Incorporating Deep Learning Techniques into Outcome Modeling in Non-Small Cell Lung Cancer Patients after Radiation Therapy

by

Sunan Cui

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied Physics)
in the University of Michigan
2020

Doctoral Committee:

Professor Issam El Naqa, Chair
Professor Roy Clarke
Professor Randall K Ten Haken
Associate Professor Ambuj Tewari

Sunan Cui

sunan@umich.edu

ORCID ID: 0000-0002-8846-9449

Acknowledgments

I have accomplished this work with the help of many people. First and foremost, I would thank my advisors Prof. Issam El Naqa and Prof. Randall K Ten Haken for their great mentorship. They introduced me to innovative and challenging topics and gave me numerous support and help during my study. And I also have a great fortune to work with two former coworkers, Dr. Huan-Hsin Tseng and Dr. Yi Luo, who collaborated with me on several research projects. They were always available to discuss insightful ideas and provided help when I needed. I am also really grateful to the Applied Physics Program which provided me with numerous resources and financial during my study. My fellow graduate students and coworkers in Prof. Issam El Naqa's group, were all an important source of support and scientific discussions during my PhD study. Last but not least, to my parents, boyfriend and friends, thank each of you for your love, understanding, support and accompanying me during my adventure.

Table of Contents

List of Tables

List of Figures

Abstract

Radiation therapy (radiotherapy) together with surgery, chemotherapy and immunotherapy are common modalities in the cancer treatment. In radiotherapy, patients are given high doses of ionizing radiation which is aimed at killing cancer cells and shrinking tumor. Conventional radiotherapy usually gives a standard prescription to all the patients, however, as patients are likely to have heterogeneous responses to the treatment due to multiple prognostic factors, personalization of radiotherapy treatment is desirable. Outcome models can serve as clinical decision-making support tools in the personalized treatment, helping evaluate patients' treatment options before the treatment or during fractionated treatment. It can further provide insights into designing of new clinical protocols. In the outcome modeling, two indices including tumor control probability (TCP) and normal tissue complication probability (NTCP) are usually investigated.

Current outcome models, e.g., analytical models and data-driven models, either fail to take into account complex interactions between physical and biological variables or require complicated feature selection procedures. Therefore, in our studies, deep learning (DL) techniques are incorporated into outcome modeling for prediction of local control (LC), which is TCP in our case, and radiation pneumonitis (RP), which is NTCP in our case, in non-small-cell lung cancer (NSCLC) patients after radiotherapy. These techniques can improve the prediction performance of outcomes and simplify model development procedures. Additionally,

longitudinal data association, actuarial prediction and multi-endpoints prediction are considered in our models. These were carried out in 3 consecutive studies.

In the first study, a composite architecture consisting of variational auto-encoder (VAE) and multi-layer perceptron (MLP) was investigated and applied to RP prediction. The architecture enabled the simultaneous dimensionality reduction and prediction. The novel VAE-MLP joint architecture with area under receiver operative characteristics (ROC) curve (AUC) [95% CIs] 0.781 [0.737-0.808] outperformed a strategy which involves separate VAEs and classifiers (AUC 0.624 [0.577-0.658]).

In the second study, composite architectures consisted of 1D convolutional layer/ locally-connected layer and MLP that took into account longitudinal associations were applied to predict LC. Composite architectures convolutional neural network (CNN)-MLP that can model both longitudinal and non-longitudinal data yielded an AUC 0.832 [0.807-0.841]. While plain MLP only yielded an AUC 0.785 [CI: 0.752-0.792] in LC control prediction.

In the third study, rather than binary classification, time-to-event information was also incorporated for actuarial prediction. DL architectures ADNN-DVH which consider dosimetric information, ADNN-com which further combined biological and imaging data, and ADNN-com-joint which realized multi-endpoints prediction were investigated. Analytical models were also conducted for comparison purpose. Among all the models, ADNN-com-joint performed the best, yielding c-indexes of 0.705 [0.676-0.734] for RP2, 0.740 [0.714-0.765] for LC and an AU-FROC 0.720 [0.671-0.801] for joint prediction. Performance of proposed models was also tested on a cohort of newly-treated patients and multi-institutional RTOG0617 datasets.

These studies taken together indicate that DL techniques can be utilized to improve the performance of outcome models and potentially provide guidance to physicians during decision

making. Specifically, a VAE-MLP joint architectures can realize simultaneous dimensionality reduction and prediction, boosting the performance of conventional outcome models. A 1D CNN-MLP joint architecture can utilize temporal-associated variables generated during the span of radiotherapy. A DL model ADNN-com-joint can realize multi-endpoint prediction, which allows considering competing risk factors. All of those contribute to a step toward enabling outcome models as real clinical decision support tools.

Chapter 1 Introduction

Personalization of Radiotherapy

Currently, most radiotherapy treatments are designed to be population-based, giving similar prescription to all the patients. However, it is well known that patients are very likely to have heterogeneous responses due to multiple clinical, physical and biological prognostic factors such as histology, stage, volume and tumor hypoxia, [1-3]. Hence, individualization and adaptation of radiotherapy (i.e., physicians may prescribe a more or less intense regimen for an individual pre-treatment or during the fractionated course of treatment of radiotherapy) are desirable and a key to optimize radiotherapy responses. This concept is illustrated in Figure 1-1.

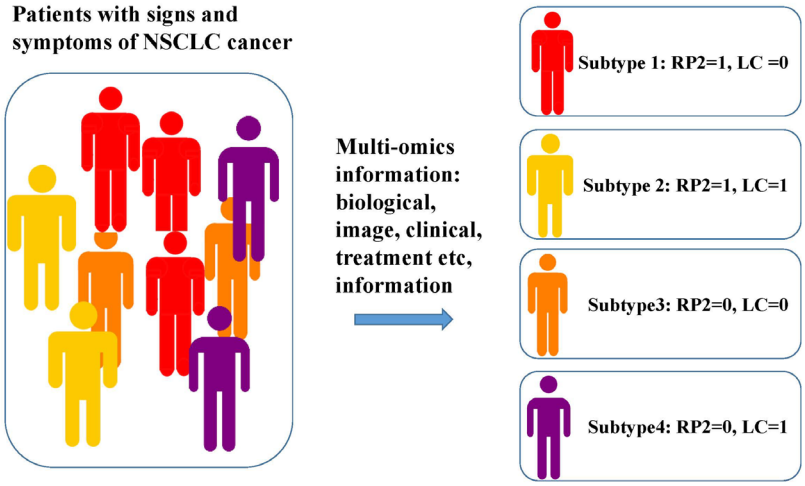

Figure 1-1. NSCLC patients of different subtypes may respond to the same radiotherapy differently. LC: local control, RP2: radiation pneumonitis grade greater or equal to 2.

Recent clinical trials focusing on treatment intensification in patients with locally advanced cancer have shown incremental improvements in local control (LC) and overall survival [4] [5]. Higher prescription doses may lead to poorer overall survival, as radiation-

induced toxicities remain major dose-limiting factors and likely culprit in treatment failure [6-8]. Therefore, there is a need for studies directed toward predicting treatment benefits versus the risk of failure. An individualized treatment would aim toward an optimized cancer treatment response while keeping in mind that a more aggressive treatment with a promised improved tumor control will not translate into improved survival unless severe toxicities are accounted for and limited during treatment planning. Therefore, improved models for predicting both LC and side effects should be considered in optimal treatment management design process.

Lung cancer, which is the most common cancer in the world, is a leading cause of cancer death in both men and women in the US. Specifically, non-small-cell lung cancer (NSCLC) accounts for 85% of lung cancer cases. In our study, locally advanced (stage III) NSCLC was considered since patients in this group will be more likely to benefit from personalized treatment compared to other NSCLC patients. Our study is aimed at realizing the potential trade-off of LC normal tissue toxicity for future personalized treatment.

Outcome modeling

One of the key components of personalization of treatment is to predict treatment outcomes during treatment planning or during a fractionated course of therapy to optimize response. Outcome models can also inform clinicians when weighing different treatment options with their patients or guiding/adapting radiotherapy fractionation subject to patient-specific variables. In the past decades, it has since tremendously evolved from simple hand calculations of dosage based on experiences and simplified understanding of cancer behavior into more advanced computer simulation models, driven by exponential growth in patient-specific data and an acute desire to have more accurate predictions of response [9].

Data resources in outcome modeling

Traditional outcome models usually only consider dosimetric information. This type of data is related to the treatment planning process which will be discussed on page 13 and includes dose-volume metrics derived from dose volume histogram (DVH) graphs [10-13].

With recent advances in quantitative multimodality imaging [14] and high throughput biotechnology (genomics [15], proteomics, transcriptomics [16], metabolomics, etc.), more patient specific information becomes available. An emerging field referred as 'radiomics' [14, 17] studies quantitative information from hybrid-imaging modalities and associate it with biological and clinical endpoints. For instance, PET/CT (positron emission tomography/computed tomography) has been utilized for staging, planning, and assessment of response to chemoradiation therapy [18, 19]. Biomarkers related to DNA damage detection and repair, oncogene, tumor suppressor, and signal transduction pathway, e.g., single-nucleotide polymorphisms (SNPs), copy number variations (CNVs)), inflammatory cytokines, anti-oxidant enzymes can also contribute to responses of treatment. Efforts of aggregating large-scale biomarkers, such as The Cancer Genome Atlas (TCGA) Data Portal have been made to collect clinical and biological information in different cancer types [20-23].

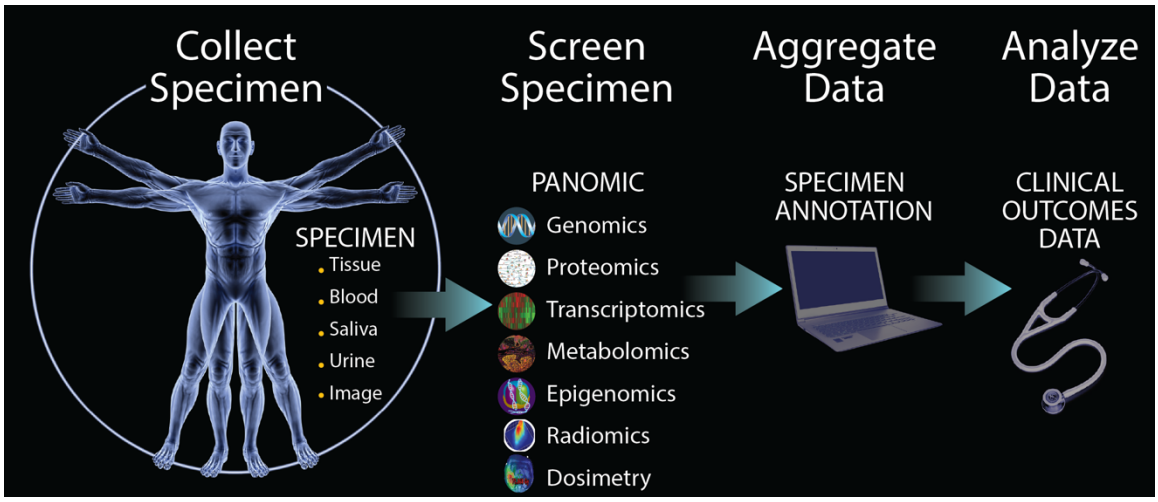

Figure 1-2. Patient specific information from all sources that could contribute to outcome modeling. Adapted from "Radiogenomics and radiotherapy response modeling" [15], by Issam El Naqa, and et al.,2017, Physics in Medicine and Biology 62 (16), p. R179-R206. Reprinted with permission.

Motivation of our study

As mentioned earlier, traditional outcome models are usually based on simple understanding of radiobiological effects. Recently, driven by advancement of quantitative multi-modality imaging and high throughput biotechnology, outcome models have been evolved into machine learning models which can provide more accurate prediction by taking into account more patient specific information. However, machine learning models usually require feature-engineering procedures which are time-consuming and may introduce selection bias, hence deep learning (DL) techniques which are known to have ability of learning complex representation from raw data are incorporated into outcome modeling to tackle this issue.

Deep learning in medicine

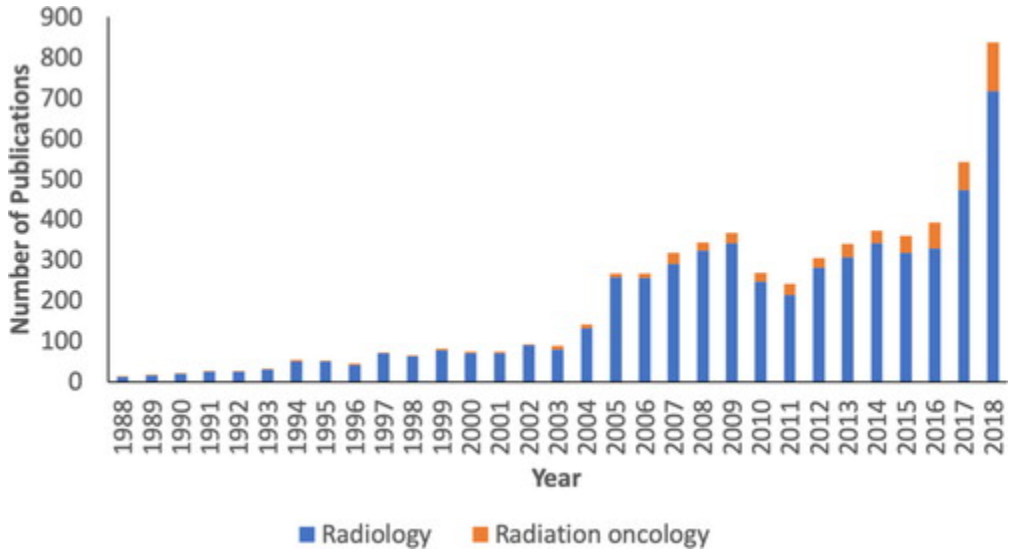

Figure 1-3. AI related publications in radiology and radiation oncology. Adapted from "Artificial intelligence: reshaping the practice of radiological sciences in 21 Century" [24], by El Naqa and et al, 2020, The British Journal of Radiology, 93 (1106), p. 20190855, Reprinted with permission.

Recent years have witnessed a great growth in AI related research in the field of radiology and radiation oncology as presented in Figure 1-3. Especially, DL which recently demonstrated tremendous success in image recognition problems [25] and natural language processing [26], have also been of interest in the medical field. DL is generally based on neural network (NN) architectures, using multiple layers to gradually extract higher-level features from the raw inputs; eliminating the necessary and typically problematic feature engineering process in classical machine learning, and hence showing superior performances. This is a key advancement in multivariable and statistical prediction modeling, where data representation and task learning can be effectively achieved in the same framework.

Motivation of applying deep learning techniques in outcome prediction

Compared to a one hidden layer multi-layer perceptron (MLP) as will be mentioned on page 18, a deeper NN may have better performance. Although, the Universal Approximation Theorem (UAT) developed by Hornik [27] states that mathematically a NN with one hidden

layer of sufficient nodes can approximate any measurable (and hence continuous) function on compact sets under certain mild conditions. Based on this, it seems that shallow (one hidden layer) MLPs will be good enough for any prediction task. However, the UAT theorem has several constraints. First, we need to have a sufficient (can be infinite) number of nodes. Secondly, it does not guarantee the theoretical performance can be achieved through optimization in practice due to local minima and convergence issues. Thus, it still depends on designing the right architecture (e.g.., activation function, regularization, number and size of layers, etc.) and adopting an appropriate training process (e.g., optimization method) in order to possibly achieve the theoretical performance estimates [28].

Adding more layers to a NN has been practically shown to provide a good architecture design versus increasing the number of nodes as suggested by UAT. An NN with more layers will show better performance than a single layer NN that has the same number of parameters. Intuitively, this is possible because each layer will transform its input, creating a new representation of the data. The multi-level abstraction that is being learned through multiple layers can be hardly coded into a single layer with the same number of nodes. Or formally speaking, the multi-layer structures enable NNs to recognize the entangled manifolds of the data more easily, so as to solve the designated task [29].

Contribution of our study

Despite the progression of machine learning applications in the outcome modeling in radiotherapy. Current models still have several limitations, to name a few: (1) they have limited predictive power for clinical implementation; (2) they have poor performance on data of limited sample size; (3) they involve tedious feature-engineering procedures; (4) they are not able to

directly utilize associations among heterogeneous data elements; (5) they overlook time-to-event information; and (6) they lack for multi-endpoint prediction mechanisms.

Hence, DL techniques are incorporated into outcome modeling in our studies to tackle all these six issues as briefly summarized below and detailed in this thesis.

In the current domain of outcome modeling, hundreds of variables are available to be explored, but with a limited sample size, which can impose a big challenge (i.e., under-power analysis) for predictive modelling. Under these circumstances, feature selection [30] [31], which constructs and selects subsets of features, is an indispensable step to build predictive models. However, the process of feature-engineering may involve lots of time and effort, and even introduce bias when model development procedures are not appropriately conducted. Hence, variational auto-encoder (VAE)-MLP joint architectures [32] were proposed in our study to realize simultaneously dimensionality reduction and prediction, which eliminated the necessity of (6) tedious feature selection in outcome modeling.

To address the issue of (2) limited sample size, it is beneficial to take into account the (4) associations among these patient specific variables. Conventional machine learning models e.g., support vector machine (SVM), random forest (RF) and MLP usually lack the inherent mechanism of modeling those associations. “Partially-connected” architectures such as CNNs [33], which are carefully designed to incorporate spatial associations, outperform an MLP that ignores such association in the prediction. In our study, a similar idea of adopting “partially-connected architecture” to model temporal data associations generated during the span of radiotherapy treatment was proposed. The composite architecture of 1D CNNs and MLPs [34] with fewer degrees of freedom compared to plain MLP can model both longitudinal and non-longitudinal data.

Conventional analytical models, e.g., Lyman models, RF, SVM and NN are usually designed for binary/multi-class classifications. However, compared to a binary endpoint which is attached to a specific follow-up time, time-to-toxicity/progression would leverage additional temporal information into outcome models and help provide better time-dependent decision support. Moreover, incorporation of time-to-censor information will also help utilize the censored information, which would be discarded otherwise. In the prediction of radiotherapy response, censored data are very common as follow-ups may be missed or patients die before the event occurs. As a result, models were proposed to (5) predict discrete-time endpoints in our study.

Unlike traditional analytical models which usually focus on predicting a single outcome, (6) multi-endpoint predictions [35] are considered in our proposed architectures, i.e., prediction of tumor control probability (TCP) and normal tissue complication probability (NTCP) can be simultaneously generated from a single architecture from a heterogeneous dataset containing multiple dosimetric, imaging and biological variables. Hence, trade-offs between competing outcomes can be possibly handled in our models, which is an important step towards establishing outcome models as easy-to-use decision-support tools.

Accomplishments

I have been awarded several fellowships and academic awards from the Rackham Graduate School and professional associations e.g., ASTRO (American Society for Radiation Oncology) AAPM (the American association of Physicists in Medicine) for my accomplishments in this thesis study. Pertained peer-viewed publications and book chapters were also listed.

Awards

- 2018-2019 Rackham Predoctoral Fellowship

- 2018 Best in Physics, ASTRO Annual Meeting (*awarded to the top abstract in the radiation oncology physics category)*
- 2018 Annual Meeting Travel Grant, ASTRO Annual Meeting
- 2017 Farrington Daniels Award AAPM (*awarded to the best paper published in Medical physics in the category of radiation therapy dosimetry, planning or delivery)*

Peer-viewed publications

1. **Sunan Cui**, Huan-Hsin Tseng, Julia Marie Pakela, Randall K. Ten Haken, Issam El Naqa, "Introduction to machine and deep learning for medical physicists", Special Issue, Medical Physics, (in production) DOI:10.1002/mp.14140
2. **Sunan Cui**, Yi Luo, Huan-Hsin Tseng, Randall K.Ten Haken, Issam El Naqa. "Combining Handcrafted Features with Latent Variables in Machine Learning for Prediction of Radiation-Induced Lung Damage", Medical Physics, 2019 March, 46(5) [32]
3. **Sunan Cui**, Yi Luo, Huan-Hsin Tseng, Randall K.Ten Haken, Issam El Naqa. "Artificial Neural Network With Composite Architectures for Prediction of Local Control in Radiotherapy, IEEE Transactions on Radiation and Plasma Medical Sciences, 2019 March, 3(2) [34]
4. Yi Luo, Huan-Hsin Tseng, **Sunan Cui**, et al. "Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling", BJROpen 2019 1:1 [36]
5. Huan-Hsin Tseng, LiseWei, **Sunan Cui**, et al. "Machine Learning and Imaging Informatics in Oncology", *Oncology*, DOI/10.1159/000493575 [37]

6. Huan-Hsin Tseng, Yi Luo, **Sunan Cui**, Jen-Tzung Chien, Randall K. Ten Hakem, Issam El Naqa. “Deep Reinforcement Learning for Automated Radiation Adaptation in Lung Cancer" (editors’ choice), *Medical Physics;* 2017 Dec; 44(12). [38]
7. **Sunan Cui**, et al, “A composite deep learning architecture for the jointly actuarial prediction of local control and radiation pneumonitis in radiotherapy for non-small cell lung cancer patients, AAPM oral presentation, manuscript in preparation

Book chapters

1. **Sunan Cui**, Randall K. Ten Haken, Issam El Naqa, “Building a Predictive Model of Toxicity: Methods", in Rancati, Tiziana & Fiorino, Claudio (Eds), “Modelling Radiotherapy Side Effects: Practical Applications for Planning Optimization", CRC Press, 2019. [39]
2. **Sunan Cui,** Issam El Naqa, “Prediction of oncology treatment outcomes”, in Machine and Deep Learning in Oncology, Medical Physics and Radiology, Springer (upcoming)

Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 provides a general overview of radiotherapy (principle, types and treatment planning), methodologies of outcome models (analytical models, machine learning models, and DL methods), validation process and radiotherapy treatment of lung cancer. A combination of conventional feature selection methods and novel VAE-MLP joint architecture [32] was presented in Chapter 3 for the prediction RP2. Chapter 4 focuses on modeling longitudinal association with novel 1D CNN-MLP joint architectures [34] and the prediction of LC. In Chapter 5, actuarial neural network architectures (AD NN) were proposed to incorporate time-to-event information to jointly predict RP2/LC. In Chapter 6, discussion of the limitation of current work and future directions are provided.

