Machine and Deep Learning in Oncology, Medical Physics and Radiology

Issam El Naqa
Martin J. Murphy
Editors

Second Edition

Springer

Machine and Deep Learning in Oncology, Medical Physics and Radiology

Issam El Naqa • Martin J. Murphy
Editors

Machine and Deep Learning in Oncology, Medical Physics and Radiology

Second Edition

Springer

Editors
Issam El Naqa
Department of Machine Learning
Moffitt Cancer Center
Tampa, FL, USA

Martin J. Murphy
Department of Radiation Oncology
Virginia Commonwealth University
Richmond, VA, USA

ISBN 978-3-030-83046-5 ISBN 978-3-030-83047-2 (eBook)
https://doi.org/10.1007/978-3-030-83047-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword to the First Edition

This book is based on one of the most consequential emergent results of the ongoing computer revolution, namely, that computers can be trained—under the right conditions—to reliably classify new data, such as patient data. This capability, called machine learning (or statistical learning), has been deployed in many areas of technology, commerce, and medicine. Data mining and statistical prediction models have already crept into many areas of modern life, including advertising, banking, sports, weather prediction, politics, science generally, and medicine in particular. The ability of computers to increasingly communicate with people in a natural way (understanding language and speaking to us), such as the famed IBM "Watson" appearance on Jeopardy, or "Siri" on iPhones, portends an accelerating role of sophisticated computer models that predict and respond to our requests. Fundamentally, these developments rely on the ability of statistical computer methods to pull (as Nate Silver puts it) "signal from the noise." While traditional statistical methods typically attempt to ascertain the role of particular variables in determining an outcome of interest (hence, needing many data points for every variable included in the prediction model), machine learning represents a different goal, to reliably predict an outcome, for example that an imaging abnormality is benign with a high degree of certainty. The statisticians and computer scientists working in this emerging area are often happy to use large numbers of variables (or previous data instances) that essentially vote together in a nonlinear fashion. Simplicity is happily traded for an improved ability to predict.

The chapters in this book comprehensively review machine learning and related modeling methods previously used in many areas of radiation oncology and diagnostic radiology. The editors and authors are explorers in this new territory, and have performed a great service by surveying and mapping the many achievements to date and outline many areas of potential application. Early chapters review the fundamental characteristics, and varieties, of machine learning methods, including difficult issues regarding evaluation of predictive model performance. The most well-developed use of machine learning reviewed is the creation of computer-aided diagnosis (CAD) models to provide a reliable "second opinion" for radiologists reading mammograms to detect breast cancer. The increasing use of wider range of imaging features referred to as "radiomics," in analogy to "genomics," presented in radiomics for disease detection, and radiomics for diagnosis, or "theragnostic" [1] chapters, which are devoted to details of image-based informatics formats and

database systems, including tools to share and learn from institutional databases. Machine learning approaches to aid in the planning, delivery, and quality assurance of radiation therapy are reviewed. Efforts to predict response to radiation therapy are also reviewed in useful detail. Obtaining enough data of sufficient quality and diversity is the biggest challenge in predictive modeling. This is only possible if data are shared across institutional and national borders, both academic and community health-care systems [2].

Machine learning—coupled with computer vision and imaging processing techniques—has been demonstrated to be useful in diagnosis, treatment planning, and outcome prediction in radiation oncology and radiology. This is of particular importance since we know that doctors have increasing difficulties to predict the outcome of modernized complex patient treatments [3]. This book provides a wonderful summary of past achievements, current challenges, and emerging approaches in this important area of medicine. Unlike many other approaches to improving medicine, the use of improved and continuously updated prediction models put together in "Decision Support Systems" holds the potential of improved clinical decision making with minimal costs to patients [4]. An intuitively attractive characteristic of this approach is the user of *all* the data available (rather than using only one type of data such as dose or gene profile). We anticipate that predictive models-based Decisions Support Systems will ease the implementation of personalized (or precision) medicine.

Despite investment in efforts to improve the skills of clinicians, patients continue to report low levels of involvement [5]. There is indeed evidence level 1 from a Cochrane systematic review evaluating 86 studies involving 20,209 participants included in published randomized controlled trials demonstrating that decision aids increase people's involvement, support informed values-based choices in patient-practitioner communication, and improve knowledge and realistic perception of outcomes. We therefore believe the next step will be to integrate, whenever possible, Shared Decision Making approaches (see, e.g., www.treatmentchoice.info; www.optiongrid.org) to include the patient perspective on the best treatment of choice [6].

We are sincerely convinced that this book will continue to advance precision medicine in oncology.

Philippe Lambin
Department of Radiation Oncology
Research Institute GROW, Maastro Clinic,
Maastricht University,
Maastricht, The Netherlands

Joseph O. Deasy
Department of Medical Physics
Memorial Sloan Kettering Cancer Center,
New York, NY, USA

References

1. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–6.
2. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, Carvalho S, Leijenaar RT, Nalbantov G, Oberije C, Scott Marshall M, Hoebers F, Troost EG, van Stiphout RG, van Elmpt W, van der Weijden T, Boersma L, Valentini V, Dekker A. Rapid Learning health care in oncology' – an approach towards decision support systems enabling customised radiotherapy. Radiother Oncol. 2013;109(1):159–64.
3. Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, van Baardwijk A, Wanders R, De Ruysscher D, Steyerberg E, Dingemans AM, Lambin P. A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. Radiother Oncol. 2014;112(1):37–43.
4. Lambin P, van Stiphout RG, Starmans MH, et al. Predicting outcomes in radiation oncology–multifactorial decision support systems. Nat Rev Clin Oncol. 2013;10:27–40.
5. Stacey D, Bennett CL, Barry MJ, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev. 2011;10:CD001431.
6. Stiggelbout AM, Van der Weijden T, De Wit MP, et al. Shared decision making: really putting patients at the centre of healthcare. BMJ. 2012;344:e256.

Preface to the First Edition

Radiotherapy is a major treatment modality for cancer and is currently the main option for treating local disease at advanced stages. More than half of all cancer patients receive irradiation as part of their treatment, with curative or palliative intent to eradicate cancer or reduce pain, respectively, while sparing uninvolved normal tissue from detrimental side effects. Despite significant technological advances in treatment planning and delivery using image-guided techniques, the complex nature of radiotherapy processes and the massive amount of structured and unstructured heterogeneous data generated during radiotherapy from early patient consultation to patient simulation, to treatment planning and delivery, to monitoring response, to follow-up visits, invite the application of more advanced computational methods that can mimic human cognition and intelligent decision making to ensure safe and effective treatment. In addition, these computational methods need to compensate for human limitations in handling a large amount of flowing information in an efficient manner, in which simple errors can make the difference between life and death.

Machine learning is a technology that aims to develop computer algorithms that are able to emulate human intelligence by incorporating ideas from neuroscience, probability and statistics, computer science, information theory, psychology, control theory, and philosophy with successful applications in computer vision, robotics, entertainment, ecology, biology, and medicine. The essence of this technology is to *humanize computers* by learning from the surrounding environment and previous experiences, with or without a teacher. The development and application of machine learning has undergone a significant surge in recent years due to the exponential growth and availability of "big data" with machine learning techniques occupying the driver's seat to steer the understanding of such data in many fields, including radiation oncology.

The growing interest in applying machine learning algorithms to radiotherapy has been highlighted by special sessions at the annual meeting of the American Association of Physicists in Medicine (AAPM) and at the International Conference on Machine Learning and Applications (ICMLA). Ensuing discussions of compiling these disparate applications of machine learning in radiotherapy into a single succinct monograph led to the idea of this book. The goal is to provide interested readers with a comprehensive and accessible text on the subject to fill in an important existing void in radiotherapy and machine learning literature. Even as these

discussions were taking place, the subject of machine learning in radiotherapy continued its growth from a peripheral subfield in radiotherapy into widespread applications that touch almost every area in radiotherapy from treatment planning, quality assurance, image guidance, and respiratory motion management to treatment response modeling and outcomes prediction. This rapid growth has driven the compilation of this textbook.

The textbook is intended to be an introductory learning guide for students and residents in medical physics and radiation oncology who are interested in exploring this new field of machine learning for their own curiosity or their research projects. In addition, the book is intended to be a useful and informative resource for more experienced practitioners, researchers, and members of both radiotherapy and applied machine learning as a two-way bridge between these communities. This is manifested by the fact that the book has been written by experts from both the radiotherapy and machine learning domains.

The book is structured into five sections:

- The first section provides an introduction to machine learning and is a must-read for individuals who are new to the field. It begins with a machine learning definition (Chap. 1), followed by a discussion of the main computational learning principles using PAC or VC theories (Chap. 2), presentation of the most commonly used supervised and unsupervised learning algorithms with demonstrative applications drawn from the radiotherapy field (Chap. 3), and descriptions of different methods and techniques used for evaluating the performance of learning methods (Chap. 4). The ever-growing role of informatics infrastructure in radiotherapy and its application to machine learning are presented in Chap. 5. Finally, given the realistic challenges related to data sharing from a global radiotherapy network, this section concludes with a discussion of how machine learning could be extended to a distributed multicenter rapid learning framework.
- The second section summarizes years of successful application of machine learning in radiological sciences—a sister field to radiotherapy—as a computational tool for computer-aided detection (Chap. 7) and computer-aided diagnosis (Chap. 8).
- The third section presents applications of machine learning in radiotherapy treatment planning as a tool for image-guided radiotherapy (Chap. 9) and a computational vehicle for knowledge-based planning (Chap. 10).
- The fourth section demonstrates the application of machine learning to respiratory motion management—a rather challenging problem for accurate delivery of irradiation to a moving target—by discussing predictive respiratory models (Chap. 11) and image-based compensation techniques (Chap. 12).
- Quality assurance is at the heart of safe delivery of radiotherapy and is a major part of a medical physicist's job. Examples for application of machine learning to QA for detection and prediction of radiotherapy errors (Chap. 13), for treatment planning (Chap. 14), and for delivery (Chap. 15) validation are presented and discussed.
- In the era of personalized evidence-based medicine, machine learning predictive analytics can play an important role in the understanding of radiotherapy

response (Chap. 16). Examples of successful machine learning applications to normal tissue complication probability (Chap. 17) and tumor control probability (Chap. 18) highlight the inherent power of this technology in deciphering complex radiobiological response.

This book is the product of a coordinated effort by the editors, authors, and publishing team to present the principles and applications of machine learning to a new generation of practitioners in radiation therapy and to present the present-day challenges of radiotherapy to the computer science community, with the hope of driving advancements in both fields.

Montreal, QC, Canada — Issam El Naqa
Stanford, CA, USA — Ruijiang Li
Richmond, VA, USA — Martin J. Murphy

Preface to the Second Edition

This is a revised and expanded edition of the original machine learning book in radiation oncology. The current expanded edition will provide a comprehensive overview of machine and deep learning and their role not only in radiation oncology and medical physics but also in the inter-related fields of general oncology and radiology. The book covers machine and deep learning from basic theory, methods, into demonstrative applications in these areas. The book is further enriched with examples and illustrations for the interested reader. The goal remains to provide the reader with a comprehensive and accessible text on the subject to fill an important existing void in the oncology, radiology, and machine learning literature.

Since the publication of the first edition in 2015, the fields of machine and deep learning have witnessed tremendous growth in medicine in general and radiological sciences in particular. This followed the watershed moment of the emergence of deep learning and its successful application across the board. The first edition described deep learning briefly; this new edition has a dedicated chapter on this topic, given its importance. In addition, the introductory chapters on machine learning have been revised and/or expanded with new chapters on deep learning, quantum computing, and software tools. The application chapters have also been revised and restructured with dedicated sections on medical image analysis, radiotherapy treatment planning and delivery, and outcomes modeling and decision support.

The targeted audience for this revised textbook is students and residents in oncology and radiology who are seeking an introductory guide and to build a solid foundation in machine and deep learning and their application. As in the first edition, the book is also intended to act as an informative resource for more experienced practitioners, researchers, and members of both radiological sciences and applied machine learning, acting as a bridge between these communities. This is manifested by the fact that the book has been written by experts from both the radiological sciences and machine learning domains.

The revised edition of the book is structured into four parts:

- The first part provides an introduction to machine and deep learning and is a must-read for individuals who are new to the field. It begins with machine and deep learning definitions (Chap. 1), followed by a discussion of the main computational learning principles (Chap. 2). Commonly used machine learning algorithms are described in the following two chapters; one focuses on traditional supervised,

unsupervised, and reinforcement learning algorithms with demonstrative applications drawn from the oncology/radiology fields (Chap. 3) and the other one is dedicated to the growing deep learning field where data representation and the learning tasks are embedded in the same framework (Chap. 4). This chapter provides a pedagogical transition from basic multi-layer neural networks to more advanced convolution, recurrent, adversarial architectures and their variants, with programming examples as well. A new chapter on the emerging subject of quantum computing in machine learning is presented in Chap. 5, which should appeal not only to information theorists but also to medical physicists, who can now see the underlying principles of their discipline through the new lens of data analytics and how it is reshaping both fields. Chapter 6 presents descriptions of different methods and techniques used for evaluating the performance of learning methods. This is followed by descriptions of software platforms, libraries, and tools for machine and deep learning that provide the reader with the many options available to jump-start their machine learning journey (Chap. 7). Finally, given the realistic challenges related to exchanging data within a global radiological sciences network, the first part concludes with a discussion of data sharing, protection, and bioethics for proper application of machine and deep learning through application of federated learning and FAIR principles in serving the medical field and advancing human welfare (Chap. 8).

- The second part summarizes the successful application of machine learning in medical image analysis, which has been recently boosted with deep learning in computer-aided detection (Chap. 9), computer-aided diagnosis (Chap. 10), and auto-contouring and segmentation (Chap. 11).
- The third part presents applications of machine and deep learning in radiotherapy treatment planning and delivery as a tool for safety and quality assurance (Chap. 12), a computational vehicle for knowledge-based planning (Chap. 13), and intelligent motion management (Chap. 14)
- The last part is dedicated to outcome modeling and decision support, which are at the heart of personalized evidence-based medicine. It starts with machine learning predictive analytics for oncology outcomes (Chap. 15), highlighting the role of imaging (radiomics) and genetics (genomics) in building these models and improving their inter-relationship (radiogenomics) (Chap. 16). Examples of successful machine learning applications in radiotherapy, normal tissue complication, and tumor control probability are demonstrated in Chap. 17. Utilization of machine and deep learning for optimizing decision making and devising smart treatment strategies is discussed in Chap. 18. This part ends with a discussion of the potential of machine/deep learning algorithms for revamping the design of the gold standard for medical practice and clinical trials, by mitigating current challenges and improving their efficacy in Chap. 19.

This book is the product of a coordinated effort by the editors, authors, and publishing team to present the principles and applications of machine and deep learning to a new generation of practitioners in oncology and radiology, with the hope of driving advancements in these fields.

Tampa, FL, USA — Issam El Naqa
Richmond, VA, USA — Martin J. Murphy

Contents

Part I

Introduction to Machine and Deep Learning Principles

What Are Machine and Deep Learning?

1

Issam El Naqa and Martin J. Murphy

1.1 Overview

A machine or a deep learning algorithm is a computational process that uses input data to achieve a desired task without being literally programmed (i.e., "hard coded") to produce a particular outcome. These algorithms are in a sense "soft coded" in that they automatically alter or adapt their architecture through repetition (i.e., experience) so that they become better and better at achieving the desired task. The process of adaptation is called *training*, in which samples of input data are provided along with desired outcomes. The algorithm then optimally configures itself so that it cannot only produce the desired outcome when presented with the training inputs, but can generalize to produce the desired outcome from new, previously unseen data. This training is the "learning" part of machine and deep learning processes. The training does not have to be limited to an initial adaptation during a finite interval. As with humans, a good algorithm can practice "lifelong" learning as it processes new data and learns from its mistakes.

There are many ways that a computational algorithm can adapt itself in response to training. The input data can be selected and weighted to provide the most decisive outcomes. The algorithm can have variable numerical parameters that are adjusted through iterative optimization. It can have a network of possible computational pathways that it arranges for optimal results. It can determine probability distributions from the input data and use them to predict outcomes.

I. El Naqa (✉)
Department of Machine Learning, Moffitt Cancer Center, Tampa, FL, USA

Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: ielnaqa@med.umich.edu; Issam.elnaqa@moffitt.org

M. J. Murphy
Department of Radiation Oncology, Virginia Commonwealth University, Richmond, VA, USA

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_1

The ideal of machine learning is to emulate the way that human beings (and other sentient creatures) learn to process sensory (input) signals in order to accomplish a goal. Traditionally, a machine learning algorithm would feed computer-extracted human-engineered patterns (features) derived from the raw data by, e.g., computer vision methods, to an algorithm to perform a designated learning task; a process colloquially referred to now as *shallow learning*. This is in contrast to a special subcategory of machine learning that allows for combined data representation (e.g., feature extraction) and task learning (e.g., classification or detection) known as deep learning. Conceptually, deep learning comprises learning methods that are provided raw data and which then automatically discover the features needed for detection or classification using the designated machine learning approach. In either learning process, the goal could be, e.g., a task in pattern recognition, in which the learner wants to distinguish apples from oranges. Every apple and orange is unique, but we are still able (usually) to tell one from the other. Rather than hard code a computer with many, many exact representations of apples and oranges, or with an exhaustive set of defining characteristics, it can be programmed to learn to distinguish them through repeated experience with actual apples and oranges. This is a good example of *supervised learning*, in which each training example of input data with features (color, shape, texture, etc.) is paired with its known classification label (apple or orange). It allows the learner to deal with similarities and differences when the objects to be classified have many variable properties within their own classes but still have fundamental qualities that identify them. Most importantly, the successful learner should be able to recognize an apple or an orange that it has never seen before.

A second type of machine learning is the so-called *unsupervised algorithm*. This might have the objective of trying to throw a dart at a bull's-eye. The device (or human) has a variety of degrees of freedom in the mechanism that controls the path of the dart. Rather than try to exactly program the kinematics *a priori*, the learner practices throwing the dart. For each trial, the kinematic degrees of freedom are adjusted so that the dart gets closer and closer to the bull's-eye. This is unsupervised in the sense that the training doesn't associate a particular kinematic input configuration with a particular outcome. The algorithm finds its own way from the training input data. Ideally, the trained dart thrower will be able to adjust the learned kinematics to accommodate, for instance, a change in the position of the target.

A third type of machine learning is *semi-supervised learning*, where part of the data is labeled, and other parts are unlabeled. In such a scenario, the labeled part can be used to aid the learning of the unlabeled part. This kind of scenario lends itself to most processes in nature and more closely emulates how humans develop their skills.

A fourth type of machine learning is *reinforcement learning*, where the algorithm learns to map inputs into optimized actions, i.e., goal-oriented tasks.

These algorithms currently represent the main categories of machine/deep learning, with supervised learning being the most common type in oncology, medical physics, and radiology with applications ranging from detection to diagnosis, drug discovery, and therapeutic interventions. However, several techniques are emerging to relieve the burden and cost of data labeling in supervised learning, including: the semi-supervised approach mentioned above, *transfer learning* (using knowledge

from other domains, such as natural images when learning medical ones), *active learning* (an interactive approach with human beings involved), and more recently *weak supervised learning*, where the labels are assumed to be imprecise or noisy.

There are two particularly important advantages to a successful algorithm. First, it can substitute for laborious and repetitive human effort. Second, and more significantly, it can potentially learn more complicated and subtle patterns in the input data than the average human observer is able to do. Both of these advantages are important to medical physics, oncology, and radiology applications. For example, the daily contouring of tumors and organs at risk during treatment planning is a time-consuming process of pattern recognition that is based on the observer's familiarity and experience with the appearance of anatomy in diagnostic images. That familiarity, though, has its limits, and consequently, there are uncertainties and inter-observer variability in the resulting contours. It is possible that an algorithm for contouring can pick up subtleties of texture or shape in one image or simultaneously incorporate data from multiple sources or blend the experience of numerous observers and thus reduce the uncertainty in the contour.

The complexity of medical physics, oncology, and radiology processes can vary and may involve several stages of sophisticated human–machine interactions and decision- making, which would naturally invite the use of machine/deep learning algorithms to optimize and automate these processes, including but not limited to computer-aided detection, diagnosis, triaging, radiation physics quality assurance, contouring and treatment planning, image-guidance, respiratory motion management, treatment response modeling, and treatment outcomes prediction.

1.2 Background

Machine learning is a category of computer algorithms that are able to emulate some aspects of human intelligence. It draws on ideas from different disciplines such as artificial intelligence, probability and statistics, computer science, information theory, psychology, control theory, and philosophy [1–3]. The relationship between artificial intelligence, machine learning, and deep learning is depicted in Fig. 1.1 [4]. This technology has been applied in such diverse fields as pattern recognition [3], computer vision [5], spacecraft engineering [6], finance [7], entertainment [8, 9], ecology [10], computational biology [11, 12], and biomedical and medical applications [13, 14]. The most important property of these algorithms is their distinctive ability to learn the surrounding environment from input data with or without a teacher [1, 2].

Historically, the inception of machine learning can be traced to the seventeenth century and the development of machines that can emulate human ability to add and subtract by Pascal and Leibniz [15]. In modern history, Arthur Samuel from IBM coined the term "machine learning" and demonstrated that computers could be programmed to learn to play checkers [16]. This was followed by the development of the perceptron by Rosenblatt as one of the early neural network architectures in 1958 [17]. However, early enthusiasm about the perceptron was dampened by the

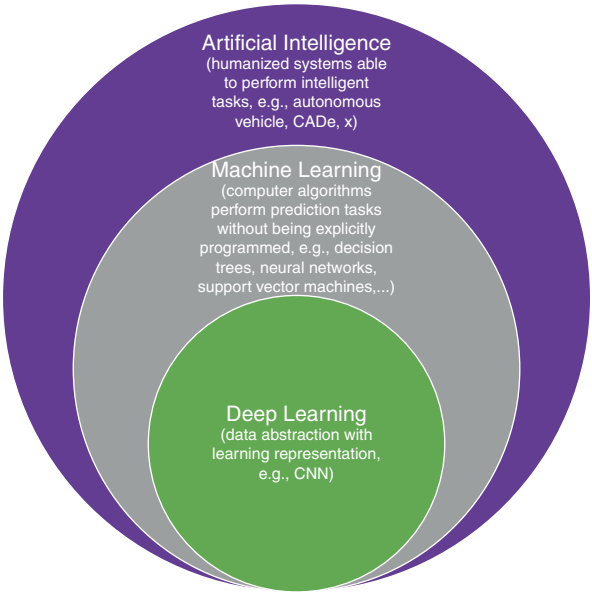


Fig. 1.1 Venn diagram of the relationship between artificial intelligence, machine learning, and deep learning from [4]

observation made by Minsky that the perceptron classification ability is limited to linearly separable problems and not common nonlinear problems such as a simple XOR logic [18]. A breakthrough was achieved in 1975 by the development of the multilayer nonlinear perceptron (MLP) by Werbos [19]. This was followed by the development of decision trees by Quinlan in 1986 [20] and support vector machines by Cortes and Vapnik [21]. Ensemble machine learning algorithms, which combine multiple learners using boosting of weak learners or bagging (model averaging), were subsequently proposed, including Adaboost [22] and random forests [23]. More recently, distributed multilayered learning algorithms such as convolutional neural networks (CNN) have emerged under the notion of deep learning [24]. These algorithms are able to learn good representations of the data that make it easier to automatically extract useful information when building classifiers or other predictors, compared to conventional machine learning algorithms [25] as discussed further below.

1.3 Machine Learning Definition

The field of machine learning has received several formal definitions in the literature. Arthur Samuel in his seminal work defined machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed" [16]. Using a computer science lexicon, Tom Mitchell presented it as "A computer program is said to learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if its performance at tasks in T, as measured by P, improves with experience E" [1]. Ethem Alpaydin in his textbook defined machine

learning as the field of "Programming computers to optimize a performance criterion using example data or past experience" [2]. These various definitions share the notion of coaching computers to intelligently perform tasks beyond traditional number crunching by learning the surrounding environment through repeated examples. The various conventional machine learning algorithms will be reviewed in Chap. 3.

1.4 Deep Learning Definition

Deep learning (DL), as noted earlier, comprises a subcategory of machine learning that deals with representation learning, where raw information or data are fed directly into the algorithm, which can then automatically discover the underlying patterns (features) needed for the detection or classification task [26]. Conceptually, it can be applied to any machine learning technology as depicted in Fig. 1.2, but has been practically shown to be most effective currently with deep neural networks methods [27, 28], which will be thoroughly discussed in Chap. 4.

1.5 Learning from Data

The ability to learn through input from the surrounding environment, whether it is playing checkers or chess games, or recognizing written patterns, or solving the daunting problems in medical physics, oncology, or radiology, is the key to a successful machine learning application. Learning is defined in this context as estimating dependencies from data [29].

The fields of data mining and machine learning are intertwined. Data mining utilizes machine learning algorithms to interrogate large databases and discover hidden knowledge in the data, while many machine learning algorithms employ data

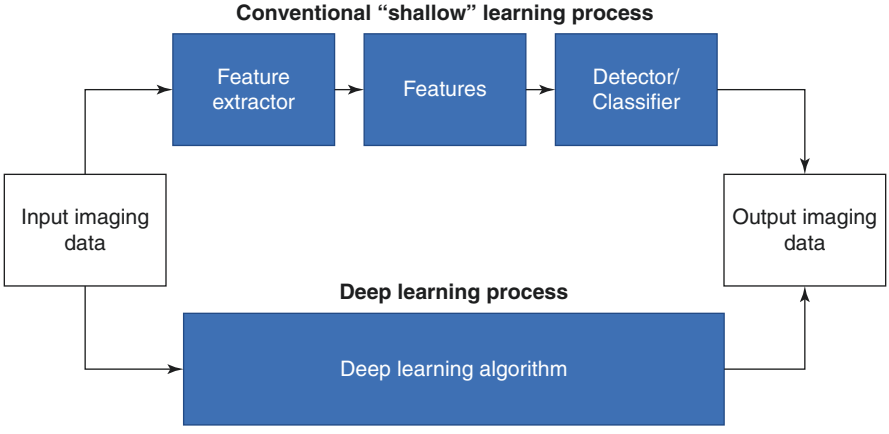


Fig. 1.2 Conventional "shallow" machine learning (top) versus deep learning algorithms, where image data representation and classification are handled within the same framework

mining methods to preprocess the data before learning the desired tasks [30]. However, it should be noted that machine learning is not limited to solving database-like problems but also extends into solving complex artificial intelligence challenges by learning and adapting to a dynamically changing situation, as is encountered in a busy radiation oncology practice, for instance.

Machine/deep learning has both engineering science aspects such as data structures, algorithms, probability and statistics, and information and control theory and social science aspects that draw on ideas from psychology and philosophy.

1.6 Overview of Machine and Deep Learning Approaches

Machine or deep learning can be divided according to the nature of the data labeling into supervised, unsupervised, semi-supervised, and reinforcement learning as shown in Fig. 1.3. Supervised learning is used to estimate an unknown input-output mapping from known input-output samples, where the output is labeled (e.g., classification and regression). In unsupervised learning, only input samples are given to the learning system (e.g., clustering and estimation of probability density function). Semi-supervised learning is a combination of both supervised and unsupervised where part of the data is partially labeled and the labeled part is used to infer the unlabeled portion (e.g., text/image retrieval systems). In reinforcement learning, the machine learning algorithm aims to control learning by accommodating a feedback system, in which an agent attempts to take a sequence of actions that may maximize a cumulative reward such as winning a game of checkers, for instance [31]. This kind of approach is particularly useful for adaptive or sequential decision-making applications as will be discussed in Chap. 19.

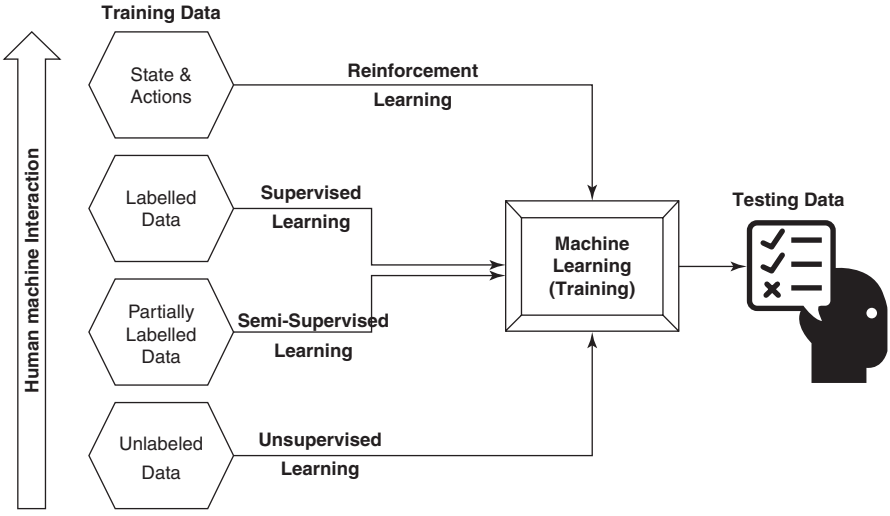


Fig. 1.3 Categories of machine learning algorithms according to training data nature

From a concept learning perspective, machine learning can be categorized into transductive and inductive learning [32]. Transductive learning involves the inference from specific training cases to specific testing cases using discrete labels as in clustering or using continuous labels as in manifold learning. On the other hand, inductive learning aims to predict outputs from inputs that the learner has not encountered before. Along these lines, Mitchell argues for the necessity of an inductive bias in the training process to allow for a machine learning algorithm to generalize beyond unseen observation [33].

From a probabilistic perspective, machine learning algorithms can be divided into discriminant or generative models. A discriminant model measures the conditional probability of an output given typically deterministic inputs, such as neural networks or a support vector machine. A generative model is fully probabilistic whether it is using a graph modeling technique such as Bayesian networks, or not, as in the case of naïve Bayes.

1.7 Quantifying the Data and Learning Objectives

The first step in the execution of a machine learning algorithm is the identification of the salient characteristics of the process to be emulated or the entity to be recognized or classified. These characteristics must necessarily be quantitative because this is, after all, a computational problem. The characteristics are extracted from the raw input data and then assembled into a "feature vector" that is presented to the algorithm. The extraction almost invariably involves data compression to avoid completely overwhelming the subsequent computational steps. For example, when we look at an image, we don't see individual pixels, we see recognizable structures. The art of feature extraction is to make the algorithm "see" structures and traits in the input data. The smaller the feature vector, the better, but it is critical that it be adequate to accurately represent the data and learning objectives. The identification and quantification of the most useful features is a fundamental part of the art of designing a machine learning algorithm, which has recently been automated in the context of deep learning.

In object classification (e.g., apples and oranges), the features could be empirical attributes that are directly quantifiable, such as dimensions, weight, density, etc., or indirectly quantifiable, such as color, texture, or smell. The indirect features need to be preprocessed further to convert them to numerical measures.

Formal features can be extracted via data transformation or reduction techniques. If the raw input data have many, many discrete elements, such as pixel values in an image, then using the entire image as the feature vector would have prohibitive computational overhead. However, if those elements are not random, then the size of the input feature vector can be dramatically reduced with minimal loss by methods of dimensionality reduction and compression such as principal component analysis (PCA) or Fourier analysis. PCA transforms a complex set of correlated data elements into a set of maximally uncorrelated principal component basis vectors and their associated coefficients. A linear combination of the basis vectors and

coefficients reproduces the original data set with an accuracy that is determined by the number of vectors that are retained from the analysis. In highly correlated data, a very small number of PCA vectors and coefficients can be sufficient to characterize its structure. The most significant coefficients are then collected into the feature vector. Fourier decomposition of the input data into a set of Fourier basis vectors and coefficients achieves the same goal, but the difference is that the PCA method requires an initial set of representative training examples to determine the principal components, while Fourier decomposition can be done case by case using fixed basis vectors. The Fourier transform method lends itself naturally to image compression, as is well known from the JPEG algorithm, but it can require many more coefficients to capture salient image content than the PCA method. Both of these methods lend themselves naturally to pattern recognition and classification algorithms such as neural networks and support vector machines. Formal feature extraction or representation also lends itself naturally to deep learning applications, which automates the process by functioning as the interface between the raw input data and the learning algorithm.

1.8 Application in Biomedicine

Machine learning algorithms have witnessed increased use in biomedicine, starting naturally in neuroscience and cognitive psychology through the seminal work of Donald Hebb in his 1949 book [34] developing the principles of associative or Hebbian learning as a mechanism of neuron adaptation and the work of Frank Rosenblatt developing the perceptron in 1958 as an intelligent agent [17]. This was shortly followed by Ledley and Lusted in their 1959 paper, where they anticipated the role of a probabilistic logic-based approach to understand and support physicians' reasoning [35]. An early major machine learning initiative was the MYCIN project at Stanford in the 1970s, which was a rule-based system to identify bacteria types that may cause infectious diseases [36], achieving an acceptability rating of 65% from a panel of experts [37]. Recent reviews of the application of machine learning in biomedicine and medicine can be found in [12, 13, 38, 39].

1.9 Applications in Radiology and Oncology

Among the earliest adoptions of machine learning algorithms was in the field of radiological and medical image analysis. Winsberg et al. reported in 1967 on a computer detection algorithm for radiographic abnormalities in mammograms [40]. Lodwick et al. presented a roentgenograms concept for analyzing bone and lung cancer images [41, 42] and Meyers et al. developed an automated computer analysis of cardiothoracic ratios [43]. However, the major thrust happened in the 1980s, when tremendous developments occurred in computer-aided detection (CADe) and computer-aided diagnosis (CADx), providing radiologists with computer output as a "second opinion" to aid in making final decisions [44–49]. These CAD systems

utilized image feature-based analysis for the detection of microcalcifications in mammogram images [50–53] and lung nodules in digital chest radiographs [54]. This expanded into every area in radiology, in the form of decision support systems. In the field of oncology and specifically, radiation oncology, early applications of machine learning have focused on treatment planning and predicting normal tissue toxicity [55–57], but its application has since branched into almost every part of the field, including tumor response modeling, radiation physics quality assurance, contouring and treatment planning, image-guided radiotherapy, and respiratory motion management. Examples of the application of machine and deep learning will be the main subject of the second half of this book.

1.10 Ethical Challenges in the Application of Machine Learning

The application of machine learning in medicine has not been without challenges and even controversies. This is understandable given the data-driven nature of these algorithms and caveats related to data sharing, provenance, patient privacy, and the nature of medical data acquisition, which not only vary in technologies and parameters but also shift over time with new developments. Moreover, issues related to learning bias [58] and adversarial examples [60, 61] need to be accounted for. For instance, a machine learning algorithm developed for predicting the risk of pneumonia counter-intuitively suggested that patients with pneumonia and asthma would be at a lower risk of death than patients with pneumonia but without asthma [59]. Similar controversial examples were noted in the case of skin cancer risk prediction, where the presence of a ruler in the image may be a cue for the ML algorithm of high risk [62] or the appearance of a tube in a chest X-ray being indicative of severe lung disease [63]. These examples and others stress the importance of data quality and context when training and applying these powerful tools.

These challenges have led the Food and Drug Administration (FDA) in the United States, the European Union, and other international bodies to advocate for lawful, ethical and robust application from technological and societal perspectives. Towards this goal there have been shifts towards developing more explainable/interpretable machine learning algorithms [64], which would allow for better transparency, oversight, and accountability.

1.11 Steps to Machine Learning Heaven

For the successful application of machine learning in general and in medical physics, radiology and oncology in particular, one first needs to properly characterize the nature of the problem, in terms of the input data and the desired outputs. Secondly, despite the robustness of machine learning to noise, a good model cannot substitute for bad data, keeping in mind that models are primarily built on approximations, and it has been stated that “All models are wrong; some models are useful (George

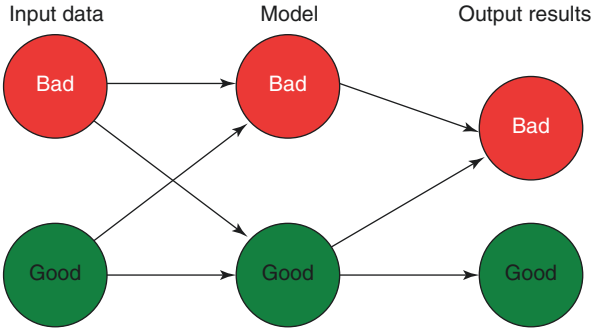


Fig. 1.4 GIGO paradigm. Learners cannot be better than the data

Box)." Additionally, this has been stated as the GIGO principle, garbage in garbage out as shown in Fig. 1.4 [65].

Thirdly, the model needs to generalize beyond the observed data into unseen data, as indicated by the inductive bias mentioned earlier. To achieve this goal, the model needs to be kept as simple as possible but not simpler, a property known as parsimony, which follows from Occam's razor that "Among competing hypotheses, the hypothesis with the fewest assumptions should be selected." Analytically, the complexity of a model could be derived using different metrics such as Vapnik–Chervonenkis (VC) dimension discussed in Chap. 2 for instance [32]. However, deep learning algorithms with their large number of layers for learning data representation and performing model prediction in the same architecture, may present a future challenge to this classical notion, but the overall objective remains the same, that is, to achieve generalizability to out-of-sample data, which should be carefully evaluated as discussed in Chap. 6. Finally, a major limitation in the adoption of machine learning in general and deep learning in particular by the larger medical community is the "black box" stigma and the inability to provide an intuitive interpretation of the learned process that could help clinical practitioners better understand their data and trust the model predictions. This is an active and necessary area of research that requires special attention from the machine learning community working in biomedicine. Solutions such as deriving proxy models, developing attention maps, providing disentangled representation or learning with known operators have been emerging to create a more interpretable/explainable machine learning paradigm [66–70].

1.12 Conclusions

Machine and deep learning present computer algorithms that are able to learn from the surrounding environment to optimize the solution for the task at hand. It builds on expertise from diverse fields such as artificial intelligence, probability and statistics, computer science, information theory, and cognitive neuropsychology. Machine learning algorithms can be categorized into different classes according to the nature of the data, its representation, the learning process, and the model type. Machine

learning has a long history in biomedicine, particularly in radiology, but its application in medical physics and oncology is in its infancy, with high potential and promising future to improve the safety and efficacy of clinical care and advance cancer research discovery.

References

1. Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.
2. Alpaydin E. Introduction to machine learning. 3rd ed. Cambridge, MA: The MIT Press; 2014.
3. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
4. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol. 2020;93:20190855.
5. Apolloni B. Machine learning and robot perception. Berlin: Springer; 2005.
6. Ao S-I, Rieger BB, Amouzegar MA. Machine learning and systems engineering. Dordrecht/New York: Springer; 2010.
7. Györfi L, Ottucsák G, Walk H. Machine learning for financial engineering. Singapore/London: World Scientific; 2012.
8. Gong Y, Xu W. Machine learning for multimedia content analysis. New York/London: Springer; 2007.
9. Yu J, Tao D. Modern machine learning techniques and their applications in cartoon animation research. 1st ed. Hoboken: Wiley; 2013.
10. Fielding A. Machine learning methods for ecological applications. Boston: Kluwer Academic; 1999.
11. Mitra S. Introduction to machine learning and bioinformatics. Boca Raton: CRC; 2008.
12. Yang ZR. Machine learning approaches to bioinformatics. Hackensack: World Scientific; 2010.
13. Cleophas TJ. Machine learning in medicine. New York: Springer; 2013.
14. Malley JD, Malley KG, Pajevic S. Statistical learning for biomedical data. Cambridge: Cambridge University Press; 2011.
15. Ifrah G. The universal history of computing: from the abacus to the quantum computer. New York: John Wiley; 2001.
16. Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev. 1959;3:210–29.
17. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65:386–408.
18. Minsky ML, Papert S. Perceptrons; an introduction to computational geometry. Cambridge, MA: MIT Press; 1969.
19. Werbos PJ. Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD thesis, Harvard University; 1974.
20. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1:81–106.
21. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
22. Schapire RE. A brief introduction to boosting. In: Proceedings of the 16th international joint conference on artificial intelligence, vol. 2. Stockholm: Morgan Kaufmann; 1999. p. 1401–6.
23. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
24. Hinton GE. Learning multiple layers of representation. Trends Cogn Sci. 2007;11:428–34.
25. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35:1798–828.
26. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6. https://doi.org/10.1038/323533a0.
27. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7. https://doi.org/10.1126/science.1127647.

28. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.
29. Cherkassky VS, Mulier F. Learning from data: concepts, theory, and methods. 2nd ed. Hoboken: IEEE Press/Wiley-Interscience; 2007.
30. Kargupta H. Next generation of data mining. Boca Raton: CRC Press; 2009.
31. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge, MA: MIT Press; 1998.
32. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
33. Mitchell TM. The need for biases in learning generalizations. New Brunswick: Rutgers University; 1980.
34. Hebb DO. The organization of behavior; a neuropsychological theory. New York: Wiley; 1949.
35. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science. 1959;130(3366):9–21.
36. Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. Math Biosci. 1975;23(3):351–79. https://doi.org/10.1016/0025-5564(75)90047-4.
37. Yu VL, Fagan LM, Wraith SM, Clancey WJ, Scott AC, Hannigan J, Blum RL, Buchanan BG, Cohen SN. Antimicrobial selection by a computer: a blinded evaluation by infectious diseases experts. JAMA. 1979;242(12):1279–82. https://doi.org/10.1001/jama.1979.03300120033020.
38. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58. https://doi.org/10.1056/NEJMra1814259.
39. Saria S, Butte A, Sheikh A. Better medicine through machine learning: what's real, and what's artificial? PLoS Med. 2018;15(12):e1002721. https://doi.org/10.1371/journal.pmed.1002721.
40. Winsberg F, et al. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. Radiology. 1967;89(2):211–5.
41. Lodwick GS, Keats TE, Dorst JP. The coding of Roentgen images for computer analysis as applied to lung cancer. Radiology. 1963;81(2):185–200.
42. Lodwick GS, et al. Computer diagnosis of primary bone tumors. Radiology. 1963;80(2):273–5.
43. Meyers PH, et al. Automated computer analysis of radiographic images. Radiology. 1964;83(6):1029–34.
44. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph. 2007;31(4-5):198–211.
45. Doi K, et al. Artificial intelligence and neural networks in radiology: Application to computer-aided diagnostic schemes. In: Hendee W, Trueblood J, editors. Digital imaging. AAPM medical physics monograph; 1993. p. 301–22.
46. Giger M, et al. Computer-aided diagnosis in mammography. In: Sonka M, Fitzpatrick M, editors. Handbook of medical imaging. Philadelphia, PA: SPIE; 2000. p. 915–1004.
47. Giger ML. Future of breast imaging. Computer-aided diagnosis. In: Haus A, Yaffe M, editors. AAPM/RSNA categorical course on the technical aspects of breast imaging; 1992. p. 257–70.
48. Giger ML. Computer-aided diagnosis in radiology. Acad Radiol. 2002;9(1):1–3.
49. Swett H, Giger M, Doi K. Computer vision and decision support. In: Hendee W, Wells P, editors. Perception of visual information. Berlin: Springer-Verlag; 1993. p. 272–315.
50. Chan HP, et al. Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography. Med Phys. 1987;14(4):538–48.
51. El-Naqa I, Yang Y, Wernick MN, Galatsanos NP, Nishikawa RM. A support vector machine approach for detection of microcalcifications. IEEE Trans Med Imaging. 2002;21:1552–63.
52. Gurcan MN, Chan HP, Sahiner B, Hadjiiski L, Petrick N, Helvie MA. Optimal neural network architecture selection: improvement in computerized detection of microcalcifications. Acad Radiol. 2002;9:420–9.
53. El-Naqa I, Yang Y, Galatsanos NP, Nishikawa RM, Wernick MN. A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans Med Imaging. 2004;23:1233–44.

54. Giger ML, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. Med Phys. 1988;15(2):158–66.
55. Gulliford SL, Webb S, Rowbottom CG, Corne DW, Dearnaley DP. Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate. Radiother Oncol. 2004;71:3–12.
56. Munley MT, Lo JY, Sibley GS, Bentel GC, Anscher MS, Marks LB. A neural network to predict symptomatic lung injury. Phys Med Biol. 1999;44:2241–9.
57. Su M, Miften M, Whiddon C, Sun X, Light K, Marks L. An artificial neural network for predicting the incidence of radiation pneumonitis. Med Phys. 2005;32:318–25.
58. Raji ID, Buolamwini J. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Honolulu, HI: ACM; 2019. p. 429–35.
59. Caruana R, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, NSW: ACM; 2015. p. 1721–30.
60. Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recognit. 2018;84:317–31.
61. Finlayson SG, et al. Adversarial attacks on medical machine learning. Science. 2019;363(6433):1287–9.
62. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
63. Rajpurkar P, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. In: arXiv e-prints; 2017.
64. Luo Y, et al. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. BJR Open. 2019;1(1):20190021.
65. Tweedie R, Mengersen K, Eccleston J. Garbage in, garbage out: can statisticians quantify the effects of poor data? Chance. 1994;7:20–7.
66. Philbrick KA, et al. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. Am J Roentgenol. 2018;211(6):1184–93.
67. Seah JCY, et al. Chest radiographs in congestive heart failure: visualizing neural network learning. Radiology. 2019;290:514–22.
68. Luna JM, et al. Building more accurate decision trees with the additive tree. Proc Natl Acad Sci U S A. 2019;116(40):19887–93.
69. Nazmul Haque K, Latif S, Rana R. Disentangled representation learning with information maximizing autoencoder. In: arXiv e-prints; 2019.
70. Maier AK, et al. Learning with known operators reduces maximum error bounds. Nat Mach Intell. 2019;1(8):373–80.

Computational Learning Theory

2

Issam El Naqa and Jen-Tzung Chien

2.1 Introduction

In many computational learning problems, we are given a relatively small number of observed data samples from the general population and asked to understand the functional dependencies and make decisions or perform tasks based on the data accordingly. In standard statistics introduced by Ronald Fisher in the 1920–1930s and in his classic textbooks [1, 2], learning dependencies are based on the concepts of sufficiency and ancillary statistics, which requires representing dependencies by a finite set of parameters and then estimating these using maximum likelihood or Bayesian techniques. However, a paradigm shift in statistical learning theory was introduced in the 1960s by Vladimir Vapnik and his colleagues in which the parameter estimation restrictions imposed by Fisher's paradigm are replaced by knowledge of some general properties of the set of functions to which the unknown dependencies belong. The determination of the general conditions for estimating the unknown data dependency, the description of the inductive learning of relationships, and the development of algorithms to implement these principles are the subjects of the modern *computational learning theory* [3].

In this framework of learning theory, the focus is on small sample size statistics, in which a machine learning algorithm is trained on a subset of the data (training

I. El Naqa (✉)
Department of Machine Learning, Moffitt Cancer Center, Tampa, FL, USA

Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: ielnaqa@med.umich.edu; Issam.elnaqa@moffitt.org

J.-T. Chien
Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan

Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
e-mail: jtchien@nctu.edu.tw

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_2

data) that is used to identify the learning function to achieve the desired response of the task at hand and is built with the goal of predicting response to unseen data (out-of-sample or test data). This is a challenging task that poses several questions regarding which learning process we should select, what is the learning capacity of the algorithm selected, what are the expected errors or their bounds, under what conditions is successful learning possible or impossible, and under what conditions is a particular learning algorithm assured of learning successfully [3, 4]. These theories are further challenged in the context of deep learning as will be presented here. In this chapter, we will start by highlighting the differences between statistical analysis and statistical modeling. We will present the theoretical background for computational learning. Specifically, the frameworks for analyzing learning algorithms, namely, the probably approximately correct (PAC) and Vapnik–Chervonenkis (VC) theory, will be discussed. This is in addition to the specific deep learning context. Finally, practical methods for estimating learning generalization ability and model complexity will be presented.

2.2 Computational Modeling Versus Statistics

There is a common mix-up between statistical analysis and computational modeling of data. The objective of statistical analysis is to use statistics to describe data and make inferences on the population for *hypothesis testing* purposes; for instance, variable x is significant while variable y is not in explaining the observed clinical endpoint of interest. In the case of computational modeling, the objective is to provide an adequate description of dependencies in observation data and summarize its latent features for *hypothesis generation* as summarized in Fig. 2.1 [5].

Machine learning is a branch of computational modeling that inherited many of its properties and utilizes the statistical modeling techniques as part of its arsenal. For instance, machine learning models of quality assurance (QA) in radiation oncology can capture many salient features in the data that may impact quality of

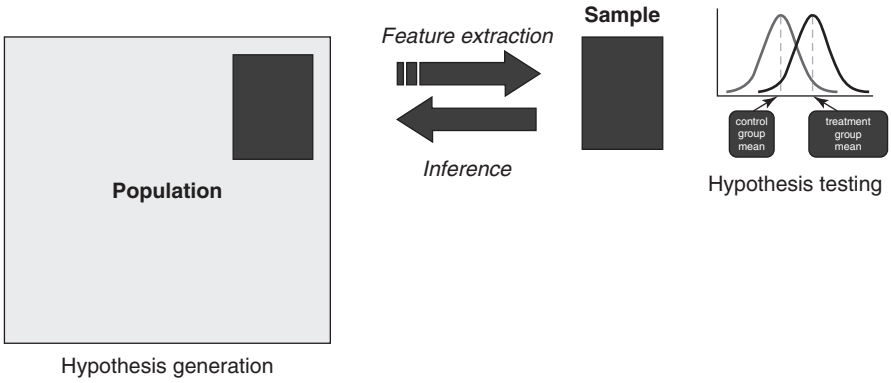


Fig. 2.1 Computational modeling vs. statistical analysis. (Adapted from [5])

delivered treatment and their possible interdependencies, which could be further tested for varying hypotheses for their severity and possible action levels to mitigate their effect. However, development of computational modeling techniques could be achieved using both deterministic and statistical methodologies. On the other hand, deep learning is a subcategory of machine learning, where the algorithm can learn the data representation from raw information (no human-engineered feature extraction is required), and subsequently conduct the learning task on these data.

2.3 Learning Capacity

Learning capacity or "learnability" defines the ability of a machine learning algorithm to learn the task at hand in terms of model complexity and the number of training samples required to optimize a performance criteria. Using formal statistical learning taxonomy [6], assuming a training set Ξ of n-dimensional vectors, $x_i^n, i = 1:m$, each labeled (by 1 or 0) according to a target function f which is unknown to the learner and called the target concept and is denoted by c, which belongs to the set of functions, C, the space of target functions as illustrated in Fig. 2.2. The probability of any given vector X being present in Ξ is $P(X)$. The goal of the training is to guess a function, $h(X)$ based on the labeled samples in Ξ, called the hypothesis. We assume that the target function is an element of a set H in the space of hypotheses. For instance, in our QA example, if we are interested in developing a treatment plan quality metric, we would have a list of input features X (e.g., energies, beam arrangements, monitor units) that is governed in our pool of treatment plans with a certain joint probability density function P. Based on clinical experience, a set of these plans are considered to be good while others are bad, which would constitute the target concept (c) of interest with an unknown functional form f that we aim to estimate. During the training process, we attempt to identify a hypothesis function $h(X)$ that would approximate the mapping to c using varying possible machine learning algorithms, and the higher the overlap between our hypothesized mapping function and the target quality metric concept, the more successful the learning process is as indicated in the Venn diagram of Fig. 2.2.

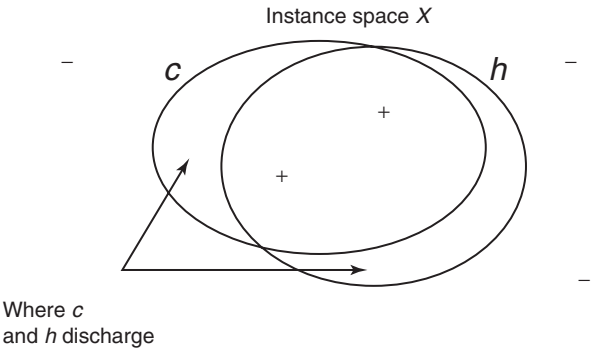


Fig. 2.2 Illustration of learning concepts. (From Nilsson [6])

There are two main theories that attempt to characterize the learnability of classical machine learning algorithms: the PAC and the VC theories as discussed below. The special context of deep learning will be discussed in a subsequent section.

2.4 PAC Learning

One method to characterize the learnability of a machine learning algorithm is by the number of training examples needed to learn a hypothesis $h(X)$ as mentioned earlier. This could be measured by the probability of learning a hypothesis that is approximately correct (PAC). Formally, this could be defined as follows. Consider the concept class C defined over a set of instances X of length m and a learner L using hypothesis space H. C is PAC learnable by L using H.

If for all $c \in C$, distributions D over X, ε such that $0<\varepsilon<1/2$ and δ such that $0<\delta<1/2$, there is a learner L with probability at least $(1-\delta)$ that will output a hypothesis $h \in H$ such that error $D(h) \leq \varepsilon$, in time that is polynomial in $1/\varepsilon$, $1/\delta$, n, and size (c) [4]. For a finite hypothesis space H, the number of training examples (m) required to reduce the probability of error below a desired level δ is given by assuming a zero training error:

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\left(1/\delta\right)\right) \tag{2.1}$$

This estimated number of training examples is sufficient to ensure that any consistent hypothesis will be probably (with probability $(1-\delta)$) approximately (within error ε) correct. In the case the training error is not necessarily zero, the number of required training examples becomes:

$$m \geq \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\left(1/\delta\right)\right) \tag{2.2}$$

It is recognized that such an estimate could be in practice an overestimate [4]. Another problem in PAC is that it includes the size of the hypothesis space H, which in practice could be unknown or infinite.

2.5 VC Dimension

An alternative approach to measure learnability that overcomes the limitations of PAC is to use Vapnik–Chervonenkis (VC) dimension [3]. The VC dimension measures the complexity of the hypothesis space H, not by the number of distinct hypotheses H as in PAC but rather by the number of distinct instances from X that can be completely discriminated using H. VC(H), of hypothesis space H defined over instance space X, is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $\mathrm{VC}(H)=\infty$. This is noted that for any finite H, $\mathrm{VC}(H) \leq \log 2|H|$. To see this, suppose that $\mathrm{VC}(H)=d$. Then, H will require $2d$ distinct hypotheses to shatter d instances. Hence, $2^d \leq \log 2|H|$. This

is illustrated in Fig. 2.2. However, if the number of hypotheses far exceeds the number of training samples available, then we are faced with the phenomenon of *curse of dimensionality.* This curse or inability to learn and generalize with high-dimensional data has motivated the development of deep learning approaches [7], which will be discussed next.

2.6 Learning with Deep Learning

Deep learning (DL) as noted earlier is a subcategory of machine learning that deals with representation learning, where raw information or data are fed directly into algorithm, which can automatically discover the underlying patterns (or latent features) needed for the detection or classification task [8]. Conceptually, this can be applied to any machine learning technology as was depicted in Fig. 1.2, but has been practically shown to be most effective currently with deep neural networks (DNNs) methods [9, 10]. DNN description and their architectures will be detailed in Chap. 4; here we focus on the underlying learning principles. One explanation is provided through the *manifold learning* hypothesis, where the interesting parts of the input occur only along a collection of manifolds with a small number of features and that the DNN learns how to represent the data in terms of the coordinates of such a manifold by transforming from one layer to the next as depicted in Fig. 2.3 [7]. Another interpretation has focused on *tensor factorization* and how a positively homogeneous regularizer of the same degree as the network can lead to a global optimal solution in non-convex problems [11].

2.7 Model Complexity Analysis in Practice

Multivariate analysis often involves a large number of variables or features in the data samples [7]. The complexity of a learning model increases with the number of input features (i.e., the dimensionality of the input feature vector); therefore, it is desirable to focus on the most important features that characterize the observations. These are usually unknown. Therefore, practical dimensionality reduction or subset selection aims to find the "significant" set of features. Finding the best subset of

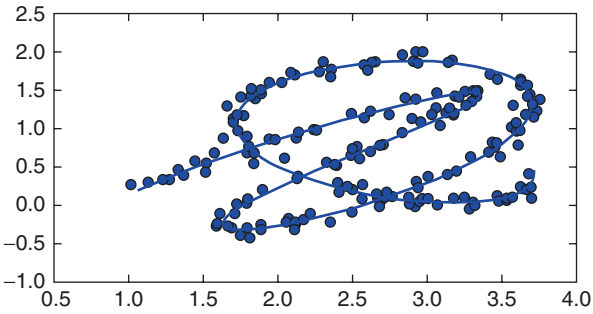


Fig. 2.3 Data sampled from a distribution in a two-dimensional space that is concentrated near a one-dimensional manifold, like a twisted string. The solid line indicates the underlying manifold that the learner (e.g., DNN) should infer [7]

features is definitely challenging, especially in the case of nonlinear models. The objective is to reduce the model complexity, decrease the computational burden, and improve the generalizability on unseen data as explained earlier. A straightforward approach is to make an educated guess based on experience and domain knowledge and then apply feature transformation (e.g., principal component analysis (PCA)) [12–14] or sensitivity analysis by using the organized searches such as sequential forward selection or sequential backward selection or a combination of both [14, 15]. A recursive elimination technique that is based on machine learning has been also suggested [16]. In this technique, the data set is initialized to contain the whole set, train the predictor (e.g., support vector machine (SVM) classifier) on the data, rank the features according to a certain criteria (e.g., the weight of the feature), and keep iterating by eliminating the lowest ranked one. It should be noted that the specific definition of model order changes depending on the functional form. It could be identified by the number of parameters in logistic regression, or by the number of neurons and layers in the case of neural networks (cf. Fig. 2.4), etc. However, in any of these forms, the model order creates a balance between the model complexity (increased model order) and the model ability to generalize to unseen data. Finding this balance is referred to in statistical learning theory as the bias–variance dilemma (see Fig. 2.5), in which an oversimple model is expected to underfit the data (large bias and small variance), whereas a too complex model is expected to overfit data (small bias and large variance) [17]. Hence, the objective is to achieve an optimally parsimonious model, i.e., a model with the correct degree of complexity to fit the data and also a maximum ability to generalize to new and unseen data sets, in other words to derive its VC dimension from the data itself. Practical approaches utilize information theoretic methods or statistical resampling as discussed below.

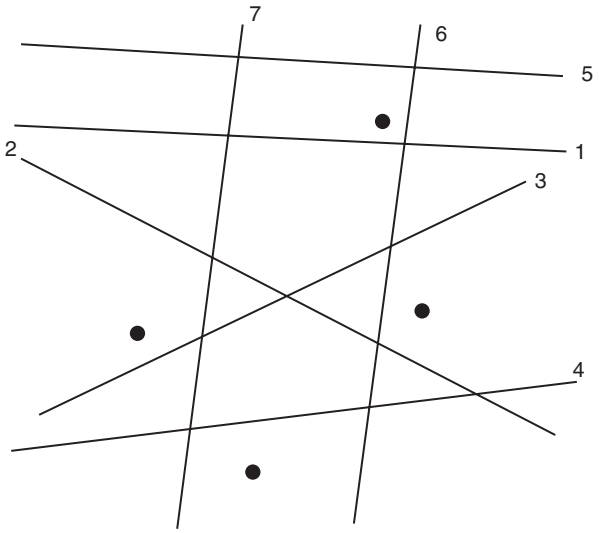


Fig. 2.4 An example of 14 dichotomies shattering 4 points in 2D. (From Nilsson [6])

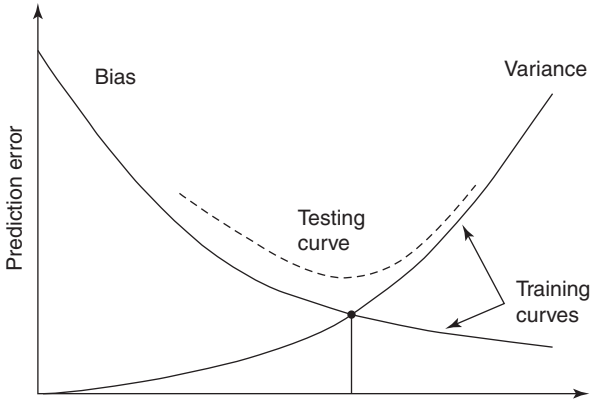


Fig. 2.5 This figure illustrates a common trade-off in model predictive power between prediction bias (average error) and prediction variance (square error). As model complexity increases, the average prediction error (bias) tends to decrease while the average square error tends to decrease. The point of optimal complexity tends to be near the point when average and square errors are of similar magnitude. (Reproduced with permission from Deasy and El Naqa [18])

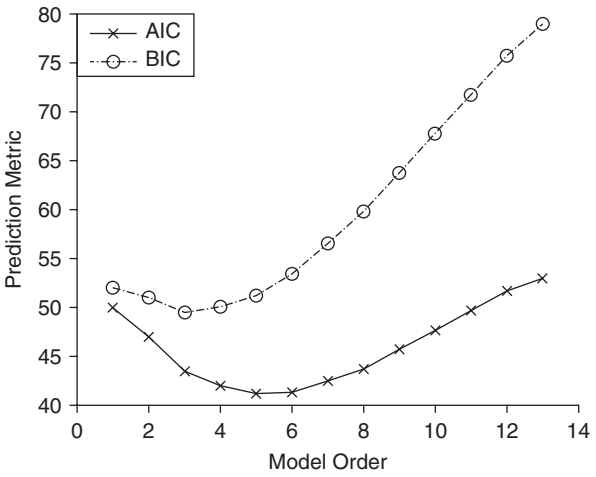


Fig. 2.6 A plot of AIC and BIC for xerostomia with logistic regression models, which correct the log-likelihood for the effects of the fitting process, with model orders of 5 and 3 variables, respectively [21]

2.7.1 Model Order Based on Information Theory

Information theory provides intuitive measures of model order optimality; among the most commonly used are Akaike information criteria (AIC) and the Bayesian information criteria (BIC) [19]. AIC is an estimate of predictive power of a model, which includes both the maximum likelihood principle and a model complexity term that penalizes models with an increasing number of parameters (to avoid overfitting the data). BIC is derived from Bayesian theory, which results in a penalty term that increases linearly with the number of parameters. An example in modeling xerostomia (dry mouth) in head and neck cancer post-radiotherapy is shown in Fig. 2.6.

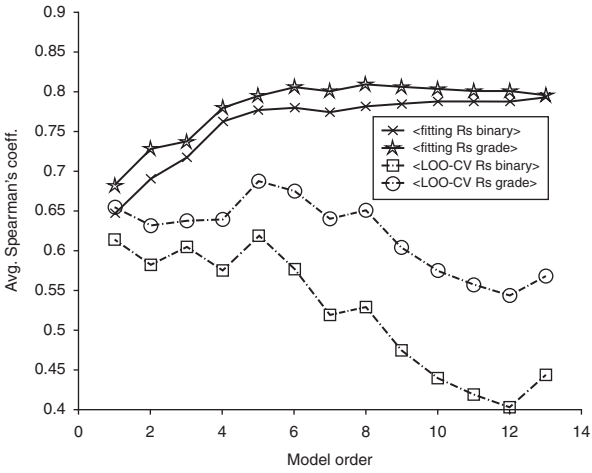


Fig. 2.7 A comparison of leave-one-out cross-validation (LOO-CV) and the training bootstrap (top two curves) as a function of model size for the xerostomia data using Spearman rank coefficient (Rs) as metric [21]

2.7.2 Model Order Based on Resampling Methods

Resampling techniques are used for model selection and performance comparison purposes to provide statistically sound results when the available data set is limited (which is almost always the case in oncology). We use two types of fit-then-validate methods: cross-validation methods and bootstrap resampling techniques. *Cross-validation* [14] uses some of the data to train the model and some of the data to test the model validity. The type we most often use is the "leave-one-out" cross-validation (LOO-CV) procedure (also known as the "jackknife"). In each LOO-CV iteration, all the data are used for training/fitting except for one data point left out for testing, and this is repeated so that each data point is left out exactly once. The overall success of predicting the left-out data is a quantitative estimate of model performance on new data sets. *Bootstrapping* [20] is an inherently computationally intensive procedure but generates more realistic results. Typically, a bootstrap pseudo-data set is generated by making copies of original data points and randomly selected with a probability of inclusion of 63%. The bootstrap often works acceptably well even when data sets are small or unevenly distributed. To achieve valid results, this process must be repeated many times, typically several hundred or thousand times. Examples of applying these methods to outcomes modeling in radiotherapy (cf. Fig. 2.7) can be found in our previous work [21] and are discussed in detail in [18].

2.8 Conclusions

In this chapter, we discussed some of the guiding principles of computational learning. Within the probably approximately correct (PAC) framework, we identify classes of hypotheses that can and cannot be learned from a polynomial number of

training examples and we define a natural measure of complexity for hypothesis spaces that allows bounding the number of training examples required for inductive learning. Within the mistake-bound framework, we examine the number of training errors that will be made by a learner before it determines the correct hypothesis [4]. The VC dimension offers an alternative approach for measuring learnability by estimating the number of instances necessary to discriminate among hypotheses. We also discussed some of the underlying principles to explain the generalizability of deep learning algorithm via manifold learning or tensor factorization. Beside these theoretical approaches, we also presented practical methods based on information theory and statistical resampling for estimating model complexity. Resampling techniques such as cross-validation and bootstrapping are among the most-used methods in the literature and will be further discussed in the context of performance evaluation in Chap. 6.

References

1. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.
2. Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd; 1935.
3. Vapnik VN. The nature of statistical learning theory. New York: Springer; 2000.
4. Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.
5. Berry MJA, Linoff G. Data mining techniques: for marketing, sales, and customer relationship management. 2nd ed. Indianapolis, IN: Wiley; 2004.
6. Nilsson NJ. The mathematical foundations of learning machines. San Mateo: Morgan Kaufmann; 1990.
7. Goodfellow I, Bengio Y, Courville A. Deep learning. Adaptive computation and machine learning. Cambridge, MA: The MIT Press; 2016.
8. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6. https://doi.org/10.1038/323533a0.
9. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.
11. Haeffele BD, Vidal R. Global optimality in tensor factorization, deep learning, and beyond. In: arXiv e-prints; 2015.
12. Guyon I, Elissee A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
13. Dawson LA, Biersack M, Lockwood G, Eisbruch A, Lawrence TS, Ten Haken RK. Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation. Int J Radiat Oncol Biol Phys. 2005;62:829–37.
14. Kennedy R, Lee Y, Van Roy B, Reed CD, Lippman RP. Solving data mining problems through pattern recognition. Upper Saddle River, NJ: Prentice Hall; 1998.
15. Härdle W, Simar L. Applied multivariate statistical analysis. Berlin/New York: Springer; 2003.
16. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.
17. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations. New York: Springer; 2001.
18. Deasy JO, El Naqa I. Image-based modeling of normal tissue complication probability for radiation therapy. Cancer Treat Res. 2008;139:215–56.

19. Burnham KP, Anderson DR. Model selection and multimodal inference: a practical information-theoretic approach. 2nd ed. New York: Springer; 2002.
20. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
21. El Naqa I, Bradley JD, Lindsay PSE, Blanco AI, Vicic M, Hope AJ, et al. Multi-variable modeling of radiotherapy outcomes including dose-volume and clinical factors. Int J Radiat Oncol Biol Phys. 2006;64:1275–86.

Conventional Machine Learning Methods

3

Sangkyu Lee and Issam El Naqa

3.1 Introduction

Learning is defined in this context as estimating statistical dependencies from data. There are three common types of learning [1]: unsupervised, supervised, and reinforcement learning. In *unsupervised learning*, only input samples are given to the learning system (e.g., clustering and estimation of probability density function). *Supervised learning* is used to estimate an unknown (input, output) mapping from known (input, output) samples (e.g., classification and regression), where a teacher provides the output samples (labels). In *reinforcement learning*, the mapping is between input data of the environment and corresponding actions rather than labels. These concepts will be detailed in this chapter using conventional machine learning techniques, where features are first extracted from the raw data and then fed into the algorithm to learn the task at hand. Discussion of deep learning, where the algorithm act directly on the raw data, will be presented in Chap. 4.

S. Lee (✉)
Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
e-mail: lees14@mskcc.org

I. El Naqa
Department of Machine Learning, Moffitt Cancer Center, Tampa, FL, USA

Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: ielnaqa@med.umich.edu; Issam.elnaqa@moffitt.org

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_3

3.2 Unsupervised Learning

3.2.1 Linear Principal Component Analysis

Suppose we have an oncology treatment data or images for a group of patients, some of whom developed a late complication and others who did not. For instance, in the case of prostate cancer radiotherapy, the data might include patient's age and weight, diagnostic factors such as Gleason score, dose delivered to the planning target volume (PTV), dose to one or more critical structures, etc. We would like to know if there are patterns in these data that can predict for the complication. The first step is to reduce the data set to its most informative elements, i.e., features. Often there will be more than one datum that measures more or less the same thing. We would like to reduce the data vector to a smaller dimension containing only components that are clearly distinctive (i.e., uncorrelated with one another). To do this, we arrange the patient data in a matrix X so that each row and column represent one patient and a variable, respectively. As a pre-processing step, each column in the matrix X is normalized to zero mean and unity variance (z-score). Principal component analysis (PCA) is then applied to the normalized X to identify a set of principal components (PCs) which are given by:

$$\mathrm{PC} = U^T \mathbf{X} = \Sigma V^T \tag{3.1}$$

where $U\Sigma V^{\mathrm{T}}$ is the singular value decomposition (SVD) of X. This is equivalent to transformation into a new coordinate system such that the greatest variance by any projection of the data would lie on the first coordinate (first PC), the second greatest variance on the second coordinate (second PC), and so on. For visualization purposes with the PCA, the heterogeneous variables are typically normalized using z-scoring (zero mean and unity variance). The term variance explained, used in PCA plots (Fig. 3.1), refers to the variance of the data model about the mean prognostic input factor values. The data model is formed as a linear combination of its principal components. Thus, if the PC representation of the data explains the spread (variance) of the data about the full data mean, it would be expected that this PC representation will capture enough information for modeling. Moreover, PCA analysis can provide an indication about class separability; however, it should be cautioned that PCA is an indicator and is not necessarily optimized for this purpose as supervised linear discriminant analysis, for instance [2].

Figure 3.1 shows three examples of PCA applied to patient data for three different prognostic challenges: prediction of xerostomia (dry mouth), esophagitis (esophagus inflammation), and pneumonitis (lung inflammation). The main purpose of PCA in this case is to visualize a degree of separation between patients with and without complications. For the case of xerostomia, PCA revealed several significant principal modes in the prognostic data, the first two of which accounted for only 60% of the total variance among the data components. However, the first two principal components already show a fairly clear distinction between the cases with and without xerostomia, meaning that the rest of the variance may not be relevant to the complication. In contrast, PCA reveals only

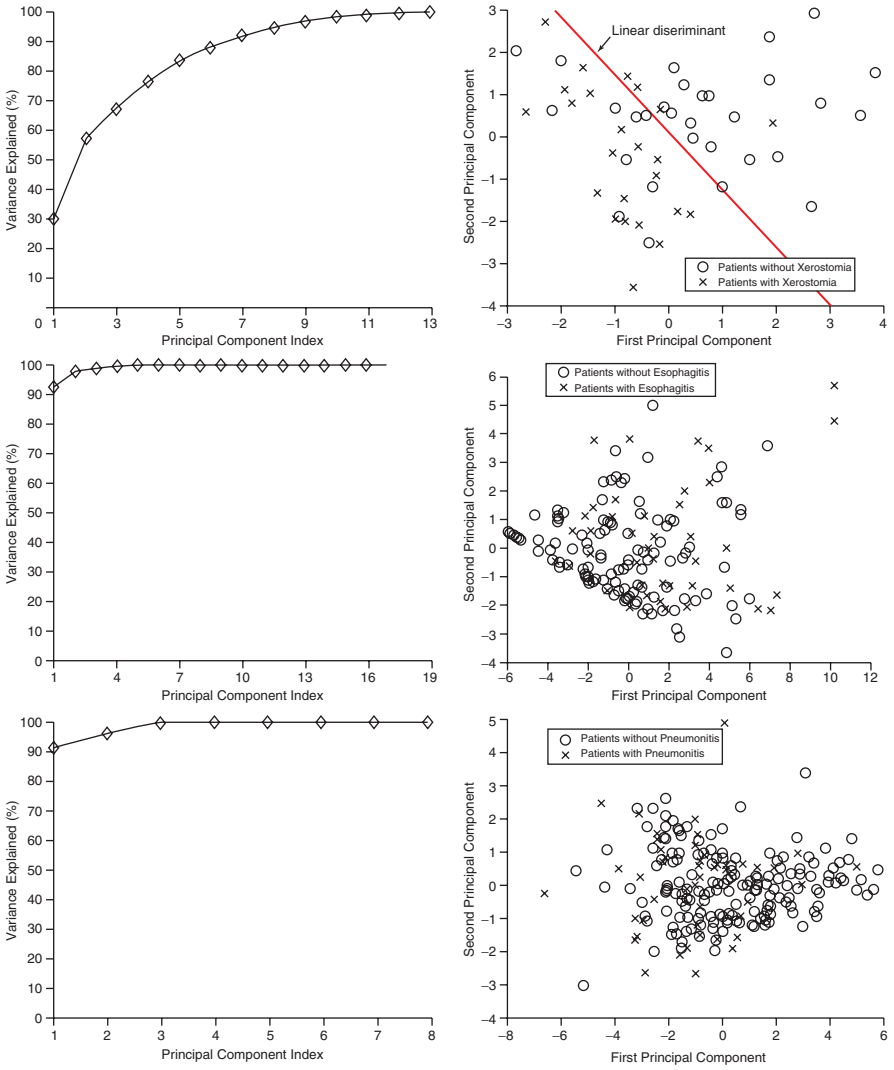


Fig. 3.1 Projection of prognostic factors for xerostomia (*top*), esophagitis (*middle*), and radiation pneumonitis (*bottom*) into a two-dimensional space consisting of the first and second principal components (*the right column*). The *left column* shows variation explanation versus principle component index. Linear separation in the xerostomia dataset is well demonstrated but not as much for the pneumonitis case (as seen from a wide class overlap). (Reproduced from El Naqa et al. [2])

one strong principal mode for esophagitis and pneumonitis, i.e., the original data components are so highly correlated that PCA reduces them to a single principal component. The projected data do not demonstrate clear separation among cases, which calls for a nonlinear modeling approach such as kernel-based methods (see Sect. 3.3.4).

3.2.2 Kernel Principal Component Analysis

Kernel PCA is a nonlinear form of the principal component analysis by use of a kernel technique (see the upcoming section on support vector machine (SVM)). It is useful for detecting nonlinear behaviors in data that cannot be represented in terms of linear combination of the existing variables. The kernel trick effectively transforms an input space into a higher-dimensional feature space in which nonlinear patterns can be made discoverable in a linear fashion (Fig. 3.2). This concept is analogous to mapping to higher dimensions in theoretical physics to identify unifying frameworks of particle and natural forces behavior (e.g., string theory). However, the input space transformation does not need to be defined explicitly, as the PCA only requires the knowledge of a covariance matrix in the transformed space. The (i,j)-th component of a covariance matrix for the data $x_1, x_2, ..., x_n$ can be computed directly from the kernel function $k(\cdot,\cdot)$:

$$K_{ij} = \left\langle \Phi\left(x_i\right), \Phi\left(x_j\right) \right\rangle = k\left(x_i, x_j\right) \tag{3.2}$$

where $\Phi(x)$ denotes the input space transformation. K is then diagonalized to extract a set of principal components (PCs) and their corresponding eigenvalues. Kernel PCA becomes more computationally expensive than linear PCA when the number of samples exceeds input dimension. Nevertheless, when applied to problems containing nonlinear patterns (e.g., handwriting), a nonlinear PCA could be more suitable than the linear one for reducing data dimension prior to a classification task [3].

3.2.3 Factor Analysis (FA)

Factor analysis (FA) aims at capturing common patterns of variance from observed variables. Similarly to PCA, factor analysis can be seen as a dimensionality reduction technique where a large number of observed variables are summarized into a much smaller number of *latent variables*, or *factors*. Both PCA and FA do so by reconstructing the variance/covariance matrix using linear combination of latent variables. However, the major difference between PCA and FA lies in the

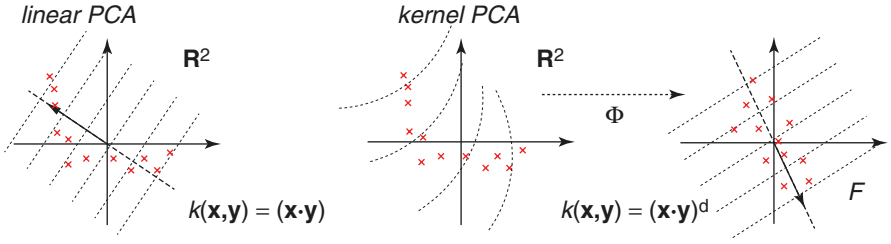


Fig. 3.2 A cartoon describing the utility of kernel PCA in linearizing a nonlinear pattern by feature transformation Φ via a polynomial kernel. The *dotted lines* are contour lines of the same value of projection to the first principal component. (Reproduced from Scholkopf et al. [3])

assumption made on the variance of variables. Factor analysis divides a variance of a variable into two components: (1) the variance that can be explained by linear combination of factors and (2) the variance unique to the respective variable and therefore cannot be explained by the factors. This can be expressed as:

$$X = WZ + \mu + \varepsilon \tag{3.3}$$

where X is a $p \times n$ matrix for 1,…,p observed variables, Z is a $l \times n$ matrix for 1,…,l factors, μ contains the means of the observed variables, and ε represents the variance unique to each variable. The $p \times l$ matrix W, called the *loading*, specifies the weighting factors for linear combination of the factors. In PCA, in its classical form, ε is assumed to be zero. In other words, it coerces the variance to be fully explained by latent variables without unique variance. Thus, PCA can be seen as a special case of FA.

There are two types of factor analysis. Exploratory factor analysis (EFA) sets the number of factors to be equal to the number of variables and allows any factors to be linked to any variables. It is performed at an exploratory stage, as the name suggests, to observe the patterns of correlations between variables and constructs. In comparison, confirmatory factor analysis (CFA) tests the validity of the existing latent structure model (constructed based on prior knowledge or the results of EFA) to observation data. Unlike EFA, CFA does not allow a variable to be associated with more than one factor. The difference between EFA and CFA is illustrated in Figure 3.3.

Factor analysis is particularly useful for identifying distinct patterns of responses from survey or questionnaire data. It has been applied to studying patient-reported radiotherapy outcomes in a form of questionnaires [4–7]. For example, Thor et al. [7] used factor analysis to identify 8 symptom domains from the Late Effects of treatment in Normal Tissue (LENT) questionnaire data from prostate patients, and identified redundant or irrelevant questions that could be removed for streamlining the toxicity reporting process.

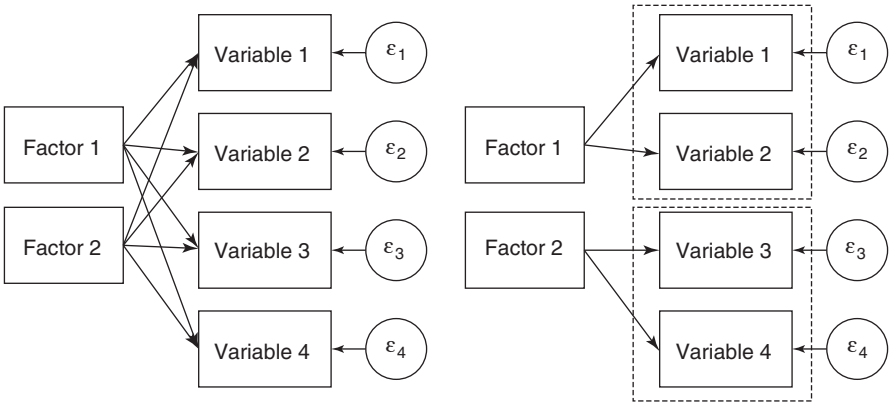


Fig. 3.3 Illustration for exploratory (left) and confirmative (right) factor analyses using 4 variables and 2 factors. Unique variance is indicated as ε

3.2.4 Clustering

Cluster analysis refers to detection of collective patterns in data based on similarity criteria. It can be performed either in a supervised or unsupervised fashion. Grouping data points into clusters is useful in several ways. First, it can provide intuitive and succinct representation of the nature of data prior to major investigation. Secondly, clustering can be applied to compressing complex data distribution into a group of vectors corresponding to cluster centroids (vector quantization).

The *K*-means clustering is one of the most popular clustering methods. It begins with randomized partitions with the given number (*K*) of clusters. The partitions are then iteratively refined by the following steps: (1) an assignment step (reassignment of the cluster membership of each data point based on a distance to cluster centroids) and (2) an update step (recalculation of cluster centroids as a geometric mean of the updated membership). The Minkowski distance between the *d*-dimensional vectors **a** and **b**, also known as a L_p norm, is used as a measure of proximity:

$$L_k\left(\mathbf{a},\mathbf{b}\right)=\left(\sum_{i=1}^{d}\left|\mathbf{a}_i-\mathbf{b}_i\right|^k\right)^{1/k} \tag{3.4}$$

The widely used Euclidean and Manhattan distance refer to the Minkowski distance at $p=2$ and $p=1$, respectively.

The *K*-means gained popularity thanks to its simplicity and fast convergence [8]. One of the drawbacks of this algorithm is its tendency to converge to local minima when initial partitions are not carefully chosen [8]. This can be partially overcome by introducing seeding heuristics such as the *K*++ means algorithm by Arthur and Vassilvitskii [9]. Furthermore, the original *K*-means requires the number of clusters (*K*) to be given *a priori*. The choice can either be made based on domain knowledge or optimized in data in a cross-validated fashion. The optimization method employs for an objective function the Bayesian information criteria (BIC) [10] or the minimum description length (MDL) [11] that penalizes the larger number of clusters.

Another clustering algorithm gaining popularity is a neural network-derived method called a self-organizing map (SOM) or a Kohonen map [12]. In a SOM, distinct patterns in input data are represented by nodes, which are typically arranged in a two-dimensional hexagonal or rectangular grid for better visualization. Each node is assigned with its location in the grid and a vector of weights on input variables. The learning algorithm begins with randomizing node weights. Then, one training example is sampled from the training set and the node at the closest distance from it (Minkowski metrics can be used) is identified as a best matching unit (BMU). The weight vectors for the BMU and the nodes in its vicinity are adjusted to decrease the distance to the training example according to the following update formula:

$$\mathbf{w}_v\left(t+1\right)=\mathbf{w}_v\left(t\right)+\alpha\left(t\right)\Lambda\left(\left|\mathbf{w}_{\mathrm{BMU}}\left(t\right)-\mathbf{w}_v\left(t\right)\right|\right)\left(\mathbf{x}_i-\mathbf{w}_v\left(t\right)\right) \tag{3.5}$$

where $\mathbf{w}_v(t)$ is a weight vector for a node v at iteration t and $\mathbf{x}_i$ is the i-th input sample. The magnitude of the update is determined by the factors that depend on the

distance from the BMU $|\mathbf{w}_{\mathrm{BMU}}(t) - \mathbf{w}_v(t)|$—and the number of iteration (t). A window function (Λ) is the highest when v = BMU and tapers off to zero as a node goes farther away from the BMU. It ensures the nodes will be topologically ordered (neighboring nodes have similar weight patterns). The learning rate, $\alpha(t)$, typically decreases with iterations to ensure convergence. After the learning is repeated through all the training samples, the nodes tend to clump toward the weights that appears in input patterns frequently (topological ordering). A SOM has been shown useful in some areas such as speech recognition, linguistics, and robot control (Fig. 3.4).

Hierarchical clustering is capable of building hierarchical layers of clusters. In contrast to the previous two algorithms, a user does not have to specify the number of clusters. Clusters can be built by setting each individual sample as a cluster and merging a pair of cluster at each iteration (agglomerative clustering). In this case, the two clusters to be merged are determined based on a user-defined similarity metric—the similarity between two clusters can be based on the most similar pair of samples (single linkage), most dissimilar pair (complete linkage), or the total variance within the merged cluster (Ward's method). The less commonly used way is to start with the entire samples in a single cluster and split one cluster into two over iterations (divisive clustering). The same principle can be applied to cluster variables, instead of samples (can be easily achieved by transposing the data matrix). The result of the hierarchical clustering is visualized in a binary tree called a dendrogram (Fig. 3.5) In this diagram, each sample is represented as a terminal node, and formation of a cluster is shown in edges. Hierarchical cluster provides visual insights on how many salient groups among samples or variables could be present. However, computational time increases quadratically in terms of sample size, which makes it slower than k-means.

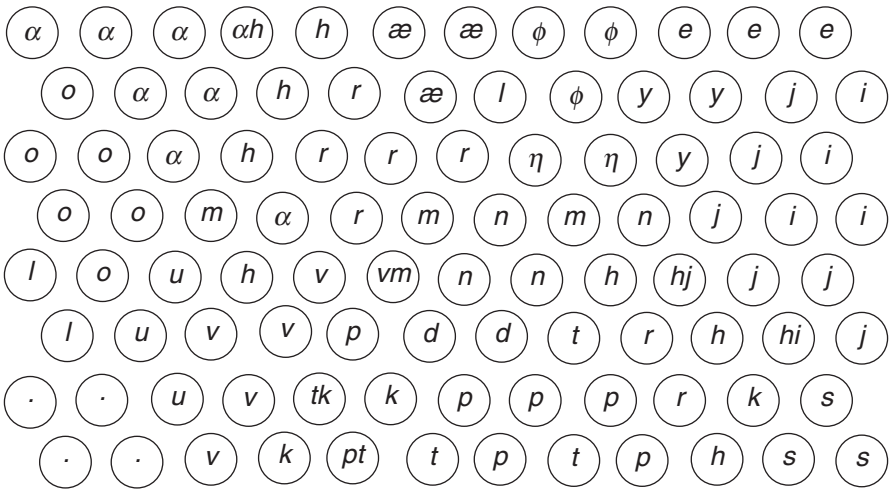


Fig. 3.4 Self-organizing map learned from natural Finnish speech analysis by Kohonen [12]. Each node represents one acoustic unit of speech called a phoneme

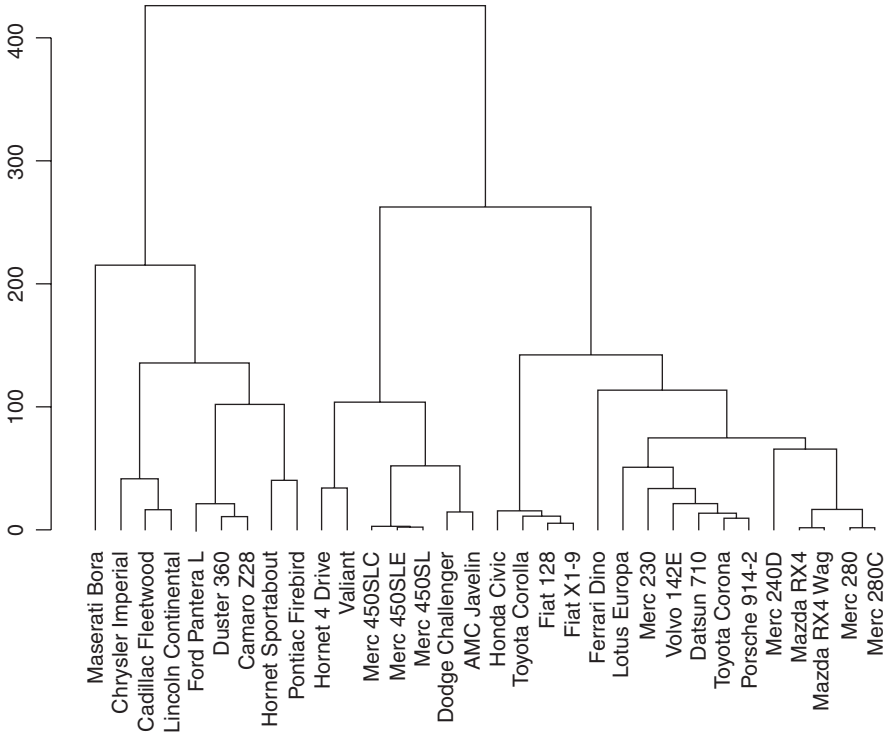


Fig. 3.5 A dendrogram for the hierarchical clustering applied to the *R* built-in dataset *mtcars*. Similarity measure is represented as the height of a branch

Another popular clustering technique is the *t*-distributed Stochastic Neighbor Embedding (t-SNE) algorithm, which allows for visualization of data in higher-dimensional space as in PCA and KPCA. However, in *t*-SNE a probability distribution is constructed where similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with corresponding probability using metrics such as Euclidean distance or more commonly the Kullback–Leibler (KL) divergence [13]. The *t*-SNE has been applied to for identifying breast lesions from ultrasound/MRI images [14].

Many challenges in bioinformatics are framed as a clustering problem, such as identifying a group of genes showing similar patterns of expression under certain conditions or diseases. A work by Svensson et al. [15] is a good example from radiotherapy toxicity modeling. They grouped 1182 candidate late toxicity marker genes into two groups using their expression patterns in lymphocytes after radiation, although the grouping did not correlate with toxicity status. In contrast, a SOM of radiation pneumonitis risk factors built by Chen et al. [16] showed that grouping patterns among the factors can be exploited for predicting the toxicity with decent accuracy (AUC = 0.73). Hierarchical clustering has been adopted in many radiomic studies to identify a group of correlated or redundant radiomic features [17–19].

3.3 Supervised Learning

3.3.1 Logistic Regression

In treatment outcomes modeling, the response will usually follow an S-shaped curve. This suggests that models with sigmoidal shape are more appropriate to use [20–27]. A commonly used sigmoidal form is the logistic model, which also has nice numerical stability properties. The logistic model is given by [28, 29]:

$$f\left(\mathbf{x}_j\right) = \frac{e^{g\left(x_j\right)}}{1+e^{g\left(x_j\right)}}, \quad i = 1,2,\ldots,n \tag{3.6}$$

where n is the number of cases (patients) and $\mathbf{x}$ is a vector of the input variable values used to predict $f(\mathbf{x}_i)$ for the outcome y_i of the i_{th} patient. The $f(\cdot)$ is referred to as the logic transformation. The "x-axis" summation $g(\mathbf{x}_i)$ is given by:

$$g\left(\mathbf{x}_i\right) = \beta_0 + \sum_{j=1}^{s} \beta_j x_{ij}, \quad i = 1,\ldots,n, \quad j = 1,\ldots,s \tag{3.7}$$

where s is the number of model variables and the β's are the set of model coefficients that are determined by maximizing the probability that the data gave rise to the observations (i.e., the likelihood function). Many commercially available software packages, such as SAS, SPSS, and Stata, provide estimates of the logistic regression model coefficients and their statistical significance. The results of this type of approach are not expressed in closed form as above, but instead, the model parameters are chosen in a stepwise fashion to define the abscissa of a regression model as shown in Fig. 3.6. However, it is the analyst's responsibility to test for interaction effects on the estimated response, which can potentially be corrected by adding cross terms to Eq. (3.6). However, this transformation suffers from limited

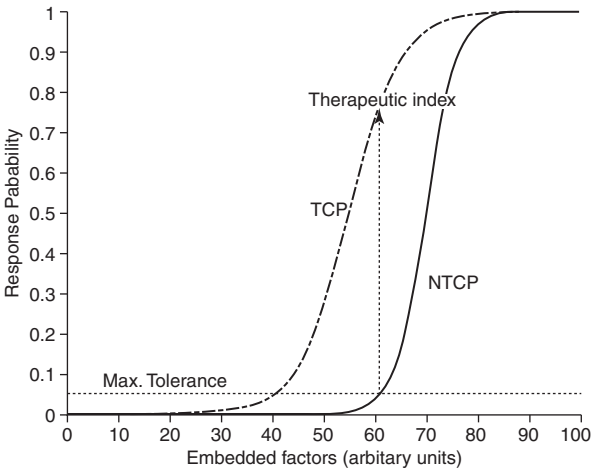


Fig. 3.6 Sigmoidally shaped response curves (for tumor control probability of normal tissue complication probability) are constructed as a function of a linear weighting of various factors, for a given dose distribution, which may include multiple dose-volume metrics as well as clinical factors. The units of the x-axis may be thought of as equivalent dose units. (Reproduced from El Naqa et al. [30])

learning capacity. In such a model, it is the user's responsibility to determine whether interaction terms or higher order terms should be added. A solution to ameliorate this problem is offered by applying artificial intelligence methods.

3.3.2 Feed-Forward Neural Networks (FFNN)

Neural networks are described as adaptive massively parallel-distributed computational models that consist of many nonlinear elements arranged in patterns similar to a simplistic biological neuron network. A typical neural network architecture is shown in Fig. 3.7.

Neural networks have been applied successfully to model many different types of complicated nonlinear processes, including many pattern recognition problems [31]. A three-layer FFNN network would have the following model for the approximated functional:

$$f\left(\mathbf{x}\right) = \mathbf{y}^T \mathbf{w}^{(2)} + b^{(2)} \tag{3.8}$$

where $\mathbf{v}$ is a vector, the elements of which are the output of the hidden neurons, i.e.,

$$v = s\left(\mathbf{x}^T \mathbf{w}_i^{(1)} + b^{(1)}\right) \tag{3.9}$$

where $\mathbf{x}$ is the input vector and $\mathbf{w}^{(j)}$ and $\mathbf{b}^{(j)}$ are the interconnect weight vector and the bias of layer j, respectively, $j = 1,2$. In the FFNN, the activation function $s(\cdot)$ is usually a sigmoid, but radial basis functions were also used [32]. The FFNN could be trained in two ways: batch mode or sequential mode. In the batch mode, all the training examples are used at once; in sequential mode, the training

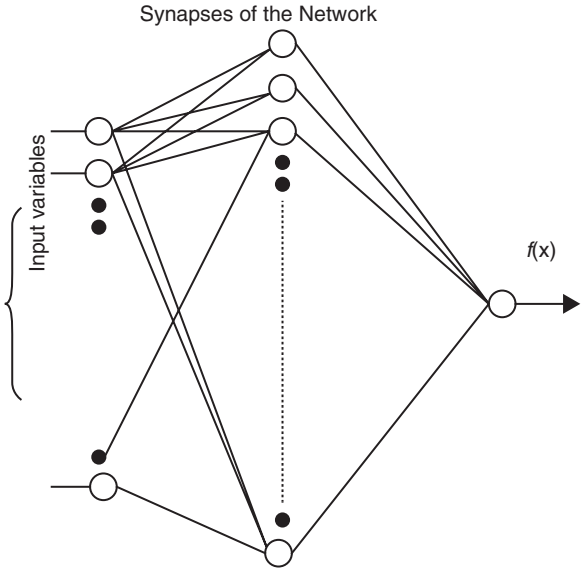


Fig. 3.7 Neural network architecture consisting of an input layer, middle (hidden) layer(s), and an output layer. The synapses of the network consist of neurons that fire depending on their chosen activation functions

examples are presented on a pattern basis, in the order that is randomized from one epoch (cycle) to another. The number of neurons is a user-defined parameter that determines the complexity of the network; the larger the number of neurons, the more complex the network would be. The number is determined during the training phase.

3.3.3 General Regression Neural Networks (GRNN)

The GRNN [33] is a probabilistic regression model based on neural network architecture. It is characterized as non-parametric, which means that it does not require any pre-determined functional form (e.g., polynomials). Instead, it estimates the joint density of input variables **x** and a target y from training data. The regression output using the GRNN is obtained by taking the expectation value of y for a given observation **X** and the joint density $g(\mathbf{x}, y)$:

$$\hat{y}(\mathbf{X}) = E(y|\mathbf{X}) = \frac{\int_{-\infty}^{\infty} y g(\mathbf{X}, y)\, dy}{\int_{-\infty}^{\infty} g(\mathbf{X}, y)\, dy} \tag{3.10}$$

The joint density $g(\mathbf{x}, y)$ is estimated from training examples $\mathbf{X_i}$ and y_i via the Parzen estimator where the density is regarded as the superposition of Gaussian kernels centered at the observation points with a spread σ. The resulting form of the regression function is:

$$\hat{y}(\mathrm{X}) = \frac{\sum_{i=1}^{n} y_i \exp\left(\frac{-D_i^2}{2s^2}\right)}{\sum_{i=1}^{n} \exp\left(\frac{-D_i^2}{2s^2}\right)} \tag{3.11}$$

where $D_i^2 = (\mathbf{X} - \mathbf{X}_i)^{\mathrm{T}} (\mathbf{X} - \mathbf{X}_i)$, denoting the Euclidean distance between the testing data **X** and the i-th training data **Xi**.

The GRNN is fairly simple to train, with only the Gaussian width σ to be tuned. Thus, implementation of the GRNN does not require an optimization solver to obtain the weights, as in the case of FFNN. However, the output is obtained as a weighted sum of all the training samples, which could make it less efficient during running time. This could be improved by performing cluster analysis on training data (see Sect. 3.2.3) to compress it into a few cluster centers so that the metric D_i can be computed only between those center points and a testing example. The computational speed can also benefit from parallelized neural network implementation since each summation can be performed independently using synapses and an exponential activation function. In our previous work, we demonstrated that GRNN can outperform traditional FFNN in radiotherapy outcomes prediction [34].

