understand the probabilistic nature of the data generation process, they are suitable method for inference. Also, a decision tree is one of the most widely used and practical methods for inductive inference. Particularly, Bayesian networks are popular for causal inference, since these models can be arranged to incorporate many assumptions about the data generation process.

Although there is no unified framework for ML interpretability, in general, the interpretability of the NIP ML methods can be improved by integrating them with the IP ML approaches. In addition to combining handcrafted features with latent variables, employing decision trees to encode all potential decision modes of the CNN in a coarse-to-fine manner as previously stated in Interpretability Improvement for Deep Learning, nomograms were employed to visualize and interpret SVM results[44,71] and to combine a DL-based radiomics signature with clinical factors to improve the accuracy and interpretability of overall survival prediction.[48] Moreover, other IP ML methods have also been used to improve the explanation of DL methods. For example, the problem of neural network structural learning was cast as a problem of Bayesian network structure learning, where a generative graph was learned, and its stochastic inverse was constructed resulting in a discriminative graph to simplify the neural network structure. Also it was proven that conditional-dependency relations among the latent variables in a generative graph are preserved in the class-conditional discriminative graph.[72] In order to handle the exhaustive and empirical neural network parameterization process, a new deep Bayesian network architecture was proposed by adopting the principle of multilayer Bayesian network in order to make use of the edges' significance, the causality, and the uncertainty of the Bayesian network for improving the meaningfulness of the hidden layers and the latent variable's connections.[73]

As more patient-specific information is becoming available due to advances in imaging and biotechnology, the classical p(variables)>>n(samples) inference problem of statistical learning will become more challenging in the areas of personalized and adaptive radiation oncology. Therefore, more advanced data analytics will be deployed and the demand to integrate accuracy and interpretability will rise to cope with clinical practice needs in the field.[74] Although different techniques are associated with distinct inherent limitations for radiation outcomes prediction, which include the independence assumption for features in logistic regression, the robustness in decision trees, the need for feature discretization in Bayesian networks, or the network configuration

Figure 5. The accuracy and interpretability of IP and NIP ML approaches in radiation outcomes prediction and the location of potential ideal approaches with more balanced accuracy and interpretability for the pART. Besides the notifications introduced in the paper, the rest of abbreviations in the figure can be described as follows, "EN", elastic net; "LR", logistic regression; "MB", MediBoost; "RR", ridge regression; "LSVM", linear support vector machine; "DT", decision tree. IP, interpretable; ML, machine learning; NIP, non-interpretable.



dependency in DL, our review shows that combining predictions among a handful of good, but different, IP and NIP models may result in better ML approaches to achieve higher accuracy and interpretability for radiation outcomes prediction.

ACKNOWLEDGMENT

This work was partly supported by National Institutes of Health (NIH) grants P01-CA059827, R37-CA222215, and R01-CA233487.

REFERENCES

1. Tseng H-H, Wei L, Cui S, Luo Y, Ten Haken RK, El Naqa I. Machine Learning and Imaging Informatics in Oncology. *Oncology* 2018;: 1–19. doi: https://doi.org/10.1159/000493575
2. Tseng H-H, Luo Y, Ten Haken RK, El Naqa I, Naqa E I. The Role of Machine Learning in Knowledge-Based Response-Adapted Radiotherapy. *Front Oncol* 2018; **8**: 266. doi: https://doi.org/10.3389/fonc.2018.00266
3. Samuel AL. Some studies in machine learning using the game of Checkers. *IBM J Res Dev* 1959; **3**: 210–29. doi: https://doi.org/10.1147/rd.33.0210
4. Naqa E I, Li R, Murphy MJ. *Machine Learning in Radiation Oncology: Theory and Applications*. Switzerland: Springer; 2015.
5. Valdes G, Simone CB, Chen J, Lin A, Yom SS, Pattison AJ, et al. Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiother Oncol* 2017; **125**: 392–7. doi: https://doi.org/10.1016/j.radonc.2017.10.014
6. Shiraishi S, Moore KL. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. *Med Phys* 2016; **43**: 378–87. doi: https://doi.org/10.1118/1.4938583
7. Valdes G, Solberg TD, Heskel M, Ungar L, Simone CB. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. *Phys Med Biol* 2016; **61**: 6105–20. doi: https://doi.org/10.1088/0031-9155/61/16/6105
8. Luo Y, El Naqa I, McShan DL, Ray D, Lohse I, Matuszak MM, et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis. *Radiother Oncol* 2017; **123**: 85–92. doi: https://doi.org/10.1016/j.radonc.2017.02.004
9. Gennatas ED, Wu A, Braunstein SE, Morin O, Chen WC, Magill ST, et al. Preoperative and postoperative prediction of long-term meningioma outcomes. *PLoS One* 2018; **13**: e0204161. doi: https://doi.org/10.1371/journal.pone.0204161
10. Luo Y, McShan D, Ray D, Matuszak M, Jolly S, Lawrence T, et al. Development of a fully Cross-Validated Bayesian network approach for local control prediction in lung cancer. *IEEE Transactions on Radiation and Plasma Medical Sciences* 2018.