Chapter 2 Background

Radiation therapy

Radiation therapy (radiotherapy), which uses high doses of ionizing radiation to eradicate tumor cells [40] is among the common cancer treatment modalities, e.g., surgery, chemotherapy, radiation therapy, immunotherapy. At high doses, radiation can kill cancer cells by damaging their DNA and hence help cure cancer. Radiation therapy is usually applied to tumors that are localized to one area of the body. It can also be combined with other treatment modalities, being used before, during or after surgery (as adjuvant therapy) or chemotherapy. The amount of radiation, i.e., dose in radiotherapy is usually measured in Gray (Gy), which is defined as the absorption of one joule of radiation energy per kilogram of matter. The goal of radiation therapy is to deliver high doses of ionizing radiation to eradicate tumor cells [2], while at the same time minimizing the risks of damaging surrounding normal tissue [3].

Types of radiotherapy

Depending on where the radiation is from, i.e., external beam or internal radionuclide, radiotherapy can be classified into two classes, external beam radiation therapy (EBRT) and brachytherapy.

Our study is focused on outcome modeling for EBRT, which is a far more prevalent case than brachytherapy. Radiation in EBRT comes from machines e.g., medical linear accelerator (LINAC) (shown in Figure 2-1), Gamma knife, Cyberknife, cyclotrons, these machines can accelerate and deliver photons, electrons or protons to patients. These particles would finally release their energy to the tumors and kill them. To allow normal cells which are generally more

efficient than tumor cells in repairing DNA to recover [41], the total dose (usually 40-80 Gy) is fractionated (spread out over time) into around 20-30 fractions and would be delivered in 4-7 weeks.

Figure 2-1. Varian Truebeam Linac. Adapted from Varian official website, https://www.varian.com/products/radiotherapy/treatment-delivery/truebeam

In brachytherapy, a radiation source is placed into the human body, in or near the tumor. It can be used in a limited number of cancer types, e.g., breast, cervix, prostate and eye.

Side effects

During radiotherapy, radiation can be unavoidably delivered to normal tissue, leading to side effects. Radiation-induced toxicity can be categorized according to its onset time into early and late effects. Early effects can occur during or a few days to weeks after irradiations, typically in the rapidly proliferating tissues. These effects include skin erythema, mucositis, esophagitis, diarrhea and immunosuppression. Late effects typically occur months to years after treatment,

usually in slowly or non-proliferating tissues. Common late effects are lung fibrosis, kidney damage, heart disease, liver disease, spinal cord injury and proctitis.[42]

Radiation-induced toxicities are usually categorized using clinical standards such as RTOG (the Radiation Therapy Oncology Group), LENT-SOMA (late effects of normal tissue-subjective, objective, management, analytic scales), or the National Cancer Institute CTCAE (common terminology criteria for adverse events).[43] Some common side effects metrics like patient symptoms (e.g., shortness of breath), formal clinical/functional assessments (e.g., quality of life tools) and laboratory tests (e.g., pulmonary function tests (PFTs)) are usually considered in these standards for evaluating toxicities. In our study, radiation pneumonitis was graded based on CTCAE criterion by radiation oncologists.

Treatment planning

In modern radiotherapy, treatment planning is carefully conducted to deliver uniform dose to the tumor while minimizing side effects on normal tissue.

Treatment planning process usually starts with simulation and image segmentation. During simulation, anatomic images of high quality e.g., computed tomography (CT) and magnetic resonance imaging (MRI) are obtained. In image segmentation, anatomic regions of interests e.g., tumor, critical normal structures, anatomic landmarks are delineated slice-by-slice on the obtained anatomic images. Specifically, several treatment volumes e.g., gross tumor volume (GTV), (i.e., visible tumor volumes) clinical target volume (CTV), (i.e., GTV+ subclinical/invisible invasion), internal target volume (ITV), (i.e., CTV+ internal margin for organ motion) and planning target volume (PTV) (i.e., ITV + set up margin and error) are defined. After contouring is done, one would find out how to design fields and arrange beams. Specifically, appropriate field apertures, beam direction, number of fields, beam weights and

intensity modifiers (e.g., wedges, compensators, dynamic multileaf collimator) needs to be determined to ensure that dose will be delivered to the entire tumor and all the normal tissues are spared, which is a step so-called plan optimization. Some common treatment techniques include three-dimensional conformal radiation therapy (3-D CRT), intensity-modulated radiation therapy (IMRT) and volumetric-modulated arc therapy (VMRT), where different groups of parameters e.g., beam directions, beam weights and intensity modifiers are to be determined. In a forward-planning system, these parameters are selected iteratively based on a trial-and-error process. In a more advanced inverse-planning system, these parameters can be optimized through minimizing the difference between actual and ideal dose distribution or achieving some clinical objectives i.e., physical endpoints and biologic endpoints. Physical endpoints are associated with optimal dose distribution within specified target volume and dose to critical organs. Biologic endpoints can be in indices e.g., TCP and NTCP generated by outcome models.

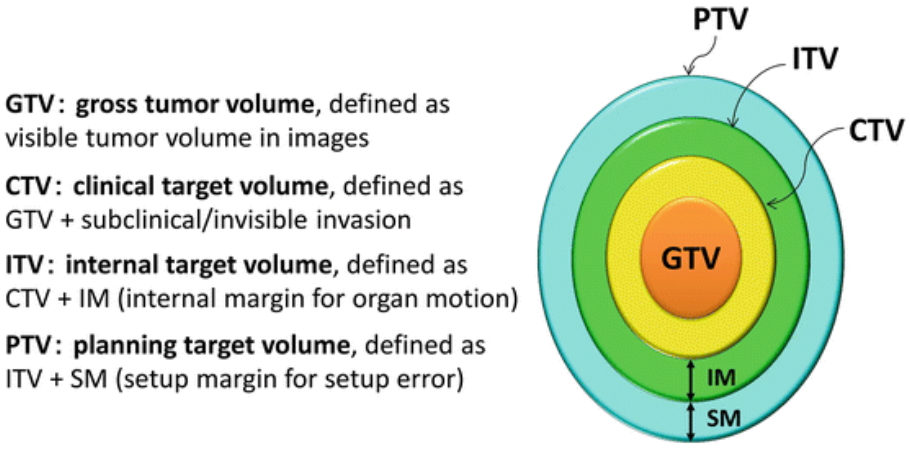

Figure 2-2. The diagram of target volumes GTV, CTV, ITV and PTV. Adapted from Hidetaka Arimura and et al. in "Computer-assisted target volume determination" [44] in Hidetaka Arimura (eds) "Image-based computer-assisted radiation therapy", Springer Singapore. Reprinted with permission

DVH is usually used when evaluating the treatment plan and predicting treatment responses. Information extracted from DVH and DVH itself is included in our study for the prediction of RP2 and LC. DVH can be represented in two forms: the cumulative integral DVH (Figure 2-3 down) which is a plot the volume of a given structure receiving a certain dose or

higher as a function of dose, and the differential DVH which is a plot of volume receiving a dose within a specific dose bin as a function of dose. DVH summarizes the entire dose distribution of a structure of interest into a single curve.

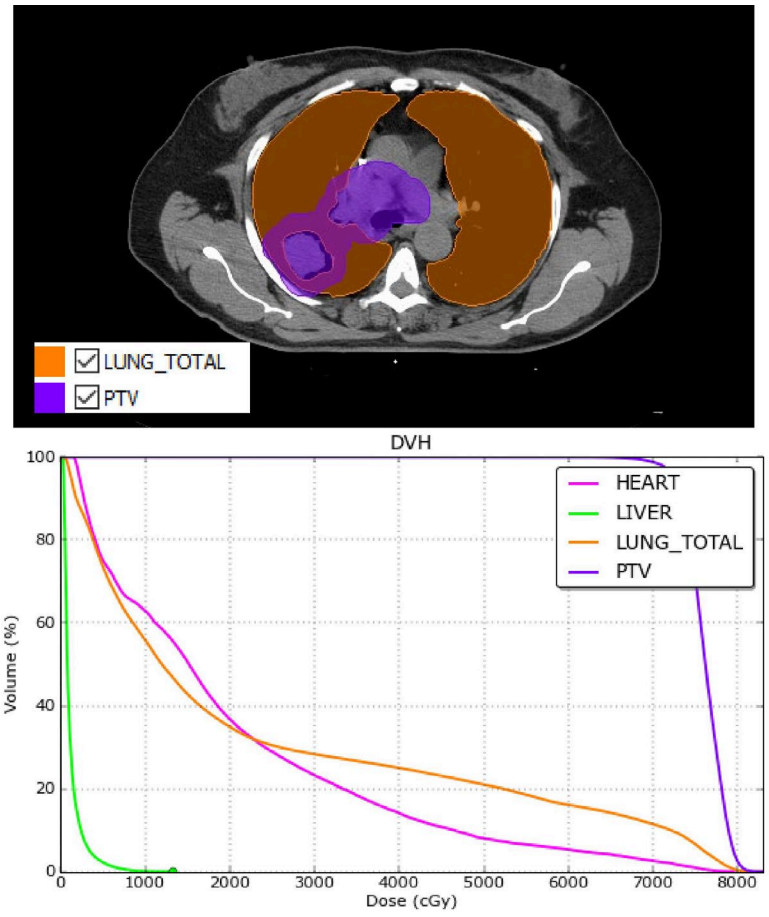

Figure 2-3. Illustration of delineation of the target volume (PTV) and lung on CT images (up). The cumulative DVH of the target volume (PTV) and critical organs (heart, liver and lung) (down).

Tumor control probability and normal tissue complication probability modeling

Radiotherapy outcomes are usually characterized by two indices: tumor control probability (TCP) [45], which is the probability of the extinction of clonogenic tumor cells after radiotherapy, and normal tissue complication probability (NTCP), which is the probability of healthy normal tissue injury [46]. The trade-off between the two indices should be carefully

examined in the personalization of treatment, as a more aggressive treatment may improve tumor control, but will not translate into improved survival due to possible severe toxicities [6-8]. Outcome modeling plays an important role in treatment personalization and adaption [47] in radiation oncology. Prevalent models include analytical models and machine learning models.

Analytical models

Traditional analytical models are categorized as mechanistic models and phenomenological models. The former approach mathematically formulates toxicity based on a simplified biophysical understanding of radiation effects on cells primarily from *in vitro* cell culture experiments. The latter attempts to fit the available dosimetric data to an empirical and parametric model.

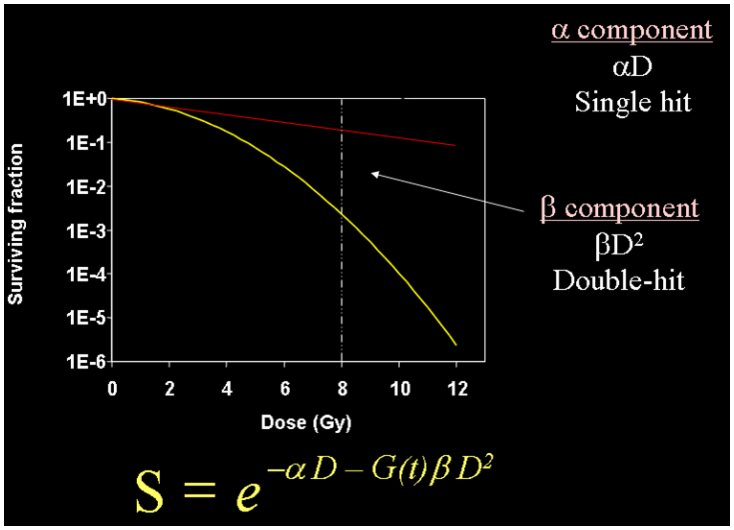

Figure 2-4. Logarithm of survival fraction in LQ models. Adapted from "Building a predictive model of toxicity: methods" [39] by Sunan Cui and et al. in Tiziana Rancati and Claudio Fiorino (eds) in " Modelling radiotherapy side effects: practical applications for planning optimization", 2019, CRC Press and Taylor&Francis Group

Mechanistic models e.g., linear quadratic (LQ) model attributes cell killing to DNA damage in the nucleus. Parameters α and β are related to radiosensitivity and their values can emphasize the difference between different responding tissues (e.g., lung tumor: $\frac{\alpha}{\beta} = 10$, lung

tissue $\frac{\alpha}{\beta} = 4$). Moreover, the model can be practicably extended to applications in fractionated radiotherapy [48]. A quantity called the biologically effective dose (BED) is defined for a very large number of fractions with very small doses and is used to simplify the conversion between different radiation fractionation regimens:

$$BED = \frac{E}{\alpha} = nd \times (1 + \frac{d}{\alpha/\beta}) \quad \text{Eq. 1}$$

To convert BED back to a physical quantity, an equivalent dose at some standard fractionation is used (e.g., EQD2 for 2 Gy fraction):

$$EQD2 = BED/[1 + 2/(\alpha/\beta)] \quad \text{Eq. 2}$$

In our study, to account for the effects of different fractionated dose (range: 2Gy-3Gy), the dose received by patients was all converted into EQD2.

In phenomenological models, TCP and NTCP can be modeled by a sigmoid-shaped function [49]. In a log-logistic model, TCP is expressed as [50]

$$TCP(D, D_{50}, k) = \frac{1}{1+\left(\frac{D_{50}}{D}\right)^k} \quad \text{Eq. 3}$$

where D is the uniform dose irradiated to the tumor, k describes the slope of the curve and D_{50} is the uniform tumor dose-related to 50% probability of tumor control. In a Lyman model [51], NTCP is expressed as a cumulative distribution function of a Gaussian distribution (a Probit function),

$$NTCP(D, D_{50}, m) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} exp\left(-\frac{u^2}{2}\right) du \quad \text{Eq. 4}$$

$$t = \frac{D - D_{50}}{m\, D_{50}}$$

, where D is the uniform dose irradiated to the organ of interest, D_{50} is the dose-related to 50% toxicity probability and m is a parameter to control the slope of a curve. The above expressions are in the condition of uniform irradiation, when the organ is irradiated with inhomogeneous

dose distribution described by a dose-volume histogram (DVH), D_{eff} can be defined and replaces *D*.

$$D_{eff} = (\sum_i v_i D_i^{\frac{1}{n}})^n \quad \text{Eq. 5}$$

Machine learning models

Recent years have witnessed the emergence of machine learning models utilizing informatics techniques, in which dose-volume metrics are combined with other patient- or disease-based prognostic factors [10-13, 52-54]. Some common machine learning models that have been applied to model TCP or NTCP include support vector machine (SVM), random forests (RF) and neural networks (NN).

An SVM [55] is a classifier formally defined by a separating hyperplane, which can categorize labeled data. In practice, as it is usually not feasible to completely separate samples from different classes, some tolerance errors ξ are allowed. In an SVM, NTCP or TCP can be modeled as,

$$f(x) = w^T \phi(x) + b, \quad \text{Eq. 6}$$

, where x ($x \in \mathbb{R}^d$) represents patient specific information, (w, b) represent model parameters. $\phi(\cdot)$ is a non-linear mapping function, which maps variables from an original space to a higher dimension space for a better separation. A so-called kernel function *K* is defined as an inner product in a feature (Hilbert) space based on $\phi(\cdot)$. Some common kernel function K include polynomial kernel, radial basis function kernel. The optimal parameters of SVM are determined by optimization of a hinge loss function,

$$\min_{w \in \mathbb{R}^d} \| w \|^2 + C \sum_{i}^{N} \max(0, 1 - y_i f(x_i)) \quad \text{Eq. 7}$$

, where the first term is correlated with the size of margins between two classes, the second term is an error-tolerance term. Parameter *C* is for regularization and is responsible for trade-offs between these two terms.

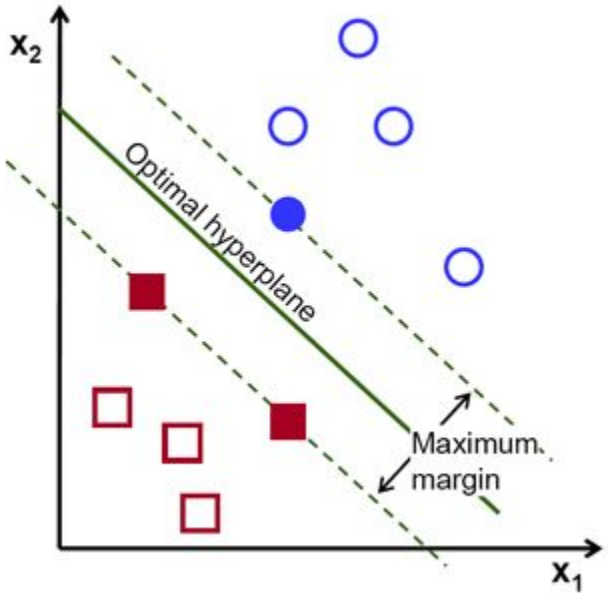

Figure 2-5. A hyperplane separates different classes (circles and squares) in SVM, and a tolerance error defined by a maximum margin is allowed

A RF [56] is another classical machine learning method. It is an ensemble learning method based on decision trees. A decision tree [57] is a flowchart-like structure where each node represents a “test" on an attribute (feature) splitting samples into different branches, nodes can be then repeatedly applied to test attributes of different branches until a decision regarding classification is done by the leaf nodes. During this process, the Gini coefficient [58] is a common measure used to decide a split (i.e., the feature applied, threshold). RF randomly selects observations and features to build several decision trees and averages the results to reduce the variance.

Multi-layer neural networks or multi-layer perceptrons (MLPs) [59] are a types of artificial neural networks that are feed-forward and fully-connected. These methods have witnessed revived interest in recent years with the advent of DL methods and their popularity particularly in computer vision applications. An MLP consists of several layers and neurons, where every neuron in the following layer is connected to all the neurons in its former layer. The connection is unidirectional and no circles exist in the network architecture.

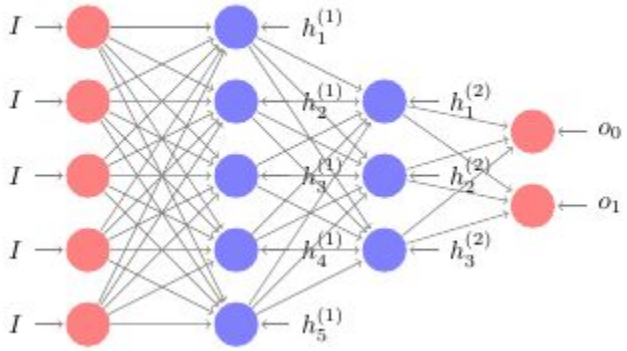

Figure 2-6. A diagram of an MLP with two hidden layers

The value of a neuron in the hidden layer and output layer is calculated by taking a weighted sum of all the neurons in its former layer followed by a non-linear activation function as shown in Figure 2-7. Some examples of activation functions [60] are, sigmoid $g(t) = \frac{1}{1+e^{-t}}$, ReLU [61] $g(t) = \max(0, t)$ and softmax $g_i(t) = \frac{e^{t_i}}{\Sigma_j e^{t_j}}$.