3.3.4 Kernel-Based Methods

Kernel-based methods and its most prominent member, support vector machines (SVMs), are universal constructive learning procedures based on the statistical learning theory [35]. For discrimination between patients who are at low risk versus patients who are at high risk of radiation therapy, the main idea of the kernel-based technique would be to separate these two classes with hyper-planes that maximizes the margin between them in the nonlinear feature space defined by implicit kernel mapping. The optimization problem is formulated as minimizing the following cost function:

$$L\left(\mathbf{x},\ \xi\right)=\frac{1}{2}\mathbf{w}^{T}\mathbf{w}+C\sum_{\mathrm{i}=1}^{n}\xi_{i} \tag{3.12}$$

subject to the constraints:

$$y_{i}\left(\mathbf{W}^{\mathrm{T}}\Phi\left(\mathbf{X}_{i}\right)+b\right)\geq 1-\zeta_{i}$$

$$3\zeta_{i}\geq 0\quad \text{for all } i$$

where **w** is a weighting vector and $\Phi(\cdot)$ is a nonlinear mapping function. The ζ_i represents the tolerance error allowed for each sample being on the wrong side of the margin. Note that minimization of the first term in Eq. (3.12) increases the separation (improves generalizability) between the two classes, whereas minimization of the second term (penalty term) improves fitting accuracy. The trade-off between complexity and fitting error is controlled by the regularization parameter *C*. However, such nonlinear formulation would suffer from the curse of dimensionality (i.e., the dimension of the problem becomes too large to solve) [1, 36]. However, computational efficiency is achieved from solving the dual optimization problem instead of the equation which is convex with a complexity that is dependent only on the number of samples [35]. The prediction function in this case is characterized only by a subset of the training data known as support vectors s_i :

$$C^{+}\sum_{i\in Z^{+}}\xi_{i}+C^{-}\sum_{i\in Z^{-}}\xi_{i} \tag{3.13}$$

where n_s is the number of support vectors, α *s* are the dual coefficients determined by quadratic programming, and $K(\cdot,\cdot)$ is the kernel function as discussed next. Typically, used nonlinear kernels include:

$$\text{Polynomials}: K\left(\mathrm{x},\mathrm{x}'\right)=\left(\mathrm{x}^{T}\mathrm{x}'+c\right)$$

$$\text{Radial basis function}\ \left(\mathrm{RBF}\right): K\left(x,x'\right)=\exp\left(\frac{\|x-x'\|}{2\sigma^{2}}\right)$$

where *c* is a constant, *q* is the order of the polynomial, and σ is the width of the radial basis functions. The kernel-based approach is very flexible, which allows for constructing a neural network by using combination of sigmoidal kernels or chooses

a logistic regression equivalent kernel by replacing the hinge loss with a binomial deviance [1].

SVM has been widely used for many radiotherapy outcome prediction cases where complex relationships between risk factors are expected. Examples include lung cancer prognosis [37–39], radiation pneumonitis [40, 41], and GI/genitourinary toxicity [42].

3.3.5 Decision Trees and Random Forests

A decision tree is suitable for generating the hypotheses that consist of multiple Boolean conditions on attributes (disjunctive hypotheses). Although it can also perform regression, we will limit the discussion to its application to classification. A decision tree divides an input space into several disjoint subregions. A testing instance falls into one of the subregions after successive tests on its attribute values. Then, the instance is given for its classification result the value that is assigned to the subregion. The tests are organized in the order specified by a tree structure (Fig. 3.8). A tree consists of nodes, branches, and leaves, each representing the following:

- Node: the attribute to be tested
- Branch: the outcome of the test, for example, is the body temperature of a patient higher than 37° (continuous attribute) or is the patient taking aspirin (categorical attribute)?
- Leaf node: the node located at the terminus of a tree representing a subset of data and a class label assigned to the subset

The tree and its parameters are learned from training data in a supervised fashion. The learning process can be thought of as dividing training instances into subgroups (corresponding to nodes) in a way that class labels in the subgroups are

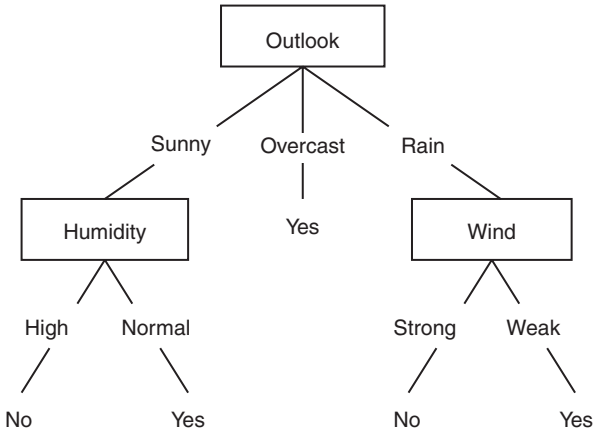


Fig. 3.8 An example decision tree that classifies whether to go play tennis or not (written in *bold*) based on three attributes (outlook, humidity, wind) shown in *box nodes*. Values of the three attributes are written on the corresponding branches (Reproduced from Mitchell [43])

made as homogeneous as possible. The major questions in decision tree learning are (1) in which order the attributes be tested, (2) what level of purity the partition class labels is desired as a result of a single test, and (3) how many nodes are needed.

The ID3 (iterative dichotomizer 3) algorithm is a primitive form of the decision tree learning algorithm that aims to arrive at an optimal decision tree via a greedy search [44]. The ID3 algorithm is initiated by identifying the first attribute (root node) to create the first set of partitions, and the tree is further branched by applying the same procedure to the resulting subsets and the remaining attributes. At each round of partitioning, the attribute to split is chosen based on how well it can predict a target class by itself. In the context of decision tree learning, the predictive value of an attribute A with respect to a class C is measured by its information gain, which is defined as:

$$\text{gain}(\text{A}) = H(A) - H(A|C) \tag{3.14}$$

where H, entropy, is a measure of information conveyed by a probability distribution. For a variable A with the distribution of c discrete states and corresponding probabilities $p_1, p_2, ..., p_c$, the entropy is:

$$H(A) = \sum_{i=1}^{c} - p_i \log_2 p_i. \tag{3.15}$$

In the case of continuous attributes, a threshold (A_{th}) is set to split the data into two subsets with proportions p_1 and p_2 where $p_1 = p(A < A_{\text{th}})$ and $p_2 = p(A > A_{\text{th}})$. The value of A_{th} is chosen so that the resulting information gain is the largest.

A branch of the tree stops growing when all the attributes have been used or all the partitions of the branch are purified to one class. However, when no regulatory measures are taken, a tree can easily overfit the data by adding more branches until every training instance is correctly classified. A number of preventive methods have been proposed to improve the generalizability of a tree. Reduced-error pruning [45] reduces the size of a tree after it was learned by applying iterative pruning to branches. The branches closer to leaves are removed first and the pruning propagates upstream until the validation performance of the pruned tree begins to decrease.

Overfitting can also be alleviated by a meta-algorithm called *ensemble learning*. The idea is to train a group of classifiers with a given dataset and combine their output in order to compensate for the high variance of an individual model. Breiman [1] [46] applied this concept to tree learning, which is dubbed as the *random forest*. In creating a bag of models, the random forest algorithm introduces two levels of randomization: First, it randomizes training samples by resampling with replacements (bootstrapping). Second, at each branching step it chooses an attribute to split among a randomly selected subset of attributes. After a bag of trees is trained, prediction is made for all the individual trees and the most frequent class selected by the trees is taken as a final result. *Boosting* is another ensemble meta-algorithm that is often used in conjunction with decision tree. In this setting, trees are learned sequentially in the following way: after a tree is learned, the incorrectly classified

training examples are assigned with larger weights and the subsequent tree is learned with the reweighed training set. The final classification result is taken as an average output of the group of trees. Detailed algorithm can be consulted in a paper by Freund and Schapire [47].

Decision trees have been a popular choice for many decision support systems, especially in the field of medicine, because their representation of hypotheses as sequential "if-then" clauses is easy to interpret and somewhat resembles human reasoning. For example, Delaney et al. [48] conducted a literature survey to construct a tree to determine recommendation for radiotherapy to melanoma patients based on several clinical attributes (Fig. 3.9). Das et al. [49] trained an ensemble of trees that combined dosimetric and non-dosimetric risk factors for radiation pneumonitis and showed that the prediction can be improved by combining a larger number of trees. A drawback of using an ensemble of trees is loss of interpretability: Valdes et al. [50] addressed this problem in their algorithm MediBoost where a decision tree is trained in a boosted fashion by introducing "soft" partitioning. In other studies, variable importance measures from random forest were utilized to interpret the predictive models for radiotherapy toxicity [51, 52].

3.3.6 Bayesian Network

Bayesian belief network, or Bayesian network, is designed to model probabilistic relationships among a set of random variables. A key feature of Bayesian network is graphical representation of the relationships via a directed acyclic graph (DAG) which encodes the presence and direction of influence between variables. In a DAG, each variable is assigned to a node and connected to each other via an edge (vertex) which originates from a variable (parent) that influences the probability of the variable it is connected to (child). Thus, probability of a random variable is set to be conditional upon its parent variable(s). The connectivity information in a DAG derives conditional independence relationships that can be stated as random variables X and Y are conditionally independent given another variable set $Z_1, Z_2, ..., Z_n$ if and only if:

$$P\left(X \mid Y, Z_1, Z_2, \ldots, Z_n\right) = P\left(X \mid Z_1, Z_2, \ldots, Z_n\right) \tag{3.16}$$

A set of conditional independence relationships specified in a DAG greatly simplifies computation of probability distributions by use of this convenient property: joint probability distribution between the entire variable set, $X = X_1, X_2, , X_n$, can be obtained by taking the product of all the conditional probabilities for each parents-child set (*the chain rule for Bayesian networks* [53]). Figure 3.10 demonstrates a network of local control of non-small-cell lung cancer (LC) in relation to the following clinical and dosimetric variables: age (A), GTV volume (G), PTV coverage ($V75$, $V60$), and pre-treatment chemo (P) [54]. Using the chain rule, a joint probability can be factorized into:

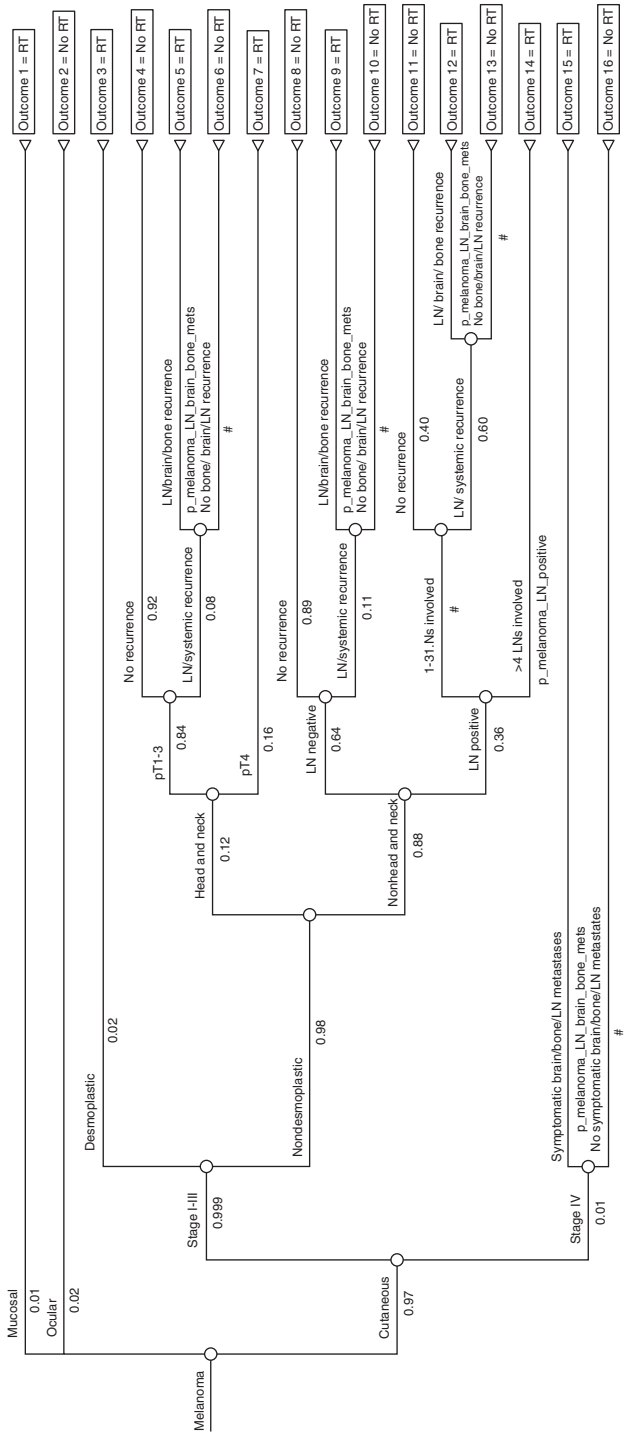


Fig. 3.9 A tree representing decision rules that determine whether radiotherapy should be used for melanoma patients based on characteristics of the disease. *RT* radiotherapy, *LN* lymph node. (Reproduced from Delaney et al. [48])

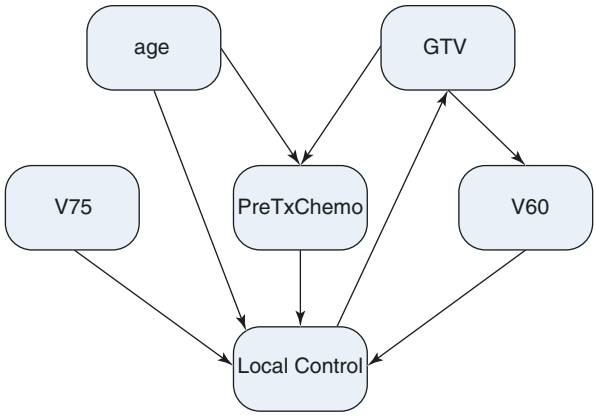


Fig. 3.10 A Bayesian network DAG for predicting local control of NSCLC using radiotherapy and clinical variables. The DAG was trained from clinical data by Oh et al. [54]

$$P(LC, A, G, V75, V60, C) = P(A)P(G)P(V75)P(C \mid A, G)P(V60 \mid G)$$
$$P(LC \mid A, G, C, V75, V60)$$

Conditional probability values are often referred to as the "parameters" of Bayesian network. The parameters can be trained from data as a maximum likelihood estimate or maximum a posteriori (MAP) which incorporates a prior probability with the likelihood obtained from observations.

A DAG can be constructed using prior knowledge on the study domain. When the domain knowledge is not sufficient, observational data can be used to search for the DAG that can best describe the data. DAG searching can be solved as an optimization problem where a predefined scoring function is maximized over a space of possible DAG configurations. Searching algorithms can vary according to a choice of the scoring function and searching procedures. Widely used scoring functions include a marginal likelihood (Bayesian) score and a Bayesian information criteria (BIC) score. Both scores aim at achieving a balance between the fitness to data (an edge is more likely to be formed between the variables with stronger correlation in data) and complexity of a graph (quantified by the number of edges or parameters), although difference exists in a degree to which complexity is penalized. Mathematical details can be consulted in a primer by Koller and Friedman [53].

Since the number of possible DAGs grows super-exponentially with the number of variables, it is impractical to search exhaustively over the entire graph space for the highest-scoring DAG. Various heuristic approaches have been suggested to reduce a computational cost. For example, a greedy search algorithm begins with the empty graph and keeps adding on edges only when it leads to a higher graph score. Also, constraints on graph topology can be imposed to the search algorithm in order to confine a search domain. For example, the search can be restricted to treelike structures (Chow-Liu trees) [55] or a certain variable ordering that permits only the edges between the variables in descending order ($K2$ algorithm) [56]. High-scoring DAGs can be discovered by a sampling method such as the Markov Chain Monte Carlo (MCMC) [57]. The MCMC algorithm generates samples of

DAGs encountered during a random walk over the graph space (Markov chain), which can be approximated as a posterior distribution of DAGs upon convergence of a chain.

The probabilistic approach of BN makes it suitable for handling uncertainties. Especially in a medical domain, missing records or test results could have a negative impact on prediction performance. Bayesian network does not require the full observation on its features for prediction, as it is capable of building and marginalizing joint probability using the conditional dependence relationships between the features. This advantage, in comparison to non-probabilistic classifiers such as SVM, was shown in survival prediction of lung cancer patients by Jayasurya et al. [38]. Other applications of the BN in radiation oncology include a prognostic network for prostate cancer [58] and lung cancer [54].

3.3.7 Naive Bayes

Naive Bayes is a simplified derivative of Bayesian network that is used solely for classification. This method makes an assumption that feature variables are considered independent given a class variable. This so-called naive independence assumption can be graphically represented by the Bayesian DAG as shown in Fig. 3.11. Inference of the most probable state for a class, C_{MAP}, is derived from the maximum a posteriori (MAP) rule, using the independence assumption:

$$
\begin{aligned}
C_{\text{MAP}} &= \arg\max_{C} P\left(C \mid A_1, A_2, \ldots, A_n\right): \\
&= \arg\max_{C} \frac{P\left(A_1, A_2, \ldots, A_N \mid C\right) P(C)}{P\left(A_1, A_2, \ldots, A_n\right)} \\
&= \arg\max_{C} \frac{P(C) \prod_{i=1}^{N} P\left(A_i \mid C\right)}{P\left(A_1, A_2, \ldots, A_n\right)} \\
&\propto \arg\max_{C} P(C) \prod_{i=1}^{N} P\left(A_i \mid C\right)
\end{aligned}
$$

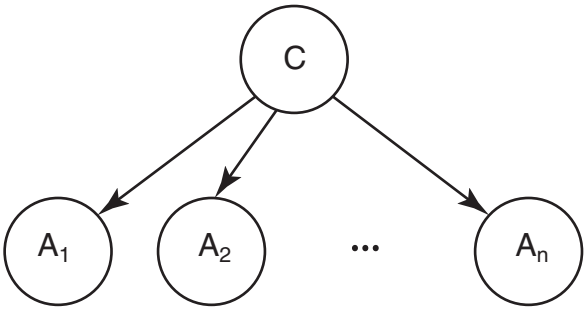


Fig. 3.11 Directed graph representation of the naive Bayes model for a class C and features A_1, A_2, A_n

Naive Bayes is effective for classification in a high-dimensional space where estimating joint probability of a full variable set is challenging. Its theoretical property is shown to be less sensitive to noisy variation in input, which contributes to its robust performance [59]. However, naive Bayes is not suitable for direct estimation of class posterior as the unrealistic independence assumption results in inaccurate probability estimate. Nevertheless, it has been applied to many medical prognostic problems where classification of a disease state is the only interest. For example, Kazmierska and Malicki predicted brain tumor relapse from a set of 96 features with naive Bayes which accuracy surpassed Bayesian network and decision tree algorithms [60].

3.4 Reinforcement Learning

Reinforcement learning (RL) is a class of machine learning algorithms in which a learner or software agent attempts to take a sequence of actions based on the underlying environment status that would maximize a cumulative reward such as winning a game of checker or chess, for instance [61]. To an extent RL mimics the way human learns by combining the fields of Markov decision processes (MDP) (e.g., dynamic programming) with supervised learning. An RL could be depicted as shown in Fig. 3.12 [62], in which at any time point (t) actions (a_t) taken by the agent lead to rewards (r_{t+1}) from the current environment state (s_t). The objective is to maximize expected discounted returns value (V) at particular state to a given policy (π):

$$V^{\pi}\left(s_t\right) = E\left\{R / s_t, \pi\right\} \tag{3.16}$$

where R is the return function $R = E\left\{\gamma r_{t+1}\right\} = \sum_{t=0}^{\infty} \gamma^t r_{t+1}$ and $0 \le \gamma \le 1$ are discounted return rates.

A known approach to solving such a discounted infinite horizon MDP is Q-learning stochastic algorithm [63] which is an iterative approach to solving the Bellman optimality of the action-selection problem using model-free

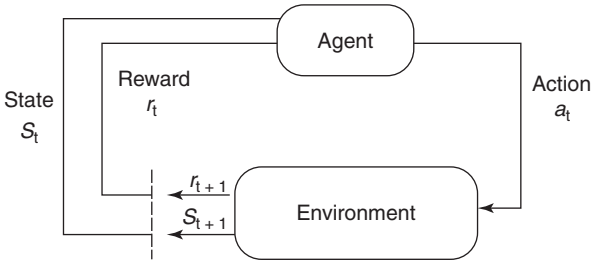


Fig. 3.12 Reinforcement learning system

(unsupervised) or model-based (supervised) methods. Techniques based on RL have been adopted in the design of adaptive clinical trials to estimate individual treatment rules [64]. For instance, the sequential multiple assignment randomized trial (SMART) has been applied for adaptive interventions in different diseases related to drug abuse, HIV/AIDS, and mental illness with promising results [65]. Examples of applying RL to radiotherapy is presented by Kim et al. [66], where they showed numerical examples of modifying dose fractionation schedules using a Markov decision process for adaptive radiotherapy applications. Another example is presented by Vincent et al. to optimize the dose per fraction using different utility functions in cell culture experiments [67].

3.4.1 Reinforcement Learning for Adaptive Liver Cancer Treatment

Adaptive radiotherapy was applied to the population of 88 liver stereotactic body radiation therapy (SBRT) patients with 35 on non-adaptive and 53 on adaptive protocols from a population of 145 patients [68]. Adaptation was based on liver function in a split course of 3 + 2 fractions with a month break. Plasma biomarkers were analyzed before and during radiotherapy. Normal tissue complication probability (NTCP) was assessed as a one-grade change in ALBI toxicity score. The radiotherapy environment was modeled as a 2-stage MDP: baseline and one month into radiotherapy states. States were represented by the patient's clinical, dosimetric, and biological covariates. Two decision-making scenarios at stage-2 were considered for evaluating RL: (1) adapting with a split course or not, and (2) delivering an additional 2 fractions after the initial 3 fractions course. The reward/regret was defined by the complication-free tumor control (P+) as a function tumor control probability (TCP) and NTCP: of [P+=TCP × (1 − NTCP)]. Q-learning with a simple regression of state-action mapping was used for strategy optimization. The performance was evaluated using an adjusted *R*-squared (aR^2) to correct for overfitting. Using a state of clinical and dosimetric (tumor size, tumor dose, mean liver dose) covariates, Q-learning at one month (stage-2) selected split-course adaptation as an optimal action with an aR^2=0.65 ($p < 0.001$). Percentage change in the cytokine TGF-β1 concentration was the only biological variable to correlate with outcomes (ALBI score, $p = 0.03$). Its addition improved the fit to aR^2=0.74. In the case of 3 versus 5 fractions determination, the delivery of 2 extra fractions provided better action with an aR^2=0.66 ($p < 0.001$) and 70.5% of the patients benefiting. The addition of TGF-β1 improved the fit to aR^2=0.74, and the percentage of patients benefiting from current clinical adaptation was found to be 65.5% as shown in Fig. 3.13.

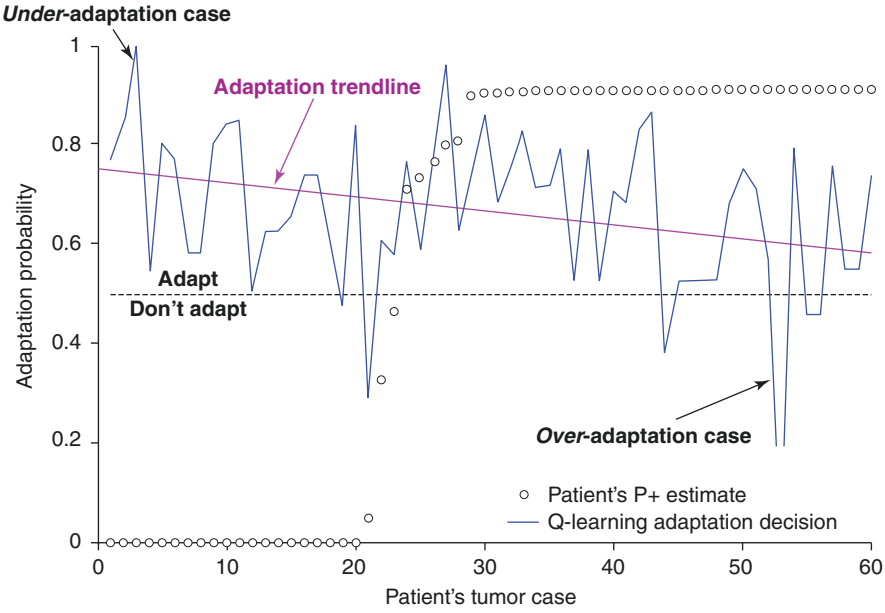


Fig. 3.13 SBRT hepatocellular cancer adaptation analysis by using a simple P+ utility. Adaptation decision estimates are based on Q-learning with a threshold of 0.5. Note that the algorithm is more optimistic about adaptation due to overestimated risks by conventional TCP/NTCP modeling

References

1. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.
2. El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. Phys Med Biol. 2009;54(18):S9.
3. Scholkopf A, Smola J, Muller KR. Kernel principal component analysis. Cambridge: MIT Press; 1999. p. 327–52.
4. Farnell DJ, Mandall P, Anandadas C, et al. Development of a patient-reported questionnaire for collecting toxicity data following prostate brachytherapy. Radiother Oncol. 2010;97(1):136–42.
5. Kuku S, Fragkos C, McCormack M, Forbes A. Radiation-induced bowel injury: the impact of radiotherapy on survivorship after treatment for gynaecological cancers. Br J Cancer. 2013;109(6):1504–12.
6. Xiao C, Hanlon A, Zhang Q, et al. Symptom clusters in patients with head and neck cancer receiving concurrent chemoradiotherapy. Oral Oncol. 2013;49(4):360–6.
7. Thor M, Olsson C, Oh JH, et al. Urinary bladder dose-response relationships for patient-reported genitourinary morbidity domains following prostate cancer radiotherapy. Radiother Oncol. 2016;119(1):117–22.
8. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31(3):264–323.
9. Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, SODA'07. Philadelphia: Society for Industrial and Applied Mathematics; 2007. p. 1027–35.

10. Pelleg D, Moore A. X-means: extending k-means with efficient estimation of the number of clusters. In: Proceedings of the 17th international conference on machine learning. San Francisco: Morgan Kaufmann; 2000. p. 727–34.
11. Bischof H, Leonardis A, Selb A. Minimum description length principle for robust vector quantisation. Pattern Anal Appl. 1999;2(1):59–72.
12. Kohonen T. The self-organizing map. Proc IEEE. 1990;78(9):1464–80.
13. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
14. Jamieson AR, Giger ML, Drukker K, Li H, Yuan Y, Bhooshan N. Exploring nonlinear feature space dimension reduction and data representation in breast Cadx with Laplacian eigenmaps and t-SNE. Med Phys. 2010;37(1):339–51.
15. Svensson JP, Stalpers LJA, Lange REEE, Franken NAP, Haveman J, Klein B, Turesson I, Vrieling H, Giphart-Gassler M. Analysis of gene expression using gene sets discriminates cancer patients with and without late radiation toxicity. PLoS Med. 2006;3(10):e422.
16. Chen S, Zhou S, Yin FF, Marks LB, Das SK. Using patient data similarities to predict radiation pneumonitis via a self-organizing map. Phys Med Biol. 2008;53(1):203.
17. Choi W, Oh JH, Riyahi S, et al. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. Med Phys. 2018;45(4):1537–49.
18. Parmar C, Leijenaar RT, Grossmann P, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Sci Rep. 2015;5:11044.
19. Hunter LA, Krafft S, Stingo F, et al. High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images. Med Phys. 2013;40(12):121916.
20. Blanco AI, Chao KSC, El Naqa I, Franklin GE, Zakarian K, Vicic M, Deasy JO. Dose-volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy. Int J Radiat Oncol Biol Phys. 2005;62(4):1055–69.
21. Bradley J, Deasy JO, Bentzen S, El-Naqa I. Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma. Int J Radiat Oncol Biol Phys. 2004;58(4):1106–13.
22. Bradley JD, Hope A, El Naqa I, Apte A, Lindsay PE, Bosch W, Matthews J, Sause W, Graham MV, Deasy JO. A nomogram to predict radiation pneumonitis, derived from a combined analysis of rtog 9311 and institutional data. Int J Radiat Oncol Biol Phys. 2007;69(4):985–92.
23. Hope AJ, Lindsay PE, Naqa IE, Alaly JR, Vicic M, Bradley JD, Deasy JO. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. Int J Radiat Oncol Biol Phys. 2006;65(1):112–24.
24. Huang EX, Bradley JD, El Naqa I, Hope AJ, Lindsay PE, Bosch WR, Matthews JW, Sause WT, Graham MV, Deasy JO. Modeling the risk of radiation-induced acute esophagitis for combined Washington University and rtog trial 93-11 lung cancer patients. Int J Radiat Oncol Biol Phys. 2012;82(5):1674–9.
25. Huang EX, Hope AJ, Lindsay PE, Trovo M, El Naqa I, Deasy JO, Bradley JD. Heart irradiation as a risk factor for radiation pneumonitis. Acta Oncol. 2011;50(1):51–60.
26. Marks LB. Dosimetric predictors of radiation-induced lung injury. Int J Radiat Oncol Biol Phys. 2002;54(2):313–6.
27. Tucker SL, Cheung R, Dong L, Liu HH, Thames HD, Huang EH, Kuban D, Mohan R. Dose-volume response analyses of late rectal bleeding after radiotherapy for prostate cancer. Int J Radiat Oncol Biol Phys. 2004;59(2):353–65.
28. Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley; 2000.
29. Vittinghoff E, Glidden D, Shiboski S, McCulloch C. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. New York: Springer; 2006.
30. El Naqa I, Bradley J, Blanco AI, Lindsay PE, Vicic M, Hope A, Deasy JO. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. Int J Radiat Oncol Biol Phys. 2006;64(4):1275–86.
31. Ripley BD. Pattern recognition and neural networks. Cambridge/New York: Cambridge University Press; 1996.
32. Su M, Miften M, Whiddon C, Sun X, Light K, Marks L. An artificial neural network for predicting the incidence of radiation pneumonitis. Med Phys. 2005;32(2):318–25.

33. Specht DF. A general regression neural network. IEEE Trans Neural Netw. 1991;2(6):568–76.
34. El Naqa I, Bradley J, Deasy J. Machine learning methods for radiobiological outcome modeling. In: Mehta M, Paliwal B, Bentzen S, editors. Physical, chemical, and biological targeting in radiation oncology. Madison: Medical Physics; 2005. p. 150–9.
35. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
36. Haykin S. Neural networks: a comprehensive foundation. Upper Saddle River: Prentice Hall PTR; 1998.
37. Dehing-Oberije C, Yu S, Ruysscher DD, Meersschout S, Beek KV, Lievens Y, Meerbeeck JV, Neve WD, Rao B, van der Weide H, Lambin P. Development and external validation of prognostic model for 2-year survival of non small cell lung cancer patients treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys. 2009;74(2):355–62.
38. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, De Neve W, Lievens Y, Lambin P, Dekker ALAJ. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys. 2010;37(4):1401–7.
39. Klement R, Allgauer M, Appold S, Dieckmann K, Ernst I, Ganswindt U, Holy R, Nestle U, Nevinny-Stickel M, Semrau S, Sterzing F, Wittig A, Andratschke N, Guckenberger M. Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer. Int J Radiat Oncol Biol Phys. 2014;88(3):732–8.
40. Chen S, Zhou S, Yin F-F, Marks LB, Das SK. Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. Med Phys. 2007;34(10):3808–14.
41. Spencer SJ, Bonnin DA, Deasy JO, Bradley JD, El Naqa I. Bioinformatics methods for learning radiation-induced lung inflammation from heterogeneous retrospective and prospective data. J Biomed Biotechnol. 2009;2009:892863. https://doi.org/10.1155/2009/892863.
42. Pella A, Cambria R, Riboldi M, Jereczek-Fossa BA, Fodor C, Zerini D, Torshabi AE, Cattani F, Garibaldi C, Pedroli G, Baroni G, Orecchia R. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. Med Phys. 2011;38(6):2859–67.
43. Mitchell TM. Machine learning. 1st ed. New York: McGraw-Hill; 1997.
44. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1:81–106.
45. Quinlan JR. Simplifying decision trees. Int J Man Mach Stud. 1987;27(3):221–34.
46. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
47. Freund Y, Schapire RE. A brief introduction to boosting. In: Proceedings of the sixteenth international joint conference on artificial intelligence. San Francisco: Morgan Kaufmann; 1999. p. 1401–6.
48. Delaney G, Barton M, Jacob S. Estimation of an optimal radiotherapy utilization rate for melanoma. Cancer. 2004;100(6):1293–301.
49. Das SK, Zhou S, Zhang J, Yin F-F, Dewhirst MW, Marks LB. Predicting lung radiotherapy-induced pneumonitis using a model combining parametric Lyman probit with nonparametric decision trees. Int J Radiat Oncol Biol Phys. 2007;68(4):1212–21.
50. Valdes G, Luna JM, Eaton E, Simone CB, Ungar LH, Solberg TD. MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. Sci Rep. 2016;6:37854.
51. Oh JH, Kerns S, Ostrer H, Powell SN, Rosenstein B, Deasy JO. Computational methods using genome-wide association studies to predict radiotherapy complications and to identify correlative molecular processes. Sci Rep. 2017;7:43381.
52. Lee S, Kerns S, Ostrer H, Rosenstein B, Deasy JO, Oh JH. Machine learning on a genome-wide association study to predict late genitourinary toxicity after prostate radiation therapy. Int J Radiat Oncol Biol Phys. 2018;101(1):128–35.
53. Koller D, Friedman N. Probabilistic graphical models: principles and techniques—adaptive computation and machine learning. Cambridge: The MIT Press; 2009.
54. Oh JH, Craft JM, Townsend R, Deasy JO, Bradley JD, El Naqa I. A bioinformatics approach for biomarker identification in radiation-induced lung inflammation from limited proteomics data. J Proteome Res. 2011;10(3):1406–15.

55. Chow C, Liu C. Approximating discrete probability distributions with dependence trees. IEEE Trans Inf Theor. 2006;14(3):462–7.
56. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn. 1992;9(4):309–47.
57. Madigan D, York J, Allard D. Bayesian graphical models for discrete data. Int Stat Rev. 1995;63(2):215–32.
58. Smith WP, Doctor J, Meyer J, Kalet IJ, Phillips MH. A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model. Artif Intell Med. 2009;46(2):119–30.
59. Friedman JH. On bias, variance, 0/1ñloss, and the curse-of-dimensionality. Data Min Knowl Discov. 1997;1(1):55–77.
60. Kazmierska J, Malicki J. Application of the nave Bayesian classifier to optimize treatment decisions. Radiother Oncol. 2008;86(2):211–6.
61. Sutton RS, Barto AG. Introduction to reinforcement learning. Cambridge: MIT Press; 1998.
62. Kulkarni P. Reinforcement and systemic machine learning for decision making. Hoboken: Wiley-IEEE Press; 2012.
63. Watkins CJCH, Dayan P. Technical note: Q-learning. Mach Learn. 1992;8(3):279–92.
64. Kosorok M, Moodie E. Adaptive treatment strategies in practice. Philadelphia, PA: SIAM; 2016.
65. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy SA. A "SMART" design for building individualized treatment sequences. Annu Rev Clin Psychol. 2012;8(1):21–48.
66. Kim M, Ghate A, Phillips MH. A Markov decision process approach to temporal modulation of dose fractions in radiation therapy planning. Phys Med Biol. 2009;54(14):4455.
67. Vincent R, Pineau J, Ybarra N, El Naqa I. Practical reinforcement learning in dynamic treatment regimes. In: Kosorok MR, Moodie EEM, editors. Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine. Philadelphia, PA: SIAM; 2016.
68. El Naqa I, Feng M, Bazzi L, et al. Reinforcement learning strategies for decision making in knowledge-based adaptive radiation therapy: application in liver cancer. Int J Radiat Oncol Biol Phys. 2016;96(2):S45.

Overview of Deep Machine Learning Methods

4

Julia Pakela and Issam El Naqa

4.1 Introduction

Deep machine learning or "deep learning" refers to a class of machine learning methods, which takes raw data as inputs and, through training, learns multiple layers of relevant latent features to map the raw inputs to the desired output space; the desired mapping is defined by either a reward or a loss function of the outputs for detection or classification tasks [1]. This is in contrast to shallow learning, where the features are manually crafted and do not contain multiple layers of abstraction. Conceptually then, deep learning can be applied to any machine learning technology as depicted in Fig. 4.1, but as of this time has been practically shown to be most effective with deep neural networks [2, 3], which will be the main subject of this chapter.

Deep learning algorithms typically take the form of artificial neural networks (ANNs) with multiple hidden layers; however, it is important to recognize that deep learning is defined by the ability to automatically learn relevant features (data representations) from raw inputs rather than any particular structure scheme [3]. This is achieved using highly parameterized networks featuring layers of nodes, where each node takes as input the weighted outputs from other nodes and produces its own output using a nonlinear transformation on a weighted sum of the inputs plus an additional bias term (as will be discussed later).

J. Pakela (✉)
Department of Radiation Oncology and Applied Physics, University of Michigan, Ann Arbor, MI, USA
e-mail: jpakela@umich.edu

I. El Naqa
Department of Machine Learning, Moffitt Cancer Center, Tampa, FL, USA

Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: ielnaqa@med.umich.edu; Issam.elnaqa@moffitt.org

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_4

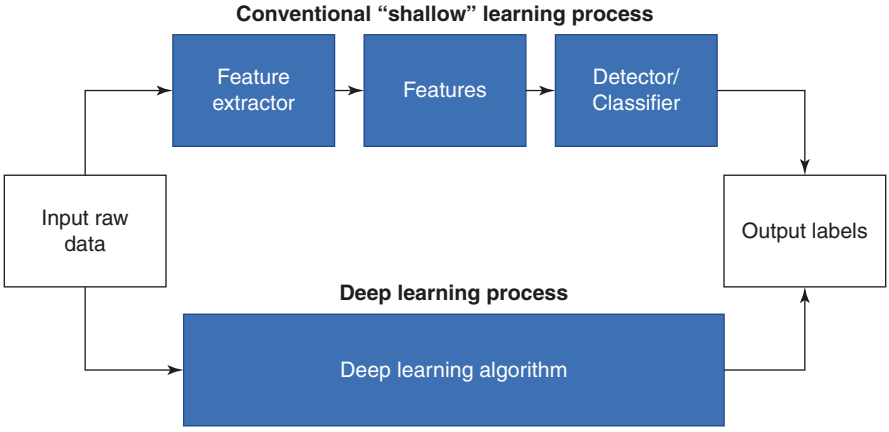


Fig. 4.1 Conventional "shallow" machine learning (top) versus deep learning algorithms, where image data representation and classification are handled within the same framework

The foundational groundwork for deep learning was laid decades prior to its rise in popularity: first in the 1950s when Rosenblatt introduced the concept of the perceptron (a precursor to the hidden nodes in today's neural networks) and later in the 1980s with Hinton and Sejnowski's invention of the Boltzmann Machine (a multilayer network similar in design to modern neural networks) [4, 5]. However, for several decades deep learning methods were not viable for solving real-world problems. This was partly due to limitations on the computer processing power necessary to train the multilayer networks envisioned by deep learning researchers and also due to a need for further algorithmic innovations (such as dropout and stochastic optimization schemes) to improve training efficiency. By the early 2010s, such discoveries, in combination with improvements in computer parallel processing power and GPUs, helped to provide the necessary conditions for deep learning methods to take the world by storm. This was a phenomenon which arguably began with the startling 2012 victory in the ImageNet competition (an annual image classification challenge which sets the bar for the state-of-the art in computer vision) with AlexNet: a deep convolutional neural network architecture [6]. Deep learning has since been applied with great success to many diverse challenges, including language translation, speech recognition, training of self-driving cars, and even stock-market predictions [7].

Since its recent rise in popularity, deep machine learning has also made a significant impact on the medical field, specifically in the areas of diagnostic (Radiology) and therapeutic (Radiation oncology) radiological sciences [8].

Applications in radiology and computer-aided diagnosis (CAD) were at the forefront of application of machine and deep learning in medicine since the 1980s [9–14]. These applications included using ANNs generally [15] as well as convolutional neural networks (CNNs) for breast cancer detection and diagnosis [16–18]. This

pioneering work has led to several FDA approved systems including the QuantX Advanced system to aid in breast cancer diagnosis (CADx), developed originally by Giger and colleagues [19–23]. Today, deep learning techniques are touching every aspect of radiology from improving image quality by improving current image reconstruction and filtering of modalities such as MRI [24], CT [25] and ultrasound [26], to image segmentation [27], registration [28], to precision medicine and the derivation of reproducible imaging biomarkers [29].

As another innovative and data-heavy field, radiation oncology has been uniquely positioned to experience an explosion of deep learning applications as well. The number of radiation oncology and medical physics papers published which feature deep learning has increased steadily over the past 5 years, with a wide variety of applications including treatment planning [30–33], adaptive radiotherapy [34–38], quality assurance [39–42], and outcomes modeling [43–49]. For the interested reader, there also exist several comprehensive reviews of deep learning, machine learning, and artificial intelligence applications in medical physics and radiation oncology [50–53]. Figure 4.2 provides a visualization of the rise in deep learning both in biomedical literature and in publications specific to medical physics and radiation oncology.

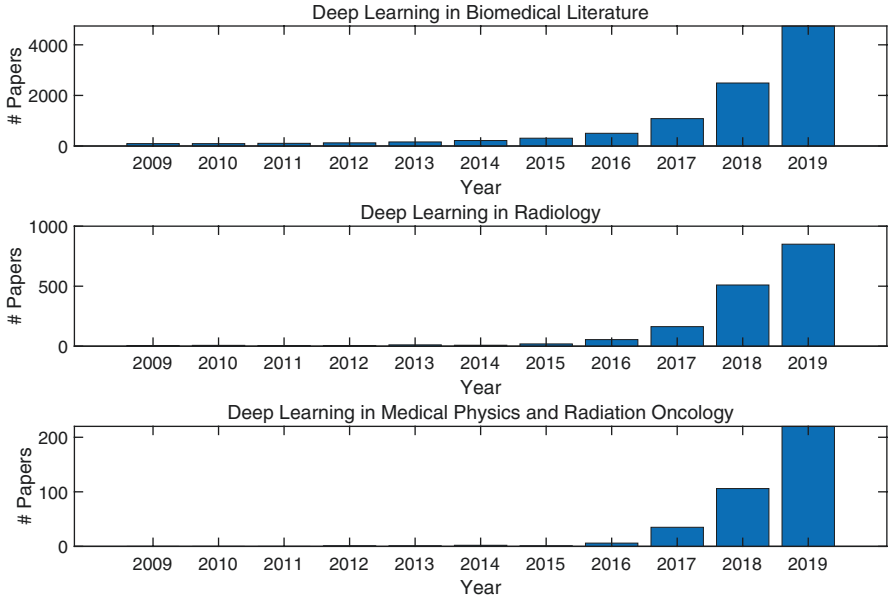


Fig. 4.2 Incidence of deep-learning-themed papers in biomedical literature as well as in medical physics and radiation oncology. Data was obtained through the advanced search function on PubMed. Search criteria used were (all fields: deep learning), (all fields: deep learning) AND (all fields: medical physics OR all fields: radiation oncology) and (all fields: deep learning) AND (all fields: radiology), respectively

4.2 The Vanilla Neural Network

A standard "fully connected" or "vanilla" neural network consists of layers of neurons (also called "units" or "nodes"). In a given layer, each node has weighted connections to every node in the previous layer and every node in the next layer. Nodes do not share connections within the same layer. This forward-directed flow of information is why vanilla neural networks are also referred to as "feedforward neural networks." In literature, one may also see the term "multilayer perceptron" used to refer to fully connected feedforward neural networks.

The structure of a three-layer vanilla neural network is visualized in Fig. 4.3. The first layer in the neural network is the input layer, $\mathbf{x}^{(1)}$. The input layer takes raw data inputs and propagates them to the next layer. The final layer, $\mathbf{x}^{(3)}$, is the output layer. The results from the output layer represent the network's output and are used to define a loss function. The width (i.e., number of nodes) for the input and output layers are typically determined by inherent characteristics of the data and the task the network performs. For example, a network designed to use greyscale images with 128 × 128 pixels will have 16,384 nodes in its input layer, where each node represents the intensity of a pixel in the image. If the task for the network is to classify handwritten digits, then the output layer will have a width of 10, with each node outputting the raw prediction score (to be later normalized into a probability) of a given digit. Hidden layers ($\mathbf{x}^{(2)}$) consist of all layers between the input and output. The width and total number of hidden layers are *hyperparameters*, meaning their

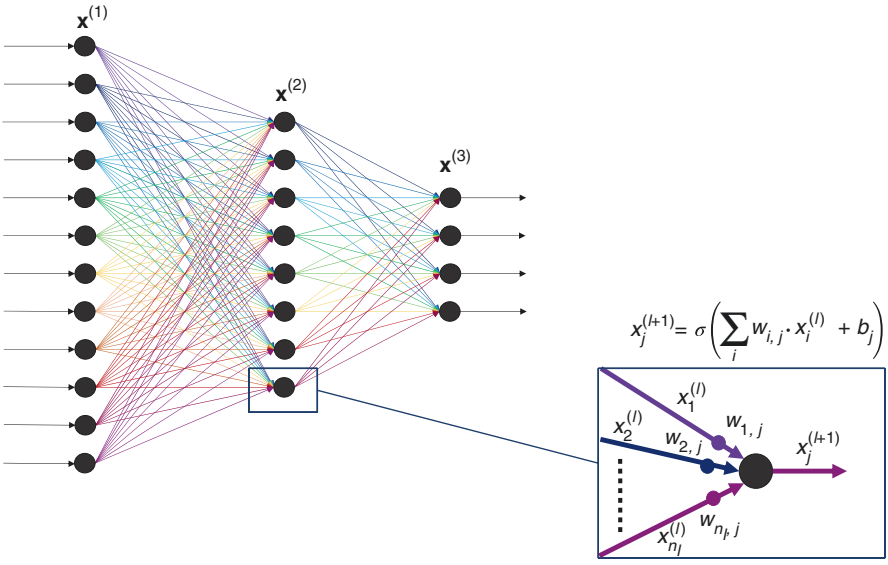


Fig. 4.3 A three-layer neural network (input layer, hidden layer, and output layer). The output of each node is determined by performing a nonlinear transformation (known as the activation function) on the sum of the weighted inputs plus an additional bias term

values are chosen by the user and are not updated during training. The act of selecting/tuning hyperparameters for a given model is an active area of research and considered an art in and of itself [54, 55].

4.2.1 Training a Neural Network

The process of training a neural network involves training a set of weights and biases which act like "knobs" to control the flow of information across the network. The number of weights and biases needed to train depends on the size of the network: the example neural network shown in Fig. 4.2 has 128 weights and 12 biases, but a typical network can have hundreds of millions of weights [3]. As mentioned in Sect. 4.2, each node in a neural network receives information from each of the nodes in the previous layer (this can be visualized through the color-coded arrows in Fig. 4.3). The output for each node is a function of the weighted sum of each incoming signal plus a bias term. We can therefore write the output for a given node, $x_j^{(l+1)}$, within layer, $l + 1$, as:

$$x_j^{(l+1)} = \sigma\left(\sum_i W_{ij} \cdot x_i^{(l)} + b_j\right) \tag{4.1}$$

where W_{ij} represents the weight associated with the connection from node $x_i^{(l)}$ to $x_j^{(l+1)}$, b_j is the bias associated with node $x_j^{(l+1)}$, and σ is the activation function. In a biological neural network, in order for a neuron to fire, it needs to receive enough electrical signal from other neurons to overcome a threshold known as the activation potential. In artificial neural networks, the *activation function* performs a similar role to the activation potential: it determines whether the node has received enough "signal" to fire. Common choices for activation functions include the hyperbolic tangent, sigmoid, softmax, and Rectified Linear Unit (ReLU) functions. Of these three, ReLU is arguably the most popular as it has been found to make networks more easily trainable [3, 56]. Figure 4.4 displays equations and graphical representations for each of these functions. Importantly, one trait that neural network activation functions share is that they are nonlinear functions: they perform a nonlinear operation or transformation on input features to produce an output. This trait is significant because it allows neural networks to model complex, nonlinear relationships between the input data and the desired output.

At the start of training, the weights and biases of the network are initialized. A common method for initialization which has been found to work well is the *Xavier* initialization, in which the biases are initially set to 0 and the weights for a given node are randomly sampled from a normal distribution and bounded by:

$$\left[-\frac{6}{\sqrt{n_j + n_{j+1}}}, \frac{6}{\sqrt{n_j + n_{j+1}}}\right] \tag{4.2}$$

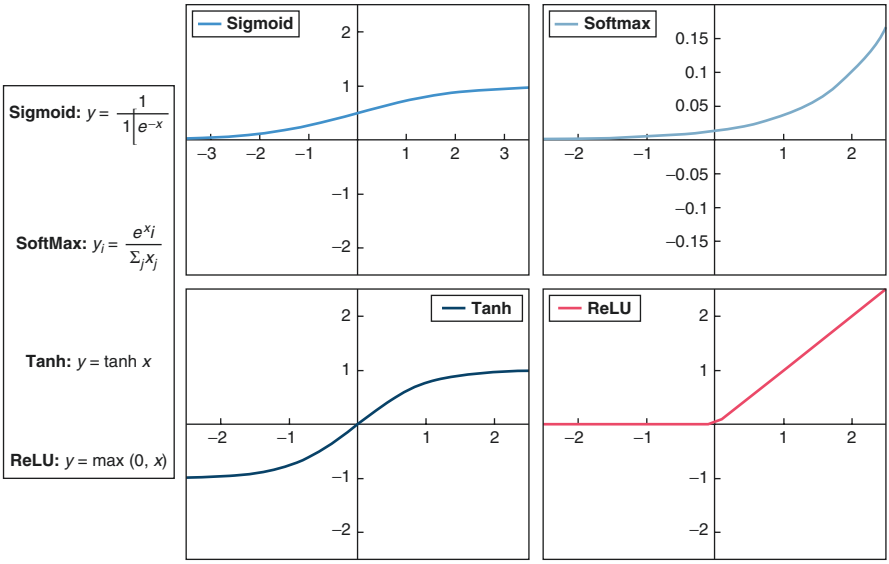


Fig. 4.4 Common nonlinear activation functions used in neural networks

where n_j and n_{j+1} are the number input and output connections to the node, respectively [57]. Another popular initialization technique is *He* initialization [58], which instead bounds uniformly sampled weights from:

$$\left[-\sqrt{\frac{2}{n_j}}, \sqrt{\frac{2}{n_j}}\right] \tag{4.3}$$

After initialization, data is fed into the network, which produces an output. This output is then fed into a cost function (also called a loss function) to calculate a loss. The weights and the values of the network are then updated via *gradient descent*. The gradient of the loss function is calculated with respect to the weights and biases of every node in the network using the chain rule. Each of these parameters is then updated by adding the gradient term multiplied by a fractional learning rate typically on the order of hundredths or thousandths. This process is then repeated in an iterative fashion until a stopping criteria is met—either the loss reaches a minimum criterion, or a max number of iterations is performed. For a large network, calculating the gradient of the loss function with respect to millions of parameters is a computationally expensive endeavor. A major breakthrough in the deep learning community was the introduction of the backpropagation algorithm, which provided an efficient means of calculating the loss gradient with respect to network parameters [1, 59].

4.2.2 Hyperparameters Associated with Training

Four significant hyperparameters associated with the training process are *batch size*, number of *epochs, learning rate,* and *dropout rate*. The *batch size* is the number of data samples fed to the network before the weights and biases are updated. Batch size can range from 1 (the network is updated after a single sample) to the size of the training dataset (the network is only updated after it has seen every possible data sample). Training schemes with a batch size of 1 are referred to as stochastic gradient descent, while training schemes with a batch size equal to the number of training samples are referred to as just gradient descent or batch gradient descent. If the batch size is a number between 1 and the training sample size, the algorithm is said to undergo minibatch gradient descent. Of these three training schemes, batch gradient descent is the least noisy because it uses every data sample when calculating the gradient—meaning it isn't going to be impacted by variations within the dataset. However, for very large datasets (on the order of millions of training examples) it is computationally expensive to pass all of the data through the network before each update, which increases the overall training time. Stochastic gradient descent leads to a much noisier learning process because the network is updated after only a single data sample, which may not be representative of the dataset as a whole. A benefit of stochastic gradient descent is that in updating after only a single sample, the loss function may approach a value close to the global minimum at a faster rate. However, this also means the network must be updated more frequently, which is also computationally expensive. Minibatch gradient descent serves as a compromise between these two extremes by using a batch size large enough to be somewhat representative of the entire dataset, which minimizes training noise, but small enough that the network is able to update more frequently, leading to faster overall training times. During training, it is standard to normalize the network inputs on a batch by batch basis—a procedure called *batch normalization*—in order to improve the training speed and model stability [60].

An *epoch* refers to an instance in which the entire training dataset has been passed through the network. The *number of epochs* is thus a hyperparameter which describes how many times the entire dataset is passed through the network during training. Typically, a network requires the entire training dataset to be passed through it multiple times before the loss function converges to a minimum.

The *learning rate* (mentioned briefly in Sect. 4.2.2) is a fractional coefficient applied to the loss function gradient and can be thought of as the step size used when updating the weights and biases to minimize the loss function. A popular choice for neural network optimization is to use an adaptive learning rate defined using the *Adam* stochastic optimization algorithm [61].

Dropout is a standard practice used during neural network training to prevent overfitting of the model. It consists of randomly selecting a set number of nodes in the network and setting their output to 0—effectively dropping them (and their connections) from the network. The fraction of neurons in a given layer which are dropped during training is called the *dropout rate* [62].

4.2.3 What Makes a Neural Network Deep?

Deep machine learning was defined in Sect. 4.1 as any machine learning method which learns multiple layers of latent features from raw data. Deep neural networks are by far the most successful deep learning architecture to date, so much so that the phrase *deep learning* is sometimes used interchangeably with *neural networks*. This subsection aims to provide some intuition as to why and under what conditions neural networks perform so well.

One important property of neural networks (already mentioned in Sect. 4.2.1) is the use of nonlinear activation functions, which allow the model to learn nonlinear relationships between data and outputs. With respect to modeling capabilities, it has been proven that neural networks are universal approximators: a feedforward neural network with as few as one hidden layer and arbitrary bounded, non-constant activation functions can be used to approximate any real, continuous function on a closed and bounded subset of R^n to any accuracy [63]. This result is often referred to as the *universal approximation theorem* and tells us that a multilayer feedforward neural network can represent nearly any function given a sufficient number of hidden nodes. The result has undergone multiple iterations to clarify what conditions the activation function must meet for the theorem to be true—notably with showing that the universal approximation theorem extends to the highly successful ReLU activation function [64, 65].

A key limitation of the universal approximation theorem is that it does not tell us how large (in terms of layers or neurons) the network needs to be in order to achieve universal approximation, nor does it guarantee that a sufficiently large network can be successfully and efficiently trained to be generalizable—i.e., a network that is large enough to represent any function may be trained to perform with high accuracy on the training dataset but still perform poorly on the validation data [65]. While a shallow neural network with only one hidden layer is capable of being a universal approximator, such a network requires so many nodes in its hidden layer that efficient and effective training is infeasible—it is too wide [65, 66]. It has been found that neural networks with more hidden layers reduce the necessary width for each layer and ultimately the total number of nodes needed for high accuracy, providing the best trade-off between training efficiency and model performance. Generally, a neural network is considered *deep* if it has multiple hidden layers.

4.2.4 Example: Neural Network for Binary Classification

In this subsection we provide a didactic example, borrowed from Cui et al., of a feedforward neural network performing a binary classification task within a radiation oncology paradigm [67].

The input data were acquired from the Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) and Lung Squamous Cell Carcinoma (TCGA-LUSC) datasets, respectively. The selection criteria included patients from these datasets who received external beam radiotherapy for a primary tumor and had

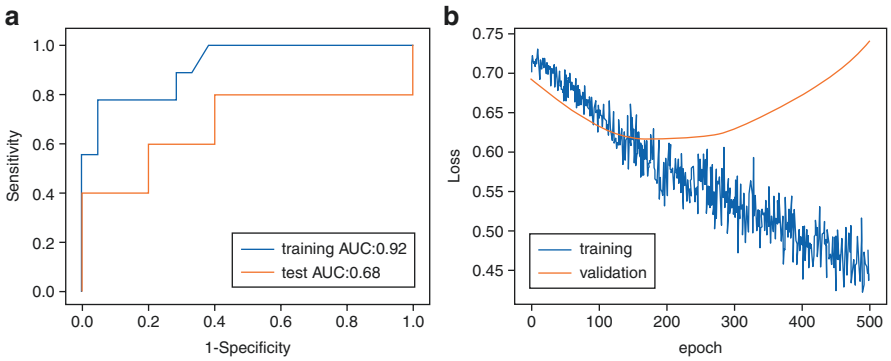


Fig. 4.5 (**a**) Final ROC scores for network trained to predict local control in lung cancer patients. (**b**) Training and validation loss functions over the course of 500 epochs

complete dose and local control information. In total 45 patients were selected and were randomly split into training and validation sets. Data collected from each patient included patient outcome (defined as either local control or tumor progression) as well as predictive variables such as gender, primary tumor stage, total radiation dose, and smoking history. These predictive variables were converted to numerical values where necessary and standardized using the z-score transformation. Any missing data was filled with median values prior to standardization.

A feedforward neural network built using the PyTorch framework was trained on the lung-patient dataset to predict patient outcomes. The network has two hidden layers; each layer features a ReLU activation function and dropout (dropout rate = 20%). The final output layer undergoes a sigmoid activation. The model was trained using Adam optimization for 500 epochs. The training and testing loss and final ROC curve and code are displayed below. The code and data can also be downloaded at: https://github.com/sunancui/lung_TCGA_prediction (Fig. 4.5).

```
import numpy as np
import pandas as pd
import os
import torch
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
# processing the data, convert to numeric values
def data_processing(data_df):
        data_df['gender']=data_df['gender'].replace(["MALE","FEM
ALE"],[0,1])
        data_df['pathologic_N']=data_df['pathologic_N'].replace
(["N0","N1","N2","N3","NX"],[0,1,2,3,np.nan])
```

```
    data_df['pathologic_stage']=data_df ['‘pathologic_stage’]
.replace(["Stage IA","Stage IB","Stage IIA","Stage IIB","Stage II
IA", "Stage IIIB","[Discrepancy]"],[1,1,2,2,3,3,np.nan])
    data_df['pathologic_T']=data_df['pathologic_T'].replace(["T1",
"T1a","T1b","T2","T2a","T2b","T3","T4"],[1,1,1,2,2,2,3,4])
    data_df['other_dx']=data_df['other_dx'].replace(["No","Yes,
History of Prior Malignancy"],[0,1])
    data_df['tobacco_smoking_history']=data_df['tobacco_smoking_
history'].replace(["1","2","3","4","[Not Available]"],[1,2,3,4,
np.nan])
    data_df['primary_outcome']=data_df['primary_outcome'].replace
(["progressive",'local'],[1,0])
    #fill missing data with median values
    data_df_fill=data_df.fillna(data_df.median())
    return data_df_fill
os.chdir("/Users/sunan/Desktop/github/lung_TCGA_prediction")
AD_SC_patient=pd.read_csv("./rd_AD_SC.csv",index_col=0)
select_column=["gender","pathologic_N","patholo
gic_stage","pathologic_T", 'other_dx', 'tobacco_smoking_
history','radiation_total_dose','primary_outcome']
data_all=AD_SC_patient.loc[:,select_column]
data_all_values=data_processing(data_all)
##features
data_X=data_all_values.iloc[:,:-1].values
scaler=StandardScaler()
scaler.fit(data_X)
X_scaled=scaler.transform(data_X)
##label
data_Y=data_all_values.iloc[:,-1].values
##train_test_split
X_train, X_test, Y_train, Y_test=train_test_split
(X_scaled,data_Y,test_size=0.33,random_state=0,stratify=data_Y)
##convert numpy array to torch tensor
X_train_torch=torch.from_numpy(X_train)
Y_train_torch=torch.from_numpy(Y_train.reshape(-1,1))
X_test_torch=torch.from_numpy(X_test)
Y_test_torch=torch.from_numpy(Y_test.reshape(-1,1))
##set some parameters
input_dim=X_train.shape[1]
output_dim=1
hidden_dim=5
learning_rate=0.001
num_epoch=500
# define a 2-hidden layer fully-connected NN
```

```
class MLP_NN(torch.nn.Module):
    def __init__(self, input_dim, hidden_dim_1,hidden_dim_2, out-
put_dim):
        super(MLP_NN, self).__init__()
        self.L1=torch.nn.Linear(input_dim,hidden_dim_1)
        self.D1=torch.nn.Dropout(0.2)
        self.L2=torch.nn.Linear(hidden_dim_1,hidden_dim_2)
        self.D2=torch.nn.Dropout(0.2)
        self.L3=torch.nn.Linear(hidden_dim_2,output_dim)
    def forward(self,x):
        a1=torch.relu(self.L1(x))
        a1=self.D1(a1)
        a2=torch.relu(self.L2(a1))
        a2=self.D2(a2)
        outputs=torch.sigmoid(self.L3(a2))
        return outputs
#initialize the model
model=MLP_NN(input_dim,hidden_dim,hidden_dim,output_dim)
#using binary cross entropy as loss
criterion=torch.nn.BCELoss(reduction='mean')
#using Adam optimizer
optimizer=torch.optim.Adam(model.parameters(),
 lr=learning_rate)
history_loss_train=[]
history_loss_test=[]
for i in range(num_epoch):
    #train data
    model.train()
    optimizer.zero_grad()
    y_pred=model(X_train_torch.float())
    #define the loss, adding l2 penalty
    loss=criterion(y_pred,Y_train_torch.float())
    loss.backward()
    optimizer.step()
    # evaluate on test data
    model.eval()
    y_pred_test=model(X_test_torch.float())
    loss_test=criterion(y_pred_test,Y_test_torch.float())
    if i%10==0:
        print(i, loss, loss_test)
    history_loss_test.append(loss_test.detach().numpy())
    history_loss_train.append(loss.detach().numpy())
# plot the history of training/test loss
plt.figure()
```

```
plt.plot(np.
arange(0,num_epoch),history_loss_train,label="training")
plt.plot(np.arange(0,num_epoch),history_loss_test,
label="validation")
plt.xlabel("epoch",fontsize=18)
plt.ylabel("Loss",fontsize=18)
plt.legend(prop={'size':16})
plt.show()
#calculate AUC for traing/test sets
auc_train=roc_auc_score(Y_train, y_pred.detach().numpy())
auc_test=roc_auc_score(Y_test, y_pred_test.detach().numpy())
fpr, tpr,thresholds=roc_curve(Y_train, y_pred.detach())
plt.figure()
fpr_t, tpr_t,thresholds=roc_curve(Y_test, y_pred_test.detach().
numpy())
plt.plot(fpr,tpr,label="training AUC:"+str(round(auc_train,2)))
plt.plot(fpr_t,tpr_t,label="test AUC:"+str(round(auc_test,2)))
plt.legend(prop={'size':16})
plt.xlabel('1-Specificity',fontsize=16)
plt.ylabel('Sensitivity',fontsize=16)
plt.show()
```

4.3 Autoencoders

Autoencoders (AEs) represent a class of unsupervised deep learning architectures which use a two-part encoder-decoder framework to learn a lower-dimensional data representation (or feature representation). An AE's encoder portion takes raw data as input and encodes it into a lower-dimensional representation of the data or latent-space representation, h. The decoder portion takes the latent-space representation, h, as input and then outputs a reconstruction which is the same dimension as the original raw data. The loss function is defined as a function of the difference between the input, x, and the final reconstruction, x'. Figure 4.6 displays an example of a simple AE which uses feedforward nodes as its building blocks. Note, however, that AEs can also utilize CNNs (Sect. 4.4) in their design. AEs have uses in both feature learning and for reducing high-dimensional datasets into their most critical components for more efficient training on other architecture types. Variational autoencoders (VAE) represent a variant of AEs which utilize the same encoder-decoder architecture, but additionally assume that the latent space of the data is subject to a Gaussian distribution with a mean, μ, and variance, σ [68]. In a VAE, the encoder receives an input and generates parameters which define the latent space distribution. This latent space distribution is then sampled to generate the latent

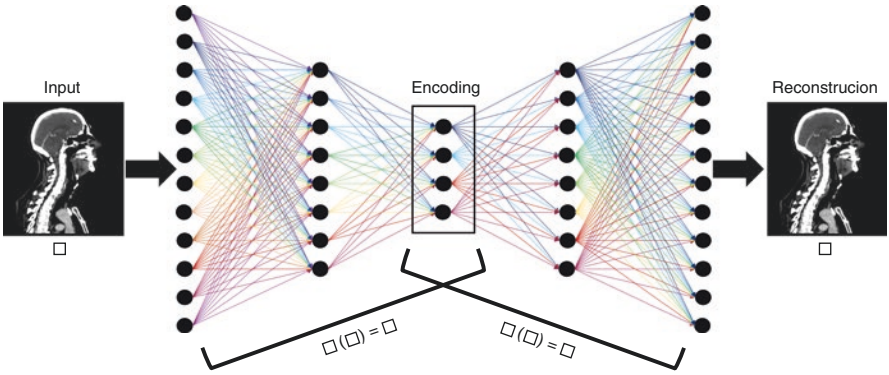


Fig. 4.6 Schematic of autoencoder architecture

representation, *h,* which is subsequently fed to the decoder to produce the reconstruction, *x'*. The loss function used for VAEs can be written as [49, 68]:

$$L = \| x - x' \|^2 + \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\left(\sigma_j^2\right) - \mu_j^2 - \sigma_j^2\right) \tag{4.4}$$

The first term in Eq. (4.4) represents the reconstruction error and is an approximation of the expected negative log likelihood of outcome of x being correctly observed given the network-generated latent vector, h. The second term serves as a regularization penalty and is an approximation of the Kullback–Leibler divergence between the network-learned latent space distribution and a standard normal distribution. Because a VAE tries to learn both the true latent distribution and the latent representation, it can be utilized as a generator to produce additional data samples, which are representative of the true dataset.

4.4 Convolutional Neural Networks (CNNs)

Among the most well-known deep learning algorithms are convolutional neural networks (CNNs). Biologically inspired by animal's visual cortex, CNNs were first proposed in 1980 [69] and have since revolutionized the field of computer vision. An important characteristic of CNNs is their translation invariant properties, which is a main attraction for their popularity in computer vision and imaging applications. CNNs differ from standard feedforward neural networks in that they are able to learn features which depend on local structural relationships within the data, making them particularly adept at tasks involving image data.

The standard building blocks of a CNN are convolutional layers, which are responsible for creating feature maps of the input data; pooling layers, which reduce the number of parameters in the model and help to prevent overfitting; and one or more fully connected layers at the end of the network. Each of these components is discussed in greater detail in the following sections. Figure 4.7 displays a simple

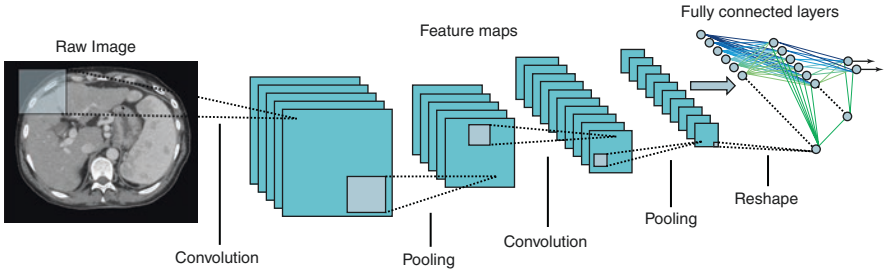


Fig. 4.7 Schematic of a CNN with 2 convolutional layers, and 2 fully connected layers

CNN with two convolutional layers, two pooling layers, and three fully connected layers.

4.4.1 Convolutions

A convolution is an operation on two functions which produces a third; it is typically denoted with an asterisk, $*$. The convolution of two real-valued, continuous functions is represented mathematically as the integral over the product of two functions, where one of the functions is flipped and shifted:

$$(f*g)(x)=\int f(t)\cdot g(x-t)\,dt \tag{4.5}$$

One known property of convolutions is that they are commutative, meaning $(f*g)=(g*f)$, allowing for the order in Eq. (4.5) to be flipped when convenient. In machine learning—particularly image processing—both the input data and the model parameters are discrete values and are therefore not represented by continuous functions but by *discrete functions* defined over a finite range. For two discrete, complex functions defined on a set from is $-M$ to M, the discrete convolution is defined as:

$$(f*g)[x]=\sum_{m=-M}^{M} f[m]\cdot g[x-m] \tag{4.6}$$

For two-dimensional discrete functions (defined on $-M$ to M and $-N$ to N, respectively), the corresponding 2D convolution can be written (using the commutative property) as:

$$(f*g)[x,y]=(g*f)[x,y]=\sum_{m=-M}^{M}\sum_{n=-N}^{N} f[x-m,y-n]\cdot g[m,n] \tag{4.7}$$

The function f is typically referred to as the input—in the case of Eq. (4.7), f would be an $M\times M$ grid of values, such as a 2D greyscale image. The function g (in Eq. (4.7), an $N\times N$ grid of values) is called a kernel. The values within the kernel are learned parameters initialized at the beginning of training and then updated

during backpropagation. One detail to note is that the operation that many machine learning libraries (including PyTorch and Tensorflow) perform in practice is actually a cross correlation operation, which is similar to convolution but does not flip the direction of the kernel [65, 70, 71]:

$$(g * f)[x, y] = \sum_{m=-M}^{M} \sum_{n=-N}^{N} f[x+m, y+n] \cdot g[m, n] \tag{4.8}$$

The use of the cross correlation in place of a true convolution is a choice of convenience which has no impact on a network's performance capabilities. The difference between a CNN trained using cross correlation vs. one trained using convolution is that the kernel learned for the latter will be flipped with respect to the former [65]—i.e., the weights learned are saved in different positions. For simplicity we will continue to refer to the operation which occurs in the convolutional layer as a "convolution," with the understanding that unless stated otherwise we are referring to the cross correlation process.

The process of convolving a kernel with an input can be visualized as the kernel sliding across the input volume, where the value stored in each index of the feature map is the sum of the products between the overlapping grids (see Fig. 4.8). The convolution operation can be generalized to any dimension—allowing it to be applied to 3D images such as CT scans, videos, and color images. A color image can be represented as a 3D image with dimensions of image height × image width × 3—where the third dimension represents the red, blue, and green channels, respectively. In the case of videos, the third dimension represents time, and the number of channels is determined by the number of timeframes in the video.

The output of a convolution in a CNN is called a *feature map* and contains higher-order information about the original sample input. The dimension of the feature map is determined by the input size, the kernel size, the stride length (how far the kernel shifts between each sub-calculation), and whether the input has been padded around its border. Figure 4.8a displays a convolution operation with no

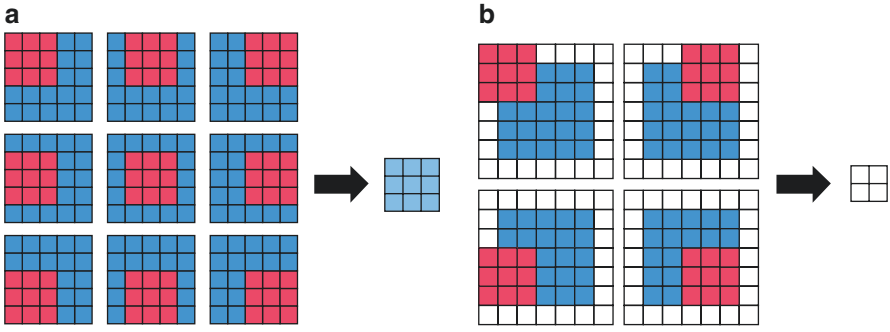


Fig. 4.8 The convolution operation can be visualized as sliding the kernel across the input and taking the sum of the products between the overlapping grid values. (**a**) displays the convolution operation on an input with no padding and a kernel stride length of 1, while (**b**) shows a convolution operation on an input with 1 layer of padding and a kernel stride length of 3

input padding and a stride length of 1, while Fig. 4.8b shows an example with 1 layer of padding and a stride length of 3. A visually rich review on convolutional operations for CNNs can be found in Dumoulin and Visin's guide to convolutional arithmetic [72], while a more mathematical description can be found in Goodfellow et al.'s deep learning textbook [73]. Feature maps are typically passed through a nonlinear function (such as ReLU) and undergo pooling after which they are either used as input into another convolution operation (which represents a new convolutional layer in the network) or reshaped into a vector to be used as input into a feedforward neural network. Most CNNs have a feedforward neural network as the backend of their framework, which is typically referred to as the fully connected layer.

An important characteristic of CNNs is that the convolutional method by which feature maps are created requires that for a convolutional layer, the parameter weights (defined by the kernel) are shared across the entire input. This, combined with the use of kernels whose dimensions are much smaller than the input, greatly improves the model training efficiency and allows the model output to be translationally invariant [65].

4.4.2 Pooling

Prior to their use as inputs for new layers, feature maps typically undergo *pooling*: an operation which reduces their dimensionality. Pooling can be thought of as splitting a feature map into uniformly sized regions and saving a single representative value from each region in a final, lower-dimensional feature map. There are several types of pooling: common variants include max pooling, which saves the maximum value in the region of interest; mean pooling, which saves the average value in the region of interest; and L2 norm pooling, which saves the Euclidean norm in the region of interest. The size of the region of interest used, and by extension the extent of the dimensional reduction, is a hyperparameter chosen by the user. Pooling is an important component of CNN models because it teaches the models to identify intrinsic patterns in the data without being sensitive to small translations. One can imagine how this property would be valuable in the context of a classification problem. For example, if the goal of a CNN is to identify the presence of malignant tumors, it is not necessary for the network to learn the exact pixel locations of the tumor in order to perform this task—simply knowing there is one somewhere in the top right corner of an image would suffice.

4.5 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a type of deep learning method designed to handle sequential data or time-dependent data such as that found in text, audio, or video. RNNs differ in structure from vanilla neural networks in that the width of the network (i.e., number of nodes per layer) is determined by the number of elements

in the sequence and the flow of information is not feedforward but rather flows both "upwards" across the layers and "horizontally" across the sequence itself. Within each layer of an RNN, a given cell at time-step t calculates a *hidden state*, h_t, which holds the memory of the network up through that time-step. For any RNN, the hidden state can be written as:

$$h_t = f\left(x_t, h_{t-1}\right) \tag{4.9}$$

where f is a function of the input state x_t and the previous hidden state h_{t-1}. For a standard or vanilla RNN, the hidden state is defined as:

$$h_t = \sigma\left(W_{ih}x_t + b_{ih} + W_{hh}h_{t-1} + b_{hh}\right) \tag{4.10}$$

where W_{ih}, b_{ih}, W_{hh}, and b_{hh} are network weights and biases and σ is an activation function, typically tanh or ReLU. RNNs share the same weights and biases for each cell in the sequence, which significantly reduces the number of parameters the network needs to train. This property is also the reason why RNNs bear the name recurrent: the operation in Eq. (4.9) is repeated throughout each node in the network with only the inputs changing. The output at each node in the final layer of the network is calculated as:

$$o_t = \sigma\left(W_{oh}h_t + b_{oh}\right) \tag{4.11}$$

where σ again represents a nonlinear activation function. For RNNs with more than one layer, the output o_t is only calculated for the set of hidden states in the final layer, and the nodes of intermediate layers take corresponding hidden states from the previous layer as their input, x_t. The general structure of an RNN is displayed in Fig. 4.9.

A significant limitation of vanilla RNNs is that as the sequence length increases, the impact of long-term memories become increasingly small, leading to performance degradation. This challenge is typically referred to as the *vanishing gradient* problem because the gradients used for updating the networks weight and bias parameters become negligibly small and the model ceases to learn [74]. It has been

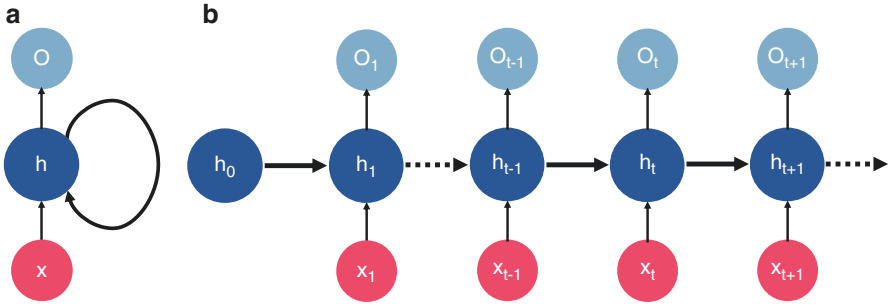


Fig. 4.9 Schematic diagram of a standard recurrent neural network (RNN). (**a**) shows the network in a folded state, representing that the final output at any point in the sequence is a product of the current input and the networks memory of previous states. (**b**) displays the network fully unfolded

found that a way to mitigate the vanishing gradient problem is to build in a way for the network to learn to "forget" some portion of its past history while holding onto memories that are essential for minimizing its particular loss function. Two leading variations on the standard RNN which incorporate this concept into their design are Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs); each are discussed in detail in the following sections.

4.5.1 Long Short-Term Memory (LSTM)

LSTM is a variant of RNNs, which adaptively learns to remember only the information from its sequence history which is relevant to the problem it is solving. The LSTM architecture was described by Hochreiter and Schmidhuber in 1997 [75]. Since its conception, many variations and evolutions of LSTM have been proposed, but the version of LSTM currently accepted as standard or "vanilla" by the deep learning community was described by Graves and Schmidhuber in 2005 [76, 77]. As such, when sources refer to an LSTM without specifying which version, one can infer the Graves and Schmidhuber implementation is implied.

Equations (4.12)–(4.17) depict the calculations which occur for each cell in a vanilla LSTM network, where σ represents the sigmoid activation function and $\odot$ denotes the Hadamard product, which performs element-wise multiplication between two variables.

$$i_t = \sigma\left(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}\right) \tag{4.12}$$

$$f_t = \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}\right) \tag{4.13}$$

$$g_t = \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}\right) \tag{4.14}$$

$$o_t = \sigma\left(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}\right) \tag{4.15}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{4.16}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{4.17}$$

LSTMs feature three states for each time-step, t: the input state, x_t; the hidden state, h_t; and the cell state, c_t. The cell state is a key component of LSTMs because it allows selective information to travel long distances across the network, giving that network a long-term memory. The variables i_t, f_t, g_t, and o_t represent the input, forget, cell, and output gates, respectively. The input and forget gates are used to control the flow of information across the cell gate, the input gate controls what information gets incorporated into the current cell state, while the forget gate allows the state to "reset" itself. The cell gate (also referred to as the block input) works with the input gate to help determine what new information is added to the cell state.

Finally, the output gate controls what information from the cell state incorporated into the hidden state.

4.5.2 Gated Recurrent Units (GRUs)

GRUs are a variant of LSTM whose success and popularity rival the vanilla LSTM. Proposed by Cho et al. in 2014, GRUs are similar in design to LSTMs but contain fewer gates and are therefore faster to train [78]. Both GRUs and LSTMs have been shown to outperform traditional RNNs; however, the question of which network is superior is ambiguous and may be dependent on the dataset at hand [79, 80]. The gates which comprise each unit of a GRU are defined as follows:

$$r_t = \sigma\left(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}\right) \tag{4.18}$$

$$z_t = \sigma\left(W_{iz}x_t + b_{iz} + W_{hz}h_{t-1} + b_{hz}\right) \tag{4.19}$$

$$n_t = \tanh\left(W_{in}x_t + b_{in} + r_t \odot \left(W_{hn}h_{t-1} + b_{hn}\right)\right) \tag{4.20}$$

$$h_t = \left(1 - z_t\right) \odot n_t + z_t \odot h_{t-1} \tag{4.21}$$

Equation (4.18) defines the reset gate, r_t. The reset gate is responsible for allowing the node to reset itself or forget some of the information from earlier hidden states. Equation (4.19) defines the update gate, z_t, which controls how much information from the previous hidden state, h_{t-1}, transfers current hidden state, h_t. The σ term in Eqs. (4.18) and (4.19) represents the sigmoid activation function. Equation (4.20) defines n_t—an intermediate state involved in the calculation of hidden state, h_t. As previously stated in Sect. 5.5.1, the $\odot$ symbol denotes the Hadamard product, which performs element-wise multiplication between two variables. A schematic diagram of an LSTM cell (a) and a GRU cell (b) is shown in Fig. 4.10.

4.6 Generative Adversarial Networks (GANs)

Generative networks represent a class of deep learning algorithms which seek to generate data samples which are representative of the training dataset. Popular generative networks include VAEs discussed in Sect. 4.3 and generative adversarial networks (GANs). GANs were first proposed by Goodfellow et al. in 2014 [81] and have since seen success in the generation of both images and videos. In medical imaging and radiation oncology, GANs have been applied to tasks such as the generating synthetic CT images from MRI scans [82], denoising low-dose CT images, [83], and reducing the sparsity of training data for a deep reinforcement treatment decision network by generating patient examples [35]. Although GANs have proven themselves to be a powerful tool for synthetic data generation, a major drawback is that they are notoriously difficult to train.

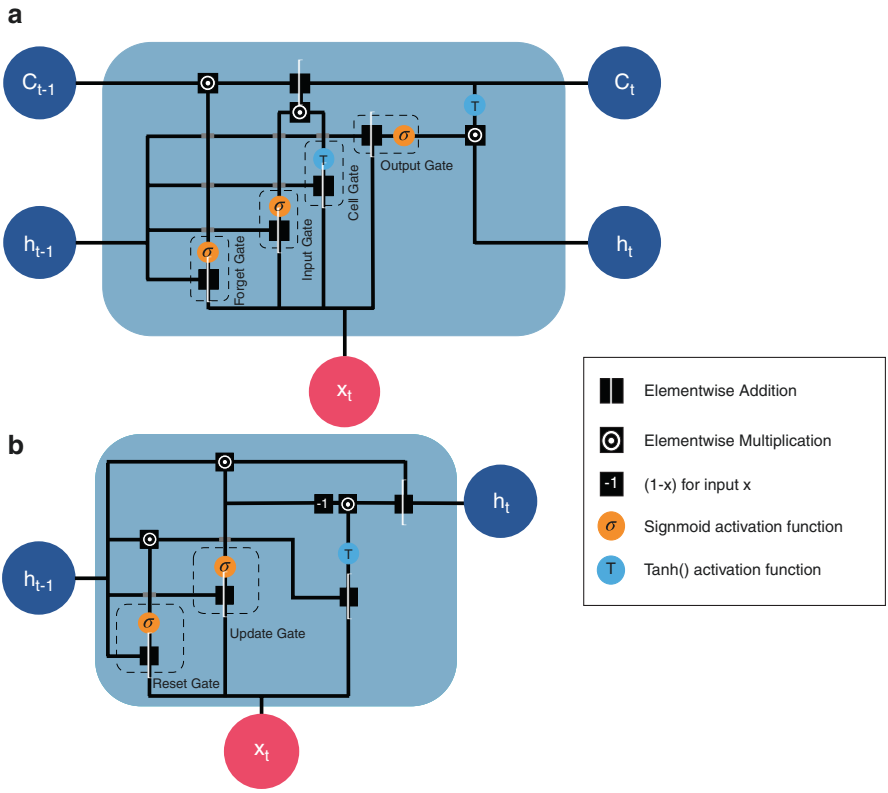


Fig. 4.10 Schematic diagram of an LSTM cell (**a**) and a GRU cell (**b**)

4.6.1 Vanilla GANs

GANs are composed of two networks that compete against each other: a generator, which generates "fake" data using random variables sampled from a user-selected distribution, and a discriminator, whose job is to differentiate between real data and fake data. Both networks are typically either feedforward networks or CNNs. The generator and discriminator are trained simultaneously under the framework of a two-player minimax game with a value function defined as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}(\log(D(\boldsymbol{x}))) + \mathbb{E}_{x\sim p_z(z)}\left(1-\log\left(D\left(G(\boldsymbol{z})\right)\right)\right) \tag{4.22}$$

During training, $V(D, G)$ is minimized with respect to a discriminator, D, while simultaneously maximizing the function with respect to a generator, G. The expression $D(\boldsymbol{x})$) represents the probability (according to the discriminator) that sample $\boldsymbol{x}$ came from the real dataset, while the expression $G(\boldsymbol{z})$ represents the mapping (by means of the generator) from a distribution $\boldsymbol{z}$ to a fake sample. The term p_{data}

represents the distribution of the data given by $\boldsymbol{x}$ but is not explicitly known by the user. The variable p_z represents the distribution (such as a uniform or Gaussian distribution) that $\mathbf{z}$ is randomly sampled from. As the discriminator learns to recognize real from fake data, the generator is simultaneously learning to generate better fakes. The action operation performed by the generator can be thought of as a mapping from the user-chosen distribution p_z to the model distribution, p_g. The goal of the generator is then to minimize the divergence between p_g and the data distribution, p_x. The optimal model is achieved when the discriminator predicts both real and generated samples to be genuine with equal probability (i.e., $D(\boldsymbol{x})) = D(G(\boldsymbol{z}))) = 0.5$) (Fig. 4.11).

4.6.2 Common GAN Variants: DCGAN, WGAN

Since their conception, there have been several variants on vanilla GANs. One such variant which is noteworthy is the Wasserstein GAN (WGAN), which was shown both in theory and practice to have more stable (and thus easier) training than the vanilla GAN [84]. WGAN achieves this improved performance by defining the distribution divergence to be minimized as an approximation of the earthmover (also known as the Wasserstein-1) distance, resulting in a value function:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}}\left[D(\boldsymbol{x})\right] - \mathbb{E}_{z\sim p_z}\left[D(G(\boldsymbol{z}))\right] \tag{4.23}$$

The earthmover distance is theoretically advantageous for training because under some weak assumptions it can be shown to be continuous everywhere and differentiable almost everywhere [84].

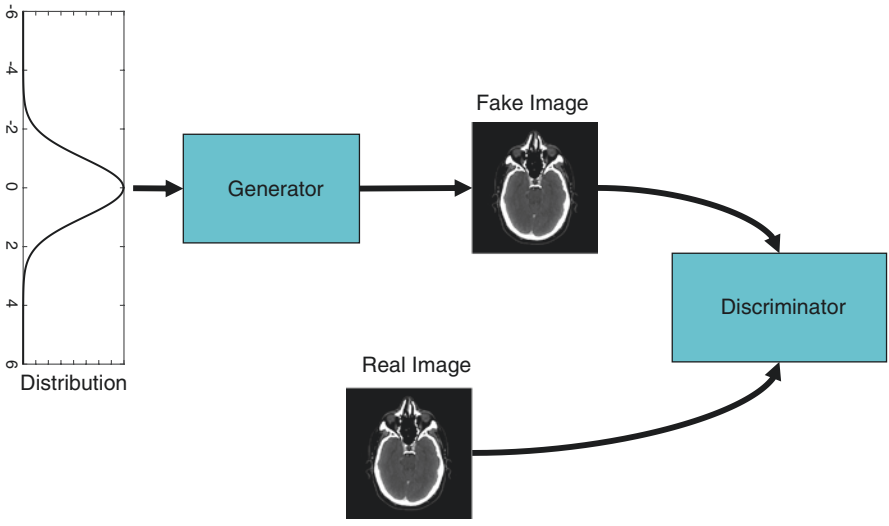


Fig. 4.11 Generative Adversarial Network (GAN)

Another type of GAN is the Deep Convolutional (DC) GAN [85]. The generator in a DCGAN is a CNN architecture. In addition, DCGAN allows for the generation of labeled data. This is performed by randomly assigning a label to the randomly sampled variables fed into the generator and teaching the discriminator to not only recognize fake data but to correctly assign labels to data.

4.7 Deep Reinforcement Learning (DRL)

DRL represents a combination of deep learning and reinforcement learning, which allows a network to perform decision-making tasks in an unsupervised manner. Reinforcement learning (RL) can be conceptualized as a type of machine learning in which an agent (i.e., an optimal search algorithm) learns to make a series of sequential decisions based on interactions with its environment (a Markov Decision Process [MDP]) in order to maximize a cumulative reward function. This method of learning based on external stimuli is interesting in its own right because it is similar to the way humans (or other biological agents) learn. DRL has been found to be particularly successful at performing a variety of decision-making tasks, managing to beat top human players in games such as poker, chess, and Go! [86]. A variant of DLR is deep Q-learning (DQN); In DQN, as the agent chooses the optimal action given the state of its environment by maximizing the Q-function, defined as the average discounted sum of rewards from the agent's current state up through all future steps. DQN has been shown to be capable of mimicking clinician dose adaption decisions in a radiotherapy artificial environment [35].

4.8 Current Challenges and Future Directions

Despite the promising role of deep learning in medicine and radiological sciences, there remain several challenges which must be addressed for the continued incorporation of deep learning applications into the clinical workflow. One challenge is the need for large, accessible datasets for training and validation of deep learning models. There is an active field of work in creating systems which allow for inter-institutional sharing of data while maintaining HIPAA and IRB compliance as well as for ensuring that data is curated or "farmed" by institutions in a standardized and easily usable fashion [87].

Another important challenge relates to the need for acceptance of deep learning methods by the greater medical community. A valid concern for the clinical use of deep learning models is that they are akin to black boxes—it is very difficult to interpret how a deep learning model "thinks" when it maps input data to an output prediction [88]. This property leaves predictions made by deep learning methods vulnerable to biases or errors which could harm the quality of patient care. In order

to gain buy-in from clinicians, further progress must be made to develop techniques, which allow medical staff to intuitively interpret the predictions made by deep learning models. Recent developments in deep learning visualization have already begun the process of opening the black box, particularly for image classification models. These deep learning visualization techniques can help users to understand what parts of an image contribute most to the network's final decision. For example, Grad-CAM is a technique which provides insight into the decision process for CNNs by creating a localization map which shows which regions of an input image are most important in determining the networks final prediction [89]. Saliency maps represent another method for visualizing classification CNNs; saliency maps use the derivative of the class score with respect to the image value at a given point to show which pixels in a given image have the largest impact on a given class prediction [90]. Another way to gain a deeper understanding of the inner workings of a deep neural network is to visualize the features that individual parts of the network have learned to identify. This can be achieved by attaching a deconvolutional neural network to each layer in a CNN to produce feature maps [91].

4.9 Conclusion

Deep machine learning methods have led to significant breakthroughs in the field of artificial intelligence. As large datasets become more readily accessible across different domains, deep learning will continue to contribute significantly to many fields. Deep learning technologies have already begun to make an impact in medicine, and data-heavy specialties such as radiology and radiation oncology are poised to be particularly impacted by these innovations. While any prediction of the future is bound to have a large degree of uncertainty, one can gain some insight into how deep learning might impact radiology and radiation oncology in the future from Amara's law: we tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run [92]. In other words, humans tend to have high expectations for emerging technologies, only to become disillusioned once the initial hype wears off and inherent limitations are more widely understood. At the time of this book's publication, it is the authors' belief that society is approaching the tail end of the deep learning hype cycle. With respect to the medical field, this observation is perhaps even more certain because despite early successes, the clinical implementation of deep learning methods still faces significant challenges with regard to interpretability and transparency—which are both necessary to ensure that these technologies follow the imperative to *first, do no harm* and are ethically accountable to the patients they serve. As deep learning continues on the arduous journey that is benchtop to bedside translation, it will become all too easy to grow pessimistic and underestimate its value. It therefore bears remembering that the future for deep learning is very bright indeed.

References

1. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6.
2. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
4. Sejnowski TJ. The unreasonable effectiveness of deep learning in artificial intelligence. Proc Natl Acad Sci. 2020;117:201907373.
5. Wang H, Raj B. On the origin of deep learning. In: arXiv preprint arXiv:170207800; 2017.
6. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25 (NIPS 2012). Red Hook, NY: Curran Associates; 2012.
7. Sejnowski TJ. The deep learning revolution. Cambridge, MA: MIT Press; 2018.
8. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol. 2020;93(1106):20190855.
9. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph. 2007;31(4-5):198–211.
10. Doi K, Giger ML, Nishikawa R, MacMahon H, Schmidt R. Artificial intelligence and neural networks in radiology: application to computer-aided diagnostic schemes. In: Hendee W, Trueblood J, editors. Digital imaging. 2: AAPM Medical Physics Monograph; 1993. p. 301–22.
11. Giger M, Huo Z, Kupinski M, Vyborny C. Computer-aided diagnosis in mammography. In: Sonka M, Fitzpatrick M, editors. Handbook of medical imaging, vol. 2. Bellingham, WA: SPIE; 2000. p. 915–1004.
12. Giger ML. Future of breast imaging. Computer-aided diagnosis. In: Haus A, Yaffe M, editors. AAPM/RSNA categorical course on the technical aspects of breast imaging; 1992. p. 257–70.
13. Giger ML. Computer-aided diagnosis in radiology. Acad Radiol. 2002;9(1):1–3.
14. Swett H, Giger M, Doi K. Computer vision and decision support. In: Hendee W, Wells P, editors. Perception of visual information. Berlin: Springer-Verlag; 1993. p. 272–315.
15. Wu Y, Doi K, Giger ML, Nishikawa RM. Computerized detection of clustered microcalcifications in digital mammograms: applications of artificial neural networks. Med Phys. 1992;19(3):555–60.
16. Chan HP, Lo SC, Sahiner B, Lam KL, Helvie MA. Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network. Med Phys. 1995;22(10):1555–67.
17. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, et al. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging. 1996;15(5):598–610.
18. Zhang W, Doi K, Giger ML, Wu Y, Nishikawa RM, Schmidt RA. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. Med Phys. 1994;21(4):517–24.
19. Chen W, Giger ML, Newstead GM, Bick U, Jansen SA, Li H, et al. Computerized assessment of breast lesion malignancy using DCE-MRI robustness study on two independent clinical datasets from two manufacturers. Acad Radiol. 2010;17(7):822–9.
20. Chen W, Giger ML, Bick U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. Acad Radiol. 2006;13(1):63–72.
21. Bhooshan N, Giger ML, Jansen SA, Li H, Lan L, Newstead GM. Cancerous breast lesions on dynamic contrast-enhanced MR images: computerized characterization for image-based prognostic markers. Radiology. 2010;254(3):680–90.
22. Yuan Y, Giger ML, Li H, Bhooshan N, Sennett CA. Multimodality computer-aided breast cancer diagnosis with FFDM and DCE-MRI. Acad Radiol. 2010;17(9):1158–67.

23. https://www.prnewswire.com/news-releases/quantitative-insights-gains-industrys-first-fda-clearance-for-machine-learning-driven-cancer-diagnosis-300495405.html.
24. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys. 2019;29(2):102–27.
25. Akagi M, Nakamura Y, Higaki T, Narita K, Honda Y, Zhou J, et al. Deep learning reconstruction improves image quality of abdominal ultra-high-resolution CT. Eur Radiol. 2019;29(11):6163–71.
26. Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, et al. Deep learning in medical ultrasound analysis: a review. Engineering. 2019;5(2):261–75.
27. Graffy PM, Sandfort V, Summers RM, Pickhardt PJ. Automated liver fat quantification at nonenhanced abdominal CT for population-based steatosis assessment. Radiology. 2019;293(2):334–42.
28. Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. In: arXiv e-prints [Internet]; 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190302026H. Accessed 1 Mar 2019.
29. Giger ML. Machine learning in medical imaging. J Am Coll Radiol. 2018;15(3 Pt B): 512–20.
30. Fan J, Wang J, Chen Z, Hu C, Zhang Z, Hu W. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. Med Phys. 2019;46(1):370–81.
31. Liu F, Yadav P, Baschnagel AM, McMillan AB. MR-based treatment planning in radiation therapy using a deep learning approach. J Appl Clin Med Phys. 2019;20(3):105–14.
32. Shen C, Gonzalez Y, Klages P, Qin N, Jung H, Chen L, et al. Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. Phys Med Biol. 2019;64(11):115013.
33. Liu Y, Lei Y, Wang T, Kayode O, Tian S, Liu T, et al. MRI-based treatment planning for liver stereotactic body radiotherapy: validation of a deep learning-based synthetic CT generation method. Br J Radiol. 2019;92(1100):20190067.
34. Elmahdy MS, Jagt T, Zinkstok RT, Qiao Y, Shahzad R, Sokooti H, et al. Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer. Med Phys. 2019;46(8):3329–43.
35. Tseng H-H, Luo Y, Cui S, Chien J-T, Ten Haken RK, Naqa IE. Deep reinforcement learning for automated radiation adaptation in lung cancer. Med Phys. 2017;44(12):6690–705.
36. Chun J, Zhang H, Gach HM, Olberg S, Mazur T, Green O, et al. MRI super-resolution reconstruction for MRI-guided adaptive radiotherapy using cascaded deep learning: In the presence of limited training data and unknown translation model. Med Phys. 2019;46(9):4148–64.
37. Kurz C, Maspero M, Savenije MHF, Landry G, Kamp F, Pinto M, et al. CBCT correction using a cycle-consistent generative adversarial network and unpaired training to enable photon and proton dose calculation. Phys Med Biol. 2019;64(22):225004.
38. Huang P, Yu G, Lu H, Liu D, Xing L, Yin Y, et al. Attention-aware fully convolutional neural network with convolutional long short-term memory network for ultrasound-based motion tracking. Med Phys. 2019;46(5):2275–85.
39. Nyflot MJ, Thammasorn P, Wootton LS, Ford EC, Chaovalitwongse WA. Deep learning for patient-specific quality assurance: identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks. Med Phys. 2019;46(2):456–64.
40. Tomori S, Kadoya N, Takayama Y, Kajikawa T, Shima K, Narazaki K, et al. A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. Med Phys. 2018;45(9):4055–65.
41. Galib SM, Lee HK, Guy CL, Riblett MJ, Hugo GD. A fast and scalable method for quality assurance of deformable image registration on lung CT scans using convolutional neural networks. Med Phys. 2020;47(1):99–109.
42. Kimura Y, Kadoya N, Tomori S, Oku Y, Jingu K. Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy. Phys Med. 2020;73:57–64.

43. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med. 2018;15(11):e1002711.
44. Wei L, Osman S, Hatt M, El Naqa I. Machine learning for radiomics-based multimodality and multiparametric modeling. Q J Nucl Med Mol Imaging. 2019;63(4):323–38.
45. Bibault J-E, Giraud P, Housset M, Durdux C, Taieb J, Berger A, et al. Deep learning and radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Sci Rep. 8:12611.
46. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, Mak RH, Aerts HJWL. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. Clin Cancer Res. 2019;25(11):3266–75. https://doi.org/10.1158/1078-0432.CCR-18-2495. Epub 2019 Apr 22. PMID: 31010833; PMCID: PMC6548658.
47. Shen W-C, Chen S-W, Wu K-C, Hsieh T-C, Liang J-A, Hung Y-C, et al. Prediction of local relapse and distant metastasis in patients with definitive chemoradiotherapy-treated cervical cancer by deep learning from [18F]-fluorodeoxyglucose positron emission tomography/computed tomography. Eur Radiol. 2019;29(12):6741–9.
48. Cui S, Luo Y, Hsin Tseng H, Ten Haken RK, El Naqa I. Artificial neural network with composite architectures for prediction of local control in radiotherapy. IEEE Trans Radiat Plasma Med Sci. 2019;3(2):242–9.
49. Cui S, Luo Y, Tseng H-H, Ten Haken RK, El Naqa I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. Med Phys. 2019;46(5):2497–511.
50. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys. 2019;46(1):e1–e36.
51. Bibault J-E, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: state of the art and future prospects. Cancer Lett. 2016;382(1):110–7.
52. Boldrini L, Bibault J-E, Masciocchi C, Shen Y, Bittner M-I. Deep learning: a review for the radiation oncologist. Front Oncol. 2019;9:977.
53. Tseng H-H, Luo Y, Ten Haken RK, El Naqa I. The role of machine learning in knowledge-based response-adapted radiotherapy. Front Oncol. 2018;8:266.
54. Hutter F, Lücke J, Schmidt-Thieme L. Beyond manual tuning of hyperparameters. Künstliche Intelligenz. 2015;29(4):329–37.
55. Luo G. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Network modeling analysis in health informatics and bioinformatics. 2016;5(1):18.
56. Xavier G, Antoine B, Yoshua B. Deep sparse rectifier neural networks. PMLR; 2011. p. 315–23. Accessed 14 Jun 2011.
57. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics; 2010.
58. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: International Conference on Computer Vision; 2015. p. 1026–34.
59. Nielsen MA. Neural networks and deep learning. San Francisco, CA: Determination Press; 2015.
60. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: arXiv preprint arXiv: 150203167; 2015.
61. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: arXiv preprint arXiv: 14126980; 2014.
62. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.
63. Hornik K. Approximation capabilities of multilayer feedforward networks. Neural Netw. 1991;4(2):251–7.
64. Leshno M, Lin VY, Pinkus A, Schocken S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw. 1993;6(6):861–7.

65. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
66. Bengio Y. Learning deep architectures for AI. Delft: Now Publishers; 2009.
67. Cui S, Tseng H-H, Pakela J, Ten Haken RK, El Naqa I. Introduction to machine and deep learning for medical physicists. Med Phys. 2020;47(5):e127–e47.
68. Kingma DP, Welling M. Auto-encoding variational bayes. In: arXiv preprint arXiv:13126114; 2013.
69. Fukushima K, Miyake S. Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets. Berlin: Springer; 1982. p. 267–85.
70. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al.. Pytorch: an imperative style, high-performance deep learning library. 2019.
71. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. 2016.
72. Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning. In: arXiv preprint arXiv:160307285; 2016.
73. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: MIT Press; 2016.
74. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertain Fuzz Knowl Based Syst. 1998;6(02):107–16.
75. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
76. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: a search space odyssey. IEEE Trans Neural Netw Learn Syst. 2016;28(10):2222–32.
77. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 2005;18(5):602–10.
78. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: arXiv preprint arXiv:14061078; 2014.
79. Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures. In: International conference on machine learning; 2015. p. 2342–50.
80. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: arXiv e-prints. arXiv:1412.3555; 2014.
81. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: arXiv:1406.2661; 2014.
82. Kazemifar S, Barragán Montero AM, Souris K, Rivas ST, Timmerman R, Park YK, et al. Dosimetric evaluation of synthetic CT generated with GANs for MRI-only proton therapy treatment planning of brain tumors. J Appl Clin Med Phys. 2020;21(5):76–86.
83. Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans Med Imaging. 2018;37(6):1348–57.
84. Arjovsky M, Chintala S, Bottou L. Wasserstein gan. In: arXiv preprint arXiv:170107875; 2017.
85. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: arXiv preprint arXiv:151106434; 2015.
86. François-Lavet V, Henderson P, Islam R, Bellemare MG, Pineau J. An introduction to deep reinforcement learning. In: arXiv preprint arXiv:181112560; 2018.
87. Mayo CS, Kessler ML, Eisbruch A, Weyburne G, Feng M, Hayman JA, et al. The big data effort in radiation oncology: data mining or data farming? Adv Radiat Oncol. 2016;1(4):260–71.
88. El Naqa I, Ruan D, Valdes G, Dekker A, McNutt T, Ge Y, et al. Machine learning and modeling: data, validation, communication challenges. Med Phys. 2018;45(10):e834–e40.
89. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. Int J Comput Vision. 2017;2:336–59.
90. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. In: arXiv preprint arXiv:13126034; 2013.
91. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Berlin: Springer; 2014.
92. Coates JF, Jarratt J. What futurists believe. Bethesda, MD: Lomond; 1989.

Quantum Computing for Machine Learning

5

Dipesh Niraula, Jamalina Jamaluddin, Julia Pakela, and Issam El Naqa

5.1 Introduction

Quantum mechanics is arguably among the most influential inventions of the twentieth century in physics. Advancement of this fundamental theory and its application have led to better understanding of the laws of nature that govern our surrounding universe as well as the invention of many modern electronic devices such as transistors, electron microscopes, lasers, magnetic resonance imaging, and LEDs. The invention of the transistor single handedly revolutionized modern industrial and consumer technologies by significantly downsizing their footprint, thus reducing their power requirements. This is particularly true in the case of personal computers, which were originally powered by cumbersome vacuum tubes. Further advancements of transistor technology, from point-contact transistors to integrated circuits to very large-scale integrated processes, gradually made computers cheaper, faster, and computationally more powerful, versatile, and fit for general purpose applications making them an indispensable component of today's modern society.

D. Niraula (✉) · I. El Naqa
Department of Machine Learning, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA
e-mail: Dipesh.Niraula@moffitt.org; Issam.ElNaqa@moffitt.org

J. Jamaluddin
Department of Nuclear Engineering and Radiological Sciences, University of Michigan, Ann Arbor, MI, USA
e-mail: jamalina@umich.edu

J. Pakela
Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: jpakela@med.umich.edu

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_5

Naturally, software technology has advanced tremendously along with the hardware components. In 1936, Alan Turing devised a conceptual cognitive system called the Turing machine formally introducing a mathematical model of cognitive computation which aided in the theoretical development of modern artificial intelligence in computer science. Right after, Claude Shannon implemented Boolean logic (binary) in programming in 1937 as a primitive electromagnetic computer and thereby showing that a machine could operate merely with "0"s and "1"s [1]. Software tools such as compilers and interpreters, which translates high-level programming language to machine level binary language, led to a rapid advancement in software technology by making it possible to program a computer without having to understand the minute hardware details. This separation of duty simplified the task of software development and programming which transformed the world into a digital era.

Current commercial (mainstream) computing paradigm is based on deterministic classical Boolean logic. Even though hardware technology extensively may apply quantum mechanical principles for its advancement, the computational process itself is purely classical and is based on binary logic. To better harness the power of quantum mechanics at the computational level, computer scientists and physicists have been actively working on merging quantum mechanics, information theory, and computation to create a new computing paradigm known as quantum computing [2, 3]. The original inspiration for this concept was presented by notable physicist and Nobel prize laureate Richard Feynman in the 1970s. He postulated that a quantum computer can solve problems that classical computers cannot. This idea followed a proposal by Paul Benioff of a quantum mechanical model of the Turing machine [4].

This new quantum computing paradigm is vastly different from the current classical one at both the software and hardware levels. Unlike the basic unit of classical information/computation, a bit, which can either be a "0" or an "1," its quantum counterpart, a qubit, can exist in a mixed state; a qubit can be represented by the linear superposition as,

$$\left\langle |\psi| \right\rangle = c_0 |0\rangle + c_1 |1\rangle, \tag{5.1}$$

where the probability amplitude c_0 and c_1 are complex numbers and thus can take infinitely many values. Only upon an event, i.e., a quantum measurement, $|\psi\rangle$ will collapse to either $|0\rangle$ or $|1\rangle$ binary quantum state with a probability of $|c_0|^2$ or $|c_1|^2$, respectively. By extension, an n qubit register can represent 2^n states simultaneously. For instance, for a 2-qubit system, a uniform superposition state, $|\psi = \frac{1}{2}\left(|00 + |01 + |10 + |11\right)$ represents four states simultaneously. This property of quantum state is known as *quantum superposition* which further allows for *quantum parallelism* providing quantum computing with a much faster computational speed than its classical counterpart.

Quantum parallelism is the ability to simultaneously operate on all the superimposed state in parallel. Since the number of states represented by qubit register grows exponentially with the number of registers, an operation performed on a quantum computer would have taken an exponentially large of operations on a classical computer with the same numbers of registers. One trade-off of quantum

parallelism is that the probability of measuring one particular state out of 2^n states also decreases exponentially. The main challenge of any quantum algorithm is to overcome this, which will be discussed in Sect. 5.4 in the context of common quantum algorithms devised by Peter Shor [5] and Grover [6].

Another quantum mechanical phenomenon relevant to quantum computing is quantum entanglement. For instance an entangled Bell State: $\frac{1}{\sqrt{2}}\left(|00\rangle + |11\rangle\right)$, has a 50% chance of being measured in either $|00\rangle$ state and 50% chance of being measured in either $|11\rangle$ state. Moreover, the two qubits are correlated, i.e., if the first qubit is measured to be in $|1\rangle$ state, the second qubit will also collapse in $|1\rangle$ state. Entanglement phenomenon provides quantum computing with a resource absent in classical computing that makes classically impossible processes, like super-dense coding and quantum teleportation, possible. An actively sought out real-life application of quantum communication is to create an eavesdrop-proof channel for key distribution in cryptography [7]. In the world of computing, this translates to an efficient representation of highly correlated information, which is still a challenge in classical computation.

While quantum algorithm and quatnum information hold promises for several improvements, quantum computers are the devices necessary to make those theoretical gains a reality. Thus there has been a growing interest in research and developvent of quantum hardwares and quantum computers from both the private sector and the government [8–10]. A quantum computer is built from quantum circuit and quantum gates, just like classical computers are made up of electrical circuits and logic gates. Physical realization of quantum circuit has been achieved through different means: trapped ions [11], nuclear magnetic resonance (NMR), and linear optical systems, and superconducting solid-state system [12]. Currently, superconducting solid-state system is the main focus in commercial sectors and companies, with Google, IBM, Microsoft, and Intel racing to build a universal quantum computer, whereas NMR, optical and trap ion computing systems are most actively researched in academic and governmental institutes.

In the era of big data analytic and machine learning, an emerging hybrid field capturing the best of quantum computing and machine learning has been generating novel ideas such as quantum neural networks [13], quantum convolutional neural network [14], quantum generative adversarial networks [15], and quantum reinforcement learning [16, 17]. The deep learning half of such hybrid models equip these tools with the ability to represent complex data patterns while the quantum information half makes them faster and less prone to noisy fluctuations in data.

Although quantum technology is still in its infancy, it has the potential to bring about a new era in the world of computing, and by extension a new era in machine learning. This chapter provides a high-level overview of quantum computing, including a summary of the current state of quantum hardware development, a review of the two major quantum computing algorithms which promise quantum supremacy over classical computers—the Shor's algorithm and the Grover's

algorithm, a review of the emerging field of quantum machine learning—including quantum deep neural networks and quantum reinforcement learning, and finally, a review of current applications of quantum computing in medical physics at the date of this publication.

5.2 Postulates of Quantum Mechanics

Some working knowledge of quantum mechanics and linear algebra may be necessary to follow through the rest of this chapter. To facilitate this, six postulates that concisely cover the fundamentals of quantum mechanics are reviewed in this section. The origin of these postulates is largely based on Dirac–von Neumann axioms, functional analysis, and is part of the standard quantum physics curriculum. Furthermore, descriptions of mathematical tools necessary to understand the postulates are subsequently presented. For a basic understanding of linear algebra, the readers may refer to [18].

1. The state of a quantum mechanical system is completely represented by a normalized ket $|\psi\rangle$.
   - Mathematically, kets are vectors that reside in an inner-product vector space over complex number field called Hilbert Space. While Hilbert Space can represent infinite dimensional space, the dimensionality of the Hilbert space is determined by the physics of a problem.
   - Vector spaces satisfy the closure property, i.e., the linear combination of any two vectors of a vector space must lie in the same vector space. This implies that the superposition of two states is also a state of the system: if $|\psi_0\rangle$ and $|\psi_1\rangle$ are two possible states of a system, then so is $|\psi\rangle = c_0|\psi_0\rangle + c_1|\psi_1\rangle$, where c_0 and c_1 are complex numbers.
   - In Dirac notation, "bra" ($\langle|$) is the dual of vector "ket" ($|\rangle$) and together in the bra-ket succession defines inner product. Mathematically, "bra" denotes a linear functional $f : V \rightarrow \mathrm{C}$ that maps vector "ket" to a number in the complex plane. The Dirac notation simplifies the inner product notation especially for states of continuous variables, which are represented in terms of wave functions, i.e.,

$$\langle|\psi|\phi\rangle = \int \mathrm{d}x\psi^*(x)\phi(x).$$

2. Every physical observable attribute of a quantum mechanical system is described by a Hermitian operator $\hat{O}$ that acts on kets describing that system.
   - Observables are physically measurable attributes such as position, linear and angular momentum, energy, and so on.
   - An operator is Hermitian if it is equal to its conjugate transpose, i.e., $O = O^{\dagger}$. This property guarantees real eigenvalues necessary for the observables to be physical.

- An operator $\hat{O}$ acting on a ket $|\psi\rangle$ is denoted by left multiplication, i.e., $|\psi' = \hat{O}|\psi$. In general, the operation changes the state of the quantum system.

3. The only possible result of a measurement of an observable O is one of the eigenvalues of the corresponding operator $\hat{O}$.
   - This postulate describes the source of the word "quantum." If the observable is of continuous spectrum, like position or momentum, then the measurement will give us classical results. However, if the observable has a discrete spectrum, like the angular momentum of an orbiting electron, the measurement will yield discrete set of values in multiples of Planck's constant.
4. Upon a measurement of the observable O on a quantum mechanical system in the state $|\psi\rangle$, the probability of obtaining the eigenvalue o_n is given by the square of the inner product of $|\psi\rangle$ with the eigenstate $|o_n\rangle$, i.e., $|\langle o_n|\psi\rangle|^2$.
   - Besides having real eigenvalues, the eigenstates of a Hermitian operator are orthogonal, i.e., $\langle o_i|o_j\rangle = \delta_{ij}$, where the Kronecker delta $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ otherwise. The orthogonal eigenstate spans the space of states and forms a basis. This means that the state of a quantum mechanical system can be expanded as a linear combination of eigenstate of a Hermitian operator with complex coefficients.

$$|\psi\rangle = \sum_n c_n |o_n\rangle.$$

   - The complex coefficients, $\langle o_n|\psi\rangle = c_n$, represent the "probability amplitude" associated with the eigenstate $|o_n\rangle$. This means that $|\psi\rangle = \sum_n \langle o_n|\psi\rangle\ |o_n\rangle$ and hence $\langle\psi|\psi\rangle = \sum_n |\langle o_n|\psi\rangle|^2$. Since $\langle\psi|\psi\rangle$ is the total probability and sums to 1, $|\langle o_n|\psi\rangle|^2$ are the individual probability associated with measuring eigenstates o_n.
   - The probabilistic interpretation requires the normalization of kets, i.e., $\langle\psi|\psi\rangle = 1$, which can be achieved by dividing $|\psi'\rangle$ by its norm, $\sqrt{\langle\psi'|\psi'\rangle}$, where prime notation represents the state before normalization.
5. Immediately after the measurement of the observable O that yielded the value o_n, the state of the system collapses to the normalized eigenstate $|o_n\rangle$.
   - This postulate is "counterintuitive" with respect to quantum mechanics. Measurements of observable O after preparing several identical quantum mechanical system in state $|\psi\rangle$ can yield different results. This is to be expected as quantum mechanical states are probabilistic in nature and can exist in superimposed state. However, carrying out a second measurement immediately on a system that yielded o_n value in the first measurement will always yield o_n.
6. The time evolution of a quantum mechanical system preserves the normalization of the associated ket. The time evolution of the state is a unitary transformation described by $|\psi(t)\rangle = \hat{U}(t,t_0)|\psi(t_0)\rangle$.
   - The preservation of normalization during time evolution of a quantum mechanical system implies conservation of probability: the probability of finding a system in an eigenstate summed over all possible eigenstate must be

1. This is ensured by setting the time evolution operator as a unitary transformation.
- Mathematically, unitary transformations preserve inner products: the inner product of kets are equal to the inner products of the transformed kets. i.e., $\langle\psi|\psi\rangle = (\langle\psi|U^\dagger)(U|\psi\rangle)$ which is true when $UU^\dagger = I$, implying $U^\dagger = U^{-1}$. By symmetry of the inner product operations, $U^\dagger U = I = UU^\dagger$.
- Practically, unitary transformations are reversible operations such as rotation and reflection, i.e., an inverse transformation of equal magnitude will nullify the forward transformation and send a state back to its original state. All physical quantum gates used in quantum circuit are unitary operators.

5.3 Quantum Hardware

Quantum hardware are devices that can implement one of the several quantum objects and/or phenomena that exists in nature. For instance, optical quantum devices use polarization of a photon as a quantum state with mirrors, beam splitters, phase shifters, and interferometers as quantum gates. Use of photons are advantageous for its stability, yet the lack of photon to photon interaction is a major drawback. Similarly, trapped ion quantum computing devices are based on trapping ions by creating a saddle shaped electric potential well via an oscillating field. The quantum state is the atomic spin (magnet), manipulated by shining laser onto the ion. The most popular and promising quantum objects currently are the superconductors. Superconducting quantum computer has been actively developed by tech giants such as IBM, Google, and Intel in pursuit of creating a universal quantum computer. In a superconductor, pairs of electrons, known as *Cooper pairs*, act as the basic charge carrier. Each pair has an integer number spin associated, therefore acting as a boson in effect. Qubits are then defined as either the phase, charge, or the flux of the pair.

Aside from the underlying quantum object, quantum hardware can be categorized in terms of their functionality. Optical quantum devices, for instance, are being developed for quantum communication, while superconducting quantum computing devices are more for general purpose. Radiation oncology, in particular, has found a use of quantum annealers and universal quantum computing (simulators) to tackle various optimization problems such as inverse treatment planning problems [19].

5.3.1 Quantum Annealers

Quantum annealers refer to a class of quantum hardware which solve optimization problems through the process of quantum annealing. First proposed by Kadowaki and Nishimori [20], quantum annealing exploits the result of the quantum adiabatic theorem, which states that when a quantum system undergoes a gradual change in its total energy from H_i to H_f, if it starts out in the nth eigenstate of H_i, it will end up in the corresponding nth eigenstate of H_f. This means that a quantum system can be constructed to start in the ground state of a known (or solvable) objective function,

and then gradually shifted into the ground state of the objective function of interest using an annealing coefficient, $\mathcal{T}(t)$:

$$H(t) = H_{\mathrm{f}} + \mathcal{T}(t) H_i.$$

At the start of the optimization process, the annealing variable $\tau(t)$ is very large such that $H \simeq \mathcal{T}(t) H_i$. By the end of the optimization process, $\mathcal{T}(t)$ approaches 0 and the system is now in the lowest energy eigenstate of the Hamiltonian defined by the objective function of interest, $H_{(t_{\mathrm{end}})} = H_{\mathrm{f}}$. Thus, similar to simulated annealing, quantum annealing is guaranteed to find the global optimal solution if allowed to run long enough, and theoretical and experimental results suggest that quantum annealing boasts performance benefits over its classical counterpart, simulated annealing [21, 22].

Initial studies on quantum annealing were performed as simulations on classical computers [23]; however, classical simulation of this quantum process is computationally expensive and therefore impractical for higher order objective functions—such as those which would be necessary for many real-world problems. Quantum annealers are devices which use quantum hardware to physically realize the quantum annealing process. The major developer of quantum annealers to date is DWave INC. DWave quantum annealers use qubits which are made up of superconducting loops of current—the direction of the resulting magnetic field (down vs. up) defines the classical binary states of the qubit (0 vs. 1), but as quantum objects they can also exist in both states simultaneously. As of this time DWave hardware supports a maximum of 2000 qubits, which can represent 2^{2000} unique states.

5.3.2 Universal Quantum Computers

Universal quantum computers are analogous to the Turing's universal machine. For a computer to be universal, it must be capable of taking any arbitrary set of inputs or instructions and convert them to an arbitrary set of corresponding outputs. This requirement is easily satisfied by classical computing, though not as straightforward in quantum computing. In the quantum world, not every observable commute: observable A and B commute if the order in which the operations are carried out does not affect the result, i.e., $AB - BA = 0$. For quantum computer to be universal, it must be able to handle non-commuting observable.

For instance as in [24], consider implementing the unitary operation, $\hat{U} = e^{i(aX+bZ)}$, where X and Z are some non-commuting observable. Note that any observable A (Hermitian) can be associated to a unitary operation by $U = e^{(iA)}$. Since X and Z do not commute, $e^{i(aX+bZ)} \neq e^{iaX} e^{ibZ}$. However, by breaking down the operation in n small slices, we can use the approximation $\hat{U} = \lim_{n\to\infty} \left(e^{\frac{iaX}{n}} e^{\frac{ibZ}{n}} \right)^n$. This yields an error that scales as $1/n^2$. Combining n slices, target U is approximated with an error that scales as $1/n$. So by increasing the number of slices, it is possible to get as close to U as needed. As long as enough slices are maintained, i.e. divide the task in hand into enough small operations, universality of the circuit-based quantum computer can be gaurenteed.

Currently commercial sectors including IBM, Google, Intel, and Microsoft are heavily invested in creating a general purpose universal quantum computer power enough to carry out practical tasks. State-of-the-art universal quantum computers have crossed the 50 plus qubit marker and will be more powerful in the future with speculations of orders of magnitude increase in the next few years. An important challenge currently faced in quantum computer development is the ability to maintain a useful number of qubits in coherence long enough to perform the calculations of interest. It is therefore important to note that at this time, many of the quantum algorithms proposed in literature are tested by simulating quantum states on classical computers rather than through an actual quantum computing hardware. As long as enough slices are maintained, i.e. divide the task in hand into enough small operations, universality of the circuit-based quantum computer can be gaurenteed.

5.4 Common Quantum Computing Algorithms

For a quantum algorithm to be useful, it has to at least theoretically outperform the best classical algorithm given their inherent complexity.

5.4.1 Grover's Algorithm

Grover's algorithm is a quantum search algorithm designed to carry out unstructured search. It can speed up an unstructured search problem quadratically faster than any classical algorithm and its probability amplitude amplification subroutine is applicable beyond a mere search problem [24]; in later section we utilize Grover iteration as an optimization procedure.

Unstructured search problem is a task to locate a marked item m from a large list of N items. On average, a classical algorithm takes $N/2$ steps to locate m and in the worst case scenario takes N steps. With the probability amplification trick, Grover's algorithm can find m in about $\sqrt{N}$ steps, providing a quadratic speedup. For instance, for a list of 300 million unstructured items, classical search algorithms on average take 150 million operations, while Grover's algorithm will take only about 17 thousands. Furthermore, it does not utilize the internal structure of the items, making it a generic algorithm, also commonly known as a black box.

Before carrying out Grover iteration, N items are represented by a uniformly distributed 2^n superimposed quantum states where n is the number of qubits. This process requires preparing quantum states, usually prepared in $|0\rangle$ state, which is then passed through a Hadamard gate to create a superimposed state. Mathematically, a Hadamard gate is a unitary operation that transforms $|0\rangle$ to $\frac{|0\rangle+|1\rangle}{\sqrt{2}}$. In matrix representation, $H=\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$ and $|0\rangle=\begin{pmatrix}1\\ 0\end{pmatrix}$. For n qubits, n Hadamard gates are required to create 2^n superimposed states. For example, in a three qubit system the eight superimposed states are

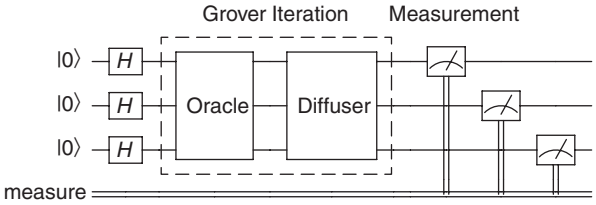


Fig. 5.1 Schematic of a Grover amplification procedure on a three qubit quantum circuit. Before carrying out the Grover amplification, the qubits are created in $|0\rangle$ state which are then passed through Hadamard gates, creating a uniform superimposed state. One Grover iteration is composed of a phase oracle and a diffuser. Quantum measurement is carried out at the end to obtain the result

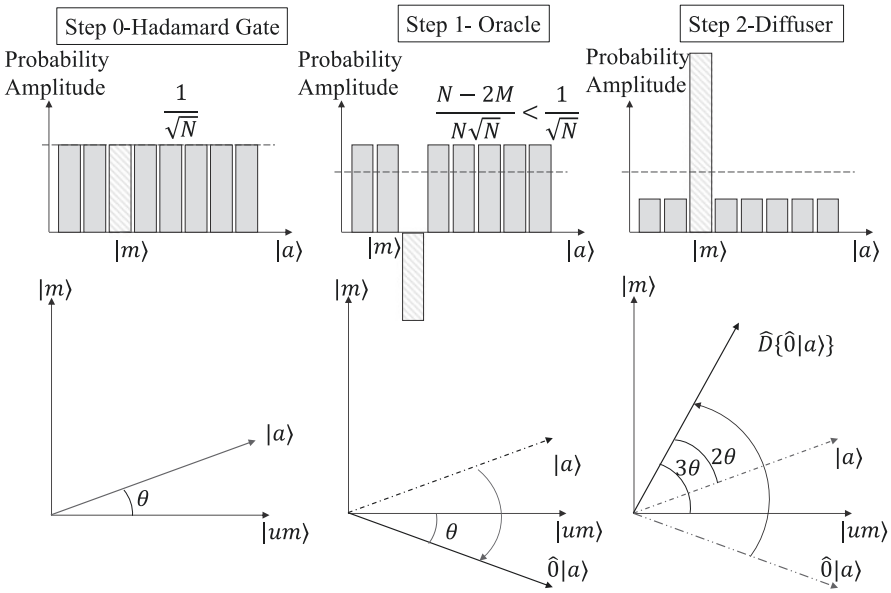


Fig. 5.2 Visualization of Grover iteration via probability amplitude graph (top row) and geometrical interpretation (bottom row) in the plane spanned by marked ($|m\rangle$) and unmarked state ($|um\rangle$). Phase oracle, $\hat{O}$, flips the probability amplitude sign of $|m\rangle$ and Diffuser, $\hat{D}$, inverts the superimposed state $|a\rangle$ about the mean $N-2M/N\sqrt{N}$ amplifying $|m\rangle$. Geometrically, $\hat{O}$ is a reflection across $|um\rangle$ and $\hat{D}$ is a reflection across $|a\rangle$. Collectively, they result in a rotation of $|a\rangle$ by an odd number multiple of the angle subtended by $|a\rangle$ and $|um\rangle$

$$|a\rangle = \frac{1}{\sqrt{8}}\left(|000\rangle + |001\rangle + |010\rangle + |011\rangle + |100\rangle + |101\rangle + |110\rangle + |111\rangle\right). \quad (5.2)$$

Grover iteration comprises a series of two operations, i.e., $\hat{G} = \hat{D}\hat{O}$, where $\hat{O}$ and $\hat{D}$ stand for phase oracle and diffuser operation. Figure 5.1 presents a schematic of one Grover iteration in a three qubit quantum circuit. In practice, a series of

Grover iteration are carried out as permitted prior to measurement. This strengthens the chance of obtaining the marked item from a measurement.

The Phase Oracle operator $\hat{O}$ flips the probability amplitude sign of the marked state, $|m\rangle$, and then Diffuser operator, $\hat{D}$, also known as inversion about mean, inverts the whole superimposed state about the mean of the inverted marked state and remaining unmarked states $|\text{um}\rangle$ as shown in Fig. 5.2. The mean on the probability amplitude, $1/\sqrt{N}$, decreases to $(N-2M)/N\sqrt{N}$ after the phase oracle operation; M is the number of marked states. Since the mean lies closer to the unmarked states, the inversion amplifies $|m\rangle$ while diminishing $|\text{um}\rangle$.

Mathematically, Phase Oracle $\hat{O} = \hat{I} - 2|m\rangle\langle m|$ and Diffuser $\hat{D} = 2|a\rangle\langle a| - \hat{I}$. Suppose the third state, $|010\rangle$, of the superimposed three qubit system of Eq. 5.2 is the marked state, then the operators and transformations from the operations are as follows:

$$\hat{O}|a\rangle = |a\rangle - 2|m\rangle\langle m|a\rangle$$

$$= \frac{1}{\sqrt{8}}\left(|000\rangle + |001\rangle - |010\rangle + |011\rangle + |100\rangle + |101\rangle + |110\rangle + |111\rangle\right)$$

$$\hat{D}\left\{\hat{O}|a\rangle\right\} = \left(2|a\rangle\langle a| - \hat{I}\right)\left(|a\rangle - 2|m\rangle\langle m|a\rangle\right) = \frac{1}{2}|a\rangle + \frac{2}{\sqrt{8}}|m\rangle$$

$$= \frac{1}{2\sqrt{8}}\left(|000\rangle + |001\rangle + 5|010\rangle + |011\rangle + |100\rangle + |101\rangle + |110\rangle + |111\rangle\right).$$

Since both of the operators are unitary, the total probability after each transformation is 1. The probability amplitude of the marked state $|010\rangle$ flips after the first operation and then amplifies to five times as much as the unmarked states. After one Grover iteration, the chances of obtaining the marked state $|010\rangle$ from a quantum measurement is $\left|5/2\sqrt{8}\right|^2 = 78.1\%$. This chance can be increased with repetition of Grover iteration, though there is a limit which when crossed, will start behaving erratically. The limit can be determined by geometrical analysis of the operators.

Geometrically, Grover iteration can be understood in terms of two unitary transformations: reflection and rotation, in the plane spanned by the basis $|m\rangle$ and $|\text{um}\rangle$; $|\text{um}\rangle$ can be constructed by removing $|m\rangle$ from $|a\rangle$ as $|\text{um}\rangle = \sqrt{\frac{N}{N-M}}|a\rangle - \sqrt{\frac{M}{N-M}}|m\rangle$. As shown in Fig. 5.2, the phase oracle is a reflection along $|\text{um}\rangle$ and the diffuser is a reflection along $|a\rangle$. These two reflections result in a rotation of the superimposed state, $|a\rangle$ by 3θ, where θ is the angle subtended by $|a\rangle$ and $|\text{um}\rangle$. It can be shown that for t Grover iterations, $|a\rangle$ will subtend an angle of $(2t + 1)\theta$. Since the goal of Grover iteration is to rotate $|a\rangle$ toward $|m\rangle$ as much as possible, t must be limited so that $(2t+1)\theta \le \frac{\pi}{2}$ where the angle $\theta = \arccos\left(\langle a|\text{um}\rangle\right) = \arccos\left(\sqrt{\frac{N-M}{N}}\right)$. Similarly, there is limit on the number of marked state that can be amplified for a given number of qubits. Since one Grover iteration rotates $|a\rangle$ by 3θ, $|a\rangle$ subtending

$\theta > 30°$ will rotate past $|m\rangle$ in to the second quadrant. This can be avoided by using more qubits.

5.4.2 Quantum Phase Estimation

Phase estimation is key in many quantum computing algorithms, acting as a subroutine to perform complex computational tasks, one of which is the Shor's factorization algorithm. Hence, it is imperative to understand phase estimation in its entirety.

Given a unitary operator U that has an eigenvector $|u\rangle$ with eigenvalue $e^{2\pi i\theta}$, the goal of the phase estimation algorithm is to estimate θ. The setup involves two registers: a counting register containing n qubits to store $2^n\theta$ values, initialized in the state $|0\rangle$, and a second register in the state $|\Psi\rangle$, containing as many qubits necessary. Mathematically, the initial setup, Ψ_0, is as follows:

$$\Psi_0 = |0\rangle^{\otimes n} |\Psi\rangle.$$

A Hadamard transformation is applied to the counting register:

$$\Psi_1 = \frac{1}{\sqrt{2^n}} \left(|0\rangle + |1\rangle\right)^{\otimes n} |\Psi\rangle.$$

Next, a set of controlled-U operations is applied on the second register, with U raised to successive powers of two. Note that, by definition of U, applying U to $|\Psi\rangle$ gives:

$$\begin{aligned} U^{2^j} |\Psi\rangle &= U^{2^{j-1}} U |\Psi\rangle \\ &= U^{2^{j-1}} e^{2\pi i\theta} |\Psi\rangle \\ &= U^{2^{j-2}} e^{2\pi i 2\theta} |\Psi\rangle \\ &= \dots \\ &= e^{2\pi i 2^j \theta} |\Psi\rangle. \end{aligned}$$

The final state of the first register is now:

$$\begin{aligned} \Psi_2 &= \frac{1}{\sqrt{2^n}} \left(|0\rangle + e^{2\pi i\theta 2^{n-1}} |1\rangle\right) \otimes \dots \otimes \left(|0\rangle + e^{2\pi i\theta 2^0} |1\rangle\right) \otimes |\Psi \\ &= \frac{1}{\sqrt{2^n}} \sum_{k=0}^{2^n-1} e^{2\pi i\theta k} |k\rangle \otimes |\Psi\rangle. \end{aligned}$$

Without going into much detail, the quantum Fourier transform (QFT) operation can be expressed as:

$$\text{QFT}|x\rangle = \frac{1}{\sqrt{2^n}} \left(|0\rangle + e^{\frac{2\pi i}{2}x} |1\rangle\right) \otimes \left(|0\rangle + e^{\frac{2\pi i}{2}x} |1\rangle\right) \otimes \dots \otimes \left(|0\rangle + e^{\frac{2\pi i}{2^n}x} |1\rangle\right).$$

Note the similarities in form with Ψ_2. Therefore, applying inverse QFT to Ψ_2 will recover the required state:

$$\left|\Psi_3\right\rangle = \frac{1}{\sqrt{2^n}} \sum_{x=0}^{2^n-1} \sum_{k=0}^{2^n-1} e^{-\frac{2\pi ik}{2^n}\left(x-2^n\theta\right)} \left|x\right\rangle \otimes \left|\Psi\right\rangle.$$

Measuring $|\Psi_3\rangle$, which peaks near $x = 2^n\theta$, obtains the phase with high probability, given the x is an integer:

$$\left|\Psi_4\right\rangle = \left|2^n\theta\right\rangle \otimes \left|\Psi\right\rangle.$$

Even if x is not an integer, the peak is still near $x = 2^n\theta$ with probability better than 40%.

5.4.3 Shor's Algorithm

Factoring a composite number into two prime numbers is the core of many cryptography algorithms. Since the security of information hinges on the fact that no classical algorithm can break an RSA encryption fast enough, successful implementation of Shor's algorithm in the real world is highly sought after.

To find the factors for a number N, the key point is to find the order of a modulo N, for a random number a such that $1 < a \leq N - 1$. The order, r, is the least positive integer r that satisfies $a^r \bmod N = 1$. If r is odd or r is even but $a^{r/2} = -1 (\bmod N)$, repeat the steps with a different random number. Otherwise, the factors for N will be $\gcd(a^{r/2} \pm 1, N)$.

For quantum factorization, a quantum register that represents the number N is first prepared in equal superposition:

$$\left|0\right\rangle^{\otimes n} \left|0\right\rangle^{\otimes m} = \sum_{x=0}^{2^n-1} \left|x\right\rangle \left|0\right\rangle^{\otimes m-1} \left|1\right\rangle.$$

It then utilizes a QPE subroutine to apply modular exponentiation, producing an entangled state:

$$\sum_{x=0}^{2^n-1} \left|x\right\rangle \left|a^x \mod N\right\rangle.$$

Finally, applying an inverse quantum Fourier transform followed by a measurement yields the desired order r. The post-processing steps to eventually obtain the factors is the same as the classical method, through the use of gcd.

On a quantum computer, the complete Shor's algorithm can factor a number in polynomial time. In contrast, the current best classical computer factors a composite number in sub-exponential time, using the general number field sieve method. That being said, with the development of a quantum computer being on hold, the factorization of large composite numbers is not yet feasible. IBM first demonstrated Shor's algorithm, factoring 15 into 3×5 with 7 qubits, using NMR implementation

[25]. Following IBM's success, two other groups have successfully implemented the algorithm through the use of photonic qubits, and one group factorized with solid-state qubits [26–28]. To date, the largest integer that has been successfully factored with Shor's algorithm experimentally is 21.

Evidently, a real-world implementation of Shor's algorithm for large enough N to be useful in breaking an RSA encryption is still an open problem, mainly due to the limitations of number of qubits readily available for use with a universal quantum computer. Currently, the largest integer factorization for a quantum algorithm is 1,005,973, through quantum annealing, using 89 qubits [29]. However, this is an evolving field and better quantum computing resources are emerging.

5.4.4 Quantum Machine Learning

In recent years, there is an emergence of interweaving quantum with machine learning concepts in solving complex data analytic problems. The intrinsic characteristics of quantum mechanics, such as quantum coherence, entanglement, and parallelism, offer valuable tools to advance the current capabilities of machine and deep learning algorithms, leading to the emergence of the new field of *quantum machine learning*. The field of quantum machine learning promises to solve current data analytic problems exponentially faster than their classical counterparts. In addition, quantum machine learning can overcome several hindrances encountered in classical machine learning algorithms, such as the sparse matrix constraint and poor performance with noisy data. As real-world data are likely to be noisy and dense, quantum machine learning is certainly an attractive prospect to overcome current barriers in classical machine and deep learning applications.

5.4.4.1 Quantum Support Vector Machines

A support vector machine (SVM) solves a supervised classification problem by finding a hyperplane that separates two classes of data with maximum margin [30] as described in Chap. 3.

Given a set of M training data points, $\{(\mathbf{x_1}, y_1), \ldots, \mathbf{x_M}, y_M)\}$, where $y_i \in \{-1, 1\}$, depends on the class where $\mathbf{x_i}$ belongs to. In the simplest solution where the hyperplane is linear, the output of the SVM is of the form $y_i = \pm(\mathbf{w}\cdot\mathbf{x_i} + b)$, where $\mathbf{w}$ is the normal vector to the hyperplane and $b/|\mathbf{w}|$ is the offset from the origin. To find the optimal hyperplane, the goal of the SVM is to minimize

$$\frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{M}\xi_i \tag{5.3}$$

subject to the constraint $y_i(\mathbf{w}\cdot\mathbf{x_i} + b) \geq 1 - \xi_i$ for all i, and $\xi_i \geq 0$. Here, ξ refers to the classification error and C is the cost parameter. The dual formulation of the optimization problem can be defined by introducing the Karush–Kuhn–Tucker multipliers. Instead, one can maxmize:

$$\sum_{i=1}^{M}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{M}\alpha_i y_i K_{j,k}\alpha_j y_j \tag{5.4}$$

subject to the constraints $\sum_{i=1}^{M}\alpha_i y_i = 0$, $\alpha_i \in [0, C]$ for all training data points. $K_{j,k}$ function is the kernel matrix, which needs to be evaluated across all training data points, and then solving the dual form by finding the optimal α_j.

There are currently two known implementations to apply quantum mechanics in SVMs. In both versions, the classical data $\mathbf{x}$ is first mapped to a quantum state $|\Phi(\mathbf{x})\rangle$. A straightforward method of a quantum SVM is to treat it as a discriminator, where the quantum data is fully trained on a quantum computer, producing the separating hyperplane upon measurement [31]. A different approach uses quantum computer to estimate the kernel function, which can then be trained like a classical SVM [32]. This method approximates the SVM as a least-squares problem and utilizes quantum matrix inversion to formulate an optimal solution. This least-squares approximation has also been proven experimentally with NMR for a handwritten digit classification problem [33].

The kernel estimation is where the quantum SVMs give an advantage over its classical counterpart. With quantum, SVMs have been proven to give exponential speedup. Introducing quantum also allows for dense training vectors to be used, which would be useful for real-world applications.

5.4.4.2 Quantum Principal Component Analysis

Principal component analysis (PCA) is often used as a dimensionality reduction tool in processing high dimensional data set. PCA finds the principal components (eigenvectors) of the covariance matrix of a zero mean dataset and then performs a change of basis. PCA then projects the high dimensional data set into the low-dimension subspace constructed from the first few principal components by discarding the eigenvectors corresponding to eigenvalues below a threshold. The main idea is that the first principle component of the covariance matrix corresponds to the largest variance and thus represents the most important behavior of the system.

The quantum PCA (qPCA) [34] utilizes the QPE algorithm and density matrix exponentiation in conducting PCA) on a quantum state. Once the quantum states (represented by density matrix ρ), corresponding to the covariance matrix of a dataset, are generated using oracles, a unitary operator $e^{-i\rho t}$ is defined by making n copies of ρ and the QPE algorithm is carried out for the spectral decomposition (finding eigenvalue and eigenvector) of the covariance matrix. qPCA is based on the principal of quantum tomography: since the state of a quantum before the measurement cannot be known, it is necessary to prepare many copies of a given state and perform measurements of different observables for analyzing the results statistically. A detailed account including the limitations of qPCA is given in Ref. [34]. The computational time required for qPCA is in the order $\mathcal{O}(R \log d)$ for the R ranked d-dimensional ρ matrix.

5.4.4.3 Quantum Bayesian Network

Bayesian Network (BN) are probabilistic directed acyclic graphs (DAG), where nodes represent random variables and edges represent the conditional dependence

between the nodes. BN are a widely used machine learning tool that can be prepared from expert knowledge or can be learned directly from data and then used to make inferences. Each node is assigned with a conditional probability distribution $P\left(x_i \middle| x_{A_i}\right)$ denoting the probability of the variable X_i conditioned on its parent's value, x_{A_i}. The conditional probability must be a non-negative real number and $\sum_{x_i} P\left(x_i \middle| x_{A_i}\right) = 1$ for all i, where the sum over x_i encompasses all the states that the random variable X_i can assume.

A quantum BN (qBN) [35] takes a BN and simply replaces probabilities of each node with complex number probability amplitudes. The conditional probability amplitude is a complex number and the probability amplitudes $\psi\left(x_i \middle| x_{A_i}\right)$ must now satisfy $\sum_{x_i} \left|\psi\left(x_i \middle| x_{A_i}\right)\right|^2 = 1$. Mathematically, qBN can represent more information than a classical BN due to the nature of the complex number system: given a qBN, a special BN can be constructed by squaring $\psi\left(x_i \middle| x_{A_i}\right)$ at each node; however, given a BN, a whole family of qBN can be constructed by taking a square root of $P\left(x_i \middle| x_{A_i}\right)$ times a phase factor, $e^{i\phi}$. The phase factor $e^{i\phi}$ is an additional parameter that needs to be tuned according to the problem at hand.

The quantum probability theory can explain certain phenomena that the classical probability theory cannot. Accordingly, qBN has been successful in correctly modeling the decision making process in situations that is known to classically violate the *Sure Thing Principle*; Moreira et al. [36] with additional modification to the previous qBN [35, 37] have successfully modeled the prisoner's dilemma and the two-stage gambling game. In addition, Borujeni et al. [38] have proposed a quantum circuit representation of Bayesian network for the realization of a fully quantum qBNs.

5.4.4.4 Quantum Neural Network and Deep Learning

Neural Networks (NNs) are the current state-of-the-art tools for machine learning purposes popularized by its wide success in deep learning tasks in the field of computer vision, computer audition, and natural language processing. NNs are scaled according to the complexity of the task at hand; a complex task requires more layers of NN than a simple task. However, the classical hardware has almost reached its size limit and will eventually cease to scale-down. This can be partly overcome by creating a quantum version of the NN (QNN) which can inherently process and store more information.

A unit of NN, perceptron or neuron, is made up of input nodes, weights, activation functions, and an output node. Training an NN is carried out via back-propagation and inference is carried out via feed forward on a trained NN. Back-propagation tunes the weights of the NN by minimizing a loss function via a gradient descent procedure. The tuned weights are then stored for inference. The QNN utilizes quantum mechanics and quantum computing in developing a quantum analogue of the perceptron and the related processes.

Since QNN is an emerging technology, a universally accepted neuron design is still missing. However, all the common designs define a neuron as a unitary operator [13]. Since quantum mechanics is based on linear algebra, defining a non-linear activation function needs extra effort; this is achieved by the use of QPE algorithm. Similarly, the feed forward, loss function, and back-propagation are defined as parameterized unitary operators where the weight parameters are updated via quantum dynamical descent procedure [39] mimicking time-evolution process. The weight parameters are defined as the quantum gate rotation angles. There also exist hybrid quantum-classical NN [24] where quantum circuits are used as the hidden layers: the weighted classical input layers are mapped to quantum circuits and the measurement statistics from the quantum circuit are then fed into a classical output layer. The parameter-shift [40] rule can be implemented for the back-propagation.

Along with the development of QNN and its realization, quantum analogue of other deep learning tools such as quantum convolution NN and quantum generative adversarial network have been actively developed. Li et al. [14] developed a QCNN framework using quantum parameterized circuit and variational quantum algorithm and demonstrated its feasibility by testing in MNIST and GTSRB datasets. Similarly, Zoufal et al. [15] implemented classical NN and quantum parameterized circuit in developing a QGAN framework; they trained and tested the framework in a simulator and in real IBM Q Experience quantum processor, showing possible application in finance application.

5.4.4.5 Quantum Reinforcement Learning

Quantum Reinforcement Learning (QRL) [16] is another QML algorithm that utilizes quantum computing in finding the optimal decision making policy in a dynamical system. Since its inception, QRL has since been used in various tasks such as Robot Navigation [41] and human decision making [17]. Specifically, QRL implements Grover's algorithm as the optimization tool in the classical RL framework reviewed in Chap. 3. In doing so, it takes the framework to the domain of quantum information which adds quadratic speed and presents a natural solution to the so-called exploration-exploitation trade-off.

As described in Chap. 3 and reviewed briefly here, classical RL [42, 43] constitutes of five elements: agent, environment, policy, reward, and value function. Assuming Markov's property, i.e., environment's response at time $t + 1$ depends only on the state and action representations at time t, the RL task is mathematically described as a Markov Decision Process (MDP). MDP is a 4-tuple (S, A, p, r), where S is a finite set of states, A is a finite set of actions, p is the transition function that maps state s and action a at time t to the next state s' and r is the reward function that assigns a reward after transiting from s to s'.

An optimal policy, π^*, corresponds to finding the optimal state-value function ($v_*(s) = \max_\pi v_\pi(s)$) or optimal action-value function ($q_*(s, a) = \max_\pi q_\pi(s, a)$). Optimal policy will lead the agent to its goal with maximum return, for instance an agent reaching the terminal grid traveling through the shortest path in a Gridworld environment. Optimal policy is achieved by following a greedy scheme in every

step of the training process. In practice, $v_*(s)$ is computed by following the on-policy iterative temporal-difference (TD) one-step update rule as,

$$V(s_t) \leftarrow V(s_t) + \alpha\left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\right], \tag{5.5}$$

while $q_*(s)$ is calculated by one-step off-policy TD control rule also known as Q-learning [43] as,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]. \tag{5.6}$$

In on-policy scheme, action selection and value update are implemented from a common policy, while in off-policy, two separate polices are followed. For instance, in Q-learning action selection (behavior policy) is carried out by ε-greedy or ε-soft policies, whereas the action-value function is updated with the best action (target policy). ε-greedy or ε-soft policy is necessary to balance the exploration-exploitation trade-off; learning the optimal decision making policy can only be obtained through sufficient exploration of the unknown environment. In general, the policy selects a random action with probability ε and the best action with probability $1 - \varepsilon$. As the training progresses, the probability ε is gradually decreased so that the agent mostly explores the environment in the initial phase and exploits the knowledge gained in the later phase. For an in-depth analysis and explicit examples of RL algorithms, the readers should refer to Ref. [42].

Dong et al. [16] implemented the Grover's search algorithm in finding the best actions, i.e., maximum value-yielding actions. To apply Grover's search algorithm, the actions are defined via quantum state. Quantum state's probabilistic nature provides a natural solution to the agent's exploration vs. exploitation trade-off without the need to define ϵ and any other related adjustable parameter. In general, Grover algorithm cannot amplify the probability amplitude of the best action (marked state) to exactly 1. Therefore, there is always a chance for a quantum measurement to yield a suboptimal (unmarked) action. The probability of yielding the best action is higher than that of suboptimal action. This is very similar to ε-greedy policy, except that all the parameters are fixed by the problem.

A training episode of the QRL algorithm can be summarized as follows: the agent starts from an initial state, and a superimposed action quantum state is prepared where each eigenstate represents an action. A quantum measurement is carried out on the action state, collapsing it into an action, which is then executed, the agent's next state is observed, reward value is obtained, and the state value is updated using either Eq. (5.5) (or 5.6). The action quantum state's probability amplitude is then updated via Grover iteration. The process is repeated starting from the quantum measurement until the goal state is reached. This training episode is repeated until convergence is reached. An application of the QRL is presented in Sect. 5.2.

5.5 Application of Quantum Computing in Medical Physics

At the date of this publication, there have been relatively few applications of quantum computing in medical physics. This is not surprising given that in many ways the field of quantum computing is still in its infancy.

5.5.1 Optimization and Planning

In optimization problems, the goal is to minimize a user-defined objective function by finding the optimal combination of parameters which the objective function depends on. A common example of such a problem in radiotherapy is treatment planning optimization, where the objective function is usually a sum of prescribed dose and dose volume constraints for the tumor and organs at risk, and the parameters to be optimized include beamlet weights, beam orientations, or MLC aperture shapes. For many real-world applications, finding the optimal solution is a non-trivial task because the solution space is non-convex—meaning there exist many potential solutions which are local minima in the objective function's energy landscape and the optimization search algorithm gets trapped in these local minima before it can find the global minimum. Some optimization algorithms such as simulated annealing are guaranteed to converge on the global minimum under certain conditions, but this often requires excessive or even infinite computation times for large problems. The development of algorithms which can find optimal (or close to optimal) solutions for challenging, non-convex optimization problems within a reasonable amount of time is therefore an active area of research.

The earliest known application of quantum computing to medical physics was a study published in 2015 which investigated quantum annealing as a method for IMRT treatment plan optimization [44]. In this study, IMRT beamlet weights for two prostate cancer cases were optimized using quantum annealer hardware and compared against two optimization methods—Tabu search and simulated annealing implemented on a standard classical computer. Each algorithm was run for the same number of objective function evaluations, and performance was assessed by computational speed (defined as wall clock time), by the final objective function value, and by the overall quality of the treatment plans generated. For both cases, quantum annealing had a wall clock time that was more than twice as fast as simulated annealing and more than three times as fast as Tabu search. Simulated annealing was found to produce the highest quality plans for both cases, while the quantum annealing technique came in second and third, respectively.

One unique aspect of performing calculations on quantum hardware is that problems must be formulated such that inputs and measured outputs are binary. In the case of the above study, each beamlet weight value was represented as a 5-bit vector ($\mathbf{w}_b \in$ {[00000], [10000], [01000],…, [11111]}), allowing for $2^5 = 32$ levels of discretization. In total, each beamlet required 7 qubits for optimization, 5 to represent the numeric value and 2 for functional smoothing. It is important to note that when performing optimization with quantum hardware, the size of the solution search

space is limited by the total number of qubits that the hardware can support. The quantum annealing device used in the above study supported 512 qubits. The authors therefore defined the beamlet dimensions such that 70 beamlet weights were optimized for each plan. For clinical applications however, a typical IMRT plan can have on the order of thousands to tens of thousands of beamlets—due to the use of more treatment beams (the study above used only 5) as well as smaller beamlet sizes. In addition, beamlet weights are usually represented as (nearly) continuous variables, meaning each beamlet weight has a much larger range of potential values it can take. The most advanced quantum annealers as of the year 2020 support up to 2000 qubits, and thus technology is not yet at the point where it can support the full complexity of these types of optimization problems in radiation oncology. However, if quantum annealers are able to double the number of qubits they support every 2–4 years (as has historically been the case), then this may cease to be an issue within the next decade.

In 2020, another study published by Pakela et al. investigated the use of a quantum-inspired optimization algorithm, quantum tunnel annealing (QTA) for IMRT treatment plan optimization [19]. QTA models the solution search as a particle performing a random walk over a 1D potential energy landscape with the potential to tunnel through energy barriers. Like its classical neighbor, simulated annealing, in QTA, as the particle explores the solution space it will always move to lower energy solutions but will also sometimes accept higher-energy (i.e., worse) solutions. This ability to occasionally accept worse solutions helps to prevent the particle from becoming trapped in local minima. The key difference between QTA and simulated annealing is in the probability for accepting worse solutions.

In simulated annealing, the solution search is modeled after a many-body system in a heat bath undergoing thermal fluctuations, where the probability of accepting a worse solution over the current solution is proportional to the ratio of their respective Boltzmann factors:

$$P(t) \propto \exp\left(\frac{-\Delta V}{T(t)}\right).$$

Here, ΔV is the change in potential energy (defined as the difference between the objective function values for the solution under consideration and the current accepted solution). $T(t)$ represents the temperature of the system; T is initially set at a large value and is gradually decreased (annealed) over the course of the optimization, gradually reducing the system's likelihood of accepting worse solutions.

In QTA, the probability of accepting a worse solution is represented by the probability of a particle traversing a 1D potential energy landscape tunneling through a barrier of width, w. Using the Wentzel–Kramers–Brillouin (WKB) approximation this probability can be approximated as:

$$P(t) \propto \exp\left(\frac{-w(t)\sqrt{\Delta V}}{\mathcal{T}(t)}\right),$$

where ΔV is again the change in potential energy, $\mathcal{T}(t)$ is the annealing variable which represents the kinetic energy of the system (analogous to T in simulated annealing), and $w(t)$ represents the width of the potential energy barrier. This barrier width is a heuristic, dynamic parameter and acts as an additional degree of freedom for the annealing schedule.

The performance of QTA was benchmarked against simulated annealing for IMRT beamlet weight optimization on two stereotactic body radiation therapy (SBRT) liver cases. Both algorithms were found to produce treatment plans of nearly identical quality as indicated by the cumulative dose volume histograms and 3D dose distributions, though for the first case, QTA exhibited greater stability—converging to the optimal plan 100% of the time while SA converged to the optimal plan 60% of the time. In addition, for the second, more challenging case, QTA converged up to 26.8% faster than SA. The results of this study indicated that the presence of the barrier width parameter could potentially serve as a valuable tool for improving both speed and robustness of treatment plan optimization.

5.5.2 Outcome Modeling/Decision Making

Cancer is a complex disease and its treatment such as radiotherapy (RT) is a complex process that involves hundreds of variables. While keeping track of all the variables and their dynamics is an extremely difficult task, taking everything into consideration in planning a treatment is virtually impossible. Thus a clinical decision support system (CDSS) that helps physicians and patients in making an informed decision about the best course of treatment is highly desirable. Tseng et al.

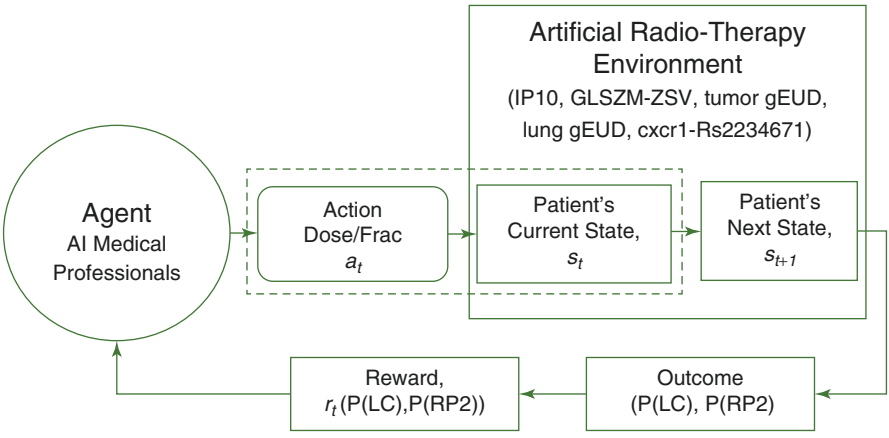


Fig. 5.3 A pictorial representation of model-based radiotherapy reinforcement framework

[45] developed a framework of such a CDSS for response-adapted RT in lung cancer based on deep RL. Response-adapted RT is a promising paradigm of personalized and precision medicine that utilizes patient-specific information such as dosimetric, imaging (radiomics), clinical, and biological markers in recognizing patient's dose response and making necessary dose adjustment mid treatment.

Biological datasets are inherently noisy, since all beings differ from each other in all levels: genetic, cellular, tissue etc. Applying quantum states and quantum computing in processing such datasets provide advantage over classical information system due to greater capacity of processing information. Accordingly, Niraula et al. [46] developed a CDSS for response-adapted RT in lung cancer based on three qubit QRL similar to Tseng et al.'s [45] work which was based on deep RL.

A model-based RL was implemented as shown in Fig. 5.3. Five patient-specific features were selected from a multi-objective Bayesian Network approach [47] in creating an artificial RT environment. The features were IP10 (cytokine), GLSZM-ZSV (radiomics-imaging), Tumor and Lung gEUD (radiation dose), and cxcr1-Rs2234671 (genetics). A three-layered Deep Neural Network (DNN) was selected as the transition function to define state dynamics and two four-layered DNN as the outcome predictor. The DNNs were trained in a dataset from a cohort of 67 non-small cell lung cancer patient. Local tumor control (LC) and radiation induced pneumonitis of grade 2 or higher (RP2) were selected as the outcome. A reward function was selected such that optimization would maximize local control while minimizing RP and the tabular q-learning approach was implemented.

Qiskit [24] quantum computing simulator was applied to define a three qubit quantum circuit and the necessary quantum gates. The eight quantum action states were assigned with dose values ranging from 1.75 to 3.5 Gy/frac. The average result

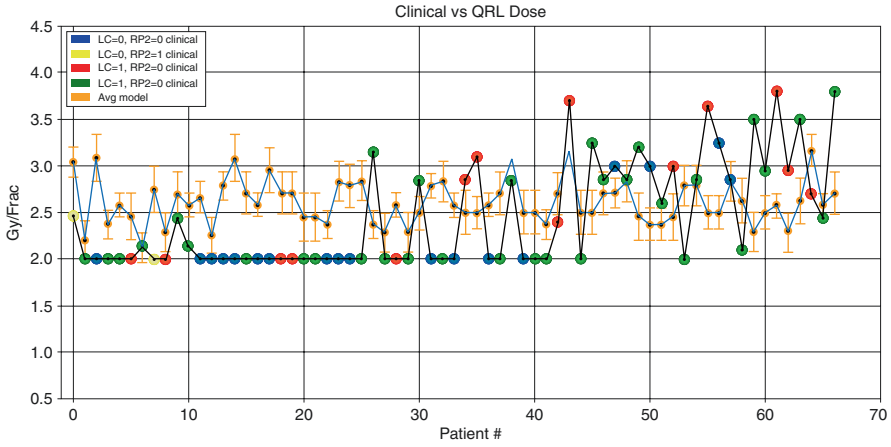


Fig. 5.4 Comparison of the clinically prescribed dose against the QRL recommended dose for week 4 to week 6 of RT treatment. The color code is based on clinical outcome

from six trained QRL models is shown in Fig. 5.4 which compares the clinical and

QRL recommended dose. This recommendation is for week 4 to week 6 of RT treatment plan; RT treatment lasts 6 weeks or 30 sessions. The RMSE deviation between the clinical and the recommended dose ranged from 0.75 to 0.84 Gy/frac which is comparable to 0.76 Gy/frac achieved by Tseng et al. via a more sophisticated Deep RL technique. The application of QRL in RT decision making is promising and need to be further explored.

5.6 Conclusion

In this chapter, we provided an overview of quantum computing and its rising role in machine learning, a new field known as quantum machine learning. Several implementations to traditional machine learning algorithms including PCA, BN, SVM, NN, RL, and deep learning are discussed. In addition, we showed example applications in medical physics treatment optimization and radiotherapy decision making.

References

1. Savage JE. Models of computation: exploring the power of computing. Boston: Addison Wesley; 1998.
2. Nielsen MA, Chuang IL. Quantum computation and quantum information. Cambridge: University Press; 2010.
3. Rieffel E, Polak W. Quantum computing: a gentle introduction. Cambridge: The MIT Press; 2011.
4. Benioff P. The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines. J Stat Phys. 1980;22:563–91.
5. Shor PW. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. arXiv:quant-ph/9508027; 1995.
6. Grover LK. A fast quantum mechanical algorithm for database search. arXiv:quantph/9605043; 1996.
7. Yin J, Li Y-H, Liao S-K, et al. Entanglement-based secure quantum cryptography over 1,120 kilometres. Nature. 2020;582:501–5.
8. Arute F, Arya K, Babbush R, et al. Quantum supremacy using a programmable superconducting processor. Nature. 2019;574:505–10.
9. U.S. Subcommitte on Quantum Information Science, Committee on Science, National Science & Technology Council. National strategic overview for quantum information science. The White House. 2018. https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/09/National-Strategic-Overview-for-Quantum-Information-Science.pdf.
10. U.S. Office of Science and Technology Policy. Artificial intelligence and quantum information science R&D summary: fiscal years 2020-2021. The White House. 2020. https://trumpwhitehouse.archives.gov/wp-content/uploads/2017/12/Artificial-Intelligence-Quantum-Information-Science-R-D-Summary-August-2020.pdf.
11. Steane AM. The ion trap quantum information processor. arXiv:quant-ph/9608011; 1996.
12. Huang H-L, Wu D, Fan D, et al. Superconducting quantum computing: a review. Sci China Inf Sci. 2020;63:180501.
13. Beer K, Bondarenko D, Farrelly T, et al. Training deep quantum neural networks. Nat Commun. 2020;11:808.

14. Li Y, Zhou R-G, Xu R, et al. A quantum deep convolutional neural network for image recognition. Quantum Sci Technol. 2020;5:044003.
15. Zoufal C, Lucchi A, Woerner S. Quantum generative adversarial networks for learning and loading random distributions. npj Quantum Inf. 2019;5:103.
16. Dong D, Chen C, Li H, Tarn T-J. Quantum reinforcement learning. IEEE Trans Syst Man Cybern B Cybern. 2008;38:1207–20.
17. Li J-A, Dong D, Wei Z, et al. Quantum reinforcement learning during human decision-making. Nat Hum Behav. 2020;4:294–307.
18. Strang G. Introduction to linear algebra. Wellesley: Cambridge Press; 2016.
19. Pakela JM, Tseng H-H, Matuszak MM, et al. Quantum - inspired algorithm for radiotherapy planning optimization. Med Phys. 2020;47:5–18.
20. Kadowaki T, Nishimori H. Quantum annealing in the transverse Ising model. Phys Rev E. 1998;58:5355.
21. Morita S, Nishimori H. Mathematical foundation of quantum annealing. J Math Phys. 2008;49:125210.
22. Mukherjee S, Chakrabarti BK. Multivariable optimization: quantum annealing and computation. Eur Phys J Special Top. 2015;224:17–24.
23. Farhi E, Goldstone J, Gutmann S, Lapan J, Lundgren A, Preda D, Quantum A. Adiabatic evolution algorithm applied to random instances of an NP-complete problem. Science. 2001;292:472–5.
24. Asfaw A, Bello L, Haim YB et al. Learn quantum computation using Qiskit. 2020. http://community.qiskit.org/textbook.
25. Vandersypen LMK, Steffen M, Breyta G, et al. Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance. Nature. 2001;414:883–7.
26. Martin-Lopez E, Laing A, Lawson T, et al. Experimental realization of Shor's quantum factoring algorithm using qubit recycling. Nat Photon. 2012;6:773–6.
27. Lanyon BP, Weinhold TJ, Langford NK, et al. Experimental demonstration of a compiled version of Shor's algorithm with quantum entanglement. Phys Rev Lett. 2007;99:250505.
28. Lu C-Y, Browne DE, Yang T, Pan J-W. Demonstration of a compiled version of Shor's quantum factoring algorithm using photonic qubits. Phys Rev Lett. 2007;99:250504.
29. Peng W-C, Wang B-N, Hu F, et al. Factoring larger integers with fewer qubits via quantum annealing with optimized parameters. Sci Chin Phys Mech Astron. 2019;62:1–8.
30. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
31. Rebentrost P, Mohseni M, Lloyd S. Quantum support vector machine for big data classification. Phys Rev Lett. 2014;113:130503.
32. Havlicek V, Corcoles AD, Temme K, et al. Supervised learning with quantum-enhanced feature spaces. Nature. 2019;567:209–12.
33. Zhaokai L, et al. Experimental realization of a quantum support vector machine. Phys Rev Lett. 2015;114:140504.
34. Lloyd S, Mohseni M, Rebentrost P. Quantum principal component analysis. Nat Phys. 2014;10:631–3.
35. Tucci R. Quantum Bayesuan Nets. Int J Mod Phys. 1995;B9:295–337.
36. Moreira C, Wichert A. Quantum-like Bayesian networks for modeling decision making. Front Psychol. 2016;7:11.
37. Leifer M, Poulin D. Quantum graphical models and belief propagation. Ann Phys J. 2008;323:1899–946.
38. Borujeni SE, Nannapanenia S, Nguyenb NH, Behrman EC, Steck JE. Quantum circuit representation of Bayesian networks. arXiv:2004.14803v1 [quant-ph]; 2020.
39. Verdon G, Pye J, Broughton M. A universal training algorithm for quantum deep learning. arXiv:1806.09729v1 [quant-ph]; 2018.
40. Crooks GE. Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition, arXiv:1905.13311v1 [quant-ph]; 2019.
41. Dong D, Chen C, Chu J, Tarn T-J. Robust quantum-inspired reinforcement learning for robot navigation. IEEE/ASME Trans Mechatronics. 2012;17:86–97.

42. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge: The MIT Press; 2018.
43. Watkins CJCH. Learning from delayed rewards. PhD Thesis, King's College, University of Cambridge, England; 1989. http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf.
44. Nazareth DP, Spaans JD. First application of quantum annealing to IMRT beamlet intensity optimization. Phys Med Biol. 2015;60:4137–48.
45. Tseng HH, Luo Y, Cui S, et al. Deep reinforcement learning for automated radiation adaptation in lung cancer. Med Phys. 2017;44:6690–705.
46. Niraula D, Jamalauddin J, Haken RT, El Naqa I. Application of quantum reinforcement learning and deep neural network for radiotherapy clinical decision support. In: AMOS 2020 Virtual Joint AAPM/COMP Meeting, Vancouver; 2020.
47. Luo Y, McShan DL, Matuszak MM, et al. A multiobjective Bayesian networks approach for joint prediction of tumor local control and radiation pneumonitis in nonsmall-cell lung cancer (NSCLC) for response-adapted radiotherapy. Med Phys. 2018;45:3980–95.

Performance Evaluation

6

Nathalie Japkowicz

As in every experimental science, advances in fundamental or applied machine learning must be strictly validated. This applies to machine learning applied to the fields of Oncology, Medical Physics, and Radiology as well. Japkowicz and Shah [1] compiled a series of techniques pertaining to the three main tenets of machine learning evaluation, focusing specifically on classification, and discussed best practices for when to choose one technique over another. In particular, the book discussed the issues of choosing an appropriate performance metric for a task; the issue of error estimation or how to sample the available data judiciously; and the issues of applying a suitable statistical test for a given situation and interpreting the results of that test. The purpose of this chapter is to review these techniques in so far as they apply to advances in Oncology, Medical Physics, and Radiology and to discuss additional evaluation techniques particularly suited for these tasks.

This chapter is divided into two sections. The first section presents the standard evaluation methods used by the machine learning community as per Japkowicz and Shah [1] and discusses some additional methods used in the medical and image processing arenas. The second section reviews current practices in the application of deep learning systems in the areas of Oncology, Medical Physics, and Radiology and makes some recommendations, based on the machine learning community practices, on how to improve them.

N. Japkowicz (✉)
Department of Computer Science, American University, Washington, DC, USA
e-mail: japkowic@american.edu

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_6

6.1 Standard Evaluation Methods for Machine Learning Systems

Japkowicz and Shah [1] describe the traditional evaluation framework used for classification systems. It consists of training a supervised machine learning algorithm on a labeled data set to learn a model. This model is then evaluated in order to assess how well it captures the relationship between the features representing the objects in need of classification and their labels. For example, an object can be a series of symptoms or absence thereof experienced by a patient and the label, or class, could be the physician's diagnostic of that patient. Similarly, an object could be a medical image, such as an X-ray or MRI scan, and the class could indicate the presence or type of tumor within that image.

Given such situations, three important questions need to be answered prior to conducting the evaluation of the classifiers at hand:

- What performance metric should be used to derive a meaningful evaluation of the different models?
- What error estimation (or, more intuitively, sampling) method is likely to yield the most reliable results? In other words, how do we divide the available data into training and testing sets, taking particular precautions when the data is scarce.
- What statistical tests should be performed to help us trust our observed results?

The first question addresses the problem of fitting the performance metric to the problem at hand. Depending on the context or domain, certain aspects of the model's performance are not that important while others are. For example while sensitivity and specificity are frequently important performance measures in the medical domain, the AUC is a more abstract concept that may not matter as much to a medical practitioner. We illustrate this on a domain in which the classifier is tasked with predicting the recurrence of breast cancer. In such a domain, while sensitivity tells us the percentage of actual recurrence cases that the classifier predicted correctly and specificity tells us the percentage of non-recurrence cases that it predicted correctly, the AUC represents the ability of a classifier to rank a randomly chosen recurrence instance higher than a non-recurrence one. While the AUC is an important intrinsic measure of a classifier's performance under different distributions, it is difficult to tie its meaning to a concrete quantity.

The second question has to do with the issue of how to divide the available data into training and testing sets, taking into consideration the fact that the relevant data available may be scarce, that training and testing on the same data set yields an unacceptable overly optimistic assessments of the models under consideration, and that it may be useful to repeat experiments on the same data set sampled in different ways in order to get as reliable an estimate of the model's performance as possible.

The third question addresses the more familiar topic of statistical significance, which attempts to assess to what extent the results obtained with the chosen metric and sampling strategy depend on chance or true ability. In particular, it asks what statistical tests are best suited to the experimental framework used in the study.

All in all these three questions are geared at finding a way to assess the usefulness of the machine learning systems designed to predict the classes of the data with unknown label. We now briefly review the most common methods used to perform this evaluation. For more detail, the reader is referred to Japkowicz and Shah [1, 2]. One important fact to keep in mind is that while there are "wrong" ways to perform classifier evaluation, there are no "right" ways. The choice of a metric, sampling technique or statistical test depends on what the researcher or practitioner is most interested in obtaining from the machine learning approach he or she uses.

6.1.1 Choosing an Appropriate Performance Measure

Many different performance measures have been proposed over the years and in various application domains. All these metrics take root in the concept of a confusion matrix, illustrated in Fig. 6.1, for a two-class problem.

A confusion matrix creates two dichotomies which it expresses simultaneously. On the one hand, it considers the dichotomy derived from the classification of the data. The data is either positive (e.g., recurrence of breast cancer) or negative (e.g., no recurrence of breast cancer). On the other hand, it considers the dichotomy of truth versus prediction: the classification is either true or it is hypothesized. Laying the two dichotomies on top of each other, we obtained four quantities:

- TP: The number of true positives, i.e., the number of instances predicted to be positives that are truly positive.
- TN: The number of true negatives, i.e., the number of instances predicted to be negatives that are truly negative.
- FP: The number of false positives, i.e., the number of instances predicted to be positives that are in fact truly negative.
- FN: The number of false negatives, i.e., the number of instances predicted to be negatives that are in fact truly positive.

True class → / Hypothesized class ↓	Pos	Neg
Yes	TP	FP
No	FN	TN
	P=TP+FN	N=FP+TN

Fig. 6.1 A Confusion Matrix

6.1.1.1 Common Metrics Used in all Machine Learning Applications

One of the simplest, most general and intuitive measure that can be obtained by combining the entries of a confusion matrix is the *accuracy* of a classifier or its opposite, the *error rate*. Accuracy is the ratio of all the instances that were correctly classified by the machine learning algorithm as either positive or negative over the total number of instances in the data set. The error rate is the opposite ratio, i.e., the ratio of all the instances that were incorrectly classified by the machine learning algorithm over the total number of instances in the data set. The formulae for both quantities are shown in Eqs. (6.1) and (6.2).

$$\text{Accuracy} = \left(TP + TN\right) / \left(P + N\right) \tag{6.1}$$

$$\text{Error Rate} = \left(FP + FN\right) / \left(P + N\right) \tag{6.2}$$

Unfortunately, while intuitive, these measures suffer from a dangerous trend: they are overly optimistic in the face of a problem that plagues nearly every domain to which machine learning is applied: the class imbalance problem [3]. The class imbalance problem is the term given to the phenomenon by which machine learning systems tend to perform poorly in the face of skewed distribution. This is common in many domains including the medical domain where occurrences of diseases such as cancer or other are minimal relative to the entire population. While classifiers have the tendency of labeling all instances with the majority class label, evaluation metrics such as Accuracy and Error Rate tend to reward this behavior rather than exposing it.

Furthermore, in the medical world, Accuracy and Error Rates are not the most interesting metrics. More interesting are Sensitivity and Specificity defined from the confusion matrix as per Eqs. (6.3) and (6.4).

$$\text{Sensitivity} = TP / P \tag{6.3}$$

$$\text{Specificity} = TN / N \tag{6.4}$$

Sensitivity can capture, for example, the proportion of people with cancer that were accurately labeled as having cancer while Specificity can point out the proportion of people who do not have cancer that were properly labeled as not having the disease.

Some other metrics of interest are the notion of Precision, as well as two widely used combinations of some of the previously mentioned metrics, the F-measure (or F-score) and the AUC. Their definitions are shown in Eqs. (6.5), (6.6), and (6.7). Precision represents the proportion of true positives within the population labeled as positive by the classifier. For example, it asks what proportion of all the people who tested positive for, say Covid-19, actually have the disease. The F-measure is a balanced combination of precision and recall (though it can be set to favor precision or recall if so desired) and the AUC, in case where the classifier threshold is set, leading to a classification rather than a score, can be expressed as a combination of sensitivity and specificity.

$$\text{Precision} = TP / \left(TP + FP\right) \tag{6.5}$$

$$\text{F} - \text{Measure}\left(\text{balanced}\right) = \left(2 * \text{Precision} * \text{Recall}\right) / \left(\text{Precision} + \text{Recall}\right) \tag{6.6}$$

$$\text{AUC}\left(\text{Classification}\right) = \left(\text{Sensitivity} + \text{Specificity}\right) / 2 \tag{6.7}$$

Two other metrics commonly used when the results of the learning system are continuous rather than discrete are the root mean squared error (RMSE) and the mean absolute error (MAE). These two metrics are defined in Eqs. (6.8) and (6.9).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \widehat{y_j}\right)^2} \tag{6.8}$$

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \widehat{y_j}\right| \tag{6.9}$$

6.1.1.2 Metrics Used Specifically in Medical Machine Learning Applications

While Sensitivity and Specificity are useful in the medical domain, they assume that the data from one class is of more interest than the data from the other. In certain cases, metrics that consider both classes as equally important are more useful. Some of these metrics, which combine Sensitivity and Specificity and are pointed out in Sokolova et al. [4], include Youden's Index, Likelihood, and Discriminant Power defined in Eqs. (6.10), (6.11), and (6.12).

$$\text{Youden's Index} = \text{Sensitivity} - \left(1 - \text{Specificity}\right) \tag{6.10}$$

$$\text{Likelihoods}\, L+ = \text{Sensitivity} / \left(1 - \text{Specificity}\right) \tag{6.11}$$

$$L- = \text{Specificity} / \left(1 - \text{Sensitivity}\right)$$

$$\text{Discriminant Power} = \frac{\sqrt{3}}{\pi}\left(\begin{array}{l}\log\left(\text{Sensitivity} / \left(1 - \text{Sensitivity}\right)\right) \\ +\log\left(\text{Specificity} / \left(1 - \text{Specificity}\right)\right)\end{array}\right) \tag{6.12}$$

Youden's Index evaluates a classifier's ability to avoid failure. It weighs equally the algorithm's performance on positive and negative examples. The positive and negative likelihoods treat sensitivity and specificity separately but equally. A higher positive likelihood and a lower negative likelihood signify a better performance on the positive and negative classes, respectively. The relation between the likelihood of two separate algorithms establishes which algorithm is preferable and in which situation [5]. The Discriminant Power evaluates how well an algorithm distinguishes between positive and negative examples.

6.1.1.3 Common Metrics Used in Computer Imaging Applications

So far, all the metrics discussed were proposed or borrowed (from other fields) by the Machine Learning community. We will now turn our attention to other metrics of interest to Oncology, Medical Physics, and Radiology. Of interest for the important task of image segmentation are the Jaccard Index and Dice Coefficient defined in Eqs. (6.13) and (6.14).

$$J(A,B) = |A \cap B| / |AUB| \tag{6.13}$$

$$D(A,B) = 2 * |A \cap B| / |A| + |B| \tag{6.14}$$

The Jaccard Index measures the area of overlap between the predicted shape and the true shape divided by the area representing their union. The Dice Coefficient is very similar and measures the overlap between the predicted shape and the true shape, but multiplies that quantity by two before dividing it by the sum of the predicted shape and the true shape. Of course, in both cases, if A = B, the ideal case, then the result is 1.

Despite the importance of these two metrics for image segmentation, it is worth noting that other more specialized supervised and unsupervised evaluation metrics for image segmentation have been proposed and are surveyed in Zhang [6, 7] among other works.

We now turn to the important question of error estimation or data sampling.

6.1.2 Choosing an Appropriate Sampling Method

Ideally, when evaluating a classifier, we would have access to the entire population or a lot of representative data from it. Unfortunately, this is usually not the case, and the limited data available has to be re-used in clever ways in order to be able to estimate the error of our classifiers as reliably as possible. This is the subfield of machine learning evaluation that we call error estimation, but two probably more intuitive terms are sampling and/or re-sampling.

The main mistake to avoid with respect to error estimation is to train and test classifiers on the same data set. Since the purpose of inducing a classifier is to use it for predictive purposes, testing the classifier on the data it was trained on is not indicative of how well it will perform on data it has never seen before. The way to estimate how useful it will be in the future is to test the classifier on data it was not trained on. If data is plentiful, that is a relatively easy task: the full data set can be divided into two large sets, a training and a testing set. The classifier gets trained on the training set and tested on the testing set. In cases where the classifier has parameters that need to be tuned, then the data must be divided into three sets: a training, a validation and a testing set. The training set is used to train multiple instantiations of the learning algorithms (each instantiation represents a different combination of parameters) which get tested on the validation set. Once the best instantiation is found, it is tested on the testing set. It is important to remember that while the

training and validation sets can be used multiple times, the testing set should be used only once in order not to bias the results.

In case where the data is scarce, dividing it into three or even two sets is not possible. In such cases, a commonly used strategy is called *k*-fold cross-validation. It consists of dividing the data into *k* disjoint subsets and training and testing *k* classifiers. Each classifier gets trained on $k-1$ subsets and tested on the subset it was not trained on. The tested subset is rotated so that eventually the k classifiers are collectively tested on all the data. The procedure is illustrated in Fig. 6.2. Of course the k classifiers are different but they have a lot in common since none of them differs from any other by more than $2*N/k$ where N is the number of instances in the data set. The performance obtained by each classifier on its dedicated testing set (called a fold) is averaged over all k folds and represents the performance of the method on that data set. Reporting the standard deviation along with the mean of the *k* folds is good practice since it indicates the stability of the classifier. It is important to note that if the classifier is parametric, then at each fold, the training set needs to be divided into a training and a validation set and the parameters must be tuned without using the testing set for that fold.

If the data set is very small, then instead of dividing the data into k subsets, it is divided into N subsets, each containing a single instance (N, again, is the number of instances in the date set). The N-fold cross-validation process is run as above but with $k = N$. This variation is called Leave-One-Out or Jacknife. It is more time consuming than k-fold cross-validation with a small enough k (e.g., $k = 10$ or $k = 5$), but it has the tendency of creating more stable classifiers since the training set used at each fold differs at most by two instances (though some instability may come from the classifier itself and not only on the data it was trained on).

Another error estimation procedure called bootstrapping is sometimes used especially on data sets that are even too small for cross-validation or leave-one-out to yield a good estimate. The idea of bootstrapping is to sample at random with replacement from the whole data set D, a very large number of new sets D_i ($i = 1$ to N, a large number, at least greater than 200) of the same size as D. The instances not selected by the random sampling procedure that created D_i represent the testing data associated with D_i. ϵ_{0i} represents the error of the classifier at run i. ϵ_0 represents the

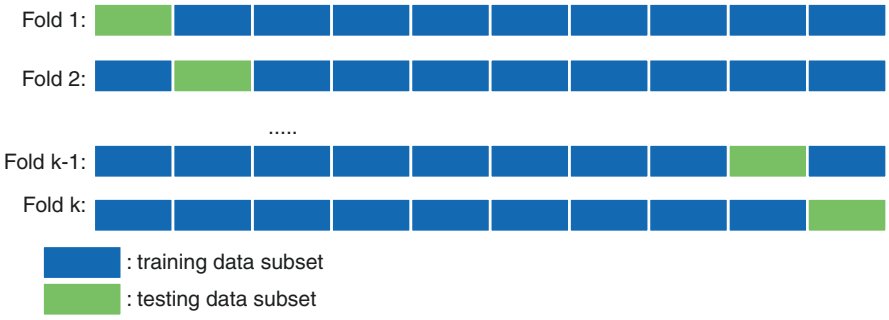


Fig. 6.2 *k*-Fold Cross-Validation

average of all the ϵ_{0i}s. It is called the ϵ_0 bootstrap. However, the ϵ_0 bootstrap tends to be pessimistic because it is only trained on 63.2% of the data in each run. The ϵ_{632} attempts to correct for this. Its formula is shown in Eq. (6.15).

$$C_{632} = 0.632 \times C_0 + 0.368 \times \text{err}(f) \tag{6.15}$$

where err(*f*) is the optimistically biased error rate obtained on the training set.

More details about cross-validation, leave-one-out, and bootstrapping can be obtained in Japkowicz and Shah [1].

We now turn to the question of statistical testing of our results to assess their reliability.

6.1.3 Choosing an Appropriate Statistical Testing Strategy

The main question asked in this section is: can the evaluation results obtained using the metrics and error estimation methods previously discussed be attributed to real characteristics of the classifiers under scrutiny or are they observed by chance? This main question can be rephrased in the context of a single classifier and in the context of several ones:

- Is the result we obtained using the methods discussed in Sects. 1.1 and 1.2 a good estimate of the true performance of our classifier?
- When comparing several classifiers on one or several domains, are the results using the methods discussed in Sects. 1.1 and 1.2 truly indicative of the best classifier?

We suggest ways to answer these questions in each of Sect. 1.3.1 and 1.3.2. More details can be found in Mitchell [8] and Japkowicz and Shah [1].

6.1.3.1 In the Context of a Single Classifier

In order to estimate, say, the true error of a classifier (but it could be any other quantity measured by one of the metrics discussed in Sect. 1.1) based on the error we observed on a testing set, we can construct a confidence interval.

This can be done if the n values being considered (either data samples themselves, if a testing set is used or averages over folds in a k-fold cross-validation regimen) were drawn independently of each other and are independent of the classifier and n is greater or equal to 30 (i.e., at least 30 data points need to be present in the testing set or 30 independent folds must have been run).

In such a case, a 95% confidence interval can be calculated around the error, *e*, (or any other metric) as per Eq. (6.16):

$$e \pm 1.96 \sqrt{\frac{e(1-e)}{n}} \tag{6.16}$$

which means that the true error lies in the interval shown in Eq. (6.16) with approximately 95% probability.

6.1.3.2 In the Context of Several Classifiers

When considering the problem of choosing the best classifier for a task or for a series of tasks, given classifiers' and their settings' sensitivity to particular and usually unknown data characteristics, there is no reliable way to select a classification method, a-priori, without comparing its performance to that of other contenders. For example, even if a Generative Adversarial Network (GAN) has been reported to do best on, say, a chest X-ray data set, it is very possible that on another chest X-Ray data set and a particular task, a convolutional neural network will actually work better. One cannot know until training both models with optimized settings for the data and comparing the results. Of course, previous studies on similar data sets are useful as they can help eliminate the classifiers that were shown not to perform well at all on that kind of data and, thus, focus on the best contenders. If, on the other hand, a type of classifier has been shown to be consistently better than others on a variety of related tasks, then it may be possible to assume that it is the best contender for the task. Even then, however, it is a good idea to verify experimentally that that result applies to the particular data set under investigation.

The purpose of statistical significance testing, in this case, is to help gather evidence of the extent to which the comparative results returned by an evaluation metric on different classifiers and possibly different data sets are representative of the general behavior of our classifiers. The principle used to establish such results is the well-known principle of hypothesis testing.

Hypothesis testing consists of stating a null hypothesis which usually is the opposite of what we wish to test (for example, classifiers A and B perform equivalently). We then choose a suitable statistical test and statistic that will be used to reject the null hypothesis. We also choose a critical region for the statistic to lie in that is extreme enough for the null hypothesis to be rejected. We calculate the observed test statistic from the data and check whether it lies in the critical region. If so, one should reject the null hypothesis. If not, one fails to reject the null hypothesis, but does not accept it either. Rejecting the null hypothesis is what gives us some degree of confidence in the belief that our observations did not occur merely by chance.

A hypothesis can be tested with a statistical test. However, there are several aspects to consider when choosing a statistical test. What kind of problem is being handled? Whether we have enough information about the underlying distributions of the classifiers' results to apply a parametric test? Regarding the type of problem, we distinguish between:

- The comparison of 2 algorithms on a single domain
- The comparison of 2 algorithms on several domains
- The comparison of multiple algorithms on multiple domains
- The comparison of multiple algorithms on a single domain

Regarding the second question, we often suggest the use of non-parametric tests since such tests make fewer assumption about the distribution of the data they are testing. On the other hand, these tests are less powerful than parametric ones and if they don't achieve statistical significance, a parametric test may be warranted after verifying that all the assumptions of the test are verified.

Table 6.1 lists the recommended test in each of the situations considered:

The tests are described in Figs. 6.3, 6.4, 6.5, 6.6, and 6.7, below, and are reprinted from the tutorial on Performance Evaluation for Learning Algorithms [9].

We will not discuss each recommended test in more detail here, but a deeper discussion can be found in Japkowicz and Shah [1].

This concludes our review of standard evaluation methods for machine learning systems. The next section questions whether these methods are appropriate in the medical imaging field.

6.2 Standard Practice in Medical Imaging and Oncology

In this section, we will review some of the studies that have been conducted in the field of medical imaging and oncology, comment on the current general evaluation practice in the community and on what could be learned from the field of machine learning's longer history of grappling with evaluation issues.

Table 6.1 Recommended Statistical Tests

Two classifiers, one domain	The sign test (non-parametric, very low power) McNemar's test (non-parametric) The t-test (parametric)
Two classifiers, multiple domains	The sign test (non-parametric) Wilcoxon's signed-rank Test (non-parametric)
Multiple classifiers, multiple domains	ANOVA (parametric) followed by Tukey's test, Nemenyi's test, etc. Friedman's test (non-parametric) followed by Nemenyi's test
Multiple classifiers, single domain	Special case of multiple classifiers multiple domains. Slightly modified ANOVA and Friedman's Test versions apply.

$$t = \frac{\bar{d} - 0}{\sigma_d / \sqrt{n}}$$

with $\bar{d} = \overline{pm}(f_1) - \overline{pm}(f_2)$ and $\bar{\sigma}_d = \sqrt{\frac{\Sigma_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$

$\bar{d}$ is the difference of the means of our performance measures obtained when applying classifiers f_1 and f_2.

d_i is the difference between the performance measures of classifiers f_1 and f_2 at trial i. n is the number of trials

Fig. 6.3 The t-test. *pm* stands for performance metric (e.g., accuracy, AUC, etc.), $\overline{pm}$ represents the average value of the performance metric over the n trials

McNemar's test is the non-parametric counterpart of the t-test. It relies on 4 values, observed on the testing set:

- The number of instances misclassified by both classifiers (c_{00})
- The number of instances misclassified by f_1 but correctly classified by f_2 (c_{01})
- The number of instances misclassified by f_2 but correctly classified by f_1 (c_{10})
- The number of instances correctly classified by both classifiers (c_{11})

The McNemar χ^2 statistics is given by $\chi^2_{MC} = \frac{(|c_{01} - c_{10}| - 1)^2}{c_{01} + c_{10}}$

If $c_{01} + c_{10} \geq 20$, then χ^2_{MC} is compared by to the χ^2 statistic. If χ^2_{MC} exceeds the $\chi^2_{1,1-\alpha}$ statistic, then we can reject the null hypothesis that assumes that f_1 and f_2 perform equally with 1-α confidence.
If $c_{01} + c_{10} < 20$, this test cannot be used, and the sign test should be used instead.

Fig. 6.4 McNemar's test

Fig. 6.5 The Sign Test

- We count the number of times that f_1 outperforms f_2, n_{f_1} and the number of times that f_2 outperforms f_1, n_{f_2}.
- The null hypothesis stating that the two classifiers perform equally well holds if the number of wins follows a binomial distribution.
- Practically speaking, a classifier should perform better on at least w_α datasets to be considered statistically significantly better at the α significance level, where w_α is the critical value for the sign test at the α significant level.

Fig. 6.6 Wilcoxon's Signed Ranked Test

- For each domain, we calculate the difference in performance of the two classifiers.
- We rank the absolute values of these differences and graft the signs in front of the ranks.
- We calculate the sum of positive and negative ranks, respectively ($W_{S1}$and W_{S2})
- $T_{Wilcoxon} = \min(W_{S1}, W_{S2})$
- Compare to ctritical value $V\alpha$. If $V\alpha \geq T_{Wilcoxon}$ we reject the null hypothesis stating that the two classifiers perform equally well at the α confidence level.

6.2.1 Review of the Current Practice in Medical Imaging and Oncology

This discussion is based on Sahiner et al. [10] where the authors review the application of deep learning strategies in medical imaging and radiation therapy. The three main types of problems they identify[1] are:

[1] An additional two tasks, Processing and reconstruction, and Imaging and treatment are also presented but are less relevant to the discussion in this chapter.

Fig. 6.7 Friedman's Test

- All the algorithms are ranked on each domain separatelty. Ties are resolve by assigning the average rank of all the classifiers involved in the tie to these classifiers.
- For each classifier, the sum of their ranks obtained on all the domains is computed and named $R_{.j}$ where j is the classifier considered.
- The Friedman statistic is calculated as:
- $\chi_F^2 = [\frac{12}{n \times k \times (k-1)} \times \Sigma_{j=1}^{k} (R_{.j})^2] \; 1 - 3 \times n \times (k-1)$

  With n representing the number of domains and k, the number of classifiers.
- If χ_F^2 exceeds the χ^2_{k-1} statistics approximated specifically for Friedman's test, then we can reject the null hypothesis that states that all classifiers perform equally well.

- Image segmentation (organ or tumor segmentation)
- Detection (organ or lesion detection)
- Characterization (lesion, tissue, diagnosis, prognosis, staging)

The issue of evaluation is discussed in a section of the paper where, in particular, warnings about testing on the same data the classifier was trained on are given and include the more subtle issue of having to create a validation set in addition to a training set. The survey also suggests the use of cross-validation if the data set is too small for a division into training validation and testing sets to be viable. Leave-one-out is implicitly mentioned (See Tables 1–6 of Sahiner et al. [10] which report the experimental settings used in a large number of studies belonging to the different types of problems identified).

While metrics are not explicitly described, many are implicitly considered (See Tables 1–6 of Sahiner et al. [10]).

Absent from the review are mentions of statistical tests to validate the results obtained, although one measure of statistical reliability is implicitly mentioned in the form of standard deviation of the measured error (See Table 3 of Sahiner et al. [10]).

6.2.2 Areas where Improvements Could Be Made

The discussion and practical methodology used by the studies reported in Tables 1–6 of Sahiner et al. [10] show good adherence to the main rules of evaluation.

On the metrics front, the authors of these studies displayed great sensitivity to some of the shortcomings of the common metrics and adopted more sophisticated ones. In particular, accuracy was most generally avoided, which is a good step to deal with its inadequacy in class imbalanced situations, a problem that is prevalent in medical settings.

On the sampling (or error estimation) front, however, a disturbing trend was observed in some of the studies reported in the paper, due mostly to the small size of the data sets available. In particular, there were a few studies considering data sets of the order of 15–20 patients for training and 8–10 patients for testing. These are concerning. Even when the pool of testing patients reaches the 20s, the results remain questionable. This is the case, statistically speaking, given that reliable confidence intervals can only start being established for 30 samples, due to the central limit theorem. Taken from a strict machine learning point of view, the problem is exacerbated in the training set by the fact that the data is of such high dimensionality to begin with, that training on a small number of instances creates greater chances of overfitting. In cases where the data sets are that limited, it would be greatly advised not to divide the data into training and testing sets, but instead, to use leave-one-out and bootstrapping. If such an evaluation regimen is too costly, the authors should at least consider k-fold cross-validation with an appropriate value for k. Asking researchers to adhere to these higher standards of evaluation would help validate the use of machine learning and deep learning methods in Oncology, Medical Physics, and Radiology.

On the statistical testing front, while Table 3 of Sahiner et al. [10], which reports organ and anatomical structure detection results using regression measures, lists both the mean error and its standard deviation, no record is made of any measure of statistical validity in any of the other studies. A look through a handful of papers indicates that indeed, such issues were often not considered. It is important to note that calculating confidence intervals or verifying the statistical significance of comparisons as discussed in Sect. 1.3 is very useful in assessing the true utility of a classifier. Along with careful sampling, including such considerations would help the field move forward faster as it could help it focus on the methods that truly work best and could be the basis for important practical advances in the field.

6.2.3 Lessons from the Past

The advent of deep learning has precipitated the use of machine learning in medical imaging, though it remains a relatively recent phenomenon. Evaluation may not seem like a priority at this point since the focus is on tuning existing architectures and coming up or trying new methods.

This is natural and happens in every field. For example, something very similar happened at the very beginning of the field of machine learning. In 1980, some papers were published with results presented as statements such as “The program works well” and “The rules developed are similar to those invented by humans playing the same game (15 complete games have been analyzed)” [11]. Despite such vague evidence, such papers were very important for the development of machine learning methods. Other articles, on the other hand, already used the notions of confusion matrices and training and testing sets [12]. However, they did not dig further into evaluation standards. By the early 1990s, Weiss and Kulikowski [13] popularized the use of cross-validation and hypothesis testing. In the late 1990s,

Provost et al. [14] started challenging accuracy as a reliable metric, while by the middle of the 2000s, Demsar decried the overuse of the t-test and introduced the machine learning community to the more appropriate statistical tests mentioned earlier [15]. As a result, despite its long history, machine learning took a long time to come to the fore.