11. Interian Y, Rideout V, Kearney VP, Gennatas E, Morin O, Cheung J, et al. Deep nets vs expert designed features in medical physics: An IMRT QA case study. *Med Phys* 2018; **45**: 2672–80. doi: https://doi.org/10.1002/mp.12890
12. Valdes G, Morin O, Valenciaga Y, Kirby N, Pouliot J, Chuang C. Use of TrueBeam developer mode for imaging QA. *J Appl Clin Med Phys* 2015; **16**: 322–33. doi: https://doi.org/10.1120/jacmp.v16i4.5363
13. Kearney V, Solberg T, Jensen S, Cheung J, Chuang C, Valdes G. Correcting TG 119 confidence limits. *Med Phys* 2018; **45**: 1001–8. doi: https://doi.org/10.1002/mp.12759
14. Alpaydin E. *Introduction to Machine Learning*. 3rd Edition; 2014. pp. 1–613.
15. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning. *Pr Int Conf Data Sc* 2018: 80–9.
16. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier. the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. *San Francisco, CA, USA: ACM* 2016.
17. El Naqa I, Pandey G, Aerts H, Chien J-T, Andreassen CN, Niemierko A, et al. Radiation Therapy Outcomes Models in the Era of Radiomics and Radiogenomics: Uncertainties and Validation. *Int J Radiat Oncol Biol Phys* 2018; **102**: 1070–3. doi: https://doi.org/10.1016/j.ijrobp.2018.08.022
18. Feng M, Valdes G, Dixit N, Solberg TD. Machine Learning in Radiation Oncology: Opportunities, Requirements, and Needs. *Front Oncol* 2018; **8**. doi: https://doi.org/10.3389/fonc.2018.00110
19. Blarer A, Intelligence EA. In: Ladetto Q, editor. defence future technologies: what we see on the horizon. *Feuerwerkerstrasse, Thun: armasuisse, Science and Technology* 2017: 41–3.
20. Ting H, Lee T, Cho M, Chao P, Chang C, Chen L, et al. Comparison of Neural Network and Logistic Regression Methods to Predict Xerostomia after Radiotherapy. *International Journal of Biomedical and Biological Engineering* 2013; **7**: 413–7.
21. Hope AJ, Lindsay PE, El Naqa I, Alaly JR, Vicic M, Bradley JD, et al. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. *Int J Radiat Oncol Biol Phys* 2006; **65**: 112–24. doi: https://doi.org/10.1016/j.ijrobp.2005.11.046
22. Landers A, Neph R, Scalzo F, Ruan D, Sheng K. Performance Comparison of Knowledge-Based Dose Prediction Techniques Based on Limited Patient Data. *Technol Cancer Res Treat* 2018; **17**: 153303381881115. doi: https://doi.org/10.1177/1533033818811150
23. Kerns SL, Kundu S, Oh JH, Singhal SK, Janelsins M, Travis LB, et al. The Prediction of Radiotherapy Toxicity Using Single Nucleotide Polymorphism-Based Models: A Step Toward Prevention. *Semin Radiat Oncol* 2015; **25**: 281–91. doi: https://doi.org/10.1016/j.semradonc.2015.05.006
24. Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu I-C, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Med Phys* 2018; **45**: 3449–59. doi: https://doi.org/10.1002/mp.12967
25. Schafer K. Nomography and Empirical Equations. *Chem-Ing-Tech* 1965; **37**: 661&.
26. Center MSKC. Prediction Tools - A Tool for Doctors and Patients. 2019. Available from: https://www.mskcc.org/nomograms.
27. Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *J Natl Cancer Inst* 1998; **90**: 766–71. doi: https://doi.org/10.1093/jnci/90.10.766
28. Kattan MW, Gönen M, Jarnagin WR, DeMatteo R, D'Angelica M, Weiser M, et al. A nomogram for predicting disease-specific survival after hepatic resection for metastatic colorectal cancer. *Ann Surg* 2008; **247**: 282–7. doi: https://doi.org/10.1097/SLA.0b013e31815ed67b