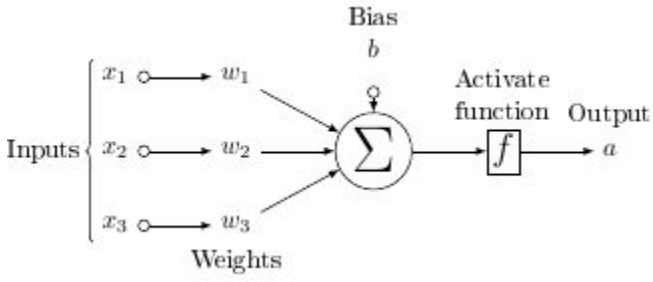

Figure 2-7. Activation function and calculation of values of nodes in NN

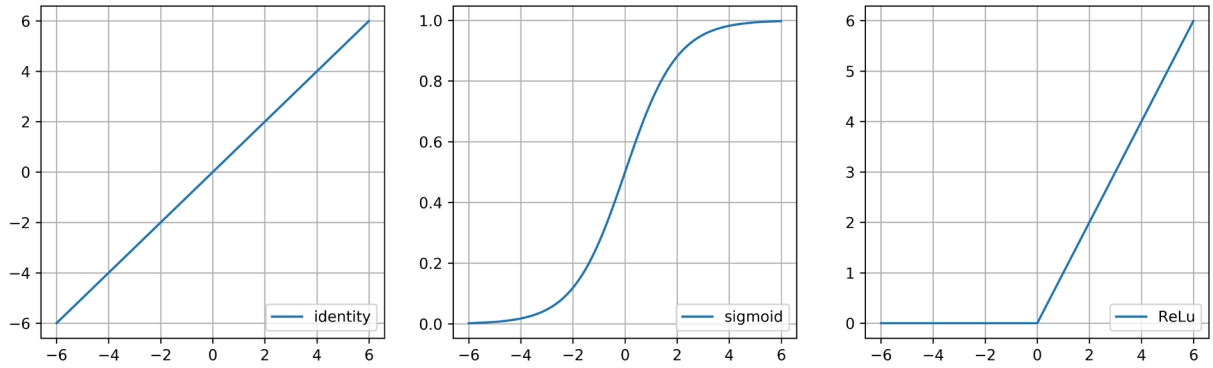

Figure 2-8. Examples of activation function

Deep learning

It has been mentioned on page 4 that adding more layers to a NN which increases levels of abstract will usually improve the performance of NN versus an increasing number of nodes in a single layer. A NN is in general referred to as a DNN if $L > 4$ (i.e., more than two hidden layers), which is the fundamental building block for DL.

Some of the most common architectures of DL include convolutional neural networks (CNNs) [62], recurrent neural networks (RNNs) [63], variational autoencoders (VAEs) [64] and generative adversarial neural networks (GANs) [65]. CNNs are typically designed for image recognition and computer vision applications. They largely reduce the number of free parameters compared to standard fully-connected NNs. They have shown competitive results in medical imaging analysis, including cancer cell classification, lesion detection [66] organ segmentation [67] and image enhancement. RNNs are usually applied for natural language processing (NLP) and audio recognition problems, as they can exhibit temporal dynamic behavior that can be exploited for sequential data analysis [68]. This property also makes RNNs valuable for aiding fractionated radiotherapy, effectively taking advantage of a variety of previously unused temporal information generated during the treatment course. A VAE is an unsupervised learning

algorithm that is able to learn the distribution of compressed data representations from a high-dimension dataset. In other words, it is the equivalent of principal component analysis (PCA) but for DL applications. It can be widely applied in radiation oncology considering the prevalence of high-dimension data due to the limitation of patient sample sizes. Similar to a VAE, a GAN is also a generative model that can learn the multivariate distribution and describe how the data are generated. GANs learn the distribution by an adversarial competition between its generator and its discriminator. They have been successfully applied in some medical imaging tasks, mapping MRI into CT images (synthetic CT) or in adaptive radiotherapy [69] for generating synthetic data and enriching the sample.

Model evaluations

To make the model evaluation meaningful to application in practice, one shouldn't evaluate the model on the dataset on which the model was trained. A complex model may describe data on which it is trained perfectly, but may not perform well on the independent (unforeseen dataset). Alternatively, resampling or cross-validation (CV) is done to evaluate the expected performance of a classifier in unseen datasets. Unless the dataset is large, one can hold out a representative portion of data reserved for testing by randomly sampling or by other criteria that are not susceptible to selection bias.

Bias variance and model complexity

Prediction error is composed of intrinsic noise, variance and bias. Assume one has a response variable Y, and a vector of features X, such that $Y = f(x) + \epsilon$, where we have introduced noise ε satisfying $\mathrm{E}(\epsilon) = 0$. One could decompose the expected prediction error of a regression fit $\hat{f}(x)$ at an input point $X = x_0$ into three terms [70].

$$\boldsymbol{Err}(\boldsymbol{x_0}) = \boldsymbol{E}\left[\left(\boldsymbol{Y} - \hat{\boldsymbol{f}}(\boldsymbol{x_0})\right)^2 \middle| \boldsymbol{X} = \boldsymbol{x_0}\right] = \boldsymbol{\sigma_\epsilon^2} + \boldsymbol{var}[\hat{\boldsymbol{f}}(\boldsymbol{x_0})] + \boldsymbol{Bias^2}[\hat{\boldsymbol{f}}(\boldsymbol{x_0})] \quad \text{Eq. 8}$$

The first term is the variance of intrinsic noise, which is not avoidable in practice. The second term $var\left[\hat{f}(x_0)\right] = E\left[\hat{f}(x_0) - E\hat{f}(x_0)\right]^2$ is called variance, which is the expected squared deviation of the learned model to its mean. The third term $bias^2\left[\hat{f}(x_0)\right] = \left[E\hat{f}(x_0) - f(x_0)\right]^2$ is the squared bias describing how much the expectation of the learned model differs from the ground truth. Typically as model $\hat{f}$ become more complex, bias will be lower but variance will be higher.

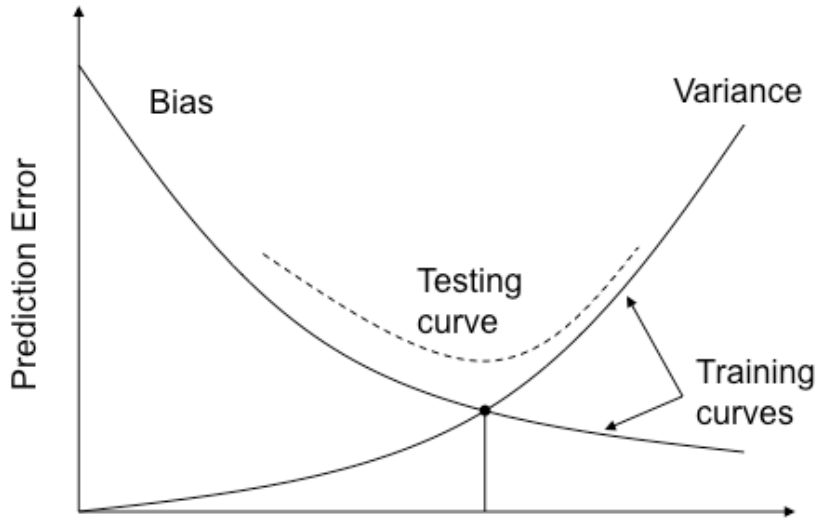

Figure 2-9. The trade-off between bias and variance, adapted from "Building a predictive model of toxicity: methods" [39] by Sunan Cui and et al. in Tiziana Rancati and Claudio Fiorino (eds) in " Modelling radiotherapy side effects: practical applications for planning optimization", 2019, CRC Press and Taylor &Francis Group

Too complex of a model is expected to overfit the data. On the contrary, too simple of a model usually under-fits the data. Generally speaking, one needs to consider the trade-off between variance and bias to choose the 'right' model.

Cross-validation

Cross-validation (CV) is the most widely used method for estimating prediction error. In *K*-fold cross-validation, one splits the data into K roughly equal-sized parts, then for each k^{th} part,

one first fits the model with the rest, $K-1$ of the parts, and then evaluates the model on the k^{th} part. Thus, the model is trained and tested for K times and the average error is calculated as:

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i)) \quad \text{Eq. 9}$$

where $\hat{f}^{-k(i)}$ is the learned classifier without the k^{th} part of the data and L is the designated loss function.

Typically, K is set to be 5 or 10; in the case of $K = N$, the method is known as leave-one-out cross-validation (LOOCV) or Jackknife.

Figure 2-10. K-fold cross-validation adapted from "Building a predictive model of toxicity: methods" [39] by Sunan Cui and et al. in Tiziana Rancati and Claudio Fiorino (eds) in " Modelling radiotherapy side effects: practical applications for planning optimization", 2019, CRC press and Taylor &Francis Group

There is a variant of K-fold cross-validation called stratified (or partitioned) K-fold cross-validation, which takes into account situations of an imbalanced dataset. In the plain K-fold cross-validation, the random division of the data may yield almost no minority data in one subset; and, the performance of the classifier on this subset can be misleading. In stratified K-fold CV, the distribution of classes in each subset is fixed to be the same as that in the whole dataset, which can guarantee a reasonable estimation of error.

Our lung cancer dataset

Lung cancer [71] is the leading cause of cancer deaths in the United States among both men and women. It begins in the lung and can spread beyond the lung in the process of metastasis. Lung cancer is classified into small cell lung cancer (SCLC) and NSCLC for therapeutic purpose. NSCLC which is the focus of our study, accounts for nearly 85% of lung cancer. The three main subtypes of NSCLC are adenocarcinoma, squamous-cell carcinoma, and large-cell carcinoma. Rare subtypes include pulmonary enteric adenocarcinoma. The survival rates for NSCLC decrease significantly due to the advancement of the disease. For stage I, the five-year survival rate is 47%, stage II is 30%, stage III is 10%, and stage IV is 1%. Our study focus on late-stage, stage III NSCLC patients, who would be more likely to have great benefit from personalized treatment.

Our study includes patients were treated with 4 different treatment protocols, in which the two protocols were dose escalation studies that had the total dose increased up to 86 Gy in 30 fractions, the other two protocols were with standard-dose fractionations, had dose up to 74 Gy, 2 Gy per fraction. The decisions regarding dose adaption in the escalating dose protocol was based on PET-CT information during radiation therapy.

Patient specific information

Multiple categories of patient specific information as listed in Table 2-1 were used in the prediction of LC/RP2 in NSCLC patients.

Radiomics features include global features and texture features. Global features are computed from histograms counting the number of gray-levels in 3D space from positron emission tomography (PET) images. Texture features are computed from the gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), gray-level size zone matrix

(GLSZM) and neighborhood gray-tone difference matrix (NGTDM). Compared to global features, these texture features can further integrate intensity and spatial information, accounting for local intensity spatial distribution.

A SNP [72] is a substitution of a single nucleotide that occurs at a specific position in the genome, where each base pair variation is present at a level of more than 1% in the population. SNPs can directly affect protein expression when falling in protein-coding regions, and affect gene splicing, transcription factor and messenger RNA degradation when falling in non-protein-coding region of genes or intergenic regions. SNPs in the human genome have been widely studied to correlate with disease, i.e., toxicity and overall survival in cancer [73] and drug response. In our study, SNPs located on several DNA repair, tumor suppressor, inflammation and transcription factor related genes are considered.

Micro RNAs (miRNA) [74] are small non-coding RNA molecules (consisting of around 22 nucleotides) that functions in RNA silencing and post-transcriptional regulation of gene expression. MiRNAs may function as oncogenes or tumor suppressors, potentially serving as biomarkers in human cancer diagnosis, prognosis and therapeutic targets [75].

Cytokines [76] are a family of signaling polypeptides that are secreted by immune cells that mediate inflammatory and immune reactions. In our studies, cytokines are considered [77] in the prediction of RP which is primarily inflammation of the lung caused by radiation therapy to the chest.

Table 2-1. Patients' information that was applied in outcome prediction in NSCLC patients

Categories	Patient specific information for LC/RP2 prediction
PET Tumor radiomics (43× 2): Global: 4 × 2 GLCM: 8 × 2 NGTDM: 5 × 2	MTV, global.variance, global. Skewness, global.kurtosis, GLCM.energy, GLCM.contrast, GLCM.entropy, GLCM.homogeneity, GLCM.IDM, GLCM. correlation, GLCM.SumMean, GLCM.variance, NGTDM.coarseness, NGTDM.contrast, NGTDM.busyness, NGTDM.complexity, NGTDM.strength,

GLRLM: 13× 2 GLSZM: 13× 2	GLRLM.SRE, GLRLM.LRE, GLRLM.GLN, GLRLM.RLN, GLRLM.RP, GLRLM.LGRE, GLRLM.HGRE, GLRLM.SRLGE, GLRLM.SRHGE, GLRLM.LRLGE, GLRLM.LRHGE, GLRLM.GLV, GLRLM.RLV, GLSZM.SZE, GLSZM.LZE, GLSZM.GLN, GLSZM.ZSN, GLSZM.ZP, GLSZM.LGZE, GLSZM.HGZE, GLSZM.SZLGE, GLSZM.SZHGE, GLSZM.LZLGE, GLSZM.LZHGE, GLSZM.GLV, GLSZM.ZSV
cytokines (30)	EGF, Eotaxin, Fractalkine, GCSF, GM-CSF, IFN-γ, IL10, IL12p40, IL12p70, IL13, IL15, IL17, IL1A, IL1B, IL1Ra, IL2, IL4, IL5, IL6, Il7, IL8, IP10, MCP1, MIP1A, MIP1B, sCD40l, TGF-α, TNFα, VEGF, TGF-β
miRNA (60)	let-7a, miR-100, miR-106b, miR-10b, miR-122, miR-124, miR-125b, miR-126, miR-134, miR-143, miR-146a, miR-150, miR-155, miR-17, miR-18a-5p, miR-192, miR-195, miR-19a, miR-19b, miR-200b, miR-200c, miR-205, miR-20a, miR-21, miR-210, miR-221, miR-222, miR-223, miR-224, miR-23a, miR-25, miR-27a, miR-296, miR-29a, miR-30d, miR-34a, miR-375, miR-423, miR-574, miR-885, miR-92a, miR-93, let-7c, miR-10a, miR-128, miR-130b, miR-145, miR-148a, miR-15a, miR-193a, miR-26b, miR-30e, miR-374a, miR-7, miR-103a, miR-15b, miR-191, miR-22, miR-24, miR-26a
SNPs (55) with its location (Gene)	Rs3857979(BMP1), Rs4988044 (ATM), Rs1800587(IL1A), Rs17561(IL1A), Rs2070874(IL4), Rs1801275(IL4R), Rs4073(CXCL8), Rs2234671(CXCR1), Rs1800896(IL10), Rs3135932(IL10RA), Rs1800872(IL10), Rs11556218(IL16), Rs4760259(GLI1), Rs1799983(NOS3), Rs689470(PTGS2), Rs12102171(SMAD3), Rs6494633(SMAD3), Rs4776342(SMAD3),Rs11615(ERCC1), Rs609261(ATM), Rs12906898(SMAD6), Rs7227023(SMAD7), Rs7333607(SMAD9), Rs664143(ATM), Rs4803455(TGFB1), Rs1061622(TNFRSF1B), Rs664677(ATM), Rs20417(PTGS2), Rs373759(ATM), Rs189037(ATM), Rs12456284(SMAD4), Rs1800057(ATM), Rs3212961(ERCC1), Rs3212948(ERCC1), Rs238406(ERCC2), Rs12917(MGMT), Rs17655(ERCC5), Rs1047768(ERCC5), Rs12913975(SMAD6), Rs1805794(NBN), Rs1625895(TP53), Rs1042522(TP53), Rs25489(XRCC1), Rs9293329(XRCC4), Rs1800469(B9D2&TGFB1), Rs2075685(TMEM167A&XRCC4), Rs25487(XRCC1), Rs1800795(IL6), Rs1799796(XRCC3), Rs1800468(B9D2&TGFB1), Rs1478486(XRCC4), Rs2228000(XPC), Rs2228001(XRC), Rs3218384(XRCC2), Rs1799793(ERCC2), Rs1803965(MGMT), Rs2279744(MDM2), Rs2308321(MGMT), Rs3218536(XRCC2), Rs2834167(IL10RB), Rs3212986(ERCC1)
dosimetric	Information from dose distribution of GTV and lung-GTV structures

Chapter 3 Combining handcrafted features with latent variables

Summary

In this study [32], a novel VAE-MLP joint architecture with multi-omics information as inputs was proposed to conduct dimensionality reduction and prediction in a single step, which potentially eliminates the necessity of feature selection in conventional machine learning. A conventional way of using a separate VAE and classifier was also applied for comparison purposes. Furthermore, the latent variables learned by VAE-MLP were used to compensate traditional feature selection methods in the prediction of RP2.

Introduction

In this study, a combination of handcrafted features and DL latent variables were investigated for the prediction of RP2 in NSCLC patients. Specifically, several MLP-based feature selection methods were investigated together with SVM- and RF-based feature selection methods. Additionally, a novel VAE-MLP joint architectures was proposed to extract latent representation of multi-omics information. Overall, four different strategies as presented in were investigated and compared in the prediction of RP2.

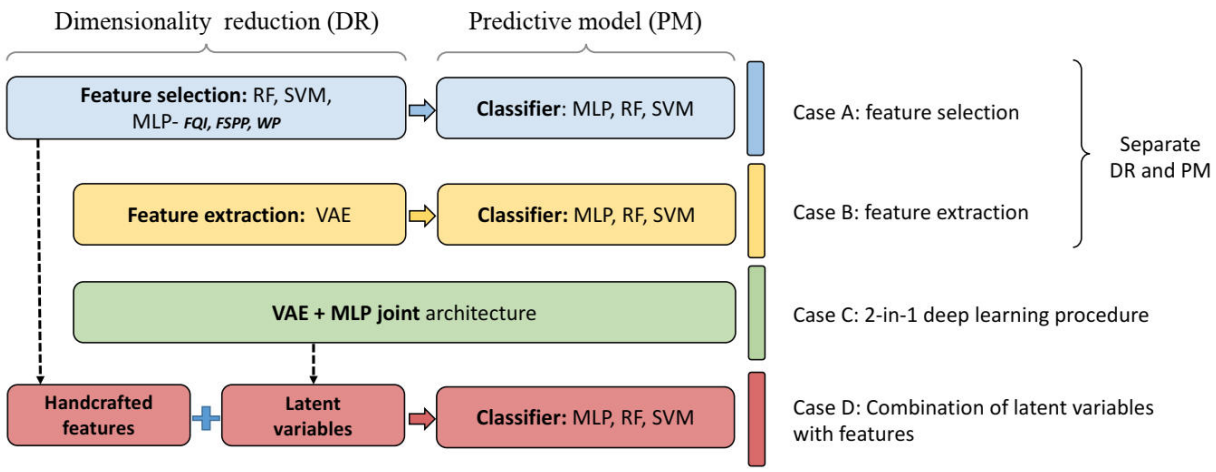

Figure 3-1. Building multivariable predictive models through standard machine learning and DL architectures

Methodology

Three MLP-based features selection methods including weight pruning (WP) [78], feature quality index (FQI) [79] and feature-based sensitivity of posterior probability (FSPP) [80] were investigated in our study.

Weight Pruning (WP)

WP exploits both the weight value and the network structure of an MLP as in Figure 2-6. The score of i^{th} features, $i = 1, \ldots, m$, is calculated by summing up the products of weights over all the paths from feature i to outputs. Specifically, in the single-hidden-layer MLP, the importance is written as

$$S_i = \sum_{j \in \mathcal{H}} \left(\frac{|w_{ji}^1|}{\sum_{i' \in \ell} |w_{ji}^1|} \sum_{k \in \mathrm{O}} \frac{|w_{kj}^2|}{\sum_{j' \in \mathrm{O}} |w_{kj'}^2|} \right) \quad \text{Eq. 10}$$

where $\ell, \mathcal{H}, \mathrm{O}$ denote nodes in the input, hidden and the output layer respectively. And the superscript of weight w denotes layer number. Eq. 10 suggests the weights to be normalized by the sum of weights that are connected to the same input for comparison reason. WP is based on the intuition that important features should result in weights of relatively large magnitude.

Feature Quality Index (FQI)

FQI considers the increase of training mean-squared error (MSE) when a feature is replaced by mean (0 if features are centered). It fixes the trained NN architecture, and replaces the value of a feature by 0, then, calculates MSE based on the output of a new feature matrix.

$$S_i = MSE\ (I_i) - MSE(I_o), MSE(I) = \frac{1}{N}\sum_{\alpha=1}^{N}\sum_{j\in O} \parallel o_{j;I}^{\alpha} - y_j^{\alpha} \parallel^2 \quad \text{Eq. 11}$$

where I_o are the original features, I_i is I_o with i^{th} feature set to be zero and $o_{j;I}^{\alpha}$ is the j^{th} output of input matrix of sample α, I^{α}.

Feature- based Sensitivity of Posterior Probability (FSPP)

FSPP considers the variation of outputs when a feature is randomly permuted among samples. One randomly permutes the i^{th} feature among *N* samples and feeds modified features to the MLP, then calculates the sum of pairwise differences between the new outputs and the original ones. It is based on the belief that "turning off" more important features will influence outputs more.

$$S_i = MSE\ (I_i) - MSE(I_o), MSE(I) = \frac{1}{N}\sum_{\alpha=1}^{N}\sum_{j\in O} \left| o_j^{\alpha} - \ o_{j;I_i}^{\alpha} \right|^2 \quad \text{Eq. 12}$$

where o_j^{α} is the j^{th} output of sample α and $o_{j;I_i}^{\alpha}$ is the j^{th} output of after I_i is randomly permuted among *N* samples.