To accelerate progress in the fields of Oncology, Medical Physics, and Radiology, it may be useful to borrow some of the advances that took place in machine learning evaluation and quickly adapt them to the needs of the community.

References

1. Japkowicz N, Shah M. Evaluating learning algorithms: a classification perspective. Cambridge: Cambridge University Press; 2011.
2. Japkowicz N, Shah M. Performance evaluation in machine learning. In: El Naqa I, et al., editors. Machine learning in radiation oncology: theory and applications. Cham: Springer; 2015.
3. Branco P, Torgo L, Ribeiro RP. A survey of predictive modelling under imbalanced distributions. ACM Comput Surv. 2016;49(2):31:1–31:50.
4. Sokolova, M., Japkowicz, N. and Szpakowicz, S. 2006, Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation, in the Proceedings of the 2006 Australian conference on artificial intelligence.
5. Biggerstaff B. Comparing diagnostic tests: a simple graphic using likelihood ratios. Stat Med. 2000;19(5):649–63.
6. Zhang YJ. A survey on evaluation methods for image segmentation. Pattern Recogn. 1996;29(8):1335–46.
7. Zhang H, Fritts JE, Goldman SA. Image segmentation evaluation: a survey of unsupervised methods. Comput Vis Image Underst. 2008;110(2):260–80.
8. Mitchell T. Machine learning. New York: McCraw Hill; 1996.
9. Japkowicz, N. Performance evaluation for learning algorithms, Canadian AI 2016, May 31, Victoria, BC. http://fs2.american.edu/japkowic/www/#Tutorials
10. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML. Deep learning in medical imaging and radiation therapy. Med Phys. January 2019;46(1):e1–e36.
11. Dietterich TG. Applying general induction methods to the card game Eleusis. In: *Proceedings of the National Conference on artificial intelligence, AAAI-80*, Stanford, California; 1980. p. 218–20.
12. Michalski RS, Chilausky RL. Learning by being told and learning from examples: an experimental comparison of two methods of knowledge acquisition. Pol Anal Inform Syst. June 1980;4(2):125–61
13. Weiss SM, Kulikowski CA. Computer systems that learn: classification and prediction methods from statistics. In: Neural nets, machine learning, and expert systems. San Maeto: Morgan Kaufmann; 1991.
14. Provost F, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. In: Shavlik J, editor. Proc. ICML-98. San Francisco: Morgan Kaufmann; 1998. p. 445–53.
15. Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.

7 Software Tools for Machine and Deep Learning

Dipesh Niraula and Issam El Naqa

7.1 Introduction

Machine learning took a huge leap with the success of deep learning in computer vision and natural language processing. Deep learning enables a computer to directly gain accurate high-level information from raw images and video mimicking human visual system, and audio mimicking human auditory system. Such ability is essential for full automation of various medical and non-medical systems. Thus, the application of machine and deep learning algorithms have drastically increased in a wide range of fields such as medical research, autonomous driving, aerospace and defense, industrial automation, electronics, business sector, etc. In turn, machine learning computer software and tools have been actively researched and rapidly developed. This chapter introduces several deep learning platforms with brief description of some of the most popular ones as depicted in Fig. 7.1.

Because Python as an interpreted, high-level, general-purpose language is easy to understand and use, most popular open-source machine and deep learning platforms are recently developed in the Python language. In fact, due to explosion of Python based deep learning community, popularity of the Python language has surpassed that of C++ and is only behind java and C [1]. So this chapter will primarily focus on Python-based libraries. Cloud computing is an emerging technology that simplifies computational process and resource access, where a user can access data-storage and computing power through internet without directly maintaining expensive hardware resources. This paradigm shift from on-premise towards cloud computing may soon become the norm and thus we have included a discussion on some of the most common cloud computing platforms that are equipped with machine learning tools.

D. Niraula (✉) · I. El Naqa
H. Lee Moffitt Cancer and Research Institute, Tampa, FL, USA
e-mail: Dipesh.Niraula@moffitt.org; Issam.ElNaqa@moffitt.org

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_7

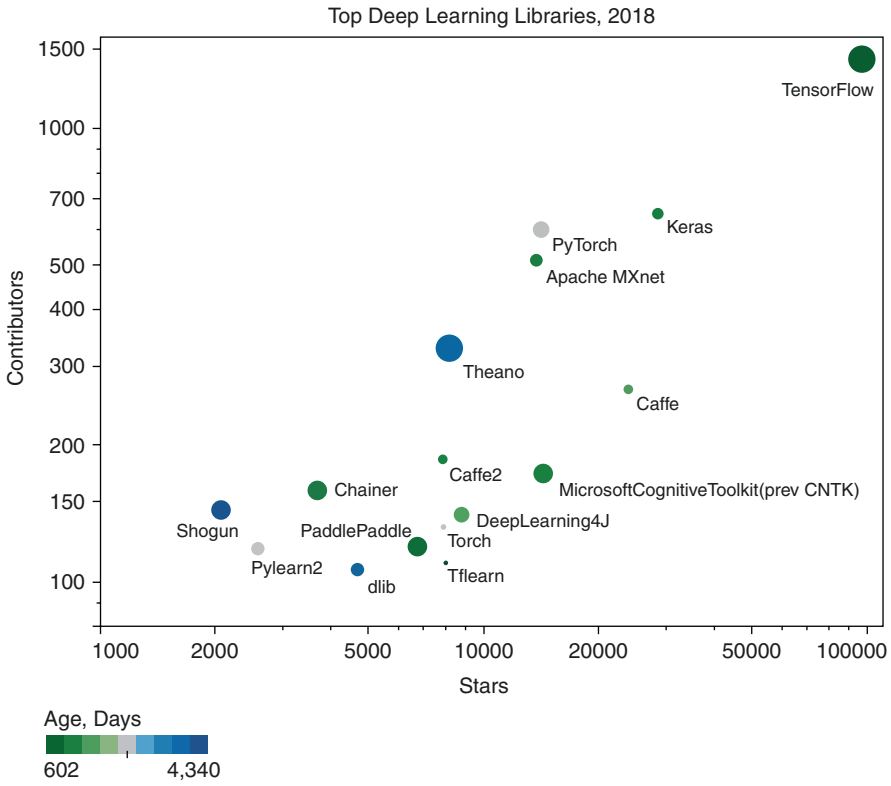


Fig. 7.1 Top Deep Learning Libraries according to KDnuggets.com, by Commits and Contributors. Circle size is proportional to number of stars

7.2 Python-Based Machine Learning Library

This section begins with introduction to tools required to download Python packages, then presents several popular Python based machine learning and deep learning libraries.

7.2.1 Pip and Conda

Pip and Conda are two software tools for installing python packages. Pip is the Python Packaging Authority's recommended tool that installs packages from Python Package Index, PyPI [2]. Conda is a cross platform package and environment manager that installs packages from the Anaconda Repository or the Anaconda Cloud (www.anaconda.com). The main difference between pip and conda is that pip installs python packages, whereas conda can install packages that might contain

software written in other languages. Thus python interpreter must be pre-installed for pip, whereas conda can install python packages as well as python interpreter. Other difference is that conda can create isolated virtual environment that can contain different version of packages, whereas secondary tools like virtualenv and venv must be utilized for pip. In practice, both of these tools are interchangeably used to get the best of both.

7.2.2 NumPy and SciPy

Although not directly related to machine or deep learning, NumPy [3, 4] and SciPy [5] packages are the most useful scientific computing packages in the Python language. NumPy provides ability to process multidimensional array (Tensors) and SciPy, built over NumPy, provides high-level mathematical functions and algorithms such as optimization, integration among many others and also can handle sparse matrices and k-dimensional trees. This makes it entirely possible to construct deep learning tools purely out of NumPy and SciPy; around 47% of all machine learning projects on GitHub used SciPy [5]. Additionally, many popular deep learning libraries are either built on NumPy or attempts to mimic NumPy.

7.2.3 Dedicated Machine Learning Libraries

Machine Learning library contains tool such as support vector machines (SVM), K-nearest-neighbor classifier, hierarchical clustering, and Principal Component Analysis (PCA) among many others for solving regression, classification, clustering, dimensionality reduction, and other problems. Following are some common machine learning libraries.

7.2.3.1 Scikit-Learn

Scikit-learn [6, 7] is an open source robust machine learning library built on NumPy, SciPy, and matplotlib, and is the most popular standalone machine learning library. It was originally created by David Cournapeau as a Google summer project in 2007. It has a range of supervised and unsupervised learning algorithms and rich resources for data preprocessing, model selection and evaluation, dimensionality reduction, and many other tools well documented in its 3000 page user-guide [8]. While the package is written in Python, it incorporates C++ libraries such as LibSVM and LibLinear. In practice, its data preprocessing tools are often used in conjunction with other deep learning libraries.

7.2.3.2 Shogun

Shogun [9] is an open-source machine learning library initiated by Soeren Sonnenburg and Gunnar Raetsch in 1999 with focus on bio-informatics. It is one of the oldest library that is written in C++ with interfaces to Python, Octave, Java/

scala, Ruby, R, Lua, and C#. It offers binding to several other libraries such as LibSVM, LibLinear, SVMLight, LibOCAS, libqp, VowpalWabbit, Tapkee, SLEP, GPML, and others. The cloud version can also be remotely accessed via Jupyter notebook. Scikit and shogun are considered to be the two comprehensive machine learning libraries currently.

7.2.3.3 mlpy

Machine Learning Python (mlpy) [10] is an open-source machine learning library developed by Davide Albanese dedicated to computational biology in general and functional genomic modeling in particular. It is built on top of NumPy, SciPy, and GNU Scientific Library and uses CPython to communicate with GNU, which is written in C. The last stable version was released in 2012.

7.2.3.4 PyMVPA

Multivariate Pattern Analysis in Python (PyMVPA) [11] is an open-source machine learning library developed as a neuroimaging software for functional magnetic resonance imaging data analysis in 2009. The library is built on NumPy and SciPy. Along with basic ML algorithms, it includes statistical tools for analysis from R via Python wrappers RPy. It also utilizes ShoGun.

7.2.3.5 MDP

Modular toolkit for Data Processing (MDP) [12] is an open-source machine learning library developed as a part of theoretical research in neural science in 2009. It is built on NumPy and SciPy. It includes supervised and unsupervised learning algorithms and other data processing units called nodes that can be combined into a sequence, a feed-forward network architecture, which is the building block of deep learning model.

7.2.3.6 PyBrain

Python-Based Reinforcement Learning, Artificial Intelligence and Neural Network Library (PyBrain) [13] is an open-source python-based machine learning toolbox that includes early versions of neural networks. It is built on SciPy that provides tools for supervised, unsupervised, and reinforcement learning. Beside machine learning algorithms such as SVM, Gaussian processes, etc., it contains deep learning elements such as Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) and deep belief networks. Pybrain is a hybrid library that forms a bridge between machine and deep learning libraries.

7.2.4 Deep Learning

Deep learning libraries contain tools pertaining to neural networks such as generic neural networks, convolution neural networks (CNN), RNN, activation functions, optimizers, weight and bias initializers and normalizers, etc. All of the libraries are efficient at matrix and tensor operations and many of them are designed to

utilize single GPU and some to utilize multiple GPUs for parallel distributed training. Several variants of stochastic gradient descent (SGD) algorithms are also included.

7.2.4.1 Theano

Theano [14] is an open-source deep learning framework developed by the Montreal Institute for Learning Algorithms (MILA) group starting from 2008. Theano efficiently compiles multi-dimensional arrays in a highly optimized fashion for both CPUs and GPUs using CUDA (Compute Unified Device Architecture), the parallel computing platform developed by Nvidia. Additionally, theano's API uses and mimics NumPy, and many deep learning frameworks have been built on top of it. Theano implements define-and-run or static-graph approach, i.e., define the fixed connection between various mathematical operations and then run the training. Static-graph approach is optimal for static neural networks architectures such as CNN.

7.2.4.2 Chainer

Chainer [15] is an open-source deep learning framework developed by Preferred Networks, a startup based in Japan, in 2015. Chainer implements define-by-run or dynamic-graph approach, i.e., the connection in the network is determined during the training. Dynamic-graph approach is better for variable-length data such as deep reinforcement learning and LSTM RNN compared to static-graph used in CNN. It utilizes CuPy, a matrix library accelerated with CUDA, for enhancing its training speed. CuPy is very similar to and highly compatible with NumPy. Chainer supports GPU (CUDA) and distributed parallel training. For computer vision task, an add-on package ChainerCV is also available.

7.2.4.3 TensorFlow

Tensorflow [16] is an open-source ML library built by the Google Brain team, focusing especially in deep learning. It is written in C++ core language with python as a binding language. It implements static-graph approach by default and switches to dynamic-graph in eager mode (Eager mode default since TensorFlow 2.0). Tensorflow also provides a visualization tool called Tensorboard that traces and illustrates the deep learning process. Keras is a high-level API now integrated with tensorflow. It has condensed syntax, as shown in Example 1, that makes it suitable for production purposes.

7.2.4.4 PyTorch

PyTorch [17] is an open-source ML library built preliminary by Facebook AI Research lab based on Torch library originally written in Lua, a simple C API. It implements dynamic-graph approach. It supports GPU (CUDA) and distributed parallel training. In practice, syntax-wise, distributed training in PyTorch is easier to set up than TensorFlow. Example 2 presents a sample PyTorch code that uses GPU when one in available.

7.2.4.5 Caffe

Caffe [18] is an open-source deep learning framework developed by Berkeley AI Research. Caffe stands for convolutional architecture for fast feature embedding. It is written in C++ core language and supports Python and MATLAB as binding languages. It implements a static-graph approach.

7.2.4.6 MXNet

MXNet, pronounced as mix-net, is an open-source deep learning software framework developed by Apache Software foundation [19] in 2015. MXNet implements mixture of static- and dynamic-graph approach, to obtain optimal performance. It is written in C++ core language and supports Python, R, Julia, and Go as binding language. It supports multiple GPU and distributed parallel training.

7.2.5 Examples

Two examples of classification model in Tensorflow and PyTorch are presented in this section for the interested reader to highlight some of the main differences between these two popular platforms and provide sample code that can be customized for other radiological applications from model building to evaluations as presented in other chapters of this book.

Example 1: Classification in Tensorflow.Keras

```
## TensorFlow and tf.keras libraries
import tensorflow as tf

##load data (x) with binary label (y)
(x_train, y_train),(x_test, y_test) = load_data()

##Build the model as dense: y=W.x+b
# Set up the layers
model = tf.keras.models.Sequential([
   tf.keras.layers.Dense(64, activation='elu'),
   tf.keras.layers.Dense(64, activation='elu'),
   tf.keras.layers.Dropout(0.2),
   tf.keras.layers.Dense(1, activation='sigmoid')
])
#Compile the model
model.compile(optimizer='adam',loss='binary_crossentropy',
              metrics=['accuracy'])
# Train the model
model.fit(x_train, y_train, epochs=100)
# Evaluate the model
model.evaluate(x_test, y_test)
```

Example 2: Classification in PyTorch

```
## Pytorch libraries and helpers
import torch
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim
import NumPy as np

##Selects GPU if available
device = torch.device("cuda:0"
         if torch.cuda.is_available() else "cpu")
class Model(nn.Module):
   def_init_(self,input_dim):
      super(Model_LCRP2, self)._init_()
##Build the model as dense: y=W.x+b
# Set up the layers
      self.linear1=nn.Linear(input_dim, 64)
      self.linear2=nn.Linear(64, 64)
      self.dropout3=nn.Dropout(p=0.2)
      self.linear4=nn.Linear(64,1)

   def forward(self,x):
      x=F.elu(self.linear1(x))
      x=F.elu(self.linear2(x))
      x=self.dropout3(x)
      return F.sigmoid(self.linear4(x))
# Define tranining and testing modules
def test(model,x,y):
   model.eval()
   test_acc = 0.0
   for x,y in zip(x,y):
      outputs = model(x)
      prediction=outputs.data
      test_acc += torch.sum(prediction.round()
                            == y.data.round())
   #Compute the average acc all test samples
   test_acc = test_acc / len(x)
   return test_acc
def train(model,optimizer,num_epochs,model_name,
                  x_train,y_train,x_test,y_test):
   Losses=np.zeros(num_epochs)
   TestAccuracy=np.zeros(num_epochs)
```

```
    TrainAccuracy=np.zeros(num_epochs)
    for epoch in range(num_epochs):
       model.train()
       train_acc=0.0
       train_loss=0.0
       for x , y in zip(x_train,y_train):
        #clear all accumulated gradients
        optimizer.zero_grad()
        outputs = model(x)
        loss = loss_fn(outputs,y)
        #Backpropagate the loss
        loss.backward()
        #Adjust parameters according to the
        #computed gradients
        optimizer.step()
        #for the performance metrics
        train_loss += loss.item()
        prediction=outputs.data
        train_acc += torch.sum(prediction.round() ==
                                          y.data.round())
       #compute averages over all training samples
       train_acc = train_acc / len(x_train)
       train_loss = train_loss / len(x_train)
       #evaluate test set
       test_acc = test(model,x_test,y_test)
       Losses[epoch]=train_loss
       TestAccuracy[epoch]=test_acc
       TrainAccuracy[epoch]=train_acc
       #Print the metrics
       print("Epoch {}, Train Accuracy: {:.2f} ,TrainLoss:
        {:.2f},Test Accuracy: {:.2f}".format(epoch,
        train_acc, train_loss,test_acc))
   #outputs metrics
   return Losses, TestAccuracy, TrainAccuracy
(x_train, y_train),(x_test, y_test) = load_data()
model = Model(input_dim=10).to(device)
optimizer = optim.Adam(model.parameters(), lr=1e-4)
loss_fn=nn.BCELoss()
epoch=100
if_name_== "_main_":
   Losses, TestAccuracy, TrainAccuracy=train(model,optimizer,
                      epoch,x_train,y_train,x_test,y_test)
```

7.2.6 Benchmark

This section provides some comparison performances of the various machine and deep learning libraries mentioned above for the reader benefit and the trade-offs employed when using them. The metrics in this section are accumulated from research papers cited in the respective table heading.

Table 7.1 compares computation speed of different tasks for all six machine learning library on Madelon data set [20].

Tables 7.2, 7.3, 7.4, 7.5, 7.6 and 7.7 compares performance of different deep learning library. The benchmarks for Tables 7.2, 7.3, 7.4, and 7.5, were run on a NVIDIA Digits DevBox, with 4 Titan X GPUs, and a Core i7-5930K CPU. Additionally, cuda 7.5.17 with cuDNN v4 and data type float 32 is used.

Table 7.1 Computation time in seconds [6]

	Scikit-Learn	mlpy	PyBrain	PyMVPA	MDP	Shogun
Support vector classification	5.2	9.47	17.5	11.52	40.48	5.63
Lasso (LARS)	1.17	105.3	–	37.35	–	–
Elastic Net						
0.52	73.7	–	1.44	–	–	
k-nearest neighbors	0.57	1.41	–	0.56	0.58	1.36
PCA(9 components)	0.18	–	–	8.93	0.47	0.33
k-means(9 clusters)	1.34	0.79	>1 h	–	35.75	0.68

Computation time for Madelon dataset [20]

Table 7.2 CNN: Forward|Backward computation time per minibatch (ms) [14]

	AlexNet	OverFeat	VGG	GoogLeNet
Theano	32|99	94|288	178|600	172|546
Theano-fast compile	44|118	111|319	263|758	250|680
Torch	27|80	90|269	164|525	132|470
Tensorflow	27|80	88|277	155|538	128|443

Processing time for various convolutional neural networks (CNN) on Imagenet dataset [21]. The specification of the CNN are: One-column variant of AlexNet [22] with batch size of 128, fast variant of Overfeat [23] with batch size of 128, model A VGG (oxfordnet) [24] with a batch size of 64, and GoogLeNet V1 [25] with a batch size of 128

Table 7.3 RNN: 1000 words/s [14]

	Small	Medium	Large
Theano	14	12	10
Theano-fast compile	11	10	8
Torch	12	8	6
Tensorflow	17	11	8

Processing speed for Long Short Term Memory (LSTM) models on the Penn Treebank dataset [26]. Here, small model corresponds to single layer, 200 hidden units, sequence length: 20, medium model corresponds to single layer, 600 hidden units, sequence length: 40, and large model corresponds to two layers, 650 hidden units each, sequence length: 50. Batch size of 20 was used

Table 7.4 Caption generating from video, Forward|Backward computation time/minibatch (ms) [14]

Batch Size	32	64	128
Theano	72|298	102|520	182|923
Tensorflow	100|323	135|520	232|850

Processing time for generating word sequences from video representations. Input video frame was preprocessed by a GoogLeNet that was pretrained for classification with ImageNet

Table 7.5 Theano data parallelism with LSTM, sync every batch|sync every 100 batch: 1000 words/s [14]

	Small	Medium	Large
1 GPU	14|14	12|12	10|10
2 GPU	23|27	20|24	15|19
4 GPU	45|55	39|47	31|38

Processing speed with multiple GPUs with Platoon on LSTM models, synchronizing after each batch and every 100 batch. Small, medium, and large models are same as Table 7.3

Table 7.6 Single-machine benchmarks: training step time (ms) [16]

	AlexNet	OverFeat	OxfordNet	GoogLeNet
Caffe	324	823	1068	1935
Neon	87	211	320	270
Torch	81	268	529	470
Tensorflow	81	279	540	445

Training step time using one GPU for 32-bits floats

Table 7.7 Throughput: sample per second [17]

	AlexNet	OverFeat	OxfordNet	GoogLeNet
Chainer	778 ± 15	N/A	219 ± 1	N/A
CNTK	845 ± 8	84 ± 3	210 ± 1	N/A
MXNet	1554 ± 22	113 ± 1	218 ± 2	444 ± 2
PaddlePaddle	933 ± 123	112 ± 2	192 ± 4	557 ± 24
TensorFlow	1422 ± 27	66 ± 2	200 ± 1	216 ± 15
PyTorch	1547 ± 316	119 ± 1	212 ± 2	463 ± 17

Training speed using 32bit float; throughput is measured in images per second

Benchmark for Table 7.6 were run on six-core Intel Core i7-5930K CPU at 3.5 GHz, and an NVIDIA Titan X GPU. Benchmark for Table 7.7 was run on 2 Intel Xeon E5-2698 v4 at 2.20 GHz with 20 cores per sockets and 1 Nvidia Quadro GP100.

7.3 Weka

Weka [27] is a conventional machine learning tool that has a graphical user interface (GUI) and enables one to apply machine learning without programming. Its development initiated in 1997 by developers at the University of Waikato and is an open source and can be freely downloaded. It is written in Java and is compatible with

most modern computing platforms. It gives access to other machine and deep learning toolboxes such as scikit-learn, R, and Deeplearning4j. Additionally, WekaDeeplearning4j is a deep learning package for Weka that adds deep learning capability to the Weka's GUI.

7.4 R

R supports machine learning such as linear discriminant analysis, classification and regression trees, k-nearest neighbors, SVM with a linear kernel, Random Forest, etc. For deep learning, keras and tensorflow has launched an R version that gives the statistically equipped R with state-of-the-art deep learning capabilities. Alternatively, python function and tools can be used in R by wrapping them with R wrapper function using the reticulate library.

7.5 Matlab

Matlab has excellent deep learning tool box (Fig. 7.2) that contains comprehensive sets of deep learning tools and models. Matlab supports single or multiple GPUs (CUDA), and can exchange models with Tensorflow and PyTorch via ONNX format and also import models from TensorFlow-keras and Caffe. The toolbox supports transfer learning with DarkNet-53, ResNet-50, NASNet, SqueezeNet, and many other pretrained model. Although matlab is not an open source, in addition to deep learning tool box, it contains computational tools from virtually every engineering field to create interdisciplinary deep learning models.

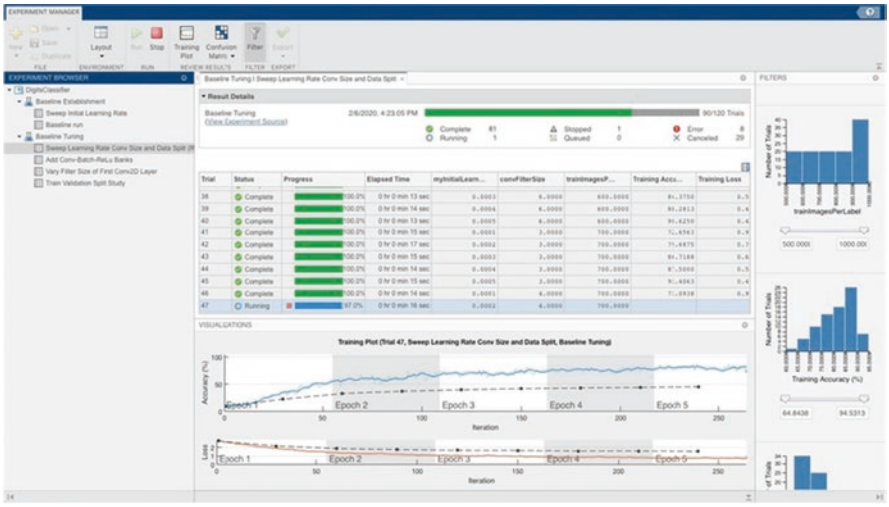


Fig. 7.2 Screenshot of Matlab deep learning tool box

7.6 Cloud-Based Platforms

With advancement of cloud and serverless systems, the computing paradigm is shifting towards web browser based computing. The main attraction of cloud computing is not having to worry about backup, storage, installing and updating software, and maintaining the server in exchange for a usage fee. However, it remains a trade-off with on-premise cluster resources for deep learning such as high performance computing (HPC) servers, where issues related to security and privacy are controlled by the local IT team versus the maintenance of such resources.

There are several sources that provides online platform services for machine and deep learning. A few popular applications are mentioned in this section. In practice, cloud services are very convenient in places with fast internet connections.

7.6.1 AWS Deep Learning AMIs and SageMaker

The Amazon Web Service (AWS) Deep Learning Amazon Machine Images (DL AMI) provides infrastructure and tool for deep learning in the cloud. It can be accessed from https://aws.amazon.com/machine-learning/amis/. Users can access multiple CPUs and GPUs for training larger models. The service is pre-installed with popular deep learning frameworks and interfaces such as TensorFlow, Keras, PyTorch, Theano, and others. For computing acceleration it includes latest NVIDIA GPU-acceleration through pre-configured CUDA and cuDNN drivers, as well as the Intel Math Kernel Library, and has pre-installed Anaconda Platform.

Amazon also offers SageMaker as a separate platform from DL AMI, which can be accessed from https://aws.amazon.com/sagemaker/. SageMaker is a development platform suitable for developing application, whereas DL AMI is suitable for research and development. Similar to DL AMI, it also supports frameworks like TensorFlow, PyTorch, Apache MXNET, Chainer, Keras, Scikit-learn, and others. SageMaker comes in three different types: SageMaker, SageMaker Studio (Fig. 7.3), and SageMaker Autopilot. Studio is a fully integrated development environment, whereas Autopilot has automated machine learning development capability for industrial purpose (Fig. 7.3).

7.6.2 Google Colab

Google Colab (Fig. 7.4) is a jupyter notebook environment that provides free deep learning computation platform running in the cloud. It can be accessed from https://colab.research.google.com/notebooks/intro.ipynb. It consists of pre-installed deep learning packages such as Keras, Tensorflow, PyTorch and other related libraries. To add new libraries one can simply pip install. It has open access platform where users can run deep learning models for free. For larger development project, users have access to powerful GPUs and even TPUs for a fee as an on-demand service.

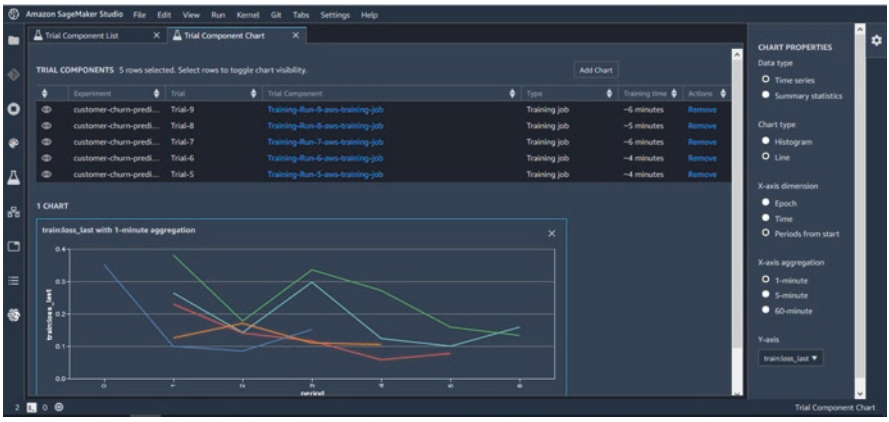


Fig. 7.3 Screenshot of AWS SageMaker Studio

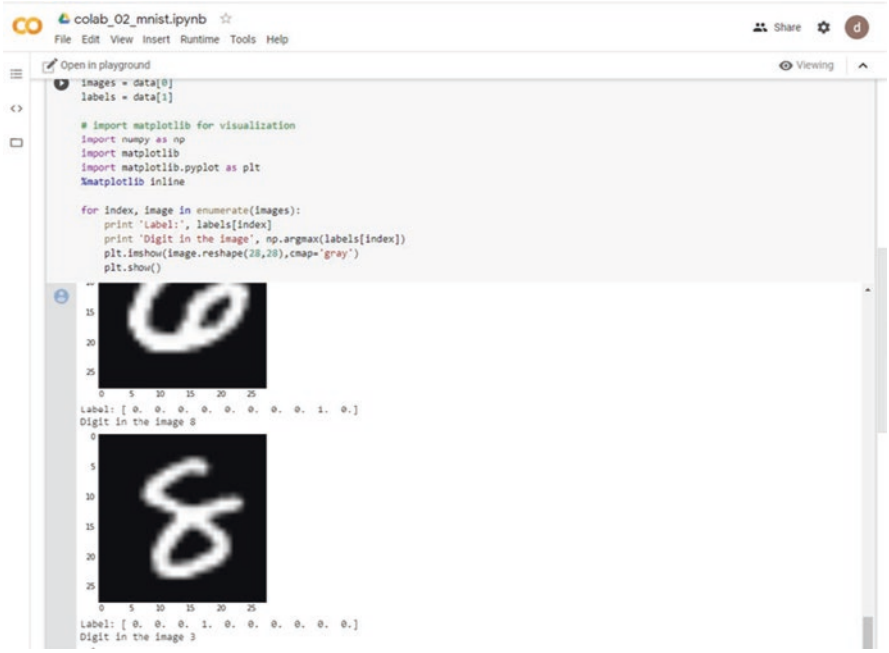


Fig. 7.4 Screenshot of Google Colab

7.6.3 Azure Machine Learning Studio

Azure Machine Learning Studio (Fig. 7.5) is a commercial machine learning software developed by Microsoft as a part of its cloud service Azure. It can be accessed from https://azure.microsoft.com/en-us/services/machine-learning/. It has an

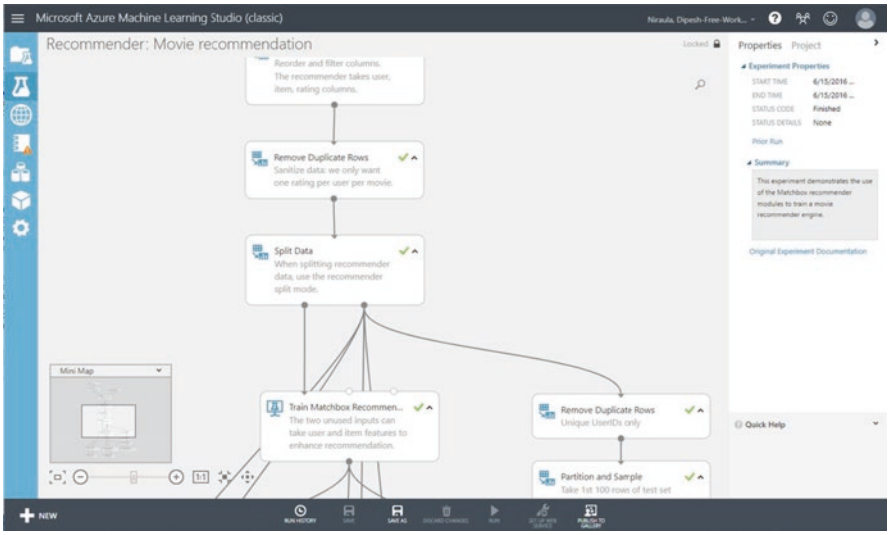


Fig. 7.5 Screenshot of Azure Machine Learning Studio

interactive GUI where one can drag-and-drop tools to create and train a deep learning model, desirable for business solutions. It supports open-source tools and frameworks like PyTorch, TensorFlow, and scikit-learn. Development tools including popular IDEs, Jupyter notebooks, and CLIs or languages such as Python and R can be chosen.

7.6.4 IBM Watson Machine Learning Studio

IBM Watson Machine Learning Studio (Fig. 7.6) is a commercial machine learning platform developed by IBM that operates via the IBM cloud. It can be accessed from https://www.ibm.com/cloud/machine-learning. Similar to Azure ML studio, it has interactive GUI. It is built on open-source platform based on Kubernetes and Docker software packaging components. It presents graphical view of the model and also provides notebooks for interactive programming environment. It also includes distributed training feature and supports GPU, it helps in optimizing hyper-parameters, and deploying trained models and python functions. It supports machine learning frameworks such as scikit-learn, XGBoost, TensorFlow, Keras, Caffe, PyTorch, IBM SPSS Modeler, and many others.

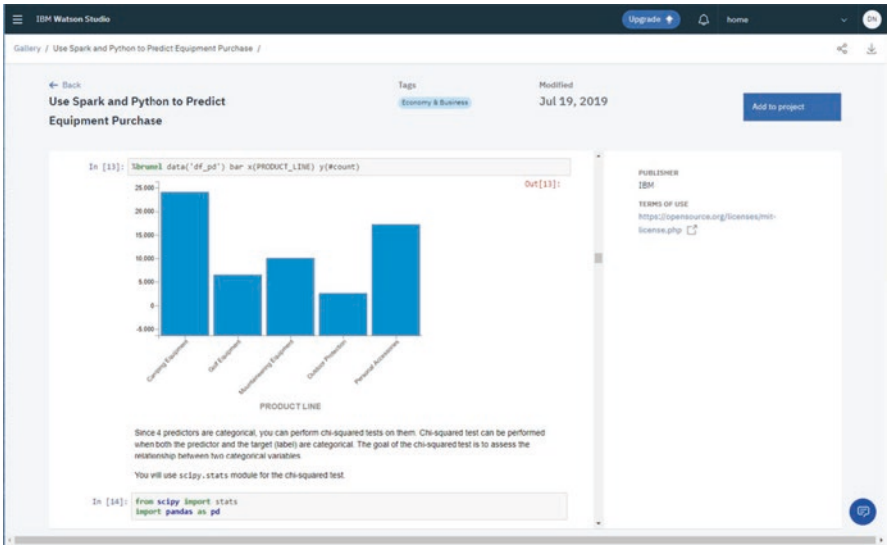


Fig. 7.6 Screenshot of IBM Watson Studio

7.7 Conclusions

Success of deep neural networks has diverted AI research community's focus towards advancing deep learning methods. Consequently, the development of many conventional machine learning libraries has been limited in favor of rapid development of deep learning platforms. Currently, there exits many deep learning platforms, out of which we only presented a few that constitute the most popular ones and likely to pass the test of time. We also presented sample code for using the most common ones: TensorFlow and PyTorch, which can help the reader jump start building their own applications using these libraries whether on-premise or in the cloud.

References

1. TIOBE Index for April 2020, Tiobe. (2020) https://www.tiobe.com/tiobe-index/ cited 2 April 2020.
2. Helmus J, Understanding Conda and Pip, Anaconda. (2018), https://www.anaconda.Com/understanding-Conda-and-pip/. Cited 25 March 2020.
3. Oliphant TE. A guide to NumPy. USA: Trelgol publishing; 2006.
4. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: a structure for efficient numerical computation. Comput Sci Eng. 2011;13:22–30.

5. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey C, Polat I, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. SciPy 1.0 Contributors, SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261âĂŞ272. https://doi.org/10.1038/s41592-019-0686-2.
6. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
7. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, Vanderplas J, Joly A, Holt B, Varo-quaux G, API design for machine learning software: experiences from the scikit-learn project. 2013, arXiv:1309.0238[cs.LG].
8. scikit-learn developers, scikit-learn user guide, Release 0.22.2. 2020. https://scikit-learn.org/stable/downloads/scikit-learn-docs.pdf
9. Sonnenburg S, Strathmann H, Lisitsyn S, Gal V, Iglesias GarcÃŋa FJ, Lin W, De S, Zhang C, Frx, Tklein23, Andreev E, Behr J, Sploving, Mazumdar P, Widmer C, Deng P, De Toni G, Mahindre S, Kislay A, Hughes K, Votyakov R, Khalednasr, Sharma S, Novik A, Panda A, Anagnostopoulos E, Pang L, Binder A, Serialhex, Björn Esser. Shogun-toolbox/shogun: Shogun 6.1.0. Zenodo; 2017. https://doi.org/10.5281/zenodo.1067840.
10. Albanese D, Kessler FB, Visintainer R, Merler S, Riccadonna S, Jurman G, Furlanello C, mlpy: Machine Learning Python. 2012, arXiv:1202.6548 [cs.MS].
11. Hanke M, Halchenko YO, Sederberg PB, Hanson SJ, Haxby JV, Pollmann S. PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI Data. Neuroinformatics. 2009;7:37âĂŞ53. https://doi.org/10.1007/s12021-008-9041-ya.
12. Zito T, Wilbert N, Wiskott L, Berkes P. Modular toolkit for Data Process- ing (MDP): a Python data processing frame work. Front Neuroinform. 2009;2:8. https://doi.org/10.3389/neuro.11.008.2008.
13. Schaul T, Bayer J, Wierstra D, Sun Y, Felder M, Sehnke F, Ruckstieb T, Schmidhuber J. PyBrain. J Mach Learn Res. 2010;11:743–6.
14. Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A, Belopolsky A, Bengio Y, Bergeron A, Bergstra J, Bisson V, Snyder JB, Bouchard N, Lewandowski NB, Bouthillier X, de Brebisson A, Breuleux O, Carrier P-L, Cho K, Chorowski J, Christiano P, Cooijmans T, Cote M-A, Cote M, Courville A, Dauphin YN, Delalleau O, Demouth J, Desjardins G, Dieleman S, Dinh L, Ducoffe M, Dumoulin V, Kahou SE, Erhan D, Fan Z, Firat O, Germain M, Glorot X, Goodfellow I, Graham M, Gulcehre C, Hamel P, Harlouchet I, Heng J-P, Hidasi B, Honari S, Jain A, Jean S, Jia K, Korobov M, Kulkarni V, Lamb A, Lamblin P, Larsen E, Laurent C, Lee S, Lefrancois S, Lemieux S, Leonard N, Lin Z, Livezey JA, Lorenz C, Lowin J, Ma Q, Manzagol P-A, Mastropietro O, McGibbon RT, Memisevic R, van Merrienboer B, Michalski V, Mirza M, Orlandi A, Pal C, Pascanu R, Pezeshki M, Raffel C, Renshaw D, Rocklin M, Romero A, Roth M, Sadowski P, Salvatier J, Savard F, Schluter J, Schulman J, Schwartz G, Serban IV, Serdyuk D, Shabanian S, Simon E, Spieckermann S, Subramanyam SR, Sygnowski J, Tanguay J, van Tulder G, Turian J, Urban S, Vincent P, Visin F, de Vries H, Farley DW, Webb DJ, Willson M, Xu K, Xue L, Yao L, Zhang S, Zhang Y, Theano: A Python framework for fast computation of mathematical expressions. 2016, arXiv:1605.02688 [cs.SC].
15. Tokui S, Okuta R, Akiba T, Niitani Y, Ogawa T, Saito S, Suzuki S, Uenishi K, Vogel B, Vincent HY, Chainer: A deep learning framework for accelerating the research cycle. 2019, arXiv:1908.00213 [cs.LG].
16. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. ManÃĺ, M. Schuster, R. Monga, S. Moore,

D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. ViÃĺgas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf
17. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, De Vito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chin-Tala S. PyTorch: An Imperative Style, High-Performance Deep Learning Library, Advances in Neural Information Processing Systems, vol. 32. 2nd ed. New York: Curran Asso-ciates, Inc.; 2019. p. 8024–35. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
18. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional Architecture for Fast Feature Embedding. 2014, arXiv:1408.5093 [cs.CV].
19. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, Xiao T, Xu B, Zhang C, Zhang Z. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. 2015, arXiv:1512.01274 [cs.DC].
20. Guyon I, Gunn S, Ben Hur A, Dror G. Result analysis of the NIPS 2003 feature selection challenge. In: NIPS'04: Proceedings of the 17th International Conference on Neural Information Processing Systems December 2004; 2004. p. 545âĂŞ552.
21. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248–255.
22. Krizhevsky A. One weird trick for parallelizing convolutional neural networks. 2014, arXiv:1404.5997 [cs.NE]
23. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, OverFeat:Integrated Recognistion, Localization, and Detection using Convolution Networks. 2014, arXiv:1312.6229 [cs.CV].
24. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv:1409.1556 [cs.CV].
25. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, A. Rabinovich, Going Deeper with Convolutions. 2014, arXiv:1409.4842 [cs.CV].
26. Marcus M, Santorini B, Marcinkiewicz MA. Building a Large Annotated Corpus of English: The Penn Treebank. 1993. https://repository.upenn.edu/cis_reports/237/.
27. Frank E, Hall MA, Witten IH. The WEKA Workbench. Online Appendix for"Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, 4th edn. 2016. https://www.cs.waikato.ac.nz/ml/weka/Witten\textunderscoreet\textunderscoreal\textunderscore2016\textunderscoreappendix.pdf.

8 Privacy-Preserving Federated Data Analysis: Data Sharing, Protection, and Bioethics in Healthcare

Ananya Choudhury, Chang Sun, Andre Dekker, Michel Dumontier, and Johan van Soest

8.1 Introduction

Technical advancements in the fields of physics, radiobiology, and engineering (and indirectly chemistry) are the main drivers for better, and thus more specific, treatment opportunities in radiation oncology. These advancements largely influence treatment methods, especially in regard to treatment planning (IGRT, IMRT, VMAT) and radiation techniques used.

In the current era of Evidence based Medicine (EBM), all of these advancements need to be validated to be sure whether a specific treatment (plan) is better than the current standard (e.g. in regard to possible patient outcome). However, we also

Adapted from
Machine Learning in Radiation Oncology: Theory and Applications, Springer International Publishing, 2015, page 71-97
Book Editors: Issam El Naqa, Ruijiang Li, Martin Murphy
ISBN: 978-3-319-18304-6
DOI: 10.1007/978-3-319-18305-3

A. Choudhury (✉) · A. Dekker
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands
e-mail: Ananya.Choudhury@maastro.nl

C. Sun · M. Dumontier
Institute of Data Science, Maastricht University, Maastricht, The Netherlands

J. van Soest
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands

Institute of Data Science, Maastricht University, Maastricht, The Netherlands

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_8

observe that new treatment options do not necessarily improve the outcome for an entire population but might only work for specific groups of patients. The standardized treatment (according to the current guidelines) might be too intense for specific groups of patients (resulting into higher toxicities and/or other radiation-induced complications) or could result in under-treatment of patients. At this point, it becomes interesting to apply machine learning to retrospectively identify prognostic factors (e.g. risk factors) and to develop predictive models to classify patients in distinct groups [1].

These groups can then be used to alter treatment options, e.g. to intensify or temper treatment.The more subgroups we can identify, the better we can optimize treatment for individual patients, leading towards the next era, called Personalized Medicine (PM).

This also imposes challenges on patient subgroup discovery and development of prognostic models as done for many years. Only several large institutions (in terms of patient turnover per year) can perform fine-grained subgroup analysis, as we need a fair number of patients with and without a specific outcome to test hypotheses regarding new treatment options for specific subgroups. Only with these large numbers of patients can we translate results of personalized medicine [2] into clinical practice by means of Clinical Decision Support Systems (CDSS) . In recent years, as newer technologies have evolved around the healthcare ecosystem, more and more data have been generated. Advanced analytics could empower the data collected from numerous sources, both from healthcare institutions, or generated by individuals themselves via apps and devices, and lead to innovations in treatment and diagnosis of diseases; improve the care given to the patient; and empower citizens to participate in the decision-making process regarding their own health and well-being. Sharing health data across institutions and individuals will tremendously benefit individual patients, health research communities, and the whole society. However, the sensitive nature of the health data prohibits healthcare organizations from sharing the data. Ethical, legal, and societal barriers to health data sharing are more impactful than the technical barriers. However, specifically designing infrastructures that cater to the ethical, legal, and societal barriers, a more sustainable radiotherapy data ecosystem can be achieved. The Personal Health Train (PHT) is a novel approach, aiming to establish a distributed data analytics infrastructure enabling the (re)use of distributed healthcare data, while data owners stay in control of their own data. The main principle of the PHT is that data remain in their original location, and analytical tasks visit data sources and execute the tasks. The PHT provides a distributed, flexible approach to use data in a network of participants, incorporating the FAIR principles. It facilitates the responsible use of sensitive and/or personal data by adopting international principles and regulations [3]. Therefore, we need to collaborate in radiation oncology research and share data to perform machine learning on larger, distributed datasets.

In this chapter, we will explain the current possibilities of machine learning in a distributed setting. We will start with the prerequisites and infrastructure

fundamentally needed for distributed machine learning. Afterwards, we will describe the concept of centralized and distributed machine learning, including the benefits and challenges. Finally, we will describe several applications/initiatives related to distributed machine learning and conclude with a summary of this chapter.

Note: In this chapter, the terms multicentre learning, distributed learning, and federated learning are used interchangeably. However, for all of the mentions of these terms, the meaning and intentions remain as federated machine learning from distributed datasets.

8.1.1 Data Landscape

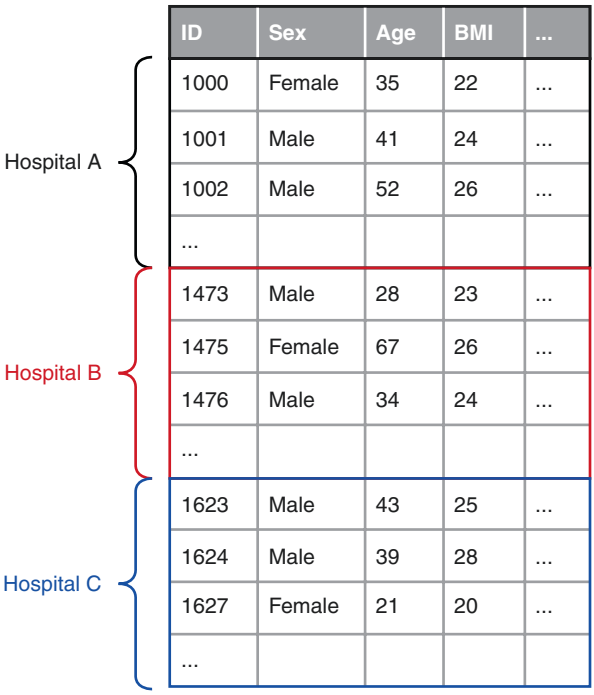
8.1.1.1 Structured Data and Unstructured Data

Radiotherapy data can be both structured and unstructured. Structured data refers to the clinical information of a patient organized in the form of "key-value" pairs and represented as csv data, relational databases, or other healthcare standards. Radiology images, digitized speeches, free texts (clinical notes), scans, etc. are unstructured information. In this chapter, we will predominantly talk keeping in mind only the structured information. Unstructured data such as clinical notes can be converted to structured information by natural language processing tools. Radiomics information extracted from radiology information can be considered structured information if suitably represented.

8.1.1.2 Horizontally Partitioned Data and Vertically Partitioned Data

In addition to data structure, data partitioning which is another key factor needs to be considered in health data sharing. In practice, data are mainly partitioned in two different situations. The first situation is referred to as horizontally partitioned data which contains the same features from different data instances (e.g. patients). For instance, patients' health data can be collected by several independent healthcare providers. Each provider collects the same features from different patients such as demographics information, hospital records, and laboratory results. Combining and analyzing data of various patients from different healthcare providers will result in a better understanding of certain diseases. Figure 8.1 shows an example of horizontally partitioned data.

The second situation is referred to as vertically partitioned data, where different data (healthcare) providers collect different features but from the same group of patients. For example, one healthcare provider collected patients' demographics information, hospital records, and laboratory results, the other one owns previous diagnosis, current medication, and treatment plan of the same group of patients. When we need to predict the best treatment plan for some patient, data from both healthcare providers are required. Combining vertically partitioned data will enlarge the observations of patients and knowledge of certain diseases. Figure 8.2 shows an example of vertically partitioned data.



ID	Sex	Age	BMI	...
1000	Female	35	22	...
1001	Male	41	24	...
1002	Male	52	26	...
...				
1473	Male	28	23	...
1475	Female	67	26	...
1476	Male	34	24	...
...				
1623	Male	43	25	...
1624	Male	39	28	...
1627	Female	21	20	...
...				

Fig. 8.1 Example of horizontally partitioned data

Hospital A Clinic C

ID	Sex	Age	BMI	...	Insulin	T2D	...
1000	Female	35	22	...	24mlU/L	No	...
1001	Male	41	24	...	None	None	None
1002	Male	52	26	...	28mlU/L	Yes	...
1004	Female	23	21	...	22mlU/L	No	...
...							

Fig. 8.2 Example of vertically partitioned data

8.2 Prerequisites

Privacy-preserving data sharing in healthcare is dependent on a number of prerequisites. These need to be addressed carefully before actually starting the machine learning process. In this paragraph, we will describe the topics of data extraction (Sect. 8.2.1), representation and FAIR data principles (Sect. 8.2.2), network infrastructures (Sect. 8.2.3), distributed learning algorithms (Sect. 8.2.4), and data protection (Sect. 8.2.5).

8.2.1 Data Extraction

Distributed data analytics on radiotherapy data requires the data to be available in a meaningful manner. Within radiation oncology, data extraction for machine learning is a labour-intensive task, as many data silos (isolated data collection within an organization) exist where data resides. In general, we need to connect to different data sources, extract data from these sources using local querying dialects, and afterwards store the extracted data in a central storage within the hospital. These steps need to be performed for different information systems used in radiation oncology. We will describe the most common systems in this paragraph. First, we need to include the Electronic Medical Record (EMR), where general patient characteristics are stored (e.g. age, gender, and diagnostic, geographical, and follow-up information such as complication and quality of life scores). Second, medical images (for diagnostic, treatment, and validation purposes) are stored in a picture archiving and communication system (PACS). Although images cannot be used directly in predictive model training, extracted information from these images can be used. Third, treatment planning-related information (e.g. radiation plan information regarding beams and dose) needs to be incorporated, as the treatment planning system (TPS) stores information in its own database, as well as in the PACS. Fourth, the record and verify system (R&V) holds information regarding the planned treatment (e.g. dose, fractionation, beams) and the actual delivery. This information is also needed during machine learning, e.g. to determine structural differences in the planned and delivered treatment. Other systems (e.g. sources containing biological data) may apply in specific or future settings; however, we've specified only the general sources of information used for machine learning.

In the distributed learning setting, extracting data from different information systems within the hospital is a challenging task. Different institutes use products from different vendors and as such need a customized approach for data extraction. The extracted data then has to be curated in a manner that the data is Findable, Accessible, Interoperable, and Reusable (FAIR) [3] (Sect. 8.2.2).

8.2.1.1 ETL Tooling and Data Warehousing

To (continuously) extract data and store it in a central location, one could consider the use of extraction, transformation, and load (ETL) tooling. This tooling can extract data from different sources (different systems), reconcile data belonging to

one patient (transformation), and store the data in a central database: the data warehouse (DWH). This could be useful for large-scale machine learning and research institutions with many smaller-sized trials. As shown by Roelofs et al. [4], implementing a data warehouse can significantly reduce the data collection time, in comparison to manual data extraction and collection. In regard to distributed settings, this also reduces the number of systems/databases a user/researcher has to include in the data request/retrieval process, thus reducing the time to merge all different datasets. Furthermore, as data are extracted and inserted into the DWH, it should be known what the data represents. The ETL process should therefore be well documented regarding queries, transformations, and the meaning of the stored data in the DWH. In comparison to the DWH, directly querying the source system for research purposes has several disadvantages. These disadvantages are mainly on the topics of query and data validity and query load on production/source systems. When a DWH is in place, query validity should not be an issue (as the data is checked before being incorporated in the DWH). Furthermore, query load issues should be mitigated, as the DWH should run on a different database/server as the production/source systems, and therefore cannot affect clinical operations.

8.2.1.2 Image Biomarker Extraction

As stated in Sect. 8.2.1, the intrinsic information of images (not just the readily available metadata) needs to be extracted from the actual image slices. Extraction of image "features" is not a standard functionality of a PACS; however, features may sporadically be available as TPS systems may store additional information in the metadata of the DICOM images. If features are stored in the metadata, these values are needed to be validated, especially in a distributed setting where different sites may use different TPS systems, which could implement different algorithms to calculate these features.

When there are no (or only a small number of) features already available, every site in the distributed setting needs to implement a feature extraction pipeline which calculates variables based on the images available in the local PACS. As the local PACS stores CT and/or PET images, delineated contours (RTSTRUCT), planned (RTPLAN), and delivered (RTDOSE) dose information, the number of features to extract becomes larger. For example, we can extract information regarding the tumour volume, maximum diameter, specific points of the dose-volume histogram (DVH) for target volumes or organs at risk, tumour activity/metabolism, and differences between planned versus delivered dose. Furthermore, radiomic analysis on these images produces more than 200 features, based on more advanced image processing algorithms (by calculating intensity distribution metrics based on, e.g. Fourier transformations and wavelets) [5]. Several of these features are potential imaging biomarkers: features which have prognostic and predictive value in terms of patient outcome or tumour response.

Preferably, this feature extraction pipeline should use common communication protocols, such as DICOM (to receive images) and SQL (to send extracted features to a local database). This increases the possibility to reuse this pipeline in all submitting centres and increases the homogeneity of applications and calculation

algorithms used by different centres. Eventually, using equal feature extraction pipelines should result in easier comparison of features/variables between centres. Although we can generalize the applications and algorithms used, including scanning and reconstruction parameters, there is still a large variability at the input of this feature extraction pipeline: differences between delineations of different centres. As shown in literature, differences in delineations may occur between individuals, even within one site [6]. These differences in delineations could result in different outcomes after feature extraction. Especially when two different structures (e.g. rectum and bladder) are close to each other, for example, it might be possible that the delineating individual accidentally delineates the bladder wall as part of the rectum. This results in a higher SUV-mean/max and therefore could compromise the prognostic value of the extracted features.

Based on the examples of delineation differences and calculation applications/algorithms used, it is important to specify the provenance of a specific variable: how did we acquire/extract this information (and which algorithms did we use)? And what are the sources used to extract the information? We will elaborate on these questions in the next section.

8.2.2 Data Representation and FAIR Data Principles

One of the prerequisites for distributed machine learning using PHT is that the data needs to be FAIR. FAIR principles emphasize on enhancing the machine interpretability of data and data reuse. Data agnostic machine learning algorithms rely on the FAIR data descriptions published on each hospital FAIR endpoints. Data stewardship at the source plays an important role for driving distributed and federated machine learning a reality. The FAIR foundational principles are explicitly described by a more detailed 15 guiding principles.

Radiotherapy data can be made Findable by assigning globally unique and persistent identifier and associating sufficient metadata. The metadata description is registered or indexed in a searchable resource. Some literature and implementation emphasize on using digital object identifiers for data and services involved in a distributed machine learning [7, 8].

FAIR repository at the radiotherapy departments in the hospital should host data in a way that is accessible using a globally acceptable, free and implementable protocol. Semantic Web technologies like the Resource Description Framework (RDF) and HL7 Fast Healthcare Interoperability Resources (FHIR) provide RESTful way for querying and accessing data.

Finally, FAIR emphasizes on the importance of data interoperability at the source. Interoperability as defined in the IEEE standard glossary "…is the ability of two or more systems or components to exchange information and use the information that has been exchanged". Each cross-institutional data exchange needs data to be *syntactically* and *semantically* interoperable. Syntactic interoperability means that the different stakeholders have to agree which (technical) protocol they use to transfer data, implying that data representation should be equal among participating

sites.One way of making data interoperable is standardizing data at the source. However, standardization alone is not sufficient and often different stakeholders may choose to implement different standards (HL7 Version 2.X and 3.X, OpenEHR and ISO 13606, HL7 CDA, XDS, OHDSI OMOP, etc.). Next to standardization of syntactical interoperability, *semantic interoperability* needs to be in place. We will use the definition of Valentini et al. [9] to describe semantic interoperability: "The ability of any communicating entity (not only computers) to share unambiguous meaning. For computers, this is the ability to exchange information and have that information properly interpreted by the receiving system in the same sense as intended by the transmitting system". In general, this means that the receiver cannot interpret information differently, as the sender uses unambiguous terms to describe that information. Therefore, we need to use terminological systems which are known by both sender and receiver. As defined by De Keizer et al. [10], a terminological system can be a thesaurus, classification, vocabulary, nomenclature, or coding system. A terminological system may pertain to more than one of these systems. For example, ICD-10 [11] is a coding system and vocabulary (as the term is accompanied by a definition); another example is the National Cancer Institute's Thesaurus (NCIT) [12], which (in addition to a vocabulary) also contains a list of synonyms or other relationships. SNOMED CT is extensively used in healthcare systems for coding concepts for diseases, findings, procedures, and substance [13]. Finally, multiple terminological systems can be embedded in an ontology, where concepts from terminological systems are reused and relations of concepts in a specific domain are described. Furthermore, an ontology can be used as a consensus model to represent data within a specific domain (e.g. radiation oncology) between different participating sites [10].

8.2.2.1 Relational Databases and Ontologies

In regard to distributed learning, we need to make sure every participating site uses the same database structure to be able to uniformly query (or federate) the data warehouse (DWH) database (Sect. 8.2.1.1). This database structure can be derived by creating a so-called entity-relationship (ER) model, based on the ontology; however, it needs to be adhered by all centres. An example to derive this ER model is the normalized universal approach described by Gali et al. [14]. Next to this database structure, it is important to use the same database system, as different database systems/vendors have different dialects. To mitigate differences in database systems/vendors, it is also possible to use automatic conversion libraries such as Hibernate (http://hibernate.org/), although these systems add another layer of complexity when performing queries and/or data federation.

When adhering to an ontology, values from local systems need to be replaced with standardized values from terminological systems as defined in the ontology. For example, the property biological sex containing the text "male" or "female" needs to be replaced by NCI Thesaurus code C20197 or C16576, respectively. Another participating site may use 0 and 1 or "m" and "f"; however, within the DWH database, all sites should use the NCI Thesaurus codes for semantic interoperability. This conversion of values is typically done in the *transform* step of the

ETL process. Therefore, the ETL process needs to be tailored per participating centre.

Although data representation is possible within relational databases, it is cumbersome to maintain in a distributed machine learning setting. As new results give new insights into biological concepts and relationships, the need for extra variables is rapidly growing. Given this fact, it is inevitable that a distributed network for machine learning will have substantial downtime. For example, when a new concept is added to the ontology, every participating site needs to update their ETL system and DWH database structure, to become up to date with the new ontology version. This may take some time, as administrators of the ETL and DWH system need to validate whether this change is valid and does not compromise patient de-identification. If one of the queried columns is not available, the Relational Database Management System (RDBMS) will result an error rather than an empty result set. Therefore, it might be that the whole federation/distributed querying system may not work (if proper error handling is not in place). In this example, we used the addition of a column, a relatively easy task which occurs frequently. However, the more complex the changes in the ontology and database structure, the more time and effort it will take to get the network up and running again.

8.2.2.2 Semantic Web, RDF, and Linked Data

One of the solutions to cope with rapidly changing ontologies in a distributed setting is to move from relational databases to Semantic Web technologies [15, 16]. In this paragraph, we will only discuss the *Resource Description Framework* (RDF), *linked data*, and the *SPARQL protocol and RDF query language* (SPARQL) as a subset of Semantic Web technologies.

Resource Description Framework

RDF is a standard, recommended by the World Wide Web Consortium (W3C) [17], and can be seen as a flexible alternative for the relational database. Where "traditional" relational databases store their data in a structure of tables and columns, the RDF specifies only one table with three columns named *subject*, *predicate*, and *object*. Each row in this single table repository is called a triple, as it only has three cells. Due to this basic difference in structure, the concept of data representation is also different. Because of this fixed table structure, the ontology becomes more important and serves as a data model consensus between centres.As an example, we have an ontology describing patients and their first name, last name, biological sex, and age. Shows the visual representation of this ontology. The RDF triples based on this ontology are represented in. Figure 8.3 shows the rdf representation of the example ontology.

Unique Resource Identifiers and Linked Data

To assure semantic interoperability, we will use the concept of unique resource identifiers (URIs), which is incorporated in the RDF specification. The RDF specification states that all resources (concepts and predicates) need to have a URI, which can be a unique resource locator (URL; "http://www.mydomain.org/

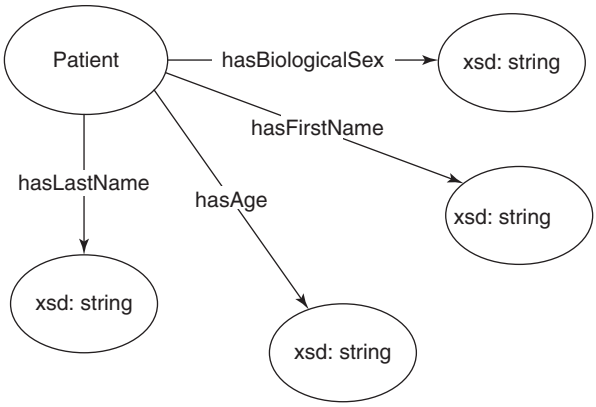


Fig. 8.3 Visual Representation of the sample ontology

ontology#hasFirstName"main.org/ontology#hasFirstName) or a unique resource name (URN; e.g. myOntology:hasFirstName). This means that someone needs to own a domain name (e.g. mydomain.org) and is administrator of this domain. If this is the case, he or she can make unique URLs for this domain, for example, to create a unique URI for patient 1001 (e.g. http://www.mydomain.org/rdf#patient1001). If the domain administrator assigns a specific sub-path of the domain to a dataset (called a *namespace*), for example, http://www.mydomain.org/rdf#, then this sub-path can also be substituted by a *prefix*, for example, "mySet". This namespace can then be used to shorten the notation of a unique patient, as shown in. This concept of unique resources also holds for ontologies, wherein the prefix "myOntology" can be used to define the namespace http://www.mydomain.org/ontology# and the prefix "ncit" refers to the unique location of the NCI thesaurus. As everyone should use the same, unique namespaces, the use of URIs enforces semantic interoperability. Therefore, semantic interoperability is enforced within the resource description framework.

Next to the enforcement of semantic interoperability, the use of URIs has a second benefit, namely, the possibility of linked data. As every resource has its unique URI, an RDF store at site A may point to a resource at site B by using the URI of the resource at point B [18]. For example, if a patient underwent a diagnostic scan at hospital A and was treated in clinic B, then clinic B can specify the treatment and link it to the patient resource with the unique URI used in hospital A.main.org/ontology#hasFirstName) or a unique resource name (URN; e.g. myOntology:hasFirstName).

Querying Using SPARQL

We have described how data can be represented in RDF, and how URIs enforce semantic interoperability and linked data. But how can we retrieve this data from an RDF store? To query these RDF stores, the W3C has adopted the *SPARQL protocol*

and RDF query language (SPARQL) [19]. Most RDF stores have integrated a SPARQL endpoint in their RDF store. A SPARQL endpoint is the public interface to receive SPARQL queries and return a result table, all using the HTTP protocol. In contrast to SQL queries, SPARQL queries do not search tables due to the underlying RDF store structure. SPARQL queries perform pattern matching on the triples in the triple store, where variables can be used to retrieve unknown values or to dynamically link values. For example, the query in Listing 8.1 will try to retrieve the first name, last name, and age for all patients. We will shortly describe the lines in this query example.

On line 1–3, the shorthand (prefix) notations for URL locations are defined. Line 5 defines the variables retrieved from the pattern matching; these variables have to start with a question mark. Lines 6–11 define the actual pattern searched for. As shown in Listing 8.1, our basic pattern is to retrieve all patient resources which have a predicate called "rdf:type", which refers to the terminological code of a patient, defined in the NCI Thesaurus (using the prefix "ncit:", which is replaced by the full URL at line 3). Afterwards, we extend our pattern match by including extra properties for every resource linked to the patient resource. If the linked resources of the patient variable have a predicate matching to our specified property (in our ontology), then the variable firstName, lastName, or age will be filled with the found value. If not found, then the query will return the patient resource URI; however, the variables firstName, lastName, or age are not filled in (due to the "OPTIONAL" keyword).

Next to querying one RDF store, a SPARQL query can also be federated to multiple stores. This is an advantage in regard to distributed learning, as a single query can retrieve data from multiple sources. Due to the structure of RDF stores, data residing in geographically separated RDF stores can easily be merged, as the data structure is the same for all stores (1 table; 3 columns) and all RDF stores should use URIs. Federation can be done both horizontally (different patients in different RDF stores) and vertically (information of a single patient stored in multiple RDF stores).

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX ontology: <http://www.mydomain.org/ontology#>
3  PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
4
5  SELECT ?patient ?firstName ?lastName ?age
6  WHERE {
7     ?patient rdf:type ncit:C16960 .
8     OPTIONAL { ?patient ontology:hasFirstName ?firstName . }
9     OPTIONAL { ?patient ontology:hasLastName ?lastName . }
10    OPTIONAL { ?patient ontology:hasAge ?age . }
11 }
```

Listing 8.1 Basic SPARQL query retrieving patient resources, related first and last names, and age of patient data stored in an RDF store, based on the ontology defined in Fig. 8.3

An application of horizontal federation in SPARQL queries is shown in Listing 8.2; an application of vertical federation is shown in Listing 8.3.

In these examples, we will use the "SERVICE" command of SPARQL to identify the execution of a subquery (or pattern match) on a different SPARQL endpoint. In Listing 8.3, we used the exact same pattern query in both services/subqueries (line 7–19). Both subqueries are sent to the respective endpoints, and the subquery results are merged at the federation endpoint. Finally, the requested variables are returned to the requesting application or user. In Listing 8.3, both services have different patterns to match. The first service (line 7–11) searches for all patients and their first/last name on SPARQL endpoint 1. The second service (line 13–15) will reuse the patient resources found in endpoint 1 and tries to find patterns matching the hasAge predicate for these given patient resources. When found, it will use the object linked to the hasAge predicate (in this case a literal of type integer) and store it in the variable "?age". Finally, the query engine will return the output as one table (using the variables of line 5 as columns), including information retrieved from both endpoints.

In this paragraph, we have presented an alternative to the widely known relational databases to represent and retrieve data. The use of Semantic Web technologies, and especially RDF, has several advantages over relational databases. Especially the meta-structure of RDF (independent of the modelled domain) and the use of URIs are useful with regard to a flexible storage solution while inherently adopting semantic interoperability and linked data.On the other hand, using

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX ontology: <http://www.mydomain.org/ontology#>
3  PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
4
5  SELECT ?patient ?firstName ?lastName ?age
6  WHERE {
7          SERVICE <http://endpoint1.mydomain.org/> {
8                  ?patient rdf:type ncit:C16960 .
9                  OPTIONAL { ?patient ontology:hasFirstName ?firstName . }
10                 OPTIONAL { ?patient ontology:hasLastName ?lastName . }
11                 OPTIONAL { ?patient ontology:hasAge ?age . }
12         }
13
14         SERVICE <http://endpoint2.mydomain.org/> {
15                 ?patient rdf:type ncit:C16960 .
16                 OPTIONAL { ?patient ontology:hasFirstName ?firstName . }
17                 OPTIONAL { ?patient ontology:hasLastName ?lastName . }
18                 OPTIONAL { ?patient ontology:hasAge ?age . }
19         }
20 }
```

Listing 8.2 An example of horizontal federation in a SPARQL query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ontology: <http://www.mydomain.org/ontology#>
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>

SELECT ?patient ?firstName ?lastName ?age
WHERE {
        SERVICE <http://endpoint1.mydomain.org/> {
                ?patient rdf:type ncit:C16960 .
                OPTIONAL { ?patient ontology:hasFirstName ?firstName . }
                OPTIONAL { ?patient ontology:hasLastName ?lastName . }
        }