29. Kidd EA, El Naqa I, Siegel BA, Dehdashti F, Grigsby PW. FDG-PET-based prognostic nomograms for locally advanced cervical cancer. *Gynecol Oncol* 2012; **127**: 136–40. doi: https://doi.org/10.1016/j.ygyno.2012.06.027
30. Lyman JT. Complication Probability as Assessed from Dose-Volume Histograms. *Radiat Res* 1985; **104**: S13–S9. doi: https://doi.org/10.2307/3576626
31. Palma DA, Senan S, Tsujino K, Barriger RB, Rengan R, Moreno M, et al. Predicting Symptomatic Radiation Pneumonitis after Concurrent Chemoradiotherapy for Non-Small Cell Lung Cancer: Results of an International Individual Patient Data Meta-analysis. *Journal of Thoracic Oncology* 2012; **7**: S267–S.
32. Cheng Z, Nakatsugawa M, Hu C, Robertson SP, Hui X, Moore JA, et al. Evaluation of classification and regression tree (CART) model in weight loss prediction following head and neck cancer radiation therapy. *Adv Radiat Oncol* 2018; **3**: 346–55. doi: https://doi.org/10.1016/j.adro.2017.11.006
33. Oermann EK, Rubinsteyn A, Ding D, Mascitelli J, Starke RM, Bederson JB, et al. Using a Machine Learning Approach to Predict Outcomes after Radiosurgery for Cerebral Arteriovenous Malformations. *Sci Rep* 2016; **6**: 21161. doi: https://doi.org/10.1038/srep21161
34. Valdes G, Luna JM, Eaton E, Simone CB, Ungar LH, Solberg TD. MediBoost: a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine. *Sci Rep* 2016; **6**. doi: https://doi.org/10.1038/srep37854
35. JH O, Naqa E I. Bayesian network learning for detecting reliable interactions of dose-volume related parameters in radiation pneumonitis. Eighth International Conference on Machine Learning and Applications. *Proceedings* 2009; 484–8.
36. Holmes DE, Jain LC. Introduction to Bayesian Networks. *Innovations in Bayesian Networks: Theory and Applications* 2008; **156**: 1–5.
37. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. *Med Phys* 2010; **37**: 1401–7. doi: https://doi.org/10.1118/1.3352709
38. JH O, Craft J, Al Lozi R, Vaidya M, Meng Y, Deasy JO, et al. A Bayesian network approach for modeling local failure in lung cancer. *Phys Med Biol* 2011; **56**: 1635–51.
39. Stojadinovic A, Bilchik A, Smith D, Eberhardt JS, Ward EB, Nissan A, et al. Clinical decision support and individualized prediction of survival in colon cancer: bayesian belief network model. *Ann Surg Oncol* 2013; **20**: 161–74. doi: https://doi.org/10.1245/s10434-012-2555-4
40. Luo Y, McShan DL, Matuszak MM, Ray D, Lawrence TS, Jolly S, et al. A multiobjective Bayesian networks approach for joint prediction of tumor local control and radiation pneumonitis in nonsmall-cell lung cancer (NSCLC) for response-adapted radiotherapy. *Med Phys* 2018; **45**: 3980–95. doi: https://doi.org/10.1002/mp.13029
41. Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T, Brady M. Bayesian networks for clinical decision support in lung cancer care. *PLoS One* 2013; **8**: e82349. doi: https://doi.org/10.1371/journal.pone.0082349
42. Kang J, Rancati T, Lee S, Oh JH, Kerns SL, Scott JG, et al. Machine Learning and Radiogenomics: Lessons Learned and Future Directions. *Front Oncol* 2018; **8**. doi: https://doi.org/10.3389/fonc.2018.00228
43. Klement RJ, Allgäuer M, Appold S, Dieckmann K, Ernst I, Ganswindt U, et al. Support vector machine-based prediction of