Random forest (RF)-based feature selections

Random forest which is mentioned **Error! Bookmark not defined.** is an ensemble learning method based on decision trees. In a decision tree, features applied at the upper split

influencing more input samples should be deemed important. As a result, one can estimate the importance of a feature by the fraction of samples the feature contributes to [81].

Support-vector machine (SVM)-based feature selections

SVM which is covered **Error! Bookmark not defined.** can be also used for feature selection. In the case of linear kernels, parameter *w* in the original optimization problem Eq. 6 can be easily recovered after solving the dual optimization problem and used as an estimator of feature importance.

Variational auto-encoder – multilayer perceptron (VAE-MLP) joint architectures

An auto-encoder (AE) is an artificial neural network designed for unsupervised learning. AE consists of two parts: an encoder (ϕ), which compresses an input (from χ) into a lower dimensional space (Z), and decoder (φ) aiming to reconstruct the input out of latent space representation. One notable variant of AE is called a VAE [64] as presented in Figure 3-2, which inherits the AE architecture but incorporates uncertainties through a stochastic variational approach into the deterministic AE. In this setting, the encoder first produces two vectors μ and σ describing the mean and the variance of the latent state distribution and then generates a latent vector by sampling from this distribution. Subsequently, the decoder receives the latent vector to reconstruct the original input. The total loss of a VAE Eq. 13 is composed of two terms, the first term stands for reconstruction error and the second term is Kullback-Leibler (KL) [82] divergence metric, which acts like penalty term. VAE can be applied to extract features that were used in subsequent classification problems.

$$L(x_{enc}, x_{dec}) = \| x_{enc} - x_{dec} \|^2 + \frac{1}{2}\sum_{j=1}^{J}[1 + \log(\sigma_j^2) - \sigma_j^2 - \mu_j^2] \quad \text{Eq. 13}$$

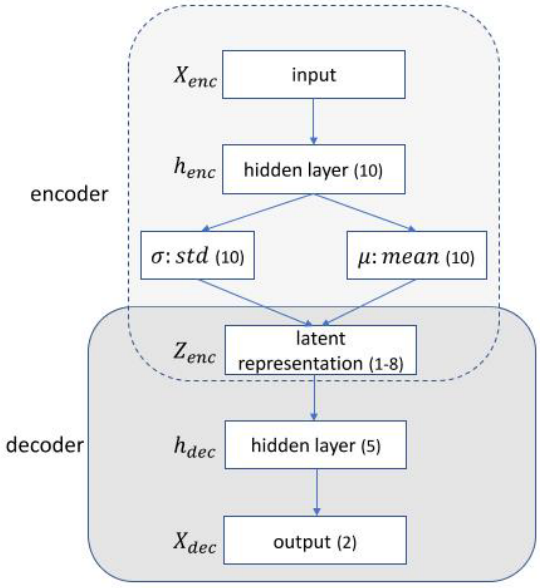

Figure 3-2. Diagram of a VAE with number of nodes (*) in the implemented architecture are denoted

A novel joint architecture of VAE and MLP as shown in Figure 3-3 was proposed to conduct dimensionality reduction and prediction tasks simultaneously, realizing efficient representation learning aided by the classification task. The total loss function of the architecture is composed of VAE loss and prediction loss (binary cross-entropy). An extra coefficient λ was used to magnify prediction loss for the trade-off between VAE loss and prediction loss.

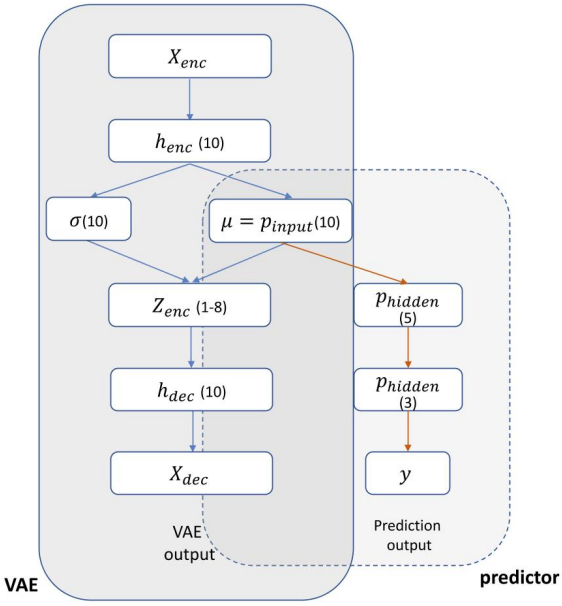

Figure 3-3. Diagram of a VAE-MLP joint architecture: μ is used as input of MLP classifier, the number of nodes in each layer is given.

Feature ranking aggregation

Due to the noisy nature of our dataset, the rankings were very sensitive to which portion of the data generated the ranking. As a result, multiple rankings (e.g., 100) based on different subsets of the data were generated and aggregated to yield a single ranking. Finally, several top features were fed into the designated classifier for evaluation of performance. Kemeny aggregation was applied in this process to summarize feature ranking. Kemeny aggregation gets an optimal ranking by minimizing a sum of Kendall τ distances $K(\tau_1, \tau_2)$, which is defined by the number of pairwise disagreements between any two ranking lists,

$$\min_{\pi} \sum_{i=1}^{B} K(\pi, \tau_i) \qquad \text{Eq. 14}$$

$$K(\tau_1, \tau_2) = |\{(i,j) | \forall\, i < j, [(\tau_1\,(i) < \tau_1\,(j)) \wedge (\tau_2\,(i) > \tau_2\,(j))] \vee [(\tau_1\,(i) > \tau_2\,(j)) \wedge (\tau_2\,(i) < \tau_2\,(j))]\}| \qquad \text{Eq. 15}$$

As one may see, the direct computation of Kemeny aggregation using Eq. 14 and Eq. 15 can be burdensome when the lists are long. In fact, it is proven to be an NP-hard problem (at least as hard as nondeterministic-polynomial-time problem). Fortunately, it can be converted into an equivalent graph problem for computational convenience [83].

Performance evaluation and TRIPOD level 2 Nested Cross-Validation (CV)

Performance evaluation

A receiver operating characteristic curve (ROC) [84] was utilized for evaluating the performance of our predictive models at various classification thresholds. The abscissa of the ROC curve is false positive rate (FPR) and the ordinate of the ROC curve is true positive rate (TPR).

$$TRP = \frac{TP}{TP + FN} \quad \text{Eq. 16}$$

$$FPR = \frac{FP}{FP + TN} \quad \text{Eq. 17}$$

, where *TP* refers to the patients with events and were classified as positive, *FN* refers to the patients with events but were classified as negative, *FP* refers to the patients without events but were classified as positive, and *TN* refers to the patients without events and were classified as negative. Each point in the ROC curve stands for a pair of TPR and FPR given a designated threshold (*S*). If the prediction of a patient $P > S$, the patient was classified as positive, otherwise, the patient was classified as negative. The area under the ROC (AUC) curve is an overall evaluation of classification performance at various thresholds. It has a range of 0.5-1. AUC of 0.5 means a random classification while AUC of 1 means a perfect classifier.

Validation

In this study, four different cases as presented in Figure 3-1 were investigated for prediction of RP2. Case A is based on traditional feature selection methods e.g., MLP-based, SVM-based and RF based-methods and subsequent classification. Case B is based on traditional features extraction methods VAE and subsequent classification problems. Case C is based on the novel VAE-MLP joint architecture which combined dimensionality reduction and prediction into a single step. Case D is based on the combination of handcrafted features and latent variables from VAE-MLP joint architectures.

```
for i in 1-10
    5-fold stratified CV on the whole dataset-> Otrain, Otest (outer-loop training and test set)
    for each pair of (Otrain, Otest) do
        on Otrain:
         1. pre-selection of feature, criterion: single variable AUC >0.6
         2.1 case A:
            rank features in multiple inner-loop CVs -> rank aggregation-> top features
            select classifier parameters in inner-loop CVs-> optimal parameters
            train model with optimal parameters and selected top features
         2.2. case B:
            train VAE-> latent variables
            select classifier parameters in inner-loop CVs-> optimal parameters
            train model with optimal parameters
         2.3. case C:
             select architecture parameters in inner-loop CVs-> optimal parameters (dropout rate)
             train model with optimal parameters
         2.4 case D:
              combine handcrafted features from case A and latent variables from case C
              select classifier parameter in inner-loop CVs-> optimal parameters
              train model with optimal parameters.
        on Otest:
             evaluate test AUC for case A, B, C, D
    end for
end for
average test AUCs
```

Figure 3-4. Nested CV in the validation process for evaluating for cases A, B, C and D

For comparison purposes and mitigating statistical bias, we implemented all four methodologies (A, B, C, D) in the same validation pipeline Figure 3-4. This is referred to as a type 2b analysis in the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement [85]. Specifically, nested CVs were performed, where the feature and parameter selection were tuned in inner-loop CVs, and the model with

optimal parameters was then identified from outer-loop training sets and evaluated on the outer-loop test sets. Multiple times of CV were performed to consolidate the results.

Results

ROC curves were obtained based on predictions by different methods. AUCs with 95% CI were then calculated and presented in Table 3-1.

Table 3-1 Summary of performance results in the four strategies

Methods	AUC	Delong test
Case B (VAE+MLP)	0.624 (95%CI: 0.577-0.658)	p-value: 1.33×10^{-7}
Case C (VAE-MLP)	0.781 (95%CI: 0.737-0.808)	
Case A (WP+MLP)	0.804 (95% CI: 0.761-0.823)	p-value: 6.42×10^{-4}
Case D ((latent Z+WP)+MLP)	0.831 (95% CI: 0.805-0.863)	

Case A. Conventional machine learning feature selection and prediction

5 different feature selection methods including WP, FQI, FSPP, RF, SVM were applied to select top features, and then 3 different classifiers MLP, RF and SVM were applied to build predictive models with results shown in Figure 3-5. Generally, cross-validated AUCs will first increase with an increasing number of features and then decrease. Particularly, WP+MLP outperformed the rest of the combinations for feature selection and prediction in the range from 23 to 36 features. With the top 29 features, WP+MLP was shown to reach the highest AUC of 0.804 (95% CI: 0.761-0.823).

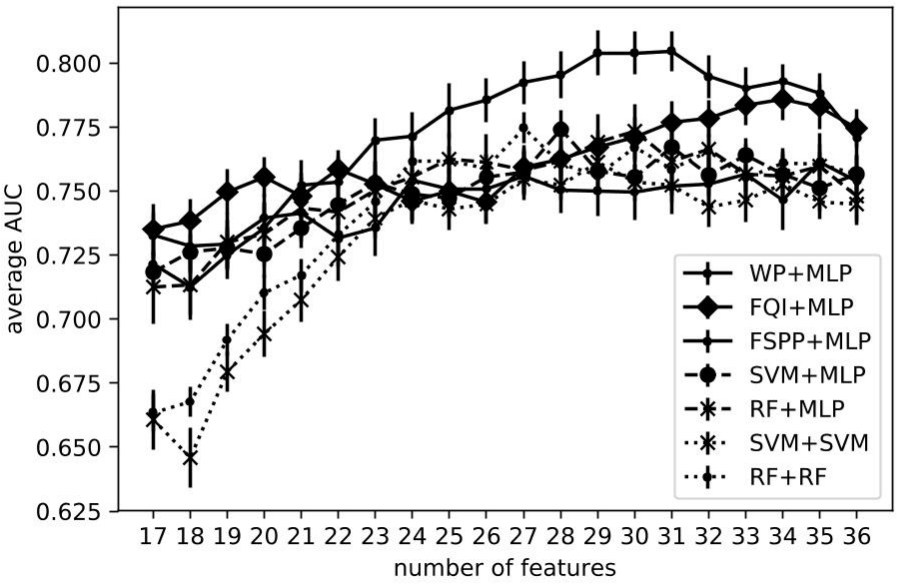

Figure 3-5. AUC trend with an increasing number of features

To analyze feature importance in this study, we considered the final ranking lists in the collection of all (50; 10 times outer-loop CV) iteration. Particularly, the frequency of a feature was included in the final set was obtained and served as an indicator of relative importance. Table 3-2 shows the summary of important features that were selected more than half of the times.

Table 3-2 Features being selected more than half of the times (the bold ones were selected every time)

Categories	Names
Dosimetric information (1)	**Mean Lung Dose**
Cytokines (10)	**2w_eotaxin**, 4w_eotaxin, pre_TNF-α, 2w_TNF-α, 4w_TNF-α, 2w_IL-8, 4w_IL-8, 2w_ MCP-1, 2w_fractalkine, pre_IFN-γ,
miRNA (8)	**hsa-miR-192-5p**, hsa-miR-22-3p, hsa-miR-128, hsa-miR-15a-5p, hsa-miR-223-3p, hsa-miR-23a-3p, hsa-miR-210, hsa-miR-100-5p
SNPs (9)	**Rs3857979(BMP1), Rs238406(ERCC2), Rs12456284(SMAD4), Rs1625895(TP53),** Rs1799983(NOS3), Rs4803455(TGFB1), Rs25487(XRCC1), Rs1800468(TGFB1), Rs2075685(XRCC4)

Case B and Case C: the comparison of separate VAE and classifiers versus VAE-MLP joint architectures.

A comparison of prediction results from case B of separate VAE and classifiers (MLP, SVM, RF) and case C of a VAE-MLP joint architecture is shown in Figure 3-6.

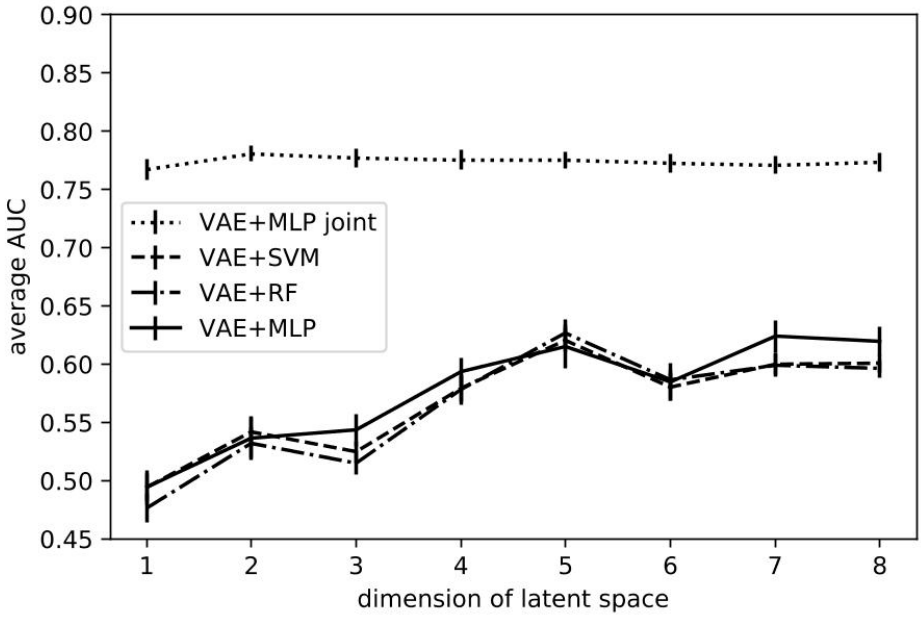

Figure 3-6. The comparison of performance by separate VAE and classifiers and VAE-MLP joint architectures

, where the dimension of the latent space varies from 1 to 8. It shows our proposed joint architectures yield better performance than conventional methods of separate VAE and classifiers with various numbers of latent space. In the joint architecture, two dimensions were sufficient to encode the original inputs for this classification problem, reaching an average AUC of 0.781 (95% CI:0.737-0.808). Patients were represented on the 2-D latent space (Z) as points. Examples from some randomly selected outer-loop CVs are shown in Figure 3-7. Two classes RP=0 and RP=1 are clearly separable in training data (red dots versus blue dots) and are partially differentiated in the test data (yellow dots versus green dots) in Figure 3-7.

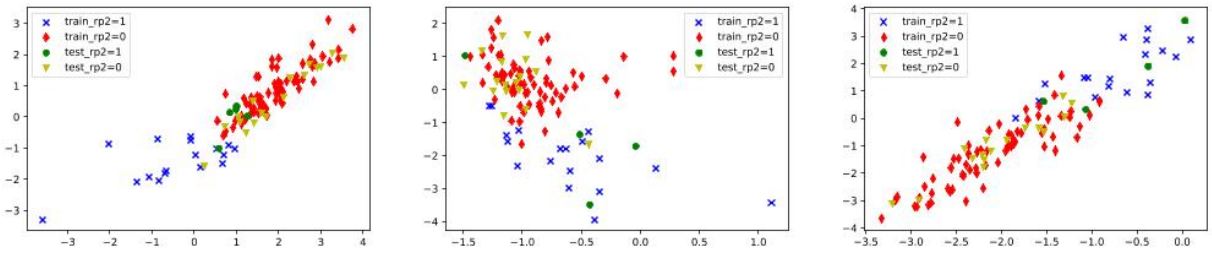

Figure 3-7. Visualization of latent variable Z of patient

Case D: Combination of features selected by WP and latent variables from VAE-MLP joint architectures

Here, the selected features by WP (case A) and the latent representation from VAE-MLP joint architecture (latent size=2) (case C) were combined and used as inputs in MLP, SVM and RF classifiers for RP2 prediction. The resulting AUCs were shown in Figure 3-8, together with AUCs of case A in for comparison purposes. Better predictive performance was achieved by combining the selected handcrafted features and the latent representation by VAEs, which is especially in the case of a small number of samples. The improvement may be because that the latent representation which takes all features into account, can compensate for the incomplete discrete representation by the handcrafted feature selection algorithms. When only a small portion of features are available for the predictive model, the complementary information was distinctively useful for such a heterogeneous data modeling problem. However, it is our conjecture that with more data samples become available, case C may supersede handcrafted features to eliminate the necessity of such a combination.

Conclusion

This work demonstrates the potential for a combination of traditional machine learning methods and DL VAE techniques in dealing with limited datasets for modeling radiotherapy toxicities. Specifically, the combination of selected features from MLP-based method WP and

latent variables from VAE-MLP joint architecture (case D) yielded the highest AUC compared to the AUCs by either handcrafted features (case A) or latent variables (cases B, C) individually.

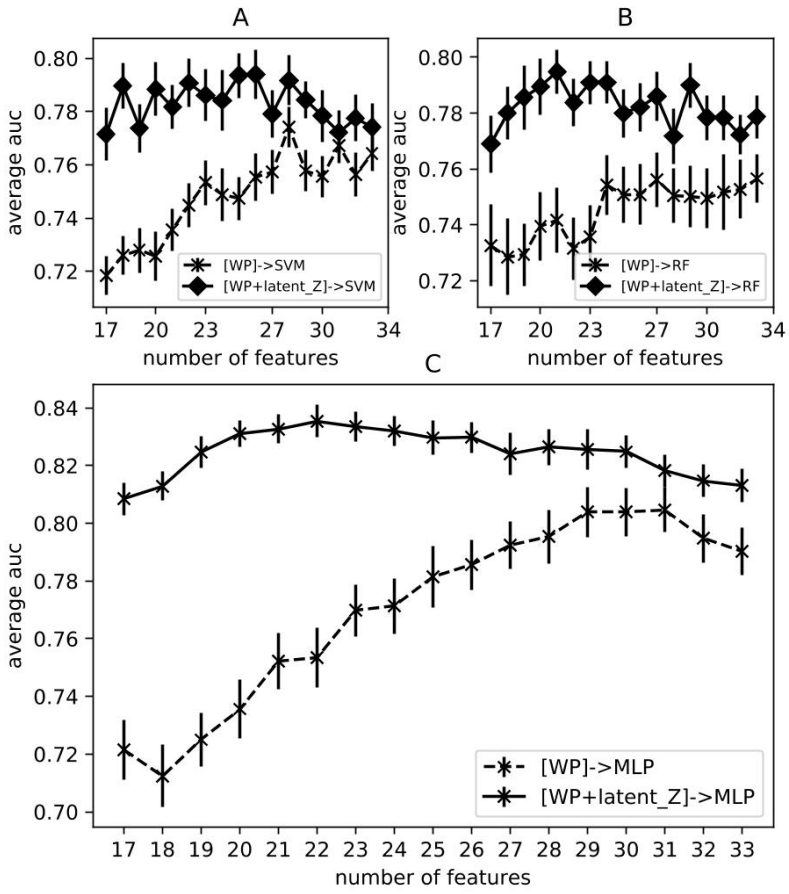

Figure 3-8. AUCs by combining WP features with latent Z in SVM (A) RF (B), MLP (C) classifiers

Chapter 4 Considering temporal associations among variables

Summary

In this study [34], temporal associations among biological and imaging information were modeled in specific neural network layers, i.e., 1D convolutional layer and 1D locally-connected layers. Compared to a fully-connected layer, these layers which fit the nature of longitudinal data can reduce the degree of freedom in the NNs and hence mitigate overfitting and improve generalization to unforeseen dataset. In order to model both longitudinal data and non-longitudinal data, a composite architecture was proposed.