        SERVICE <http://endpoint2.mydomain.org/> {
                OPTIONAL { ?patient ontology:hasAge ?age . }
        }
}
```

Listing 8.3 An example of vertical federation in a SPARQL query

Semantic Web technology has some downsides when used in distributed machine learning. The main downside is that local institute staff needs to be introduced to Semantic Web technologies, in order to maintain these data repositories and endpoints. Furthermore, development in the field of RDF stores/repositories is an ongoing process and is not yet comparable to relational databases in terms of reliability and performance, especially in daily clinical practice. On the contrary, for research projects (where uptime is less critical), the Semantic Web is more favourable because of its flexibility in storage and data structures.

8.2.2.3 HL7 FHIR and REST-APIs

Another way to ensure that the "I" part of the FAIR principles is met is to use HL7 FHIR data standardization. FHIR describes exchangeable information in the form of resources. Resources are the building blocks of FHIR, well defined by a set of metadata description and a human readable part. Data elements and metadata within a resource are associated with a suitable coding terminology, thus emphasizing also on the semantic interoperability aspect. Each FHIR resource is a unique entity identified by a unique identifier (URI) and represented in XML or JSON format. This means that all FHIR resources can be accessed in a RESTful way. In addition to describing the resources, FHIR also provides REST application programming interface to query these resources. The central resource of FHIR is the patient resource. All other resources are built on top of this resource and are linked by URIs. Figure 8.4 shows the relationship of the Patient resource to other resources in FHIR.

Figure 8.5 shows the visual representation of the information from Fig. 8.3 and and Table 8.1: RDF representation of a patient based on the ontology of Table 8.1.

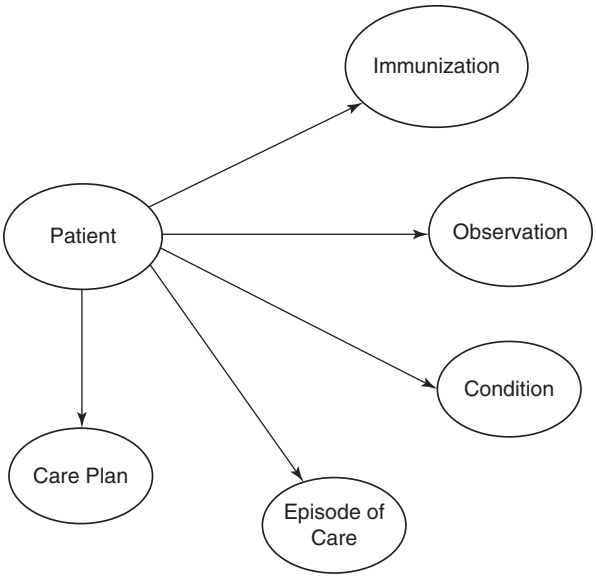


Fig. 8.4 Snapshot of the relationship of the Patient resource to other resources in FHIR

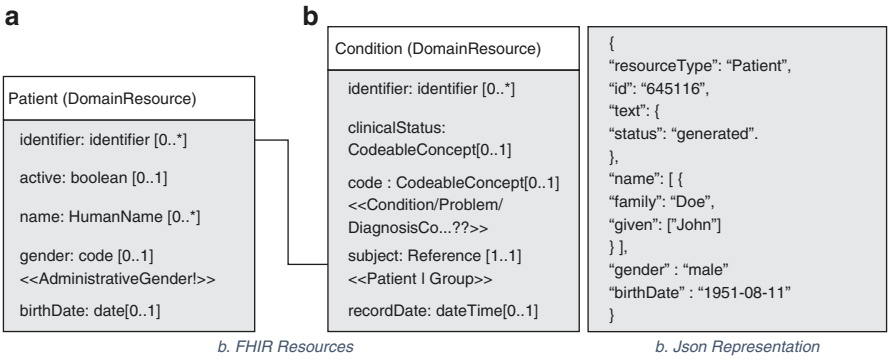


Fig. 8.5 (**a**) shows the visual representation of the information from Fig. 8.3 and Table 8.1. Additionally, we also show how the *Patient* resource is linked to the *Condition* resource through the resource identifier of the *Patient* resource. (**b**) shows the json representation of an instance of the patient resource

Table 8.1 RDF representation of a patient based on the ontology of Fig. 8.3

Subject	Predicate	Object
mySet:patient1001	rdf:type	ncit:C16960
mySet:patient1001	myOntology:hasFirstName	“John”^^xsd:string
mySet:patient1001	myOntology:hasLastName	“Doe”^^xsd:string
mySet:patient1001	myOntology:hasBiologicalSex	ncit:C20197
mySet:patient1001	myOntology:hasAge	“67”^^xsd:integer

Additionaly, we also show how the *Patient* resource is linked to the *Condition* resource through the resource identifier of the *Patient* resource.

The patient records in the form of resources can be queried with the FHIR REST API. Below we show an example FHIR query for retreiving all patients born after 1st January, 1970 who are diagnosed with primary neoplasm of lung (lung cancer). The `<base_url>` is the address of the FHIR server hosting the resources.

Example: `<base_url>Condition?_include=Condition:patient.birthDate=le1970-01-1&code=http://snomed.info/sct|93880001`

8.2.3 Network Infrastructure

The previous section describes how to extract information from multiple sources (databases, image archives) and to apply standardized terminological systems on the data extracted from these sources. Furthermore, we have described the importance of FAIR data representation in a privacy-preserving federated machine learning setup. In this paragraph, we will combine the topics of the previous paragraphs and explain how we can use them together. First, we will describe the institutional infrastructure, after which we will describe the privacy-preserving distributed/federated machine learning infrastructure, i.e. Personal Health Train (PHT) [8].

8.2.3.1 Institutional Infrastructure

The institutional infrastructure lays the foundation of fetching data from many different systems within the hospital and provides a single access point for the outside world (e.g. participating sites in the distributed machine learning setting). We will describe two different approaches:

- Traditional ETL and DWH.
- FAIR data repository.

Traditional ETL and DWH

In the approach using relational databases (Sect. 8.2.2.1), records from different source systems (e.g. EMR, PACS, TPS, and R&V) are merged using an ETL tool and converted into the requested data formats following standards used by all collaboration. The merged and transformed data are being saved in the DWH database. This database will afterwards be queried when requesting data for machine learning purposes. Therefore, this database needs to be compliant to the ontological structure (among all participating centres). When the ontology is altered, all participating centres need to update the DWH database structure, as well as the transform and/or storage scripts in the ETL tooling. Figure 8.6 shows the institutional infrastructure for getting data from multiple sources into a single DWH within the hospital.

FAIR Data Store

Different hospitals use products from different vendors and follow different healthcare standards. This means that the data representation at the source may be different for different centres (Sect. 8.2.2). Instead of hosting data in a SQL DWH with

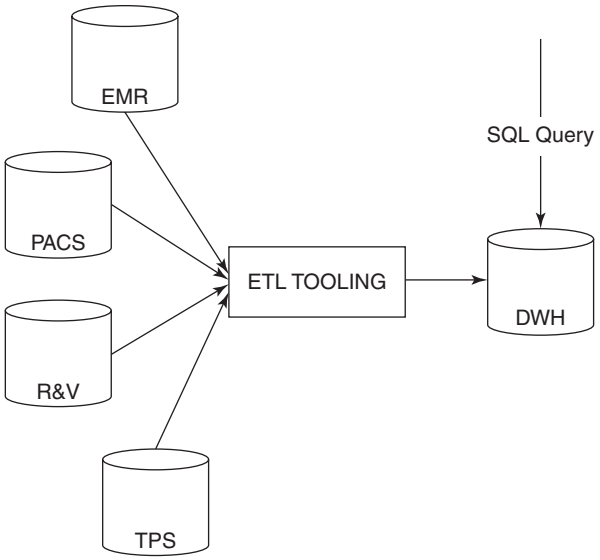


Fig. 8.6 Infrastructure of the Traditional ETL and DWH approach

exactly same schema, hospitals may store the data in a FAIR data endpoint with enough metadata description. Each centre may choose different representations locally while following the FAIR data principles. We describe four different approaches for hosting FAIR data endpoint within the centres.

- Traditional ETL and DWH with a FAIR repository.
- Traditional ETL and DWH with a virtual FAIR repository.
- Virtual FAIR repository per institute.
- Virtual FAIR repository per source and institute.

Traditional ETL and DWH with a FAIR Store

This approach uses an FAIR data endpoint on top of the traditional ETL and DWH approach (Fig. 8.7). It enables the possibility to create an institutional DWH instead of a DWH dedicated for the study. Afterwards, the "Database to FAIR data" conversion application reads the DWH database and transforms the data, taking into account a given ontology. This FAIR endpoint will afterwards be queried when requesting data for machine learning purposes. Only the "Database to FAIR endpoint" application needs to follow the rules and data structure defined in the ontology. When the ontology is altered (e.g. adding an extra data element), only this database-to-FAIR application needs to be altered (when the information is already available in the DWH). Updating the FAIR store is done by clearing and repopulation and is performed at specific time intervals.

Traditional ETL and DWH with a Virtual FAIR Store

This approach uses only the database-to-FAIR conversion application on top of the traditional ETL and DWH approach (Fig. 8.6). This approach is almost equal to the

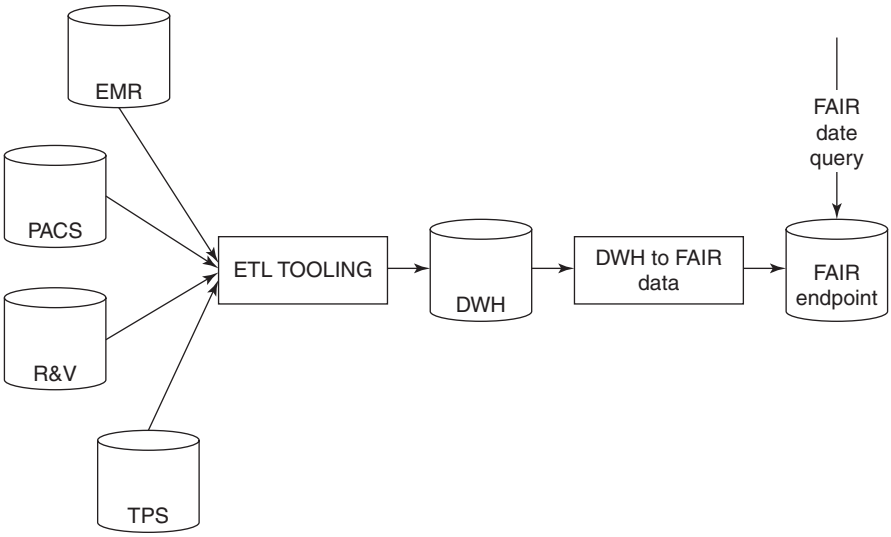


Fig. 8.7 Infrastructure of the approach using a traditional DWH with a FAIR store

physical FAIR store approach (Fig. 8.7); however, it has one difference in converting data from relational databases to FAIR.

In this case, the "Database to FAIR" application acts as a SPARQL endpoint or FHIR endpoint, accepting SPARQL queries or FHIR queries and returning the result of these queries. There is no data stored, as there is no RDF store, only a SPARQL endpoint or FHIR endpoint. When performing a SPARQL query, the database-to-FAIR application will transform SPARQL queries and/or FHIR queries into SQL queries and executes these SQL queries on the DWH. In regard to maintenance, this option holds the same requirements as using the physical RDF store. The only difference is the absence of an intermediate RDF store, resulting in real-time results of the data available in the DWH (Fig. 8.8).

Virtual FAIR Store per Institute

As the DWH usually is not a real-time representation of the clinically available data, this approach removes the DWH and directly queries the source systems. In this approach, the database-to-FAIR application is functioning as a SPARQL endpoint without an RDF store and converts SPARQL queries into SQL queries for the different source systems (Fig. 8.9). It therefore creates challenges for the database-to-FAIR application, as it needs to transform data (to convert local terms to standardized terms), which was previously done by the ETL tooling. If multiple source systems are involved, the database-to-FAIR application merges the results from all sources and presents them as a SPARQL query result. The main benefit of this approach is that we can query for real-time data, rather than have to wait before the data is added to the DWH. Furthermore, data redundancy of the intermediate

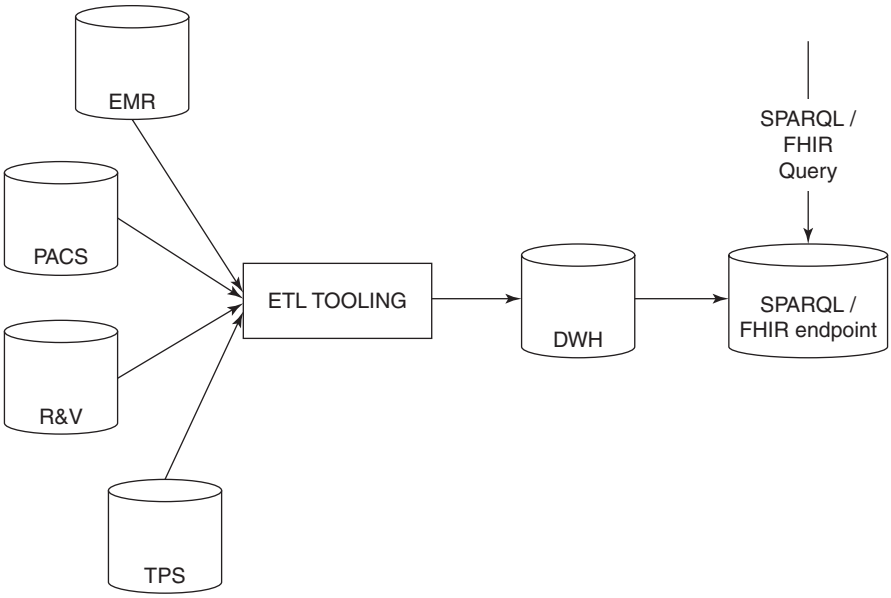


Fig. 8.8 Infrastructure of the approach using a traditional ETL and DWH with virtual RDF store

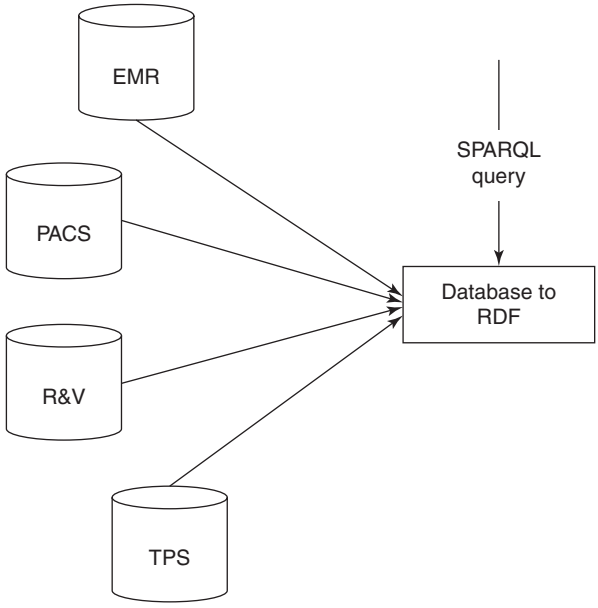


Fig. 8.9 Infrastructure using only a virtual FAIR store

storage (the DWH) is not needed, reducing the need for storage resources. However, the main disadvantage is with regard to performance, as data and queries are transformed on the fly.

Virtual FAIR Store per Source and Institute

This approach is almost similar to the "Virtual RDF store per institute" approach, however, with differences in data transformation and federation (Fig. 8.10). First, every local data source will get a FAIR endpoint, using, for example, the database-to-FAIR application. This application will convert the data from the source system into RDF, compliant with the ontology used in the distributed setting. Afterwards, the central federation endpoint will be used to merge all data elements from all database-to-FAIR applications/sources (vertical federation). In this setting, one SPARQL query will be sent to the federation endpoint. This federation endpoint will split the SPARQL query into several sub-SPARQL queries and execute these SPARQL queries on the SPARQL endpoints placed on top of the data sources. Afterwards, the federation endpoint will merge the results and return the merged result set to the application/user performing the query. The benefit of this approach is the distribution of computational resources to reduce the query execution time. The drawback is that $n + 1$ application (where n is the number of database-to-FAIR applications) needs to be maintained and updated when the ontology changes.

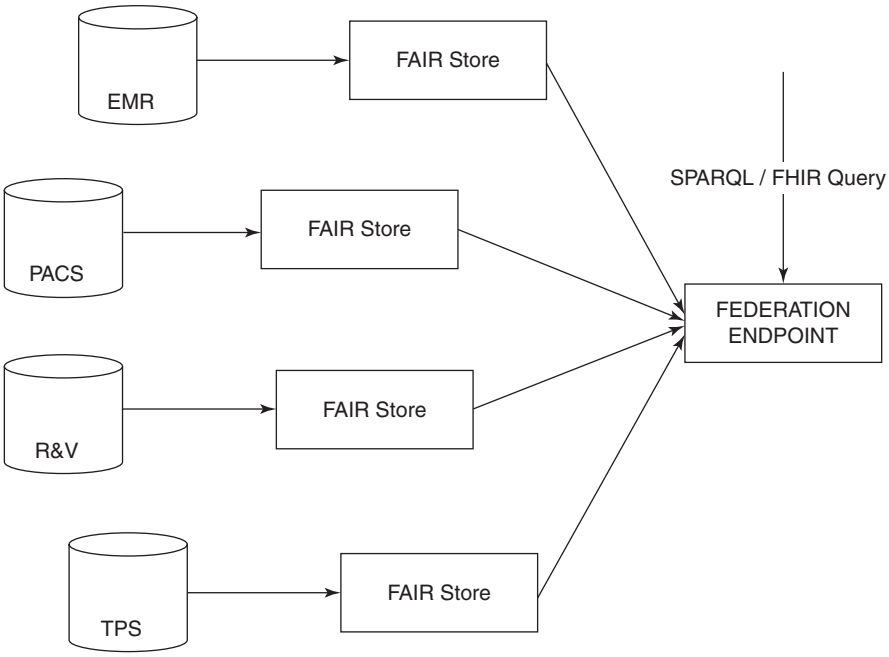


Fig. 8.10 Infrastructure using a virtual FAIR store per source per institute

8.2.3.2 Machine Learning Infrastructure

In the previous paragraph, we described the institutional infrastructure options to create one façade or data query endpoint for every centre. It depends on whether we are using centralized or distributed machine learning and whether we need an additional computation unit (e.g. a dedicated or virtual server) in each centre. Both distributed and centralized approaches can be implemented using relational databases or FAIR data store. The participating centres may choose different data representations at the source, but must agree to publish the metadata descriptions publically. In this paragraph, we will first describe the centralized machine learning infrastructure and afterwards move towards PHT.

Centralized Machine Learning Infrastructure

The general overview for the centralized multicentre infrastructure is shown in Fig. 8.11. The participating sites are displayed as a data store, as we do not need to know what the institutional infrastructure looks like. This approach gives participating centres the opportunity to establish the institutional infrastructure according to local policies. Additional to all institutional entry points, a central machine learning server (performing the computations) and a central federation point need to be set up. The central federation point will perform the horizontal federation between participating centres. To ensure privacy, the data stores of the participating centres may limit external access by only allowing access from the central federation point. The central machine learning server will accept and execute algorithms (including queries to execute on the central federation point). After the algorithm has finished, it

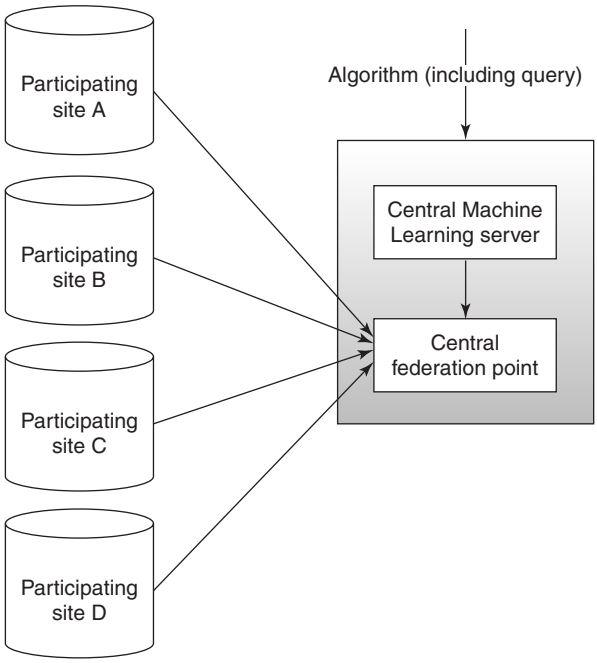


Fig. 8.11 Centralized Machine Learning Infrastructure

will return the outcome of the computation to the external source which sent the job (algorithm + query). The researcher initiating the machine learning process will have no direct access to the data. However, it is important to mention that the data will be located outside the centres and as such may pose privacy and confidentiality issues related to sharing of patient data.

Distributed Machine Learning Infrastructure: The Personal Health Train

In this section, we will discuss the privacy-preserving machine learning in a distributed setting using PHT. The core rationale of PHT is to ensure data privacy by keeping data at the source. This means that, unlike traditional approaches where data is collected centrally, PHT sends algorithm to the data source and fetches only the result. Each participating site or centre generates one or more set of results that are aggregated to obtain a global analysis result. Since no centre needs to share data, PHT stands as a privacy by design infrastructure. PHT is different from the centralized version with respect to computational locations. As shown in Fig. 8.12, the central federation point has been removed, and local computation units (machine learning slaves/agents) have been introduced. In this infrastructural setting, the central machine learning server is a coordinating server. When a job (algorithm + query) is submitted to the central ML master, the algorithm is being split into smaller sub-algorithms. These sub-algorithms and queries are packed into sub-jobs and sent towards the local computation units. They will query the local endpoint and execute the sub-algorithm. After finishing the sub-algorithm, the results are sent back to the central ML master, which gathers the results from all local endpoints. The central master will then determine whether it will perform a new sub-job on all endpoints or aggregate values and sends the final (aggregated) result back to the job-submitter.

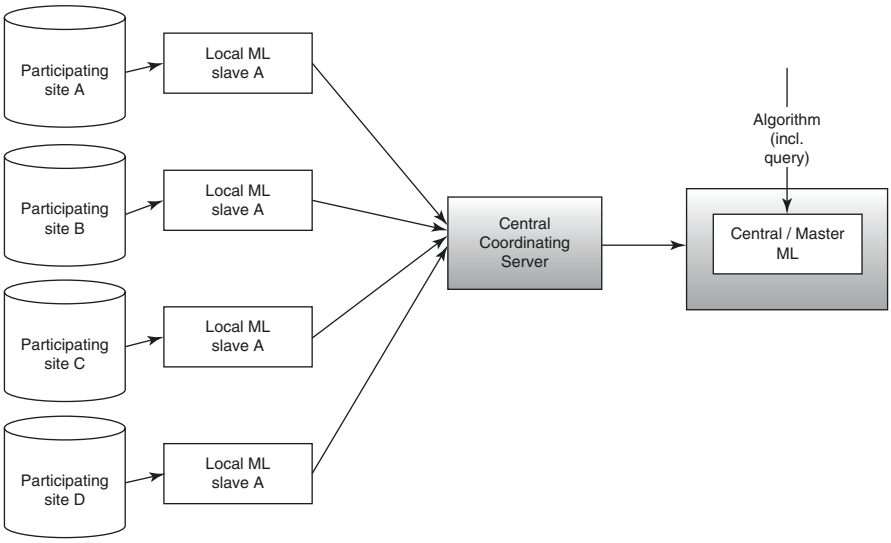


Fig. 8.12 Distributed Machine Learning Infrastructure

Since the researcher sending the algorithm is unable to know the exact data schema at the source, PHT depends on the FAIR data principles and FAIR data descriptions available at each centre. PHT leverages containerization technologies for packaging and sending algorithms to the distributed sources [20]. This ensures that the algorithms are executed in system isolated and platform independent manner, thereby reducing the task of IT maintenance at the hospital side.

The different privacy concerns and how they are handled in a distributed manner are discussed below:

Data Privacy: Based on how individual patient records are distributed among different centres, data privacy mechanisms may vary (Sect. 8.1.1). For horizontally partitioned data, PHT ensures data privacy by keeping the data at the source and only fetching analysis from the data [21–23]. For vertically partitioned data, cryptographic methods such as homomorphic encryption, secure multiparty computations, or differential privacy mechanisms are adopted to ensure privacy [24–26].

Model Privacy: Attackers and malicious users can reverse engineer the trained machine learning model to regenerate patterns of the original patient data [27, 28]. Using differential privacy techniques and cryptographic mechanisms for storing and communicating the model parameters, model privacy can be achieved.

System and User Privacy: Adoption of proper authentication and authorization techniques secure malicious attackers from accessing the infrastructure. The researcher, data provider, and the server applications are granted access to the infrastructure only after proper authentication. All communication between the parties are encrypted using suitable cryptographic mechanisms.

8.2.4 Centralized and Distributed Machine Learning Algorithms

When the prerequisites regarding semantic interoperability, data structure, infrastructure, and privacy preservation are in place, we can start performing machine learning. In this section, we merely touch upon centralized machine learning in favour of describing distributed machine learning approaches in full, which are considered superior for future, large-scope implementations.

8.2.4.1 Centralized Machine Learning

As described above, the centralized approach only needs one machine learning unit (Fig. 8.11). In this case, the machine learning system will query and retrieve data from the federation data store, irrespectively of knowing where the actual data comes from (except when provenance variables are included in this dataset). As the retrieved dataset is not different in comparison to traditional machine learning approaches, we can use standard machine learning toolboxes such as Weka [29], RapidMiner [30], or others [31]. The disadvantage is that data, with/without privacy preservation in place, is transferred to a central location at time of machine learning algorithm execution. This might contradict the policy of centres regarding data sharing.

8.2.4.2 Distributed Machine Learning

The major difference between distributed and centralized machine learning is the transfer of data versus the transfer of training models. In the centralized approach, data is transferred to the machine learning system, whereas in the distributed approach the data stays within the institute. Rather than requesting a dataset, the distributed approach dispatches a sub-process of the machine learning algorithm towards the institutional machine learning unit and returns the result of this sub-process. In this setting, the amount of data per transfer diminishes; however, the data transfer frequency increases.

Federated machine learning is defined as process where N data owners ($\mathcal{F}_1, \mathcal{F}_2, ...\mathcal{F}_N$) collaboratively train a machine learning model $\mathcal{M}_{\text{FED}}$, in which any data owner does not expose its data $\mathcal{D}_i$ to others [32]. Distributed learning or federated machine learning algorithm ideally should follow the four points as discussed below:

- The algorithm should be mathematically split in a manner that the global approximation over distributed datasets is comparable to a centralized model trained on a single data source containing all the training data.
- The training process, in addition to the infrastructure should ensure that no data will ever leave the source.
- Re-engineering of the original training data from the trained model should not be possible and as such proper encryption mechanisms should be adopted.

As described in Sect. 8.1.1.2, patient data is partitioned horizontally or vertically among different centres. Based on how data is distributed across the centres, machine learning algorithms come in two different flavours: horizontal distributed (federated) algorithms and vertical distributed (federated) algorithms. The following section describes the two categories briefly. We also introduce a third category, popularly known as Federated Transfer Learning.

Horizontal Distributed (Federated) Machine Learning

Horizontally partitioned data means that each centre contains exactly the same set of patient variables to train the algorithm. Patient data privacy in this scenario is ensured by keeping data at the source at every point of time during the training process. This type of learning is also known as sample-partitioned federated learning and assumes honest participants and security against an honest-but-curious (semi-honest) server. This means that only the server can compromise the user privacy and data security of the participants [33, 34]. The training process of a horizontal federated machine learning with PHT or similar client server-based infrastructure usually consists of the following steps:

Step 1: The researcher initiates the training process by signalling the central server. The central server requests all sites to train a machine learning algorithm. In case of PHT, the machine learning algorithm can be wrapped in a Docker image and

stored in a private Docker registry. The private Docker registry is accessible only to the authorized users of the infrastructure.

Step 2: The data centres participating in the machine learning process train the algorithms locally, compute the training gradients or model weights, encrypt and send it to the server.

Step 3: The server invokes the central averaging algorithm and performs a secure aggregation, e.g. via weighted averaging.

Step 4: The server sends the aggregated results to the participating centres.

Step 5: The centres update their respective models with the received gradients.

The iterations through the above steps continue until, the loss function converges, or the maximum allowed iterations are reached. Both non-linear and linear federated machine learning models are available in the literature. Table 8.2 lists some of the available linear and non-linear algorithms for horizontally portioned data:

A thorough explanation how distributed machine learning algorithms work is given by Boyd et al. [40] and Wu et al. [38]. Recent years has shown many interesting studies involving distributed machine learning applied to radiotherapy or oncology in general. Deist et al. [22] showed how machine learning with PHT can be scaled up to include a large distributed cohort (20,000+ patients). They trained a distributed logistic regression and optimized via the ADMM to predict 2-year post-treatment survival in lung cancer patients. Deist et al. [23] earlier presented a similar study by training SVM via ADMM to predict post-radiotherapy dyspnoea in lung cancer patients. Shi et al. [21] used PHT to train a distributed cox regression algorithm to develop a radiomics signature for 2-year survival in lung cancer patients. Bayesian network models were studied by Jochems et al. [42, 43]. Traditional machine learning with horizontal data partitioning has been widely explored and many algorithms and applications developed. Recent advancement in computing power has opened the gateway to neural networks and deep learning both in a centralized and distributed setup. Balachandran et al. [44] proposed a distributed deep learning algorithm to learn from medical images. The model has been trained by Cyclic Weight Transfer method, where one centre trains the model for a number

Table 8.2 List of machine learning models for horizontal federated learning

Algorithms	Linear/ non-linear	Privacy mechanism	Optimizer/remarks
Linear Regression [35, 36]	Linear		
Logistic Regression [22, 37]	Linear	PHT, by design	Grid binary logistic regression (GLORE) [38]
Cox Regression [21, 39]	Linear	By infrastructure design	
Support Vector Machines [23]	Non-linear	PHT, by design	Alternating direction method of multipliers (ADMM) [40]
FedAVG [41]	Non-linear, neural network		Stochastic gradient descent (SGD)

iterations and transfers it to the next centre. All centres train the model in a cyclic fashion until convergence is achieved. McMahan et al. [41] proposed a different approach, where weights from all centres are aggregated by an aggregating algorithm (FedAvg). While this method is a communication efficient way for training SGD based or DNN, it exposes the gradients or model weights and can be a privacy threat for the patient data. FedAvg with homomorphic encryption (AHE) or learning with errors (LWE) can be used to enhance the privacy and security features of the algorithm.

Vertical Distributed (Federated) Learning

Vertical distributed (federated) learning is conducted when each data centre hosts different sets of features of the same group of patients. Before training the algorithms, records of patients have to be matched and linked among data centres. The most common way to link multiple vertically partitioned data is using unique identifiers (e.g. social security number). The unique identifiers should be pseudonymized every time when vertical distributed (federated) learning starts at each data centre in order to prevent individuals from being re-identified. Unfortunately, very few existing works studied record linkage problem in vertically partitioned data scenario in the real-life use cases. As in the horizontal distributed (federated) learning, data centres in vertical distributed setting are also assumed to be honest-but-curious (semi-honest) to each other's data.

Table 8.3 lists some recent privacy-preserving machine learning models and infrastructure learning from vertically partitioned data. Sun et al. [45] deployed Personal Health Train architecture to learn vertically partitioned data of people with Type 2 Diabetes Mellitus using linear regression model. A trusted third party is required by their method to link multiple datasets and execute data analysis in a secure way. This is the only one study in the list which describes record linkage in vertical federated learning. Another linear regression model for vertically partitioned data was proposed by [46]. Their approach combined multiparty computation techniques (Yao's garbled circuits protocol) and conjugate gradient descent

Table 8.3 List of machine learning models for vertical federated learning

Algorithms	Linear/ non-linear	Privacy mechanism	Optimizer/remarks
Linear Regression [45]	Linear	PHT, by design	With trusted third party
Linear Regression [46]	Linear	Conjugate gradient descent	For high-dimensional data
Logistic Regression [25]	Linear	Global gram matrix	Fixed-Hessian Newton method to improve scalability
Generalized Linear Model [47]	Linear	Distributed block coordinate descent	Can be extended for a hybrid partitioning situation
Support Vector Machines [48, 49]	Non-linear		Specially designed for clinical/medical use
Regression and neural networks model [50]	Non-linear	By infrastructure design	Stochastic gradient descent

algorithm to solve high-dimensional data problem. The efficiency and scalability are the two major advantages of this study. Li et al. [25] proposed a new approach to solve binary logistic regression problem by using the kernel trick to obtain the global gram matrix. They also applied Fixed-Hessian Newton method to solve the time-complexity problem for large-size datasets. Van Kesteren et al. [47] proposed a method using distributed block coordinate descent which is applicable for all generalized linear models. This method is not only suitable for vertically partitioned data but also for the hybrid scenario of horizontal and vertical federated learning setting. In addition to linear models, Rahulamathavan et al. [48] and Zhu et al. [49] developed privacy-preserving SVM algorithms particularly for clinical and medical purposes. Rahulamathavan et al. [48] studied SVM Gaussian kernel-based classification for a clinical decision support system, while Zhu et al. [49] developed an online medical pre-diagnosis framework using non-linear kernel SVM algorithm. Both studies applied cryptographic techniques to encrypt patients' data. Mohassel and Zhang [50] proposed linear regression, logistic regression and neural networks models combining stochastic gradient descent methods and secure multiparty computation techniques. The method is outstanding in efficiency and scalability but limited by the number of participating parties (only 2 parties).

8.2.5 Bioethics and Data Protection

Bioethical issues have been taken into consideration in the whole data lifecycle including data collection, data processing, data querying, data publishing, and data sharing. Individuals, data entities, and society involve in different part of data lifecycle (Table 8.2). When the individual data are being collected by the data entities, data entities must explicitly state what data they are collecting, how long and where they will keep the data, for what purpose, the use of the data, if any other entities request access to the data and other information generated by the data, what is restricted by the data protection regulations in the area of data entities. As individuals, we should be aware of what data are being collected from us under which conditions. Individuals are expected to provide reliable data which will be used for health research. It requires mutual efforts to ethically gather individual health data for scientific health studies. As the collected data are maintained and governed by data entities, the ethical issues in the data processing and querying have to be addressed by them. There are several existing methods of anonymization or pseudonymization to de-identify individuals which will be discussed below. Moreover, the actual data can be also generalized or modified to protect individual's privacy after de-identification. In the data querying, data entities must ensure the query result is aggregated so that no individual data will be exposed. With a scientific purpose, data publishing is normally requested to create values and benefits for the population or the whole society. The ethical concerns need to be taken care of by the data entities and our society. It requires mutual trust between the data entity and the public. As we know, there is a trade-off between data privacy and data utility. When the data is strictly protected, its utility for research or clinical use will drop significantly due to

Table 8.4 Data protection involvement in data lifecycle

	Data collection	Data processing	Data querying	Data publishing	Data sharing
Individuals	✓				✓
Data entities	✓	✓	✓	✓	✓
Society				✓	✓

information loss caused by privacy-preserving methods. This balance needs to be made by all parties involved with trust and the urgency of the problem. The most complex situation happens in the data sharing stage which normally involves all parties—individuals, data entities, and society (Table 8.4). This situation is illustrated in the following subsections.

8.2.5.1 Bioethics and Data Protection: Individuals

Data sharing involves two or normally more than two data entities to collaboratively solve some data problems. For using both distributed and centralized multicentre infrastructures for data sharing, preserving individuals' privacy is a major topic to consider. Data can be shared only when individuals get informed and grant the permission to do so. In the most common case, individuals could give specific informed consent stating the time of data collection, the period of data usage, which data for what purpose, and how data will be processed. For a broader use, broad informed consent or dynamic informed consent can be applied under some conditions [51]. No matter which type of informed consents is given, individuals should be able to easily withdraw the informed consents at any time to restrict the use of their data.

If the distributed multicentre infrastructure is correctly implemented, it respects individual's privacy more than the centralized setting as the results of the algorithm (e.g. a predictive model) are transferred instead of the original data. Individuals have full control of their own data so that they can withdraw or re-grant their permission anytime even during data sharing and analysis process as the data never leave or be copied from the data entities. This is rather more difficult to do in a centralized manner. In addition, since the algorithms are transferred to the data, data entities and individuals could check and be informed what and how the data are being processed. The distributed multicentre infrastructure could outperform a centralized setting regarding respecting individuals' privacy.

8.2.5.2 Bioethics and Data Protection: Data Entity

Although the distributed infrastructure is a better option considering preserving individuals' privacy, this does not mean that the issues concerning privacy preservation are solved. For example, it is still possible to retrieve metadata about a dataset of a small group patient who is easily being re-identified. In this section, we will address several options for privacy preservation, ranging from pseudonymization to irreversibly modifying the original datasets. Despite all the options described below, we must state that, in our opinion, there is no standard method to ensure privacy

preservation. The researcher/designer of the infrastructure will always have to find a balance between the loss of information (data utility) and the anonymity of participating patients (data privacy).

Pseudonymization

The first option for privacy preservation is bidirectional pseudonymization of patient identifiers, for example, replacing patient names and hospital's patient identification numbers by study-specific alternatives. This can be achieved by maintaining a two-column table, where one column contains the patients' identification number and the second column contains the study identification number for this patient. Variations to this concept may apply, for example, using an extra column to maintain the study where this mapping applies to. Typically, the pseudonymization of hospital to study identification numbers is done during the transform part of the ETL process. Other patient identifying information (e.g. first and last names) can be replaced by the same study ID or may not be incorporated and thus removed during the ETL process.

The second option is to use a unidirectional pseudonymization algorithm, for example, by hashing patient identifiers (e.g. using an SHA-{1–3} algorithm). This hash should be unidirectional, meaning that the pseudonymized patient identifiers cannot be reversed to the original identifiers. Unidirectional pseudonymization might be more appropriate than bidirectional pseudonymization, however might introduce problems when study data are needed to be linked to the actual patients. For example, when study results show a worse outcome for specific patients and when it is immoral to withhold this information to these patients.

Data Obfuscation

When using strict inclusion criteria with rare variables, it might be that the resulting dataset is very small and patients might become identifiable by combination. For example, if only two patients match some inclusion criteria and the biological sex (which is a requested variable) is different in both patients, we can identify these patients when querying local source systems. This issue holds for both the centralized and distributed infrastructures. To reduce the chance of compromising the anonymity of patients, Murphy and Chueh [52] introduced a method for data obfuscation where (especially in the case of a small number of events/patients) results are obfuscated by returning a random value within a specific range based on the actual value. This method does not circumvent the problem completely, as someone with bad intentions is able to approximate the original value by sending the same request multiple times. To circumvent these actions, Murphy and Chueh proposed to implement an audit system, where performing the same query multiple times within a specific time span will result in a request denial. In this way, the system returns a value not completely representing the actual value, however returns a value within a tolerable margin (when not exceeding the maximum number of requests).

Data Perturbation

The downside on obfuscation is that it does change the distance (e.g. Euclidian distance) between points (e.g. patients or observations) in a k-dimensional space,

where every dimension may be a specific variable in the dataset. As the distance changes, it may influence the prediction model training algorithm and train a model that does not represent the actual data and values. This can lead to problems during validation, especially when the validation data is obfuscated, however in another way (due to the randomness in the obfuscation algorithm). Therefore, transformation of data might be a solution, as the whole dataset is transformed while maintaining the distance between points. As shown by Liu et al. [53], this transformation is still not good enough for privacy preservation, as the original data can be derived using independent component analysis (ICA) or overcomplete ICA. To overcome this issue, Liu et al. advise to use their random projection-based multiplicative perturbation (RPBMP) method, which reduces the number of dimensions and transforms the dataset while maintaining statistical information regarding the distance between variables. Using this method, it should not be possible to retrieve the original values and would therefore obstruct the possibility to match variables to individual patients. This RPBMP method is afterwards reused by Yu et al. [54], where they explored differences in dimension reduction options and applied it to a non-small cell lung cancer (NSCLC) dataset. Data perturbation and dimension reduction are potential solutions to preserve privacy in a multicentre setting, although they could lead to issues when performing a risk analysis (identifying variables which influence a specific outcome). The risk analysis then can only determine which *compressed* dimensions are of influence; however, it cannot determine which biological (or source) variables/features are responsible for this influence in patient outcome.

8.2.5.3 Bioethics and Data Protection: Society

Comparing to individuals and data entities, our society is playing a big role in gaining and maintaining the public trust and involvement for health data sharing and usage. Our society needs to find the balance between respecting individual's privacy and creating maximal societal benefits and values by mining shared data. To achieve this goal, data itself is expected to be Findable, Accessible, Interoperable, and Reusable (FAIR), while data models and algorithms are required to be Fair, Accurate, Confidential, and Transparent (FACT).

8.2.6 Applications and Initiatives

In the previous paragraphs, we defined the prerequisites and described how to perform distributed machine learning. In this paragraph, we will discuss several initiatives and applications of distributed machine learning. It is not mandatory that all applications use the complete set of prerequisites described previously in this chapter.

8.2.6.1 Datashield

Datashield is an initiative for upscaling biomedical data science research by taking analysis to the data instead of bringing data to the analysis. The

infrastructure enables researcher to conduct collaborative data analysis on multiple centres, without having to bring data out of the centres. Similar to PHT, Datashield protects patient privacy by keeping data at the source. The infrastructure comprises of three components: a computer server at each source study hosting an Opal database, the statistical programming environment (R), and Datashield specific R libraries installed on the data servers and client computers (federated analysis point) [55]. KETOS is a clinical decision support and a machine learning service based on Datashield. It leverages containerization technology, Docker and interfaces with web services using the HL7 FHIR standard to access patient data stored in OMOP (OMOP common data model—OHDSI) database [56]. Successful studies regarding pregnancy outcomes, epidemiological data management, ageing and mental well-being, cardiovascular risk estimation, and metabolic syndrome prediction showed that federated or distributed analysis is indeed a viable solution where data sharing directly is not possible [57–61]. However, unlike PHT, Datashield constraints researcher by mandating use of dedicated software packages and databases. PHT is more flexible with the choice of algorithm implementation and also with the data representation at the source. This enhances the diversity of patient cohorts by including more number of centres and reducing IT administration burden on the hospital side.

8.2.6.2 I2B2

The Informatics for Integrating Biology and the Bedside (I2B2; http://www.i2b2.org) project aims at integrating data from different biomedical disciplines and delivering this data to researchers. The project delivers tools to translate genomic and biologic findings to clinical findings (e.g. diseases or disorders). To be able to achieve this *translational medicine* approach, institutional data sources are federated in the I2B2 DWH using ETL tooling. The DWH database structure, called the Clinical Research Chart (CRC), is generic for medical purposes, as it does not define specific data fields. The database structure is basically a "star schema" where only patient information and observations are stored [62]. To describe all information in an observation-centred storage, local terminologies, or standardized terminological systems, are needed to define different types of observations. Afterwards, researchers can query/request data. When a specific dataset has been queried, this dataset can be stored in a separate database, using the same CRC database structure. In this separate database, researchers can clean/modify the dataset to their needs and execute machine learning algorithms on this dataset.In regard to multicentre machine learning, I2B2 supports merging multiple research databases using the Shared Health Research Information Network (SHRINE) tool [63], resulting in a federated research database of multiple institute research databases. Therefore, it enables the opportunity for centralized multicentre learning. In this approach, the terminology to define observations can be aligned when merging databases or can be kept separate [64]. In the latter approach, the researcher has to put in more effort in data alignment during the analysis, which is not favourable as it is prone to causing mistakes in the analysis.

8.2.6.3 VATE

The VATE ("VAlidation of High TEchnology based on large database analysis by learning machine") project shares the aim of the EuroCAT project. The major difference is that this project is based on open standards (regarding IT infrastructure) and uses Semantic Web technologies (e.g. RDF and ontologies) as a basis for data representation. Prior to this project, the involved institutes had developed a data infrastructure for research purposes using open standards [65]. Equal to the EuroCAT project, the VATE project has developed an umbrella protocol for rectal cancer [66]. Different from the EuroCAT project, the variables to record are classified into several levels regarding the completeness of datasets and are maintained in a publicly available ontology (http://www.github.com/RadiationOncologyOntology/ROO). The rationale behind these rankings and this public umbrella protocol is that everyone who has data regarding rectal cancer patients can join this linked data network when the data is specified according to the ontological rules, irrespective to the number of available variables. Due to the chosen aim of training a Bayesian Network for rectal cancer on the VATE infrastructure, missing data could be imputed or ignored during training, as shown by Jayasurya et al. [67].

8.2.6.4 PCORnet

The Patient-Centered Outcomes Research Network (PCORnet) is a programme aiming at building a national research network linking datasets from clinical production systems from multiple centres, using a standardized data platform [68]. The programme comprises 11 clinical data research networks (CDRN) and 18 patient-powered research networks (PPRN). The aim of the CDRNs is comparable to the previously described EuroCAT and VATE projects. The PPRN projects aim at the empowerment of patients. In these PPRNs, patients would supply the data instead of retrieving data from clinical systems. Therefore, the gathered data and research questions addressed by these projects are different from the CDRN projects [69]. The first (short-term) aims for the programme are to build and implement the network in all the CDRNs and PPRNs and include one million patients in 18 months after the start of the project. Long-term aims are to perform (distributed) machine learning on the network.

8.2.6.5 FAIRHealth

The FAIRHealth project is one of the projects under the programme of Value Creation through Responsible Access and Use of Big Data (VWData, https://commit2data.nl/vwdata) funded by the Dutch National Research Agenda. The goal of this project is to study annual healthcare costs in relation to the incidence of Type 2 Diabetes Mellitus (T2D) without revealing any original data [24, 45]. We used 3283 patients' health data from De Maastricht Studie (www.demaastrichtstudie.nl), which is characterized by extensive phenotyping and provides information on the aetiology, pathophysiology, complications, and comorbidities of T2DM. All participants are aged between 40 and 75 years and live in the southern part of The Netherlands. We requested those attributes which were complete and consented. We linked these data to their health insurance reimbursement data from Statistics

Netherlands (www.cbs.nl). We extended Personal Health Train architecture and built up a FAIR data station at each site. The original data files (SAV, CSV) were automatically transformed to RDF by defining SPARQL construct queries, and subsequently loaded into a triple store and made available as per the FAIR data station specification. To execute the analysis on the combined datasets, we used the Trust Secure Environment where also has a FAIR station. Briefly, in this infrastructure, data access is regulated by the data provider hosting the stations. If access is granted, the data providers encrypt the data and send these to the TSE. The TSE executes the researchers' application and allows aggregated results to be returned to the researcher.

8.2.6.6 Personal Health Train Initiatives

Improving healthcare with PHT has gained popularity within the healthcare research community. There have been many initiatives within the European Region and also other parts of the world (https://pht.health-ri.nl/). Many different research organizations and hospitals joined hands to improve cancer by enhancing data sharing using PHT. AMICUS aims to improve cancer screening, detection and predict treatment outcome by utilizing distributed deep learning on medical images. My Best Treatment is a project which aims to make the best decision support system for the treatment of patients with terminal lung cancer. PROTRAIT aims to set up an infrastructure for automatic data registration related to novel proton therapy in the Netherlands. CONVINCED aims to enable survival analysis on vertically partitioned data while securing privacy using multiparty computation (Sect. 8.2.4.2.2). The aim of the RARECAREnet Asia is to study patterns of incidence and survival of rare cancers in Europe and Asia. The project "Understanding Oral Cavity cancer survival in the Netherlands and Taiwan" aims to perform federated analysis of survival of patients with oral cavity cancer based on cancer registry data. One of the notable projects, the EuroCAT umbrella project and the 20K Challenge are described below:

EuroCAT

The Euregional Computer-Aided Theragnostics (EuroCAT; http://www.eurocat.info) project aims at reuse of clinical data for research purposes and to improve the speed and quality of clinical research. The project uses a distributed learning approach as described before, targeted at prediction models for lung cancer. To be able to perform this distributed learning approach, a so-called umbrella protocol was developed by the participating partners. This protocol describes the standardized data collection, including the variables to record (and terminological systems to use), questionnaires, and informed consent document templates. The first version of the EuroCAT system used a DWH and ETL infrastructure at the local institutes. Afterwards, the DWH was replaced by an RDF store. The EuroCAT system has shown that distributed multicentre machine learning works and produces the same results as centralized learning when implemented correctly [70]. Furthermore, the project has shown that distributed multicentre learning does improve the robustness of prediction models when validating on an external dataset [71].

Similarly other projects under the same umbrella like The "Dutch Network of Computer Assisted Theragnostics (duCAT)", "Rapid Learning Infrastructure for outcome prediction models in rectal cancer (chinaCAT)", and "A survival prediction Model for NSCLC Patients Through Distributed Learning Across 3 countries (meerCAT)" all aim for improving cancer care outcome by leveraging distributed machine learning technologies.

20 K Challenge
The project aimed to show that PHT can be scaled up to many thousands of patients from several hospitals. FAIR data of eight healthcare data providers from five different countries are connected using PHT. A total of 23,203 patient cases across 8 centres (Amsterdam, Cardiff, Maastricht, Manchester, Nijmegen, Rome, Rotterdam, and Shanghai) were connected. A logistic regression model predicting post-treatment 2-year survival was trained on 14,810 patients (clinical information only) treated between 1978 and 2011 and validated on 8393 patients treated between 2012 and 2015. The project successfully showed that PHT can be scaled up to many thousand patients from multiple centres [22].

8.2.7 Summary

In this chapter, we have seen that machine learning with distributed dataset is possible for both a centralized and federated approach. To be able to set up a distributed machine learning environment, several biomedical informatics-related issues need to be addressed. The most important issue is semantic interoperability among participating centres. If the participating centres cannot agree on definitions, how do we know whether all data are equally formatted? Second, the infrastructure (both institutional and central) needs to be implemented, together with the chosen data representation. The choice for an infrastructure comes with the choice of a centralized or distributed approach. Third, privacy preservation needs to be addressed and may influence the choice for a centralized or distributed approach and the preservation measures implemented (e.g. uni- versus bidirectional pseudonymization or data perturbation versus transformation). When all prerequisites are met, the actual machine learning can be performed. In this part, a centralized approach should not be different from traditional machine learning. The distributed machine learning approach needs some modifications to traditional machine learning algorithms, as local outcomes need to be aggregated and combined at a central location. Therefore, in distributed machine learning, traditional algorithms need to be split into two parts: a central node performing the general algorithm and institutional nodes performing delegated tasks requested by the central node. Finally, we have shown that distributed machine learning is possible in practice. Showing several projects and/or initiatives where data from different locations are used to develop prediction models.

In general, we have shown that distributed machine learning is not only a task for the "traditional" machine learning expert (which is already not the case in

healthcare and radiation oncology); however, it also needs other disciplines, such as expertise from the fields of terminology/ontology development, network/infrastructure, and security/privacy.

References

1. Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, Roelofs E, van Elmpt W, Boutros PC, Granone P, Valentini V, Begg AC, De Ruysscher D, Dekker A. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol. 2013;10:27–40. https://doi.org/10.1038/nrclinonc.2012.196.
2. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, Bach PB, Murphy SB. Rapid-learning system for cancer care. J Clin Oncol. 2010;28:4268–74.
3. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://www.nature.com/articles/sdata201618. Last accessed 14 Jan 2019.
4. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol. 2013;108:174–9. https://doi.org/10.1016/j.radonc.2012.09.019.
5. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, Zegers CML, Gillies R, Boellard R, Dekker A, Aerts HJWL. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–6. https://doi.org/10.1016/j.ejca.2011.11.036.
6. Leijenaar RTH, Carvalho S, Velazquez ER, van Elmpt WJC, Parmar C, Hoekstra OS, Hoekstra CJ, Boellaard R, Dekker ALAJ, Gillies RJ, Aerts HJWL, Lambin P. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013;52:1391–7. https://doi.org/10.3109/0284186X.2013.812798.
7. Juty N, Wimalaratne SM, Soiland-Reyes S, Kunze J, Goble CA, Clark T. Unique, persistent, resolvable: identifiers as the foundation of FAIR. Data Intell. 2019;2:30–9. https://doi.org/10.1162/dint_a_00025.
8. Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, Karim MdR, Dumontier M, Decker S, da Silva Santos LOB, Dekker A. Distributed analytics on sensitive medical data: the personal health train. Data Intell. 2019;96–107. https://doi.org/10.1162/dint_a_00032.
9. Multidisciplinary management of rectal cancer - questions and answers. Vincenzo Valentini. Springer. https://www.springer.com/gp/book/9783319432151. Last accessed 9 Apr 2020.
10. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: terminology and typology. Methods Inf Med. 2000;39:16–21.
11. WHO. International Classification of Diseases, 11th revision (ICD-11). http://www.who.int/classifications/icd/en/. Last accessed 9 Apr 2020.
12. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40:30–43. https://doi.org/10.1016/j.jbi.2006.02.013.
13. SNOMED Home page. Last accessed 22 Oct 2018.
14. Gali A, Chen CX, Claypool KT, Uceda-Sosa R. From ontology to relational databases. In: Conceptual modeling for advanced application domains. Berlin: Springer; 2004. p. 278–89. https://doi.org/10.1007/978-3-540-30466-1_26.
15. Allemang D, Hendler J. Semantic web for the working ontologist: effective modeling in RDFS and OWL. Amsterdam: Morgan Kaufmann; 2008.
16. Berners-Lee T, Hendler J, Lassila O. The semantic web. Sci Am. 2001;284:28–37.
17. Brickley D, Guha RV. RDF schema 1.1. W3C Recomm; 2014.
18. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. Int J Semantic Web Inf Syst. 2009;5:1–22.

19. Prud'Hommeaux E, Seaborne A. SPARQL query language for RDF. W3C Recomm. 2008;15.
20. What is a container. https://www.docker.com/resources/what-container. Last accessed 22 Oct 2018.
21. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Scientific Data. https://www.nature.com/articles/s41597-019-0241-0. Last accessed 9 Mar 2020.
22. Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, Masciocchi C, Valentini V, Wang J, Chen J, Zhang Z, Spezi E, Button M, Jan Nuyttens J, Vernhout R, van Soest J, Jochems A, Monshouwer R, Bussink J, Price G, Lambin P, Dekker A. Distributed learning on 20 000+ lung cancer patients – the personal health train. Radiother Oncol. 2020;144:189–200. https://doi.org/10.1016/j.radonc.2019.11.019.
23. Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, Eble M, Bulens P, Coucke P, Dries W, Dekker A, Lambin P. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol. 2017;4:24–31. https://doi.org/10.1016/j.ctro.2016.12.004.
24. van Soest J, Sun C, Mussmann O, Puts M, van den Berg B, Malic A, van Oppen C, Towend D, Dekker A, Dumontier M. Using the personal health train for automated and privacy-preserving analytics on vertically partitioned data. Stud Health Technol Inform. 2018;247:581–5.
25. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc. 2016;23:570–9. https://doi.org/10.1093/jamia/ocv146.
26. Li Q, Wen Z, Wu Z, Hu S, Wang N, He B. A survey on federated learning systems: vision, hype and reality for data privacy and protection. ArXiv190709693 Cs Stat.; 2020.
27. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. ArXiv161005820 Cs Stat.; 2017.
28. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: stand-alone and federated learning under passive and active white-box inference attacks. ArXiv181200910 Cs Stat.; 2018.
29. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl. 2009;11:10. https://doi.org/10.1145/1656274.1656278.
30. Hofmann M, Klinkenberg R, editors. RapidMiner: data mining use cases and business analytics applications. Boca Raton: CRC Press; 2013.
31. Ramamohan Y, Vasantharao K, Chakravarti CK, Ratnam ASK. A study of data mining tools in knowledge discovery process. Int J Soft Comput Eng. 2012;2:4.
32. Yang Q, Liu Y, Chen T, Tong,Y. Federated machine learning: concept and applications; 2019. https://doi.org/10.1145/3298981.
33. Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. https://dl.acm.org/doi/10.1145/3133956.3133982. Last accessed 15 Apr 2020.
34. Phong LT, Aono Y, Hayashi T, Wang L, Moriai S. Privacy-preserving deep learning via additively homomorphic encryption; 2017.
35. Dobriban E, Sheng Y. Distributed linear regression by averaging. ArXiv181000412 Math Stat.; 2019.
36. Yuan D, Proutiere A, Shi G. Distributed online linear regression. ArXiv190204774 Cs Math Stat.; 2019.
37. Bogowicz M, Jochems A, Deist TM, Tanadini-Lang S, Huang SH, Chan B, Waldron JN, Bratman S, O'Sullivan B, Riesterer O, Studer G, Unkelbach J, Barakat S, Brakenhoff RH, Nauta I, Gazzani SE, Calareso G, Scheckenbach K, Hoebers F, Wesseling FWR, Keek S, Sanduleanu S, Leijenaar RTH, Vergeer MR, Leemans CR, Terhaard CHJ, van den Brekel MWM, Hamming-Vrieze O, van der Heijden MA, Elhalawani HM, Fuller CD, Guckenberger M, Lambin P. Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer. Sci Rep. 2020;10:1–10. https://doi.org/10.1038/s41598-020-61297-4.

38. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc. 2012;19:758–64. https://doi.org/10.1136/amiajnl-2012-000862.
39. Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, Ohno-Machado L. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc. 2015;22:1212–9. https://doi.org/10.1093/jamia/ocv083.
40. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning. https://dl.acm.org/doi/10.1561/2200000016. Last accessed 15 Apr 2020.
41. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA. y: communication-efficient learning of deep networks from decentralized data. ArXiv160205629 Cs; 2017.
42. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital – a real life proof of concept. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0167814016343365. Last accessed 9 Mar 2020.
43. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, Jolly S, Matuszak M, Ten Haken R, van Soest J, Oberije C, Faivre-Finn C, Price G, de Ruysscher D, Lambin P, Dekker A. Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. Int J Radiat Oncol. 2017;99:344–52. https://doi.org/10.1016/j.ijrobp.2017.04.021.
44. Balachandar N, Chang K, Kalpathy-Cramer J, Rubin DL. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. J Am Med Inform Assoc. 2020;27(5):700–8. https://doi.org/10.1093/jamia/ocaa017.
45. Sun C, Ippel L, van Soest J, Wouters B, Malic A, Adekunle O, van den Berg B, Mussmann O, Koster A, van der Kallen C, van Oppen C, Townend D, Dekker A, Dumontier M. A privacy-preserving infrastructure for analyzing personal health data in a vertically partitioned scenario. Stud Health Technol Inform. 2019;264:373–7. https://doi.org/10.3233/SHTI190246.
46. Gascón A, Schoppmann P, Balle B, Raykova M, Doerner J, Zahur S, Evans D. Privacy-preserving distributed linear regression on high-dimensional data. Proc Priv Enhancing Technol. 2017;2017:345–64. https://doi.org/10.1515/popets-2017-0053.
47. van Kesteren E-J, Sun C, Oberski DL, Dumontier M, Ippel L. Privacy-preserving generalized linear models using distributed block coordinate descent. ArXiv191103183 Cs Stat; 2019.
48. Rahulamathavan Y, Veluru S, Phan RC-W, Chambers JA, Rajarajan M. Privacy-preserving clinical decision support system using Gaussian kernel-based classification. IEEE J Biomed Health Inform. 2014;18:56–66. https://doi.org/10.1109/JBHI.2013.2274899.
49. Zhu H, Liu X, Lu R, Li H. Efficient and privacy-preserving online medical prediagnosis framework using nonlinear SVM. IEEE J Biomed Health Inform. 2017;21:838–50. https://doi.org/10.1109/JBHI.2016.2548248.
50. Mohassel P, Zhang Y. SecureML: a system for scalable privacy-preserving machine learning. In: 2017 IEEE symposium on security and privacy (SP); 2017. p. 19–38. https://doi.org/10.1109/SP.2017.12.
51. Biological and health data. https://www.nuffieldbioethics.org/publications/biological-and-health-data. Last accessed 15 Apr 2020.
52. Murphy SN, Chueh HC. A security architecture for query tools used to access large biomedical databases. In: Proceedings of AMIA symposium; 2002. p. 552–56.
53. Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans Knowl Data Eng. 2006;18(1):92–106.
54. Yu S, Fung G, Rosales R, Krishnan S, Rao RB, Dehing-Oberije C, Lambin P. Privacy-preserving cox regression for survival analysis. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008;1034–42. https://doi.org/10.1145/1401890.1402013.
55. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J, Boyd AW, Newby CJ, Nuotio M-L, Wilson R, Butters O, Murtagh B, Demir I, Doiron D, Giepmans L, Wallace SE, Budin-Ljøsne I, Oliver Schmidt C, Boffetta P, Boniol M, Bota M, Carter KW, deKlerk N, Dibben C, Francis RW, Hiekkalinna T, Hveem K, Kvaløy K, Millar S, Perry IJ, Peters A,

Phillips CM, Popham F, Raab G, Reischl E, Sheehan N, Waldenberger M, Perola M, van den Heuvel E, Macleod J, Knoppers BM, Stolk RP, Fortier I, Harris JR, Woffenbuttel BH, Murtagh MJ, Ferretti V, Burton PR. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014;43:1929–1944. https://doi.org/10.1093/ije/dyu188.

56. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, Kapsner LA, Zierk J, Mate S, Stürzl M, Croner R, Prokosch H-U, Toddenroth D. KETOS: clinical decision support and machine learning as a service – a training and deployment platform based on Docker, OMOP-CDM, and FHIR web services. PLoS One. 2019;14:e0223010. https://doi.org/10.1371/journal.pone.0223010.
57. Associations between maternal physical activity in early and late pregnancy and offspring birth size: remote federated individual level meta-analysis from eight cohort studies – Pastorino. BJOG. 2019. https://obgyn.onlinelibrary.wiley.com/doi/full/10.1111/1471-0528.15476. Last accessed 15 Apr 2020.
58. MINDMAP: establishing an integrated database infrastructure for research in ageing, mental well-being, and the urban environment. BMC Public Health | Full Text. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-018-5031-7. Last accessed 15 Apr 2020.
59. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. International Journal of Epidemiology. Oxford Academic. https://academic.oup.com/ije/article/46/5/1372/4102813. Last accessed 15 Apr 2020.
60. Long-term exposure to road traffic noise, ambient air pollution, and cardiovascular risk factors in the HUNT and lifelines cohorts. European Heart Journal. Oxford Academic. https://academic.oup.com/eurheartj/article/38/29/2290/3858093. Last accessed 15 Apr 2020.
61. van Vliet-Ostaptchouk JV, Nuotio M-L, Slagter SN, Doiron D, Fischer K, Foco L, Gaye A, Gögele M, Heier M, Hiekkalinna T, Joensuu A, Newby C, Pang C, Partinen E, Reischl E, Schwienbacher C, Tammesoo M-L, Swertz MA, Burton P, Ferretti V, Fortier I, Giepmans L, Harris JR, Hillege HL, Holmen J, Jula A, Kootstra-Ros JE, Kvaløy K, Holmen TL, Männistö S, Metspalu A, Midthjell K, Murtagh MJ, Peters A, Pramstaller PP, Saaristo T, Salomaa V, Stolk RP, Uusitupa M, van der Harst P, van der Klauw MM, Waldenberger M, Perola M, Wolffenbuttel BH. The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. BMC Endocr Disord. 2014;14:9. https://doi.org/10.1186/1472-6823-14-9.
62. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, Gainer V, Berkowicz D, Glaser JP, Kohane I, Chueh HC. Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. AMIA Annu Symp Proc. 2007;2007:548–52.
63. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16:624–30. https://doi.org/10.1197/jamia.M3191.
64. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc JAMIA. 2010;17:124–30. https://doi.org/10.1136/jamia.2009.000893.
65. Roelofs E, Dekker A, Meldolesi E, van Stiphout RGPM, Valentini V, Lambin P. International data-sharing for radiotherapy research: an open-source based infrastructure for multicentric clinical data mining. Radiother Oncol. 2014;110:370–4. https://doi.org/10.1016/j.radonc.2013.11.001.
66. Meldolesi E, van Soest J, Dinapoli N, Dekker A, Damiani A, Gambacorta MA, Valentini V. An umbrella protocol for standardized data collection (SDC) in rectal cancer: a prospective uniform naming and procedure convention to support personalized medicine. Radiother Oncol. 2014;112:59–62.
67. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, De Neve W, Lievens Y, Lambin P, Dekker ALAJ. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy:

Bayesian network for survival prediction in lung cancer. Med Phys. 2010;37:1401–7. https://doi.org/10.1118/1.3352709.
68. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The Greater Plains Collaborative: a PCORnet Clinical Research Data Network. J Am Med Inform Assoc. 2014;21:637–41. https://doi.org/10.1136/amiajnl-2014-002756.
69. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014;21:578–82. https://doi.org/10.1136/amiajnl-2014-002747.
70. Wiessler W, Dekker A, Nalbantov G, Oberije C, Eble M, Dries W, Janvary L, Bulens P, Balaji K, Lambin P. Privacy-preserving, multi-centric machine learning across institutions and countries: does it work? Presented at the Geneva, April 2013.
71. Dekker A, Nalbantov G, Oberije C, Wiessler W, Eble M, Dries W, Janvary L, Bulens P, Krishnapuram B, Lambin P. Multi-centric learning with a federated IT infrastructure: application to 2-year lung-cancer survival prediction. In: 2nd ESTRO FORUM; 2013. p. S35. Geneva: Elsevier.

Part II

Machine Learning for Medical Image Analysis in Radiology and Oncology

9 Computerized Detection of Lesions in Diagnostic Images with Early Deep Learning Models

Kenji Suzuki

9.1 Introduction

Computer-aided detection (CADe) and diagnosis (CADx) [1–4] have been active research areas in medical imaging. CADe/CADx is defined as detection/diagnosis made by a physician/radiologist who takes into account the computer output as a "second opinion" [1]. CADe focuses on a detection task, namely localization of lesions in medical images. CADx focuses on a diagnosis (characterization) task, for example, distinction between benign and malignant lesions. Computer-aided diagnosis without distinction between CADe and CADx is abbreviated as CAD. As imaging technologies advance, a large number of medical images are produced which physicians/radiologists must read. They may overlook lesions from such a large number of medical images. Thus, CAD is becoming indispensable in physicians' decision-making. Evidence suggests that CAD can help improve the diagnostic performance of physicians/radiologists [5–11]. Consequently, many investigators have developed CAD schemes such as those for the detection of lung nodules in chest radiographs [12–14] and in thoracic CT [15–17], those for the detection of microcalcifications/masses in mammography [18], breast MRI [19] and breast ultrasound (US) [20], and those for the detection of polyps in CT colonography (CTC) [21–23].

Machine learning (ML) plays an essential role in CAD, because objects such as lesions and organs in medical images may be too complex to be represented accurately by a simple equation; modeling of such complex objects often requires a

K. Suzuki (✉)
Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan
e-mail: suzuki.k.di@m.titech.ac.jp

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_9

number of parameters that have to be determined by data [24–26]. For example, a lung nodule is generally modeled as a solid sphere, but there are spiculated nodules and ground-glass nodules [27]. Although a polyp in the colon is modeled as a bulbous object, there are polyps that exhibit a flat shape [28, 29]. Thus, diagnostic tasks in medical images essentially require "learning from examples (or data)" to determine a number of parameters in a complex model. Because of its importance and significance, the field of ML in medical imaging became very active [30–32]. The special issues on ML in medical imaging were published in various journals [33–37]; a series of international workshops on this topic were held from 2010 [38–41].

One of the most popular uses of ML in CAD is the classification of lesion candidates into certain classes (e.g., abnormal or normal, lesions or non-lesions, and malignant or benign) based on input features (e.g., area, contrast, and circularity) obtained from segmented candidates (this class of ML is referred to feature-based ML or segmented-object-based ML). The task of ML is to determine "optimal" boundaries for separating classes in the multidimensional feature space which is formed by the input features [42]. The ML algorithms for classification include linear discriminant analysis [43], quadratic discriminant analysis [43], multilayer perceptron [44, 45], and support vector machines [46, 47]. Such ML algorithms were applied to lung nodule detection in chest radiography [48–51] and thoracic CT [15, 52–54], detection of microcalcifications in mammography [55–58], detection of masses in mammography [59], polyp detection in CT colonography [60–62], determining subjective similarity measure of mammographic images [63–65], and detection of aneurysms in brain MRI [66].

Recently, an ML area called deep learning [67–69] emerged in the computer vision field and became very popular in virtually all fields. This research "boom" started from an event in late 2012. A deep learning approach based on a convolutional neural network (CNN) [70] won an overwhelming victory in a worldwide computer-vision competition, ImageNet Classification, with the error rate smaller by 11% than that in the second place of 26% [71]. Consequently, the MIT Technology Review named it one of the top 10 breakthrough technologies in 2013. Since then, researchers in virtually all fields, including medical imaging, have started actively participating in the explosively growing field of deep learning [67]. Details on deep learning algorithms can be found in Chap. 4.

More than a decade before this deep learning research boom started, an early deep learning model, called massive-training artificial neural networks (MTANNs), had been invented and developed in the field of medical imaging in 2002 [72]. The MTANNs were applied for the detection of lung nodules in chest CT in 2003 [73], end-to-end detection of lung nodules in 2009 [74], separation of bones from soft tissue in chest radiographs in 2004 [75, 76], and reduction of noise and artifacts on CT images in 2013 [77].

In this chapter, ML techniques and early deep learning models used in CADe and CADx schemes of the thorax and colon are described, including CADe schemes for lung nodules in chest radiography and thoracic CT, and those for the detection of polyps in CTC.

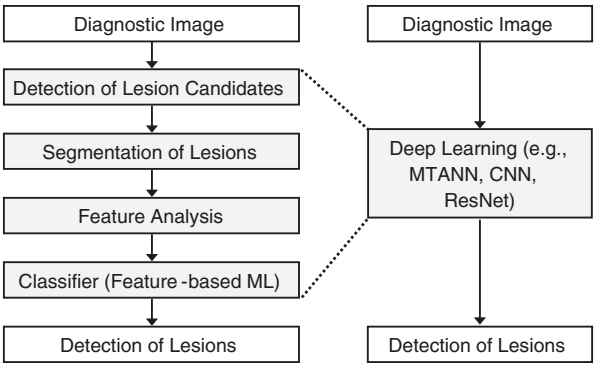


Fig. 9.1 Flowchart for a generic CAD scheme for detection of lesions in diagnostic images (left) and that for a deep-learning-based CAD scheme

9.2 Overview of Architecture of a CADe Scheme

A flowchart for a generic CADe scheme of lesions in diagnostic images is shown in Fig. 9.1. A CADe scheme generally consists of four core steps: (1) detection of lesion candidates, (2) segmentation of the detected lesion candidates, (3) feature analysis of the segmented lesion candidates, and (4) classification of the lesion candidates by use of a classifier with features (feature-based ML). The development of the detection of lesion candidates generally aims to obtain a high sensitivity level, because the sensitivity lost in this step cannot be recovered in the later steps. In the next step, the detected (or localized) lesion candidates are segmented, and connected-component labeling [78–86] is performed to identify each segmented candidate as an individual isolated object. Pattern features such as gray-level-based features, texture features, and morphologic features are extracted from the segmented candidates. Finally, the detected lesion candidates are classified into lesions or non-lesions by using a classifier (or feature-based ML). This final step is very important, because it determines the final performance of a CADe scheme when the additional step of FP reduction is not employed. The development of the classification step aims to remove as many non-lesions (i.e., FPs) as possible while minimizing the removal of lesions (i.e., true-positive detections).

After the development of a CADe scheme, the evaluation of the stand-alone performance of the developed scheme is the last step in CADe *development*. CADe *research* does not end by this step: the evaluation of radiologists' performance with the use of the developed CADe scheme is the important last step in CADe *research*.

9.3 Machine Learning (ML) in CADe

9.3.1 Feature-Based (Segmented-Object-Based) ML (Classifiers)

An ML technique is generally used in the step of classification of lesion candidates. The ML technique is trained with sets of input features and correct class labels. This

class of ML is referred to as feature-based ML, segmented-object-based ML, or simply as a classifier. The task of ML here is to determine "optimal" boundaries for separating classes in the multidimensional feature space which is formed by input features [42]. A standard classification approach is illustrated in Fig. 9.2a. First, lesions (lesion candidates) are segmented by use of a segmentation method. Next, features are extracted from the segmented lesions. Then, extracted features are entered as input to an ML model such as linear discriminant analysis [43], quadratic discriminant analysis [43], a multilayer perceptron (or artificial neural network) [44, 45], and a support-vector machine [46, 47]. When an artificial neural network is used as a classifier, the structure of the artificial neural network may be designed by use of an automated design method such as sensitivity analysis [87, 88]. The ML model is trained with sets of input features and correct class labels. Feature selection is often used to select "effective" features for a given task. One of the most recent, promising feature selection methods is feature selection under the criterion of the maximal area under the receiver-operating-characteristic curve [89]. For details of feature-based classifiers, refer to one of many textbooks in pattern recognition such as Bishop [90], Duda et al. [42], Fukunaga [43], Rumelhart et al. [45], and Vapnik [46].

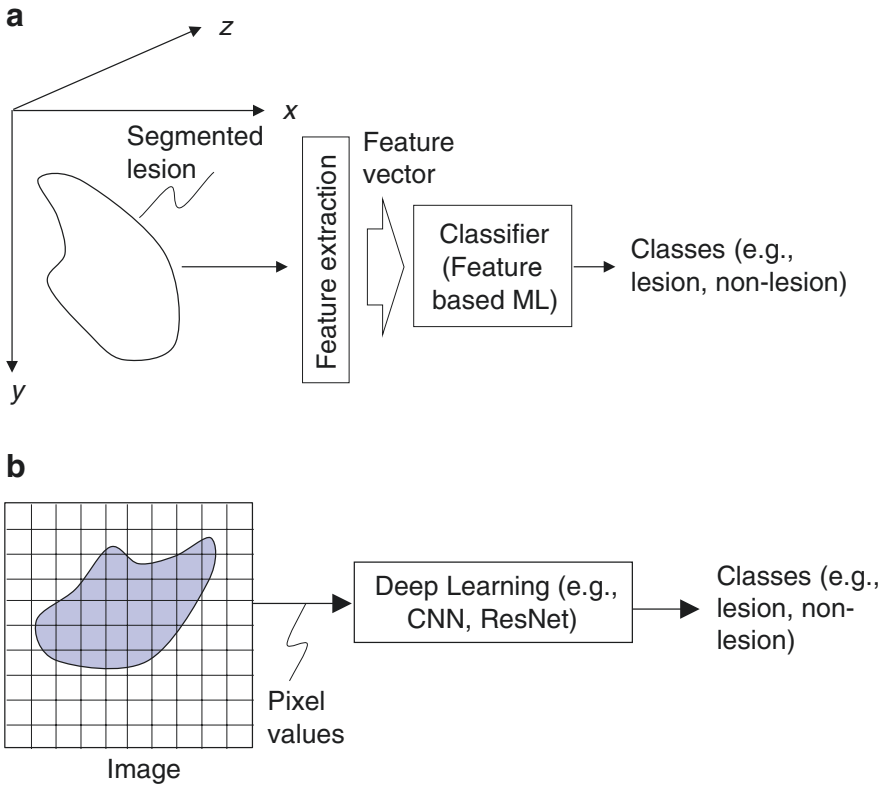


Fig. 9.2 Difference between (**a**) feature-based (segmented-object-based) ML (classifier) and (**b**) deep learning

9.3.2 Early Deep Learning Models

9.3.2.1 Overview

In 2002, an early deep learning model, called massive-training artificial neural networks (MTANNs), was invented and developed in the field of medical imaging [72]. The MTANNs were applied for the detection of lung nodules in chest CT in 2003 [73], end-to-end detection of lung nodules in 2009 [74], separation of bones from soft tissue in chest radiographs in 2004 [75, 76], and reduction of noise and artifacts on CT images in 2013 [77]. CNNs were also applied for CAD in medical imaging before the start of the deep learning boom [70, 91–99].

Predecessor of the MTANNs was developed for tasks in medical image processing/analysis and computer vision: (1) neural filters [100, 101] and (2) neural edge enhancers [102, 103]. Improved models of the MTANNs [23, 73, 74, 76, 104] were developed such as multiple MTANNs [13, 15, 73, 100, 101, 105], a mixture of expert MTANNs [22, 106], a multi-resolution MTANN [76], a Laplacian eigenfunction MTANN (LAP-MTANN) [107], and a massive-training support vector regression (MTSVR) [108]. The class of neural filters was used for image-processing tasks such as edge-preserving noise reduction in fluoroscopy, radiographs, and other digital pictures [100, 101]. The class of neural edge enhancers was used for edge enhancement from noisy images [102] and enhancement of subjective edges traced by a physician in cardiac images [103], which is recently called "semantic segmentation" [109]. The class of MTANNs was used for classification, such as false-positive (FP) reduction in CAD schemes for the detection of lung nodules in chest radiographs (chest X-ray: CXR) [13] and thoracic CT [9, 15, 73], distinction between benign and malignant lung nodules in CT [105], and FP reduction in a CAD scheme for polyp detection in CT colonography [22, 23, 106–108]. The MTANNs were also applied to pattern enhancement and suppression such as separation of bones from soft tissue in CXR [76, 104, 110–112] and enhancement of lung nodules in CT [74].

9.3.2.2 Difference Between Deep Learning and Feature-Based ML (Classifiers)

A major difference between deep learning models and ordinary classifiers (i.e., feature-based ML or segmented-object-based ML) is the input information, as illustrated in Fig. 9.2a, b. Ordinary classifiers use features extracted from a segmented object in a given image, whereas deep learning models use pixel values in an image patch in a given image as the input information. Although the input information to deep learning models can be features (see addition of features to the input information to neural filters in [101], for example), these features are obtained from an image patch pixel by pixel (as opposed to ones from a segmented object or by object). In other words, features for deep learning models are features at each pixel in a given image, whereas features for ordinary classifiers are features from a segmented object. In that sense, feature-based classifiers can be referred to as segmented-object-based classifiers. Because deep learning models use pixel/voxel values in image patches in images directly instead of features calculated from

segmented objects as the input information, segmentation or feature extraction from the segmentation results is not required. Although the development of segmentation techniques has been studied for a long time, segmentation of objects is still challenging, especially for complicated objects, subtle objects, and objects in a complex background. Thus, segmentation errors may occur for such complicated objects. Because with deep learning models, errors caused by inaccurate segmentation and inaccurate feature calculation from the segmentation results can be avoided, the performance of deep learning models can be higher than that of ordinary classifiers for some cases, such as complicated objects. Thus, deep learning does not require steps of segmentation and feature analysis, and it can obtain the final result of lesion detection from the input diagnostic images, as shown in Fig. 9.1, which is called an end-to-end ML paradigm.

A major difference between MTANN deep learning models and ordinary classifiers (or feature-based ML) and other deep learning models is the output information. The output information from ordinary classifiers and other deep learning models such as CNNs is nominal class labels such as normal or abnormal (e.g., 0 or 1), whereas that from neural filters, neural edge enhancers, and MTANNs is pixels in patches (local windows) or images, namely continuous values. Recently, deep learning models that can learn and output images have been proposed such as fully convolutional networks [109] and U-Net [113]. With the scoring method in MTANNs, output images of the MTANNs are converted to likelihood scores for distinguishing among classes, which allow MTANNs to do classification. In addition to classification, MTANNs can perform pattern enhancement and suppression as well as object detection, whereas the other deep learning models cannot.

9.3.2.3 Early Deep Learning Model: Massive-Training Artificial Neural Network (MTANN)

An early deep learning model, an MTANN was developed by extension of the neural filters and neural edge enhancers to accommodate various pattern-recognition tasks [73]. A two-dimensional (2D) MTANN was first developed for distinguishing a specific opacity from other opacities in 2D images [73]. The 2D MTANN was applied to the reduction of FPs in computerized detection of lung nodules on 2D CT images in a slice-by-slice way [9, 15, 73] and in CXR [13], the separation of ribs from soft tissue in CXR [75, 76, 104], and the distinction between benign and malignant lung nodules on 2D CT slices [105]. For processing of three-dimensional (3D) volume data, a 3D MTANN was developed by extending the structure of the 2D MTANN, and it was applied to 3D CT colonography data [22, 23, 106–108] in CADe of polyps.

The generalized architecture of an MTANN is shown in Fig. 9.3. An MTANN consists of an ML model (typically a regression model) such as a linear-output ANN regression model [102] and a support vector regression model [108], which is capable of operating on pixel/voxel data directly [102]. The linear-output ANN regression model uses a linear function instead of a sigmoid function as the activation function of the output-layer unit because the characteristics of an ANN were

improved significantly with a linear function when applied to the continuous mapping of values in image processing [102]. Note that the activation functions of the hidden layer units are a sigmoid function for nonlinear processing, and those of the input layer units an identity function, as usual. The pixel/voxel values of the input images/volumes may be normalized from 0 to 1. The input to the MTANN consists of pixel/voxel values in a subregion/sub-volume (image patch or local window), R, extracted from an input image/volume. The output of the MTANN is a continuous scalar value, which is generally associated with the center voxel in the subregion (image patch), and is represented by

$$O(x,y,z, \text{ or } t) = \mathrm{ML}\{I(x-i,y-j,z-k \text{ or } t-k) | (i,j,k) \in R\}, \tag{9.1}$$

where x, y, and z or t are the coordinate indices, ML(·) is the output of the ML model, and $I(x,y,z$ or $t)$ is a pixel/voxel value of the input image/volume. The structure of input units and the number of hidden units in the ANN may be designed by use of sensitivity-based unit-pruning methods [87, 88]. Other ML models such as support vector regression [46, 47] can be used as a core part of the MTANN. ML regression models rather than ML classification models would be suited for the MTANN framework, because the output of the MTANN is continuous scalar values (as opposed to nominal categories or classes, e.g., 0 or 1). The entire output image/

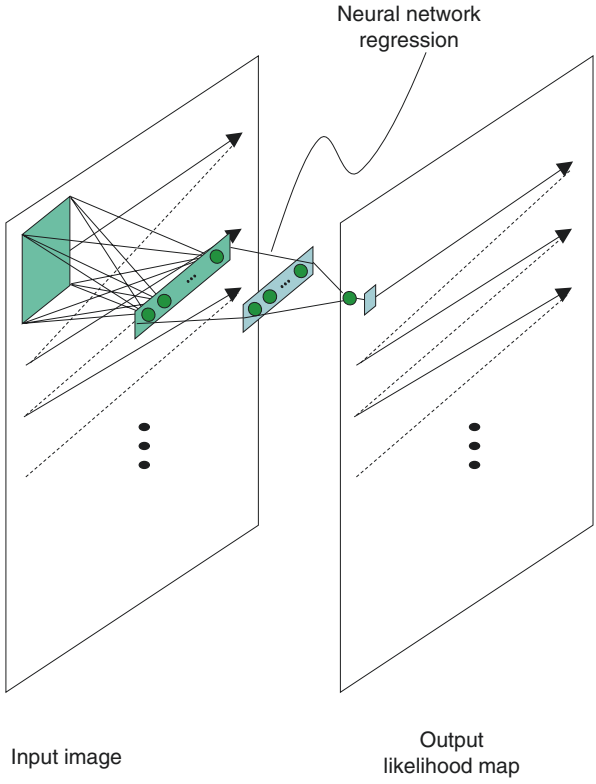


Fig. 9.3 Architecture of an MTANN which is an early deep learning model

volume is obtained by scanning with the input subvolume (local window) of the MTANN on the entire input image/volume. The input subregion/subvolume and the scanning with the MTANN can be analogous to the kernel of a convolution filter and the convolutional operation of the filter, respectively.

The MTANN is trained with input images/volumes and the corresponding "teaching" (designed) images/volumes for enhancement of a specific pattern and suppression of other patterns in images/volumes. The "teaching" images/volumes are ideal or desired images for the corresponding input images/volumes. For enhancement of lesions and suppression of non-lesions, the teaching volume contains a map for the "likelihood of being lesions," represented by

$$T(x,y,z \text{ or } t) = \begin{cases} \text{a certain distribution} & \text{for a lesion} \\ 0 & \text{otherwise.} \end{cases} \tag{9.2}$$

To enrich the training samples, a training region, R_T, extracted from the input images is divided pixel by pixel into a large number of overlapping subregions. Single pixels are extracted from the corresponding teaching images as teaching values. The MTANN is massively trained by use of each of a large number of input subregions (image patches) together with each of the corresponding teaching single pixels, hence the term "massive-training ANN." The error to be minimized by training of the MTANN is represented by

$$E = \frac{1}{P}\sum_{c}\sum_{(x,y,z \text{ or } t)\in R_T}\left\{T_c(x,y,z \text{ or } t) - O_c(x,y,z \text{ or } t)\right\}^2, \tag{9.3}$$

where c is a training case number, O_c is the output of the MTANN for the cth case, T_c is the teaching value for the MTANN for the cth case, and P is the number of total training voxels in the training region for the MTANN, R_T. The expert 3D MTANN is trained by a linear-output back-propagation (BP) algorithm [102] which was derived for the linear-output ANN model by use of the generalized delta rule [45]. After training, the MTANN is expected to output the highest value when a lesion is located at the center of the subregion of the MTANN, a lower value as the distance from the subregion center increases, and zero when the input subregion contains a non-lesion.

9.4 CADe in Thoracic Imaging

9.4.1 Thoracic Imaging for Lung Cancer Detection

Lung cancer continues to rank as the leading cause of cancer deaths in the United States and in other countries such as Japan. Because CT is more sensitive than chest radiography in the detection of small nodules and of lung carcinoma at an early stage [114–117], lung cancer screening programs are being investigated in the United States [118, 119], Japan [115, 117], and other countries with low-dose (LD) CT as the screening modality. Evidence suggests that early detection of lung cancer

may allow more timely therapeutic intervention for patients [117, 120]. Helical CT, however, generates a large number of images that must be interpreted by radiologists/physicians. This may lead to "information overload" for the radiologists/physicians. Furthermore, they may miss some cancers during their interpretation of CT images [27, 121]. Therefore, a CADe scheme for the detection of lung nodules in CT images has been investigated as a tool for lung cancer screening.

9.4.2 CADe of Lung Nodules in Thoracic CT

9.4.2.1 Overview

In 1994, Giger et al. [122] developed a CADe scheme for the detection of lung nodules in CT based on the comparison of geometric features. They applied their CADe scheme to a database of thick-slice diagnostic CT scans. In 1999, Armato et al. [52, 123] extended the method to include 3D feature analysis, a rule-based scheme, and LDA for classification. They tested their CADe scheme with a database of thick-slice (10 mm) diagnostic CT scans. They achieved a sensitivity of 70% with 42.2 FPs per case in a leave-one-out cross-validation test. Gurcan et al. [124] employed a similar approach, i.e., a rule-based scheme based on 2D and 3D features, followed by LDA for classification. They achieved a sensitivity of 84% with 74.4 FPs per case for a database of thick-slice (2.5–5 mm, mostly 5 mm) diagnostic CT scans in a leave-one-out test. Lee et al. [125] employed a simpler approach which is a rule-based scheme based on 13 features for classification. They achieved a sensitivity of 72% with 30.6 FPs per case for a database of thick-slice (10 mm) diagnostic CT scans.

Suzuki et al. [73] developed a deep learning technique called an MTANN for the reduction of a single source of FPs and a multiple MTANN scheme for the reduction of multiple sources of FPs that had not been removed by LDA. They achieved a sensitivity of 80.3% with 4.8 FPs per case for a database of thick-slice (10 mm) screening LDCT scans of 63 patients with 71 nodules with solid, part-solid, and non-solid patterns, including 66 cancers in a validation test. This MTANN approach did not require a large number of training cases: the MTANN was able to be trained with 10 positive and 10 negative cases [126–128], whereas feature-based classifiers generally require 400–800 training cases [126–128]. Arimura et al. [15] employed a rule-based scheme followed by LDA or by the MTANN [73] for classification. They tested their scheme with a database of 106 thick-slice (10 mm) screening LDCT scans of 73 patients with 109 cancers, and they achieved a sensitivity of 83% with 5.8 FPs per case in a validation test (or a leave-one-patient-out test for LDA). Farag et al. [129] developed a template-modeling approach that uses level sets for classification. They achieved a sensitivity of 93.3% with an FP rate of 3.4% for a database of thin-slice screening LDCT scans of 16 patients with 119 nodules and 34 normal patients. Ge et al. [130] incorporated 3D gradient field descriptors and ellipsoid features in LDA for classification. They employed Wilks' lambda stepwise feature selection for selecting features before the LDA classification. They achieved a sensitivity of 80% with 14.7 FPs per case for a database of 82 thin-slice CT scans of 56

patients with 116 solid nodules in a leave-one-patient-out test. Matsumoto et al. [131] employed LDA with eight features for classification. They achieved a sensitivity of 90% with 64.1 FPs per case for a database of thick-slice diagnostic CT scans of five patients with 50 nodules in a leave-one-out test.

Yuan et al. [132] tested a commercially available CADe system (ImageChecker CT, LN-1000, by R2 Technology, Sunnyvale, CA; Hologic now). They achieved a sensitivity of 73% with 3.2 FPs per case for a database of thin-slice (1.25 mm) CT scans of 150 patients with 628 nodules in an independent test. Pu et al. [133] developed a scoring method based on the similarity distance of medial axis-like shapes for classification. They achieved a sensitivity of 81.5% with 6.5 FPs per case for a database of thin-slice screening CT scans of 52 patients with 184 nodules, including 16 non-solid nodules. Retico et al. [134] used a voxel-based neural approach (i.e., a class of the MTANN approach) with pixel values in a subvolume as input for classification. They obtained sensitivities of 80–85% with 10–13 FPs per case for a database of thin-slice screening CT scans of 39 patients with 102 nodules. Ye et al. [54] used a rule-based scheme followed by a weighted SVM for classification. They achieved a sensitivity of 90.2% with 8.2 FPs per case for a database of thin-slice screening CT scans of 54 patients with 118 nodules including 17 non-solid nodules in an independent test. Golosio et al. [135] used a fixed-topology ANN for classification, and they evaluated their CADe scheme with a publicly available database from the Lung Image Database Consortium (LIDC) [136]. They achieved a sensitivity of 79% with 4 FPs per case for a database of thin-slice CT scans of 83 patients with 148 nodules that one radiologist detected from an LIDC database in an independent test.

Murphy et al. [137] used a *k*-nearest-neighbor classifier with features selected from 135 features for classification. They achieved a sensitivity of 80 with 4.2 FPs per case for a large database of thin-slice screening CT scans of 813 patients with 1525 nodules in an independent test. Tan et al. [138] developed a feature-selective classifier based on a genetic algorithm and ANNs for classification. They achieved a sensitivity of 87.5% with 4 FPs per case for a database of thin-slice CT scans of 125 patients with 80 nodules that four radiologists agreed from the LIDC database in an independent test. Messay et al. [139] developed a sequential forward selection process for selecting the optimum features for LDA and quadratic discriminant analysis (QDA). They obtained a sensitivity of 83% with 3 FPs per case for a database of thin-slice CT scans of 84 patients with 143 nodules from the LIDC database in a sevenfold cross-validation test. Riccardi et al. [140] used a heuristic approach based on geometric features, followed by an SVM for classification. They achieved a sensitivity of 71% with 6.5 FPs per case for a database of thin-slice CT scans of 154 patients with 117 nodules that four radiologists agreed on from the LIDC database in a twofold cross-validation test.

Thus, various approaches have been proposed for CADe schemes for lung nodules in CT. Sensitivities for the detection of lung nodules in CT range from 70% to 95%, with from a few to 70 FPs per case. Major sources of FPs are various-sized lung vessels. Major sources of false negatives are ground glass nodules, nodules attached to vessels, and nodules attached to the lung wall (i.e., juxtapleural

nodules). Ground glass nodules are difficult to detect, because they are subtle, of low-contrast, and have ill-defined boundaries. The MTANN approach was able to enhance and thus detect ground-glass nodules [73]. The cause of false negatives due to vessel-attached nodules and juxtapleural nodules is mis-segmentation and thus inaccurate feature calculation. Because the MTANN approach does not require segmentation or feature calculation, it was able to detect such nodules [73].

9.4.2.2 Illustration of a CADe Scheme

Figure 9.4a illustrates an axial slice of a CT scan of the lungs with a lung cancer. The lung cancer on the CT image is the target that we want to detect with a CADe scheme. Figure 9.4b illustrates lung segmentation by simple thresholding followed by mathematical morphology filtering.

Suzuki [11] developed a supervised "lesion enhancement" filter based on an MTANN for enhancing lesions and suppressing non-lesions in medical images. Figure 9.5b illustrates the enhancement of a lung nodule in a CT image by means of a trained MTANN lesion-enhancement filter for the original axial CT slice shown in Fig. 9.5a. In the output image, the lung nodule in the original CT image is enhanced, while normal structures such as lung vessels are suppressed substantially. Figure 9.5c shows the detection and segmentation result for the lung nodule by using simple thresholding followed by the removal of small regions. After thresholding, connected-component labeling [79, 84, 86] was performed to calculate the area of each isolated region (i.e., connected component). By removing small regions, the lung nodule was detected correctly with no FP detection. By use of the MTANN lesion-enhancement filter, detection and segmentation of lung nodules can be realized in an end-to-end fashion.

To reduce remaining FPs, Suzuki et al. developed an FP reduction technique based on MTANNs [73]. The architecture of the MTANN for FP reduction is shown in Fig. 9.6. For enhancement of nodules (i.e., true positives) and suppression of non-nodules (i.e., FPs) on CT images, the teaching image contains a distribution of values that represent the "likelihood of being a nodule." For example, the teaching volume contains a 3D Gaussian distribution with standard deviation σ_T for a lesion

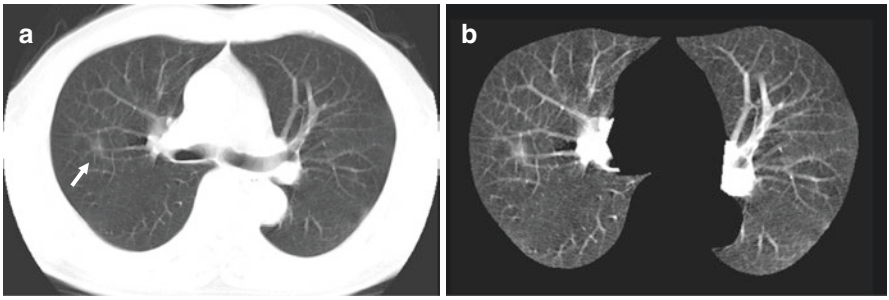


Fig. 9.4 (**a**) Axial slice of a CT scan of the lungs with a lung cancer (indicated by an arrow) and (**b**) a lung segmentation result

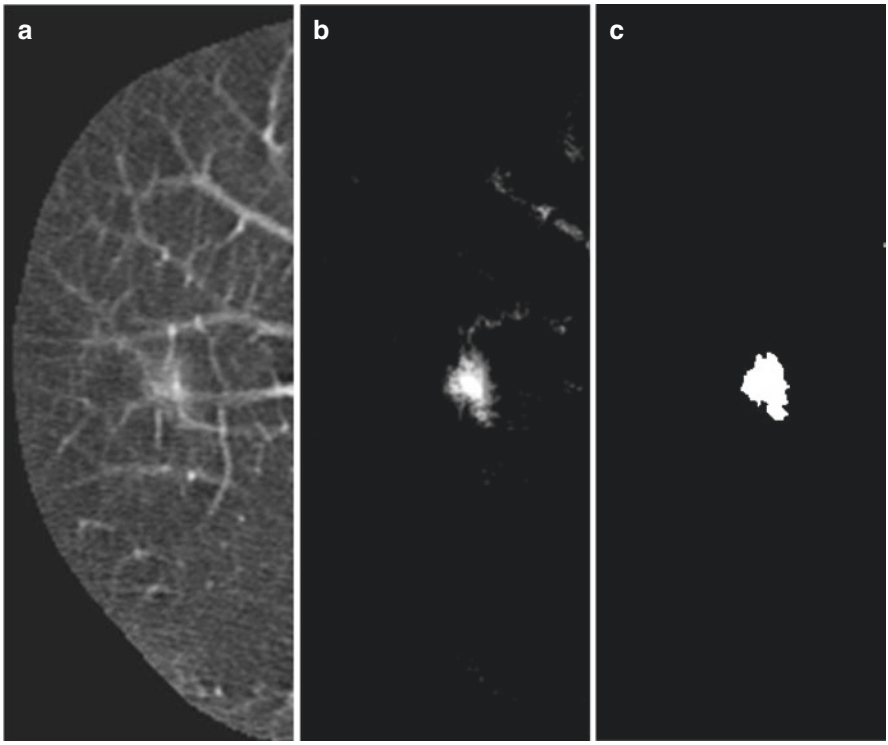


Fig. 9.5 Lesion enhancement by means of a supervised MTANN lesion-enhancement filter. (**a**) Original axial CT slice with a lung nodule. (**b**) Output image of the trained MTANN nodule-enhancement filter. In the output image (**b**), the lung nodule in the original CT image (**a**) is enhanced, whereas normal structures such as lung vessels are suppressed substantially. (**c**) Detection and segmentation of the nodule by using thresholding followed by removal of small regions

and zero (i.e., completely dark) for non-lesions, as illustrated in Fig. 9.6. This distribution represents the "likelihood of being a lesion":

$$T(x,y,z \text{ or } t) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma_T} \exp\left\{-\dfrac{\left(x^2+y^2+z^2 \text{ or } t\right)}{2\sigma_T{}^2}\right\} & \textit{for a} \text{ lesion} \\ 0 & \text{otherwise.} \end{cases} \tag{9.4}$$

A scoring method is used for combining output voxels from the trained MTANNs, as illustrated in Fig. 9.7. A score for a given region-of-interest (ROI) from the MTANN is defined as

$$S = \sum_{(x,y,z \text{ or } t) \in R_E} f_W(x,y,z \text{ or } t) \times O(x,y,z \text{ or } t), \tag{9.5}$$

where

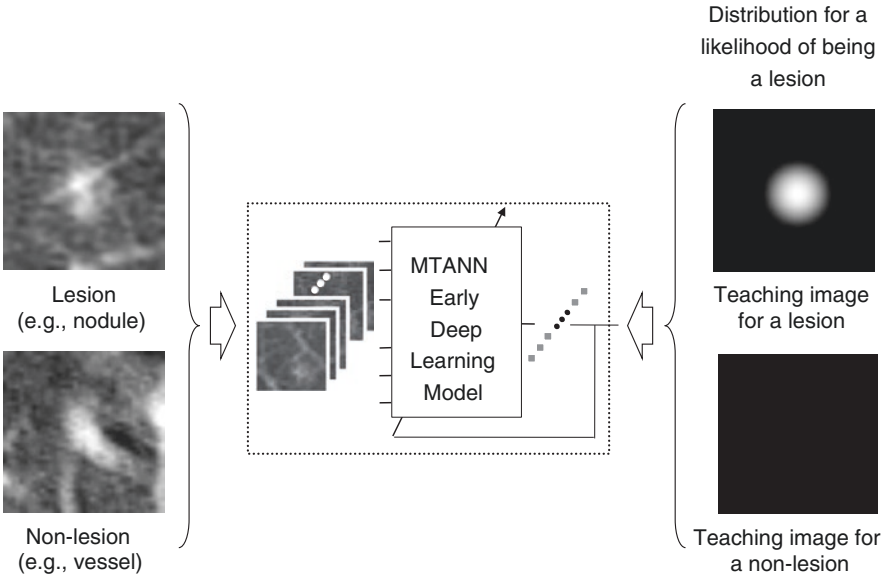