local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer. *Int J Radiat Oncol Biol Phys* 2014; **88**: 732–8. doi: https://doi.org/10.1016/j.ijrobp.2013.11.216
44. Cho BH, Yu H, Lee J, Chee YJ, Kim IY, Kim SI. Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *Ieee T Inf Technol B* 2008; **12**: 247–56.
45. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. *Med Phys* 2019; **46**: e1–36. doi: https://doi.org/10.1002/mp.13264
46. Huynh BQ, Antropova N, Giger ML. Comparison of Breast DCE-MRI Contrast Time Points for Predicting Response to Neoadjuvant Chemotherapy Using Deep Convolutional Neural Network Features with Transfer Learning. *Medical Imaging 2017: Computer-Aided Diagnosis* 2017; **10134**.
47. Cha KH, Hadjiiski L, Chan H-P, Weizer AZ, Alva A, Cohan RH, et al. Bladder Cancer Treatment Response Assessment in CT using Radiomics with Deep-Learning. *Sci Rep* 2017; **7**. doi: https://doi.org/10.1038/s41598-017-09315-w
48. Lao J, Chen Y, Li Z-C, Li Q, Zhang J, Liu J, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci Rep* 2017; **7**. doi: https://doi.org/10.1038/s41598-017-10649-8
49. Bibault JE, Giraud P, Durdux C, Taieb J, Berger A, Coriat R, et al. Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. *Sci Rep-Uk* 2018; **8**.
50. Ibragimov B, Toesca D, Chang D, Yuan Y, Koong A, Xing L. Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT. *Med Phys* 2018; **45**: 4763–74. doi: https://doi.org/10.1002/mp.13122
51. Valdes G, Interian Y. Comment on 'Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study'. *Phys Med Biol* 2018; **63**: 068001. doi: https://doi.org/10.1088/1361-6560/aaae23
52. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med* 2018; **15**: e1002711. doi: https://doi.org/10.1371/journal.pmed.1002711
53. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996; **49**: 1225–31.
54. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? 2017;.
55. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys* 2017; **44**: 5162–71. doi: https://doi.org/10.1002/mp.12453
56. Cui S, Luo Y, Tseng H-H, Ten Haken RK, El Naqa I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. *Med Phys* 2019; **46**: 2497–511Accepted.. doi: https://doi.org/10.1002/mp.13497
57. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci U S A* 2018; **115**: E2970–E2979. doi: https://doi.org/10.1073/pnas.1717139115
58. Faust K, Xie Q, Han D, Goyle K, Volynskaya Z, Djuric U, et al. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. *BMC Bioinformatics* 2018; **19**: 173. doi: https://doi.org/10.1186/s12859-018-2184-4
59. Cho K, Courville A, Bengio Y. Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks. *IEEE Transactions on Multimedia* 2017; **17**: 1875–86. doi: https://doi.org/10.1109/TMM.2015.2477044
60. Sha Y, Wang MD. Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network. In: *Acm-Bcb' 2017: Proceedings of the 8th Acm International Conference on Bioinformatics, Computational Biology,and Health Informatics*; 2017. pp. 233–40.
61. Zhang Q-shi, Zhu S-chun, Zhu SC. Visual interpretability for deep learning: a survey. *Frontiers Inf Technol Electronic Eng* 2018; **19**: 27–39. doi: https://doi.org/10.1631/FITEE.1700808
62. Zhang Q, Cao R, Shi F, YN W, Zhu S-C. Interpreting cnn knowledge via an explanatory graph. *The Thirty-Second AAAI Conference on Artificial Intelligence* 2018;.
63. Zhang Q, Yang Y, YN W, Zhu SC. Interpreting CNNs via decision trees. 2018;.
64. Sabour S, Frosst N, Hinton GE. Dynamic Routing Between Capsules. *Advances in Neural Information Processing Systems* 2017; **30**(Nips 2017)30.
65. TF W, Xia GS, Zhu SC. Compositional boosting for computing hierarchical image structures. *2007 Ieee Conference on Computer Vision and Pattern Recognition* 2007; **s 1-8**: 492.
66. Zhang QS, YN W, Zhu SC. Interpretable Convolutional Neural Networks. *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition* 2018;: 8827–36.
67. Gall R. Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI. 2019. Available from: https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html.
68. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. *Sydney, Australia* 2015;.
69. Ribeiro T. M, Singh S, Guestrin C. eds. *the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: ACM New York; 2016 .
70. Gunning D, Intelligence EA. Gunning)%20IJCAI-16%20DLAI%20WS.pdf. 2016;Available fromXAI.
71. Belle V V, Van Calster B, Van Huffel S, Suykens JAK, Lisboa P. Explaining Support Vector Machines: A Color Based Nomogram. *Plos One* 2016; **11**.
72. Rohekar RY, Nisimov S, Gurwicz Y, Koren G, Novik G. Constructing Deep Neural Networks by Bayesian Network Structure Learning. *Adv Neur In* 2018; **31**.
73. Njah H, Jamoussi S, Mahdi W. Deep Bayesian network architecture for Big Data mining. *Concurrency Computat Pract Exper* 2019; **31**: e4418. doi: https://doi.org/10.1002/cpe.4418
74. Naqa IE, Kosorok MR, Jin J, Mierzwa M, Ten Haken RK. Prospects and challenges for clinical decision support in the era of big data. *JCO Clin Cancer Inform* 2018; **2**.