Introduction

The application of artificial neural networks (ANNs) with composite architectures into the prediction of local control (LC) of lung cancer patients after radiotherapy was investigated. The motivation of this study was to take advantage of the temporal associations among longitudinal (sequential) data to improve the predictive performance of outcome models under the circumstance of limited sample sizes. Two composite architectures: (1) a one dimension (1D) convolutional + fully connected and (2) a locally-connected+ fully connected architectures were implemented for this purpose.

Methodology

Convolutional layer and locally-connected layer

Unlike fully-connected layers in MLP, where all the nodes in adjacent layers are connected, there is another class of NN where nodes are “partially-connected”, e.g., locally-connected layer and convolutional layer as presented in Figure 4-1 and Figure 4-2. In a locally-connected layer, when calculating the value of a node, only the value of adjacent nodes in the previous layer will be summed up. In a convolutional layer [33], a weight sharing scheme will be further applied to make it shift-invariance.

One can understand that using kernels in NNs as being equivalent to template matching or seeing" local information of a neighborhood while blocking information from far apart or less related regions, as depicted in Figure 4-2 (left) This can be also visualized by vectorizing the input and the output as in Figure 4-2 (right) from which one can realize locally-connected layer only connect to certain nodes within a layer when compared to a fully-connected neural network. Overall, a locally connected neural network only considers local relations (receptive field) while it decouples information far away in space and/or time allowing for efficient data representation and improved task learning. Such “partially-connected” properties will help consider associations among biological, imaging dosimetric and physical data, reducing free parameters in the architectures. Specifically, in our study, these partially-connected architectures were applied to account for longitudinal associations.

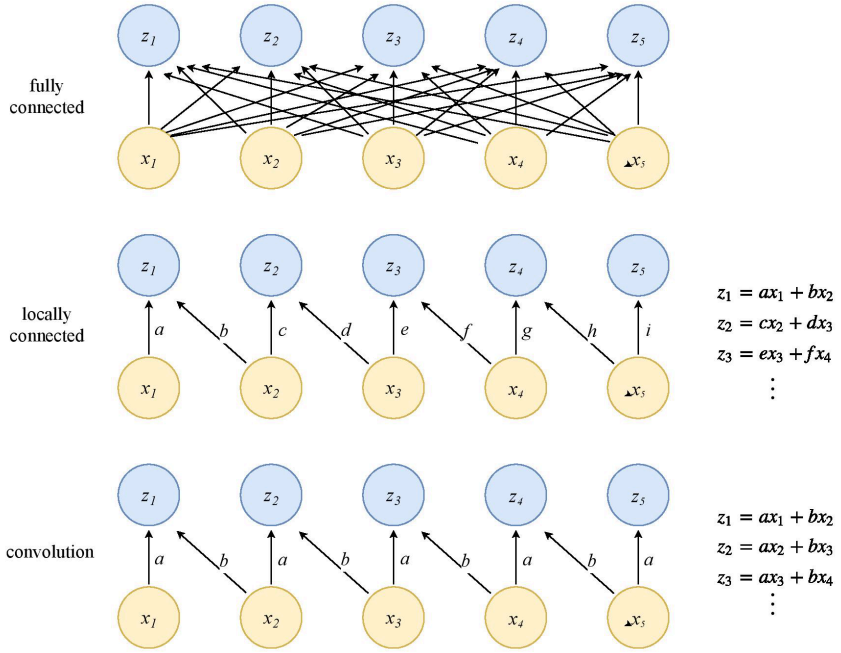

Figure 4-1 Diagram of connection in fully-connected, 1D locally-connected and 1D convolutional layers

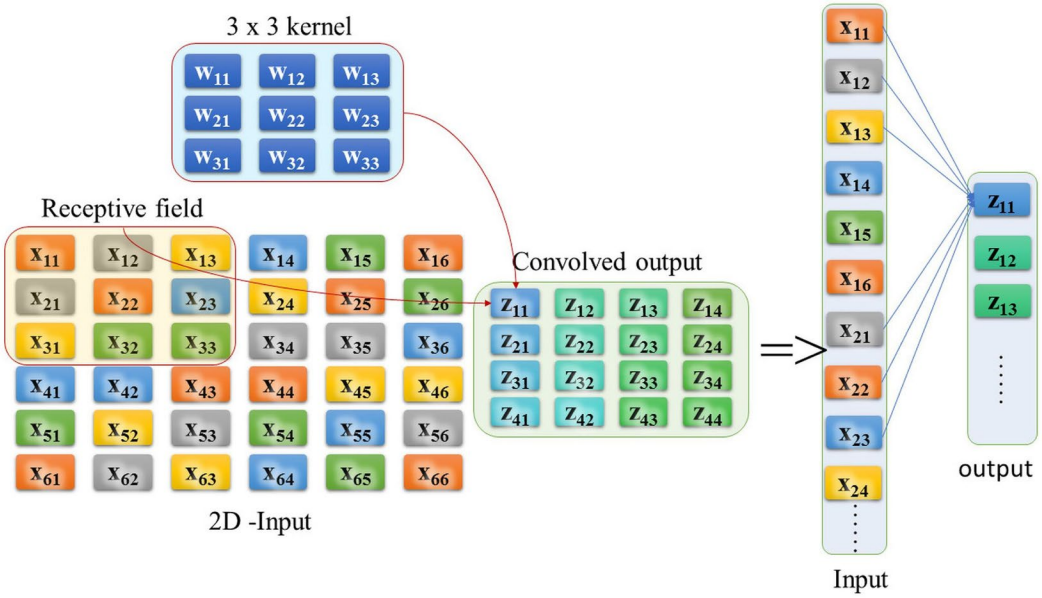

Figure 4-2. The diagram of connection in a 2D locally-connected architectures

In a 1D locally-connected layer Figure 4-3, a, b, c denotes three different variables, and each was measured for three times during the span of the treatment. In this layer, only the inputs

that are corresponding to measurement of the same variable will be connected to the same output. In this case, the kernel size is 3×1 and stride size is 3 which the number of variables. One can also add more number of channels to increase the versatility of the locally-connected layers.

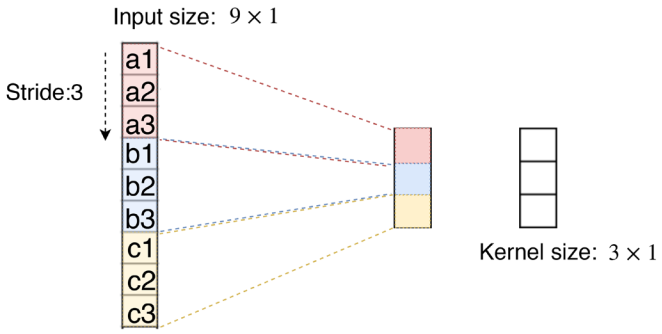

Figure 4-3 An illustration of a locally-connected layers with stride=3 for a 1D input

Figure 4-4 illustrates an example of applying a 1D convolutional layer to a 2D fictitious input whose size is 11×5, where 11 and 5 are lengths in temporal (vertical) and non-temporal axis (horizontal) respectively. It is worth noting that "1D" means a convolution kernel is convolved with the input over 1D to produce outputs, it does not put any constraint to the input size dimension. For a 2D input, a 2D convolutional layer is automatically reduced to 1D convolution when the width of a kernel equals to the width of inputs. Consequently, the width of a filter should be fixed to 5 in order to perform 1D convolution, while the length of a filter (say 3) can be arbitrarily assigned. This layer has a filter of size 3×5 sliding along the time-axis by 1 pixel (stride) at a time. In each position, an element-wise multiplication of the input patch and the filter is carried out. The corresponding results would then be one of the output neurons.

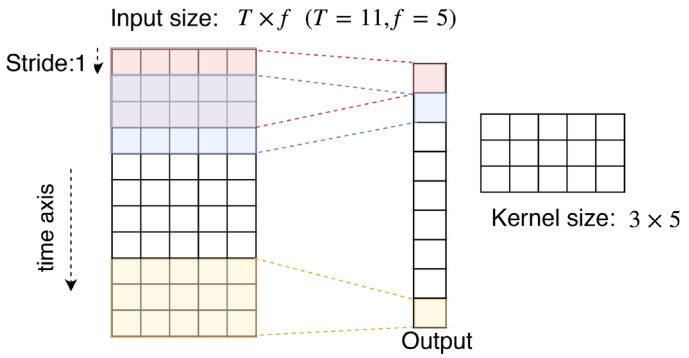

Figure 4-4. An illustration of a 1D convolutional layer for a 2D input, to which kernels of size 3x5 were applied where T represents the number of time points and f denotes the number of variables

Dropout

As the sample size is limited, overfitting can be a main concern. Intuitively, regularization techniques can resolve this issue by suppressing the noise in the training data. As a result, dropout, a neuroscience-inspired trick [86] was applied in our study. During the training, dropout will randomly select some portion (dropout rate, e.g., 20-50%) of nodes being ignored. They will not affect updated weights, as their contribution to the activation of downstream neurons is temporally removed. Indeed, dropout is currently a very effective ensemble method, performing averaging with NNs while mitigating the risk of memorizing the data. Hence, the resulting NN is capable of better generalization to unseen data and is less likely to over-fit (memorize) the training data.

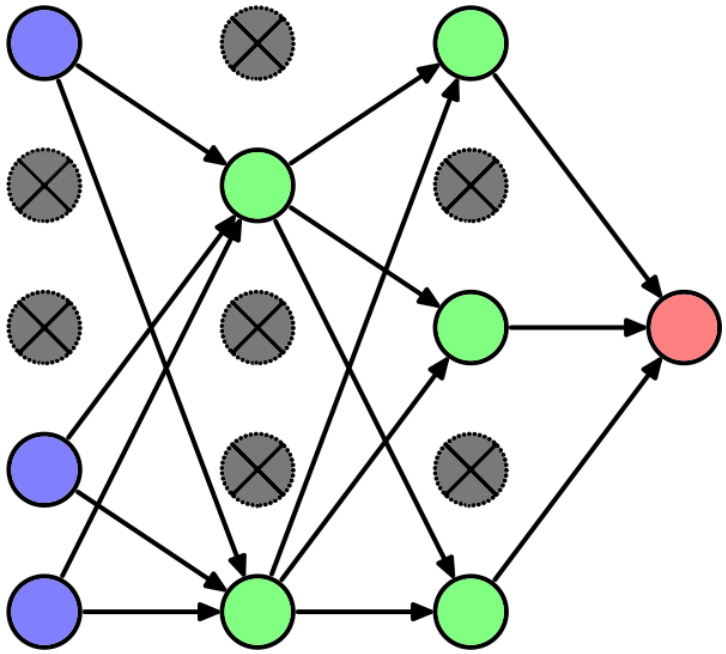

Figure 4-5 The diagram of dropout technique applied in a fully-connected layer

Composite architectures A, B designed for modeling both longitudinal and non-longitudinal data

In our study, 18 features as in Table 4-1 are considered in predictive models including cytokines, SNPs, miRNA, dosimetric data and PET radiomics. Level of cytokines were measured pre-treatment, 2-week and 4-week during the treatment. PET radiomics were collected pre-treatment and mid-treatment. Mean doses before and after the adaptation were both recorded. Composite architecture of 1D locally-connected layer/ 1D convolutional layers and MLP was proposed to take account of both longitudinal and non-longitudinal data.

In the architecture A Figure 4-6, which is composed of convolutional layers and MLPs, the longitudinal data were first fed into 1D locally-connected layer, and then the reduced representation was concatenated with the rest of non-longitudinal data into a single vector which was subsequently fed into 2 fully-connected layer for the prediction of LC. A similar architecture B that composed of 1D locally-connected layer and MLP was also applied for the LC prediction.

Note that separate convolutional layers were considered for cytokine and other data because their time steps were not consistent. However, this separation did make sense, since variables of the same categories were usually more correlated with each other than variables of different categories. To better visualize this, a heat map of the correlation matrix of longitudinal data is shown in Figure 4-7. From the heat map, one can easily find distinguishable groups that would rationalize the separation assumption applied here.

Table 4-1 Features that were applied for LC prediction.

Categories	**Variables**
Biological data: cytokines (3*) SNPs (1) MiRNA(1)	Eotaxin, interlukin-1-α
	SMAD9
	145-5p, 574-5p, 122-5p
Dosimetric data (2)	Mean tumor doses
PET image data (2)	GLSZM.ZP,GLSZM.LGZE, GLRLM.GLN

* The number in the brackets following the categories of variables denotes how many times the same variable was measured over in the treatment. 3* denotes longitudinal data.

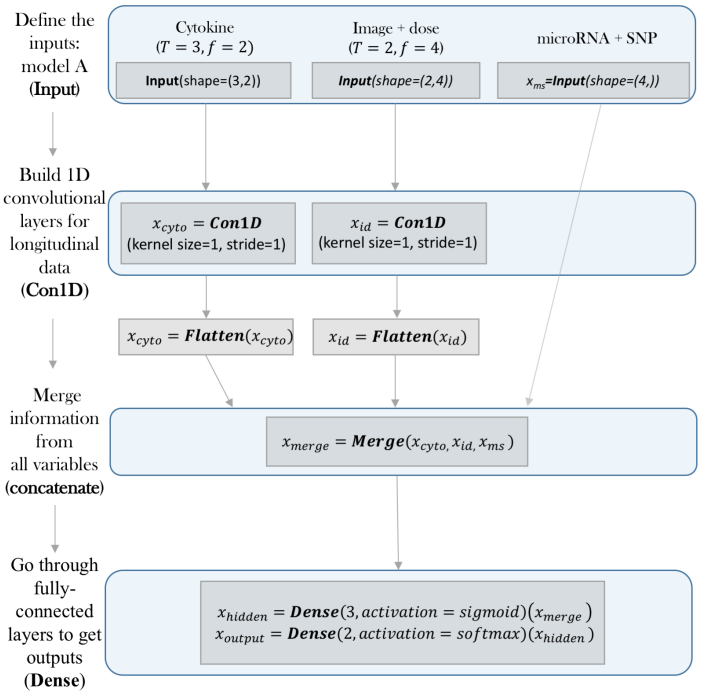

Figure 4-6. The diagram of architecture A built base on Keras (built-in functions applied: Input, Con1D, Flatten, Concatenate, Dense). In the input bock, T and f represent the number of time points and the number of variables at each time point respectively.

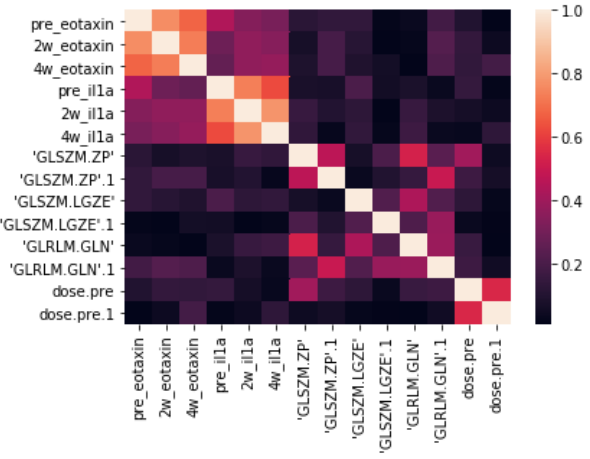

Figure 4-7. Absolute value of Spearman correlation among predictive features

Performance evaluation

In this study, AUCs as mentioned on page 34 as a discrimination measurement and Brier scores which is a calibration criterion was used to evaluate the performance predictive models. The most common formulation of the Brier score is defined as

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \qquad \text{Eq. 18}$$

, where p_i and o_i are the predictive result and true label of patient i respectively.

Results

The proposed neural network architectures were implemented with the Python DL library Keras [87]. The RMSprop optimization [88] method was implemented for the estimation of the network weights. 20 times of five-fold CVs were performed to evaluate the predictive models' performance and assess their generalizability. The oversampling technique synthetic minority over-sampling (SMOTE) [89] was applied to the training data to mitigate the class imbalance problem. Architecture C, an MLP were also evaluated for a comparison purpose.

Predictive performance

Table 4-2 is a summary of predictive performances of the three architectures A, B and C with their corresponding number of parameters (*n*). A single AUC in this table is an average of test AUCs from 25 times of 5-fold cross-validation (the random seed for a partition of folds was changed each time). Brier scores (BS) [90], as a measure of the accuracy of prediction performance based on the test prediction and test labels in those cross-validations without SMOTE (for a fair comparison), were also evaluated.

Table 4-2 Cross-validated AUC predictions of LC in architectures A, B and C. The activation applied for the convolutional layers in architecture A and the size of a kernel of the convolutional layer for cytokines were shown in the table. For reference, Brier score of null models where all the patients were given an LC probability as population LC rates is 0.209.

Architecture	Number of free parameters	AUC with 95%CI	Brier score
A: "ReLU" 1 × 2	54	0.785 (0.752-0.792)	0.189
A: "ReLU" 2 × 2	56	0.786 (0.757-0.796)	0.189
A: "linear" 1 × 2	54	0.814 (0.787-0.823)	0.182
A: "linear" 2 × 2	56	0.812 (0.779-0.820)	0.182
A: "sigmoid" 1 × 2	54	0.832 (0.807-0.841)	0.157
A: "sigmoid" 2 × 2	56	0.829 (0.804-0.838)	0.160
B	60	0.802 (0.775-0.811)	0.161
C	231	0.778 (0.751-0.790)	0.190

*A lower Brier score indicate a better performance.

In general, architecture A which applied a 1D convolutional layer for longitudinal data yielded the best performance. Among the three implemented activations, "sigmoid" activation showed the best AUC of 0.83, "linear" activation also achieved a decent AUC of 0.81. However, the "ReLU" activation did not work well in our case, achieving an AUC of only 0.79. The size of the kernel, specifically the length of the kernel which determined how many time steps would be considered in a single kernel almost did not affect the results. The AUCs under the two different

settings of the length of a kernel were roughly the same (e.g., DeLong test of architecture A with "sigmoid" activation showed a p-value 0.814). Architecture B, which applied locally-connected layers to the longitudinal data yielded an AUC of 0.80. This slightly outperformed the architecture C MLP in the prediction performance (AUC of 0.78).

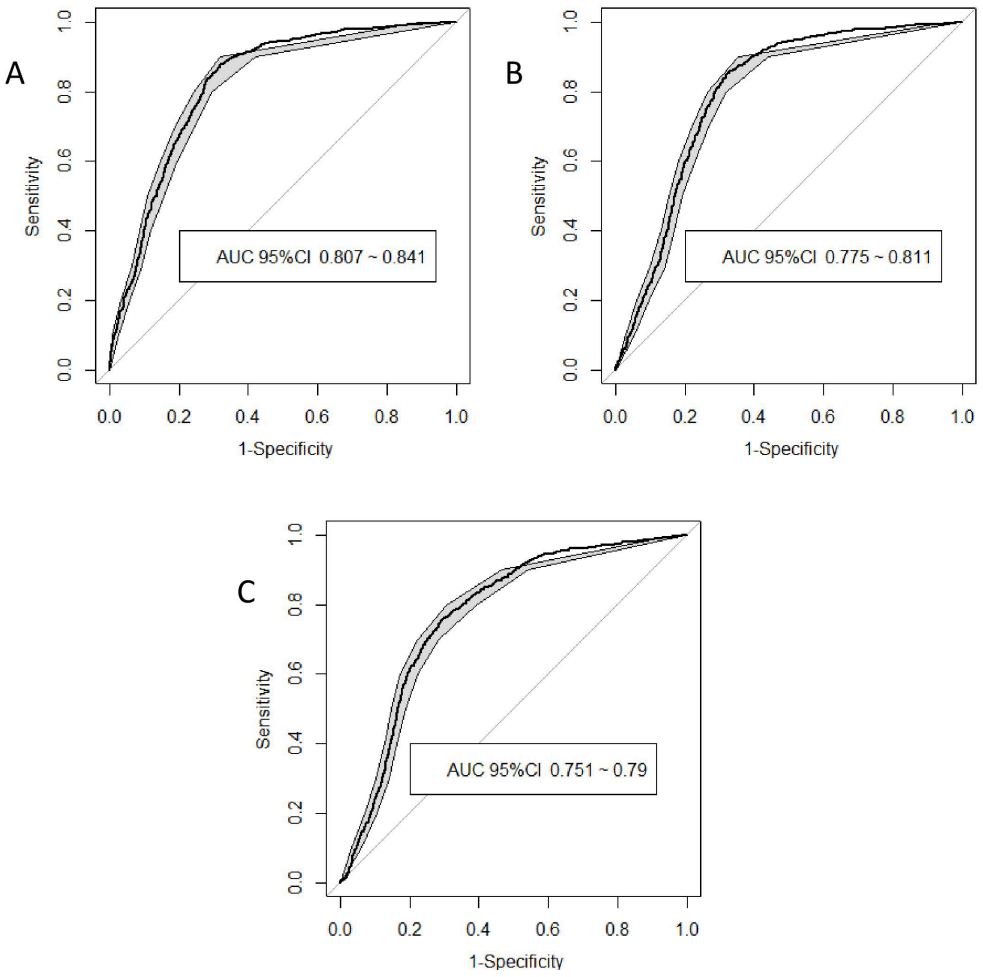

Figure 4-8. Cross-validated AUC of architecture A, B and C

Analysis of trained architecture

Further analysis of weights gives one some insights about how the architecture extract information for prediction. The weights of locally-connected layers in architecture B for cytokines are shown in Table 4-3. as well as Spearman correlation coefficients between LC and those variables. Clearly, the effect of the locally-connected layer is to weigh the cytokines at

different time points, likely giving more weight to the more relevant (to LC) time point. The relative change of cytokines is not as considered as all the weights are with the same sign. This is consistent with the fact, the change of cytokines showed a weak correlation with LC, as in Table 4-3. "Change_corr" row, where the correlation between LC and $2\_week - pre$, $4\_week - 2\_week$ cytokines are shown. Similarly, weights corresponding to image features and dose inputs were also shown in Table 4-3. For variables GLSZM.ZP and GLRLM.GLN, clearly, weights corresponding to pre-measurement are far larger than those of mid-measurement, which is consistent with results of Spearman correlation, where the mid-measurement of the two variables showed weaker correlation with LC (correlation coefficient -0.07 and 0.008, respectively.) For dose and GLSZM.ZP whose pre- and mid-measurement both yield relatively higher correlations, weights seem to be roughly equally given to the two time points. In general, locally-connected layers were able to extract important and relevant information from the inputs to eventually aid the prediction task, which are consistent with individual correlations in this data.