Fig. 9.6 Architecture of an MTANN for FP reduction. The teaching image for a lesion contains a Gaussian distribution; that for a non-lesion contains zero (completely dark). After the training, the MTANN expects to enhance lesions and suppress non-lesions

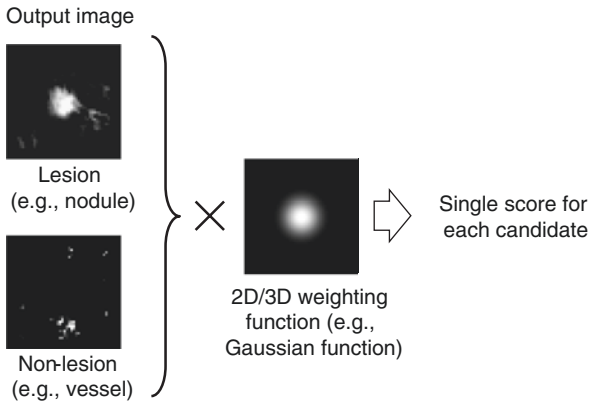


Fig. 9.7 Scoring method for combining pixel-based output responses from the trained MTANN into a single score for each ROI

$$f_{\mathrm{W}}(x,y,z \text{ or } t) = f_{\mathrm{G}}(x,;y,;z \text{ or } t,;\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2+y^2+z^2 \text{or} t^2}{2\sigma^2}} \tag{9.6}$$

is a 3D Gaussian weighting function with standard deviation σ, and with its center corresponding to the center of the volume for evaluation, R_{E}; and O is the output image of the trained MTANN, where its center corresponds to the center of R_{E}. The use of the 3D Gaussian weighting function allows us to combine the

responses (outputs) of a trained MTANN as a 3D distribution. A 3D Gaussian function is used for scoring, because the output of a trained MTANN is expected to be similar to the 3D Gaussian distribution used in the teaching images. This score represents the weighted sum of the estimates for the likelihood that the ROI (lesion candidate) contains a lesion near the center, i.e., a higher score would indicate a lesion, and a lower score would indicate a non-lesion. Thresholding is then performed on the scores for distinction between lesions and non-lesions.

The MTANNs were trained to enhance lung nodules and suppress various types of FPs (i.e., non-nodules) such as lung vessels. Figure 9.8 shows the results of the enhancement of various lung nodules such as non-solid (ground-glass), part-solid (mixed-ground-glass), and solid nodules (a) and those of the suppression of various-sized lung vessels (b). Figure 9.9 shows a free-response receiver operating

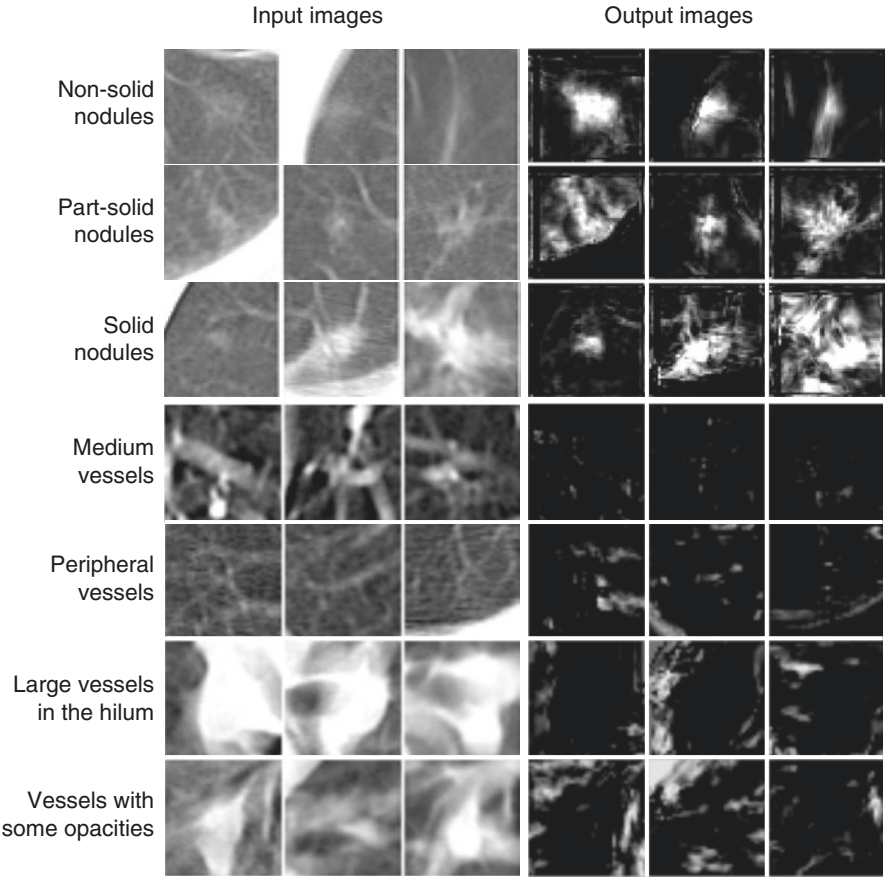


Fig. 9.8 Enhancement of lung nodules and suppression of FPs (i.e., lung vessels) by use of MTANNs for FP reduction. Once lung nodules are enhanced, and FPs are suppressed, FPs can be distinguished from lung nodules by use of scores obtained from the output images

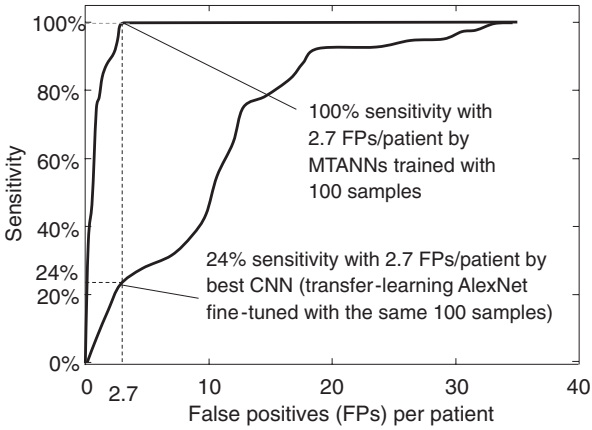


Fig. 9.9 Comparison of the performance of the MTANNs with that of the best CNN in a CADe scheme for detection of lung nodules in CT

characteristic (FROC) curve [141], indicating the performance of the trained MTANNs in the CADe scheme. The performance of well-known CNNs (including the AlexNet, the LeNet, a relatively deep CNN, a shallow CNN, and a fine-tuned AlexNet which used to transfer learning from a computer-vision-trained AlexNet) and MTANNs was compared extensively [142]. Comparison experiments were done for detection of lung nodules in CT with the same databases. The experiments demonstrated that the performance of MTANNs was substantially higher than that of the best-performing CNN under the same condition, as demonstrated in Fig. 9.9. The MTANNs generated 2.7 FPs per patient at 100% sensitivity, which was significantly ($p < 0.05$) lower than that for the best-performing CNN model (Fine-tuned AlexNet), with 22.7 FPs per patient at the same level of sensitivity.

Figure 9.10 shows an example of CADe outputs on a CT image of the lungs. A CADe scheme detected a lung nodule correctly with one FP which was a branch of the lung vessels.

9.4.3 CADe of Lung Nodules in CXR

Chest radiograph (CXR) is the most commonly used imaging examination for chest diseases because it is the most cost-effective, routinely available, and dose-effective diagnostic examination [143, 144]. Because CXRs are widely used, improvements in the detection of lung nodules in CXRs could have a significant impact on early detection of lung cancer. Studies have shown that, however, 30% of nodules in CXRs were missed by radiologists in which nodules were visible in retrospect. Therefore, CADe schemes [12, 14] for nodules in CXRs have been investigated for assisting radiologists in improving their sensitivity. A wide variety of approaches in CADe schemes for nodule detection in CXRs have been developed. Giger et al. developed a difference-image technique to reduce complex anatomic background structures while enhancing nodule-like structures for initial nodule candidate detection [12, 145]. Lo et al. used a technique similar to the difference-image technique

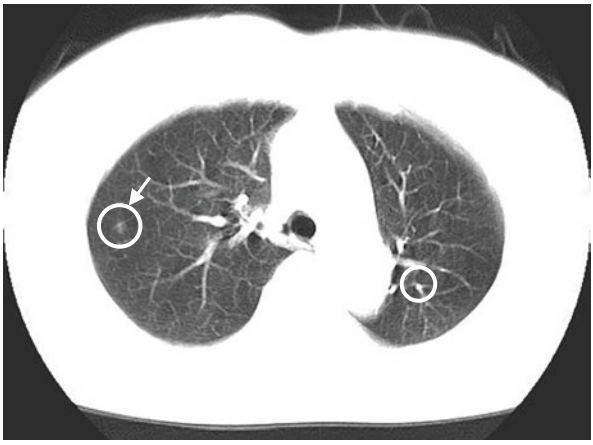


Fig. 9.10 CADe outputs (indicated by circles) on an axial CT slice of the lungs. A lung nodule (indicated by an arrow) was detected correctly by a CADe scheme with one FP detection (branch of lung vessels) on the right

to create nodule-enhanced images, which were then processed by a feature-extraction technique based on edge detection, gray-level thresholding, and sphere profile matching [146, 147]. Then a convolution neural network was employed in the classification step. Penedo et al. then improved the performance of the scheme by incorporating two-level ANNs that employed cross-correlation teaching images and input images in the curvature peak space [148]. Coppini et al. developed a CADe scheme based on biologically inspired ANNs with fuzzy coding [49]. Shiraishi et al. incorporated a localized searching method based on anatomical classification and automated techniques for the parameter setting of three types of ANNs into a CADe scheme [51].

Studies showed that 82–95% of the missed lung cancers in CXR were partly obscured by overlying bones such as ribs and/or a clavicle [149, 150]. To address this issue, Suzuki et al. [76, 151] developed a multiresolution MTANN for the separation of bones such as ribs and clavicles from soft tissue in CXRs. They employed multiresolution decomposition/composition techniques [152, 153] to decompose an original high-resolution image into different-resolution images. First, one obtains a medium-resolution image $g_M(x,y)$ from an original high-resolution image $g_H(x,y)$ by performing down-sampling with averaging, i.e., four pixels in the original image are replaced by a pixel having the mean value for the four pixel values, represented by.

$$g_{\mathrm{M}}(x,y) = \frac{1}{4}\sum_{i,j\in R_{22}} g_{\mathrm{H}}(2x-i,2y-j), \tag{9.7}$$

where R_{22} is a 2-by-2-pixel region. The medium-resolution image is enlarged by up-sampling with pixel substitution, i.e., a pixel in the medium-resolution image is replaced by four pixels with the same pixel value, as follows:

$$g_{\mathrm{M}}^{\mathrm{U}}(x,y) = g_{\mathrm{M}}(x/2, y/2). \tag{9.8}$$

Then, a high-resolution difference image $d_{\mathrm{H}}(x,y)$ is obtained by subtraction of the enlarged medium-resolution image from the high-resolution image, represented by

$$d_H(x,y) = g_H(x,y) - g_M^U(x,y). \tag{9.9}$$

These procedures are performed repeatedly, producing further lower-resolution images. Thus, multiresolution images having various frequencies are obtained by use of the multiresolution decomposition technique.

An important property of this technique is that exactly the same original-resolution image $g_H(x,y)$ can be obtained from the multi-resolution images, $d_H(x,y)$ and $g_M(x,y)$, by performing the inverse procedures, called a multi-resolution composition technique, as follows:

$$g_H(x,y) = g_M(x/2, y/2) + d_H(x,y). \tag{9.10}$$

Therefore, we can process multiresolution images independently instead of processing original high-resolution images directly; i.e., with these techniques, the processed original high-resolution image can be obtained by composing of the processed multiresolution images. Each of multiple MTANNs only needs to support a limited spatial frequency range in each resolution image instead of the entire spatial frequencies in the original image.

First, input CXRs and the corresponding teaching bone images are decomposed into sets of different-resolution images, and then these sets of images are used for training three MTANNs in the multiresolution MTANN. Each MTANN is an expert for a certain resolution, i.e., a low-resolution MTANN is in charge of low-frequency components of ribs, a medium-resolution MTANN is for medium-frequency components, and a high-resolution MTANN for high-frequency components. Each resolution MTANN is trained independently with the corresponding resolution images. After training, the MTANN produce different-resolution images, and then these images are composed to provide a complete high-resolution image by use of the multiresolution composition technique. The complete high-resolution image is expected to be similar to the teaching bone image; therefore, the multiresolution MTANN would provide a "bone-image-like" image in which ribs and clavicles are separated from soft tissues. Chen and Suzuki [111, 112] improved the performance of the MTANN "virtual" dual-energy chest radiography by means of anatomically specific multiple MTANNs. Zarshenas and Suzuki improved the MTANN by incorporating the wavelet transform [154]. Figure 9.11 illustrates suppression of bones from soft tissue in CXR by using the MTANNs [111].

Suzuki et al. developed an FP reduction technique based on MTANNs in a CADe scheme of nodules in CXR. They removed 68% of the FPs that had not been removed by feature-based ML, and the performance of the CADe scheme was substantially improved from 4.5 to 1.4 FPs per image, while maintaining the original sensitivity of 81.3%.

Chen and Suzuki developed a CADe scheme of lung nodules in CXRs based on feature-based SVM [155]. They improved the performance by using the MTANN virtual dual-energy imaging [110]. They improved the performance substantially from the original sensitivity of 79% with 5 FPs per image to a sensitivity of 85% with the same FP rate. Figure 9.12 illustrates computer outputs from their CADe scheme without and with the MTANN virtual dual-energy imaging [110].

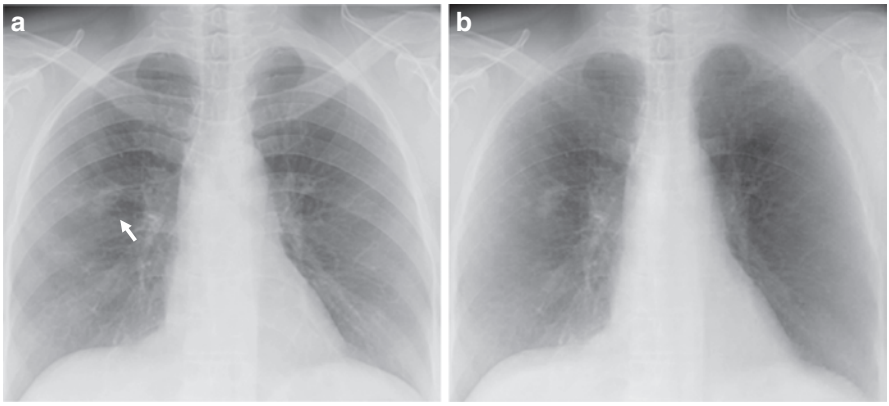


Fig. 9.11 Suppression of bones such as ribs and clavicles from soft tissue in CXR. (**a**) Original CXR with a lung nodule (indicated by an arrow). (**b**) Bone suppression imaging (or "virtual" dual-energy radiography) result by means of a multiresolution MTANN

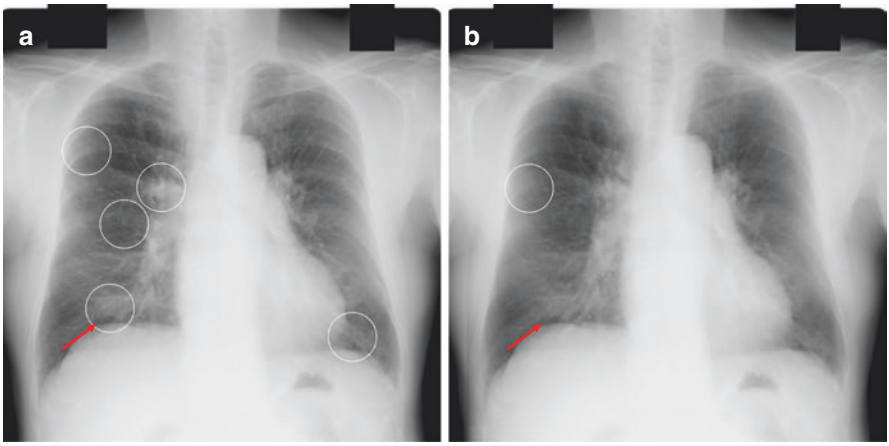


Fig. 9.12 Illustration of the improvement in nodule detection by CADe scheme with our VDE technology. CADe marks are indicated by circles. (**a**) False negatives (arrow) and false positives of the original CADe scheme. (**b**) True positives (arrow) and false positives of the VDE-based CADe scheme with the VDE technology

They compared the performance of their CADe scheme with that of an FDA-approved CADe product with the same database. Their CADe scheme achieved a sensitivity of 81% with 2.0 FPs per image, whereas the FDA-approved product achieved a substantially inferior performance that had a sensitivity of 67% at the same FP rate. They also compared the performance with other CADe schemes in literature by using the same publicly available database of the JSRT [156]. Wei et al. reported that their CAD scheme achieved a sensitivity of 80% with 5.4 FPs per image. Hardie et al. reported that their scheme marked 80% of nodules with 5 FPs

per image [50]. The performance of Chen Suzuki CADe scheme was substantially higher than that of Hardie's CADe scheme, i.e., it achieved a sensitivity of 78% at an FP rate of 2.0 per image, whereas Hardie's CADe scheme achieved a sensitivity of 63% at the FP rate.

9.5 CADe in Colonic Imaging

9.5.1 Colonic Imaging for Colorectal Cancer Detection

Colorectal cancer is the second leading cause of cancer deaths in the United States [157]. Evidence suggests that early detection and removal of polyps (i.e., precursors of colorectal cancer) can reduce the incidence of colorectal cancer [158, 159]. Consequently, the American Cancer Society (ACS) recommends that an individual who is at average risk for developing colorectal cancer, beginning at the age of 50 years, should have colorectal cancer screening with examinations including optical colonoscopy and CTC. CTC (or virtual colonoscopy) is a technique for detecting colorectal neoplasms by using CT scans of the colon [160]. The diagnostic performance of CTC in detecting polyps, however, varies by experience of radiologists, hospitals, and protocols [161]. Therefore, CADe of polyps has been investigated to address this issue with CTC [162–164].

9.5.2 Overview of CADe of Polyps in CTC

CADe has the potential to (a) increase radiologists' sensitivity in the detection of polyps, (b) decrease reader variability, and (c) reduce radiologists' reading time when CADe is used during the primary read [163, 164]. A number of researchers have developed CADe schemes for the detection of polyps in CTC [62, 165–170]. Figure 9.13 shows an example of a CADe output for detection of polyps in CTC. A CADe scheme detected the polyp correctly.

In 2000, Summers et al. [21] developed a CADe scheme for the detection of polyps in CTC based on curvature analysis. In 2001, Yoshida and Nappi [62] developed a CADe scheme based on curvature analysis called a shape index. In 2001, Gokturk et al. [171] employed an SVM with histogram input that is used as a shape signature for classification. Näppi and Yoshida [172] developed a CADe scheme based on LDA or QDA with 54 volumetric features (9 statistics of 6 features). Acar et al. [173] used edge-displacement fields and QDA for classification. Jerebko et al. [60] used a multilayer perceptron to classify polyp candidates in their CADe scheme and improved the performance by incorporating a committee of multilayer perceptrons [174] and a committee of SVMs [175]. Wang et al. [176] developed a classification method based on LDA with internal features (geometric, morphologic, and textural) of polyps.

Suzuki et al. [23] developed a 3D MTANN by extending the structure of a 2D MTANN [17] to process 3D volume data in CTC. Their CADe scheme was based

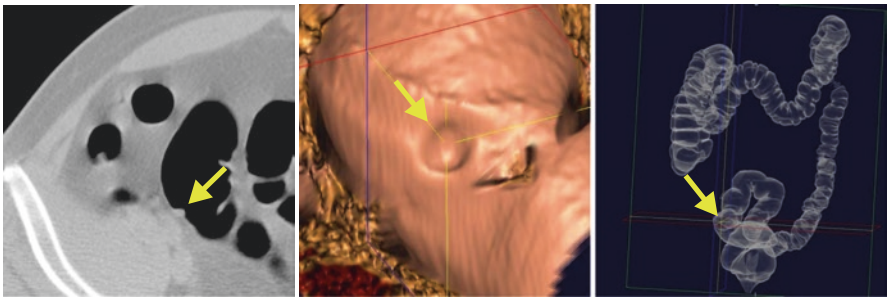


Fig. 9.13 CADe output (indicated by an arrow) for the detection of polyps in an axial slice, an endoluminal view, and a 3D colon view in CTC. A polyp (indicated by an arrow) was detected correctly by a CADe scheme

on a Bayesian ANN with texture and geometric features, followed by 3D MTANNs. They removed FPs due to rectal tubes by using a single 3D MTANN [23] and multiple sources of FPs by developing and using a mixture of expert 3D MTANNs [22].

Li et al. [177] developed a classification method based on an SVM classifier with wavelet-based features. Wang et al. [61] improved the SVM performance by using nonlinear dimensionality reduction (i.e., a diffusion map and locally linear embedding). Yao et al. [178] employed a topographic height map for calculating features for an SVM classifier.

Suzuki et al. [106] tested a CADe scheme based on a Bayesian ANN and MTANNs. They used CTC data of 24 patients, including 23 polyps (6–25 mm) and a mass (35 mm), that had been "missed" by radiologists [179] in a multicenter clinical trial [180]. They achieved a by-polyp (by-patient) sensitivity of 96.4% (100%) with 1.1 FPs/patient in a leave-one-lesion-out cross-validation test of the classification part. Suzuki et al. [107, 181] also improved the efficiency of the MTANN approach by incorporating principal-component analysis-based and Laplacian eigenmap-based dimension reduction techniques. Xu and Suzuki [108] showed that other nonlinear regression models such as support vector and nonlinear Gaussian process regression models instead of the ANN regression model could be used as the core model in the MTANN framework.

Zhou et al. [182] developed projection features for an SVM classifier. Wang et al. [183] improved the performance of a CAD scheme by adding statistical curvature features in multiple-kernel learning. They obtained a sensitivity of 83% with 5 FPs/patient in a leave-one-out cross-validation test of the classification part.

Thus, various ML approaches have been proposed in CADe schemes for polyps in CTC, which include LDA, QDA, an SVM, ANNs, and a Bayesian ANN.

Existing CADe schemes tend to miss superficially elevated neoplasms (often called flat lesions) [28, 184]. Suzuki et al. developed a CADe scheme for the detection of superficially elevated neoplasms [185]. Detection of superficially elevated neoplasms is very important, because they are histologically aggressive, and because they are often missed by radiologists in CTC as well as by gastroenterologists in optical colonoscopy.

9.6 Summary

In this chapter, ML techniques and early deep learning models used in CAD schemes for detection of lung nodules in CXR and thoracic CT and those for detection of polyps in CTC are described. Before deep learning was introduced, feature-based (segmented-object-based) ML (classifiers) had been dominant and used in one of steps in a CADe scheme. Deep learning is an end-to-end ML paradigm which skips multiple steps in a CADe scheme. In CAD and medical imaging fields, early deep learning models called neural filters, neural edge enhancers, and MTANNs were developed for detection, classification, and image processing tasks in medical imaging. MTANNs have advantages of small-sample-size training, high performance, stable training, and efficient training and computation over other deep learning models.

Acknowledgments This work would not have been possible without the help and support of countless people. The author is grateful to all members in the Suzuki laboratory, i.e., postdoctoral scholars, computer scientists, visiting scholars/professors, medical students, graduate/undergraduate students, research technicians, research volunteers, and support staff, in the Department of Radiology at the University of Chicago, in the Medical Imaging Research Center at the Illinois Institute of Technology, for their invaluable assistance in the studies, to colleagues and collaborators for their valuable suggestions. CAD technologies, MTANNs technologies, the bone separation technology, and their source code developed at the University of Chicago have been licensed to companies including R2 Technology (Hologic), Riverain Medical (Riverain Technologies), Median Technologies, and AlgoMedica.

References

1. Doi K. Current status and future potential of computer-aided diagnosis in medical imaging. Br J Radiol. 2005;78(Spec No 1): S3–19.
2. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph. 2007;31:198–211.
3. Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. Med Phys. 2008;35:5799–820.
4. Giger ML, Suzuki K. Computer-aided diagnosis (CAD). In: Feng DD, editor. Biomedical information technology. San Diego: Academic Press; 2007. p. 359–74.
5. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Sanjay-Gopal S. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology. 1999;212:817–27.
6. Dachman AH, Obuchowski NA, Hoffmeister JW, Hinshaw JL, Frew MI, Winter TC, Van Uitert RL, Periaswamy S, Summers RM, Hillman BJ. Effect of computer-aided detection for CT colonography in a multireader, multicase trial. Radiology. 2010;256:827–35. https://doi.org/10.1148/radiol.10091890.
7. Dean JC, Ilvento CC. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. AJR Am J Roentgenol. 2006;187:20–8.
8. Li F, Aoyama M, Shiraishi J, Abe H, Li Q, Suzuki K, Engelmann R, Sone S, Macmahon H, Doi K. Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy. AJR Am J Roentgenol. 2004;183:1209–15.

9. Li F, Arimura H, Suzuki K, Shiraishi J, Li Q, Abe H, Engelmann R, Sone S, MacMahon H, Doi K. Computer-aided detection of peripheral lung cancers missed at CT: ROC analyses without and with localization. Radiology. 2005;237:684–90.
10. Petrick N, Haider M, Summers RM, Yeshwant SC, Brown L, Iuliano EM, Louie A, Choi JR, Pickhardt PJ. CT colonography with computer-aided detection as a second reader: observer performance study. Radiology. 2008;246:148–56.
11. Suzuki K, Hori M, McFarland E, Friedman AC, Rockey DC, Dachman AH. Can CAD help improve the performance of radiologists in detection of difficult polyps in CT colonography? In: Proceedings of RSNA annual meeting, Chicago; 2009. p. 872.
12. Giger ML, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. Med Phys. 1988;15:158–66.
13. Suzuki K, Shiraishi J, Abe H, MacMahon H, Doi K. False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network. Acad Radiol. 2005;12:191–201.
14. van Ginneken B, ter Haar Romeny BM, Viergever MA. Computer-aided diagnosis in chest radiography: a survey. IEEE Trans Med Imaging. 2001;20:1228–41.
15. Arimura H, Katsuragawa S, Suzuki K, Li F, Shiraishi J, Sone S, Doi K. Computerized scheme for automated detection of lung nodules in low-dose computed tomography images for lung cancer screening. Acad Radiol. 2004;11:617–29.
16. Armato SG 3rd, Li F, Giger ML, MacMahon H, Sone S, Doi K. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. Radiology. 2002;225:685–92.
17. Suzuki K, Armato SG, Li F, Sone S, Doi K. Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose CT. Med Phys. 2003;30:1602–17.
18. Chan HP, Doi K, Galhotra S, Vyborny CJ, MacMahon H, Jokich PM. Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography. Med Phys. 1987;14:538–48.
19. Gilhuijs KG, Giger ML, Bick U. Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. Med Phys. 1998;25:1647–54.
20. Drukker K, Giger ML, Metz CE. Robustness of computerized lesion detection and classification scheme across different breast US platforms. Radiology. 2005;237:834–40.
21. Summers RM, Beaulieu CF, Pusanik LM, Malley JD, Jeffrey RB Jr, Glazer DI, Napel S. Automated polyp detector for CT colonography: feasibility study. Radiology. 2000;216:284–90.
22. Suzuki K, Yoshida H, Nappi J, Armato SG 3rd, Dachman AH. Mixture of expert 3D massive-training ANNs for reduction of multiple types of false positives in CAD for detection of polyps in CT colonography. Med Phys. 2008;35:694–703.
23. Suzuki K, Yoshida H, Nappi J, Dachman AH. Massive-training artificial neural network (MTANN) for reduction of false positives in computer-aided detection of polyps: Suppression of rectal tubes. Med Phys. 2006;33:3814–24.
24. Suzuki K. Pixel-based Machine Learning (PML) in medical imaging. Int J Biomed Imaging. 2012;2012:792079, 18p.
25. Suzuki K. A review of computer-aided diagnosis in thoracic and colonic imaging. Quant Imaging Med Surg. 2012;2:163–76. https://doi.org/10.3978/j.issn.2223-4292.2012.09.02.
26. Suzuki K. Machine learning in computer-aided diagnosis of the thorax and colon in CT: a survey. IEICE Trans Inf Syst. 2013;E96-D:772–83.
27. Li F, Sone S, Abe H, MacMahon H, Armato SG 3rd, Doi K. Lung cancers missed at low-dose helical CT screening in a general population: comparison of clinical, histopathologic, and imaging findings. Radiology. 2002;225:673–83.
28. Lostumbo A, Wanamaker C, Tsai J, Suzuki K, Dachman AH. Comparison of 2D and 3D views for evaluation of flat lesions in CT colonography. Acad Radiol. 2010;17:39–47. https://doi.org/10.1016/j.acra.2009.07.004. pii: S1076-6332(09)00400-0.

29. Soetikno RM, Kaltenbach T, Rouse RV, Park W, Maheshwari A, Sato T, Matsui S, Friedland S. Prevalence of nonpolypoid (flat and depressed) colorectal neoplasms in asymptomatic and symptomatic adults. JAMA. 2008;299:1027–35.
30. Shen D, Wu G, Zhang D, Suzuki K, Wang F, Yan P. Machine learning in medical imaging. Comput Med Imaging Graph. 2015;41:1–2. https://doi.org/10.1016/j.compmedimag.2015.02.001.
31. Suzuki K, Zhou L, Wang Q. Machine learning in medical imaging. Pattern Recognit. 2017;63:465–7. https://doi.org/10.1016/j.patcog.2016.10.020.
32. Yan P, Suzuki K, Wang F, Shen D. Machine learning in medical imaging. Mach Vis Appl. 2013;24:1327–9. https://doi.org/10.1007/s00138-013-0543-8.
33. Shen D, Wu G, Zhang D, Yan P, Suzuki K, Wang F. Machine learning in medical imaging. Comput Med Imaging Graph. 2014;41:1–2.
34. Suzuki K. Machine learning for medical imaging. Algorithms; 2010.
35. Suzuki K. Machine learning for medical imaging 2012. Algorithms; 2012.
36. Suzuki K, Yan P, Wang F, Shen D. Machine learning in medical imaging. Int J Biomed Imaging. 2012;2012:123727. https://doi.org/10.1155/2012/123727.
37. Yan P, Suzuki K, Wang F, Shen D. Machine learning in medical imaging. Mach Vis Appl. 2012;24:1327.
38. Suzuki K, Wang F, Shen D, Yan P. Machine learning in medical imaging (MLMI). Lecture notes in computer science, vol. 7009. Berlin: Springer; 2011. p. 355.
39. Wang F, Shen D, Yan P, Suzuki K. Machine learning in medical imaging (MLMI). Lecture notes in computer science, vol. 7588. Berlin: Springer; 2012. p. 276.
40. Wang F, Yan P, Suzuki K, Shen D. Machine learning in medical imaging (MLMI). Lecture notes in computer science, vol. 6357. Berlin: Springer; 2010. p. 192.
41. Wu G, Zhang D, Shen D, Yan P, Suzuki K, Wang F. Machine learning in medical imaging (MLMI). Lecture notes in computer science, vol. 8184. Berlin: Springer; 2013. p. 262.
42. Duda RO, Hart PE, Stork DG. Pattern recognition. 2nd ed. Hoboken: Wiley Interscience; 2001.
43. Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. San Diego: Academic Press; 1990.
44. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. Parallel Distrib Process. 1986;1:318–62.
45. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6.
46. Vapnik VN. The nature of statistical learning theory. Berlin: Springer; 1995.
47. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
48. Chen S, Suzuki K, MacMahon H. A computer-aided diagnostic scheme for lung nodule detection in chest radiographs by means of two-stage nodule-enhancement with support vector classification. Med Phys. 2011;38:1844–58.
49. Coppini G, Diciotti S, Falchini M, Villari N, Valli G. Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms. IEEE Trans Inf Technol Biomed. 2003;7:344–57.
50. Hardie RC, Rogers SK, Wilson T, Rogers A. Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. Med Image Anal. 2008;12:240–58. https://doi.org/10.1016/j.media.2007.10.004. pii: S1361-8415(07)00103-X.
51. Shiraishi J, Li Q, Suzuki K, Engelmann R, Doi K. Computer-aided diagnostic scheme for the detection of lung nodules on chest radiographs: localized search method based on anatomical classification. Med Phys. 2006;33:2642–53.
52. Armato SG 3rd, Giger ML, MacMahon H. Automated detection of lung nodules in CT scans: preliminary results. Med Phys. 2001;28:1552–61.
53. Way TW, Sahiner B, Chan HP, Hadjiiski L, Cascade PN, Chughtai A, Bogot N, Kazerooni E. Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. Med Phys. 2009;36:3086–98.

54. Ye X, Lin X, Dehmeshki J, Slabaugh G, Beddoe G. Shape-based computer-aided detection of lung nodules in thoracic CT images. IEEE Trans Biomed Eng. 2009;56:1810–20. https://doi.org/10.1109/TBME.2009.2017027.
55. El-Naqa I, Yang Y, Wernick MN, Galatsanos NP, Nishikawa RM. A support vector machine approach for detection of microcalcifications. IEEE Trans Med Imaging. 2002;21:1552–63.
56. Ge J, Sahiner B, Hadjiiski LM, Chan HP, Wei J, Helvie MA, Zhou C. Computer aided detection of clusters of microcalcifications on full field digital mammograms. Med Phys. 2006;33:2975–88.
57. Wu Y, Doi K, Giger ML, Nishikawa RM. Computerized detection of clustered microcalcifications in digital mammograms: applications of artificial neural networks. Med Phys. 1992;19:555–60.
58. Yu SN, Li KY, Huang YK. Detection of microcalcifications in digital mammograms using wavelet filter and Markov random field model. Comput Med Imaging Graph. 2006;30:163–73.
59. Wu YT, Wei J, Hadjiiski LM, Sahiner B, Zhou C, Ge J, Shi J, Zhang Y, Chan HP. Bilateral analysis based false positive reduction for computer-aided mass detection. Med Phys. 2007;34:3334–44.
60. Jerebko AK, Summers RM, Malley JD, Franaszek M, Johnson CD. Computer-assisted detection of colonic polyps with CT colonography using neural networks and binary classification trees. Med Phys. 2003;30:52–60.
61. Wang S, Yao J, Summers RM. Improved classifier for computer-aided polyp detection in CT colonography by nonlinear dimensionality reduction. Med Phys. 2008;35:1377–86.
62. Yoshida H, Nappi J. Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps. IEEE Trans Med Imaging. 2001;20:1261–74.
63. Muramatsu C, Li Q, Schmidt R, Suzuki K, Shiraishi J, Newstead G, Doi K. Experimental determination of subjective similarity for pairs of clustered microcalcifications on mammograms: observer study results. Med Phys. 2006;33:3460–8.
64. Muramatsu C, Li Q, Schmidt RA, Shiraishi J, Suzuki K, Newstead GM, Doi K. Determination of subjective similarity for pairs of masses and pairs of clustered microcalcifications on mammograms: comparison of similarity ranking scores and absolute similarity ratings. Med Phys. 2007;34:2890–5.
65. Muramatsu C, Li Q, Suzuki K, Schmidt RA, Shiraishi J, Newstead GM, Doi K. Investigation of psychophysical measure for evaluation of similar images for mammographic masses: preliminary results. Med Phys. 2005;32:2295–304.
66. Arimura H, Li Q, Korogi Y, Hirai T, Katsuragawa S, Yamashita Y, Tsuchiya K, Doi K. Computerized detection of intracranial aneurysms for three-dimensional MR angiography: feature extraction of small protrusions based on a shape-based difference image technique. Med Phys. 2006;33:394–401.
67. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. https://doi.org/10.1038/nature14539.
68. Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol. 2017;10:257–73. https://doi.org/10.1007/s12194-017-0406-5.
69. Suzuki K. Survey of deep learning applications to medical image analysis. Med Imaging Technol. 2017;35:212–26.
70. Lawrence S, Giles CL, Tsoi AC, Back AD. Face recognition: a convolutional neural-network approach. IEEE Trans Neural Netw. 1997;8:98–113.
71. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–105.
72. Suzuki K, Doi K. Massive training artificial neural network (MTANN) for detecting abnormalities in medical images. United States Patent; 2002.
73. Suzuki K, Armato SG 3rd, Li F, Sone S, Doi K. Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. Med Phys. 2003;30:1602–17.

74. Suzuki K. A supervised 'lesion-enhancement' filter by use of a massive-training artificial neural network (MTANN) in computer-aided diagnosis (CAD). Phys Med Biol. 2009;54:S31–45. https://doi.org/10.1088/0031-9155/54/18/S03. pii: S0031-9155(09)14266-5.
75. Suzuki K, Abe H, Li F, Doi K. Suppression of the contrast of ribs in chest radiographs by means of massive training artificial neural network. In: Proceedings of SPIE medical imaging (SPIE MI), San Diego; 2004, p. 1109–9.
76. Suzuki K, Abe H, MacMahon H, Doi K. Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN). IEEE Trans Med Imaging. 2006;25:406–16. https://doi.org/10.1109/TMI.2006.871549.
77. Suzuki K, Liu Y, Higaki T, Funama Y, Awai K. Supervised conversion of ultra-low-dose to higher-dose CT images by using pixel-based machine learning: phantom and initial patient studies. In: Program of scientific assembly and annual meeting of Radiological Society of North America (RSNA), SST14-06, Chicago, vol. SST14-06; 2013.
78. He L, Chao YKS, Yu Q, Tang W, Shi Z. An algorithm for labeling connected components and holes. Am J Eng Technol Res. 2011;11:2149–54.
79. He L, Chao Y, Suzuki K. A run-based two-scan labeling algorithm. IEEE Trans Image Process. 2008;17:749–56. https://doi.org/10.1109/TIP.2008.919369.
80. He L, Chao Y, Suzuki K. Two efficient label-equivalence-based connected-component labeling algorithms for 3-D binary images. IEEE Trans Image Process. 2011;20:2122–34. https://doi.org/10.1109/TIP.2011.2114352.
81. He L, Chao Y, Suzuki K. A new first-scan method for two-scan labeling algorithms. In: Computer-aided diagnosis systems for lung cancer, vol. E95-D; 2012. p. 2142–5.
82. He L, Chao Y, Suzuki K. Configuration-transition-based connected-component labeling. IEEE Trans Image Process. 2014;23:943–51. https://doi.org/10.1109/TIP.2013.2289968.
83. He L, Chao Y, Suzuki K, Nakamura T. A new first-scan strategy for raster-scan-based labeling algorithms. J Inf Process Soc Jpn. 2011;52:1813–9.
84. He L, Chao Y, Suzuki K, Wu K. Fast connected-component labeling. Pattern Recognit. 2009;42:1977–87.
85. He L, Chao Y, Yang Y, Li S, Zhao X, Suzuki K. A novel two-scan connected-component labeling algorithm. IAENG Trans Eng Technol. 2013:445–59.
86. Suzuki K, Horiba I, Sugie N. Linear-time connected-component labeling based on sequential local operations. Comput Vis Image Understand. 2003;89:1–23.
87. Suzuki K. Determining the receptive field of a neural filter. J Neural Eng. 2004;1:228–37. https://doi.org/10.1088/1741-2560/1/4/006. pii: S1741-2560(04)85485-5.
88. Suzuki K, Horiba I, Sugie N. A simple neural network pruning algorithm with application to filter synthesis. Neural Process Lett. 2001;13:43–53.
89. Xu J-W, Suzuki K. Max-AUC feature selection in computer-aided detection of polyps in CT colonography. IEEE J Biomed Health Inform. 2014;18:585–93. https://doi.org/10.1109/JBHI.2013.2278023.
90. Bishop CM. Neural networks for pattern recognition. New York: Oxford University Press; 1995.
91. Lin JS, Lo SB, Hasegawa A, Freedman MT, Mun SK. Reduction of false positives in lung nodule detection using a two-level neural classification. IEEE Trans Med Imaging. 1996;15:206–17. https://doi.org/10.1109/42.491422.
92. Lo SB, Lou SA, Lin JS, Freedman MT, Chien MV, Mun SK. Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans Med Imaging. 1995;14:711–8. https://doi.org/10.1109/42.476112.
93. Lo SC, Li H, Wang Y, Kinnard L, Freedman MT. A multiple circular path convolution neural network system for detection of mammographic masses. IEEE Trans Med Imaging. 2002;21:150–8. https://doi.org/10.1109/42.993133.
94. Lo SCB, Chan HP, Lin JS, Li H, Freedman MT, Mun SK. Artificial convolution neural network for medical image pattern recognition. Neural Netw. 1995;8:1201–14.
95. Neubauer C. Evaluation of convolutional neural networks for visual recognition. IEEE Trans Neural Netw. 1998;9:685–96.

96. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, Goodsitt MM. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging. 1996;15:598–610. https://doi.org/10.1109/42.538937.
97. Wei D, Nishikawa RM, Doi K. Application of texture analysis and shift-invariant artificial neural network to microcalcification cluster detection. Radiology. 1996;201:696.
98. Zhang W, Doi K, Giger ML, Nishikawa RM, Schmidt RA. An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. Med Phys. 1996;23:595–601.
99. Zhang W, Doi K, Giger ML, Wu Y, Nishikawa RM, Schmidt RA. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. Med Phys. 1994;21:517–24.
100. Suzuki K, Horiba I, Sugie N. Efficient approximation of neural filters for removing quantum noise from images. IEEE Trans Signal Process. 2002;50:1787–99.
101. Suzuki K, Horiba I, Sugie N, Nanki M. Neural filter with selection of input features and its application to image quality improvement of medical image sequences. IEICE Trans Inf Syst. 2002;E85-D:1710–8.
102. Suzuki K, Horiba I, Sugie N. Neural edge enhancer for supervised edge enhancement from noisy images. IEEE Trans Pattern Anal Mach Intell. 2003;25:1582–96.
103. Suzuki K, Horiba I, Sugie N, Nanki M. Extraction of left ventricular contours from left ventriculograms by means of a neural edge detector. IEEE Trans Med Imaging. 2004;23:330–9.
104. Oda S, Awai K, Suzuki K, Yanaga Y, Funama Y, MacMahon H, Yamashita Y. Performance of radiologists in detection of small pulmonary nodules on chest radiographs: effect of rib suppression with a massive-training artificial neural network. AJR Am J Roentgenol. 2009;193:W397–402. https://doi.org/10.2214/AJR.09.2431. pii: 193/5/W397.
105. Suzuki K, Li F, Sone S, Doi K. Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. IEEE Trans Med Imaging. 2005;24:1138–50.
106. Suzuki K, Rockey DC, Dachman AH. CT colonography: advanced computer-aided detection scheme utilizing MTANNs for detection of "missed" polyps in a multicenter clinical trial. Med Phys. 2010;37:12–21.
107. Suzuki K, Zhang J, Xu J. Massive-training artificial neural network coupled with Laplacian-eigenfunction-based dimensionality reduction for computer-aided detection of polyps in CT colonography. IEEE Trans Med Imaging. 2010;29:1907–17. https://doi.org/10.1109/TMI.2010.2053213.
108. Xu J, Suzuki K. Massive-training support vector regression and Gaussian process for false-positive reduction in computer-aided detection of polyps in CT colonography. Med Phys. 2011;38:1888–902.
109. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3431–40.
110. Chen S, Suzuki K. Computerized detection of lung nodules by means of "virtual dual-energy" radiography. IEEE Trans Biomed Eng. 2013;60:369–78. https://doi.org/10.1109/TBME.2012.2226583.
111. Chen S, Suzuki K. Separation of bones from chest radiographs by means of anatomically specific multiple massive-training ANNs combined with total variation minimization smoothing. IEEE Trans Med Imaging. 2014;33:246–57. https://doi.org/10.1109/TMI.2013.2284016.
112. Chen S, Zhong S, Yao L, Shang Y, Suzuki K. Enhancement of chest radiographs obtained in the intensive care unit through bone suppression and consistent processing. Phys Med Biol. 2016;61:2283–301. https://doi.org/10.1088/0031-9155/61/6/2283.
113. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI international conference on medical image computing and computer-assisted intervention. New York: Springer; 2015. p. 234–41.

114. Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, Libby DM, Pasmantier MW, Koizumi J, Altorki NK, Smith JP. Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet. 1999;354:99–105.
115. Kaneko M, Eguchi K, Ohmatsu H, Kakinuma R, Naruke T, Suemasu K, Moriyama N. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. Radiology. 1996;201:798–802.
116. Miettinen OS, Henschke CI. CT screening for lung cancer: coping with nihilistic recommendations. Radiology. 2001;221:592–6.
117. Sone S, Takashima S, Li F, Yang Z, Honda T, Maruyama Y, Hasegawa M, Yamanda T, Kubo K, Hanamura K, Asakura K. Mass screening for lung cancer with mobile spiral computed tomography scanner. Lancet. 1998;351:1242–5.
118. Henschke CI, Yankelevitz DF, Naidich DP, McCauley DI, McGuinness G, Libby DM, Smith JP, Pasmantier MW, Miettinen OS. CT screening for lung cancer: suspiciousness of nodules according to size on baseline scans. Radiology. 2004;231:164–8.
119. Swensen SJ, Jett JR, Hartman TE, Midthun DE, Sloan JA, Sykes AM, Aughenbaugh GL, Clemens MA. Lung cancer screening with CT: Mayo Clinic experience. Radiology. 2003;226:756–61.
120. Heelan RT, Flehinger BJ, Melamed MR, Zaman MB, Perchick WB, Caravelli JF, Martini N. Non-small-cell lung cancer: results of the New York screening program. Radiology. 1984;151:289–93.
121. Gurney JW. Missed lung cancer at CT: imaging findings in nine patients. Radiology. 1996;199:117–22.
122. Giger ML, Bae KT, MacMahon H. Computerized detection of pulmonary nodules in computed tomography images. Invest Radiol. 1994;29:459–65.
123. Armato SG 3rd, Giger ML, Moran CJ, Blackburn JT, Doi K, MacMahon H. Computerized detection of pulmonary nodules on CT scans. Radiographics. 1999;19:1303–11.
124. Gurcan MN, Sahiner B, Petrick N, Chan HP, Kazerooni EA, Cascade PN, Hadjiiski L. Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system. Med Phys. 2002;29:2552–8.
125. Lee Y, Hara T, Fujita H, Itoh S, Ishigaki T. Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique. IEEE Trans Med Imaging. 2001;20:595–604.
126. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. Med Phys. 1999;26:2654–68.
127. Sahiner B, Chan HP, Hadjiiski L. Classifier performance prediction for computer-aided diagnosis using a limited dataset. Med Phys. 2008;35:1559–70.
128. Suzuki K, Doi K. How can a massive training artificial neural network (MTANN) be trained with a small number of cases in the distinction between nodules and vessels in thoracic CT? Acad Radiol. 2005;12:1333–41.
129. Farag AA, El-Baz A, Gimelfarb G, El-Ghar MA, Eldiasty T. Quantitative nodule detection in low dose chest CT scans: new template modeling and evaluation for CAD system design. Med Image Comput Comput Assist Interv. 2005;8:720–8.
130. Ge Z, Sahiner B, Chan HP, Hadjiiski LM, Cascade PN, Bogot N, Kazerooni EA, Wei J, Zhou C. Computer-aided detection of lung nodules: false positive reduction using a 3D gradient field method and 3D ellipsoid fitting. Med Phys. 2005;32:2443–54.
131. Matsumoto S, Kundel HL, Gee JC, Gefter WB, Hatabu H. Pulmonary nodule detection in CT images with quantized convergence index filter. Med Image Anal. 2006;10:343–52. https://doi.org/10.1016/j.media.2005.07.001.
132. Yuan R, Vos PM, Cooperberg PL. Computer-aided detection in screening CT for pulmonary nodules. AJR Am J Roentgenol. 2006;186:1280–7. https://doi.org/10.2214/AJR.04.1969.
133. Pu J, Zheng B, Leader JK, Wang XH, Gur D. An automated CT based lung nodule detection scheme using geometric analysis of signed distance field. Med Phys. 2008;35:3453–61.

134. Retico A, Delogu P, Fantacci ME, Gori I, Preite Martinez A. Lung nodule detection in low-dose and thin-slice computed tomography. Comput Biol Med. 2008;38:525–34. https://doi.org/10.1016/j.compbiomed.2008.02.001.
135. Golosio B, Masala GL, Piccioli A, Oliva P, Carpinelli M, Cataldo R, Cerello P, De Carlo F, Falaschi F, Fantacci ME, Gargano G, Kasae P, Torsello M. A novel multithreshold method for nodule detection in lung CT. Med Phys. 2009;36:3607–18.
136. Armato SG III, McLennan G, McNitt-Gray MF, Meyer CR, Yankelevitz D, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Reeves AP, Croft BY, Clarke LP. Lung image database consortium: developing a resource for the medical imaging research community. Radiology. 2004;232:739–48.
137. Murphy K, van Ginneken B, Schilham AM, de Hoop BJ, Gietema HA, Prokop M. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. Med Image Anal. 2009;13:757–70. https://doi.org/10.1016/j.media.2009.07.001.
138. Tan M, Deklerck R, Jansen B, Bister M, Cornelis J. A novel computer-aided lung nodule detection system for CT images. Med Phys. 2011;38:5630–45. https://doi.org/10.1118/1.3633941.
139. Messay T, Hardie RC, Rogers SK. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Med Image Anal. 2010;14:390–406. https://doi.org/10.1016/j.media.2010.02.004.
140. Riccardi A, Petkov TS, Ferri G, Masotti M, Campanini R. Computer-aided detection of lung nodules via 3D fast radial transform, scale space representation, and Zernike MIP classification. Med Phys. 2011;38:1962–71.
141. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A free-response approach to the measurement and characterization of radiographic-observer performance. J Appl Photogr Eng. 1978;4:166–71.
142. Tajbakhsh N, Suzuki K. Comparing two classes of end-to-end learning machines for lung nodule detection and classification: MTANNs vs. CNNs. Pattern Recognit. 2017;63:476–86.
143. Murphy GP, Lawrence W, Lenhard RE, American Cancer Society. American Cancer Society textbook of clinical oncology. 2nd ed. Atlanta: The Society; 1995.
144. Zhao H, Lo SC, Freedman M, Wang Y. Enhanced lung cancer detection in temporal subtraction chest radiography using directional edge filtering techniques. In: Proceedings of SPIE medical imaging: image processing, San Diego, vol. 4684; 2002
145. Giger ML, Ahn N, Doi K, MacMahon H, Metz CE. Computerized detection of pulmonary nodules in digital chest images: use of morphological filters in reducing false-positive detections. Med Phys. 1990;17:861–5.
146. Lo SC, Freedman MT, Lin JS, Mun SK. Automatic lung nodule detection using profile matching and back-propagation neural network techniques. J Digit Imaging. 1993;6:48–54.
147. Lo SC, Lou SL, Lin JS, Freedman MT, Chien MV, Mun SK. Artificial convolution neural network techniques and applications to lung nodule detection. IEEE Trans Med Imaging. 1995;14:711–8.
148. Penedo MG, Carreira MJ, Mosquera A, Cabello D. Computer-aided diagnosis: a neural-network-based approach to lung nodule detection. IEEE Trans Med Imaging. 1998;17:872–80.
149. Austin JH, Romney BM, Goldsmith LS. Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect. Radiology. 1992;182:115–22.
150. Shah PK, Austin JH, White CS, Patel P, Haramati LB, Pearson GD, Shiau MC, Berkmen YM. Missed non-small cell lung cancer: radiographic findings of potentially resectable lesions evident only in retrospect. Radiology. 2003;226:235–41.
151. Suzuki K, Abe H, Li F, Doi K. Suppression of the contrast of ribs in chest radiographs by means of massive training artificial neural network. Proc SPIE Med Imaing (SPIE MI). 2004;5370:1109.
152. Akansu AN, Haddad RA. Multiresolution signal decomposition. Boston: Academic Press; 1992.

153. Stephane GM. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell. 1989;11:674–93.
154. Zarshenas A, Liu J, Forti P, Suzuki K. Separation of bones from soft tissue in chest radiographs: anatomy-specific orientation-frequency-specific deep neural network convolution. Med Phys. 2019;46:2232–42. https://doi.org/10.1002/mp.13468.
155. Chen S, Suzuki K, MacMahon H. Development and evaluation of a computer-aided diagnostic scheme for lung nodule detection in chest radiographs by means of two-stage nodule enhancement with support vector classification. Med Phys. 2011;38:1844–58.
156. Wei J, Hagihara Y, Shimizu A, Kobatake H. Optimal image feature set for detecting lung nodules on chest X-ray images. In: Computer assisted radiology and surgery; 2002. p. 706–11.
157. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics, 2005. CA Cancer J Clin. 2005;55:10–30.
158. Dachman AH. Atlas of virtual colonoscopy. New York: Springer; 2003.
159. Winawer SJ, Fletcher RH, Miller L, Godlee F, Stolar MH, Mulrow CD, Woolf SH, Glick SN, Ganiats TG, Bond JH, Rosen L, Zapka JG, Olsen SJ, Giardiello FM, Sisk JE, Van Antwerp R, Brown-Davis C, Marciniak DA, Mayer RJ. Colorectal cancer screening: clinical guidelines and rationale. Gastroenterology. 1997;112:594–642.
160. Macari M, Bini EJ. CT colonography: where have we been and where are we going? Radiology. 2005;237:819–33.
161. Fletcher JG, Booya F, Johnson CD, Ahlquist D. CT colonography: unraveling the twists and turns. Curr Opin Gastroenterol. 2005;21:90–8.
162. Suzuki K, Dachman AH. Computer-aided diagnosis in CT colonography. In: Dachman AH, Laghi A, editors. Atlas of virtual colonoscopy. 2nd ed. New York: Springer; 2011. p. 163–82.
163. Yoshida H, Dachman AH. Computer-aided diagnosis for CT colonography. Semin Ultrasound CT MR. 2004;25:419–31.
164. Yoshida H, Dachman AH. CAD techniques, challenges, and controversies in computed tomographic colonography. Abdom Imaging. 2005;30:26–41.
165. Kiss G, Van Cleynenbreugel J, Thomeer M, Suetens P, Marchal G. Computer-aided diagnosis in virtual colonography via combination of surface normal and sphere fitting methods. Eur Radiol. 2002;12:77–81.
166. Paik DS, Beaulieu CF, Rubin GD, Acar B, Jeffrey RB Jr, Yee J, Dey J, Napel S. Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT. IEEE Trans Med Imaging. 2004;23:661–75.
167. Summers RM, Johnson CD, Pusanik LM, Malley JD, Youssef AM, Reed JE. Automated polyp detection at CT colonography: feasibility assessment in a human population. Radiology. 2001;219:51–9.
168. Summers RM, Yao J, Pickhardt PJ, Franaszek M, Bitter I, Brickman D, Krishna V, Choi JR. Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population. Gastroenterology. 2005;129:1832–44.
169. Yoshida H, Masutani Y, MacEneaney P, Rubin DT, Dachman AH. Computerized detection of colonic polyps at CT colonography on the basis of volumetric features: pilot study. Radiology. 2002;222:327–36.
170. Yoshida H, Nappi J, MacEneaney P, Rubin DT, Dachman AH. Computer-aided diagnosis scheme for detection of polyps at CT colonography. Radiographics. 2002;22:963–79.
171. Gokturk SB, Tomasi C, Acar B, Beaulieu CF, Paik DS, Jeffrey RB Jr, Yee J, Napel S. A statistical 3-D pattern processing method for computer-aided detection of polyps in CT colonography. IEEE Trans Med Imaging. 2001;20:1251–60.
172. Nappi J, Yoshida H. Automated detection of polyps with CT colonography: evaluation of volumetric features for reduction of false-positive findings. Acad Radiol. 2002;9:386–97.
173. Acar B, Beaulieu CF, Gokturk SB, Tomasi C, Paik DS, Jeffrey RB Jr, Yee J, Napel S. Edge displacement field-based classification for improved detection of polyps in CT colonography. IEEE Trans Med Imaging. 2002;21:1461–7.

174. Jerebko AK, Malley JD, Franaszek M, Summers RM. Multiple neural network classification scheme for detection of colonic polyps in CT colonography data sets. Acad Radiol. 2003;10:154–60.
175. Jerebko AK, Malley JD, Franaszek M, Summers RM. Support vector machines committee classification method for computer-aided polyp detection in CT colonography. Acad Radiol. 2005;12:479–86.
176. Wang Z, Liang Z, Li L, Li X, Li B, Anderson J, Harrington D. Reduction of false positives by internal features for polyp detection in CT-based virtual colonoscopy. Med Phys. 2005;32:3602–16.
177. Li J, Van Uitert R, Yao J, Petrick N, Franaszek M, Huang A, Summers RM. Wavelet method for CT colonography computer-aided polyp detection. Med Phys. 2008;35:3527–38.
178. Yao J, Li J, Summers RM. Employing topographical height map in colonic polyp measurement and false positive reduction. Pattern Recognit. 2009;42:1029–40. https://doi.org/10.1016/j.patcog.2008.09.034.
179. Doshi T, Rusinak D, Halvorsen RA, Rockey DC, Suzuki K, Dachman AH. CT colonography: false-negative interpretations. Radiology. 2007;244:165–73.
180. Rockey DC, Paulson E, Niedzwiecki D, Davis W, Bosworth HB, Sanders L, Yee J, Henderson J, Hatten P, Burdick S, Sanyal A, Rubin DT, Sterling M, Akerkar G, Bhutani MS, Binmoeller K, Garvie J, Bini EJ, McQuaid K, Foster WL, Thompson WM, Dachman A, Halvorsen R. Analysis of air contrast barium enema, computed tomographic colonography, and colonoscopy: prospective comparison. Lancet. 2005;365:305–11. https://doi.org/10.1016/S0140-6736(05)17784-8. pii: S0140673605177848.
181. Suzuki K, Wu J, Sheu I. Principal-component massive-training machine-learning regression for false-positive reduction in computer-aided detection of polyps in CT colonography. Lecture notes in computer science, machine learning in medical imaging (MLMI), vol. 6357. Beijing: Springer; 2010. p. 182–9.
182. Zhu H, Liang Z, Pickhardt PJ, Barish MA, You J, Fan Y, Lu H, Posniak EJ, Richards RJ, Cohen HL. Increasing computer-aided detection specificity by projection features for CT colonography. Med Phys. 2010;37:1468–81.
183. Wang S, Yao J, Petrick N, Summers RM. Combining statistical and geometric features for colonic polyp detection in CTC based on multiple kernel learning. Int J Comput Intell Appl. 2010;9:1–15. https://doi.org/10.1142/S1469026810002744.
184. Lostumbo A, Suzuki K, Dachman AH. Flat lesions in CT colonography. Abdom Imaging. 2010;35:578–83. https://doi.org/10.1007/s00261-009-9562-3.
185. Suzuki K, Sheu I, Kawaler E, Ferraro F, Rockey DC, Dachman AH. Computer-aided detection (CADe) of flat lesions in CT colonography (CTC) by means of a spinning-tangent technique. In: Program of RSNA; 2010. p. 319.

Classification of Malignant and Benign Tumors

10

Juan Wang, Issam El Naqa, and Yongyi Yang

10.1 Introduction

In recent years, there have been significant interests and efforts in the development of computerized methods for automatically classifying a tumor or lesion being malignant or benign. These methods are collectively known as computer-aided diagnosis (CADx), the purpose of which is to provide a second opinion to assist radiologists in their diagnosis of the detected tumors. Indeed, in the literature, CADx techniques have been studied both for various disease types and for different imaging modalities, ranging from CT in oncology, magnetic resonance imaging (MRI) for brain tumors, to mammography and ultrasound for breast cancer and many others. For instance, the application of CT to early lung cancer has generated significant interests. In a recent randomized clinical trial referred to as the NELSON trial with 15,822 enrolled participants [1], it was shown that low-dose CT screening can improve the sensitivity and specificity of lung cancer detection [2]. However, this situation has been more challenging in cases of head and neck cancer, where the combination with positron emission tomography (PET) has overcome some shortages of CT and revolutionized the management of this cancer [3]. On the other hand, MRI, which is more financially expensive but with better soft tissue discrimination and avoiding exposure to ionizing radiation, has risen in recent years in the diagnosis of difficult cases such as prostate [4], brain [5, 6], and breast cancers [7].

J. Wang · Y. Yang (✉)
Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, USA
e-mail: yangyo@iit.edu

I. El Naqa
Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: ielnaqa@med.umich.edu

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_10

In mammography, many CADx techniques have been developed for the classification of suspicious breast tumors in mammogram images, including both masses and clustered microcalcifications (MCs). For example, in the early work [8], a three-layer, feed-forward neural network was trained with a back-propagation algorithm for mammographic lesion (including MCs and masses) interpretation. Subsequently, various supervised learning techniques were studied for diagnosis of MC lesions (e.g., [9–15]) and mass lesions (e.g., [16–20]). There also exist several laboratory studies which demonstrate that CADx techniques can either be more accurate than the human readers or help improve their diagnosis accuracy [11, 21–26].

In the rest of this chapter, we will first provide an overview of the major components involved in the development of a CADx framework for tumor classification (Sect. 10.2). Afterwards, we will illustrate this framework with some examples of CADx techniques for breast lesions in mammograms (Sect. 10.3). In addition, we will also introduce the use of a visualization tool—based on the technique of multi-dimensional scaling (MDS)—for exploring the similarity among a set of tumors (Sect. 10.4). Such a tool potentially can be useful for one to compare a case under consideration against some similar, known cases in a reference library. We will also discuss some issues and challenges in the development and application of CADx techniques (Sect. 10.5).

10.2 Overview of Classification Framework

When in operation, a CADx framework for tumor classification functions as follows: For a given tumor under consideration, a set of image features is first computed from the tumor to quantify its underlying characteristics. These features are mathematically represented by a vector $\mathbf{x}$ in an n-dimensional space R^n. Afterward, a mathematical function $f(\mathbf{x})$ is applied to the feature vector $\mathbf{x}$, the value of which is used to reflect the likelihood that the tumor is either malignant or benign. The function $f(\mathbf{x})$ is called the decision function or the classifier function.

The development of a CADx framework involves the following key components: (1) determine what features $\mathbf{x}$ to use that are relevant for classification of the tumor, (2) design the classifier function $f(\mathbf{x})$ that is appropriate for the task, and (3) evaluate the accuracy level (i.e., performance metric) of the classifier output, which is key to the confidence level on the "second opinion." It is noted that with the development of deep neural networks in recent years, the first two components above may be accomplished in a so-called "end-to-end" fashion within a single representative learning framework using deep neural networks, for instance.

10.2.1 Perception Modeling

There have been significant improvements with respect to developing image quantitative imaging measures, objective image interpretations, feature extraction, and semantic descriptors over the past decades [27, 28]. However, some major

difficulties still remain pertaining to CADx applications. First, it is understood that quantitative measures can vary with the different aspects of perceptual similarity of images by radiologists; the selection of an appropriate similarity measure thus becomes problem-dependent. Second, the relation between the low-level visual features and the high-level expert human interpretation of similarity is not well defined when comparing two images; it is thus not exactly clear what features or combination of them are relevant for such judgment [29, 30]. We have been developing perceptual similarity metrics for application in content-based image retrieval (CBIR) of mammogram images [31]. In this approach, the notion of similarity is modeled as a nonlinear function of the image space (features) in a pair of mammogram images containing lesions of interest, e.g., microcalcification clusters (MCCs). If we let vectors **u** and **v** denote the features of two MCCs at issue, the following regression model is used to determine their similarity coefficient (*SC*):

$$SC(\mathbf{u},\mathbf{v}) = f(\mathbf{u},\mathbf{v}) + \zeta, \tag{10.1}$$

where $f(\mathbf{u}, \mathbf{v})$ is a function determined using a machine learning approach, which we choose to be support vector machine (SVM) learning [32], and ζ is the modeling error. The similarity function $f(\mathbf{u}, \mathbf{v})$ in Eq. (10.1) is trained using data samples collected in an observer study.

10.2.2 Feature Extraction for Tumor Quantification

The purpose of feature extraction is to describe the content of a tumor under consideration by a set of quantitative descriptors, called features, denoted by vector **x**. Conceptually, these features should be relevant to the disease condition of the tumor. For example, they may be used to quantify the size of the tumor, the geometric shape of the tumor, the density of the tissue, etc., depending on the tumor type and specific application.

In the literature, there have been many types of features studied for classification of benign and malignant tumors. For example, in [33], effective thickness and effective volume were defined on the physical properties of MCs in mammogram images and were demonstrated to be useful for diagnosis. In [34], image intensity and texture features were extracted from post-contrast T1-weighted MR images and were shown to be helpful for brain tumor classification. In [35], wavelet features were compared with Haralick features [36] for MC classification.

While the reported features are many, they can be divided into two broad categories: (1) boundary-based features, and (2) region-based features. Boundary-based features are used to describe the properties of the geometric boundary of a tumor. They include, for example, the perimeter, Fourier descriptors, and boundary moments [37]. In contrast, region-based features are derived from within a tumor region, which include the shape, texture, or the frequency domain information of the tumor. Some examples of region-based features are the tumor size, image moment features [37], wavelet-based features [35], and texture features [34].

To ensure good classification performance, the features extracted from a tumor are desired to have certain properties pertinent to the application. For example, a common requirement is that the features should be shift invariant to any translation or rotation in a tumor image. Other considerations in extracting or designing quantitative features include the effects of the image resolution and gray-level quantization used for the image. The image resolution can affect those features related to the size of a tumor, such as its area and perimeter. The quantization level in an image can affect those features related to the image intensity, such as image moments and features derived from the gray-level co-occurrence matrix (GLCM) [34]. Therefore, prior to feature extraction, the tumor images need to be preprocessed properly in order to avoid any discrepancy in resolution and quantization.

With a great number of features available, as described above, an important task in a CADx framework is how to determine a set of discriminative features in a tumor classification problem. These features are desired to have good differentiating power between benign and malignant tumors. One approach is to exploit the working knowledge of the clinicians and select those features that are closely associated with what the clinicians use in their diagnosis of the lesions [11]. For example, for MC lesions, the size and shape of the MCs and their spatial distribution are all known to be important, because the MCs tend to be more irregular and have a bigger cluster in a malignant lesion [15]. Alternatively, to determine the most salient features for use in the classification, one may employ a systematic feature selection procedure during the training stage of the classifier. The commonly used feature selection procedures in the literature include the filter algorithm [38], wrapper algorithm [39], and embedded algorithm [40].

10.2.3 Design of Decision Function Using Machine Learning

The problem of classifying benign or malignant tumors is a classical two-class classification problem, with benign tumors being one class and malignant ones being the other class. For a given tumor characterized by its feature vector $\mathbf{x}$, a decision function $f(\mathbf{x})$ is designed to determine which class (malignant or benign), $\mathbf{x}$ belongs to. Naturally, a fundamental problem is how to design the decision function for a given tumor type. A common approach to this problem is to apply supervised learning, in which a pattern classifier is first trained on a set of known cases, denoted as $\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$, where a training sample is described by its feature vector $\mathbf{x}_i$, and y_i is its known class-label (1 for malignant tumor and -1 for benign tumor). Once trained, the classifier is applied subsequently to classify other cases (unseen during training).

Broadly speaking, depending on its mathematical form, the decision function $f(\mathbf{x})$ is categorized into linear and nonlinear classifiers. A linear classifier is represented as

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \tag{10.2}$$

where **w** is the discriminant vector and b is the bias, which are parameters determined from the training samples. In contrast, a nonlinear classifier $f(\mathbf{x})$ has a more complex mathematical form and is no longer a linear function in terms of the feature variables **x**. One such example is the feed-forward neural network, in which (nonlinear) sigmoid activation functions are used at the individual nodes within the network.

Because of their simpler form, linear classifiers are easier to train and less prone to over-fitting compared to their nonlinear counterpart. Moreover, it is often easier to examine and interpret the relationship between the classifier output and the individual feature variables in a linear classifier than that in a nonlinear one. Thus, linear classifiers can be favored for certain applications. On the other hand, because of their more complex form, nonlinear classifiers can be more versatile and achieve better performance than linear ones when the underlying decision surface between the two classes is inherently nonlinear in a given problem.

Regardless of their specific form, classifier functions typically involve a number of parameters, which need to be determined before they can be applied to classifying an unknown case. There have been many different algorithms designed for determining these parameters from a set of training samples, which are collectively known as supervised machine learning algorithms. These are discussed in detail in Chap. 3 and reviewed here briefly.

Consider, for example, the case of linear classifiers in Eq. 10.2. The parameters **w** and b can be determined according to the following different optimum principles: (1) logistic regression [41], in which the log-likelihood function of the training data samples is maximized under a logistic probability model; (2) linear discriminant analysis (LDA) [42], in which the optimal decision boundary is determined under the assumption of multivariate Gaussian distributions for the data samples from the two classes; and (3) support vector machine (SVM) [43], in which the parameters are designed to achieve the maximum separation between the two classes (among the training samples).

Similarly, there also exist many methods for designing nonlinear classifiers. One popular type of nonlinear classifiers is the kernel-based methods [44]. In a kernel-based method, a so-called kernel trick is used to first map the input vector **x** into a higher-dimensional space via a nonlinear mapping; afterward, a linear classifier is applied in this mapped space, which in the end is a nonlinear classifier in the original feature space. One such example is the popular nonlinear SVM classifier. Other kernel-based methods include kernel Fisher discriminant (KFD), kernel principle component analysis (KPCA), and relevance vector machine (RVM) [45].

Another type of commonly used nonlinear CADx classifiers is the committee-based methods. These methods are based on the idea of systematically aggregating the output of a series of individual weak classifiers to form a (more powerful) decision function. Adaboost [46] and random forests [47] are well-known examples of such committee-based methods. For example, in Adaboost, the training set is applied successively to obtain a sequence of weak classifiers; the output of each weak classifier is adjusted by a weight factor according to its classification error on the training set to form an aggregated decision function [46] while random forests

using a bagging (averaging) approach to develop the committee classifier [47]. More recently, gradient boosting methods seemed to provide superior performance to other methods in this category [48].

10.2.4 Deep Learning Methods

Deep learning is a special family of machine learning methods generally based on artificial neural networks. A popular deep learning architecture in medical imaging is deep convolutional neural network (CNN). Details on deep learning can be found in Chap. 4. For example, in the context of CADx, in [13], it was used to discriminate malignant and benign MC lesions in mammography, and in [16], it was applied for mass lesion diagnosis. By design, a CNN is typically comprised of a cascade of multiple convolutional, batch normalization, pooling, fully connected layers, and other layers, which are described as follows.

Convolutional layers are used to extract the image features at varying spatial scales in an input image. Within a convolutional layer, a set of filters is used to operate on the input, from which the output is fed into a subsequent layer. The output of each filter is called a *feature map*. Specifically, for a given layer, let $\boldsymbol{x}_k$ denote its k-th input feature map, and $\boldsymbol{h}_j^k$ denote its corresponding convolutional filter for output feature map j. Then, the output can be represented as

$$\boldsymbol{y}_j = f\left(\sum_{k=1}^{K} \boldsymbol{x}_k * \boldsymbol{h}_j^k + b\right), \tag{10.3}$$

where $*$ denotes the convolution operation, K is the number of channels in the input feature map, b is a bias constant, and $f(\cdot)$ is an activation function, which is a nonlinear transformation from the input to the output. The recent development of deep learning is partly facilitated by the design of activation functions to address the problem of saturation associated with the traditional sigmoid function. Some of the commonly used activation functions are as follows:

1. Rectified linear unit (ReLU)

$$f(x) = \begin{cases} 0 & \text{for } x \le 0 \\ x & \text{for } x > 0 \end{cases} \tag{10.4}$$

2. Leaky rectified linear unit (Leaky ReLU)

$$f(x) = \begin{cases} 0.01x & \text{for } x \le 0 \\ x & \text{for } x > 0 \end{cases} \tag{10.5}$$

3. Parametric rectified linear unit (PReLU)

$$f(x) = \begin{cases} \alpha x & \text{for } x \le 0 \\ x & \text{for } x > 0 \end{cases} \tag{10.6}$$

Batch normalization layers are used to deal with the issue of internal covariate shifts during training, a phenomenon that the distribution of a layer's input varies with the change in parameters of its preceding layers. Batch normalization has been shown to speed up learning and improve classification performance. It is achieved through a normalization step which controls the mean and variance values of each layer's input. For each feature in the feature maps z-score normalization is applied such that each resulting feature has zero mean and unit standard deviation. Mathematically, let x_k^i be the k-th feature of the i-th training sample in a mini-batch, then the normalized output can be represented as

$$y_k^i = \gamma_k Z\left(x_k^i\right) + \beta_k \tag{10.7}$$

where γ_k and β_k are parameters to restore the representation power of the network, which are learned during training, and $Z\left(x_k^i\right)$ denotes the z-score normalization of k-th feature in the feature map, in which the mean and standard deviation are estimated from all the samples in the mini-batch during training.

Pooling layers are used to summarize the features of neighboring regions in a feature map by reducing the spatial dimensions of the data. A pooling operation can be done either locally or globally. A commonly used local pooling is to combine data in small regions of 2 × 2 or 3 × 3. Global pooling acts on the entire spatial dimension of a feature map. During pooling, the max or average of a region can be computed. Max pooling calculates the maximum of its input features, and average pooling uses the average of its input features.

Fully connected layers are used to connect each neuron from one layer to every neuron in another layer, just as in a traditional feed-forward neural network. Fully connected layers are usually used as the last layers in a deep neural network to achieve the classification task.

While deep learning methods have found great success in many applications, a major challenge is the need for acquiring a large number of samples for training. Transfer learning has been employed to deal with this problem. Transfer learning is a technique of creating high-performance classifiers with data more easily available from different application domains [49]. For example, in [50], it was used to classify pulmonary nodules of thoracic CT images with the classic LeNet-5 model. In [51], it was used for brain tumor classification with the GoogLeNet.

Another approach to deal with the issue of insufficient data samples in training is to combine both deep learning and traditional machine learning techniques for classification. It typically uses either pretrained CNN or unsupervised deep learning models for feature extraction and then applies machine learning methods for classification. For example, in [52], a pretrained CNN was used to extract different levels of features (along with handcrafted features) for classification by SVM. In

[53], a variational autoencoder was employed for feature learning. In [54], a CNN pretrained on the ImageNet dataset was used for feature extraction and an SVM was used to predict the likelihood of a breast lesion detected on mammograms being malignant.

10.2.5 CADx Classifier Training and Performance Evaluation

Conceptually, a CADx classifier should be trained and evaluated by using the following three sets of data samples: a training set, a validation set, and a testing set. The training set is used to obtain the model parameters of a classifier (such as $\mathbf{w}$ and b in the linear classifier in Eq. 10.2). The validation set is usually independent from the training set and is used to determine the tuning parameters of a classifier if it has any. For example, in kernel SVM, one may need to decide the type of the kernel function to use. Finally, the testing set is used to evaluate the performance of the resulting classifier. It must be independent from both the training and validation sets in order to avoid any potential bias.

Ideally, when the number of available data samples is large enough, the training, validation, and testing sets in the above should be kept to be mutually exclusive. However, in practice, the data samples are often scarce, making it impossible to obtain independent training, validation, and testing sets, which is often true when clinical cases are used. To deal with this difficulty, a k-fold cross-validation procedure is often used instead. The procedure works as follows: first, the available n data samples are divided randomly into k roughly equal-sized subsets; subsequently, each of the k subsets is held out in turn for testing while the rest $(k-1)$ subsets are used together for training. In the end, the performance is averaged over the k held-out testing subsets to obtain the overall performance. A special case of the k-fold cross-validation procedure is when $k = n$, which is also called a leave-one-out procedure (LOO). It is known that a smaller k yields a lower variance but also a larger bias in the estimated performance. In practice, $k = 5$ or 10 is often used as a good compromise in cross validation [55, 56].

When there are (hyper)-parameters needed to be tuned in a classifier model, a double loop (nested) cross-validation procedure [57] can be applied to avoid any potential bias. A double-loop cross-validation procedure has a nested structure of two loops (the inner and outer loops). The outer loop is the same as the standard k-fold cross validation above, which is used to evaluate the performance of the classifier. The inner loop is to further perform a standard k'-fold cross-validation using only the training portion of samples in each iteration of the outer loop, which is used to select the tuning parameters.

For evaluating the performance of a CADx classifier, a receiver-operating characteristic (ROC) analysis is now routinely used. An ROC curve is a plot of the classification sensitivity (i.e., true-positive fraction) as the ordinate versus the specificity (i.e., false-positive fraction) as the abscissa. For a given classifier, an ROC curve is obtained by continuously varying the threshold associated with its decision function over its operating range. As a summary measure of overall diagnostic performance,

the area under an ROC curve (denoted by AUC) is often used. A larger AUC value means better classification performance.

10.3 Application Examples in Mammography

10.3.1 Mammography

Mammography is an imaging procedure in which low-energy X-ray images of the breast are taken. Typically, they are in the order of 0.7 mSv. A mammogram can detect a cancerous or precancerous tumor in the breast even before the tumor is large enough to feel. Despite advances in imaging technology, mammography remains the most cost-effective strategy for early detection of breast cancer in clinical practice. The sensitivity of mammography could be up to approximately 90% for patients without symptoms [58]. However, this sensitivity is highly dependent on the patient's age, the size and conspicuity of the lesion, the hormone status of the tumor, the density of a woman's breasts, and the overall image quality and the interpretative skills of the radiologist [59]. Therefore, the overall sensitivity of mammography could vary from 90% to 70% only [60]. Moreover, it is very difficult to distinguish mammographically benign lesions from malignant ones. It has been estimated that one third of regularly screened women experience at least one false-positive (benign lesions being biopsied) screening mammogram over a period of 10 years [61]. A population-based study included about 27,394 screening mammograms that were interpreted by 1067 radiologists showed that the radiologists had substantial variations in the false-positive rates ranging from 1.5 to 24.1% [62]. Unnecessary biopsy is often cited as one of the "risks" of screening mammography. Surgical, needle-core, and fine-needle aspiration biopsies are expensive, invasive, and traumatic for the patient.

10.3.2 Detection of Clustered Microcalcifications in Mammograms

Clustered microcalcifications (MCs) can be an important early sign of breast cancer in women. They are found in 30–50% of mammographically diagnosed cases. MCs are calcium deposits of very small dimension and appear as a group of granular bright spots in a mammogram (e.g., Fig. 10.1). While often seen, accurate detection MCs in mammograms can be difficult, because of their subtlety in appearance, variation in shape and size, and inhomogeneity in surrounding tissue. In computer-aided diagnosis, accurately detecting the individual MCs in a cluster is important, because the image features of the detected MCs are further analyzed for classification as being benign or malignant [9, 63]. Studies have shown that the accuracy of detected individual MCs can impact on the CADx performance [24, 25, 64, 65].

In the literature, computerized methods have been investigated for accurate detection of clustered MCs. For example, in [66], a difference-of-Gaussians (DoG)

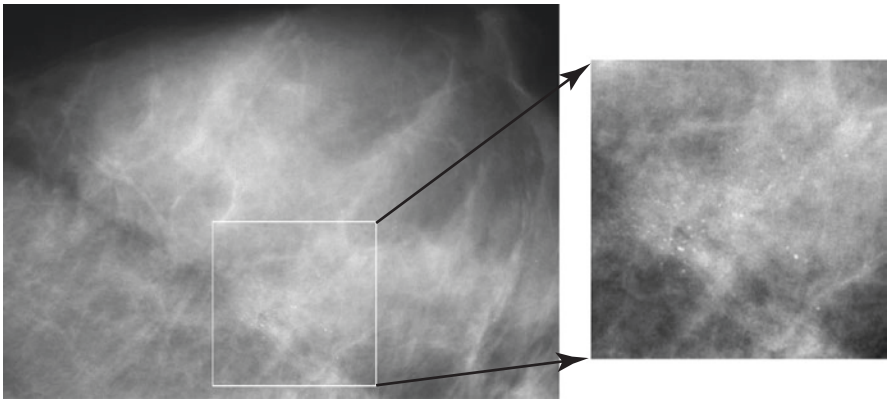


Fig. 10.1 A mammogram image (left) and its magnified view (right), where MCs are visible as granular bright spots

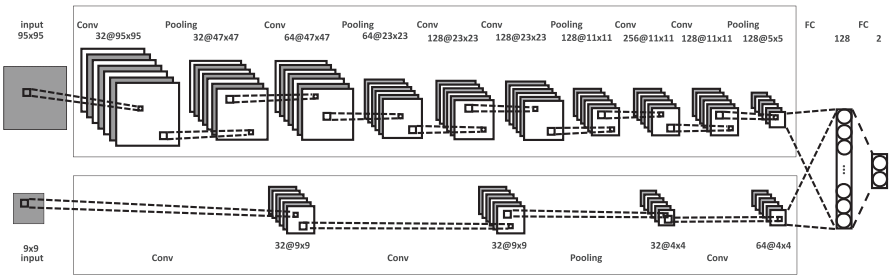


Fig. 10.2 Illustration of a context-sensitive DNN classifier architecture. It consists of two subnetworks, one for processing the large image context window (called global subnetwork) and one for processing the small MC image window (called local subnetwork). A batch normalization layer and a nonlinearity layer are included after each Convolutional (Conv) layer, which are not shown in the figure for brevity

filter, wherein the filter consisted of two kernels of limited width parameters, was applied for MC detection. In [67], an SVM was adopted using a local image window centered at a location as input. In [68], two CNNs were considered for MC detection in multi-vendor mammography, in which one CNN was used to remove easy samples and another was used for classifying the survived samples; the input sample consisted of a local window at a detection location. In Wang et al. [69], a CNN was developed to detect the presence of clustered MCs in mammograms.

MC detection example using deep learning: In this section we demonstrate the use of a deep neural network (DNN) for MC detection [70]. We formulate MC detection as a two-class classification problem, wherein a classifier is employed to

determine whether an MC object is present (class 1) or absent (class 0) at a location under consideration in a mammogram image. The DNN architecture is shown in Fig. 10.2. It takes into account not only the local image features of an MC but also its surrounding image context for MC detection. Specifically, the detector network is formed by two subnetworks, one for extracting the local image features and one for learning the image features of its surrounding background. The extracted features by the two subnetworks are combined subsequently for classifying whether an MC is present or not at a detection location. Consequently, the detector response is automatically adapted to the image background at an MC; the proposed detector is termed as context-sensitive DNN accordingly.

To evaluate the context-sensitive DNN detector, 300 mammograms were used for training, 117 for validation, and 125 for testing. Two image ROIs (500×500 or 1000×1000 pixels) were cropped from each testing image for performance evaluation, one containing clustered MCs and one without any MCs. To summarize the detection performance, a free-response receiver-operating characteristic (FROC) analysis was conducted. An FROC curve is a plot of the true-positive (TP) fraction of the MCs detected versus the average number of FPs per unit image region (1 cm^2 in area) with the decision threshold varied over an operating range. In the FROC analysis, the TP fraction was computed from the average of the TP fractions of the ROIs with clustered MCs, whereas the FP rate was computed from both the ROIs with and without any MCs. In the detector output, a detected object was treated as a TP when at least 40% of its area overlaps with that of a true MC or its distance to the center of a true MC is not larger than 0.3 mm; otherwise it was counted as an FP.

Figure 10.3 shows the FROC curves obtained by the context-sensitive DNN classifier [70], unified SVM detector [71], and a local DNN (with the local network only). The FROC curve of the context-sensitive DNN classifier is notably higher

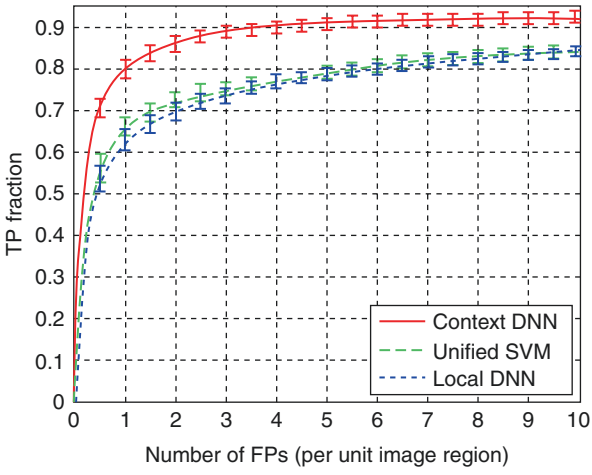


Fig. 10.3 FROC curves obtained by different classifiers in detecting individual MCs: (1) context-sensitive DNN (Context DNN) [70], (2) unified SVM [71], and (3) local DNN

(hence better detection performance) than those of both unified SVM and local DNN. In particular, with TPF at 80%, the context-sensitive DNN classifier achieved an FP rate of 1.03 FPs/cm^2, compared to 5.69 FPs/cm^2 by the unified SVM (a reduction of 81.9%) and 6.00 FPs/cm^2 by the local DNN (a reduction of 82.8%).

10.3.3 Computer-Aided Diagnosis (CADx) of Microcalcification Lesions in Mammograms

Because of the subtlety of microcalcifications in appearance in mammogram images, accurate diagnosis of MC lesions as benign or malignant is a very challenging problem for radiologists. Studies show that a false-positive diagnostic imaging study leads to unnecessary biopsy of benign lesions, yielding a positive predictive value of only 20–40% [72]. There has been intensive research in the development of CADx techniques for clustered MCs, of which the purpose is to provide a second opinion to radiologists in their diagnosis to improve the performance and efficiency [21]. In the literature, various machine learning methods such as LDA, logistic regression, ANN, and SVM have been used in the development of CADx classifiers for clustered MCs. For example, in [73], an LDA classifier was used for classification of benign and malignant MCs based on their visibility and shape features. This approach was subsequently extended to morphology and texture features in [21]. In [74], it was demonstrated that an ANN-based approach could improve the diagnosis performance of radiologists for MCs. In [15], FKD, ANN, SVM, RVM, and committee machines were explored in a comparison study, wherein the SVM was shown to yield improved performance over the others. Collectively, the reported research results demonstrate that CADx has the potential to improve the radiologists' performance in breast cancer diagnosis [75].

In the development of CADx techniques in the literature, various types of features have been investigated for characterizing MC lesions [10, 19, 65, 76–78]. These features are defined to characterize the gray-level properties (e.g., the brightness, contrast and gradient of individual MCs, the texture in the lesion region), or geometric properties of the MC lesions (e.g., the size and shape of the individual MCs, the number of MCs, the area, shape, and spatial distribution of a cluster). They are extracted either from the individual MCs or the entire lesion region. The features from individual MCs are often summarized using statistics to characterize an MC cluster.

CADx example: Machine learning methods for MC classification. In this section, we demonstrate the use of two CADx classifiers for clustered MCs, one is a linear classifier based on logistic regression, and the other is a nonlinear SVM classifier with a RBF kernel [15]. In logistic regression, the parameters $\mathbf{w}$ and b in (10.2) are determined through maximization of the following log-likelihood function:

$$L(\mathbf{w},b)=\sum_{i=1}^{N}\log p\left(y_i,\mathbf{x}_i;\mathbf{w},b\right) \tag{10.8}$$

where the probability term is given by

$$p\left(y_i = 1, \mathbf{x}_i; \mathbf{w}, b\right) = \left[1 + \exp\left(-\mathbf{w}^T \mathbf{x}_i - b\right)\right]^{-1}. \tag{10.9}$$

For the nonlinear SVM classifier, it can be represented as:

$$f\left(\mathbf{x}\right) = \mathbf{w}^T \Phi\left(\mathbf{x}\right) + b \tag{10.10}$$

where $\mathbf{w}$ is the discriminant vector, b is the bias, and $\Phi(\mathbf{x})$ is a nonlinear mapping function which is implicitly defined by a kernel function (RBF in our case).

Based on the maximum marginal criterion, the parameters $\mathbf{w}$ and b in Eq. (10.10) are determined as follows:

$$\begin{aligned} &\min J\left(\mathbf{w}, \xi\right) = \frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{i=1}^{N} \xi_i \\ &\text{s.t. } y_i f\left(\mathbf{x}_i\right) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, 2, \cdots, N \end{aligned} \tag{10.11}$$

For testing these classifiers, we used a dataset of 104 cases (46 malignant, 58 benign), all containing clustered MCs. This dataset was collected at the University of Chicago. It contains some cases that are difficult to classify; the average classification performance by a group of five attending radiologists on this dataset yielded a value of only 0.62 in the area under the ROC curve [11]. The MCs in these mammograms were marked by a group of expert readers.

For this dataset, a set of eight features were extracted to characterize MC clusters [11]: (1) the number of MCs in the cluster, (2) the mean effective volume (area times effective thickness) of individual MCs, (3) the area of the cluster, (4) the circularity of the cluster, (5) the relative standard deviation of the effective thickness, (6) the relative standard deviation of the effective volume, (7) the mean area of MCs, and (8) the second highest shape-irregularity measure. These features were selected such that they have meanings that are closely associated with features used by radiologists in clinical diagnosis of MC lesions.

To evaluate the classifiers, a leave-one-out (LOO) procedure was applied to the 104 cases, and the ROCKIT software was used to calculate the performance AUC. The logistic regression classifier achieved AUC = 0.7174. In contrast, the SVM achieved AUC = 0.7373. These results indicate that the classification performance of the classifiers is far from being perfect, which illustrates the difficulty in diagnosis of MC lesions in mammograms.

10.3.4 Adaptive CADx Boosted with Content-Based Image Retrieval (CBIR)

In recent years, CBIR has been studied as a diagnostic aid in tumor classification [79, 80], of which the goal is to provide radiologists with examples of lesions with known pathology that are similar to the lesion being evaluated. A CBIR system can be viewed as a CADx tool to provide evidence for case-based reasoning. With

CBIR, the system first retrieves a set of cases similar to a query, which can be used to assist a decision for the query [81]. For example, in [82] and [83], the ratio of malignant cases among all retrieved cases was used as a prediction for the query. In [84], the similarity levels between the query and retrieval cases were used as weighting factors for prediction.

We have been investigating an approach of using retrieved images to boost the classification of a CADx classifier [85–87]. In conventional CADx, a pattern classifier was first trained on a set of training cases, and then applied to subsequent testing cases. Deviating from approach, for a given case to be classified (i.e., query), we first obtain a set of known cases with similar features to that of the query case from a reference database and use these retrieved cases to adapt the CADx classifier so as to improve its classification accuracy on the query case. Below we illustrate this approach using a linear classifier with logistic regression [85].

Assume that a baseline classifier $f(\mathbf{x})$ in the form of Eq. 10.2 has been trained with logistic regression as in Eq. 10.8 on a set of training samples: $\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$. Now, consider a query lesion $\mathbf{x}$ to be classified. Let $\left\{\left(\mathbf{x}_i^{(r)}, y_i^{(r)}\right), i = 1, \ldots, N_r\right\}$ be a set of N_r retrieved cases which are similar to $\mathbf{x}$. In our case-adaptive approach, we use the retrieved samples $\left\{\left(\mathbf{x}_i^{(r)}, y_i^{(r)}\right), i = 1, \ldots, N_r\right\}$ to adapt the classifier $f(\mathbf{x})$. Specifically, the objective function in (10.8) is modified as

$$L(\mathbf{w}, b) = \sum_{i=1}^{N} \log p\left(y_i, \mathbf{x}_i; \mathbf{w}, b\right) + \sum_{i=1}^{N_r} \beta_i \log p\left(y_i^{(r)}, \mathbf{x}_i^{(r)}; \mathbf{w}, b\right) \tag{10.12}$$

In Eq. 10.12, the weighting factors β_i are adjusted according to the similarity of $\mathbf{x}_i^{(r)}$ to the query $\mathbf{x}$. The idea is to put more emphasis on those retrieved samples that are more similar to the query, with the goal of refining the decision boundary of the classifier in the neighborhood of the query. Indeed, the first term in Eq. 10.12 simply corresponds to the log-likelihood function in Eq. 10.8, while the second term can be viewed as a weighted likelihood of those retrieved similar samples. Intuitively, the retrieved samples are used to steer the pretrained classifier from Eq. 10.8 to achieve more emphasis in the neighborhood of the query $\mathbf{x}$. Note that the objective function in Eq. 10.12 has the same mathematical form as that in the original optimization problem in Eq. 10.8, which can be solved efficiently by the method of iteratively reweighted least square (IRLS) [41].

In our study, we implemented the following strategy for adjusting β_i according to the similarity level of a retrieved sample $\mathbf{x}_i^{(r)}$ to the query $\mathbf{x}$:

$$\beta_i = 1 + k \frac{\alpha_i}{\max\limits_{j=1,\ldots,N_r} \{\alpha_j\}}, \; i = 1, \ldots, N_r \tag{10.13}$$

where α_i denotes the similarity measure between $\mathbf{x}_i^{(r)}$ and $\mathbf{x}$, and $k > 0$ is a parameter used to control the degree of emphasis on the retrieved samples relative to other training samples. The choice of the form in Eq. (10.13) is such that the weighting factor increases linearly with the similarity level of a retrieved case to $\mathbf{x}$, with the most similar case among the retrieved receiving maximum weight $1 + k$, which

corresponds to k times more influence than the existing training samples in the objective function in Eq. (10.13).

As a similarity measure for retrieved cases, we used the Gaussian RBF kernel function:

$$\alpha_i = \exp\left(-\left\| \mathbf{x}_i^{(r)} - \mathbf{x} \right\|^2 / \gamma^2 \right),\ i = 1,\ldots,N_r \tag{10.14}$$

where γ is a scaling factor controlling the sensitivity of α_i with respect to the distance between the query and a retrieved case. In our experiments, the parameter γ was set to the tenth percentile of the distance between every possible image pairs in the training set. Such a choice is out of the consideration that most of the cases in a database are typically not similar to each other. Those cases with a large distance away from query $\mathbf{x}$ will receive a low similarity measure consequently.

To demonstrate this approach, a set of 589 cases (331 benign, 258 malignant), all containing MC lesions, were extracted from the benign and cancer volumes in the DDSM database maintained at the University of South Florida [88]. The extracted mammogram images were adjusted to correspond to the same optical density and to have a uniform resolution of 0.05 mm/pixel. To quantify the MC lesions in these mammogram images, we first applied an MC detection algorithm using an SVM classifier [31, 89] to automatically locate the MCs in each lesion region provided by the dataset. To help suppress the false-positives in the detection, the images were first processed with the isotropic normalization technique prior to the detection [90]. The detected MCs were grouped into clusters.

Afterward, a set of descriptive features were computed for the clustered MCs in the dataset; the following nine features were used [85]: (1) area of the cluster, (2) compactness of the cluster, (3) density of the cluster represented by the number of MCs in a unit area, (4) standard deviation of the inter-distance between neighboring MCs, (5) number of MCs in the cluster, (6) sum of the size of all MC objects in the cluster, (7) mean of the average brightness in each MC object, (8) mean of the intensity standard deviation in each MC object, and (9) the compactness of the second most irregular MC object in the cluster. These features were used to form a vector $\mathbf{x}$ for each lesion in the dataset.

To evaluate the classification performance, a subset of 120 cases (70 benign, 50 malignant) was randomly selected from the dataset for training the baseline classifier, and the remaining 469 cases were used for testing the adaptive classifier. An LOO procedure was applied for each testing case, for which all the remaining cases were used for retrieval. In Fig. 10.4, we show the performance results achieved by the case-adaptive classifier and the baseline classifier; for the adaptive classifier, the AUC value is shown with different number of retrieved cases N_r. From Fig. 10.4, it can be seen that the best performance (AUC = 0.7755) was obtained by the adaptive classifier when N_r = 20, compared to AUC = 0.6848 for the baseline classifier (p-value<0.0001). The performance is also noted to deteriorate somewhat with increased N_r. This is because the number of similar cases for a given query is typically small due to the limited number of cases in the reference library. With large N_r,

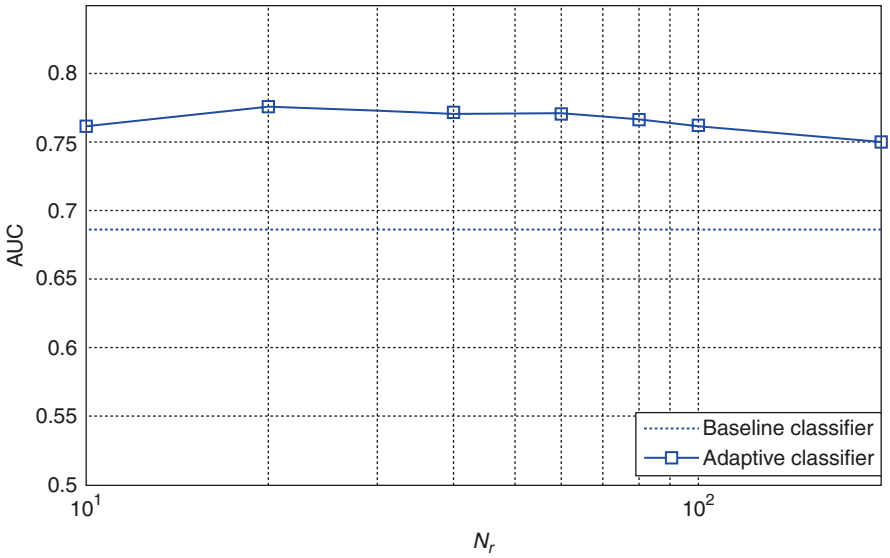


Fig. 10.4 Classification performance (AUC) achieved by the case-adaptive linear classifier. The number of retrieved cases N_r varied from 10 to 200. For comparison, results are also showed for the baseline classifier

some of the retrieved cases will become less similar to the query and will not help the classification on the query.

10.4 MDS as a Visualization Tool of Example Lesions

As an alternative approach to CADx, retrieving a set of known lesion similar to the one being evaluated might be of value in assisting radiologists in their diagnosis. In recent years, such an approach has been studied by researchers and applied for different lesion types and imaging modalities [31, 84, 86, 89, 91–94]. For this purpose, we have been studying the use of multidimensional scaling (MDS) for representation and analysis of similar lesions in a large dataset. In a retrieval framework, MDS can be used to study how a query tumor might be related to a set of similar images retrieved from a reference library [86]. When used as a visualization tool, MDS allows one to browse and explore intuitively the distribution of benign and malignant MC lesions in a dataset and to examine how this distribution might be related to the features of the tumors [14, 15, 92, 95].

10.4.1 Multidimensional Scaling (MDS) Technique

MDS is a data embedding technique for representation and analysis of a set of objects based on their mutual similarity (or dissimilarity) measurements [96]. The

basic idea of MDS is to represent the objects of interest as points in a low-dimensional (typically 2D or 3D) space such that the geometric distances between the points in this space are in accordance with the similarity measurements between the corresponding objects. The resulting representation in this lower dimensional space enables one to visualize the relationship among the objects in a rather intuitive manner.

Specifically, consider a set of N objects. The MDS seeks to embed these objects in a lower-dimensional space (R^2 or R^3) as a set of data points $\mathbf{x}_i$, $i = 1, \cdots, N$, such that the Euclidean distance $d(\mathbf{x}_i, \mathbf{x}_j)$ between a pair of points $\mathbf{x}_i$ and $\mathbf{x}_j$ is proportional to their pairwise proximity measure δ_{ij}. This is accomplished by minimizing the following objective function:

$$\sigma^2 = \frac{\sum w_{ij} \left[d\left(\mathbf{x}_i, \mathbf{x}_j\right) - \delta_{ij} \right]^2}{\sum w_{ij} \delta_{ij}} \tag{10.15}$$

where w_{ij} are weight factors (specified by users). The quantity σ is known as Stress-1, which measures the goodness of fit of the MDS model.

In our application, we use MDS to represent tumors from mammogram images as points in a 2D plane, wherein the similarity between a pair of tumor images is defined according to their perceptual similarity. Thus, tumors that are in close vicinity of each other in the MDS plot correspond to those that are perceptually similar.

10.4.2 Exploring Similar MC Lesions with MDS