Table 4-3. Weights corresponding to cytokines, dose and image features as well as Spearman rank correlation between raw values of cytokines (Corr); and the change of cytokines (Change_corr)

		pre	2_week	4_week
Eotaxin	Weights	-0.25	-0.05	-0.21
	Corr	-0.176	-0.136	-0.175
	Change_corr	-0.013	- 0.035	
Interlukin-1-α	Weights	-0.10	-0.07	-0.18
	Corr	-0.210	-0.175	-0.216
	Change_corr	0.05	-0.076	
		pre	mid	
Mean tumor doses	Weights	0.28	0.22	
	Corr	0.273	0.231	
	Change_corr	- 0.016		
GLSZM.ZP	Weights	0.42	0.52	
	Corr	0.148	0.123	
	Change_corr	-0.122		

GLSZM.LGZE	W	-0.44	-0.01
	Corr	-0.18	-0.07
	Change_corr	-0.070	
GLRLM.GLN	Weights	0.84	0.11
	Corr	0.10	0.008
	Change_corr	0.072	

Survival analysis of local control

It is worth noting that although we regarded the outcome prediction task here of LC as a binary classification problem, our proposed method is still valuable for survival analysis, where time-to-event is considered. In the medical field, survival analysis is a desirable tool to estimate the time of death, relapse or development of an adverse reaction for a group of patients. In this study, we adopted Kaplan-Meier estimator using the R package prodlim [91] to estimate local progression (LP) free (e.g., LC=0) survival probability for patients in the cross-validated test sets as predicted by architecture A outputs. The threshold between the two groups was set such that the number of high-risk patients equals to patients with LP in the original dataset to mitigate data imbalance issues. As shown in Figure 4-9, as expected, the patient group with the higher risks of LP has poorer performance over time. The difference between the two groups is distinguishable across all time points and was confirmed by log-rank test (p-value <0.0001).

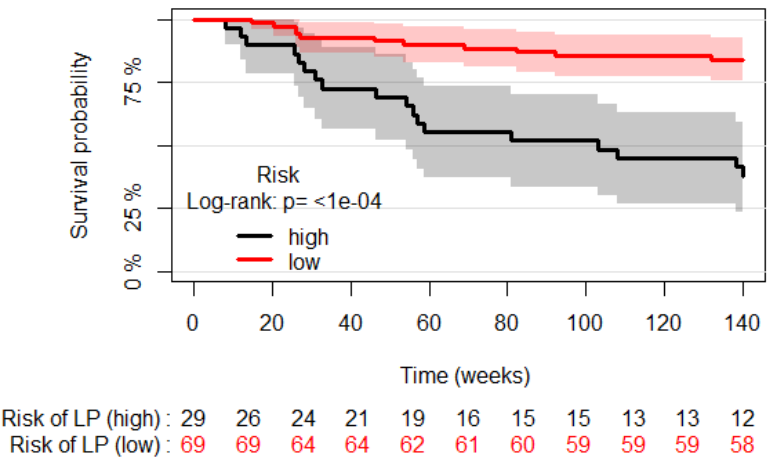

Figure 4-9. Survival curves with 95% CI for patient groups defined by 1DCNN-MLP outputs (threshold=0.397). (Based on survival time, true labels and predictions by architecture C for test patients in the cross validation

Conclusion

This work demonstrates the potential of 1D convolutional layers and 1D locally-connected layer in modeling temporal associations. The proposed composite architectures managed to model both longitudinal data and non-longitudinal data with a reduced number of parameters. The proposed models yield better performance than plain MLPs which ignore such associations.

Chapter 5 Joint actuarial prediction

Summary

In this study, proposed architectures took into account complex interactions among biological, imaging and physical variables, conducting dimensionality reduction and prediction simultaneously. Moreover, time-to-event information is considered for actuarial prediction. Multi-endpoints prediction i.e., joint prediction of RP2 and LC was considered in a single architecture.

Introduction

In this study, (a) an actuarial DNN (ADNN) architecture based on dosimetric information *ADNN-DVH* was proposed to extract morphologic characteristics of DVH and predict LC and RP2. Compared with analytical Lyman models and log-logistic models, it further accounts for information other than "average" doses (gEUD) and can reveal non-linear associations between DVH and endpoints. (b) An ADNN architecture *ADNN-com* that integrates complex interactions among biological, imaging and dosimetric information was proposed to predict LC and RP2 respectively. (c) An ADNN architecture *ADNN-com-joint* was proposed to integrate different categories of patient specific data as well as to do multi-endpoints prediction, i.e., prediction of LC and RP2 were generated simultaneously from a single architecture. Additionally, temporal information was considered in all of our proposed outcome models. This allows the prediction of time-to-event in addition to only classifying events, as commonly practiced in the outcome modeling literature.

Methodology

Generalized Lyman and log-logistical models

For comparison purposes, analytical Lyman and log-logistic models were also conducted to predict RP2 and LC. As described on page 16, in these models, simple functions e.g., CDF of a Gaussian distribution and a rational polynomial function are chosen to represent the dependence of NTCP and TCP on normal tissue dose and tumor dose. A generalized Lyman models [92] can further take into account time-to-event/censored time. The probability that a complication is observed during a given follow up time τ is calculated as a product of a patient's NTCP (as in standard Lyman model Eq. 4) and the conditional probability that the patient experiences toxicity at time τ given that toxicity will eventually occur, as shown in Eq. 19.

$$NTCP(D, D_{50}, m, n)F(\tau) \quad \text{Eq. 19}$$

, where $F(\tau)$ is a CDF of a log-normal distribution $f(\tau)$. Similarly, a lognormal distribution can be also adopted to describe time to progression. The probability that a progression is observed can be calculated as a product of a patient's TCP (as in log-logistic model Eq. 3) and the conditional probability that the patient experiences progression given the progression will eventually occur.

Model ADNN-DVH

In *ADNN-DVH* as shown in Figure 5-1A, 3 blocks of 1D convolutional layers and average pooling layers are applied to differential DVHs. Reduced representations of DVH are then concatenated with structure volume and mean dose, serving as inputs of two fully-connected layers. Outputs of *ADNN-DVH* are conditional event-free probabilities through different time intervals. Specifically, in a situation where $T = 3$ and the output is denoted as $(P_{T_1}, P_{T_2}, P_{T_3})$ as shown in Figure 5-2, the probabilities that an event happens in time interval T_1, T_2 and T_3 are

$1 - P_{T_1}, P_{T_1}(1 - P_{T_2})$ and $P_{T_1}P_{T_2}(1 - P_{T_3})$, respectively. Generally, in the situation of T intervals, log-likelihood function for an individual with failure in interval j is defined in Eq. 20.

$$l = (1 - P_{T_j}) \prod_{i=1}^{j-1} P_{T_i} \quad \text{Eq. 20}$$

The log-likelihood function for an individual without experiencing events through interval j (either censored or event-free during follow-up) is defined in Eq. 21.

$$l = \prod_{i=1}^{j} P_{T_i} \quad \text{Eq. 21}$$

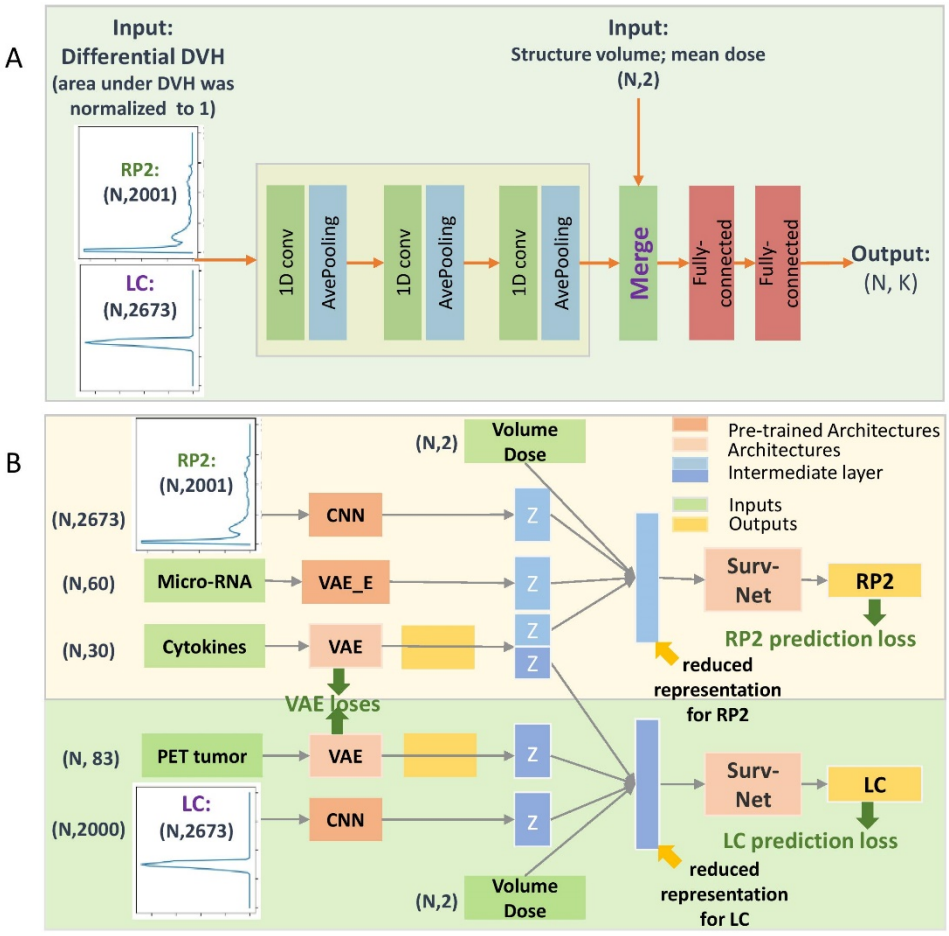

Figure 5-1. Architecture of model ADNN-DVH (A) and ADNN-com (B)

Total loss function is defined as a negative average of log-likelihood function over all the patients. Two $N \times T$ matrices $SurvE$, $SurvS$ in Eq. 22 and Eq. 23 are defined for convenience of calculation, where N denotes the number of patients and T denotes the number of time intervals. Hence, the loss function can be calculated by Eq. 24 and Eq. 25.

$$SurvE_{i,j} = \begin{cases} 1 & \text{if an event occured in time interval } j \text{ for patient } i \\ 0 & \text{otherwise} \end{cases} \quad \text{Eq. 22}$$

$$SurvS_{i,j} = \begin{cases} 1 & \text{if an event have not occured up to time interval } j \text{ for patient } i \\ 0 & \text{if an event occured in or before time interval } j \text{ for patient } i \end{cases} \quad \text{Eq. 23}$$

$$A = \log[1 - SurvS \odot (1 - P)] + \log[1 - SurvE \odot P], \quad \text{a } N \times T \text{ matrix} \quad \text{Eq. 24}$$

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{T} A_{i,j} \quad \text{Eq. 25}$$

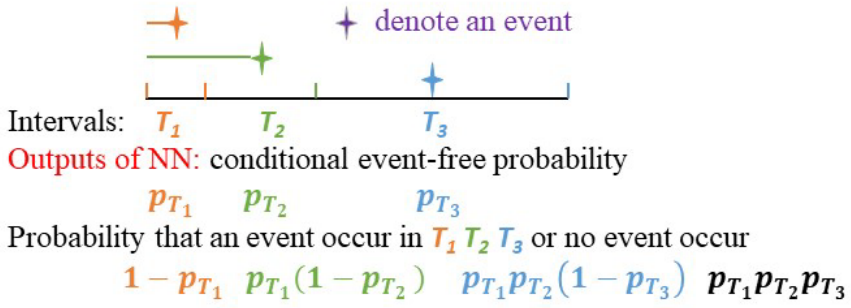

Figure 5-2. Calculating the probability that an event occurred in each time interval from output

Model ADNN-com

In *ADNN-com*, biological and imaging information is considered together with dosimetric information for the prediction of RP2 and LC. Three VAEs were applied to conduct dimensionality reduction for PET radiomics, cytokines and miRNA data, respectively. Specifically, a trained VAE in an architecture *ADNN-miRNA* in Figure 5-3 were used for miRNA data. 80 patients from TCGA-LUAD and TCGA-LUSC who had follow-up information (i.e.,

overall survival and LC) and treated with adjuvant radiotherapy on primary tumor sites were included in source learning task in transfer learning [93], which allows more reliable and generalizable representation learning. The latent representations from these VAEs were then merged with structure volumes, mean doses and reduced representations of DVH from *ADNN-DVH* into concatenated vectors, which were then fed into Surv-Net.

For RP2 prediction, inputs of subsequent Surv-Net were composed of lung mean dose, lung volume, reduced representation of lung DVH, miRNA and cytokines. For LC prediction, inputs of Surv-Net are composed of tumor mean dose, tumor volume, reduced representation of tumor DVH, PET tumor radiomics, miRNA and cytokines. The total loss of architecture *ADNN-com* is the sum of VAE losses and loss of Surv-Net.

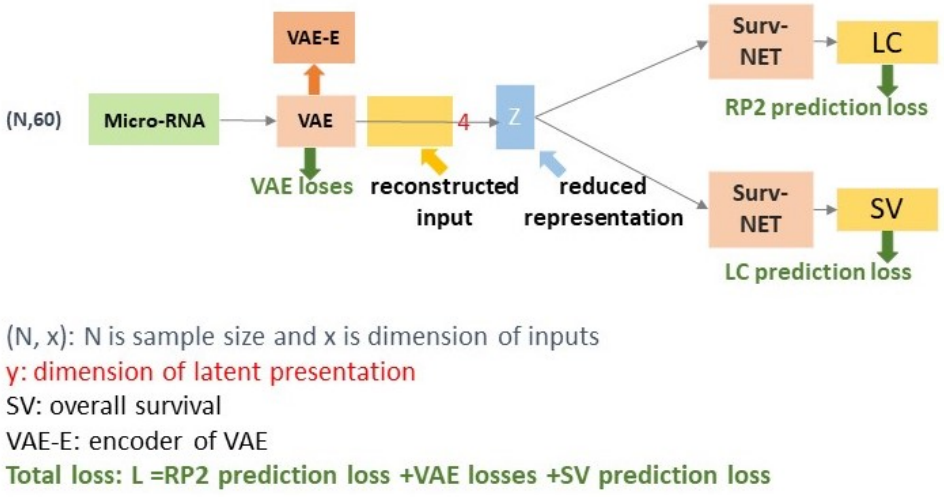

Figure 5-3. Architecture ADNN-miRNA which is applied on TCGA data that realize joint prediction of LC and overall survival (SV)

Model ADNN-com-joint

An *ADNN-com-joint* model as presented in Figure 5-2B realized joint prediction of RP2 and LC. It can be regarded as a combination of architectures ADNN-com for RP2 and LC. However, the VAEs that are applied for dimensionality reduction of cytokines and miRNA are

shared between RP2 and LC. Additionally, Surv-Nets for prediction of RP2 and LC are trained simultaneously.

Table 5-1. Details of optimal parameters in analytical models and parameters of ADNN architectures

		D_{50} Median (95% CI)	*m/k* Median (95% CI)	*n* Median (95% CI)
Analytical Models	Lyman	20.00 (11.90-47.24)	0.55 (0.26-0.890	2.51 (0.541-35.38)
	Log-logistic	51.79 (35.22-62.61)	3.21 (1.61-5.85)	-12.41 (-27.81- -9.93)
		1D-CNN: 3 conv layers (kernel size, num of channels) 3 pooling layers (kernel size)	VAE: Num of nodes in (hidden layers, Latent variables)	Survival net 2 dense layers (num of nodes)
ADNN-DVH	RP2	Conv (12, 2), (6, 4), (3, 1) Pooling 10, 9, 5	N/A	(5, 2)
	LC	Conv (11, 2), (6,4), (3,1) Pooling 9, 9, 8	N/A	(5, 2)
ADNN-com	RP2	Conv (12, 2), (6, 4), (3, 1) Pooling 10, 9, 5	Cytokines (10, 2)	(5, 2)
	LC	Conv (11, 2), (6,4), (3,1) Pooling 9, 9, 8	Cytokines (10, 1) PET (10, 2)	(5, 2)
ADNN-com-joint	RP2	Conv (12, 2), (6, 4), (3, 1) Pooling 10, 9, 5	Cytokines (10, 3=2[RP]+1[LP])	(5, 2),
	LC	Conv (11, 2), (6,4), (3,1) Pooling 9, 9, 8	PET (10, 2)	(5, 2)
ADNN-miRNA	RP2	N/A	miRNA (8, 4)	(5, 2)
	Surv	N/A		(5, 2)

Performance evaluation and validation

Performance evaluation

Harrell's c-index (c-index) [94] can evaluate the goodness of fit for models that produce time-dependent risk scores. Supposing, a pair of patients (i, j) have risk scores (S_i, S_j) and time-to-event (T_i, T_j), with $S_i > S_j$. If $T_i < T_j$, the pair is regarded as a concordant pair, and if $T_i > T_j$,

it is a discordant pair. C-index is then defined as a ratio of concordant pairs to a sum of concordant pairs and discordant pairs. This concept can be easily adopted for Cox models, which produce a single risk score for each patient. However, as our models allow hazard probability's dependence on input data to vary with time, there is no single score (different scores in different intervals) existing for an individual. Alternatively, c-index for binary data (Eq. 26**),** which is based on risk score S and event L attaching to certain follow-up time τ is adopted, and patients with censored time before τ (censored data: $d(\tau) = 0$) were excluded from calculation. Specifically, c-index in Eq. 26 can be regarded as an AUC [95], except it additionally considers time-to-event and censored time. In our study, as events i.e., local progression and RP2 are relatively sparse, the performance was designated to be evaluated at $\tau =$maximal event time in each dataset to cover all events, accordingly, any patient with follow-up less than τ (censored) are excluded from the calculation:

$$c = \frac{\sum_{i \neq j} 1\{S_i(\tau) > S_j(\tau)\}\, 1\{L_i(\tau) = 1,\ L_j = 0\} d_j(\tau)}{\sum_{i \neq j} 1\{L_i(\tau) = 1,\ L_j = 0\}\, d_j(\tau)} \qquad \text{Eq. 26}$$

To further evaluate the performance of *ADNN-com*, the area under a free-response ROC (AU-FROC) curve [96], widely used in the diagnostic classification where cases might contain two or more task-related lesions, was adopted. Similar to ordinary ROC curves, the ordinate in a FROC plot is true positive rate (TPR), however, this TPR is defined over all endpoints. The abscissa of the FROC plot is the average (over all endpoints) false positive rate per case. AU-FROC is able to summarize performance and gives an overall evaluation of both RP2 and LC prediction.

Validation

Stratified 5-fold CV was conducted on the 117 patients in accordance with TRIPOD level 2 type a criterion [85], i.e., data were randomly split into two groups: one for model development,

the other for evaluation of performance. 100 different splits were conducted to consolidate the evaluation. 95% confidence intervals (CIs) were deduced with the quantile function of the norm distribution after the variance of c-index was calculated as defined by DeLong [97]. Additionally, a dataset containing 25 newly treated patients was used in the independent test following TRIPOP level 2 type b criterion, i.e., data were split based on time, which is thought as a stronger design compared to random splits. Moreover, 327 patients from RTOG0617 protocol were considered for external validation (TRIPOD level 3).

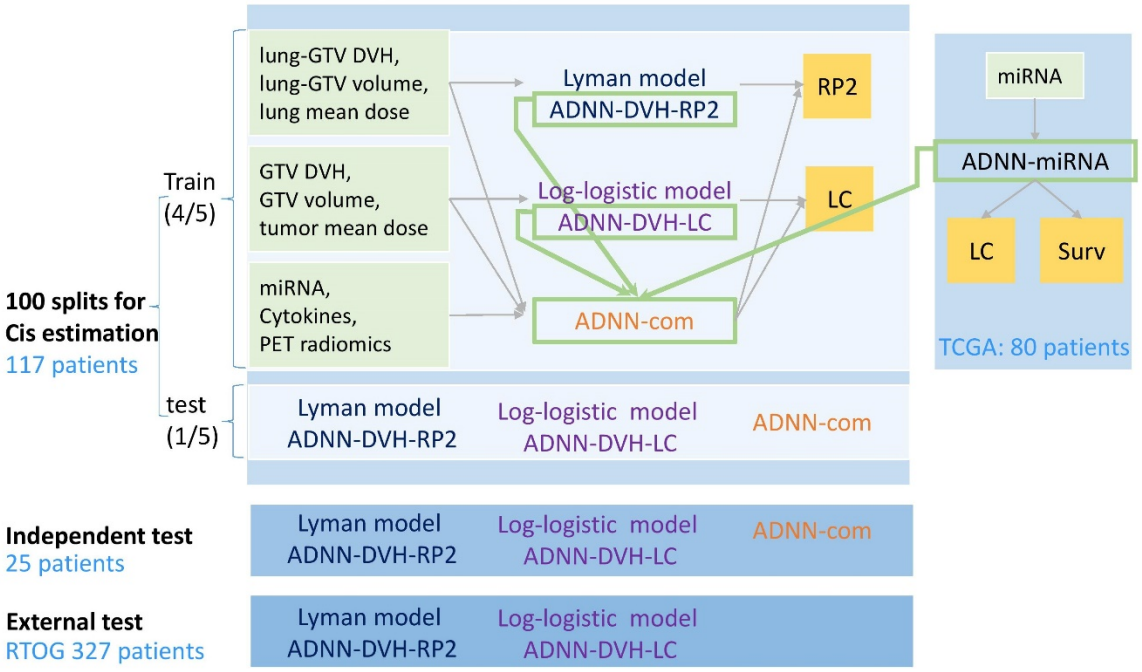

Figure 5-4. Training, validation and test processes of proposed models and analytical models.

Results

Following TRIPOD criteria level 2 type a (random split), 20 times of stratified 5-fold cross-validation in UM 117 patient dataset were conducted. In each split, a range of time intervals was determined in a way to ensure there were the same number of events that happened in each interval in the training data. Lyman/log-logistic models were trained and tested in the

same way as proposed models for comparison purposes. Their optimal values of free parameters are presented in Table 5-1. Cross-validated c-indexes with 95% confidence intervals for both analytical and ADNN models were calculated and summarized in Table 5-2. The corresponding ROC curves of RP2 and LC prediction by the best models ADNN-com-joint are presented in Table 5-2.

Cross-validated results and independent test results

Our models were independently tested on an independent dataset of 25 prospectively treated patients at the University of Michigan following TRIPOD level 2b, and the corresponding results were shown in Table 2. For external validation (TRIPOD level 3), as RP2 was not available in the RTOG0617, RP3 prediction was tested instead of RP2, which provides higher sensitivity to toxicity. 327 patients were used to validate the proposed models with results shown in Table 5-2.

Table 5-2. Cross-validated and independent testing C-index results

Model evaluation on UM 117 patients			
C-index (95% CI)	RP2	LC	RP2&LC
Lyman/log-logistic	0.613 (0.583-0.643)	0.569 (0.545-0.594)	N/A
ADNN-DVH	0.660 (0.630-0.690)	0.727 (0.700-0.753)	N/A
ADNN-com	0.691(0.661-0.722)	0.735(0.710-0.761)	N/A
ADNN-com-joint	**0.705 (0.676-0.734)**	**0.740 (0.715 - 0.765)**	**0.720 (0.671-0.801)**
Independent test on 25 newly-treated patients			
Lyman/log-logistic	0.588	0.573	N/A
ADNN-DVH	0.667	0.706	N/A
ADNN-com	0.683	0.713	N/A
ADNN-com-joint	**0.691**	**0.721**	**0.709**
RTOG 0617			
C-index	RP3	LC	N/A
Lyman/log-logistic	0.736	0.554	N/A
ADNN-DVH	0.762	0.618	N/A

.

Performance of proposed models and analytical models

Generally, *ADNN-DVH* models which were based solely on dosimetric information outperformed traditional Lyman/log-logistic models in RP2/LC prediction. Specifically, *ADNN-DVH* models showed a cross-validated c-index of 0.660 (95% CI: 0.630~0.690) on RP2 prediction and 0.727 (95% CI: 0.700~0.753) on LC prediction. While Lyman model showed a cross-validated C-index 0.613 (95% CI: 0.583~0.643) on RP2 prediction and log-logistic model showed a cross-validated C-index 0.569 (95% CI: 0.545~0.594) on LC prediction. In both independent and external tests, *ADNN-DVH* models yielded better performance than Lyman/log-logistic models. Specifically, *ADNN-DVH* yielded a C-index 0.736 (RP3) on RP prediction, and 0.618 on LC prediction on the external datasets.

Architectures *ADNN-com* which incorporated image and biological information further improved the performance over *ADNN-DVH*. And architectures *ADNN-com-joint* which are based on *ADNN-com* and realized joint prediction showed the best performance, with a cross-validated C-index 0.705 (95% CI: 0.676~0.734) and test C-index 0.691 on RP2 prediction, and a cross-validated C-index 0.740 (95% CI: 0.715 ~0.765) and test C-index 0.721 on LC prediction. It also yielded a cross-validated joint-prediction AU-FROC 0.720 (95% CI: 0.671 ~0.801) and test AU-FROC 0.709.

Visualization of convolutional layer by Grad-cam

Gradient-weighted class activation mapping (Grad-CAM) [98], a class-discriminative localization technique was applied to generate visual explanation for convolutional layers in our model. Specifically, a Grad-CAM as defined in Eq. 27 was calculated for each convolutional layer to understand the importance of each neuron for a decision of interest. In Eq. 27, c denotes an arbitrary output; $A^k \in \mathbb{R}^{u \times v}$ is the k^{th} feature map with height u and width v; α_k^c is the

weight of the k^{th} feature map in discriminating class *c*. The weight α as shown in Eq. 28, is defined as gradients of score for class c, y^c with respect to feature maps A^k of a convolutional layer followed by a global average pooling. The weight α captures the importance of feature map k for a target class c.

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k) \quad \text{Eq. 27}$$

$$\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \quad \text{Eq. 28}$$

Grad-CAM can highlight (assign higher values to) regions in an activation map that are important for a decision of interest. By interpolation, one can re-size Grad-CAM to the size of original inputs. By comparing the interpolated Grad-CAM and original inputs, one would clearly know which regions of original input contribute most to a decision of interest.

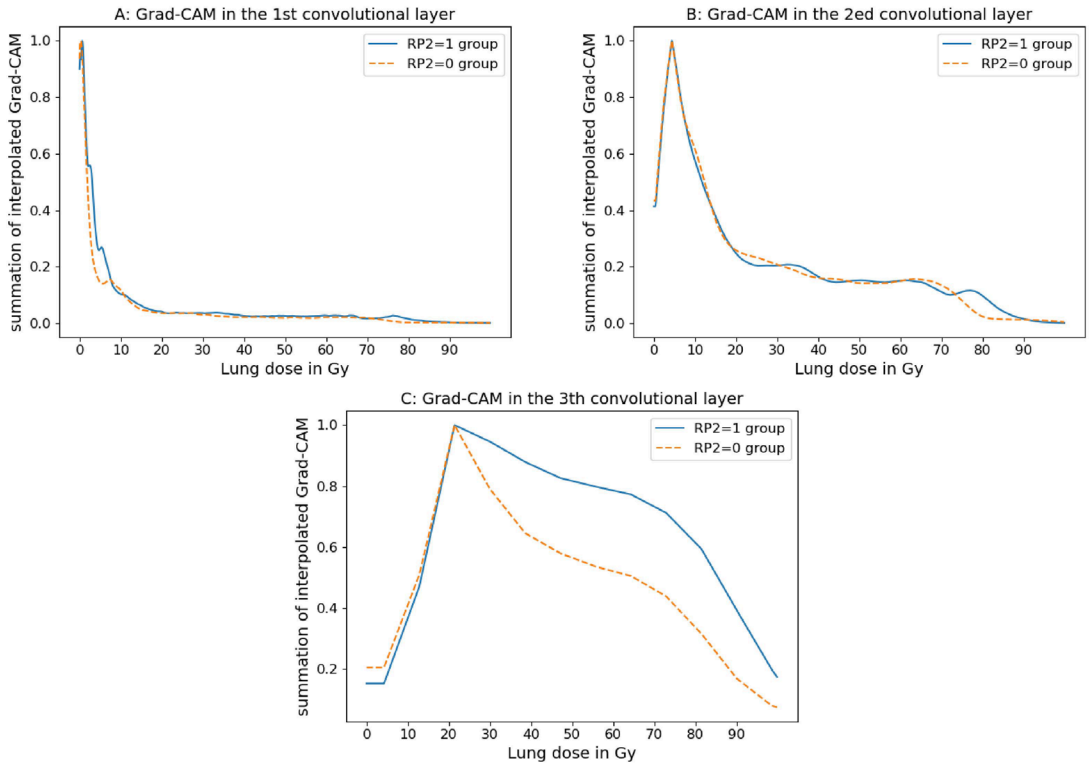

Figure 5-5. Summation of Grad-CAMs for patients in different toxicity groups in convolutional layer 1 (A), 2 (B) and 3 (C)

Figure 5-5 shows the summation of Grad-CAMs for patients in RP2=0 and RP2=1 group in different convolutional layers. In the 1st layer, Grad-CAM assigned relatively higher values to the lower-dose region, which may relate to the distribution in the original inputs. In the 2ed layer, the highlighted region slightly shifts to the right. In the 3th convolutionally layer, Grad-CAM becomes more focused on dose higher than 20 Gy for the determination of toxicity. Specifically, compared to RP2=0 group, Grad-CAM for RP2=1 group has relatively higher values in the the higher dose region (dose>20Gy). In Figure 5-6, regions of differential DVHs that are correponding to highlighted parts (> 90% of maximal values) in Grad-CAMs in two toxicity groups are compared. It shows that the plot of RP2=1 group in the 3th convolutional layer are more intense in the higher dose region compared to that of RP2=0 group. Generally, the visualization by Grad-CAMs indicates CNN graduately become more focused on higher-dose regions for the determination of toxicity from lower-level to higher-level layers.

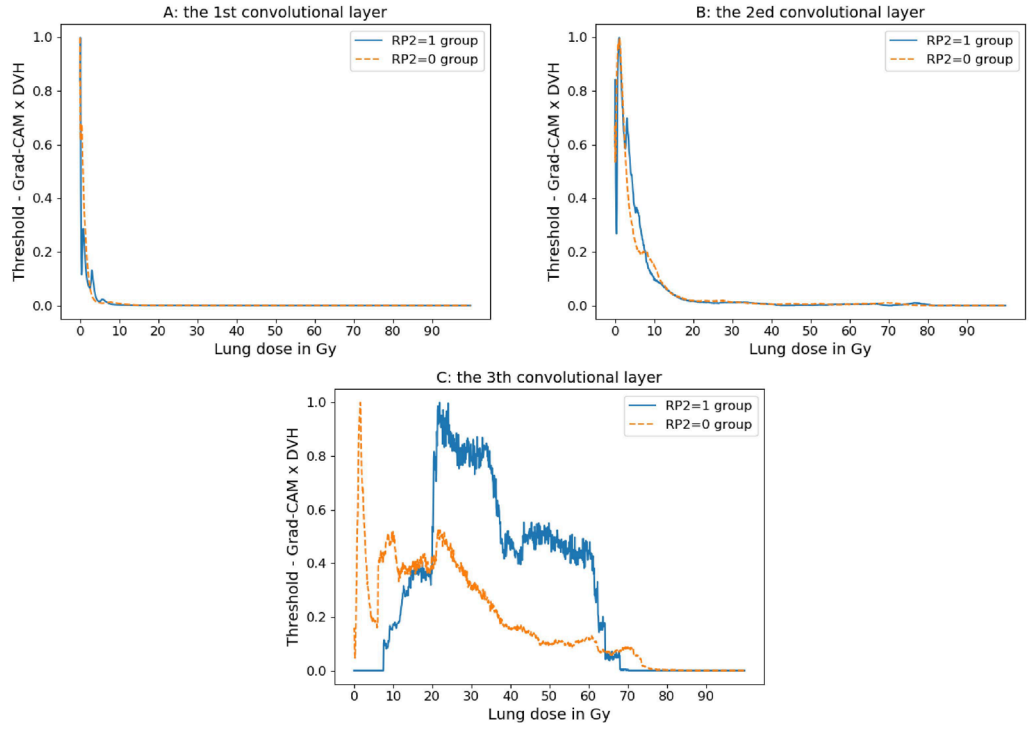

Figure 5-6. Summation of Grad-CAM (high-intensity regions with 90% threshold)-weighted differential DVHs for patients in RP2=0 and RP2=1 groups in convolutional layer 1 (A), 2 (B) and 3 (C)

Conclusion

This study proposes several deep learning architectures that outperform analytical models in the actuarial prediction of LC and RP2. An ADNN-DVH further accounts for information other than the average dose (as in analytical models), i.e., morphological characteristics extracted from DVH in the prediction of LC and RP2. An ADNN-com further integrates complex interactions among biological, imaging and dosimetric information. An ADNN-com-joint realizes multi-endpoints prediction which make outcome models more realistic clinical decision support tools.

.

Chapter 6 Discussion and future perspectives

Discussion

In our study, DL techniques have been incorporated into outcome modeling for LC and RP2 prediction in NSCLC patients. Compared with analytical models and traditional machine learning models, our proposed outcome models, e.g., (1) VAE-MLP joint architectures can combine feature engineering and prediction into a single procedure, (2) VAE-MLP-joint and *ADNN-com* architectures consider complex radiotherapy interaction among physical, biological and imaging variables, (3) 1D CNN-MLP architectures model longitudinal associations among sequential measurements, (4) ADNN-DVH and ADNN-com architectures incorporate time-to-event information for actuarial prediction, (5) Model ADNN-com-joint integrate multi-endpoints (joint) prediction. It has been shown that (1) latent variables from VAE-MLP were able to compensate traditional handcrafted features in RP2 prediction, (2) 1D CNN-MLP joint architectures outperformed plain MLPs that do not consider longitudinal association in LC prediction, and (3) ADNN architectures yielded better results in LC and RP2 prediction compared to classical Lyman and log-logistic models.

Limitation of current work and future perspectives

Other cancer sites and treatment modalities

In our study, proposed methodologies are applied only on NSCLC patients as a proof of concept. However, as temporal and spatial heterogeneity [99] exists in tumor microenvironment universally, outcome modeling that can help personalized treatment and adaptive therapy is expected regardless of cancer sites and treatment modality for improvement of treatment

response. For instance, breast cancer is a common cancer type worldwide. As breast cancer survivors frequently experience a repertoire of symptoms that are detrimental to their quality of life. Outcome modeling may be a key to consider the trade-off between minimizing side effects and maximizing tumor eradication. Furthermore, currently, breast cancer patients are usually classified into different subtypes [100] according to their gene expression profiling, which provides useful information for the decision of treatment choice. This probably indicates outcome models that can account for biological information may contribute to new personalized treatments. In the case of head and neck cancer [101], large anatomic variations including bodyweight loss, tumor shrinkage and parotid gland displacement can be commonly observed during the course of radiotherapy. As a result, outcome modeling is a desirable tool for deciding whether to adapt treatment or guiding the selection of adapted treatment options. To extend our methodologies to other cancer sites, different sets of imaging, biological, dosimetric and clinical information may be considered. The choice of information would depend on the availability of information, clinical routines or any prior knowledge for a specific site of cancer.

Also, radiation therapy is usually combined with other treatment modalities e.g., chemotherapy, surgery and most recently, immunotherapy. Outcome models that can incorporate treatment information of different modalities and study the interactions among radiation in radiotherapy, chemical agents in chemotherapy, and biological substances (e.g., immune checkpoint inhibitors, antibodies, treatment vaccine) in immunotherapy [102] can potentially provide guidance on how to combine different treatment modalities to achieve better treatment than using any one of them alone.

Methodology

As mentioned, with the current advances of high throughput biotechnology, more patient specific information e.g., genomics, proteomics transcriptomics, metabolomics patient specific information becomes available. However, current outcome models are only able to learn interactions among those variables based on available datasets, but are not able to incorporate any prior domain knowledge e.g., metabolic, genetic and signal transduction pathways [103] into these models. It is found that many genetic and epigenetic alterations affecting cell proliferation, death and migration can map to signaling pathways [41]. Changes in the tumor microenvironment, angiogenesis and inflammation are also reflected in signaling networks. Graph neural networks (GNNs) [104] are DL methods in the non-Euclidean domain, which take data in the form of graphs as input. GNNs may be applied directly to networks of biological pathways to learn efficient representation from it and predict treatment response. GNNs may also help incorporate findings from bottom-up approaches [9] of outcome modeling, which utilize first basic principles of physics, chemistry and biology to model cellular damage temporally and spatially in response to treatment.

Furthermore, current outcome models only produce a predictive result of treatment response. They can provide guidance but are not able to generate a recommendation for treatment options or prescriptions directly. Hence, a scheme, e.g., reinforcement learning (RL) [38] that can figure out optimal strategies will potentially offer direct aid to clinical decision making. RL can be applied to adaptive treatment planning, e.g., how to optimize prescriptions for patients by learning from during treatment information. It is designed to achieve a definite goal by optimizing a reward function. The learning process of an RL algorithm is through interaction with an environment so that the RL user (also called an agent) tries to earn the most

reward to obtain the designated goal. One of the obstacles of applying RL into radiotherapy treatment is no accurate reward function based on ground truth exists, however, outcome modeling that can generate robust prediction results of endpoints information as developed in this thesis can be incorporated into the RL framework to provide better guidance when searching for the optimal treatment strategy [38].

Clinical application

To make data-driven outcome models real clinical tools, efforts should be made to improve the interpretability of machine learning algorithms. Interpretability is particularly important as it can help act as a fail-safe against a scenario where algorithms may produce flawed results due to unforeseen bugs. Existing machine learning algorithms, specifically DL algorithms are known to suffer from a tradeoff between accuracy and interpretability [36].

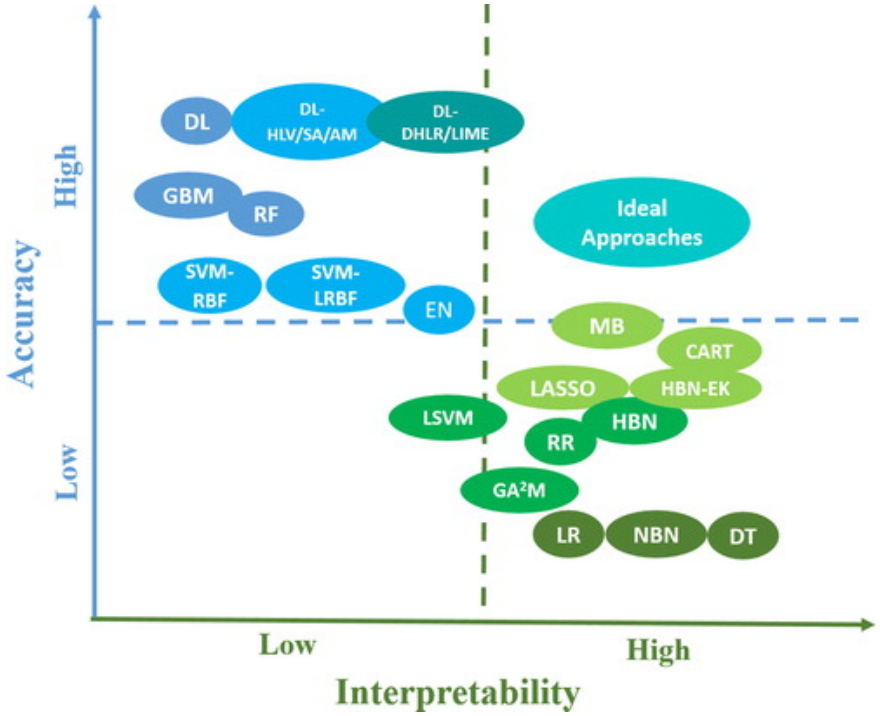

Figure 6-1. The accuracy and interpretability of approaches in radiation outcomes prediction and the location of potential ideal approaches with more balanced accuracy and interpretability for the outcome modeling. Besides the notations introduced in the paper, the rest of abbreviations in the figure can be described as follows, "GAM": generalized additive models; "HBN": hierarchical Bayesian Network; "NBN": naïve Bayesian network; "CART": classification and regression trees; "EN", elastic net; "LR", logistic regression; "MB", MediBoost; "RR", ridge regression; "LSVM", linear support vector machine; "DT", decision

tree; "GBM": gradient boosting machine. Adapted from "Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling" by Yi Luo and et al., 2019, BJR open 1(1), p.20190021. Reprinted with permission.