In order to explore how perceptually similar cases with clustered MCs may relate to one another in terms of their underlying characteristics (from disease condition to image features), we conducted an observer study to collect similarity scores from a group of readers on a set of 2000 image pairs, which were selected from 222 cases based on their images features. Afterward, we applied MDS to embed all the cases in a 2D plot, in which the potential relationship among the different cases is exhibited according to their similarity ratings. Such a plot allows one to study how neighboring cases (i.e., cases similar to each other) may relate to one another. In particular, we will examine the relationships among the cases in several aspects, including: (1) case pathology, (2) spatial distribution patterns of their clustered MCs, and (3) image pairs of clustered MCs that are highly similar.

Dataset: The dataset used in this study was collected by the Department of Radiology at the University of Chicago. It consists of 365 mammogram images from 222 cases (110 malignant, 112 malignant), of which all have been proven by biopsy containing lesions with MCs. These images are of dimension 1024 × 1024 or 512 × 512 pixels, digitized with a spatial resolution of 0.1 mm/pixel. Among the 222 cases, 143 have images in both carniocaudal (CC) and mediolateral-oblique (MLO) views. The MCs in each mammogram were manually identified by a group of experienced radiologists. These MCs were used as ground truth in our study.

Since we are mainly interested in the pairs of images that are similar, we first apply a selection procedure based on the image features of the MCs in these cases to identify those potentially similar image pairs for reader scoring. For this purpose, a set of nine image features [11, 97] are used for quantifying the MCs; these features are commonly used for the classification of MC lesions in computer-aided diagnosis (CADx). Specifically, they are: (1) image features describing individual MCs, including the standard deviation of the image contrast values of MCs, and the maximum and the standard deviation of the sizes of MCs, (2) spatial clustering features of MCs, including the number of MCs in a cluster, the area of the cluster, and the compactness of the cluster, and (3) texture-based features, including the energy, contrast, and correlation derived from the gray-level co-occurrence matrices. The cases in the dataset are then selected for pairing based on the feature values (Euclidean distance) of their MCs. In the end, a total of 2000 image pairs were selected.

Subsequently, based on the similarity scores collected on the 2000 image pairs (described below), we further select a subset of 1000 image pairs from them, the purpose being to refine the set of potentially similar pairs for further reader scoring. These pairs are selected based on both the similarity scores from the readers and the Euclidean distances of the all nine features.

Reader study: The reader study was carried out by a group of five radiologists for the 1000 image pairs, based on their perceptual similarity, using a discrete scale from 0 (most dissimilar) to 10 (most similar). These five radiologists are MQSA-qualified breast imagers with between 2 and 20 years of experience. To reduce the effects of reader fatigue, the set of image pairs is randomly divided into four separate scoring sessions. Similarly, a separate reader study was carried out by a group of five non-radiologists for the 2000 image pairs (which were used for further pair selection as described above). These readers were researchers in breast imaging with a minimum of 5 years of experience. A total of 10 separate sessions were used in scoring.

Because of the subjective nature in interpretation of clustered MCs in mammogram images, readers can vary in their similarity scores. To suppress such apparent differences, we first transformed the similarity scores from individual readers into z-scores. Afterward, the scores were averaged among the readers for the set of 1000 image pairs (denoted by S_1), which were scored by both radiologists and non-radiologists, and similarly for the other set of 1000 image pairs (denoted by S_2), which were scored by only non-radiologists. The average scores are further transformed into z-scores.

MDS plot: To explore how perceptually similar cases with clustered MCs relate to each other, we apply the MDS technique to embed the different cases in the dataset in a 2D plot based on their similarity scores.

Consider a pair of cases i and j with similarity score SC_{ij}. In the MDS placement, they will be separated by proximity

$$\delta_{ij} = \frac{1}{3.75 + SC_{ij}} \tag{10.16}$$

where a constant offset 3.75 (over three standard deviation) is added to ensure that d_{ij} is positive.

Due to the fact that similarity scores are available for only those image pairs scored by the readers, the weighted MDS technique is used, in which those image pairs not scored are assigned a weight of 0; for the scored image pairs, the weight is adjusted according to the level of similarity and the readers for scoring as following: for image pair p consisting of cases i and j,

$$w_p = \begin{cases} 1 & \text{if } p \in S_2 \text{ and } SC_{ij} \le 0 \\ 1.5 & \text{if } p \in S_1 \text{ and } SC_{ij} \le 0 \\ 2 & \text{if } p \in S_2 \text{ and } SC_{ij} > 0 \\ 3 & \text{if } p \in S_1 \text{ and } SC_{ij} > 0 \end{cases} \tag{10.17}$$

The rationale for such a choice is to assign a higher weight value to pairs that are more similar and scored by more readers.

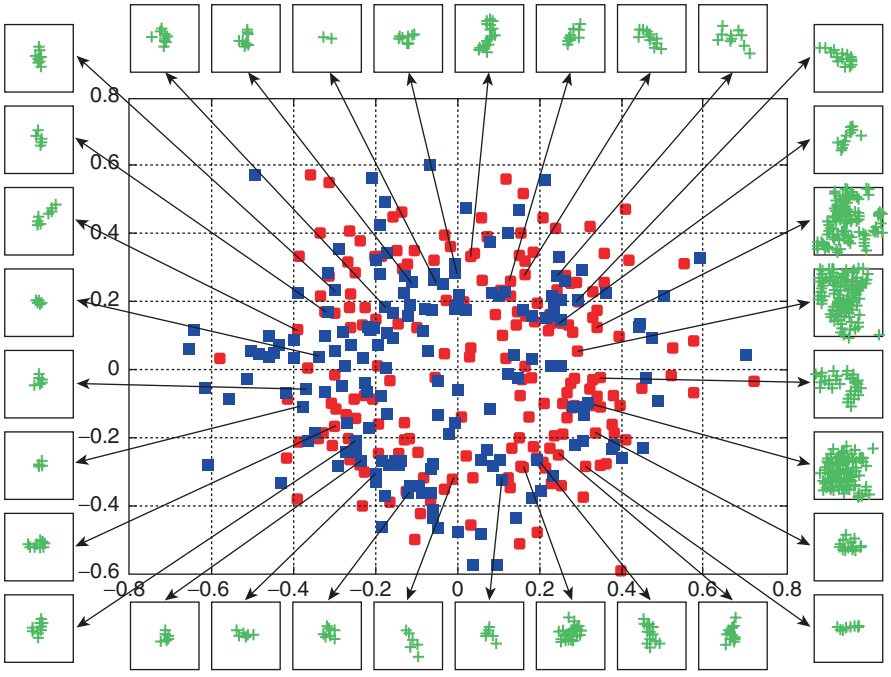


Fig. 10.5 MDS embedding of perceptually similar cases in the dataset, wherein cancer cases are denoted by "red dots" and benign cases are represented by "blue squares." The spatial distribution patterns of clustered MCs are shown for some sample cases, where the spatial MC locations are indicated by the "green plus" signs

In Fig. 10.5, we show the MDS embedding of all the 222 cases in the dataset according to their similar scores. While at the first sight there is no apparent separation between cancer and benign cases, it is evident that there are more cancer cases (and fewer benign cases) in the right half of the plot than in the left half. More importantly, cases of same disease tend to be clustered together locally. For example, while cancer cases are scattered in different regions throughout the plot, they are also distributed in small clusters in which a cancer case is closely surrounded by other cancer cases; the same is true for benign cases.

Furthermore, to explore how the readers' notion of similarity may relate to the image features of the clustered MCs. We also show in Fig. 10.5 the spatial distribution patterns of the clustered MCs for some sample cases, where the spatial locations of the individual MCs are indicated by "+" signs. It can be seen that the neighboring cases tend to have MC clusters similar in size and shape and that the MC clusters in the right half of the plot tend to be larger and irregular.

10.5 Issues and Recommendations

Despite that there have been great many computerized methods developed for use in CADx schemes as a diagnostic aid to improving radiologists' diagnostic accuracy, some significant, challenging issues still remain to be addressed. Below we discuss a few of them, which are by no means meant to be complete.

Thanks to intense research and development efforts, multiple laboratory observer studies have shown that CADx schemes can help improve the diagnostic accuracy in differentiating between benign and malignant tumors. For example, in mammography, radiologists with CADx can improve their biopsy recommendation by sending more cancer cases and fewer benign cases to biopsy [20–22, 75, 98].

In CADx, the computer predicts the likelihood that a lesion is malignant, which is presented to the radiologist as a second opinion. One difficulty in implementing CADx clinically is that a CADx classifier is often criticized for being a "black box" approach in its decision. When presented with a numerical value, such as the likelihood of malignancy, but without additional supporting evidence, it may be difficult for a radiologist to incorporate optimally this number into his or her decision. As an alternative aid, image retrieval has been studied as a CADx tool in recent years. We conjecture that by integrating a retrieval system with the CADx classifier, the retrieved images could serve as supporting evidence to the CADx classifier, which may facilitate the interpretation of the likelihood of malignancy by the radiologists.

In the literature, the CADx schemes are often, if not always, developed with different datasets which are limited by the number of cases available. The heterogeneity among the different datasets will inevitably lead to variability when evaluating the performance of a CADx scheme. Thus, it is desirable to establish common benchmark databases which are large enough to be representative of a disease population. In practice, this can be an expensive process. It will ensure that a CADx

scheme can be optimized and tested without any bias so that it can generalize well when applied to cases outside the database.

Finally, while research and development has led to improvement in CADx performance, as a diagnostic aid, the accuracy level achieved by CADx classifiers is rather moderate for certain tumor types due to the inherent difficulty of the problem (e.g., MC lesions). There is still need for development of more salient features and CADx algorithms in order to improve the classification accuracy. This may include the use of additional features acquired from multi-modality imaging.

10.6 Conclusions

In this chapter, we presented the application of machine learning algorithms in CADx systems. Particularly, we presented examples of its application in mammography to differentiate benign and malignant cases. A main critique of traditional CADx approaches when implemented clinically is that a CADx classifier could be perceived as a "black box" approach in its decision. As an alternative aid, CBIR has been studied as a CADx tool in recent years. We conjecture that with the integration of a retrieval system and a CADx classifier, the retrieved images could serve as supporting evidence to the CADx classifier, which may facilitate the interpretation of the likelihood of malignancy by the radiologists in clinical practice. In this chapter, we presented the process of developing and validating such system exploiting both supervised and unsupervised machine learning algorithms.

References

1. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395–4010.
2. Horeweg N, Scholten ET, de Jong PA, van der Aalst CM, Weenink C, Lammers J-WJ, et al. Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. Lancet Oncol. 15:1342–50. https://doi.org/10.1016/S1470-2045(14)70387-0.
3. Agarwal V, Branstetter Iv BF, Johnson JT. Indications for PET/CT in the head and neck. Otolaryngol Clin N Am. 2008;41:23–410. https://doi.org/10.1016/j.otc.2007.10.005.
4. Thompson J, Lawrentschuk N, Frydenberg M, Thompson L, Stricker P, Usanz V. The role of magnetic resonance imaging in the diagnosis and management of prostate cancer. BJU Int. 2013;112(Suppl 2):6–20. https://doi.org/10.1111/bju.12381.
5. Leung D, Han X, Mikkelsen T, Nabors LB. Role of MRI in primary brain tumor evaluation. J Nat Compreh Cancer Netw. 2014;12:1561–8.
6. Young RJ, Knopp EA. Brain MRI: tumor evaluation. J Mgnt Reson Img. 2006;24:709–24. https://doi.org/10.1002/jmri.20704.
7. Pilewskie M, King TA. Magnetic resonance imaging in patients with newly diagnosed breast cancer: a review of the literature. Cancer. 2014;120:2080–10. https://doi.org/10.1002/cncr.28700.
8. Wu Y, Giger M, Doi K, Vyborny C, Schmidt R, Metz C. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology. 1993;187(1):81–7.

9. Andreadis II, Spyrou GM, Nikita KS. A comparative study of image features for classification of breast microcalcifications. Meas Sci Technol. 2011;22(11):114005.
10. Cheng HD, Cai X, Chen X, Hu L, Lou X. Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern Recogn. 2003;36:2967–91.
11. Jiang Y, Nishikawa RM, Wolverton EE, Metz CE, Giger ML, Schmidt RA, Vyborny CJ. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology. 1996;198:671–8.
12. Sakka E, Prentza A, Koutsouris D. Classification algorithms for microcalcifications in mammograms (review). Oncol Rep. 2006;15(4):1049–55.
13. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep. 2016c;6(1):1–10.
14. Wei L, Yang Y, Nishikawa RM, Wenick MN, Edwards A. Relevance vector machine for automatic detection of clustered microcalcifications. IEEE Trans Med Imaging. 2005a;24(10):1278–85.
15. Wei L, Yang Y, Nishikawa RM, Jiang Y. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. IEEE Trans Med Imaging. 2005b;24(3):371–80.
16. Al-antari MA, Al-masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. Int J Med Inform. 2018;117:44–54.
17. Bozek J, Mustra M, Delac K, Grgic M. A survey of image processing algorithms in digital mammography. Rec Adv Multimedia Signal Process Commun. 2010;15:631–57.
18. Cheng HD, Shi XJ, Min R, Hu LM, Cai XP, Du HN. Approaches for automated detection and classification of masses in mammograms. Pattern Recogn. 2006;310(4):646–68.
19. Elter M, Horsch A. CADx of mammographic masses and clustered microcalcifications: a review. Med Phys. 2001;36:2052–68.
20. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Metz CE. Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness. Acad Radiol. 2000;7(12):1077–84.
21. Chan H, Sahiner B, Lam KL, Petrick N, Helvie MA, Goodsitt MM, Adler DD. Computerized analysis of mammographic microcalcifications in morphological and texture feature space. Med Phys. 1998;25:2007–20110.
22. Horsch K, Giger ML, Vyborny CJ, Lan L, Mendelson EB, Hendrick RE. Classification of breast lesions with multimodality computer-aided diagnosis: observer study results on an independent clinical data set. Radiology. 2006;240:357–68.
23. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schimidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol. 1998;5(3):155–68.
24. Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. Potential of computer-adied diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. Radiology. 2001a;220:787–94.
25. Jiang Y, Nishikawa RM, Papaioannou J. Dependence of computer classification of clustered microcalcifications on the correct detection of microcalcifications. Med Phys. 2001b;28(9):1949–57.
26. Wang J, Yang Y, Wernick MN, Nishikawa RM. An image-retrieval aided diagnosis system for clustered microcalcifications. IEEE 13th Int Symp Biomed Imag (ISBI). 2016b;10:1076–9.
27. Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications--clinical benefits and future directions. Int J Med Inform. 2004;73:1–23.
28. Bustos B, Keim D, Saupe D, Schreck T. Content-based 3D object retrieval. IEEE Comput Graph Appl. 2007;27:22–7.
29. Bhanu B, Peng J, Qing S. Learning feature relevance and similarity metrics in image database. IEEE Workshop Proceedings on Content-Based Access of Image and Video Libraries. 1998;12:14–8.

30. El Naqa I, Yang Y, Galatsanos NP, Nishikawa RM, Wernick MN. A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans Med Imaging. 2004;23:1233–44.
31. El-Naqa I, Yang Y, Galatsanos NP, Nishikawa RM, Wernick MN. A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans Med Imaging. 2004a;23:1233–44. https://doi.org/10.1109/TMI.2004.834601.
32. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
33. Jiang Y, Nishikawa RM, Giger ML, Doi K, Schmidt R, Vyborny C. Method of extracting signal area and signal thickness of microcalcifications from digital mammograms. Proc SPIE. 1992;1778:28–36.
34. Sachdeva J, Kumar V, Gupta I, Khandelwal N, Ahuja CK. Segmentation, feature extraction, and multiclass brain tumor classification. J Digit Imaging. 2013;26(6):1141–50.
35. Soltanian-Zadeh H, Rafiee-Rad F, Pourabdollah-Nejad S. Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms. Pattern Recogn. 2004;37:1973–86.
36. Haralick R, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans On Systems Man and Cybernetics. 1973;3:610–21.
37. Gonzalez RC, Woods RE. Digital image processing. Princeton: Prentice Hall; 2002.
38. Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. ICML. 2003;3:856–63.
39. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997;97(1):273–324.
40. Perkins S, Lacker K, Theiler J. Grafting: fast, incremental feature selection by gradient descent in function space. The Journal of Machine Learning Research. 2003;3:1333–56.
41. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
42. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer; 2001.
43. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
44. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge university press; 2000.
45. Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res. 2001;1:211–44.
46. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–1310.
47. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn. 2000;40(2):139–57.
48. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21.
49. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big data. 2016;3(1):10.
50. Zhang S, Sun F, Wang N, Zhang C, Yu Q, Zhang M, Babyn P, Zhong H. Computer-aided diagnosis (CAD) of pulmonary nodule of thoracic CT image using transfer learning. J Dig Imag. 2011;32(6):995–1007.
51. Deepak S, Ameer PM. Brain tumor classification using deep CNN features via transfer learning. Comput Biol Med. 2011;111:103345.
52. Natalia A, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys. 2017;44(10):5162–71.
53. Cui S, Luo Y, Tseng H-H, Ten Haken RK, El Naqa I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. Med Phys. 2011;46(5):2497–511.
54. Wang Y, Heidari M, Mirniaharikandehei S, Gong J, Qian W, Qiu Y, Zheng B. A hybrid deep learning approach to predict malignancy of breast lesions using mammograms. Med Imag. 2018;12:10579.
55. Breiman L, Spector P. Submodel selection and evaluation in regression. The x-random case. Int Stat Rev. 1992;14:291–3110.

56. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI. 1995;14(2):1137–45.
57. Mertens BJ, de Noo ME, Tollenaar RAEM, Dcclder AM. Mass spectrometry proteomic diagnosis: enacting the double cross-validatory paradigm. J Comput Biol. 2006;13(9):1591–605.
58. Mushlin AI, Kouides RW, Shapiro DE. Estimating the accuracy of screening mammography: a meta-analysis. Am J Prev Med. 1998;14:143–53.
59. Urbain JL. Breast cancer screening, diagnostic accuracy and health care policies. CMAJ. 2005;172:210–1.
60. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology. 2002;225:165–75.
61. Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med. 1998;338(16):1089–96.
62. Tan A, Freeman DH Jr, Goodwin JS, Freeman JL. Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment. Breast Cancer Res Treat. 2006;100:309–18.
63. Wang J, Yang Y. Boosted classification of breast cancer by retrieval of cases having similar disease likelihood. ICASSP. 2016;12:908–11.
64. de Cea MVS, Nishikawa RM, Yang Y. Estimating the accuracy level among individual detections in clustered microcalcifications. IEEE Trans Med Imaging. 2017;36(5):1162–71.
65. Wang J, Yang Y. Spatial density modeling for discriminating between benign and malignant microcalcification lesions. ICIP. 2013:133–6.
66. Salfity MF, Nishikawa RM, Jiang Y, Papaioannou J. The use of a priori infor- mation in the detection of mammographic microcalcifications to improve their classification. Med Phys. 2003;30(5):823–31.
67. El-Naqa I, Yang Y, Wernick MN, Galatsanos NP, Nishikawa RM. A support vector machine approach for detection of microcalcifications. IEEE Trans Med Imaging. 2002;21(12):1552–63.
68. Mordang JJ, Gubern-Mérida A, Bria A, et al. The importance of early detection of calcifications associated with breast cancer in screening. Breast Cancer Res Treat. 2018;167(2):451–58. https://doi.org/10.1007/s10549-017-4527-7.
69. Wang J, Nishikawa RM, Yang Y. Global detection approach for clustered microcalcifications in mammograms using a deep learning network. J Med Imag. 2017;4(2):024501.
70. Wang J, Yang Y. A context-sensitive deep learning approach for microcalcification detection in mammograms. Pattern Recogn. 2018;78:12–22.
71. Wang J, Nishikawa RM, Yang Y. Improving the accuracy in detection of clustered microcalcifications with a context-sensitive classification model. Med Phys. 2016a;43(1):159–70.
72. Sickles EA, Miglioretti DL, Ballard-Barbash R, Geller BM, Leung JWT, Rosenberg RD, Smith-Bindman R, Yankaskas BC. Performance benchmarks for diagnostic mammography. Radiology. 2005;235:775–90.
73. Chan H, Wei D, Lam K, Lo S, Sahiner B, Helvie M, Adler D. Computerized detection and classification of microcalcifications on mammograms. SPIE. 1995;2434:612–20.
74. Markopoulos C, Kouskos E, Koufopoulos K, Kyriakou V, Gogas J. Use of artificial neural networks (computer analysis) in the diagnosis of microcalcifications on mammography. Eur J Radiol. 2001;39(1):60–5.
75. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol. 1991;6:22–33.
76. Nishikawa RM. Current status and future directions of computer-aided diagnosis in mammography. Comput Med Imaging Graph. 2007;31:224–35.
77. Rangayyan RM, Fabio JA, Desautels JL. A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs. J Franklin Instit. 2007;344:312–48.
78. Sampat MP, Markey MK, Bovik AC. Computer-aided detection and diagnosis in mammography. Chap. 10.4, Handbook of image & video processing, 2nd ed., Elsevier Academic Press New York; 2005.

79. Muller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval system in medical applications-clinical benefits and future directions. Int J Med Informat. 2004;73:1–23.
80. Rahman M, Want T, Desai B. Medical image retrieval and registration: towards computer assisted diagnostic approach. In Proc. IDEAS Workshop on Medical Information Systems: The Digital Hospital. 2004; 78–810.
81. Holt A, Bichindaritz I, Schmidt R, Perner P. Medical applications in case-based reasoning. Knowl Eng Rev. 2005;20:289–92.
82. Floyd CE, Lo J, Tourassi GD. Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. Am J Roentgenol. 2000;175(5):1347–52.
83. Bilska-Wolak A, Floyd E. Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADS™ lexicon. Med Phys. 2002;29:2090.
84. Zheng B, Lu A, Hardesty LA, Sumkin JH, Hakim CM, Ganott MA, Gur D. A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment. Med Phys. 2006;33:111–7.
85. Jing H, Yang Y. Case-adaptive classification based on image retrieval for computer-aided diagnosis. ICIP. 2010;12:4333–6.
86. Wei L, Yang Y, Nishikawa RM, Jiang Y. Learning of perceptual similarity from expert readers for mammogram retrieval. ISBI. 2006:1356–13510.
87. Wei L, Yang Y, Nishikawa RM. Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis. Pattern Recogn. 2001;42:1126–32.
88. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer WP. The digital database for screening mammography. Fifth Int Workshop on Digital Mammogr. 2001;15:212–8.
89. El-Naqa I, Yang Y, Galasanos NP, Nishikawa RM, Wernick MN. A similarity learning approach to content based image retrieval: application to digital mammography. IEEE Trans Med Imag. 2004b;23:1233–44.
90. Mcloughlin KJ, Bones PJ, Karssemeijer N. Noise equalization for detection of microcalcification clusters in direct digital mammogram images. IEEE Trans. Med. Imag. 2004;23(3):313–20.
91. Aisen A, Broderick L, Winer-Muram H, Brodley C, Kak A, Pavlopoulou C, Dy J, Shyu C, Marchiori A. Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment. Radiology. 2003;228:265–70.
92. Muramatsu C, Nishimura K, Endo T, Oiwa M, Shiraiwa M, Doi K, Fujita H. Representation of lesion similarity by use of multidimensional scaling for breast masses on mammograms. J Digit Imaging. 2013;26(4):740–7.
93. Tourassi GD, Harrawood B, Singh S, Lo JY, Floyd CE. Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms. Med Phys. 2007;34:140–50.
94. Yang L, Jin R, Mummert L, Sukthankar R, Goode A, Zheng B, Hoi SCH, Satyanarayanan M. A boosting framework for visuality-perserving distance metric learning and its application to medical image retrieval. IEEE Trans Pattern Aana Mach Intell. 2010;32(1):30–44.
95. Wang J, Jing H, Wernick MN, Nishikawa RM, Yang Y. Analysis of perceived similarity between pairs of microcalcification clusters in mammograms. Med Phys. 2014;41(5):051904.
96. Borg I, Groenen PJF. Modern multidimensional scaling: theory and application. New York: Springer; 2005.
97. Karahahiou AN, Boniatis IS, Skiadopoulos SG, Sakellaropoulos FN, Arikidis NS, Likaki EA, Panayiotakis GS, Costaridou LI. Breast cancer diagnosis: analyzing texture of tissue surrounding microcalcifications. IEEE Trans Inf Technol Biomed. 2008;12:731–8.
98. Hadjiiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Alder D, Nees A, Shen J. Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study. Radiology. 2004;233:255–65.

Auto-contouring for Image-Guidance and Treatment Planning

11

Rachel B. Ger, Tucker J. Netherton, Dong Joo Rhee, Laurence E. Court, Jinzhong Yang, and Carlos E. Cardenas

11.1 Introduction

Image segmentation is an important task routinely performed in radiotherapy to identify treatment targets and anatomical structures (organs-at-risk, OARs). In a typical clinical workflow, a radiation oncologist or dosimetrist manually segments these regions of interest (ROI) on a radiotherapy simulation scan. Traditionally, computed tomography (CT) scans have been used for radiotherapy simulation; however, with the increasing role of magnetic resonance imaging in brachytherapy procedures and with the advent of magnetic resonance(MR)-guided external beam RT, MR simulation scans are being more rapidly adopted for radiotherapy planning in clinics worldwide. The manual segmentation of these ROIs is a time-consuming process with some studies reporting several hours of physician time per patient [1–3]. This could lead to significant delays in start of radiotherapy treatment, particularly in clinics with limited resources, which has been correlated to worse loco-regional control and overall

Excerpts from this chapter were previously published in "Advances in auto-segmentation," (2019) *Seminars in Radiation Oncology*. Vol. 29, No. 3, pp. 185–197. by Cardenas et al.

R. B. Ger
Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA

T. J. Netherton · D. J. Rhee · L. E. Court · J. Yang
Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

C. E. Cardenas (✉)
Department of Radiation Oncology, The University of Alabama at Birmingham, Birmingham, AL, USA
e-mail: cecardenas@uabmc.edu

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_11

survival rates [4, 5]. Furthermore, the significant time commitment required to segment each patient's ROIs has been regarded as a rate-limiting step for adaptive RT, as it is necessary for the ROIs to be segmented on new imaging reflecting patient's anatomical changes to ensure accurate dose accumulation estimates for the radiotherapy treatment.

The efficacy and safety of the radiotherapy plan require accurate segmentations as these regions of interest are generally used to optimize and assess the quality of the plan. However, inconsistencies in target and OAR segmentations have been reported in studies assessing inter- and intra-observer segmentation variability [1, 6–8]. These inconsistencies may arise from the fact that the segmentation task can be subjective in nature as the expert performing the segmentations evaluates the available imaging and then makes the decision, based on prior knowledge and/or experience, of what voxels to include as part of the ROI being segmented. Subsequently, the inherent variability observed in manual segmentations could have a significant impact on quantitative [9–13] (e.g., radiomics) and dosimetric analyses [1, 14–16]. Automatic segmentation (or auto-segmentation) is, therefore, preferable as it would address these challenges.

Auto-segmentation algorithms must overcome several image-related problems to ensure accurate predictions. First, medical images are subject to noise that can affect the intensity of a voxel. Second, tissues within a patient typically exhibit intensity nonuniformity, meaning that voxel intensities within a single tissue may vary over the extent of the image. Lastly, medical images are reconstructed during acquisition to have a predefined voxel size which leads to partial volume averaging. Limited by a finite image resolution, voxels may contain more than one tissue such that the voxel intensity may not be representative of either tissue class. Furthermore, there are imaging modality-related challenges that may be specific to individual modalities. While MR scans provide exquisite soft tissue contrast, image intensities tend to vary between acquisitions due to magnetic susceptibility artifacts. These problems, along with the large anatomical presentation and tissue distribution among different individuals in a population, suggest that some degree of uncertainty is expected for both manual and auto-segmentations.

The field of medical image auto-segmentation has rapidly evolved over the past two decades. Previously, auto-segmentation techniques have been grouped into first-, second-, and third-generation algorithms, representing a new standard in algorithm development [17]. However, more recently, deep learning-based auto-segmentation techniques have been shown to provide significant improvements over more traditional approaches, suggesting we have entered the fourth generation of auto-segmentation algorithm development.

The field of deep learning became more mainstream after the seminal paper by Krizhevsky et al. showed that using a deep convolutional neural network architecture (AlexNet) could significantly improve predictions in image classification and recognition tasks [18]. In their work, the authors employed graphical processing units (GPU) to perform convolutional computations significantly reducing the time required to train their classification model on the ImageNet dataset [19]. Shortly after, research showed that using convolutional neural networks (CNN) for image segmentation tasks could outperform previously preferred algorithms, resulting in the swift adaptation of these architectures for medical image auto-segmentation.

This chapter provides a brief overview of traditional (pre-deep learning era) auto-segmentation techniques, introduces concepts behind deep learning-based

auto-segmentation algorithms and commonly used architectures, presents considerations for clinical implementation of auto-segmentation tools, and provides a brief overview of the state-of-the-art results in radiotherapy image segmentation.

11.2 Traditional Auto-Segmentation Techniques

The development of auto-segmentation algorithms has been accompanied by the capability of the algorithms to use prior knowledge for new segmentation tasks. In an early stage, limited by the computer power and the availability of segmented data, most segmentation techniques used no or little prior knowledge, referred to as low-level segmentation approaches. More advanced techniques were developed in an attempt to avoid heuristic approaches leading to the introduction of uncertainty models and optimization methods [17, 20].

11.2.1 First-Generation Auto-Segmentation Techniques

The first generation of auto-segmentation techniques are low-level techniques that include little to no prior information. These include thresholds, region growing, and edge tracing. These techniques are common and available in most commercial contouring solutions. Thresholds are simply applied with cutoff image intensity units to identify contrasting regions of image intensity. These are commonly used within commercial systems to contour contrasting organs from their surrounding areas, such as the lungs or brain. Region growing involves picking a seed location and identifying a homogeneity criteria. Pixels or voxels that meet these criteria are included. This process is applied outward from the seed location to determine the contour. Region growing is often used in contouring lesions on PET images, while tools exist for CT contouring but are less consistent in their results due to difficulty in determining the homogeneity criteria to be used. First-generation techniques often suffer from all main issues that plague auto-segmentation techniques: noise in images that affects intensity of pixels, intensity nonuniformity where a given tissue gradually varies in intensity over the image, and partial volume averaging. Users always have to make significant edits to structures generated using these techniques.

11.2.2 Second-Generation Auto-Segmentation Techniques

The second-generation techniques attempt to handle these issues that affect auto-segmentation with uncertainty models and optimization methods. Techniques that fall into this group are statistical pattern recognition, c-means clustering, deformable models, graph search, multiresolution methods, minimal path, and target tracking.

Region-based techniques, such as active contours, level-sets, graph cuts, and watershed algorithms, have been used in medical imaging auto-segmentation. Active contours and level-set algorithms are considered deformable models as they use closed surfaces that are able to contract or expand to conform to distinct image features within an image, whereas graph cuts and the watershed algorithm employ

principles behind graph theory to maximize interconnections between image voxels. Probability-based auto-segmentation techniques, such as Gaussian mixture models, clustering, k-nearest neighbor, Bayesian classifiers, and shallow artificial neural networks, rose in popularity with the turn of the century thanks to advances in the statistical community and the availability of higher computing power. These approaches are characterized by their ability to classify individual voxels in an image as belonging to one of a known set of classes; however, these typically lack contextual information from neighboring voxels (as voxels are classified independently) which is often mediated by implementing hidden Markov random fields [21].

11.2.3 Third-Generation Auto-Segmentation Techniques

Third-generation techniques further build upon the advances of the second generation to avoid the image based-issues of segmentation by incorporating higher-level knowledge. This is done through a priori information, defined rules, and models of the desired contour. These techniques include shape models, appearance models, atlas-based segmentation, rule-based segmentation, and coupled surfaces.

In the last two decades, a large amount of exploratory work has been invested in making use of prior knowledge. An example is the use of shape and appearance characteristics of anatomical structures to compensate for insufficient soft tissue contrast of CT data which prevents accurate definition of the anatomical boundary. Depending on how much prior knowledge is used in the algorithms, the approaches can be grouped as (multi)atlas-based segmentation, model-based segmentation, and machine learning-based segmentation [22].

Single atlas-based segmentation uses one reference image with structures of interest already segmented, referred to as an atlas, as prior knowledge for new segmentation tasks [23]. The segmentation of a new image relies on deformable registration finding the optimal transformation between the atlas and the new image to map the atlas contours onto the new image (Fig. 11.1). Varied deformable registration algorithms have been used for this purpose [24–29], and most of them are intensity-based algorithms. The segmentation performance solely depends on the performance of deformable registration, which is influenced by the similarity of the morphology of organs of interest between atlas and the new image. To achieve good segmentation results, varied atlas selection strategies have been proposed [30–36]. Alternatively, using an atlas that reflects an average patient anatomy can potentially improve segmentation performance [37, 38].

Atlas-based segmentation is often impacted by inter-subject variability. Instead of using a single-atlas, multi-atlas approaches use a number of atlases (usually around 10) as prior knowledge for new segmentation tasks [39–44]. Similar to single-atlas-based approaches, deformable registration is the enabling technique to map individual atlas contours to the new image (Fig. 11.1). An additional step, frequently referred to as label/contour fusion, is performed to combine the individual segmentations from multiple atlases to produce a final segmentation that is the best estimate of the true segmentation [36, 45–48]. Multi-atlas segmentation has been shown to minimize the effects of inter-subject variability and improve

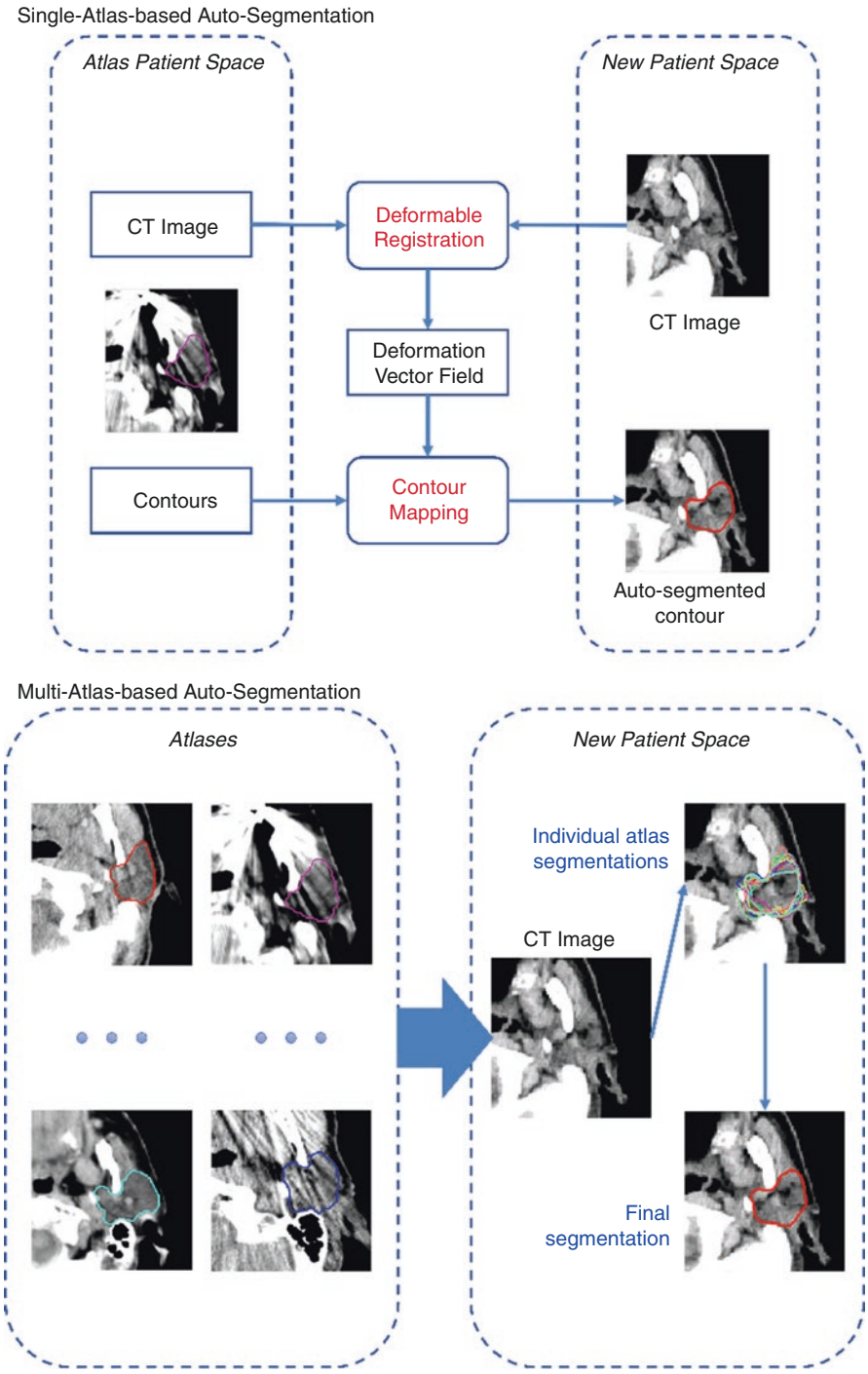


Fig. 11.1 Comparison of single-atlas- and multi-atlas-based auto-segmentation. For multi-atlas-based auto-segmentation, multiple atlases are used to generate contours on the new CT image; the resultant individual atlas segmentations are then combined to derive the final auto-segmented contour

segmentation accuracy from single-atlas approaches. The atlases define the prior knowledge, which therefore can affect the accuracy of contours based on representation within the atlases; for example, if only post-operational head and neck patients are used to define head and neck atlases, then contours on a pre-operational patient will not perform well as this patient is not represented in the prior knowledge. In the past decade, multi-atlas segmentation has been shown as one of the most effective segmentation approaches in several grand challenges [49–51]. This approach has been validated for clinical radiation oncology applications in contouring head and neck normal tissue [52], cardiac substructures [53], brachial plexus [41], and others.

When more contoured images are available, characteristic variations of shape or appearance of structures of interest could be used to train statistical shape models (SSM) or statistical appearance models (SAM) for auto-segmentation. SSM uses deformable models while limiting the extent of allowed model deformation. Landmark points on an object boundary are identified and analyzed on training images to determine a statistical representation. Then in contouring, deformation occurs but is restricted within the bounds of the model. SAM is an extension of SSM where both shape and intensity of an object are incorporated into a statistical model. These approaches can restrict the final segmentation results to anatomically plausible shapes described by the models [54]. However, model-based segmentation is less flexible due to the limitation of specific shapes characterized by the statistical models. Also, size and content of the training data limit the segmentation performance. In radiation oncology applications, model-based segmentation is mostly used for the segmentation of structures in the pelvic region [55–57].

On the other hand, when more contoured images are available, machine learning approaches can aid in segmentation by learning correspondent features for structures and organs or image context and tissue appearance for voxel classification [58–60]. Support vector machines [61–65] and tree ensemble (i.e., random forests) [66–72] algorithms have shown promising results in thoracic, abdominal, and pelvic tumor and normal tissue segmentation. These generally employ human-engineered features (i.e., radiomics) usually derived from the image intensity histograms, from a large patient database as inputs to train the segmentation model (Fig. 11.2). Conventional radiomics features were originally designed for use with satellite images to determine objects within the image, such as the gray level co-occurrence matrix from Haralick [73]. The medical community has adopted these conventional features although their biological meaning can be difficult to understand in comparison to intensity histogram features, such as kurtosis. Additionally, studies employing radiomics features often use different filters before extraction of the features which creates a large number of features to be analyzed. Therefore, techniques must be employed to efficiently reduce the number of features before proceeding with model building to ensure adequate power is preserved. Typically for contouring, the simple intensity and gradient features are used. Radiomics (*see* Fig. 11.2) has been used to contour brain tumors [58, 61, 62], liver [63], lung nodules [65], cardiac structures [66], the prostate and prostatic lesions [57, 60, 74], pelvic organs [70], kidney sub-structures [68], chest lymph nodes [71], and others.

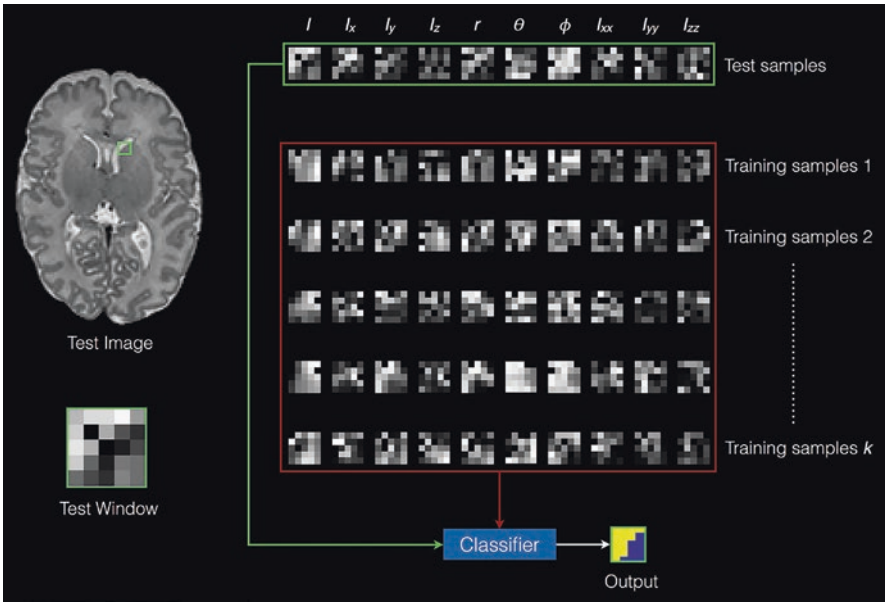


Fig. 11.2 Illustration from Serag et al. [72] demonstrating how imaging features can be used for medical image segmentation. The green box represents a small window from the test image from which different feature maps are calculated (shown by the green rectangle). Similarly, features from other patients are used to train a model that can be used as classifier to classify individual pixels on the test window

The traditional auto-segmentation techniques have advanced significantly over the last decade due to advancements in computing systems. The first-generation techniques are simple and do not require much computing power. These techniques are standard tools in most commercial contouring systems. With the advancement in computers, the second- and third-generation techniques came into being and allowed the use of a priori information. The most significant of these techniques, as it is more and more readily available in commercial systems, is atlas-based segmentation. Recently, there have been further advancements in computing systems that have moved auto-segmentation into deep learning and transitioned into the fourth generation of auto-segmentation techniques.

11.3 Deep Learning-Based Auto-Segmentation

Over the past few years, deep learning networks have made tremendous advances in image segmentation, resulting in the birth of a new generation of segmentation models. Paving the way as the fourth generation of auto-segmentation algorithms, deep learning models often achieve the highest segmentation accuracy on public segmentation challenges, suggesting we are entering a paradigm shift in medical image auto-segmentation algorithm development.

11.3.1 Convolutional Neural Networks and Fully Convolutional Networks

Convolutional neural networks (CNN) are of interest in computer vision tasks (i.e., segmentation, detection, classification) as these learn the filters or kernels that were previously engineered for use in traditional approaches. These architectures are usually formed by stacking several types of layers (convolutional layers, pooling layers, fully connected layers, etc.) that transform the input (image) into the desired output. Convolutional, pooling, and fully connected layers are further defined in the following paragraphs (Fig. 11.3).

A convolutional layer contains a set of learnable filters or kernels. Each filter has a predefined size (width, height, and depth) which is generally spatially small when compared with the input image size. For example, a 3D image segmentation network may use an initial filter size of $3 \times 3 \times 3$ which corresponds to 3 voxels for width, 3 voxels for height, and 3 voxels for depth where width and height can be thought of pixels in an axial slice and depth provides information from slices below and above the central slice. During training, filters within convolutional layers are convolved about the input image computing dot products between the values of the filter and the input image at any position. As the filters are convolved throughout the whole input image, a convolutional layer generates an activation map that provides the responses of that filter at every spatial position. Through the training process, the network will initially learn filters that activate when they see specific visual features such as edges or textures, eventually learning more abstract patterns on higher layers of the network. Usually, many filters are used in a single convolutional layer, and each of them will produce individual activation maps which are then concatenated to produce the output for an individual convolutional layer. An important feature of convolutional layers is that they provide local connectivity between neurons of adjacent layers. This allows the networks to learn features both globally and locally, allowing the network to detect subtle variations in the input data.

Parameters typically defined for a convolutional layer include padding, stride, and dilation. Image padding (i.e., adding a new pixel(s) around the edges of an image) can be used to ensure that the input and output volumes of a convolutional layer have the same size spatially. The stride is the step the kernels take as they convolve about an input volume. For example, using a stride of $S = 1$ ensures that a filter is convolved on every pixel in an input volume, while using a stride $S = 2$ would only calculate activations for every other pixel. Using a stride $S > 1$ will result in a reduction of the output volume. To determine the effect of padding/stride on the output volume O, one can use the following relationship: $O = (W - F + 2P)/S + 1$, where W is the input volume size, F is the filter size, P is the padding used, and S is the stride. For a 17×17 input, a filter size of 3, padding of 1, and stride of 2, O would result in an output image of 9×9 [$(17 - 3 + 2\,(1))/2 + 1 = 9$]. Another hyperparameter for the convolutional layer is the dilation of the filters. Traditionally, filters used in imaging are continuous (dilation $D = 1$), meaning that the pixels used in the dot product of the convolution are all spatially next to each other. Filters using dilations that are greater than 1 ($D > 1$) can quickly increase the receptive field of

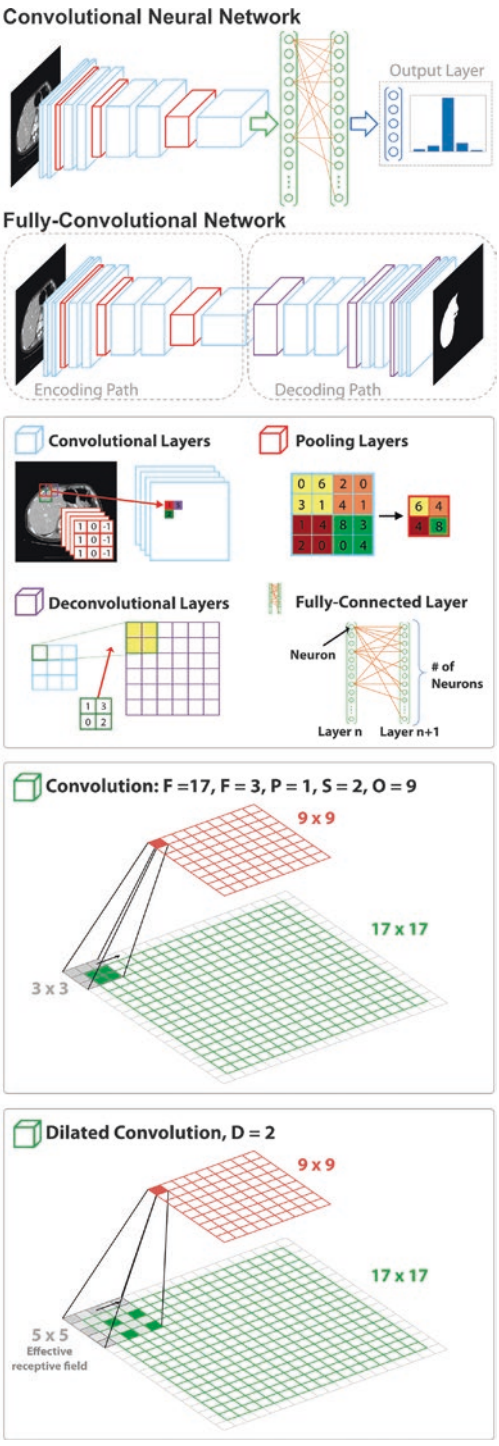


Fig. 11.3 Illustration adapted from Cardenas et al. [20] demonstrating the difference between convolutional neural networks and fully convolutional networks (encoder–decoder). The middle panel shows commonly used components of convolutional networks, whereas the bottom panels provide examples of how convolutional layer parameters affect the size of the layer's output size (including increasing receptive field for the layer)

the layer by using filters that have spaces between pixels during the dot product. Here, the receptive field is defined as the region in the input space that a particular CNN's feature is looking at (i.e., be affected by). For example, a 3 × 3 filter with dilation $D = 2$ has the effective receptive field of a 5 × 5 filter (*see* Fig. 11.3).

Pooling layers usually follow successive convolutional layers in a typical convolutional neural network. These layers reduce the spatial size of convolutional layers' output volumes to decrease the number of parameters and computations through the network. Pooling layers operate on each individual activation map from the previous layer to spatially resize resulting maps, typically using the max pooling operation. To further explain this process, we present the following example: a pooling layer with filters of size 2 × 2 convolved using a stride of 2 results in the downsampling of the input volume by 2 on each dimension. In this 2D scenario, the pooling layer removes 75% of the activations as the max pooling operation takes the maximum over 4 numbers (2 × 2 filter size). Other pooling operations such as average pooling or L2-norm pooling are available, but max pooling is often preferred due to its mathematical simplicity and reliable use in practice.

Fully connected layers are not commonly used in image segmentation nowadays but are important to understand as they are commonly used in CNNs for classification (e.g., feed forward/perceptron). All neurons in these layers are connected to all activations in the previous layer and can be computed with a matrix multiplication, making it faster to compute than convolutional layers. Fully connected layers are used in many popular CNN classification architectures as the last layers prior to the final output of the network; after feature extraction (via convolutional layers), the fully connected layers are able to classify the data into N defined classes by learning a function between the high-level features given as an output from the convolutional layers.

Initially, CNN architectures like AlexNet [18] and VGG [75] were investigated for segmentation purposes by generating individual pixel classifications. This approach was computationally ineffective as the same convolutions are computed several times due to the large overlap between input patches from neighboring pixels and resulted in pixel predictions which lacked correlation to neighboring pixels. Fully convolutional networks (FCNs) were introduced by Long et al. to overcome the loss of spatial information resulting from the implementation of fully connected layers as final layers of classification CNNs [76]. In this seminal work, the authors modified the VGG16 [75] and GoogLeNet [77] and replaced all fully connected layers with the fully convolutional layers. Using this approach results in the model producing a spatial segmentation map instead of individual voxel classification scores. Long et al. used skipped connections to concatenate feature maps from earlier layers and combined these with the final layers of the model which are upsampled back to the original image size to produce accurate and detailed segmentations. This work was one of the first to demonstrate that a deep learning network can be trained for semantic segmentation in an end-to-end manner.

Most architectures used for medical image segmentation are based on 2D or 3D variants of successful methods adapted from computer vision. Improvements in 3D convolution computation efficiency and hardware, in particular the fast increase in

available GPU memory, have promoted the extension of these methods to 3D imaging. Patch-wise architectures, those using 2D ($N_x \times N_y$) or 3D ($N_x \times N_y \times N_z$) patches centered around the voxels in an image, were introduced to address these bottlenecks. In this simple approach, patches extracted from the whole image, along with their corresponding label maps, are used to train the segmentation network. Several approaches (shift-and-stitch, fusion, etc.) are being used to combine individual patch segmentation probability maps to create dense outputs [76, 78, 79]. Some results have suggested that the performance of patch-wise architectures can be improved by using multi-scale inputs (multiple inputs with different patch sizes) which provide the network with global and local context [80, 81].

11.3.2 Popular Deep Learning Auto-Segmentation Architectures

The most popular medical image segmentation deep learning architecture is the U-Net [82] (Fig. 11.4). While previous works had already proposed the use of encoding and decoding paths to create dense outputs, Ronneberger et al. combined this approach with skip connections, which concatenate features from the encoding to the decoding layers [82]. Thus, higher resolution features from the encoding path along with the up-sampled features from the decoding path allow for the architecture to better localize and learn representations from the input image. Furthermore, the U-Net allowed for efficient end-to-end training, meaning that it did not require a pre-trained network as others had previously proposed, and showed that the network could be trained to produce accurate segmentations with very little labeled training data. The original 2D application of the U-Net was extended by Cicek et al. to allow the use of 3D images to train this network [83].

Other groups have introduced variants of the original U-Net architecture. Milletari et al. proposed the V-Net [84], a 3D version of the U-Net architecture that introduced the use of a Dice coefficient loss function and implemented residual learning [85] at each resolution stage (Fig. 11.4). Other variations of the U-Net include Hybrid Densely Connected U-Net (H-DenseUNet) [86], Res-UNet [87], and UNet++ [88]. Each adaptation from the original U-Net implements novel strategies which allow for better segmentation accuracy when compared to the original U-Net architecture. Dense U-Nets take advantage of dense (or "skip") connections between architecture layers; for example, H-DenseUNet uses repetitive densely connected building blocks (a building block is a group of stacked convolutional layers) where each block has connections to all subsequent layers [86]. Using dense connections has some advantages: the dense connectivity between layers results in fewer output dimensions than traditional networks avoiding learning unnecessary features. Also, having a dense path results in each layer receiving all the information learned by previous layers improving gradient flow which is essential when training and searching for an optimal solution in deep neural networks. Other dense U-Nets include Multi-scale Densely Connected U-Net (MDU-Net) [89], Distributed Dense U-Net (DDU-Net) [90], among others. For the Res-UNet, Diakogiannis et al. [87] replace the building blocks of the U-Net architecture with

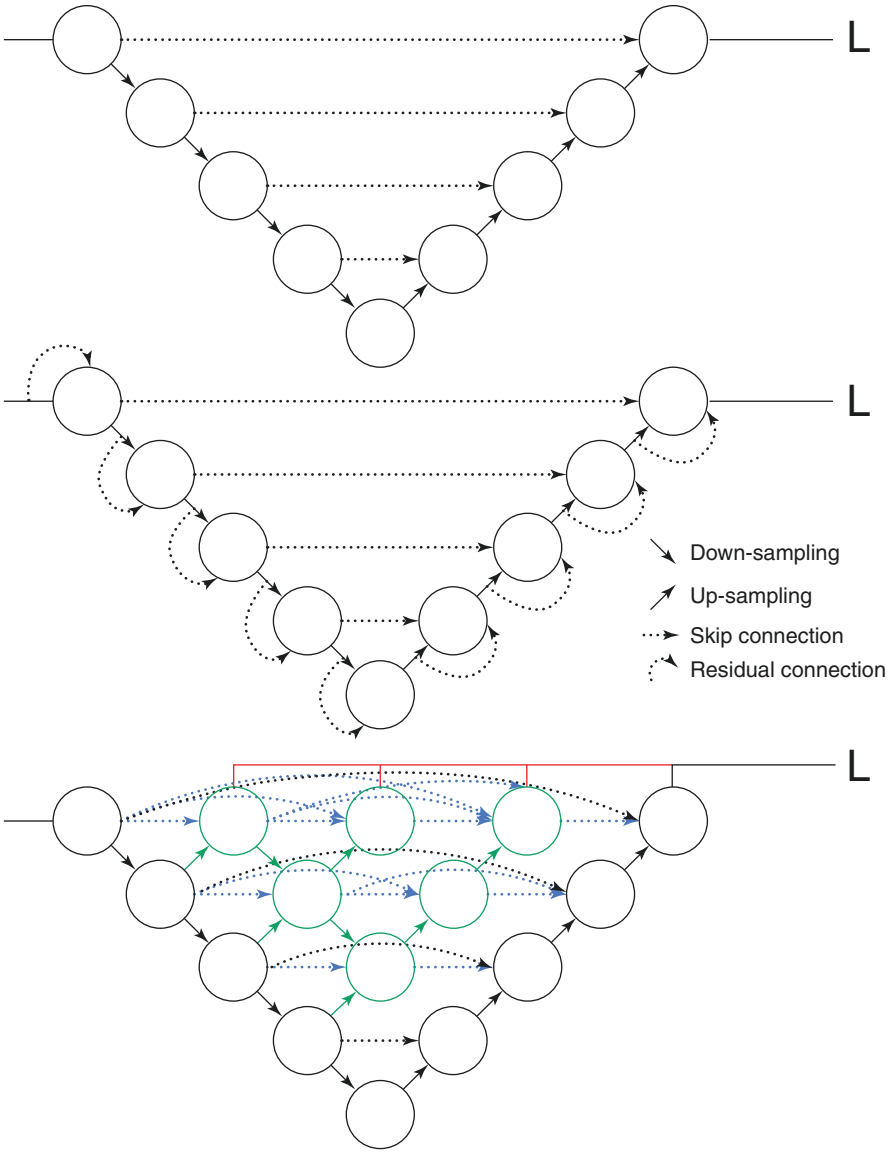


Fig. 11.4 Comparison between the vanilla U-Net [82] (top), the V-Net [91] (center), and the UNet++ [88] (bottom) deep learning architectures. The circles in this figure represent individual convolutional blocks. The V-Net builds on the vanilla U-Net by adding residual connections at each convolutional block. The UNet++ architecture uses dense connections (i.e., every convolutional block is connected to convolutional blocks down-stream) and deep supervision (shown by red lines). Here the UNet++ generates four predictions, and their individual losses are taken into consideration when training the model

modified residual blocks as those introduced by He et al. [85]; residual blocks are useful as they reduce inherent vanishing and exploding gradients which are commonly present in deep architectures. In addition, they use dilated (atrous) convolutions within each residual block to increase the receptive field of each layer. UNet++ combines previously introduced dense connections and deep supervision to further improve the traditional U-Net. In UNet++, the connections from the encoding path go through a densely connected block where the number of convolution layers depends on the pyramid level (green circles in Fig. 11.4). Deep supervision is a technique in deep learning where a model learns to generate representations of the output at multiple levels of the network and results in faster convergence. Deep supervision can be thought of as a way to improve the learning process; here, previous layers in the network are checked via deep supervision (i.e., evaluated against the physician-approved segmentations) to ensure that the learned features are passed on to subsequent layers and are useful in identifying additional features as the network becomes deeper. In this work, the authors design the UNet++ to generate full-resolution feature maps at multiple levels (shown by red lines in Fig. 11.4) each with individual loss metrics that focus learning at each level in the network.

Kamnitsas and collaborators [80, 92] introduced the DeepMedic architecture which used multi-scale 3D CNNs with fully connected conditional random fields [93] for brain lesion segmentation (Fig. 11.5). Their dual pathway architecture provided the network with local and more global context from the input images by using image patches at multiple scales simultaneously. Other groups have used this multi-scale approach to improve auto-segmentation results. For example, Roth et al. used a multi-scale pyramid of 3D FCNs for abdominal organ segmentation [94]. In this work, the authors use inputs from two scales (low- and high-resolution inputs) centered about the same voxel to improve normal tissue auto-segmentation [94]; here, two FCNs are used where the first network uses the low-resolution image and its output is resized and used as an additional input for the high-resolution image. Using this approach, the authors are able to provide global context from the low-resolution image while being able to perform fine-detail segmentations on the high-resolution input using end-to-end training.

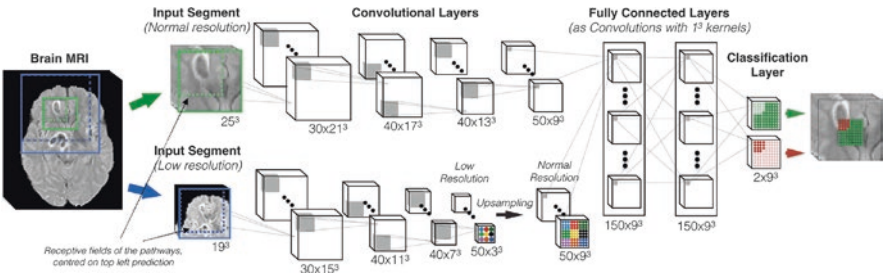


Fig. 11.5 DeepMedic architecture from Kamnitsas et al. [92] Multi-resolution inputs (normal [green] and low resolution [blue]) are used to learn local and global information about the region of interest

Dilated convolutional networks such as those in the DeepLab family and the Densely Connected Atrous Spatial Pyramid Pooling (DenseASPP) [95] have been popular in image segmentation. Developed by Chen et al., DeepLabV1 [96], DeepLabV2 [97], DeepLabV3 [98], and DeepLabV3+ [99] (the DeepLab family) are deep convolutional neural networks which, at the time of their publication, have been considered state-of-art segmentation architectures. In DeepLabV1, the authors used atrous convolutions to explicitly control the resolution at which feature maps are computed within the architecture. DeepLabV2 adds the use of atrous spatial pyramid pooling (ASPP) which allows for the network to segment objects at multiple scales with filters at multiple sampling rates which provide additional context for the segmentation task. The DeepLabV3 uses image-level features within the ASPP module to capture long-range information; furthermore, the authors also include batch normalization layers to improve training. DeepLabV3+ implements an encoder–decoder architecture which uses the DeepLabV3 framework as the encoder. In addition, dilated depthwise separable convolutions and pointwise convolutions are used throughout the network. Similarly, the DenseASPP [95] architecture takes advantage of densely connected atrous convolutional layers to effectively generate densely spatial-sampled and scale-sampled features in a deep network. An advantage of this network is that there are no pooling operations (no encoding/decoding paths) so each layer in the network captures features in the original resolution of the input image.

While generative adversarial networks (GANs) have found a niche in medical imaging in the generation of synthetic images such as generating synthetic CT images from MR scans, GANs have also been shown to produce high-quality results for segmentation tasks [100]. Auto-segmentation models that adopt GANs usually train a segmentation network (generator) which is coupled with an adversarial network that discriminates physician-approved segmentation maps from those segmentations generated by the segmentation network. The discrimination network here is used to see if the generator network can generate segmentations that "fool" the discriminator by creating segmentations that are indistinguishable from the manual contours (Fig. 11.6). Coupling these networks results in an end-to-end solution which can lead to more realistic segmentations. While GANs are very promising for image segmentation, they are unstable during training often leading to the model to not converge, the generator's collapse, or discriminator being too successful causing the generator's gradient to disappear resulting in a model that does not learn during the training process.

Recurrent neural networks (RNNs) have been previously explored to leverage contextual information for 3D image segmentation. RNNs are often used in junction with a convolutional neural network (e.g., U-Net) which generates segmentations on a 2D image providing intra-slice context where the RNN exploits inter-slice context to improve the accuracy of the automatically generated segmentations. Previously, Chen et al. [102] used a bidirectional contextual long short-term memory (BDC-LSTM) network, a type of RNN, which applied a sequence of 2D feature maps from adjacent 2D slices (i.e., $z - 1$, z, and $z + 1$) extracted from a series of 2D U-Nets which generated segmentation probability maps for each slice (i.e., partial

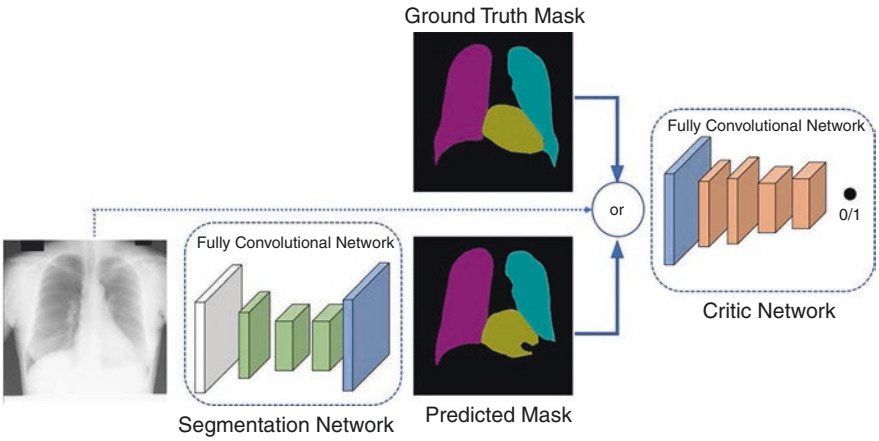


Fig. 11.6 Illustration by Dai et al. [101] of how GANs can be used for image segmentation. Here a fully convolutional network is used as a generator and the resulting predicted masks are compared to the physician-approved truth segmentations using an adversarial network to further refine the auto-segmentations

3D volume). This approach was popular in the early days of deep learning-based auto-segmentation as they overcame the need to use 3D architectures, which were computationally expensive with available GPU hardware, providing sequential object-connectivity (morphology) information from previously segmented 2D slices. RNNs found a niche in the automatic segmentation of serial (i.e., contiguous) structures and organs such as the heart vessels, gastrointestinal organs, and spine, where information from adjacent slices provides better connectivity through the 3D volume segmentation. Image segmentation RNNs are computationally expensive to train, and for this reason, their use has recently been declined for medical image segmentation.

Attention-based image segmentation architectures have recently gained popularity. Attention gates are useful as they can reduce redundant model parameters by allowing the network to focus on portions of an image which are more relevant while suppressing unimportant regions of the image. How attention is derived can vary among architectures; a key variant is the use of soft or hard attention. Soft attention is when the network calculates the context vector as a weighted sum of the encoder hidden states. Hard attention differs in that attention scores are used to select a single hidden state (e.g., using argmax function). Both hard and soft attention approaches have been explored, but soft attention mechanisms are often preferred as they are differentiable and can be updated via back propagation making them relatively easy to train. Attention gates incorporated into existing architectures like the U-Net (Fig. 11.7) can improve model sensitivity and accuracy to the resulting segmentations without significantly increasing the computational overhead. When using the U-Net as an example, attention gates are implemented before concatenation of skip connections and up-convolutional layers; by merging only relevant activations, gradients carrying information from the background class are

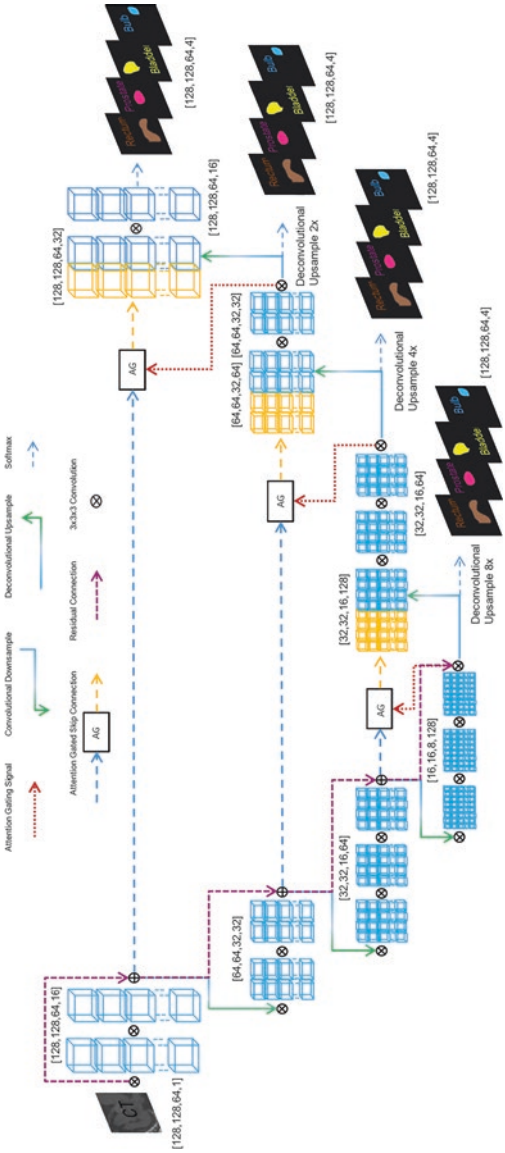


Fig. 11.7 3-D U-Net implementation by Kearney et al. [103] using attention gated skip connections and deep supervision

down-weighted during backpropagation allowing parameters in prior layers to be updated focusing on spatial regions of an image that are relevant to the segmentation task. More recently, channel attention modules have been explored. Here, the channels are the feature maps generated by a convolutional layer block. One can think of these channels (or feature maps) to have class-specific responses where different features are associated with individual segmentation classes/organs. Furthermore, combining spatial and channel self-attention modules has been shown to further improve segmentation accuracy compared to individual modules alone.

11.4 Image Segmentation Packages and Publicly Available Datasets

Open-source software and publicly available datasets have contributed and supported medical image research for decades advancing our knowledge and understanding of computational algorithms leading to breakthroughs in algorithm development and improvements for a variety of medical image analysis tasks including image segmentation. In this section, we provide an overview of medical image segmentation open-source software, highlight the role of publicly available datasets in driving medical image segmentation research, and discuss the automatic segmentation tools currently available in commercial systems.

11.4.1 Open-Source Image Segmentation Packages

Several open-source software tools have been developed for medical image segmentation over the past decade. The software tools described in the following paragraphs provide automatic segmentation algorithms that vary in levels of complexity including basic intensity thresholding techniques, semi-automatic approaches (active contours with user defined seeds), atlas-based algorithms, and, more recently, deep learning-based algorithms. The tools described in this section have been developed for research and are not intended for clinical use.

3D Slicer is a software platform which was developed for the analysis and visualization of medical images [104]. It supports multi-modality imaging including MRI, CT, ultrasound, nuclear medicine, and microscopy images. 3D Slicer is widely used in the medical imaging community as it provides plug-in capabilities for adding new algorithms and applications. Some examples of these plug-ins include the semi-automatic PET tumor segmentation [105], a fast implementation of the GrowCut method [106], and DeepInfer [107] which is a deep learning deployment toolkit extension. ITK-SNAP is a software tool which was designed with a focus on medical image segmentation [108]. Its interface provides tools for manual segmentation and image navigation which are complemented with a semi-automatic segmentation tool which uses active contour segmentation algorithms. The latest release of ITK-SNAP (3.8.0) introduced Distributed Segmentation Services (DSS) which is an architecture that allows segmentation tasks to be implemented as

services on the internet [109]. DSS provides a platform for segmentation algorithm developers to make their algorithms publicly available to ITK-SNAP users.

Plastimatch [110] is a software tool that was developed to perform volumetric registration of medical images (i.e., CT, MR, PET) and offers a system to perform multi-atlas-based segmentation (MABS). Within Plastimatch, a user can use MABS to prepare their own atlas, perform the segmentation, and tune parameters within MABS to further improve the quality of the automatic segmentations. A feature of MABS is that the user can select the registration strategy including algorithm and registration metric. Seg3D is another open-source software tool which was developed specifically for medical image segmentation [111]. Its interface allows the user to use basic semi-automatic segmentation tools such as intensity thresholding and level set segmentation, which uses seeds to find regions in a data volume that are similar to the original seed using statistics calculated in the seed region to determine the extent of volume expansion about individual seeds.

As deep learning-based segmentation research has ramped up over the past few years, several open-source platforms have emerged to increase the access and availability to these algorithms and trained models. The Deep Learning Tool Kit [112] (DLTK) is a neural network toolkit written in Python which uses TensorFlow [113], an open source library for development and training of machine learning models, as its backend. DLTK has a Model Zoo which offers a limited number of segmentation models. NiftyNet [114] is another open-source convolutional neural network platform built-on TensorFlow for research in medical image analysis and segmentation. NiftyNet provided an easy way to share networks and pre-trained models, as well as implementation of commonly used architectures (e.g., U-Net, DeepMedic), loss functions, and a comprehensive list of evaluation metrics for image segmentation. In early 2020, the developers of NiftyNet decided to transition developing efforts from NiftyNet toward Project MONAI (Medical Open Network for AI). MONAI [115] is a framework that uses PyTorch [116] for deep learning in healthcare imaging. While in its early days, MONAI provides tutorial examples for training volumetric image segmentation models. Eisen is another framework which implements an API that builds directly on PyTorch to enable simple and quick development and experimentation with deep learning models. Eisen [117] provides a feature where users can design experiments and define deep learning workflows by visually mixing (drag/drop) Eisen building blocks. Medical Image Segmentation with Convolutional Neural Networks [118] (MIScnn) is an open-source Python library which offers an API for medical image segmentation pipelines based on Keras with Tensorflow as its backend. It offers a variety of 2D and 3D segmentation architectures and commonly used loss functions for medical image segmentation model training. NVIDIA's Clara [119] is a healthcare application framework for AI-powered imaging, genomics, and for development and deployment of smart sensors. Through the Clara Train SDK, Clara allows for AI-assisted annotations and collaborative learning using techniques such as federated learning and transfer learning which enables researchers to collaborate and build AI models while keeping data private and secure. Furthermore, Clara offers Clara Deploy SDK which allows for the creation of application workflows for inference of AI-models.

11.4.2 Publicly Available Datasets

The availability of publicly available datasets with manually labeled segmentations has promoted advances in segmentation algorithm development. These datasets are generally published as part of "grand challenges" that are usually hosted by organizations such as the American Association of Physicists in Medicine (AAPM), the Medical Image Computing and Computer-Assisted Intervention (MICCAI) Society, and the International Society for Optics and Photonics (SPIE). These segmentation challenges allow participants to evaluate their algorithm's performance on a common benchmark image dataset. The Cancer Imaging Archive [120] (TCIA) has played a tremendous role in making medical imaging datasets available to the public. TCIA data is organized in patient cohorts or "collections" with many publicly available datasets providing RTSTRUCT files containing clinical contours or contours generated for a specific purpose (i.e., segmentation challenge). Similarly, the website https://grand-challenge.org (accessed 7/27/2020) provides a platform for the medical imaging community to upload medical imaging data easily and securely. Through this site, researchers around the globe are able to share medical imaging challenge data to promote algorithm development through scholarly competitions. Several popular segmentation challenge datasets such as the LiTS [121], BraTS [122], KiTS [123], SegTHOR [124], RT-MAC[125], PROMISE12 [126], Head And Neck Auto-Segmentation Challenge [50] are advertised through this website.

At this time, radiation oncology datasets are relatively sparse due to numerous barriers of data sharing. As a community, we should strive for the promotion and adoption of data sharing solutions across institutional and international borders. In regard to data sharing barriers in radiation oncology, Thompson et al. [127] suggests that "just as open-access publication is mandated for publicly funded research, perhaps FAIR-compliant DICOM-RT and matched clinical data publication could be a requirement of any publicly funded radiotherapy trials." Access to standardized datasets could further promote algorithm development and serve as a common ground for algorithm evaluation; however, several factors should be considered to ensure that curated datasets are representative of diverse clinical populations to reduce implicit biases. For image segmentation, it is important to have a representative distribution of patients within the training data and to clearly indicate this. For widely dispersed models, these should include patients with a variety of anatomical presentations (i.e., post-operational), patient setup positions (supine, prone, etc.), as well as demographic and clinical backgrounds. For models that are specific to one patient population, training data can be limited to that specific population, but this limitation should be denoted so the model is not incorrectly applied to other patient populations later. In addition, to achieve high robustness in auto-segmentations, it is important to consider the quality of the medical images. For example, factors such as image noise, spatial resolution, and low contrast could impact the quality of the resulting segmentations.

11.4.3 Commercial Systems

Tools from the first and second generation of auto-segmentation, such as intensity thresholding and region growing algorithms, are included in almost every commercial contouring software. The third- and fourth-generation techniques are more novel and still growing in their availability across platforms.

Raystation, Velocity, Elekta, and Philips offer model-based segmentation which works for a small number of vendor defined structures with adjustable shape, size, and property parameters for different organs at risk. There are many commercial systems that currently offer atlas-based contouring [128]. Mirada Medical offers WorkFlow Box which uses deformable image registration to automatically apply contours to CTs based on multiple expert atlases. MIM Software has MIM Maestro where one can use auto expert atlases or user defined atlases. These can be sorted by TNM status, lesion laterality, or physician. Multiple atlases can be selected to start the auto-segmentation, in which case, a structure set is created for each atlas and STAPLE is used for each contour. Elekta has Atlas-Based Autosegmentation (ABAS) which approximates the contours by scanning a library of reference images. The user can select an atlas from the library or choose to use the STAPLE algorithm. The user cannot see or edit the contours within ABAS, but the contours can be imported into other contouring software. Philips offers Smart Probabilistic Image Contouring Engine (SPICE) which is within the treatment planning system Pinnacle. SPICE uses a combination of rigid and deformable registrations together with probability-based structure refinements. Raystation offers the Anatomically Constrained Deformation Algorithm (ANACONDA) which uses a hybrid model that combines information from image intensities and anatomical information. Velocity has an atlas-based model that uses b-spline deformable image registration. It has both expert atlases and allows user-defined atlases [129]. Velocity allows a local deformable registration for individual structures which can allow for a better match before structure creation. Velocity also allows for user-defined exclusion areas, such as high-contrast artifacts due to dental work or arms, to allow for better registration. Varian offers SmartSegmentation within the treatment planning system Eclipse which is focused on CT-based auto-segmentation. SmartSegmentation provides an auto-segmentation solution based on a deformable image registration algorithm (enhanced form of the Demons algorithm combined with a multi-resolution approach [24]). Brainlab provides atlas-based contouring within its treatment planning software which offers contouring of anatomical and surgical structures [130, 131]. The reference atlas is built on the CT of a single patient with segmentation based on deformable image registration with active post-processing. Additional commercial atlas-based segmentation systems are IMAgo from Dosisoft, OnQ RTS from OSL, and MultiPlan from Accuray.

Currently, there are fewer commercially available deep learning-based segmentation options, although this is rapidly changing as many vendors are investing resources in developing these tools. Most of these are black-boxes systems with minimal information on their designs. Mirada's DLCExpert uses convolutional neural networks for the automatic segmentation of organs for various anatomical sites

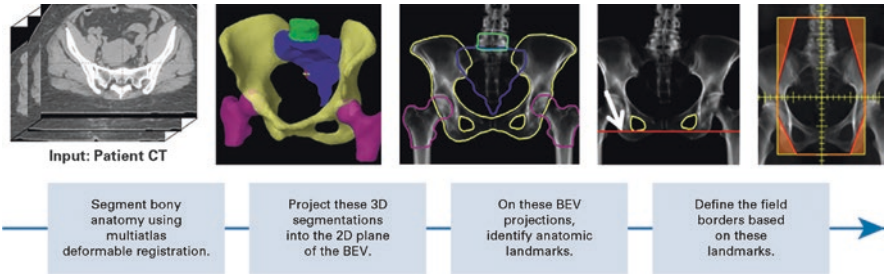


Fig. 11.8 Illustration from Kisling et al. [138] where auto-segmented bony structures are used to automatically define treatment fields for cervical cancer radiotherapy

[132, 133]. Raystation offers a deep learning solution for the thoracic region and male pelvis with the option for the user to develop their own model for other sites using transfer learning. Manteia offers a model that mixes a convolutional neural network with other traditional machine learning techniques and image registration. Microsoft's Project "InnerEye" uses decision forests and adversarial neural networks. Limbus Contouring (Limbus AI, Regina, Canada) offers deep learning-based contouring for some treatment sites including CNS, head and neck, thorax/breast and pelvis [134, 135]. AI-based automatic contouring is also being offered by some CT vendors (e.g., SOMATOM go.Sim by Siemens [136]).

The Radiation Planning Assistant (RPA, https://rpa.mdanderson.org) is a cloud-based service being developed at The University of Texas MD Anderson Cancer Center to offer a suite of fully automated contouring and radiotherapy treatment planning tools (Fig. 11.8). While the RPA is currently available for retrospective review of contours and radiotherapy plans [137–140], it is seeking its medical device clearance by the Food and Drug Administration to offer the RPA to low- and middle-income countries where access to radiotherapy is limited and less accessible. The RPA uses deep learning-based auto-segmentation models to automatically generate contours for a variety of treatment sites [141–146].

11.5 Auto-Segmentation Software Commissioning and Quality Assurance

When contours are used in the radiotherapy treatment planning process, any errors in the segmentation can have a serious impact on the patient treatment. According to AAPM TG 275, "Strategies for Effective Physics Plan and Chart Review in Radiation Therapy," two of the top ten failure modes in radiotherapy treatment planning result from "wrong" or "inaccurate" target contours [147]. Depending on the location and extent of the error, normal tissues (e.g., cord) could receive unintended doses, or targets could be under-treated. Thus, it is important to perform appropriate evaluation and commissioning of the auto-segmentation algorithm, routine procedural maintenance of the system, and patient-specific verification of the auto-segmentations.

11.5.1 Auto-Segmentation Evaluation

The commissioning process involves testing of the functions of a given piece of software and documentation of its different capabilities. The most obvious test for segmentation software is an evaluation of the accuracy of the segmentation, probably by comparison with manually drawn contours using overlap and distance metrics. An extensive review of overlap and distance metrics used for medical imaging segmentation quantitative analysis can be found in the publication by Taha and Hanbury [148]. In the next paragraphs, we provide a brief summary of the most commonly used segmentation metrics in the literature and present some recently proposed metrics (Fig. 11.9). It should be mentioned that quantitative metrics can be used to evaluate similarity between automatic segmentations and their corresponding physician-approved (manual) segmentations, but they could also be used to evaluate manual segmentations between multiple observers to quantify inter-observer variability and to evaluate multiple contours from a single observer to measure intra-observer variability (i.e., reproducibility).

The Dice Similarity Coefficient [149] (DSC) is the most commonly used overlap metric in medical image segmentation. The DSC is defined in Eq. 11.1,

$$\mathrm{DSC} = \frac{2\left|A \cap B\right|}{\left|A\right| + \left|B\right|} \tag{11.1}$$

where $|A|$ and $|B|$ are the volumes from contours A and B, respectively, and $|A \cap B|$ denotes the intersection volume between contours A and B. The DSC has values that range between 0 and 1, with a score of 1 meaning there is perfect overlap and a score of 0 meaning that there is no overlap between the segmentations. The Jaccard Index [150] (JI) is similar to the DSC where values range from 0 to 1 (1 for perfect overlap); the JI measures the intersection ($|A \cap B|$) over the union ($|A \cup B|$) between two contours (Eq. 11.2).

$$\mathrm{JI} = \frac{\left|A \cap B\right|}{\left|A \cup B\right|} \tag{11.2}$$

The JI is also referred to as the conformity index (CI) or the Intersection over Union (IoU). Both the DSC and JI are sensitive to volume sizes with larger volumes generally resulting in DSC and JI values closer to 1.

When considering surface distance metrics used to compare two segmentations, the mean surface distance (MSD), often referred to as the Average Symmetric Surface Distance (ASSD) [54], and the Hausdorff distance [151] (HD), often referred to as the Maximum Symmetric Surface Distance (MSSD), are routinely used in medical image segmentation studies. These metrics calculate the minimum distance (typically the Euclidean distance) between finite points on two closed surfaces (Eq. 11.3). Here, for a volume A, distances are calculated from each point (a) to each point (b) in volume B, resulting in a vector of distances ($d(A,B)$) equal in length to the number of points on the surface of volume A.

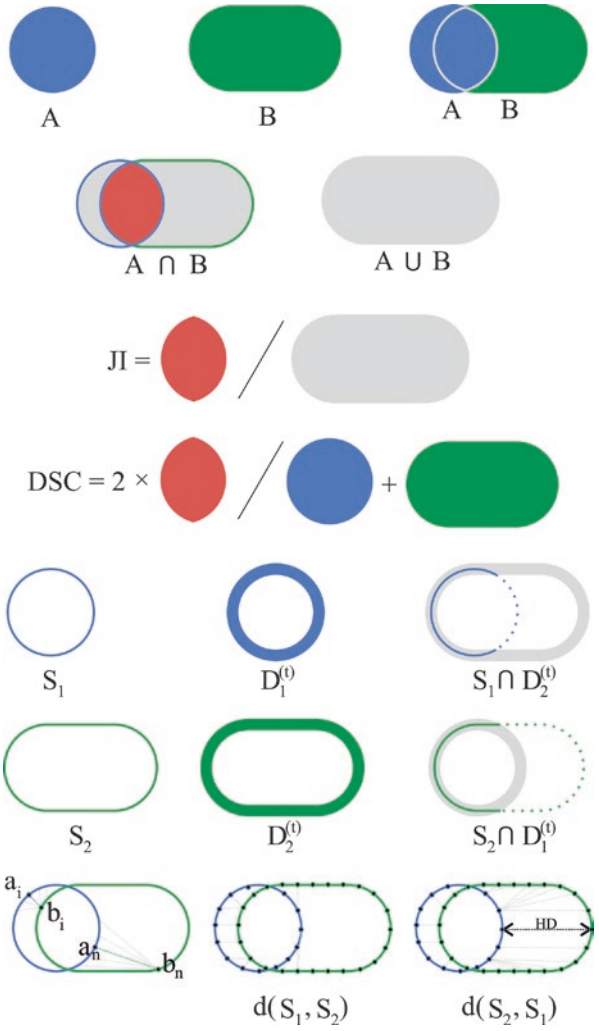


Fig. 11.9 Visual representation of the overlap and distance metrics defined in Eqs. 11.1–11.5, as well as the surface Dice metric

$$d(A,B) = \min a - b \tag{11.3}$$

Generally, these distances are calculated in both directions (i.e., $A \rightarrow B$ and $B \rightarrow A$) and the resulting list of distances from both surfaces are used to calculate both MSD and HD. The MSD and HD are defined in Eqs. 11.4 and 11.5, respectively.

$$\text{MSD} = \frac{1}{2}\left(\frac{1}{N_a}\sum d(A,B) + \frac{1}{N_b}\sum d(B,A)\right) \tag{11.4}$$

$$\text{HD} = \max\left(\max\left(d(A,B)\right), \max\left(d(B,A)\right)\right) \tag{11.5}$$

A limitation of the HD is that it is sensitive to outliers. For this reason, the 95th percentile HD (95HD) is often used in the literature (Eq. 11.6).

$$95\text{HD} = \max\left(\text{percentile}\left(d\left(A,B\right),95\right),\text{percentile}\left(d\left(B,A\right),95\right)\right) \tag{11.6}$$

Here, the 95th percentile is calculated on the vector of distances calculated for both surfaces.

More recently, Nikolov et al. introduced the surface Dice as a new metric to compare two segmented volumes [152]. Here, each volume's surface (S_1 and S_2, Fig. 11.9) is dilated using a predefined tolerance value (τ). The dilated surfaces ($D_1^{(\tau)}$ and $D_2^{(\tau)}$) are then used to calculate the overlap between the non-dilated and opposite dilated surfaces (i.e., $S_1 \cap D_2^{(\tau)}$). An advantage to using the surface Dice is that it allows for generating organ-specific tolerance values which can be generated using manual contours from multiple observers (Nikolov et al. used 95th percentile of the distances collected across multiple segmentations [152]); however, there is the requirement of generating the tolerance values for each organ, and these may be dependent on a variety of factors such as interpreter's experience, image modality, image quality, etc. which may be a limiting factor when using this metric to compare studies in the literature.

Lastly, the added path length (APL) was recently introduced by Vaassen et al. as a metric to evaluate manual adjustments made to an auto-segmentation [133]. In this study, the authors show that the path length of a contour that had to be added (whether by shrinking or expanding a volume) during manual edits of an auto-segmented organ required to meet institutional contouring guidelines was closely correlated with the required time necessary for the manual adjustments.

11.5.2 Patient-Specific Evaluations

Currently, there is no specific guidance for patient-specific evaluations of auto-contours as the integration of auto-segmentation systems in clinics has not been widely adopted. As the use of these tools increases, there is a need for guidance until such AAPM report is published. Previous AAPM reports can be utilized to fill this gap. AAPM TG-132 [153], which is on image registration in radiotherapy, employs concepts that are directly applicable to auto-segmentation. AAPM TG-132 states, "For initial commissioning of an image registration system, quantitative validation is required; however, for patient-specific evaluation of image registration, quantitative verification is not always possible due to limited time and resources and difficulty in determining the ground truth." Similarly, for auto-contours, the ground truth contour is not available for reference. Therefore, these contours must be qualitatively evaluated by the dosimetrist, physicist, or physician. During the commissioning process specific guidance on the amount of deviation each contour can have should be established. Auto-contours may differ from manual contours, but these differences should be minimal, resulting in only cosmetic change of the contour with little to no dosimetric impact for most contours, otherwise the tool is not an

appropriate fit for the clinic. It can be a time-consuming activity to edit the auto-contours; therefore, the guidance established at commissioning will determine which contours need to be manually edited during the qualitative review.

Automatic quality assurance of auto-segmentations has also been investigated [141, 154–156]. These measure ROI-specific characteristics (centroid, volume, shape, etc.) and use statistical approaches to determine any large deviations in segmented volumes. Another suggested approach, for example, could use the results of a primary segmentation algorithm and compare these to a secondary verification algorithm [141]. This approach requires the two algorithms to be independent, as the assumption is that they will fail in different ways. Although this approach does not replace the need for careful review of contours by the attending physician, it may help flag cases that will require extra attention.

11.5.3 Commissioning and QA

Evaluation for clinical use, however, involves a more comprehensive evaluation than quantitative analyses alone. The commissioning process should include extensive testing with patient data from the local institution, to ensure that the software works as expected for their range of image types, patient anatomies, etc. Additionally, it is important to ensure that segmentations created within one software tool are exported/imported properly to other systems, with all segmentations' information being transferred consistently and accurately to the treatment planning system. If the segmentation does not work sufficiently, accurately, or reliably for any of these combinations, then this limitation should be clearly documented so that the users are aware, and vendors can address these issues. Lastly, it is important to train users so that they understand the potential and limitations (risks) of the auto-segmentation software.

AAPM TG-53 provides information on QA for the TPS [157]. It details many QA checks for the treatment planning process and not just the system itself. It specifically addresses tests that should be performed in Tables 3–4 and 3–5 (see AAPM TG-53). MPPG 5a, which discusses commissioning of the TPS, emphasizes that representative patient cases should be utilized in commissioning [158]. Similarly, for auto-segmentation, representative patients should be selected for use in commissioning with a subset of these used for routine QA, as suggested by MPPG 5a. Such site-specific (e.g., abdomen, thorax) uses of auto-segmentation tools should be commissioned individually. Likewise, if the auto-segmentation tool is intended for single or multi-imaging modality use, the tool should be evaluated on the appropriate imaging modality or modalities (e.g., CT, CBCT, MR). Contour similarity and dosimetric comparisons should be performed during commissioning to establish guidance for clinical use. Two of the quantitative metrics posed by TG-132 for commissioning and QA involve the use of contours which makes them directly translatable to auto-segmentation evaluation [153]. These are mean distance to agreement (which is the monodirectional equivalent of the MSD, i.e., $A \rightarrow B$ only) and DSC which should be used to determine the contour similarity between manual

contours and auto-contours. The manual contours used for this should be performed in the typical workflow of the clinic, whether that is dosimetrists, physicists, or physicians contouring the anatomy. An easy way to achieve this is to select previously treated patients. The tolerance suggested by TG-132 for these two quantitative metrics can be applied here: the mean distance to agreement should be "within the contouring uncertainty of the structure or maximum volume dimension (~2–3 mm)," and the DSC should be "within the contouring uncertainty of the structure (~0.80–0.90)."

For dosimetric comparisons, plans generated using the manual contours and the typical workflow should be used and then re-calculated on the auto-contours. Qualitative DVH analysis can be performed along with specific clinical end points for each OAR (e.g., mean dose to parotid) to determine if the auto-contours provide an accurate and reasonable representation. Plans should also be created using the auto-contours and then compared to determine if there is an impactful difference in the optimization due to their use. Using previously treated patients will also make this step easier as there will be no need to create the treatment plan on the manual contours as this will already be present. The results from the contour and dosimetric evaluations should allow for specific guidance on necessary edits to auto-contours (e.g., identification of cosmetic changes that would not need to be used). This should be clearly documented and presented during training to those who will be clinically using the auto-segmentation tools. One method to ensure accurate performance is end-to-end testing. If auto-segmentation tools are to be used within the clinical care path (either during treatment planning or online during image-guidance or adaptive treatment), incorporating the use of such tools into end-to-end testing is desirable, yet not currently available for auto-segmentation evaluation.

After commissioning of the auto-segmentation tool has been accomplished, appropriate tolerance and action levels should be established (as a function of site, imaging modality, and clinical endpoint). The evaluation of these tools should take place through an ongoing, periodic QA program. Further evaluation should take place upon update of the auto-contouring software, as changes in performance may impact the quality of clinical segmentations. Also, it is important to note that changes in simulation procedure (e.g., patient positioning, use of new simulation devices) may have unintended consequences upon the performance of auto-segmentation tools. Thus, any changes in imaging or simulation protocols should be accompanied with the evaluation of tool performance using established tests from the QA program. The results of the QA tests may indicate whether the software update or procedural modification yields results within acceptable levels. Because contouring can cause major failures modes in radiotherapy, performing a failure modes and effects analysis as described in AAPM TG-100 can further elucidate potential failures modes [159]. Such a process is highly encouraged when incorporating new operations into the clinical care path, as new technology may incorporate new failures modes.

Once the commissioning process is complete, and the physicist has established a QA program, the software can then be released for clinical use. It is essential to train users on the auto-segmentation software and to highlight potential limitations and

risks identified during the commissioning process. Some routine maintenance is necessary afterwards to ensure that the software continues to perform in a consistent manner. The focus on quality is, however, now performed on an individual patient-by-patient basis. All segmentations should be carefully reviewed and approved by the local clinical staff (e.g., radiation oncologists) before use in a treatment plan. During the initial stages of deployment, the output of the automatic segmentation software should be treated as if a trainee had performed the contouring—that is, it is probably a reasonable starting point, but careful review is essential. The benefits of peer-review assessment through quality assurance contouring rounds have been previously reported [160–162], and establishing similar practices to assess auto-segmentation results, even for algorithms that have been shown to give excellent segmentations, could ensure the overall safety of the radiotherapy treatment.

11.5.4 Current Limitations to Auto-Segmentation Algorithm Development and Implementation

There are several important limitations to auto-segmentation algorithm development and use. Many data-related challenges are presented in auto-segmentation applications, especially the requirement of high-quality segmented datasets. Auto-segmentation algorithm performance approaches depend not only on the quantity but also on the quality of the segmentations (i.e., prior knowledge) used to train or develop a model. This limitation could be addressed through standardization of manual contours via the adoption of established international consensus guidelines. A reduction in inter- and intra-observer contouring variability could further improve the prediction accuracy of an existing model.

Many algorithms provide very little interpretability to understand what features (anatomical and/or image intensity-based) play an important role in generating an automatic segmentation. Multi-atlas-based segmentation and deep learning-based segmentations suffer from this limitation which may hinder the ability of researchers or end-users to fully understand and identify the cause behind inaccurate segmentations.

Individual algorithms are subject to their own limitations. For the first- and second-generation techniques, the contrast and noise within an image can play a large role in the performance of segmentation. For atlas-based approaches, the quality of the automatic segmentations is closely related to the performance of the registration algorithm. If the registrations between atlas patients and a reference patient's image are poor, one would expect the atlas-based approach to generate poor quality segmentations. Deep learning approaches can be subject to overfitting, which often happens when a model captures patterns in the training set with much higher accuracy compared to the accuracy of the model's predictions on unseen data. One of the leading causes of overfitting in deep learning-based image segmentation is the use of sample datasets which are not representative of the larger patient population. There may be a wide variety in position and shape of the internal organ which may be dependent on how the patient is set up during image acquisition,

patient's clinical presentation (e.g., anatomical and density changes due to pathologies), and/or prior interventions such as surgeries.

Lastly, variations in image acquisition protocols could potentially affect the performance of an auto-segmentation algorithm. Ger et al. scanned a radiomics phantom on 100 CT scanners using the local head protocol, local lung protocol, and a controlled protocol [163]. While only 20% of the scanners were within the radiation therapy department, there was a large variety in the settings used for the head and lung protocol scans across which would lead to differences in noise and partial volume effects. The controlled protocol could reduce variability in imaging features by over 50%. However, such a standardized protocol for patients is not something that will be seen in the larger community as each institution optimizes scanning protocols for their scanner, thus the variability in the local protocol scans is a realistic snapshot of the variability. This could potentially make some models not transferable due to using training data only at one institution that uses a specific scanning protocol that is significantly different than another institution. Recently, Huang et al. demonstrated the potential impact of several of these imaging protocol settings on atlas-based and deep learning-based auto-contouring [164]. This work demonstrated that deep learning-based auto-segmentations may be more robust to changes in imaging protocol than atlas-based approaches.