More work regarding interpreting and explaining machine learning algorisms' decisions [105] is expected. In our study mentioned in Chapter 3 above, MLP-based methods WP, FQI and FSPP were investigated to select important features, those methods can be also regarded as a means to interpret an MLP model. Other methods e.g., sensitivity analysis which is based on a local measure of variation such as local gradient, Taylor decomposition which decompose the learned function as a sum of relevance scores according to Taylor expansion, and relevance propagation which moves in the reverse direction of generating prediction progressively redistributing the prediction score until the input have been previously applied into fields such as ecological model, mutagenicity and imaging recognition [106]. These methods can potentially provide insights into the logic behind predictions made by outcome models by revealing important variables for clinical decision making. Alternatively, techniques that tackle this issue by building a prototype in the input domain, which is interpretable and representatives of learned concepts can be also considered. Activation maximization (AM) searches for an input pattern that produces maximum response in the outputs [107]. When applied to image recognition, those prototypes take the form of synthetic images that would be classified as one class with high probability. In our third study, Grad-CAM techniques have been applied for the visualization of CNN models. This technique can also provide insights into which regions of original input contribute most to a decision of interest. One can also enforce expert knowledge into AM to focus on more probable input space and generate a more realistic prototype. In the outcome models, AM can potentially help generate synthetic data of patients that are more or less likely to experience toxicity or progression.

Additionally, human-in-to-loop (HITL) [108] concepts, which can guide to optimize the entire learning process by introducing human-computer interaction into the system may be used in model development. Machines are recognized for their capabilities of learning from a vast dataset, while humans can make descent decisions even with scare information. Incorporating experts' intelligence into AI systems may improve both accuracy and interpretability for practical decision making in the radiation oncology clinic. Moreover, it would increase physicians' confidence in applying computational tools, hence, make outcome models a more viable clinical tool.

Appendices

Abbreviations

3D-CRT: three-dimensional conformal radiation therapy

ADNN: Actuarial deep neural network

ANN: artificial neural network

AM: Activation maximization

AUC: area under ROC curve

BED: biological effective dose

BS: Brier scores

CI: confidence interval

CNN: convolutional neural network

CNV: copy number variations

CT: computed tomography

CTCAE: thee National Cancer Institute

CTV: clinical tumor volume

CV: Cross-validation

DL: deep learning

DVH: dose volume histogram

EBRT: external beam radiation therapy

FPR: false positive rate

FQI: feature quality index

FSPP: feature-based sensitivity of posterior probability

GLCM: gray level co-occurrence matrix

GLRLM: gray level run length matrix

GLSZM: gray level size zone matrix

GNN: Graph neural networks

GTV: gross tumor volume

Gy: gray (dose unit)

HITL: human-in-to-loop

IMRT: intensity modulated radiation therapy

ITV: internal target volume

KL: Kullback–Leibler

LC: local control

LINAC: medical linear accelerator

LOOCV: leave-one-out cross validation

LQ: model: linear quadratic model

MLP: multi-layer perceptron

MRI: magnetic resonance imaging

MSE: mean squared error

NGTDM: neighborhood gray-tone difference matrix

NN: neural networks

NTCP: normal tissue complication probability

PCA: principal component analysis

PFTs: pulmonary function tests

PTV: planning target volume

RF: random forest

RL: reinforcement learning

RNN: recurrent neural network

ROC: receiver operative characteristic

RP: radiation pneumonitis

RTOG: Radiation Therapy Oncology Group

SCLC: small cell lung cancer

SMOTE: synthetic minority over-sampling techniques

SNP: single-nucleotide polymorphism

SVM: support vector machine

TCGA: The Cancer Genome Atlas

TCP: tumor control probability

TPR: true positive rate

TRIPOD: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

UAT: Universal Approximation Theorem

VAE: variational auto-encoder

VMART: volumetric-modulated arc therapy

WP: weight pruning

miRNA: micro RNA

Bibliography

1. Weinstein, M.C., et al., *Modeling for health care and other policy decisions: uses, roles, and validity.* Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research, 2001. **4**(5): p. 348-361.
2. Moiseenko V, K.T., Van Dyk J. *Biologically-based treatment plan optimization: a systematic comparison of NTCP models for tomotherapy treatment plans*. in *14th international conference on the use of computers in radiation therapy*. 2004. Seoul.
3. Hope, A.J., et al. *Clinical, Dosimetric, and Location-Related Factors to Predict Local Control in Non-Small Cell Lung Cancer*. in *ASTRO 47TH ANNUAL MEETING*. 2005. Denver, CO.
4. Halperin EC, P.C., Brady LW, *Perez and Brady's principles and practice of radiation oncology. 5th ed*. 2008, Philadelphia, PA, US: Wolters Kluwer Health/Lippincott Williams&Wilkins.
5. Bradley, J.D., et al., *Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study.* The Lancet. Oncology, 2015. **16**(2): p. 187-199.
6. Bentzen, S.M., et al., *Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues.* International journal of radiation oncology, biology, physics, 2010. **76**(3 Suppl): p. S3-S9.
7. Jackson, A., et al., *The lessons of QUANTEC: recommendations for reporting and gathering data on dose-volume dependencies of treatment outcome.* International journal of radiation oncology, biology, physics, 2010. **76**(3 Suppl): p. S155-S160.
8. Bradley, J.D., et al., *Long-Term Results of RTOG 0617: A Randomized Phase 3 Comparison of Standard Dose Versus High Dose Conformal Chemoradiation Therapy +/- Cetuximab for Stage III NSCLC.* International Journal of Radiation Oncology • Biology • Physics, 2017. **99**(2): p. S105.
9. El Naqa, I., *A Guide to Outcome Modeling In Radiotherapy and Oncology: Listening to the Data*. 2018, Boca Raton, FL: CRC Press. Taylor & Francis Group
10. Allen Li, X., et al., *The use and QA of biologically related models for treatment planning: short report of the TG-166 of the therapy physics committee of the AAPM.* Medical physics, 2012. **39**(3): p. 1386-1409.
11. Brahme, A., *Optimized radiation therapy based on radiobiological objectives.* Seminars in radiation oncology, 1999. **9**(1): p. 35-47.
12. Choi, N., et al., *Predictive factors in radiotherapy for non-small cell lung cancer: present status.* Lung cancer (Amsterdam, Netherlands), 2001. **31**(1): p. 43-56.
13. Fu, X.L., et al., *Study of prognostic predictors for non-small cell lung cancer.* Lung cancer (Amsterdam, Netherlands), 1999. **23**(2): p. 143-152.
14. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis.* Eur J Cancer, 2012. **48**(4): p. 441-6.

15. El Naqa, I., et al., *Radiogenomics and radiotherapy response modeling.* Physics in medicine and biology, 2017. **62**(16): p. R179-R206.
16. Kirk, S., et al., *Radiology Data from The Cancer Genome Atlas Lung Squamous Cell Carcinoma [TCGA-LUSC] collection. The Cancer Imaging Archive.* 2016.
17. Kumar, V., et al., *Radiomics: the process and the challenges.* Magn Reson Imaging, 2012. **30**(9): p. 1234-48.
18. Bussink, J., et al., *PET-CT for radiotherapy treatment planning and response monitoring in solid tumors.* Nat Rev Clin Oncol, 2011. **8**(4): p. 233-242.
19. Zaidi, H. and I. El Naqa, *PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques.* Eur J Nucl Med Mol Imaging, 2010. **37**(11): p. 2165-87.
20. The Cancer Genome Atlas, N., *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**(7536): p. 576-582.
21. Abeshouse, A., et al., *The Molecular Taxonomy of Primary Prostate Cancer.* Cell. **163**(4): p. 1011-1025.
22. The Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-550.
23. Brennan, Cameron W., et al., *The Somatic Genomic Landscape of Glioblastoma.* Cell. **155**(2): p. 462-477.
24. El Naqa, I., et al., *Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century.* The British Journal of Radiology, 2020. **93**(1106): p. 20190855.
25. He, K., et al. *Deep Residual Learning for Image Recognition*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
26. Graves, A., A.r. Mohamed, and G. Hinton. *Speech recognition with deep recurrent neural networks*. in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013.
27. Hornik, K., M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators.* Neural Network, 1989. **2**: p. 359-366.
28. Vidal, R., et al., *Mathematics of deep learning.* CoRR 2017. **abs/1712.04741**
29. Brahma, P.P., D. Wu, and Y. She, *Why deep learning works: A manifold disentanglement perspective.* IEEE Transactions on Neural Networks and Learning Systems, 2006. **27**: p. 1997-2008.
30. Reshef, D.N., et al., *Detecting Novel Associations in Large Data Sets.* Science, 2011. **334**(6062): p. 1518-1524.
31. Lal, T.N., et al., *Embedded Methods*, in *Feature Extraction: Foundations and Applications*, I. Guyon, et al., Editors. 2006, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 137-165.
32. Cui, S., et al., *Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage.* Medical Physics, 2019. **46**(5): p. 2497-2511.
33. Alex Krizhevsky and Sutskever, I.a.H., Geoffrey E, *ImageNet Classification with Deep Convolutional Neural Networks*, in *Advances in Neural Information Processing Systems 25*, F.P.a.C.J.C.B.a.L.B.a.K.Q. Weinberger, Editor. 2012, Curran Associates, Inc. p. 1097-1105.

34. Cui, S., et al., *Artificial Neural Network With Composite Architectures for Prediction of Local Control in Radiotherapy.* IEEE Transactions on Radiation and Plasma Medical Sciences, 2019. **3**(2): p. 242-249.
35. Luo, Y., 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.* Medical physics, 2018: p. 10.1002/mp.13029.
36. Luo, Y., et al., *Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling.* BJR|Open, 2019. **1**(1): p. 20190021.
37. Tseng, H.H., et al., *Machine Learning and Imaging Informatics in Oncology.* Oncology, 2018: p. 1-19.
38. Tseng, H.H., et al., *Deep reinforcement learning for automated radiation adaptation in lung cancer.* Med Phys, 2017. **44**(12): p. 6690-6705.
39. Cui, S., R.K.T. Haken, and I.E. Naqa, *Building a Predictive Model of Toxicity: Methods*, in *Modelling Radiotherapy Side Effects: Practical Applications for Planning Optimisation*, T. Rancati and C. Fiorino, Editors. 2019, CRC Press, Taylor&Francis Group Boca Raton, FL.
40. Hall, E.J. and A.J. Giaccia, *Radiobiology for the Radiologist*. 2006: Lippincott Williams & Wilkins.
41. Hall, E.J. and A.J. Giaccia, *Radiobiology for the Radiologist*. 2011, Philadelphia, UNITED STATES: Wolters Kluwer Health.
42. Halperin, E.C., C.A. Perez, and L.W. Brady, *Perez and Brady's principles and practice of radiation oncology*. 5th ed. 2008, Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins. xxxii, 2106 p.
43. Jackson, A., et al., *The lessons of QUANTEC: recommendations for reporting and gathering data on dose-volume dependencies of treatment outcome.* Int J Radiat Oncol Biol Phys, 2010. **76**(3 Suppl): p. S155-60.
44. Arimura, H., et al., *Computer-Assisted Target Volume Determination*, in *Image-Based Computer-Assisted Radiation Therapy*, H. Arimura, Editor. 2017, Springer Singapore: Singapore. p. 87-109.
45. Zaider, M. and L. Hanin, *Tumor control probability in radiation treatment.* Med Phys, 2011. **38**(2): p. 574-83.
46. Steel, G.G., *Basic clinical radiobiology*. 3rd ed. 2002, London, New York: Arnold, Oxford University Press. viii, 262 p.
47. Luo, Y., 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.* Medical Physics, 2018. **45**(8): p. 3980-3995.
48. Joiner, M.C., B. Marples, and H. Johns. *The Response of Tissues to Very Low Doses per Fraction: A Reflection of Induced Repair?* in *Acute and Long-Term Side-Effects of Radiotherapy*. 1993. Berlin, Heidelberg: Springer Berlin Heidelberg.
49. Willers, H. and K.D. Held, *Introduction to clinical radiation biology.* Hematol Oncol Clin North Am, 2006. **20**(1): p. 1-24.
50. Allen Li, X., et al., *The use and QA of biologically related models for treatment planning: short report of the TG-166 of the therapy physics committee of the AAPM.* Med Phys, 2012. **39**(3): p. 1386-409.
51. Lyman, J.T., *Complication probability as assessed from dose-volume histograms.* Radiat Res Suppl, 1985. **8**: p. S13-9.

52. Blanco, A.I., et al., *Dose-volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy.* Int J Radiat Oncol Biol Phys, 2005. **62**(4): p. 1055-69.
53. Bradley, J., et al., *Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma.* Int J Radiat Oncol Biol Phys, 2004. **58**(4): p. 1106-13.
54. Marks, L.B., *Dosimetric predictors of radiation-induced lung injury.* International Journal of Radiation Oncology Biology Physics, 2002. **54**(2): p. 313-316.
55. Steinwart, I. and A. Christmann, *Support Vector Machines*. 2008: Springer Publishing Company, Incorporated. 602.
56. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
57. Rahman, M.G. and M.Z. Islam, *Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques.* Knowledge-Based Systems, 2013. **53**: p. 51-65.
58. Nembrini, S., I.R. König, and M.N. Wright, *The revival of the Gini importance?* Bioinformatics, 2018. **34**(21): p. 3711-3718.
59. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.
60. Ding, B., H. Qian, and J. Zhou. *Activation functions and their characteristics in deep neural networks*. in *2018 Chinese Control And Decision Conference (CCDC)*. 2018.
61. Glorot, X., A. Bordes, and Y. Bengio, *Deep Sparse Rectifier Neural Networks*, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, G. Geoffrey, D. David, and D. Miroslav, Editors. 2011, PMLR: Proceedings of Machine Learning Research. p. 315--323.
62. Huang, G., et al. *Densely Connected Convolutional Networks*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
63. Junyoung, C.C., Gulcehre; KyungHyun, Cho; Yoshua, Bengio, *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.* CoRR, 2014. **abs/1412.3555**.
64. Kingma, D.P. and M. Welling, *Auto-Encoding Variational Bayes*. 2013.
65. Goodfellow, I.J., et al., *Generative Adversarial Networks.*
66. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* Nature. **542**(7639): p. 115-118.
67. Dalmis, M.U., et al., *Using deep learning to segment breast and fibroglandular tissue in MRI volumes.* Medical Physics, 2017. **44**(2): p. 533-546.
68. Zachary Chase, L., *A Critical Review of Recurrent Neural Networks for Sequence Learning.* CoRR, 2015. **abs/1506.00019**.
69. Wolterink, J.M., et al. *Deep MR to CT synthesis using unpaired data*. 2017. Springer Verlag.
70. Tibshirani, R. and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. 2009: Springer.
71. Baker, S., et al., *A critical review of recent developments in radiotherapy for non-small cell lung cancer.* Radiation Oncology, 2016. **11**(1): p. 115.
72. Vignal, A., et al., *A review on SNP and other types of molecular markers and their use in animal genetics.* Genet Sel Evol, 2002. **34**(3): p. 275-305.
73. Deng, N., et al., *Single nucleotide polymorphisms and cancer susceptibility.* Oncotarget, 2017. **8**(66): p. 110635-110649.

74. O'Brien, J., et al., *Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation.* Frontiers in Endocrinology, 2018. **9**(402).
75. Kraemer, A., et al., *MicroRNA-Mediated Processes are Essential for the Cellular Radiation Response.* Radiation Research, 2011. **176**(5): p. 575-586.
76. Dinarello, C.A., *Historical insights into cytokines.* European journal of immunology, 2007. **37 Suppl 1**(Suppl 1): p. S34-S45.
77. Di Maggio, F.M., et al., *Portrait of inflammatory response to ionizing radiation treatment.* Journal of inflammation (London, England), 2015. **12**: p. 14-14.
78. Yacoub, M. and Y. Bennani, *HVS: A Heuristic for variable selection in multilayer artificial neural network classifier* in *Intelligent Engineering Systems Through Artificial Neural Networks*, D. CH, Editor. 1997, ASME press. p. 1323-1335.
79. Verikas, A. and M. Bacauskiene, *Feature selection with neural networks* Pattern Recogn. Lett 2002. **23**(11): p. 1323-1335.
80. Yang, J., et al., *Feature selection for MLP neural network: the use of random permutation of probabilistic outputs* IEE transaction on Neural Network 2009. **20** p. 1911-1922.
81. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research 2011. **12**: p. 2825-2830.
82. Amari, S.-i., R. Karakida, and M. Oizumi, *Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem.* Information Geometry, 2018. **1**(1): p. 13-37.
83. Conitzer, V., A. Davenport, and J. Kalagnanam, *Improved bounds for computing kemeny rankings.* Proceedings of the 21st national conference on Artificial intelligence, 2006: p. 620-626.
84. Hajian-Tilaki, K., *Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.* Caspian journal of internal medicine, 2013. **4**(2): p. 627-635.
85. Collins, G.S., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement.* Annals of Internal Medicine, 2015. **162**(1): p. 55-63.
86. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* J. Mach. Learn. Res., 2014. **15**(1): p. 1929-1958.
87. Chollet, F. *Keras*. 2015.
88. Kingma, D.P. and J. Ba, *Adam: A Method for Stochastic Optimization.* CoRR, 2014. **abs/1412.6980**.
89. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* J. Artif. Int. Res., 2002. **16**(1): p. 321-357.
90. Steyerberg, E., *Evaluation of Perfomrance*, in *Clinical Prediction Models A Practical Approach to Development, Validation, and Updating*, E. Steyerberg, Editor. 2009, Springer-Verlag New York.
91. Thomas A. Gerds, *R:Fast and user friendly implementation of nonparametric estimators for censored event history (survival) analysis. Kaplan-Meier and Aalen-Johansen method.* 2018.
92. Tucker, S.L., et al., *Analysis of radiation pneumonitis risk using a generalized Lyman model.* International journal of radiation oncology, biology, physics, 2008. **72**(2): p. 568-574.

93. Pan, S.J. and Q. Yang, *A Survey on Transfer Learning.* IEEE Transactions on Knowledge and Data Engineering, 2010. **22**(10): p. 1345-1359.
94. Uno, H., et al., *On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.* Statistics in medicine, 2011. **30**(10): p. 1105-1117.
95. Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 1982. **143**(1): p. 29-36.
96. *5. Extensions to Conventional ROC Methodology: LROC, FROC, and AFROC.* J icru, 2008. **8**(1): p. 31-5.
97. DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.* Biometrics, 1988. **44**(3): p. 837-845.
98. Selvaraju, R.R., et al. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization*. 2017. IEEE.
99. McQuerry, J.A., et al., *Mechanisms and clinical implications of tumor heterogeneity and convergence on recurrent phenotypes.* Journal of molecular medicine (Berlin, Germany), 2017. **95**(11): p. 1167-1178.
100. Bettaieb, A., et al., *Precision medicine in breast cancer: reality or utopia?* Journal of Translational Medicine, 2017. **15**(1): p. 139.
101. Castadot, P., et al., *Adaptive radiotherapy of head and neck cancer.* Semin Radiat Oncol, 2010. **20**(2): p. 84-93.
102. Wang, Y., et al., *Combining Immunotherapy and Radiotherapy for Cancer Treatment: Current Challenges and Future Directions.* Frontiers in pharmacology, 2018. **9**: p. 185-185.
103. Sever, R. and J.S. Brugge, *Signal transduction in cancer.* Cold Spring Harbor perspectives in medicine, 2015. **5**(4): p. a006098.
104. Wu, Z., et al., *A Comprehensive Survey on Graph Neural Networks*. 2019.
105. Montavon, G., W. Samek, and K.-R. Müller, *Methods for interpreting and understanding deep neural networks.* Digital Signal Processing, 2018. **73**: p. 1-15.
106. Montavon, G., W. Samek, and K.-R. Müller, *Methods for interpreting and understanding deep neural networks.* Digital Signal Processing, 2018. **73**(2): p. 1-15.
107. Simonyan, K., A. Vedaldi, and A. Zisserman, *T Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.* J CoRR, 2013. **abs/1312.6034**.
108. Zanzotto, F.M., *Viewpoint: Human-in-the-loop Artificial Intelligence.* Journal of Artificial Intelligence Research, 2019. **64**: p. 243-252.