11.6 Overview of State-of-the-Art Results in Medical Image Auto-Segmentation

The following section provides a summary of the state-of-the-art performance from recently published auto-segmentation works focusing on normal tissues, as well as tumors and clinical target volumes.

11.6.1 Normal Tissues

11.6.1.1 Craniospinal

The human central nervous system is composed of the brain and spinal cord. Two of the most prominent volumetric imaging modalities used for anatomic imaging of the central nervous system are MRI and CT. In general, MRI has superior soft tissue contrast to that of CT and can distinguish between different tissues by exploiting differences in relaxation times. To isolate the brain from non-brain structures (e.g., skin, skull, eyes), "skull stripping," or brain extraction, is an essential step for many neuroimaging workflows, morphological studies, and studies involving clinical diagnoses. Many prominent automated approaches for skull stripping have been developed to include mathematical morphology-based methods, intensity-based methods, deformable surface-based methods, atlas-based methods, and hybrid methods and are discussed at length by Kalavathi et al. [165] However, such methods can fail depending on the type of MRI scan and pathology present in the scan. To address such failures, three-dimensional deep learning methods using CNNs are

emerging to remedy skull stripping failures to improve the robustness of skull stripping across various imaging modalities and pathologies. When these deep learning methods are evaluated on publicly available skull-stripped databases from the Preprocessed Connectomes Project [166], mean DSC values, sensitivity, and specificity exceed 0.985 for extracted brains [167, 168].

In 2013, the MRBrainS13 workshop was launched through MICCAI to segment white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) using 3T multi-sequence (T1w, T2-FLAIR, and T1-IR) MRI scans. The leading team was able to segment WM, GM, and CSF with DSC mean values 0.88, 0.84, and 0.78, respectively [169]. To segment other sub-structures within the brain, Moeskops et al. used a CNN architecture with 9 branches to input T1w, T2-FAIR, and T1-IR scans and output segmentations for WM, cortical gray matter (cGM), basal ganglia and thalami (BGT), cerebellum (CB), brain stem (BS), lateral ventricular cerebrospinal fluid (lvCSF), peripheral cerebrospinal fluid (pCSF), and white matter hyperintensities with presumed vascular origin (WMH) [170]. Using this approach, mean DSC values were 0.87, 0.85, 0.82, 0.93, 0.92, 0.93, 0.76 for each structure, respectively. Later in 2018, the MRBrainS18 Challenge was won by Luna et al. who designed a three-dimensional patchwise U-Net with transitional layers to segment GM, BGT, WM, WMH, CSF, ventricles, cerebellum, and brain stem with mean DSC values of 0.86, 0.83, 0.88, 0.65, 0.84, 0.93, 0.94, and 0.91 respectively [171].

When radiotherapy treatment planning of the brain is performed, MRI is advantageous and allows for accurate contouring of the brain and its sub-structures. However, since MRI image voxels do not contain the relevant electron density information necessary for heterogeneous dose calculation, CT images must be registered to MRI scans to facilitate dose calculation. To evaluate the dose to the brain and its sub-structures, contours made using the MRI can be mapped to the CT, or contours can be generated using only the CT. In scenarios where MRI scans are contraindicated or cost prohibitive, cross modality synthesis using GANs has been investigated by many authors to create synthetic MRI scans from CT scans or vice versa [100]. Such work may soon allow for contours made on one imaging modality to be directly mapped to synthetic scans for use in adaptive radiation therapy or image guidance. Although interest in automatically contouring cranial structures from CT has historically been limited due to poor soft tissue contrast, recent studies using atlas-based approaches have been performed [172, 173] (Fig. 11.10). Current deep learning approaches that automatically segment the brain and brainstem directly from CT images are described in the following section (*see* Sect. 11.6.1.2).

Segmentation of the spinal cord in MRI images has many applications in the study of neurological diseases. Changes of size or shape in the spinal cord are known to be correlated with changes in cortical activity and can also gauge disability [174, 175]. To characterize the state-of-the-art performance in gray and white matter segmentation within MRI images of the spinal cord, the Spinal Cord Grey Matter Segmentation Challenge was organized [176]. A recent approach from Perone et al. using deep dilated convolutions to segment gray matter from the spinal cord brings mean DSC on this public dataset to 0.85 [177]. Spinal cord is also readily contoured on CT images for the purposes of radiotherapy treatment

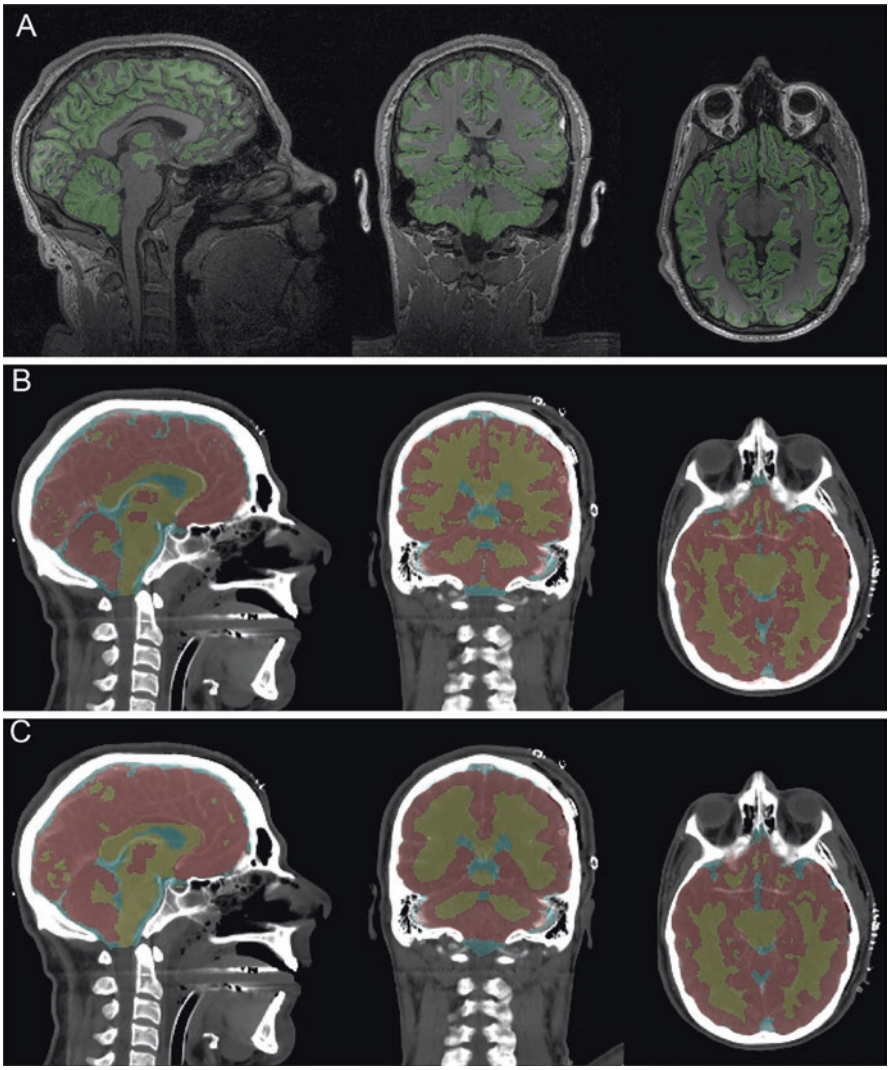


Fig. 11.10 Illustration from Irimia et al. [172] showcasing the impact of image resolution on automatic segmentations (row B vs. row C). The authors evaluated models using full resolution ($1 \times 1 \times 1.25$ mm^3, row B) and down-sampled resolution ($1 \times 1 \times 3.75$ mm^3, row C) CT scans. As expected, the segmentations on the down-sampled images were limited due to the lower spatial resolution resulting in lower quality segmentations. To further visualize the segmented gray matter, the authors provide T1-weighted MR images (row A); here, the white matter segmentations are not outlined to appreciate regions of contrast between the white/gray matter

planning. Because the spinal cord is a serial organ and has dire consequences for patients when dose tolerances are exceeded, accurate contouring of the spinal cord is essential. According to a review on auto-segmentation by Cardenas et al., numerous deep learning studies report DSC values for spinal cord spanning from 0.82 to 0.96 [20].

Surrounding and protecting the spinal cord is the vertebral column. To perform medical image-related tasks such as radiotherapy treatment planning, oncologic surgical staging, image-guided intervention, or clinical diagnoses, specific regions of the spine must be localized (i.e., labeled) and/or segmented. Assessment of soft tissue complications or involvement of the spine can be visualized with MRI, while assessment of osseous integrity or spatially accurate three-dimensional morphology is best visualized with CT imaging [178, 179]. In addition to localization and segmentation being time-intensive tasks, challenges related to anatomic abnormalities (e.g., scoliosis, atypical vertebral counts) and variations in imaging protocol (e.g., field-of-view, slice thickness, patient orientation) make automation of localization and segmentation difficult. Recent advancement in these areas have been driven by the availability of high-quality datasets, international challenges, multi-center studies, and innovations in machine and deep learning.

Datasets and challenges archived by SpineWeb (http://spineweb.digitalimaginggroup.ca/) contain those from CT, MRI, planar X-ray, and other imaging modalities. Such challenges address vertebra localization and segmentation on CT [180, 181], intervertebral disk localization and segmentation using MR, and numerous others. Many of these competitions have been hosted though MICCAI or annual CSI Workshops. To provide a common benchmark for researchers to develop accurate algorithms, Sekuboyina et al. organized the Large-Scale Vertebrae Segmentation Challenge (VerSe) to be held in conjunction with MICCAI 2019. VerSe provides the largest collection of publicly available CT scans (160 image series of 141 patients; annotated masks of 1725 vertebrae) to date [182]. From this competition, the approach by Payer et al. [183] using a spatial configuration-net plus U-Net obtained a DSC of 89.8% and identification rate of 94.2% across all vertebral bodies in the spine for the test set [184].

One challenge of deep learning-based localization is finding an effective methodology to encode long-range contextual information from distant anatomic landmarks into the model. To automatically label vertebral bodies in CT images, authors have used a myriad of approaches by incorporating multi-view frameworks [142, 185, 186] (Fig. 11.11), recurrent networks [187, 188], hidden Markov models [189, 190], or other classification techniques [191–193] into CNN for localization. Many localization networks, such as the approach by Payer et al., are used in two-step approaches, where localization, and then segmentation, is performed [183]. In other instances, networks perform both localization and segmentation using one framework [189, 192, 194]. One example of such an approach was by Lessmann et al. where their FCN performs iterative instance segmentation, regression of anatomic labels, and visibility prediction [192]. This approach was validated on both MRI and CT images and obtained second place in the VerSe 2019 challenge with DSC of 85.85% and identification rate of 89.9%.

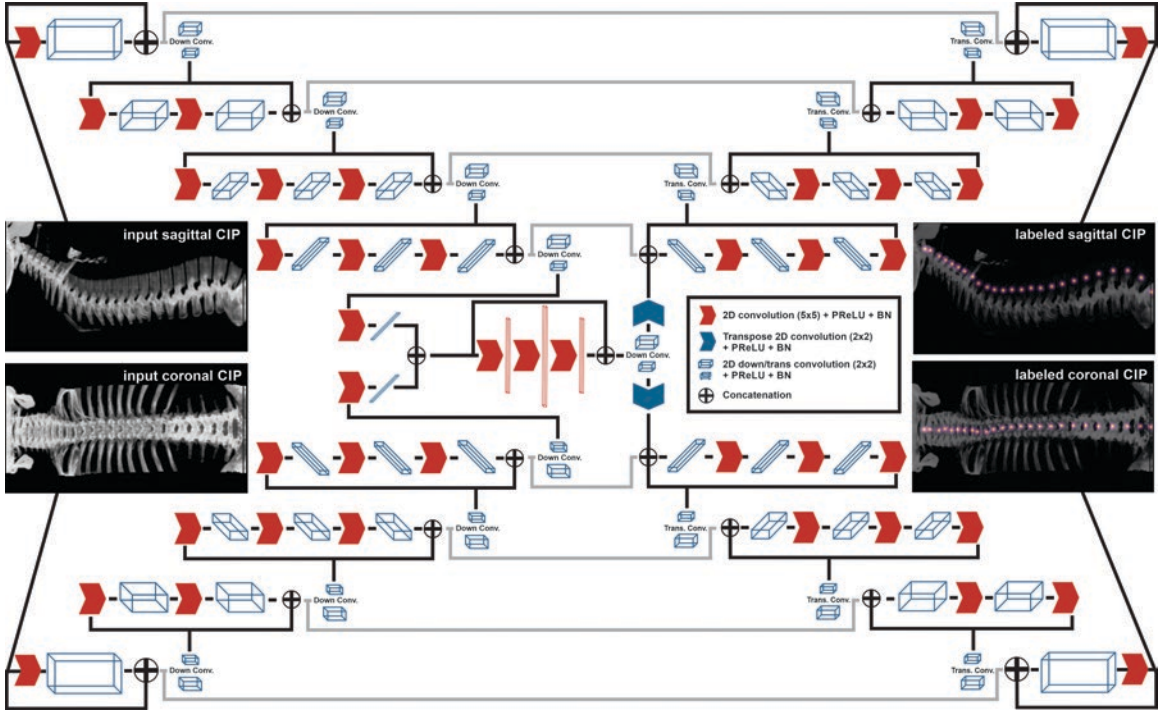


Fig. 11.11 X-Net architecture by Netherton et al. [142] used to automatically label vertebral bodies by automatically segmenting individual vertebral bodies' center of mass. This model uses a multiview input where sagittal and coronal intensity projection images are generated from a patient's CT scan

11.6.1.2 Head and Neck

Head and neck cancer radiotherapy treatment planning is considered one of the most complex sites due to the large number of organs at risk found in the head and neck region. The development of an auto-contouring tool can alleviate the complexity of the process by saving a significant amount of time contouring these tissues. Because of this, many types of research on automation of contouring head and neck normal tissues have been conducted by various groups. Furthermore, there were numerous challenges and competitions to investigate the best approach for developing an auto-contouring tool through the AAPM or MICCAI annual meetings [49, 50, 195].

CT is the primary imaging modality used for the head and neck radiotherapy planning process, and therefore, the majority of the auto-contouring tools for head and neck normal tissues have been developed for CT images. Nikolov et al. [152] trained a 3D U-Net to auto-contour 21 head and neck normal tissues (brain, brainstem, cochleae, lacrimal glands, lenses, lungs, mandible, optic nerves, orbits, parotids, spinal canal, spinal cord, and submandibular glands) and achieved mean DSC values between 0.57 and 0.99 (Fig. 11.12). They acquired CT images from TCIA [120] and created manual contours from radiation oncologists and radiation therapy technologists to test their model, and then made these contours publicly available so that other researchers can use them as benchmarking data. Rhee et al. [141] used a CNN-based classification architecture to limit the extent of CT slices in the craniocaudal direction and applied a segmentation model to auto-contour 16 normal tissues with the 3D V-Net [84] and achieved mean DSC values of 0.41–0.98. Wang et al. used a two-stage 3D U-Net framework [196] (Fig. 11.13), where the bounding box of the organ-of-interest was located on the first stage, and the fine segmentation was performed on the bounding box on the second stage. They achieved mean DSC values of 0.93, 0.86, 0.88, 0.76, 0.74, 0.4, and 0.45 for the mandible, parotids, brainstem, submandibular glands, optic nerves, and chiasm, respectively. Iyer et al. [197] developed an auto-contouring tool specifically for swallowing and chewing structures using the DeepLabV3+ in 2.5D, where the model input consists of three consecutive slices, and achieved mean DSC values of 0.88, 0.87, 0.83, 0.81, 0.80, and 0.67 for the right masseter muscle, left masseter muscle, larynx, left medial pterygoid muscle, right medial pterygoid muscle, and constrictor muscles, respectively.

Although head and neck radiotherapy planning fully on MR images is not common, the availability of online adaptive therapy with MR-LINAC increases the desire to develop an MR-based auto-contouring tool for head and neck normal tissues. Lei et al. [198] used a 3D Faster R-CNN [199] to segment eight normal tissues, and achieved mean DSC values of 0.89, 0.89, 0.85, 0.85, 0.84, 0.82, 0.81, and 0.79 for oral cavity, spinal cord, mandible, pharynx, esophagus, left parotid, right parotid, and larynx, respectively.

11.6.1.3 Thoracic

Examples of normal tissue structures in the thoracic region of the body include the breast, lymph nodes (e.g., supraclavicular, axillary, internal mammary), spinal cord, lungs, esophagus, and heart. These structures are commonly segmented on CT or

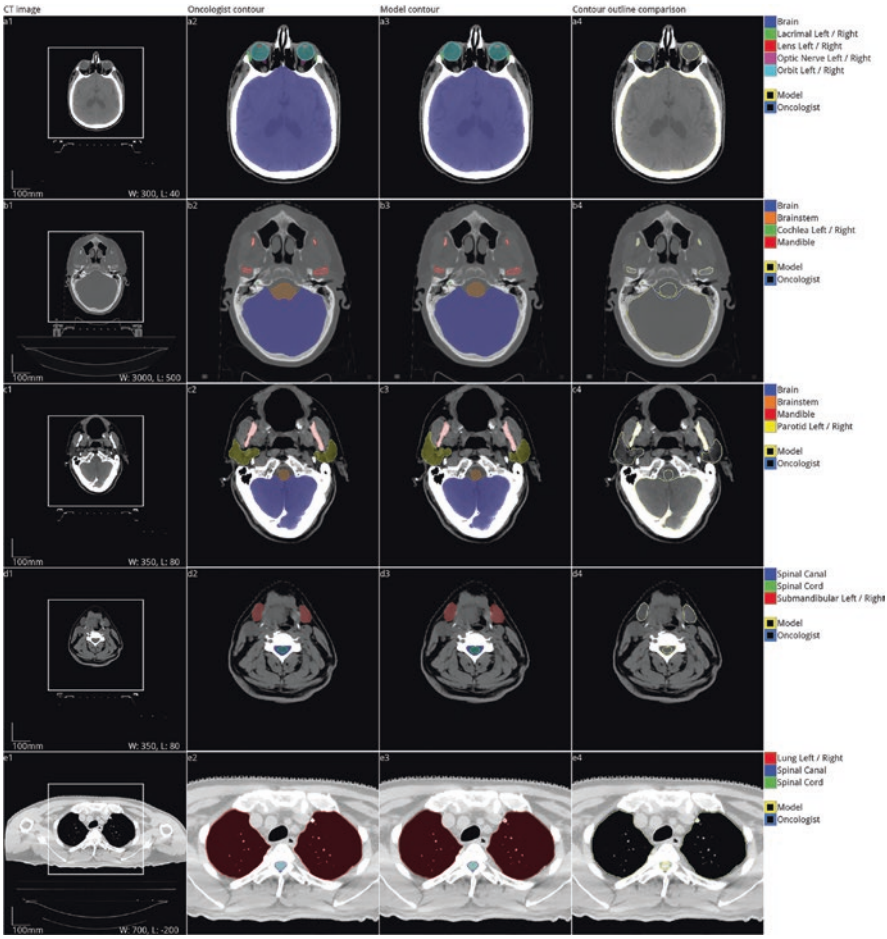


Fig. 11.12 Comparison between physician and auto-segmented contours from the work of Nikolov et al. [152] The authors can achieve human performance when evaluating head and neck normal tissue auto-segmentations

MRI images for the purposes of radiotherapy treatment planning. Guidelines for manually contouring these organs are often found in international consensus guidelines such as those from the North American-based Radiation Therapy Oncology Group (RTOG) and European Society for Radiotherapy and Oncology (ESTRO). To further increase consistency and efficiency in contouring, atlas and deep learning-based methods can be used to automatically contour normal tissue structures in the thoracic region for the purposes of radiation treatment planning. Although no gold standard library exists in any imaging modality for thoracic structures, various grand challenges hosted through professional organizations have established benchmark datasets for the evaluation of accurate auto-segmentation techniques. One such example was the AAPM 2017 Thoracic Auto-segmentation Challenge, in

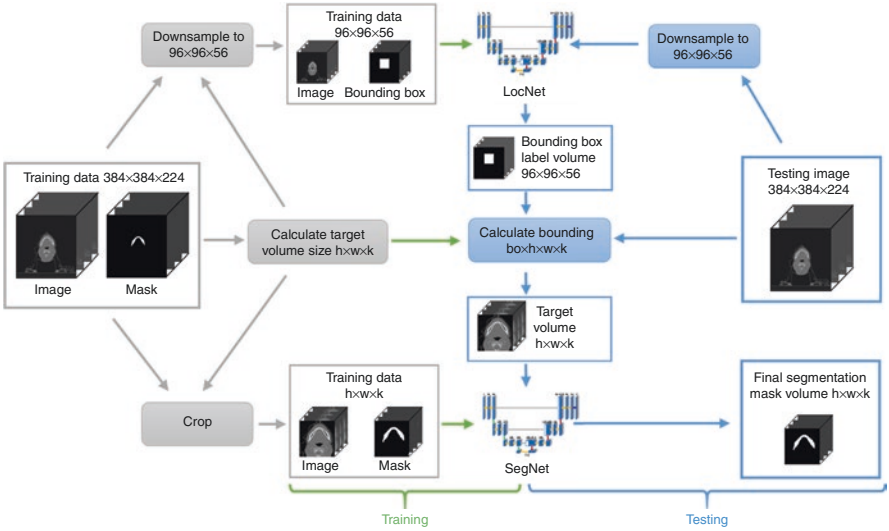


Fig. 11.13 Two-step approach used by Wang et al. [196] to auto-segment head and neck normal tissues. Here the authors used a localization network (LocNet) to find a bounding box volume to generate full-resolution segmentations (SegNet) used as final segmentations for individual organs at risk

which CT data and RTOG-guided segmentations (left and right lungs, heart, esophagus, and spinal cord) from multiple clinics were made available from The Cancer Imaging Archive (TCIA) for participants to train and test their auto-segmentation techniques [51, 120]. From the results of the competition it was found that deep learning outperformed atlas-based segmentation in accuracy and prediction time [51]. Other evidence for how deep learning-based segmentation methods outperform atlas-based segmentation methods has also been noted by other commercial tools such as MIRADA's DLC Expert for organs-at-risk in lung cancer [132]. The winning team from the Thoracic Segmentation Competition (Elekta) used a hierarchical segmentation approach to segment the lungs with a fast 2.5D residual U-Net, crop the area surrounding the lungs, and then segment the thoracic structures using a 3D model. Mean DSC and HD95% values were 0.97 and 2.9 mm, 0.97 and 4.7 mm, 0.93 and 5.8 mm, 0.72 and 7.3 mm, and 0.88 and 2.0 mm for left lung, right lung, heart, esophagus, and spinal cord, respectively. Using this same training and testing data from the AAPM Thoracic Segmentation Challenge, Dong et al. improved on previous U-Net-based methods by incorporating the use of GANs in a U-Net-GAN for multi-organ segmentation [200]. This method involved the use of a discriminator, was superior to U-Net alone, and produced mean DSC values for esophagus superior to those from the 2017 challenge as well as MSD and HD95 values superior to all previous contenders for all other structures [200].

Although the above-mentioned models performed well on test sets from the competition, out-of-sample input complications can be caused by predicting contours on images when differences in simulation procedure and patient positioning

are different from those used in the training set. This was demonstrated by Feng et al. when differences in patient abdominal position caused incorrect segmentations when using a previously trained network with the AAPM Thoracic Segmentation Challenge dataset [201] (Fig. 11.14). By incorporating as few as 10 cases from the other dataset, the accuracy and robustness of the model were restored to a performance level previously attained [202]. In addition, Schreier et al. has investigated how the use of multi-institution, single-single institution, and third-party trained models for contours used in breast cancer radiotherapy can be leveraged to create accurately trained deep learning models. The model used for this study was BibNet, a U-Net inspired architecture with residual and skip connections at every resolution level which demonstrated mean DSC values of 0.924, 0.929, and 0.951 for left breast, right breast, and heart, respectively [203].

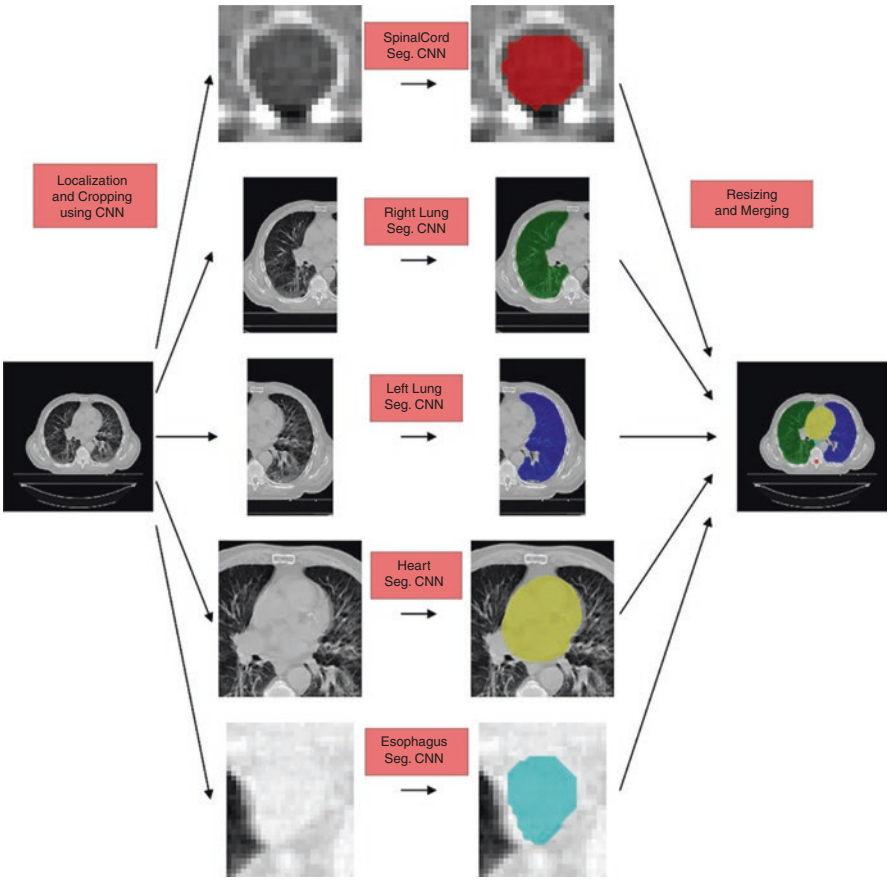


Fig. 11.14 Illustration of the two-step approach used by Feng et al. [201] Here the authors first identify a bounding volume within the CT scan which contains individual organs at risk, then using individual organ at risk models, the authors generate individual ROI auto-segmentations

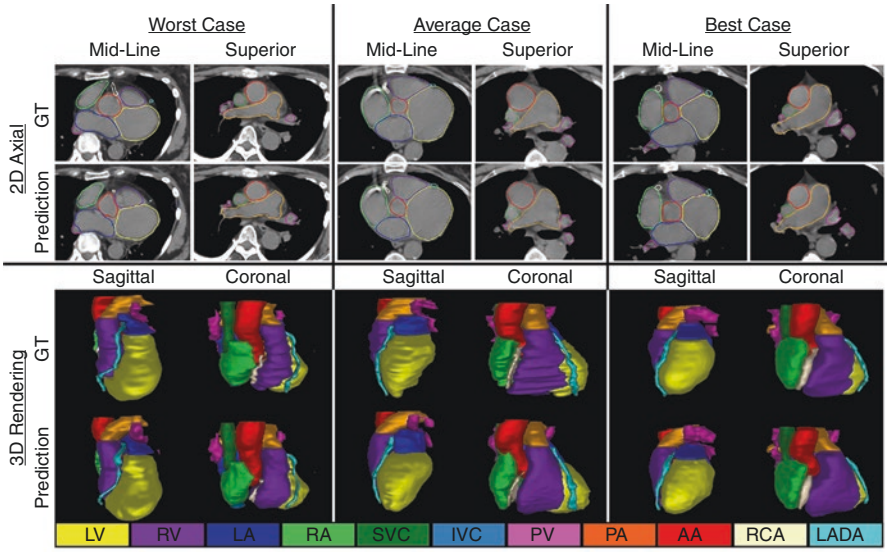


Fig. 11.15 Auto-segmented cardiac structures by Morris et al. [205] From left to right, the authors present the cases where the auto-segmentations were classified as worst, average, and best quality based on overlap and distance metrics when compared to the manual contours

The Multi-Modality Whole Heart Segmentation (MM-WHS) challenge, hosted in conjunction with MICCAI 2017, has provided researchers with multi-modality scans (60 CT and 60 MRI) with manually contour cardiac substructures. Of the twelve groups that submitted entries, it was found that CT-based models were, in general, better than MRI-based approaches [204]. The best algorithms were able to achieve mean DSC values of 0.908 and 0.870 for CT- and MR-based approaches, respectively, on the challenge's final test dataset. For datasets not incorporating contrast enhanced scans, Morris et al. developed 3D deep learning techniques to segment sub-structures of the heart (Fig. 11.15) that were superior to multi-atlas techniques [205].

11.6.1.4 Pelvis and Abdomen

Normal tissues in abdominal and pelvic regions are largely duplicated for various radiotherapy sites such as liver, pancreas, cervix, prostate, and rectum [206]. However, different sites have various treatment protocols (e.g., full vs. empty bladder), various patient orientation (e.g., supine vs. prone), and different organs for male/female patients (e.g., prostate, uterus). This makes the development of a ubiquitous auto-segmentation tool very complicated. Instead, the auto-segmentation tools have been developed for each site/protocol [20].

In the abdominal region, liver and pancreatic cancers are commonly treated with radiotherapy. For liver cancer, auto-segmentation models were mostly developed for CT images as it is a major modality for radiotherapy. Ahn et al. [207] used FusionNet [208] to auto-segment normal tissues for radiotherapy for liver cancer,

and achieved mean DSC values of 0.92, 0.93, 0.86, 0.85, and 0.60 for heart, liver, right kidney, left kidney, and stomach, respectively. Kim et al. [209] used a 3D U-Net to auto-segment organs in the abdominal region and achieved mean DSC values of 0.96, 0.81, 0.60, 0.90, and 0.91 for liver, stomach, duodenum, right kidney, and left kidney, respectively. Tong et al. [210] proposed a self-paced DenseNet architecture to develop an auto-segmentation model for eight abdominal structures, and achieved mean DSC values of 0.96, 0.95, 0.95, 0.89, 0.81, 0.79, 0.72, and 0.69 for liver, spleen, left kidney, stomach, gallbladder, pancreas, esophagus, and duodenum, respectively. Anderson et al. [211] developed an auto-segmentation model for the liver using DeepLabV3+ and achieved a mean DSC value of 0.96 for both contrast and non-contrast CT images. For pancreatic cancer, on the other hand, MRI helps delineate the tumors due to better soft-tissue contrast [212–214]; therefore, most of the auto-segmentation systems were developed for MR images, especially for online adaptive radiotherapy using MR-LINAC. Fu et al. [215] proposed a CNN-based architecture with the conditional random field as a post-processing method. They trained the architecture to auto-contour five normal tissues in the abdominal region for MRI-guided adaptive radiotherapy, and achieved mean DSC values of 0.95, 0.93, 0.87, 0.85, and 0.66 for liver, kidneys, bowel, stomach, and duodenum, respectively (Fig. 11.16). Liang et al. [216] used a multi-level fusion approach to auto-contour liver, left kidney, and right kidney for online adaptive MR-guided radiotherapy and achieved mean DSC values of 0.97, 0.90, and 0.86, respectively.

In the pelvic region, prostate, cervical, and rectal cancers are the most common cancer sites that are treated with radiotherapy. For prostate cancer, both CT and MRI are commonly used to delineate the tumors and normal tissues. Elguindi et al. [217] trained DeepLabV3+ to auto-segment organs on MRI of the male pelvis, and

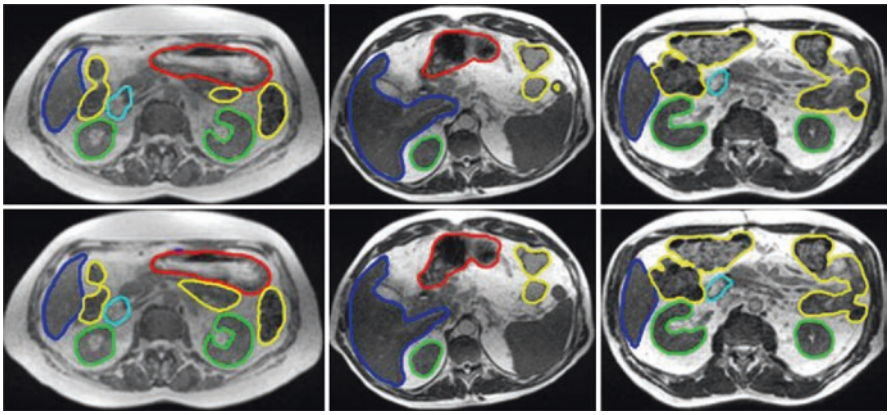


Fig. 11.16 MR-based auto-segmentation of abdominal organs at risk by Fu et al. [215] The top row shows auto-segmentations, whereas the bottom row shows the manually contoured organs at risk. ROIs contour include liver (blue), kidneys (green), duodenum (light blue), stomach (red), and bowels (yellow)

achieved mean DSC values of 0.93, 0.83, 0.82, 0.81, 0.74, and 0.69 for bladder, prostate and seminal vesicles, rectum, rectal spacer, penile bulb, and urethra, respectively. Dong et al. [218] synthetically generated MR images from CT images using cCycleGAN, then trained a deep attention U-Net with the synthetic MR images. They achieved mean DSC values of 0.95, 0.87, and 0.89 for bladder, prostate, and rectum, respectively. Balagopal et al. [219] used a 2D U-Net for organ localization and a 3D U-Net with ResNeXt blocks for organ segmentation on CT images, and achieved mean DSC values of 0.96, 0.95, 0.95, 0.90 and 0.84 for left femur, right femur, bladder, prostate, and rectum, respectively. Compared to prostate cancer, auto-segmentation studies on cervical and rectal cancer are relatively less common. Liu et al. [220] proposed a new CNN-based architecture that modified the U-Net by replacing the convolutional layers to Context Aggregation Blocks. They trained the modified U-Net to auto-segment six normal tissues for cervical cancer radiotherapy on CT images and achieved mean DSC values of 0.92, 0.91, 0.85, 0.83, 0.83, 0.79, for the bladder, femurs, bone marrow, small intestine, spinal cord, and rectum, respectively. Song et al. [221] trained DeepLabV3+ to segment normal tissues for rectal cancer on CT images and achieved mean DSC values of 0.90, 0.90, 0.76 for bladder, femurs, and small intestine, respectively. Furthermore, Men et al. [222] studied the impact of patient orientation (i.e., prone or supine) on CNN-based auto-segmentation models for rectal cancer patients and demonstrated that a model trained from data combining both orientations works as good as a model trained on data from a single orientation on that specific orientation (Fig. 11.17).

11.6.2 Tumors and Clinical Target Volumes

11.6.2.1 Tumors

The gross tumor volume (GTV) is the gross demonstrable extent and location of the malignant growth where the tumor cell density is the highest [223]. The shape, size, and location of a GTV are often determined from a combination of various imaging techniques, such as X-ray, CT, various MRI sequences, and PET images. Defining the true GTV accurately and consistently is a major challenge in radiation treatment planning, because the GTV may appear in different size and shape on different imaging modalities. Depending on the availability of imaging examinations and the experience of the radiation oncologists, the defined GTV contours can vary greatly from one physician to the other. Even with multiple imaging examinations, it is still not rare to see significant inter-observer variation in GTV definition [224, 225].

The success of normal tissue auto-segmentation sparked the research of GTV auto-segmentation. However, auto-segmentation of the GTV is much more challenging than the normal tissue. Current state-of-the-art segmentation algorithms, such as the atlas-based, model-based, or deep learning-based methods, rely on the prior knowledge of the underlying structures/anatomy to be segmented [20]. The fundamental assumption here is that the structure to be segmented is similar from patient to patient, so that the existing knowledge (i.e., contoured patients) can be learned by the algorithm. This assumption is generally true for normal tissues.

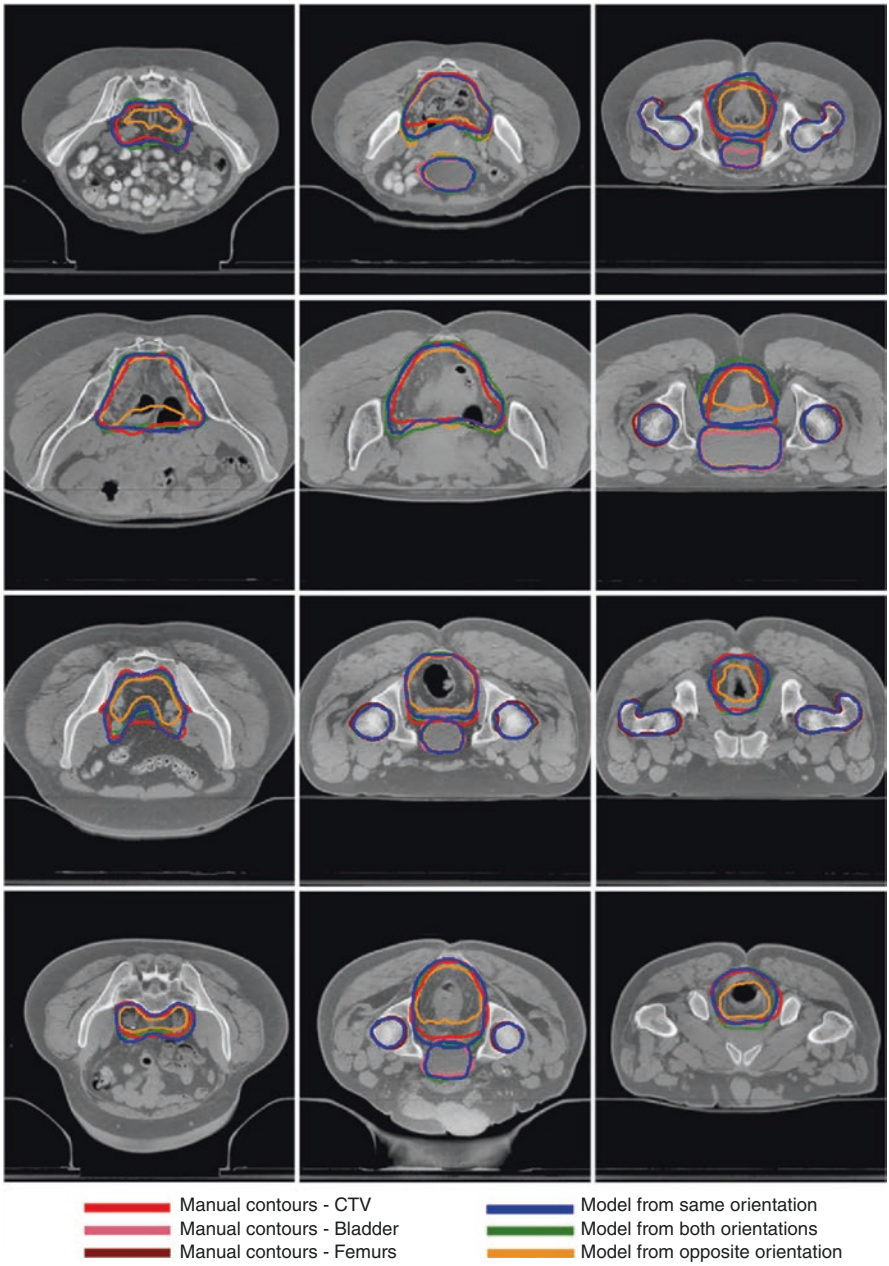


Fig. 11.17 Illustration by Men et al. [222] highlighting under-performance of auto-segmentations when trained models used scans where patients were positioned supine, prone, or both. Including patients in both prone and supine positions resulted in better predictions than individual position (i.e., only supine cases) alone

However, unlike the normal tissue, tumor is different from patient to patient in terms of size, shape, and location. In addition, the size and shape of the tumor can keep changing from day to day. This variation requires a complex model being built to segment the GTV, and training and building such a complex model is not straightforward. Furthermore, multiple modalities of images are often needed to accurately define the GTV. Auto-segmentation from multiple modality images adds additional complexity in creating the segmentation models [226, 227]. Nevertheless, active research has been carried out on GTV auto-segmentation and the current status is summarized according to anatomical site as follows.

Brain tumor segmentation is possibly the most widely investigated of all tumor sites, mainly due to the widely available benchmark datasets. A typical example is the BRATS, the multimodal brain tumor image segmentation benchmark [122]. Brain tumor segmentation is commonly performed using multiple MRI sequences. Occasionally, the CT image may be included. A nice review article has summarized the brain tumor segmentation of MRI images in recent years [228]. Of the methods summarized therein, the best performance was achieved by the DeepMedic with conditional random field (CRF) method, with a mean DSC value of 0.90 for whole tumor segmentation [92].

More recently, a lot more brain tumor segmentation methods have been proposed, almost all using deep learning methods. Some most recent publications have shown a mean DSC value of 0.91, achieved by different groups using different deep learning architectures [229–231] (Fig. 11.18). On the other hand, researchers also investigated the detection and segmentation of brain metastases for stereotactic radiosurgery [232, 233]. The most recent study reported the detection accuracy with the area under the receiver operating characteristic curve (AUC) of 0.98 and the detection and segmentation accuracy with a mean DSC value of 0.79 [233].

Head and neck tumors are generally not visible on non-contrast CT. To define the GTV accurately for treatment planning, radiation oncologists often use a combination of MR (T1, T1c, and T2), CT, PET/CT, and/or dual energy CT images. In the last decades, most development for GTV auto-segmentation has been focused on PET or PET/CT image pair. Varied approaches have been proposed, including graph-cut [234, 235], Markov random field [236], random walk [237], decision tree [238], and k-nearest neighbor [239]. Two recent studies also proposed a 2D CNN and a 3D CNN to segment the GTV from PET-CT pair [240, 241]. Deng et al. proposed a support vector machine method to segment the GTV from contrast-enhanced MRI [242]. The only multi-modality segmentation was proposed by Yang et al. [243], by using the Markov random field and expectation-maximization algorithm to segment from PET-CT image pair, T1-weighted contrast MRI, and simulation CT image. A mean DSC value of 0.74 was achieved in the multi-modality segmentation.

Lung tumors are generally visible on CT, but its intensity is similar to normal tissue. Depending on the tumor location, segmentation directly from a CT image alone could be a straightforward process if the tumor is located in the middle of the lung. However, if the tumor attaches to the chest wall or mediastinum, it is often difficult to separate it from surrounding normal tissues. In this situation, using PET

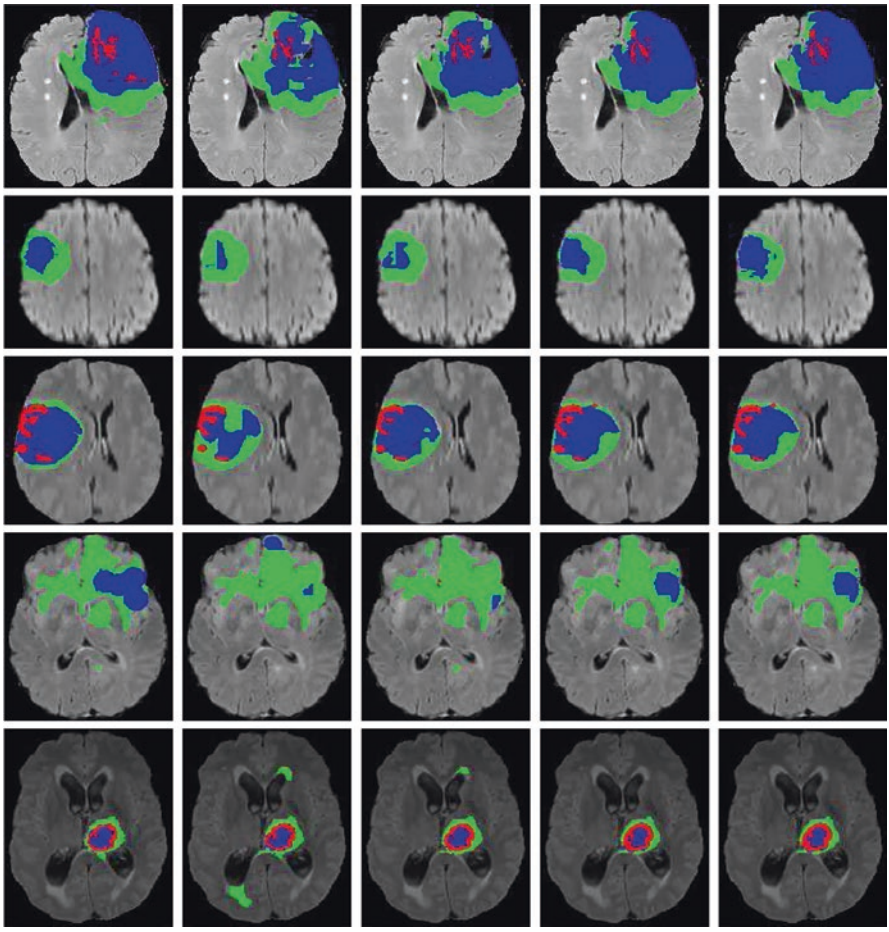


Fig. 11.18 Brain tumor segmentation by Zhou et al. [230] on FLAIR MR scans. From left to right, the authors present the physician-approved segmentations with results from four different auto-segmentation models, with the furthest right column displaying auto-segmentations using the authors' proposed approach. Contours include edema (green), necrosis and non-enhancing tumor (red), and the tumor core (blue)

together with CT can greatly improve the segmentation accuracy. Varied traditional segmentation approaches, mostly semi-automatic, have been developed to segment the lung tumor from PET or CT images. Representative approaches include single-click ensemble methods [244] and marker controlled watershed methods [245]. Most recently, there has been concerted effort on developing deep learning methods for lung tumor segmentation. For example, a multiple-resolution residual network (MRRN) was developed to segment lung tumors from CT images with a reported mean DSC value of 0.75 [246] (Fig. 11.19). A multimodal spatial attention module in combination with a CNN was developed to segment lung tumors from PET–CT

pairs with a reported mean DSC value of 0.71 [247]. Some other recently developed approaches have been presented at the 2020 SPIE Medical Imaging conferences [248, 249]. While MR is not often used for lung tumor segmentation, a recent study developed a deep learning approach to combine CT and MR for lung tumor segmentation with a reported mean DSC value of 0.75 [250].

Not many studies investigated abdominal tumor segmentation or pelvis tumor segmentation, possibly due to the limited clinical usability, the difficulty to achieve a desirable result, and the limited availability of benchmark imaging datasets. A recent development applied a radiomics-guided GAN for segmentation of liver tumor from MR images [251]. On the other hand, a weakly supervised CNN method was proposed to segment renal tumor from abdominal CT angiographic images [252]. It is worth mentioning that a unified and end-to-end adversarial learning

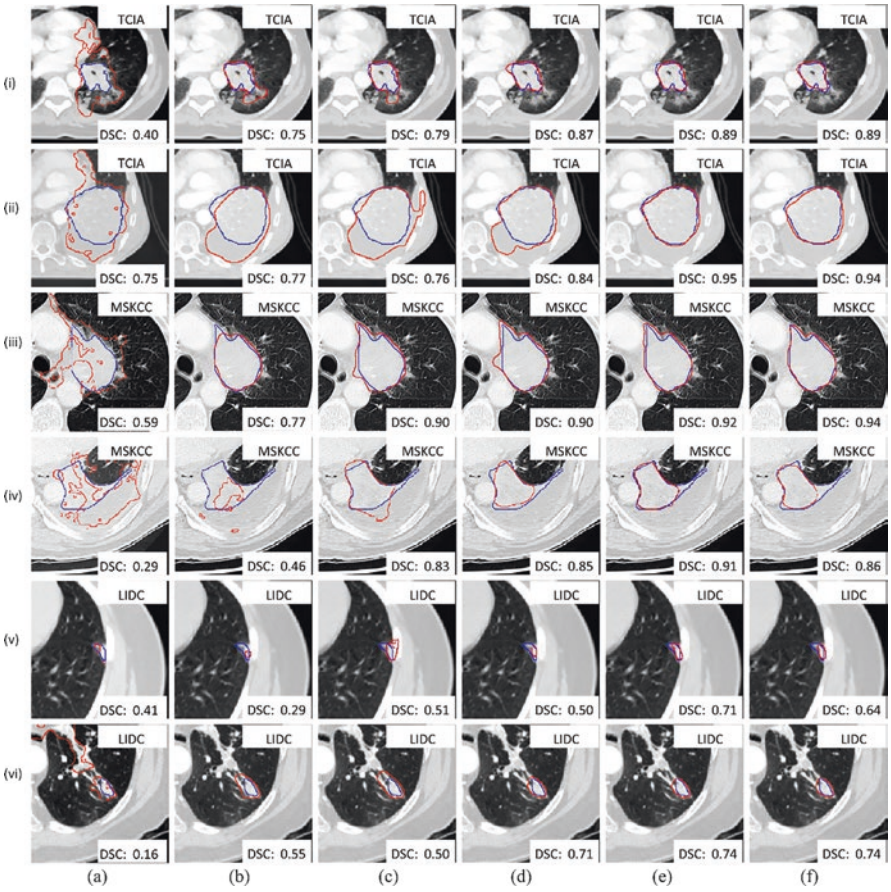


Fig. 11.19 CT-based lung tumor auto-segmentation by Jiang et al. [246] Manual contours (blue) and auto-segmentations (red) are displayed for six cases (rows) and six models (columns). Results from the authors' proposed multiple-resolution residual network (MRRN) are displayed in column (f)

framework named CTumorGAN, consisting of a Generator network and a Discriminator network, has been developed specifically for tumor segmentation from CT images [253]. This segmentation method was tested with lung tumors, liver tumors, and kidney tumors and achieved mean DSC values of 0.71, 0.80, and 0.84, respectively.

11.6.2.2 Clinical Target Volumes

Clinical target volume (CTV) definition is a complex task where the radiation oncologist decides which regions about the tumor would need to be targeted to receive radiation. Generally, the CTV consists of the GTV plus a margin extension which aims to cover any microscopic disease present that cannot be seen with currently available medical imaging devices [223]. Delineating the CTV is considered a more difficult task than normal tissue or GTV segmentation since it requires detailed knowledge of the surrounding anatomy and pathways of tumor spread. Adding to this complexity, several CTV levels are often used for specific treatment sites to deliver reduced doses to intermediate-, and/or low-risk disease spreading regions, such as lymph node levels of the head and neck and pelvic regions. Accurate and reproducible CTV delineation is very important in radiation oncology. As physicians rely on training and experience to manually delineate CTVs, this process can become subjective, and contours have the potential to greatly differ between physicians [1, 254]. Auto-segmenting CTVs could address these issues by increasing consistency through systematic definition of these volumes [255].

When considering most cancer sites, CTVs can be defined based on their primary (CTVp) and/or nodal disease (CTVn). For many cases, CTVp can be defined as the GTV plus some margin extensions which may need to be corrected for anatomical barriers of tumor invasion. Automatic CTV definition approaches should mirror this approach. Belshi et al. proposed a GTV-to-CTV margin expansion following rules posed by anatomical barriers where the barriers were manually defined [256]. More recently, Shusharina et al. developed a fully automatic GTV-to-CTV expansion approach for glioblastoma using deep learning-based auto-segmentations of adjacent anatomical barriers resulting in an automated workflow that allows for CTV definition that is consistent with neuroanatomy [257].

As discussed in Sect. 11.6.1, deep learning has been shown as a promising tool to auto-segment organs at risk for a variety of anatomical sites. These same models could be leveraged for CTVp delineation for radiotherapy treatment sites such as the prostate. Liu et al. [258] showed that auto-segmentation of the prostate on CT images was possible reporting mean DSC values of 0.85 between the auto-segmented and physician-drawn volumes. Elguindi et al. [217] showed similar results on MR images where they achieved mean DSC values of 0.83 for prostate radiotherapy CTVs (prostate + seminal vesicles). Anas et al. showed feasibility of prostate brachytherapy CTV auto-delineation using transrectal ultrasound images achieving a mean DSC value of 0.94 [259]. Men et al. used a deep dilated residual network to automatically define intact breast for breast cancer radiotherapy achieving mean DSC values of 0.91 [260]. More recently, Schreier et al. used a densely connected 3D U-Net to auto-segment the breast achieving median DSC values of 0.93 [203].

Lymph node level segmentation through atlas-based approaches has been previously explored for head and neck [27, 261, 262], breast [263], pelvis [263], and the thorax [264]. More recently, deep learning-based approaches have been investigated to further improve the accuracy of the resulting auto-segmentations. For head and neck, Cardenas et al. [145] developed a model that can auto-delineate lymph node level target volumes (Fig. 11.20) providing physicians with several target coverage options (i.e., coverage options include combinations of lymph node levels: Ia through V, Ib through V, II through IV, and retropharyngeal nodes) with high agreement to the physician contours (mean DSC value of 0.89 for all target volumes). Furthermore, the authors showed through a multi-institutional evaluation of the auto-delineated target volumes that 99% of target volumes reviewed could be used without risk by treating physicians. Rhee et al. [144] developed a deep learning model to auto-segment cervical cancer CTVs and organs at risk (Fig. 11.21) and achieved mean DSC values of 0.81 and 0.76 for pelvic and paraaortic lymph nodes, respectively. Upon physician evaluation, the authors report that 70% and 87% of auto-delineated pelvic and paraaortic lymph node CTVs, respectively, were clinically acceptable without requiring significant manual edits. Several works report promising results for deep learning-based rectal cancer CTV definition using a single ROI to provide coverage for CTVp and CTVn. Men et al. used a 2D deep dilated CNN to auto-delineate rectal cancer CTV and reported mean DSC values of 0.88 for target volumes [265]. Similar results (mean DSC value of 0.88) were achieve by Song et al. using the DeepLabV3+ architecture [221].

The above paragraphs introduce segmentation problems where the tasks focus on anatomical structures rather than tumors. For many sites, tumor location, size, and invasion of surrounding anatomy mandate potential pathways of disease spread. Deep learning offers an advantage over previously developed techniques as it could potentially identify CTV delineation patterns from prior treated volumes based on the location of the GTV. Men et al. and Cardenas et al. demonstrated that delineation patterns could be captured and used to automatically define low-risk CTVs for nasopharyngeal [266] and oropharyngeal [267] cancer patients, respectively. Similarly, Cardenas et al. [143] developed a deep learning model to auto-delineate high-risk CTVs for oropharyngeal cancer patients achieving mean DSC values of 0.81. Jin et al. investigated the use of adjacent anatomical structures to provide spatial context to a deep network to auto-delineate esophageal CTVs [268]; in this work, the authors report a mean DSC value of 0.84 for these CTVs. Post-operative definition of CTVs brings additional challenges as the original tumor is no longer present, and there is the possibility of drastic anatomical changes due to the surgical procedure. Balagopal et al. developed a 3D deep network which localizes and then segments post-operative CTVs for prostate cancer radiotherapy reporting mean DSC values of 0.87 for the auto-delineated CTVs [269]. Bi et al. showed improved consistency in CTV delineation for post-operative non-small lung cancer patients when compared to junior faculty delineations achieving a mean DSC value of 0.75 for auto-delineated CTVs (vs. 0.72 for manual contours) [270]. The ability of deep learning-based auto-segmentation approaches to identify and replicate delineation patterns from prior radiotherapy cases is promising in advancing the use of computational models for CTV delineation.

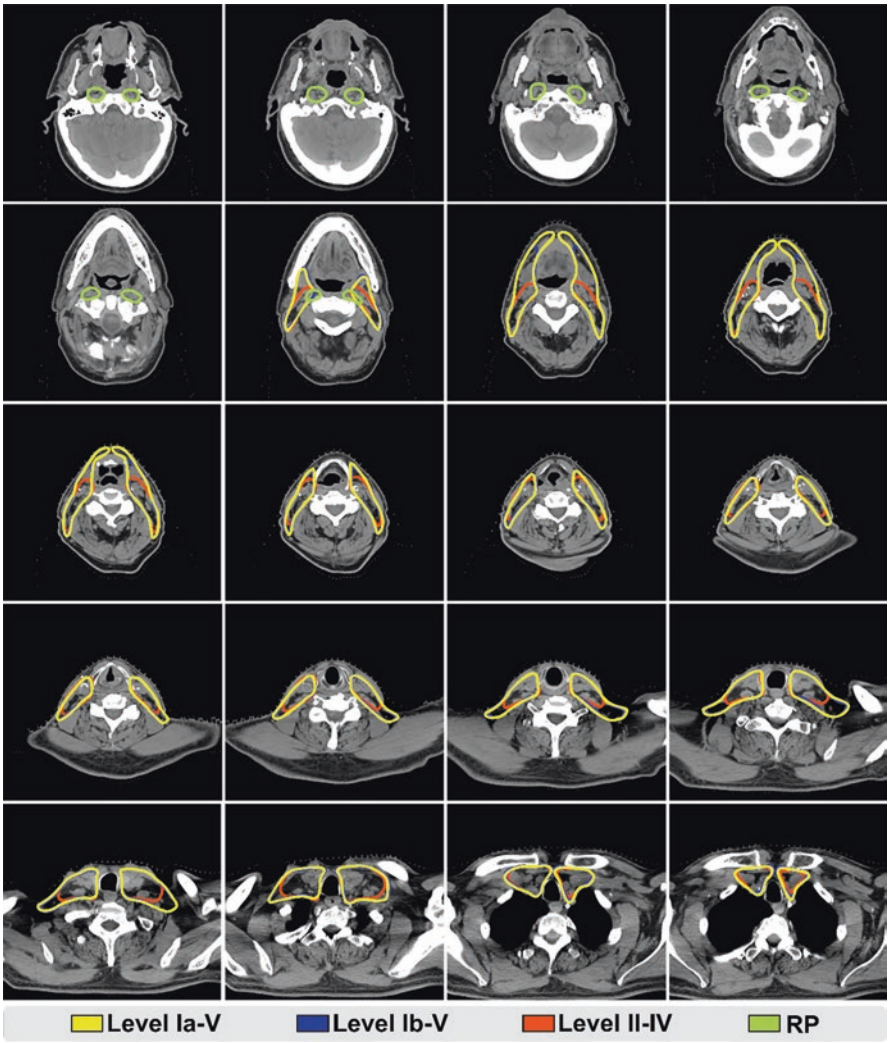


Fig. 11.20 Deep learning-based auto-segmentation of lymph node level target volumes by Cardenas et al. [145] Each panel displays a CT slice on a test patient; here, target volumes for lymph node levels Ia-V (yellow), levels Ib-V (blue), levels II-IV (red), and retropharyngeal nodes (RP, green) are shown for each slice

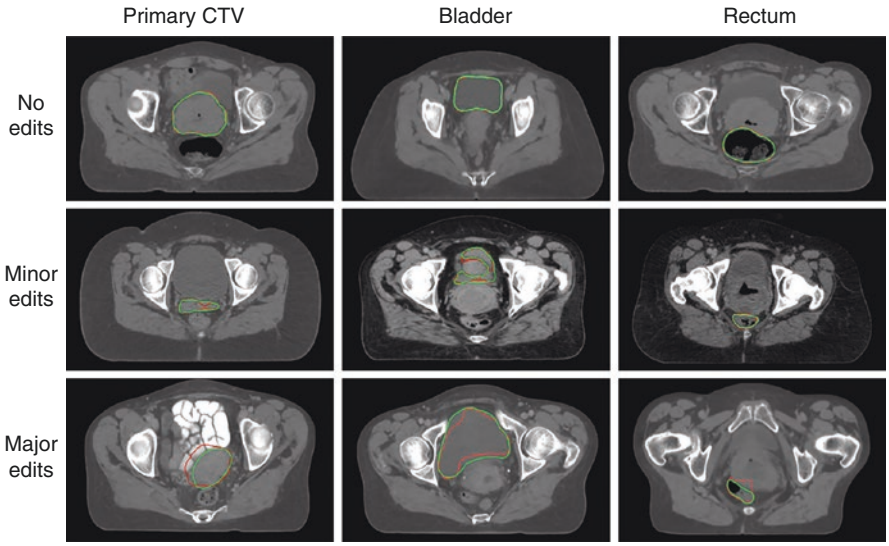


Fig. 11.21 Illustration from Rhee et al. [144] highlighting cases whose auto-segmentations required no edits, minor edits, or major edits after physician visual inspection. For primary CTVs, the authors report that 83% of auto-segmentations were scored as clinically-acceptable when considering minor/stylistic edits

11.7 Conclusion

Auto-segmentation has gone through many advances over recent decades. Many of the techniques from the first- and second-generation algorithms are still in use today due to their simplicity and ease of use. The third-generation techniques, particularly atlas-based contouring, have become almost ubiquitous, with most commercial and many open-source systems offering atlas-based auto-contouring solutions. Recently, with computing and algorithmic advances, deep learning techniques have become the state-of-the-art in medical image segmentation moving the field into the fourth generation of auto-segmentation technique development. Deep learning techniques vary from CNNs to FCNs to GANs and more. Algorithm development has been very rapid with novel architectures and auto-segmentation strategies emerging often and beating out others in auto-contouring challenges. Relatively recently, image segmentation moved from using 2D inputs to 3D volumes with deep learning. The deep learning techniques have been applied to auto-segment targets and normal tissues in many anatomical sites including the thorax, abdomen, pelvis, head and neck, and brain with some applications producing better results than the measured inter- and intra-observer contouring variability. Additionally, deep learning has been shown to contour tumors and CTVs showing where treatment planning in radiotherapy is likely headed. These deep learning systems are still in their relative infancy compared to other techniques, but their impressive results and ever increasing use will lead to increased availability (commercial and open-source) of deep learning-based auto-segmentation tools for radiotherapy treatment planning, as well as increased acceptance and implementation of auto-segmentation tools in clinical practice. For all types of

auto-segmentation, commissioning and periodic QA of these systems is vital in ensuring patient safety and proper use of the systems. Particularly with the performance of deep learning-based auto-contouring techniques, we may be approaching the end of manual segmentation as auto-contouring solutions provide much more reliable contours compared to the manual inter- and intra-observer contouring variability and significantly reduce the time to produce a radiotherapy treatment plan making online adaptive radiotherapy more feasible.

References

1. Hong TS, Tome WA, Harari PM. Heterogeneity in head and neck IMRT target design and clinical practice. Radiother Oncol. 2012;103(1):92–8. https://doi.org/10.1016/j.radonc.2012.02.010.
2. Multi-Institutional Target Delineation in Oncology. Human—Computer Interaction in Radiotherapy Target Volume Delineation: A Prospective , Multi-institutional Comparison of User Input Devices. 2011;24:794–803. https://doi.org/10.1007/s10278-010-9341-2.
3. Harari PM, Song S, Tomé WA. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010;77(3):950–8. https://doi.org/10.1016/j.ijrobp.2009.09.062.
4. Chen Z, King W, Pearcey R, Kerba M, Mackillop WJ. The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature. Radiother Oncol. 2008;87(1):3–16. https://doi.org/10.1016/j.radonc.2007.11.016.
5. Stefoski Mikeljevic J, Haward R, Johnston C, et al. Trends in postoperative radiotherapy delay and the effect on survival in breast cancer patients treated with conservation surgery. Br J Cancer. 2004;90(7):1343–8. https://doi.org/10.1038/sj.bjc.6601693.
6. Li XA, Ph D, Tai A, et al. Variability of target and normal structure delineation for breast-cancer radiotherapy: a RTOG multi-institutional and multi- observer study. Int J Radiat Oncol Biol Phys. 2009;73(3):944–51. https://doi.org/10.1016/j.ijrobp.2008.10.034.Variability.
7. Eminowicz G, McCormack M. Variability of clinical target volume delineation for definitive radiotherapy in cervix cancer. Radiother Oncol. 2015;117(3):542–7. https://doi.org/10.1016/j.radonc.2015.10.007.
8. Ng SP, Dyer BA, Kalpathy-Cramer J, et al. A prospective in silico analysis of interdisciplinary and interobserver spatial variability in post-operative target delineation of high-risk oral cavity cancers: does physician specialty matter? Clin Transl Radiat Oncol. 2018;12:40–6. https://doi.org/10.1016/j.ctro.2018.07.006.
9. Owens CA, Peterson CB, Tang C, et al. Lung tumor segmentation methods : Impact on the uncertainty of radiomics features for non-small cell lung cancer. PLoS One. 2018;13:1–23. https://doi.org/10.1371/journal.pone.0205003.
10. Parmar C, Velazquez ER, Leijenaar R, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One. 2014;9(7):1–8. https://doi.org/10.1371/journal.pone.0102107.
11. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and prognosis of quantitative features extracted from CT images. Transl Oncol. 2014;7(1):72–87. https://doi.org/10.1593/tlo.13844.
12. Lee M, Woo B, Kuo MD, Jamshidi N, Kim JH. Quality of radiomic features in glioblastoma multiforme: impact of semi-automated tumor segmentation software. Korean J Radiol. 2017;18(3):498–509. https://doi.org/10.3348/kjr.2017.18.3.498.
13. Kalpathy-cramer J, Mamomov A, Zhao B, et al. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. Tomography. 2016;2(4):430–7. https://doi.org/10.18383/j.tom.2016.00235.
14. Rasch C, Steenbakkers R, Van Herk M. Target definition in prostate, head, and neck. Semin Radiat Oncol. 2005;15(3):136–45. https://doi.org/10.1016/j.semradonc.2005.01.005.

15. Weiss E, Hess CF. The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy theoretical aspects and practical experiences. Strahlenther Onkol. 2003;179:21–30. https://doi.org/10.1007/s00066-003-0976-5.
16. Saarnak AE, Boersma M, Van Bunningen BNFM, Wolterink R, Steggerda MJ. Inter-observer variation in delineation of bladder and rectum contours for brachytherapy of cervical cancer. Radiother Oncol. 2000;56(1):37–42. https://doi.org/10.1016/S0167-8140(00)00185-7.
17. Withey DJ, Koles ZJ. Medical image segmentation: methods and software. Proc NFSI ICFBI. 2007;2007:140–3. https://doi.org/10.1109/NFSI-ICFBI.2007.4387709.
18. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012:1–9. https://doi.org/10.1016/j.protcy.2014.09.007.
19. Deng J, Dong W, Socher R, Li L-J, Li K. Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Piscataway, New Jersey: IEEE; 2009. p. 248–55. https://doi.org/10.1109/CVPR.2009.5206848.
20. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. Semin Radiat Oncol. 2019;29(3):185–97. https://doi.org/10.1016/j.semradonc.2019.02.001.
21. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging. 2001;20(1):45–57. https://doi.org/10.1109/42.906424.
22. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Med Phys. 2014;41(5):050902.
23. Rohlfing T, Brandt R, Menzel R, Russakoff DB, Maurer CR. Quo Vadis, atlas-based segmentation? In: Handbook of biomedical image analysis. Boston, MA: Springer US; 2005. p. 435–86. https://doi.org/10.1007/0-306-48608-3_11.
24. Thirion J-P. Image matching as a diffusion process: an analogy with Maxwell's demons. Med Image Anal. 1998;2(3):243–60. https://doi.org/10.1016/S1361-8415(98)80022-4.
25. Rueckert D, Sonoda LI, Hayes C, Hill DLG, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging. 1999;18(8):712–21. https://doi.org/10.1109/42.796284.
26. Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach. Med Phys. 2011;38(11):6160–70. https://doi.org/10.1118/1.3654160.
27. Han X, Hoogeman MS, Levendag PC, et al. Atlas-based auto-segmentation of head and neck CT images. Med Image Comput Comput Assist Interv. 2008;11:434–41. https://doi.org/10.1007/978-3-540-85990-1-52.
28. Klein S, van der Heide UA, Lips IM, van Vulpen M, Staring M, Pluim JPW. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Med Phys. 2008;35(4):1407–17. https://doi.org/10.1118/1.2842076.
29. Wang H, Dong L, Lii MF, et al. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. Int J Radiat Oncol. 2005;61(3):725–35. https://doi.org/10.1016/j.ijrobp.2004.07.677.
30. Commowick O, Malandain G. In: Ayache N, Ourselin S, Maeder A, editors. Efficient selection of the Most similar image in a database for critical structures segmentation, vol. 4792. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. https://doi.org/10.1007/978-3-540-75759-7.
31. Rohlfing T, Brandt R, Menzel R, Maurer CR. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage. 2004;21(4):1428–42. https://doi.org/10.1016/j.neuroimage.2003.11.010.
32. Blezek DJ, Miller JV. Atlas stratification. Med Image Anal. 2007;11(5):443–57. https://doi.org/10.1016/j.media.2007.07.001.
33. Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage. 2009;46(3):726–38. https://doi.org/10.1016/j.neuroimage.2009.02.018.
34. Wu M, Rosano C, Lopez-Garcia P, Carter CS, Aizenstein HJ. Optimum template selection for atlas-based segmentation. NeuroImage. 2007;34(4):1612–8. https://doi.org/10.1016/j.neuroimage.2006.07.050.

35. Jia H, Wu G, Wang Q, Shen D. ABSORB: atlas building by self-organized registration and bundling. In: 2010 IEEE computer society conference on computer vision and pattern recognition, vol 51. Piscataway, New Jersey: IEEE; 2010. p. 2785–90. https://doi.org/10.1109/CVPR.2010.5540007.
36. Yang J, Haas B, Fang R, et al. Atlas ranking and selection for automatic segmentation of the esophagus from CT scans. Phys Med Biol. 2017;62(23):9140–58. https://doi.org/10.1088/1361-6560/aa94ba.
37. Yang J, Zhang Y, Zhang L, Dong L. Automatic segmentation of parotids from CT scans using multiple atlases. Med Image Anal Clin A Gd Chall. 2010:323–30.
38. Commowick O, Warfield SK, Malandain G. Using Frankenstein's creature paradigm to build a patient specific atlas. Med Image Comput Comput Assist Interv. 2009;12:993–1000. https://doi.org/10.1007/978-3-642-04271-3_120.
39. Iglesias JE, Sabuncu MR. Multi-atlas segmentation of biomedical images: a survey. Med Image Anal. 2015;24(1):205–19. https://doi.org/10.1016/j.media.2015.06.012.
40. Chen A, Niermann KJ, Deeley MA, Dawant BM. Evaluation of multiple-atlas-based strategies for segmentation of the thyroid gland in head and neck CT images for IMRT. Phys Med Biol. 2012;57(1):93–111. https://doi.org/10.1088/0031-9155/57/1/93.
41. Yang J, Amini A, Williamson R, et al. Automatic contouring of brachial plexus using a multi-atlas approach for lung cancer radiation therapy. Pract Radiat Oncol. 2013;3(4):e139–47. https://doi.org/10.1016/j.prro.2013.01.002.
42. Sjöberg C, Lundmark M, Granberg C, Johansson S, Ahnesjö A, Montelius A. Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients. Radiat Oncol. 2013;8(1):1–7. https://doi.org/10.1186/1748-717X-8-229.
43. Kirişli HA, Schaap M, Klein S, et al. Evaluation of a multi-atlas based method for segmentation of cardiac CTA data: a large-scale, multicenter, and multivendor study. Med Phys. 2010;37(12):6279–91. https://doi.org/10.1118/1.3512795.
44. Isgum I, Staring M, Rutten A, Prokop M, Viergever MA, van Ginneken B. Multi-atlas-based segmentation with local decision fusion—application to cardiac and aortic segmentation in CT scans. IEEE Trans Med Imaging. 2009;28(7):1000–10. https://doi.org/10.1109/TMI.2008.2011480.
45. Sabuncu MR, Yeo BTT, Van Leemput K, Fischl B, Golland P. A generative model for image segmentation based on label fusion. IEEE Trans Med Imaging. 2010;29(10):1714–29. https://doi.org/10.1109/TMI.2010.2050897.
46. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004;23(7):903–21.
47. Langerak TR, van der Heide UA, Kotte ANTJ, Viergever MA, van Vulpen M, Pluim JPW. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Trans Med Imaging. 2010;29(12):2000–8. https://doi.org/10.1109/TMI.2010.2057442.
48. Ramus L, Malandain G. Multi-atlas based segmentation: application to the head and neck region for radiotherapy planning. Med Image Anal Clin. 2010:281–8. http://www.diagnijmegen.nl/~bram/grandchallenge2010/281.pdf
49. Pekar V, Allaire S, Qazi A. Head and neck auto-segmentation challenge: Segmentation of the parotid glands. MICCAI 2010 A Gd Chall Clin. 2010;(October 2015):273–280. http://www.diagnijmegen.nl/~bram/grandchallenge2010/273.pdf.
50. Raudaschl PF, Zaffino P, Sharp GC, et al. Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. Med Phys. 2017;44(5):2020–36. https://doi.org/10.1002/mp.12197.
51. Yang J, Veeraraghavan H, Armato SG, et al. Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. Med Phys. 2018;45(10):4568–81. https://doi.org/10.1002/mp.13141.
52. McCarroll RE, Beadle BM, Balter PA, et al. Retrospective validation and clinical implementation of automated contouring of organs at risk in the head and neck: a step toward

automated radiation treatment planning for low- and middle-income countries. J Glob Oncol. 2018;4:1–11. https://doi.org/10.1200/JGO.18.00055.
53. Zhou R, Liao Z, Pan T, et al. Cardiac atlas development and validation for automatic segmentation of cardiac substructures. Radiother Oncol. 2017;122(1):66–71. https://doi.org/10.1016/j.radonc.2016.11.016.
54. Heimann T, Meinzer H-P. Statistical shape models for 3D medical image segmentation: a review. Med Image Anal. 2009;13(4):543–63. https://doi.org/10.1016/j.media.2009.05.004.
55. Pekar V, McNutt TR, Kaus MR. Automated model-based organ delineation for radiotherapy planning in prostatic region. Int J Radiat Oncol. 2004;60(3):973–80. https://doi.org/10.1016/j.ijrobp.2004.06.004.
56. Freedman D, Radke RJ. Tao Zhang, Yongwon Jeong, lovelock DM, Chen GTY. Model-based segmentation of medical imagery by matching distributions. IEEE Trans Med Imaging. 2005;24(3):281–92. https://doi.org/10.1109/TMI.2004.841228.
57. Feng Q, Foskey M, Chen W, Shen D. Segmenting CT prostate images using population and patient-specific statistics for radiotherapy. Med Phys. 2010;37(8):4121–32. https://doi.org/10.1118/1.3464799.
58. Geremia E, Clatz O, Menze BH, Konukoglu E, Criminisi A, Ayache N. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. NeuroImage. 2011;57(2):378–90. https://doi.org/10.1016/j.neuroimage.2011.03.080.
59. Criminisi A, Shotton J, Konukoglu E. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found Trends® Comput Graph Vis. 2011;7(2–3):81–227. https://doi.org/10.1561/0600000035.
60. Li W, Liao S, Feng Q, Chen W, Shen D. Learning image context for segmentation of prostate in CT-guided radiotherapy. Med Image Comput Comput Assist Interv. 2011;14(Pt 3):570–8. http://www.ncbi.nlm.nih.gov/pubmed/22003745
61. Bauer S, Nolte L-P, Reyes M. Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. In: Medical Image Computing and Computer-Assisted Intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, vol 14; 2011. p. 354–61. https://doi.org/10.1007/978-3-642-23626-6_44.
62. Vaishnavee KB, Amshakala K. An automated MRI brain image segmentation and tumor detection using SOM-clustering and Proximal Support Vector Machine classifier. In: 2015 IEEE international conference on engineering and technology (ICETECH). Piscataway, New Jersey: IEEE; 2015. p. 1–6. https://doi.org/10.1109/ICETECH.2015.7275030.
63. Lu J, Wang D, Lin S, Heng PA. Automatic liver segmentation in CT images based on Support Vector Machine. In: Proceedings of 2012 IEEE-EMBS international conference on biomedical and health informatics, vol 25. Piscataway, New Jersey: IEEE; 2012. p. 333–6. https://doi.org/10.1109/BHI.2012.6211581.
64. Zhang X, Tian J, Xiang D, Li X, Deng K. Interactive liver tumor segmentation from ct scans using support vector classification with watershed. In: 2011 Annual international conference of the IEEE engineering in medicine and biology society, vol 2011. Piscataway, New Jersey: IEEE; 2011. p. 6005–8. https://doi.org/10.1109/IEMBS.2011.6091484.
65. Rendon-Gonzalez E, Ponomaryov V. Automatic Lung nodule segmentation and classification in CT images based on SVM. In: 2016 9th international Kharkiv symposium on physics and engineering of microwaves, millimeter and submillimeter waves (MSMW). Piscataway, New Jersey: IEEE; 2016. p. 1–4. https://doi.org/10.1109/MSMW.2016.7537995.
66. Mahapatra D. Automatic cardiac segmentation using semantic information from random forests. J Digit Imaging. 2014;27(6):794–804. https://doi.org/10.1007/s10278-014-9705-0.
67. Pereira S, Pinto A, Oliveira J, Mendrik AM, Correia JH, Silva CA. Automatic brain tissue segmentation in MR images using random forests and conditional random fields. J Neurosci Methods. 2016;270:111–23. https://doi.org/10.1016/j.jneumeth.2016.06.017.
68. Jin C, Shi F, Xiang D, et al. 3D fast automatic segmentation of kidney based on modified AAM and random Forest. IEEE Trans Med Imaging. 2016;35(6):1395–407. https://doi.org/10.1109/TMI.2015.2512606.

69. Chang KW, Summers RM, Narayanan D, et al. Automated segmentation of the thyroid gland on thoracic CT scans by multiatlas label fusion and random forest classification random forest classification. Med Imaging. 2017;3(2):044006. https://doi.org/10.1117/1.JMI.2.4.044006.
70. Gao Y. Accurate segmentation of CT pelvic organs via incremental cascade learning and regression-based deformable models. ProQuest Diss Theses. 2016;35(6):153. http://libproxy.library.wmich.edu/login?url=https://search.proquest.com/docview/1828255901?accountid=15099%0Ahttp://primo-pmtna01.hosted.exlibrisgroup.com/openurl/01WMU/01WMU_SERVICES??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genr
71. Liu J, Hoffman J, Zhao J, et al. Mediastinal lymph node detection and station mapping on chest CT using spatial priors and random forest. Med Phys. 2016;43(7):4362–74. https://doi.org/10.1118/1.4954009.
72. Serag A, Wilkinson AG, Telford EJ, et al. SEGMA: an automatic SEGMentation approach for human brain MRI using sliding window and random forests. Front Neuroinform. 2017;11(January):1–11. https://doi.org/10.3389/fninf.2017.00002.
73. Haralick RM, Dinstein I, Shanmugam K. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973;SMC-3(6):610–21. https://doi.org/10.1109/TSMC.1973.4309314.
74. Shiradkar R, Podder TK, Algohary A, Viswanath S, Ellis RJ, Madabhushi A. Radiomics based targeted radiotherapy planning (rad-TRaP): a computational framework for prostate cancer treatment planning with MRI. Radiat Oncol 2016;11(1):1–14. doi:https://doi.org/10.1186/s13014-016-0718-3.
75. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd Int Conf Learn Represent ICLR 2015 - Conf track proc. September 2014:1–14. http://arxiv.org/abs/1409.1556.
76. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2015;349:3431–40. https://doi.org/10.1109/CVPR.2015.7298965.
77. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2014:1–9. https://doi.org/10.1109/CVPR.2015.7298594.
78. Liu H, Yan M, Song E, et al. Label fusion method based on sparse patch representation for the brain MRI image segmentation. IET Image Process. 2017;11(7):502–11. https://doi.org/10.1049/iet-ipr.2016.0988.
79. Zhu Y, Wang L, Liu M, et al. MRI-based prostate cancer detection with high-level representation and hierarchical classification. Med Phys. 2017;44(3):1028–39. https://doi.org/10.1002/mp.12116.
80. Kamnitsas K, Ferrante E, Parisot S, et al. DeepMedic for brain tumor segmentation. In: Brainlesion: glioma, multiple sclerosis, Stroke and Traumatic Brain Injuries BrainLes, vol 2016. New York: Springer Publishing; 2016. p. 138–49. https://doi.org/10.1007/978-3-319-55524-9_14.
81. Roth HR, Oda H, Zhou X, et al. An application of cascaded 3D fully convolutional networks for medical image segmentation. Comput Med Imaging Graph. 2018;66:90–9. https://doi.org/10.1016/j.compmedimag.2018.03.001.
82. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28.
83. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. Lect Notes Comput Sci. 2016;9901:424–32. https://doi.org/10.1007/978-3-319-46723-8_49.
84. Milletari F, Navab N, Ahmadi S-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth international conference on 3D vision (3DV). Piscataway, New Jersey: IEEE; 2016. p. 565–71. https://doi.org/10.1109/3DV.2016.79.
85. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–8. https://doi.org/10.1109/CVPR.2016.90.

86. Li X, Chen H, Qi X, Dou Q, Fu C-W, Heng P-A. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans Med Imaging. 2018;37(12):2663–74. https://doi.org/10.1109/TMI.2018.2845918.
87. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS J Photogramm Remote Sens. 2020;162:94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013.
88. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. Lect Notes Comput Sci. 2018;11045:3–11. https://doi.org/10.1007/978-3-030-00889-5_1.
89. Zhang J, Jin Y, Xu J, Xu X, Zhang Y. MDU-Net: Multi-scale Densely Connected U-Net for biomedical image segmentation. 2018. http://arxiv.org/abs/1812.00352.
90. Zhang H, Li J, Shen M, Wang Y, Yang GZ. DDU-nets: distributed dense model for 3D MRI brain tumor segmentation, vol 11993. New York: Springer International Publishing; 2020. https://doi.org/10.1007/978-3-030-46643-5_20.
91. Milletari F, Navab N, Ahmadi S. "V-Net: fully convolutional neural networks for volumetric medical image segmentation," 2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 565–71. https://doi.org/10.1109/3DV.2016.79.
92. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal. 2017;36:61–78. https://doi.org/10.1016/j.media.2016.10.004.
93. Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with gaussian edge potentials. Adv Neural Inform Proc Syst. 2012;24:109–17.
94. Roth HR, et al. A Multi-scale Pyramid of 3D Fully Convolutional Networks for Abdominal Multi-organ Segmentation. In: Frangi A., Schnabel J., Davatzikos C., Alberola-López C., Fichtinger G. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science, vol 11073. Springer, Cham. https://doi.org/10.1007/978-3-030-00937-3_48.
95. Yang M, Yu K, Zhang C, Li Z, Yang K. DenseASPP for semantic segmentation in street scenes. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2018:3684–92. https://doi.org/10.1109/CVPR.2018.00388.
96. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2014;40(4):834–48. http://arxiv.org/abs/1412.7062
97. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2018;40(4):834–48. https://doi.org/10.1109/TPAMI.2017.2699184.
98. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking Atrous convolution for semantic image segmentation. 2017. http://arxiv.org/abs/1706.05587.
99. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11211. Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_49.
100. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal. 2019;58 https://doi.org/10.1016/j.media.2019.101552.
101. Dai W, Dong N, Wang Z, Liang X, Zhang H, Xing EP. SCAN: structure correcting adversarial network for organ segmentation in chest X-rays. In: Stoyanov D, Taylor Z, Carneiro G, et al., editors. Lecture Notes in Computer Science, vol 11045. Cham: Springer International Publishing; 2018. p. 263–73. https://doi.org/10.1007/978-3-030-00889-5_30.
102. Chen J, Yang L, Zhang Y, Alber M, Chen DZ. Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation. Adv Neural Inf Process Syst. 2016:3044–52.
103. Kearney V, Chan JW, Wang T, Perry A, Yom SS, Solberg TD. Attention-enabled 3D boosted convolutional neural networks for semantic CT segmentation using deep supervision. Phys Med Biol. 2019;64(13):135001. https://doi.org/10.1088/1361-6560/ab2818.

104. Kikinis R, Pieper SD, Vosburgh KG. 3D slicer: a platform for subject- specific image analysis, visualization, and clinical support. Intraoperative Imaging Image-Guided Ther. 2014:277–89. https://doi.org/10.1007/978-1-4614-7657-3.
105. PET Tumor Segmentation Extension—3D Slicer. https://www.slicer.org/wiki/Documentation/Nightly/Extensions/PETTumorSegmentation.
106. Fast Grow Cut—3D Slicer. https://www.slicer.org/wiki/Documentation/4.3/Modules/FastGrowCut.
107. DeepInfer—3D Slicer. https://www.slicer.org/wiki/Documentation/Nightly/Modules/DeepInfer.
108. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage. 2006;31(3):1116–28. https://doi.org/10.1016/j.neuroimage.2006.01.015.
109. Distributed Segmentation Services - ITK-SNAP. https://alfabis-server.readthedocs.io/en/latest/.
110. Zaffino P, Raudaschl P, Fritscher K, Sharp GC, Spadea MF. Technical note: Plastimatch MABS, an open source tool for automatic image segmentation. Med Phys. 2016;43(9):5155–60. https://doi.org/10.1118/1.4961121.
111. CIBC. Seg3D: Volumetric Image Segmentation and Visualization. Scientific Computing and Imaging Institute (SCI). http://www.seg3d.org.
112. Pawlowski N, Ktena SI, Lee MCH, et al. DLTK: State of the Art Reference Implementations for Deep Learning on Medical Images 2017:1–4. http://arxiv.org/abs/1711.06853.
113. Abadi M, Barham P, Chen J, et al. TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning. 12th USENIX Symp Oper Syst Des Implement (OSDI '16). 2016:265–284. https://doi.org/10.1038/nn.3331
114. Gibson E, Li W, Sudre C, et al. NiftyNet: a deep-learning platform for medical imaging. Comput Methods Prog Biomed. 2018;158:113–22. https://doi.org/10.1016/j.cmpb.2018.01.025.
115. MONAI: Medical Open Network for AI. https://monai.io/.
116. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019;(NeurIPS). http://arxiv.org/abs/1912.01703.
117. EISEN.ai. https://eisen.ai/.
118. Müller D, Kramer F. MIScnn: A Framework for Medical Image Segmentation with Convolutional Neural Networks and Deep Learning. October 2019. http://arxiv.org/abs/1910.09308.
119. NVIDIA Clara for Medical Imaging. https://developer.nvidia.com/clara-medical-imaging.
120. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26(6):1045–57. https://doi.org/10.1007/s10278-013-9622-7.
121. Bilic P, Christ PF, Vorontsov E, et al. The liver tumor segmentation benchmark (LiTS). January 2019:1–43. http://arxiv.org/abs/1901.04056.
122. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 2015;34(10):1993–2024. https://doi.org/10.1109/TMI.2014.2377694.
123. Heller N, Sathianathen N, Kalapara A, et al. The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. March 2019. http://arxiv.org/abs/1904.00445.
124. Trullo R, Petitjean C, Dubray B, Ruan S. Multiorgan segmentation using distance-aware adversarial networks. J Med Imaging. 2019;6(01):1. https://doi.org/10.1117/1.JMI.6.1.014001.
125. Yang J, Veeraraghavan H, van Elmpt W, Dekker A, Gooding M, Sharp G. CT images with expert manual contours of thoracic cancer for benchmarking auto-segmentation accuracy. Med Phys. 2020:1–6. https://doi.org/10.1002/mp.14107.
126. Litjens G, Toth R, van de Ven W, et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med Image Anal. 2014;18(2):359–73. https://doi.org/10.1016/j.media.2013.12.002.

127. Thompson RF, Valdes G, Fuller CD, et al. Artificial intelligence in radiation oncology: a specialty-wide disruptive transformation? Radiother Oncol. 2018;129(3):421–6. https://doi.org/10.1016/j.radonc.2018.05.030.
128. Delpon G, Escande A, Ruef T, et al. Comparison of automated atlas-based segmentation software for postoperative prostate cancer radiotherapy. Front Oncol. 2016;6:1–6. https://doi.org/10.3389/fonc.2016.00178.
129. Hu Y, Byrne M, Archibald-Heeren B, et al. Implementing user-defined atlas-based auto-segmentation for a large multi-Centre organisation: the Australian experience. J Med Radiat Sci. 2019;66(4):238–49. https://doi.org/10.1002/jmrs.359.
130. Wittenstein O, Hiepe P, Sowa LH, Karsten E, Fandrich I, Dunst J. Automatic image segmentation based on synthetic tissue model for delineating organs at risk in spinal metastasis treatment planning. Strahlenther Onkol. 2019;195(12):1094–103. https://doi.org/10.1007/s00066-019-01463-4.
131. Daisne J-F, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. Radiat Oncol. 2013;8(1):154. https://doi.org/10.1186/1748-717X-8-154.
132. Lustberg T, van Soest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol. 2018;126(2):312–7. https://doi.org/10.1016/j.radonc.2017.11.012.
133. Vaassen F, Hazelaar C, Vaniqui A, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol. 2020;13:1–6. https://doi.org/10.1016/j.phro.2019.12.001.
134. Wong J, Fong A, McVicar N, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother Oncol. 2020;144:152–8. https://doi.org/10.1016/j.radonc.2019.10.019.
135. Zabel WJ, Conway JL, Gladwish A, et al. Clinical evaluation of deep learning and atlas-based auto-contouring of bladder and rectum for prostate radiation therapy. Pract Radiat Oncol. 2020;11:1–10. https://doi.org/10.1016/j.prro.2020.05.013.
136. SOMATOM go.Sim by Siemens. https://www.siemens-healthineers.com/radiotherapy/ct-for-rt/somatom-go-sim. Accessed March 8, 2020.
137. Court LE, Kisling K, McCarroll R, et al. Radiation planning assistant—a streamlined, fully automated radiotherapy treatment planning system. J Vis Exp. 2018;134:e57411. https://doi.org/10.3791/57411.
138. Kisling K, Zhang L, Simonds H, et al. Fully automatic treatment planning for external-beam radiation therapy of locally advanced cervical cancer: a tool for low-resource clinics. J Glob Oncol. 2019;5:1–9. https://doi.org/10.1200/jgo.18.00107.
139. Kisling K, Zhang L, Shaitelman SF, et al. Automated treatment planning of postmastectomy radiotherapy. Med Phys. 2019;46(9):3767–75. https://doi.org/10.1002/mp.13586.
140. Kisling K, Johnson JL, Simonds H, et al. A risk assessment of automated treatment planning and recommendations for clinical deployment. Med Phys. 2019;46(6):2567–74. https://doi.org/10.1002/mp.13552.
141. Rhee DJ, Cardenas CE, Elhalawani H, et al. Automatic detection of contouring errors using convolutional neural networks. Med Phys. 2019;46:5086–97. https://doi.org/10.1002/mp.13814.
142. Netherton T, Joo Rhee D, Cardenas C, et al. Evaluation of a multiview architecture for automatic vertebral labeling of palliative radiotherapy simulation CT images. Med Phys. 2020;47(11):5592–608. https://doi.org/10.1002/mp.14415.
143. Cardenas CE, McCarroll RE, Court LE, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in Dice similarity coefficient parameter optimization function. Int J Radiat Oncol Biol Phys. 2018;101(2):468–78. https://doi.org/10.1016/j.ijrobp.2018.01.114.

144. Rhee DJ, Jhingran A, Rigaud B, et al. Automatic contouring system for cervical cancer using convolutional neural networks. Med Phys. 2020;47(11):5648–58. https://doi.org/10.1002/mp.14467.
145. Cardenas CE, Beadle BM, Garden AS, et al. Generating high-quality lymph node clinical target volumes for head and neck cancer radiotherapy using a fully automated deep learning-based approach. Int J Radiat Oncol. 2020;109(3):801–12. https://doi.org/10.1016/j.ijrobp.2020.10.005.
146. Kisling K, Cardenas C, Anderson BM, et al. Automatic verification of beam apertures for cervical cancer radiation therapy. Pract Radiat Oncol. May 2020:1–10. https://doi.org/10.1016/j.prro.2020.05.001.
147. Ford E, Conroy L, Dong L, et al. Strategies for effective physics plan and chart review in radiation therapy: report of AAPM task group 275. Med Phys. 2020;47(6):e236–72. https://doi.org/10.1002/mp.14030.
148. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging. 2015;15(1):29. https://doi.org/10.1186/s12880-015-0068-x.
149. Dice LR. Measures of the amount of ecologic association between species. Ecology. 1945;26(3):297–302. https://doi.org/10.2307/1932409.
150. Jaccard P. The distribution of the flora in the alphine zone. New Phytol. 1912;XI(2):37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.
151. Huttenlocher DP, Rucklidge WJ, Klanderman GA. Comparing images using the Hausdorff distance under translation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 1992:654–6. https://doi.org/10.1109/CVPR.1992.223209.
152. Nikolov S, Blackwell S, Mendes R, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy 2018:1–31. http://arxiv.org/abs/1809.04430.
153. Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM radiation therapy committee task group no. 132: report. Med Phys. 2017;44(7):e43–76. https://doi.org/10.1002/mp.12256.
154. Chen HC, Tan J, Dolly S, et al. Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: a general strategy. Med Phys. 2015;42(2):1048–59. https://doi.org/10.1118/1.4906197.
155. McCarroll R, Yang J, Cardenas CE, et al. Machine learning for the prediction of physician edits to clinical auto-contours in the head-and-neck. Med Phys. 2017;44(6):3160.
156. Hui CB, Nourzadeh H, Watkins WT, et al. Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach. Med Phys. 2018;45(5):2089–96. https://doi.org/10.1002/mp.12835.
157. Fraass B, Doppke K, Hunt M, et al. American association of physicists in medicine radiation therapy committee task group 53: quality assurance for clinical radiotherapy treatment planning. Med Phys. 1998;25(10):1773–829. https://doi.org/10.1118/1.598373.
158. Smilowitz JB, Das IJ, Feygelman V, et al. AAPM medical physics practice guideline 5.a.: commissioning and qa of treatment planning dose calculations—megavoltage photon and electron beams. J Appl Clin Med Phys. 2016;17(1):6166. https://doi.org/10.1120/jacmp.v17i1.6166.
159. Huq MS, Fraass BA, Dunscombe PB, et al. The report of task group 100 of the AAPM: application of risk analysis methods to radiation therapy quality management. Med Phys. 2016;43(7):4209–62. https://doi.org/10.1118/1.4947547.
160. Cardenas CE, Mohamed ASR, Tao R, et al. Prospective qualitative and quantitative analysis of real-time peer review quality assurance rounds incorporating direct physical examination for head and neck cancer radiation therapy. Int J Radiat Oncol Biol Phys. 2017;98(3):532–40. https://doi.org/10.1016/j.ijrobp.2016.11.019.
161. Marks LB, Adams RD, Pawlicki T, et al. Enhancing the role of case-oriented peer review to improve quality and safety in radiation oncology: executive summary. Pract Radiat Oncol. 2013;3(3):149–56. https://doi.org/10.1016/j.prro.2012.11.010.

162. Cox BW, Kapur A, Sharma A, et al. Prospective contouring rounds: a novel, high-impact tool for optimizing quality assurance. Pract Radiat Oncol. 2015;5(5):e431–6. https://doi.org/10.1016/j.prro.2015.05.005.
163. Ger RB, Zhou S, Chi PCM, et al. Comprehensive investigation on controlling for CT imaging variabilities in Radiomics studies. Sci Rep. 2018;8(1):1–14. https://doi.org/10.1038/s41598-018-31509-z.
164. Huang K, Rhee DJ, Ger RB, et al. Effects of CT image acquisition and reconstruction parameters on automatic contouring algorithms. Med Phys. 2019;46(6):E138–9.
165. Kalavathi P, Prasath VBS. Methods on skull stripping of MRI head scan images—a review. J Digit Imaging. 2016;29(3):365–79. https://doi.org/10.1007/s10278-015-9847-8.
166. Puccio B, Pooley JP, Pellman JS, Taverna EC, Craddock RC. The preprocessed connectomes project repository of manually corrected skull-stripped T1-weighted anatomical MRI data. Gigascience. 2016;5(1):45. https://doi.org/10.1186/s13742-016-0150-5.
167. Hwang H, Rehman HZU, Lee S. 3D U-net for skull stripping in brain MRI. Appl Sci. 2019;9(3):569. https://doi.org/10.3390/app9030569.
168. Kleesiek J, Urban G, Hubert A, et al. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. NeuroImage. 2016;129:460–9. https://doi.org/10.1016/j.neuroimage.2016.01.024.
169. Mendrik AM, Vincken KL, Kuijf HJ, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. Comput Intell Neurosci. 2015;2015:1–16. https://doi.org/10.1155/2015/813696.
170. Moeskops P, de Bresser J, Kuijf HJ, et al. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. NeuroImage Clin. 2018;17:251–62. https://doi.org/10.1016/j.nicl.2017.10.007.
171. Luna M, Park SH. 3D Patchwise U-Net with transition layers for mr brain segmentation. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science, vol 11383. Springer, Cham. https://doi.org/10.1007/978-3-030-11723-8_40.
172. Irimia A, Maher AS, Rostowsky KA, Chowdhury NF, Hwang DH, Law EM. Brain segmentation from computed tomography of healthy aging and geriatric concussion at variable spatial resolutions. Front Neuroinform. 2019;13:9. https://doi.org/10.3389/fninf.2019.00009.
173. Manniesing R, Oei MTH, Oostveen LJ, et al. White matter and gray matter segmentation in 4D computed tomography. Sci Rep. 2017;7(1):119. https://doi.org/10.1038/s41598-017-00239-z.
174. Losseff NA, Webb SL, O'Riordan JI, et al. Spinal cord atrophy and disability in multiple sclerosis. Brain. 1996;119(3):701–8. https://doi.org/10.1093/brain/119.3.701.
175. Freund P, Weiskopf N, Ward NS, et al. Disability, atrophy and cortical reorganization following spinal cord injury. Brain. 2011;134(6):1610–22. https://doi.org/10.1093/brain/awr093.
176. Prados F, Ashburner J, Blaiotta C, et al. Spinal cord grey matter segmentation challenge. NeuroImage. 2017;152:312–29. https://doi.org/10.1016/j.neuroimage.2017.03.010.
177. Perone CS, Calabrese E, Cohen-Adad J. Spinal cord gray matter segmentation using deep dilated convolutions. Sci Rep. 2018;8(1):5966. https://doi.org/10.1038/s41598-018-24304-3.
178. Tins B. Technical aspects of CT imaging of the spine. Insights Imaging. 2010;1(5–6):349–59. https://doi.org/10.1007/s13244-010-0047-2.
179. Shah LM, Salzman KL. Imaging of spinal metastatic disease. Int J Surg Oncol. 2011;2011:1–12. https://doi.org/10.1155/2011/769753.
180. Yao J, Burns JE, Forsberg D, et al. A multi-center milestone study of clinical vertebral CT segmentation. Comput Med Imaging Graph. 2016;49:16–28. https://doi.org/10.1016/j.compmedimag.2015.12.006.
181. Glocker B, Feulner J, Criminisi A, Haynor DR, Konukoglu E. Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In. 2012:590–8. https://doi.org/10.1007/978-3-642-33454-2_73.

182. Löffler MT, Sekuboyina A, Jacob A, et al. A vertebral segmentation dataset with fracture grading. Radiol Artif Intell. 2020;2(4):e190138. https://doi.org/10.1148/ryai.2020190138.
183. Payer C, Štern D, Bischof H, Urschler M. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and U-net. VISIGRAPP 2020 - proc 15th Int Jt Conf Comput vision. Imaging Comput Graph Theory Appl. 2020;5:124–33. https://doi.org/10.5220/0008975201240133.
184. Sekuboyina A, Bayat A, Husseini ME, et al. VerSe: A Vertebrae Labelling and Segmentation Benchmark. January 2020. http://arxiv.org/abs/2001.09193.
185. Sekuboyina A, Rempfler M, Valentinitsch A, Kirschke JS, Menze BH. Adversarially learning a local anatomical prior: vertebrae labelling with 2D reformations. February 2019. http://arxiv.org/abs/1902.02205.
186. Chen J, et al. LSRC: a long-short range context-fusing framework for automatic 3D vertebra localization. In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. Springer, Cham. https://doi.org/10.1007/978-3-030-32226-7_11.
187. Qin C, Yao D, Zhuang H, Wang H, Shi Y, Song Z. Residual Block-based Multi-Label Classification and Localization Network with Integral Regression for Vertebrae Labeling. January 2020. http://arxiv.org/abs/2001.00170.
188. Yang D, et al. Deep image-to-image recurrent network with shape basis learning for automatic vertebra labeling in large-scale 3D CT volumes. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins D, Duchesne S. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science, vol 10435. Springer, Cham. https://doi.org/10.1007/978-3-319-66179-7_57.
189. Chu C, Belavý DL, Armbrecht G, Bansmann M, Felsenberg D, Zheng G. Fully Automatic Localization and Segmentation of 3D Vertebral Bodies from CT/MR Images via a Learning-Based Method. PLoS One. 2015;10(11):e0143327. https://doi.org/10.1371/journal.pone.0143327.
190. Chen Y, Gao Y, Li K, Zhao L, Zhao J. Vertebrae identification and localization utilizing fully convolutional networks and a hidden Markov model. IEEE Trans Med Imaging. 2020;39(2):387–99. https://doi.org/10.1109/TMI.2019.2927289.
191. Korez R, Ibragimov B, Likar B, Pernus F, Vrtovec T. A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation. IEEE Trans Med Imaging. 2015;34(8):1649–62. https://doi.org/10.1109/TMI.2015.2389334.
192. Lessmann N, van Ginneken B, de Jong PA, Išgum I. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. Med Image Anal. 2019;53:142–55. https://doi.org/10.1016/j.media.2019.02.005.
193. Jakubicek R, Chmelik J, Jan J, Ourednicek P, Lambert L, Gavelli G. Learning–based vertebra localization and labeling in 3D CT data of possibly incomplete and pathological spines. Comput Methods Prog Biomed. 2020;183:105081. https://doi.org/10.1016/j.cmpb.2019.105081.
194. Janssens R, Zeng G, Zheng G. Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), vol 2018. Piscataway, New Jersey: IEEE. p. 893–7. https://doi.org/10.1109/ISBI.2018.8363715.
195. Cardenas CE, Mohamed ASR, Yang J, et al. Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. Med Phys. 2020;47(5):2317–22. https://doi.org/10.1002/mp.13942.
196. Wang Y, Zhao L, Wang M, Song Z. Organ at risk segmentation in head and neck CT images using a two-stage segmentation framework based on 3D U-net. IEEE Access. 2019;7:144591–602. https://doi.org/10.1109/ACCESS.2019.2944958.
197. Iyer A, Thor M, Haq R, Deasy JO, Apte AP. Deep learning-based auto-segmentation of swallowing and chewing structures in CT (2020). bioRxiv 772178; https://doi.org/10.1101/772178.

198. Lei Y, Zhou J, Dong X, et al. Multi-organ segmentation in head and neck MRI using U-Faster-RCNN. In: Landman BA, Išgum I, editors. Medical imaging 2020: image processing. Bellingham, Washington: SPIE; 2020. p. 117. https://doi.org/10.1117/12.2549596.
199. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
200. Dong X, Lei Y, Wang T, et al. Automatic multiorgan segmentation in thorax CT images using U-net- GAN. Med Phys. 2019;46(5):2157–68. https://doi.org/10.1002/mp.13458.
201. Feng X, Qing K, Tustison NJ, Meyer CH, Chen Q. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. Med Phys. 2019;46(5):2169–80. https://doi.org/10.1002/mp.13466.
202. Feng X, Bernard ME, Hunter T, Chen Q. Improving accuracy and robustness of deep convolutional neural network based thoracic OAR segmentation. Phys Med Biol. 2020;65(7):07NT01. https://doi.org/10.1088/1361-6560/ab7877.
203. Schreier J, Attanasi F, Laaksonen H. A full-image deep Segmenter for CT images in breast cancer radiotherapy treatment. Front Oncol. 2019;9:677. https://doi.org/10.3389/fonc.2019.00677.
204. Zhuang X, Li L, Payer C, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med Image Anal. 2019;58:101537. https://doi.org/10.1016/j.media.2019.101537.
205. Morris ED, Ghanem AI, Dong M, Pantelic MV, Walker EM, Glide-Hurst CK. Cardiac substructure segmentation with deep learning for improved cardiac sparing. Med Phys. 2020;47(2):576–86. https://doi.org/10.1002/mp.13940.
206. Rhee DJ, Jhingran A, Kisling K, Cardenas C, Simonds H, Court L. Automated radiation treatment planning for cervical cancer. Semin Radiat Oncol. 2020;30(4):340–7. https://doi.org/10.1016/j.semradonc.2020.05.006.
207. Ahn SH, Yeo AU, Kim KH, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. Radiat Oncol. 2019;14(1):213. https://doi.org/10.1186/s13014-019-1392-z.
208. Quan TM, Hildebrand DGC, Jeong W-K. FusionNet: A deep fully residual convolutional neural network for image segmentation in connectomics. December 2016. http://arxiv.org/abs/1612.05360.
209. Kim H, Jung J, Kim J, et al. Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network. Sci Rep. 2020;10(1):6204. https://doi.org/10.1038/s41598-020-63285-0.
210. Tong N, Gou S, Niu T, Yang S, Sheng K. Self-paced DenseNet with boundary constraint for automated multi-organ segmentation on abdominal CT images. Phys Med Biol. 2020;65(13):135011. https://doi.org/10.1088/1361-6560/ab9b57.
211. Anderson BM, Lin EY, Cardenas C, et al. Automated contouring of variable contrast CT liver images. Adv Radiat Oncol. 2020;6(1):100464. https://doi.org/10.1016/j.adro.2020.04.023.
212. Kim JH, Park SH, Yu ES, et al. Visually Isoattenuating pancreatic adenocarcinoma at dynamic-enhanced CT: frequency, clinical and pathologic characteristics, and diagnosis at imaging examinations. Radiology. 2010;257(1):87–96. https://doi.org/10.1148/radiol.10100015.
213. Koay EJ, Hall W, Park PC, Erickson B, Herman JM. The role of imaging in the clinical practice of radiation oncology for pancreatic cancer. Abdom Radiol. 2018;43(2):393–403. https://doi.org/10.1007/s00261-017-1373-3.
214. Park HS, Lee JM, Choi HK, Hong SH, Han JK, Choi BI. Preoperative evaluation of pancreatic cancer: comparison of gadolinium-enhanced dynamic MRI with MR cholangiopancreatography versus MDCT. J Magn Reson Imaging. 2009;30(3):586–95. https://doi.org/10.1002/jmri.21889.
215. Fu Y, Mazur TR, Wu X, et al. A novel MRI segmentation method using CNN based correction network for MRI guided adaptive radiotherapy. Med Phys. 2018;45(11):5129–37. https://doi.org/10.1002/mp.13221.

216. Liang F, Qian P, Su K-H, et al. Abdominal, multi-organ, auto-contouring method for online adaptive magnetic resonance guided radiotherapy: an intelligent, multi-level fusion approach. Artif Intell Med. 2018;90:34–41. https://doi.org/10.1016/j.artmed.2018.07.001.
217. Elguindi S, Zelefsky MJ, Jiang J, et al. Deep learning-based auto-segmentation of targets and organs-at-risk for magnetic resonance imaging only planning of prostate radiotherapy. Phys Imaging Radiat Oncol. 2019;12:80–6. https://doi.org/10.1016/j.phro.2019.11.006.
218. Dong X, Lei Y, Tian S, et al. Synthetic MRI-aided multi-organ segmentation on male pelvic CT using cycle consistent deep attention network. Radiother Oncol. 2019;141:192–9. https://doi.org/10.1016/j.radonc.2019.09.028.
219. Balagopal A, Kazemifar S, Nguyen D, et al. Fully automated organ segmentation in male pelvic CT images. Phys Med Biol. 2018;63(24):245015. https://doi.org/10.1088/1361-6560/aaf11c.
220. Liu Z, Liu X, Xiao B, et al. Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network. Phys Med. 2020;69:184–91. https://doi.org/10.1016/j.ejmp.2019.12.008.
221. Song Y, Hu J, Wu Q, et al. Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy. Radiother Oncol. 2020;145:186–92. https://doi.org/10.1016/j.radonc.2020.01.020.
222. Men K, Boimel P, Janopaul-Naylor J, et al. A study of positioning orientation effect on segmentation accuracy using convolutional neural networks for rectal cancer. J Appl Clin Med Phys. 2019;20(1):110–7. https://doi.org/10.1002/acm2.12494.
223. International Commission on Radiation Units and Measurements (ICRU). Report 62: Prescribing, Recording and Reporting Photon Beam Therapy (Supplement to ICRU Report 50). Bethesda, MD; 1999.
224. Riegel AC, Berson AM, Destian S, et al. Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion. Int J Radiat Oncol. 2006;65(3):726–32. https://doi.org/10.1016/j.ijrobp.2006.01.014.
225. Breen SL, Publicover J, De Silva S, et al. Intraobserver and Interobserver variability in GTV delineation on FDG-PET-CT images of head and neck cancers. Int J Radiat Oncol. 2007;68(3):763–70. https://doi.org/10.1016/j.ijrobp.2006.12.039.
226. Zhou T, Ruan S, Canu S. A review: Deep learning for medical image segmentation using multi-modality fusion. Array. 2019;3–4:100004. https://doi.org/10.1016/j.array.2019.100004.
227. Guo Z, Li X, Huang H, Guo N, Li Q. Deep learning-based image segmentation on multimodal medical imaging. IEEE Trans Radiat Plasma Med Sci. 2019;3(2):162–9. https://doi.org/10.1109/TRPMS.2018.2890359.
228. Wadhwa A, Bhardwaj A, Singh VV. A review on brain tumor segmentation of MRI images. Magn Reson Imaging. 2019;61:247–59. https://doi.org/10.1016/j.mri.2019.05.043.
229. Feng X, Tustison NJ, Patel SH, Meyer CH. Brain tumor segmentation using an ensemble of 3D U-nets and overall survival prediction using Radiomic features. Front Comput Neurosci. 2020;14:25. https://doi.org/10.3389/fncom.2020.00025.
230. Zhou C, Ding C, Wang X, Lu Z, Tao D. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. IEEE Trans Image Process. 2020;29:4516–29. https://doi.org/10.1109/TIP.2020.2973510.
231. Ben Naceur M, Akil M, Saouli R, Kachouri R. Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. Med Image Anal. 2020;63:101692. https://doi.org/10.1016/j.media.2020.101692.
232. Liu Y, Stojadinovic S, Hrycushko B, et al. A deep convolutional neural network-based automatic delineation strategy for multiple brain metastases stereotactic radiosurgery. PLoS One. 2017;12(10):e0185844. https://doi.org/10.1371/journal.pone.0185844.
233. Grøvik E, Yi D, Iv M, Tong E, Rubin D, Zaharchuk G. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. J Magn Reson Imaging. 2020;51(1):175–82. https://doi.org/10.1002/jmri.26766.

234. Song Q, Bai J, Han D, et al. Optimal co-segmentation of tumor in PET-CT images with context information. IEEE Trans Med Imaging. 2013;32(9):1685–97. https://doi.org/10.1109/TMI.2013.2263388.
235. Beichel RR, Van Tol M, Ulrich EJ, et al. Semiautomated segmentation of head and neck cancers in 18F-FDG PET scans: A just-enough-interaction approach. Med Phys. 2016;43:2948–64. https://doi.org/10.1118/1.4948679.
236. Zeng Z, Wang J, Tiddeman B, Zwiggelaar R. Unsupervised tumour segmentation in PET using local and global intensity-fitting active surface and alpha matting. Comput Biol Med. 2013;43(10):1530–44. https://doi.org/10.1016/j.compbiomed.2013.07.027.
237. Stefano A, Vitabile S, Russo G, et al. An enhanced random walk algorithm for delineation of head and neck cancers in PET studies. Med Biol Eng Comput. 2017;55(6):897–908. https://doi.org/10.1007/s11517-016-1571-0.
238. Berthon B, Evans M, Marshall C, et al. Head and neck target delineation using a novel PET automatic segmentation algorithm. Radiother Oncol. 2017;122(2):242–7. https://doi.org/10.1016/j.radonc.2016.12.008.
239. Comelli A, Stefano A, Benfante V, Russo G. Normal and abnormal tissue classification in positron emission tomography oncological studies. Pattern Recognit Image Anal. 2018;28(1):106–13. https://doi.org/10.1134/S1054661818010054.
240. Huang B, Chen Z, Wu P-M, et al. Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: a dual-center study. Contrast Media Mol Imaging. 2018;2018:1–12. https://doi.org/10.1155/2018/8923028.
241. Guo Z, Guo N, Gong K, Zhong S, Li Q. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys Med Biol. 2019;64(20):205015. https://doi.org/10.1088/1361-6560/ab440d.
242. Deng W, Luo L, Lin X, et al. Head and neck cancer tumor segmentation using support vector machine in dynamic contrast-enhanced MRI. Contrast Media Mol Imaging. 2017;2017:1–5. https://doi.org/10.1155/2017/8612519.
243. Yang J, Beadle BM, Garden AS, Schwartz DL, Aristophanous M. A multimodality segmentation framework for automatic target delineation in head and neck radiotherapy. Med Phys. 2015;42(9):5310–20. https://doi.org/10.1118/1.4928485.
244. Gu Y, Kumar V, Hall LO, et al. Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. Pattern Recogn. 2013;46(3):692–702. https://doi.org/10.1016/j.patcog.2012.10.005.
245. Tan Y, Schwartz LH, Zhao B. Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field. Med Phys. 2013;40(4):043502. https://doi.org/10.1118/1.4793409.
246. Jiang J, Hu Y-C, Liu C-J, et al. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. IEEE Trans Med Imaging. 2019;38(1):134–44. https://doi.org/10.1109/TMI.2018.2857800.
247. Fu X, Bi L, Kumar A, Fulham M, Kim J. Multimodal Spatial Attention Module for Targeting Multimodal PET-CT Lung Tumor Segmentation. July 2020. http://arxiv.org/abs/2007.14728.
248. Byun S, Jung J, Hong H, Oh H, Kim BS. Lung tumor segmentation using coupling-net with shape-focused prior on chest CT images of non-small cell lung cancer patients. In: Hahn HK, Mazurowski MA, editors. Medical Imaging 2020: Computer-aided diagnosis. Bellingham, Washington: SPIE; 2020. p. 90. https://doi.org/10.1117/12.2551280.
249. Tian H, Xiang D, Zhu W, Shi F, Chen X. Fully convolutional network with sparse feature-maps composition for automatic lung tumor segmentation from PET images. In: Landman BA, Išgum I, editors. Medical Imaging 2020: Image processing. Bellingham, Washington: SPIE; 2020. p. 59. https://doi.org/10.1117/12.2548670.
250. Jiang J, Hu Y, Tyagi N, et al. Cross-modality (CT-MRI) prior augmented deep learning for robust lung tumor segmentation from small MR datasets. Med Phys. 2019;46(10):4392–404. https://doi.org/10.1002/mp.13695.

251. Xiao X, et al. Radiomics-guided GAN for segmentation of liver tumor without contrast agents. In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11765. Springer, Cham. https://doi.org/10.1007/978-3-030-32245-8_27.
252. Yang G, Wang C, Yang J, et al. Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal CTA images. BMC Med Imaging. 2020;20(1):37. https://doi.org/10.1186/s12880-020-00435-w.
253. Pang S, Du A, Orgun MA, et al. CTumorGAN: a unified framework for automatic computed tomography tumor segmentation. Eur J Nucl Med Mol Imaging. 2020;47(10):2248–68. https://doi.org/10.1007/s00259-020-04781-3.
254. Cloak K, Jameson MG, Paneghel A, et al. Contour variation is a primary source of error when delivering post prostatectomy radiotherapy: results of the trans-Tasman radiation oncology group 08.03 radiotherapy adjuvant versus early salvage (RAVES) benchmarking exercise. J Med Imaging Radiat Oncol. 2019;63(3):390–8. https://doi.org/10.1111/1754-9485.12884.
255. Unkelbach J, Bortfeld T, Cardenas CE, et al. The role of computational methods for automating and improving clinical target volume definition. Radiother Oncol. 2020;153:15–25. https://doi.org/10.1016/j.radonc.2020.10.002.
256. Belshi R, Pontvert D, Rosenwald J-C, Gaboriaud G. Automatic three-dimensional expansion of structures applied to determination of the clinical target volume in confomal radiotherapy. Radiat Oncol. 1997;37(3):731–6.
257. Shusharina N, Söderberg J, Edmunds D, Löfman F, Shih H, Bortfeld T. Automated delineation of the clinical target volume using anatomically constrained 3D expansion of the gross tumor volume. Radiother Oncol. 2020;146:37–43. https://doi.org/10.1016/j.radonc.2020.01.028.
258. Liu C, Gardner SJ, Wen N, et al. Automatic segmentation of the prostate on CT images using deep neural networks (DNN). Int J Radiat Oncol. 2019;104(4):924–32. https://doi.org/10.1016/j.ijrobp.2019.03.017.
259. Anas EMA, Nouranian S, Mahdavi SS, et al. Clinical target-volume delineation in prostate brachytherapy using residual neural networks. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Lecture Notes in Computer Science, vol 10435. Cham: Springer International Publishing; 2017. p. 365–73. https://doi.org/10.1007/978-3-319-66179-7_42.
260. Men K, Zhang T, Chen X, et al. Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. Phys Med. 2018;50:13–9. https://doi.org/10.1016/j.ejmp.2018.05.006.
261. Teguh DN, Levendag PC, Voet PWJ, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. Int J Radiat Oncol Biol Phys. 2011;81(4):950–7. https://doi.org/10.1016/j.ijrobp.2010.07.009.
262. Yang J, Beadle BM, Garden AS, et al. Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. Pract Radiat Oncol. 2014;4(1):e31–7. https://doi.org/10.1016/j.prro.2013.03.003.
263. Anders LC, Stieler F, Siebenlist K, Schäfer J, Lohr F, Wenz F. Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. Radiother Oncol. 2012;102(1):68–73. https://doi.org/10.1016/j.radonc.2011.08.043.
264. Sarrut D, Claude L, Rit S, Pinho R, Pitson G, Lynch R. Investigating mediastinal lymph node stations segmentation on thoracic CT following experts guidelines. MICCAI, Proc First Int Work Image-Guidance Multimodal Dose Plan Radiat Ther. 2012. http://hal.archives-ouvertes.fr/docs/00/75/52/22/ANNEX/Sarrut.pdf.
265. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. Med Phys. 2017;44(12):6377–89. https://doi.org/10.1002/mp.12602.

266. Men K, Chen X, Zhang Y, et al. Deep Deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. Front Oncol. 2017;7:1–9. https://doi.org/10.3389/fonc.2017.00315.
267. Cardenas CE, Anderson BM, Aristophanous M, et al. Auto-delineation of oropharyngeal clinical target volumes using 3D convolutional neural networks. Phys Med Biol. 2018;63(21):215026. https://doi.org/10.1088/1361-6560/aae8a9.
268. Jin D, et al. Deep esophageal clinical target volume delineation using encoded 3D spatial context of tumors, lymph nodes, and organs at risk. In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. Springer, Cham. https://doi.org/10.1007/978-3-030-32226-7_67.
269. Balagopal A, Nguyen D, Morgan H, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. 2020. http://arxiv.org/abs/2004.13294.
270. Bi N, Wang J, Zhang T, et al. Deep learning improved clinical target volume contouring quality and efficiency for postoperative radiation therapy in non-small cell lung cancer. Front Oncol. 2019;9:1192. https://doi.org/10.3389/fonc.2019.01192.

Part III

Machine Learning for Radiation Oncology Workflow

12 Machine Learning Applications in Quality Assurance of Radiation Delivery

Gilmer Valdes, Alon Witztum, and Maria F. Chan

12.1 Introduction

Radiotherapy delivery is a complex process which requires QA testing at each step in order to prevent errors and ensure the prescribed treatment is received correctly by the patient. This chapter will review the application of machine learning to radiotherapy QA. Machine learning, as it refers to the automated detection of meaningful patterns in data, has become a major area of research and a common tool in many processes in radiotherapy [1]. Simultaneously, the number of daily QA tasks performed by medical physicists has increased, and high importance has been set to prioritizing those most associated with delivering the safest treatment by the American Association of Physicists in Medicine (AAPM) Task Group 100 [2]. It therefore follows that as more complex treatments are adopted, learning from QA data to identify those tasks that require early intervention becomes increasingly crucial to our profession. However, QA data are currently only utilized to decide whether a specific test result falls within the given specifications. This chapter hopes to demonstrate the power of machine learning and the advantages it can offer in QA programs.

G. Valdes (✉) · A. Witztum
University of California at San Francisco, San Francisco, CA, USA
e-mail: Gilmer.Valdes@ucsf.edu

M. F. Chan
Department of Medical Physics, Memorial Sloan Keterring, NY, NY, USA

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_12

12.2 Overview of the Use of Machine Learning in Quality Assurance and Treatment Delivery

Machine learning is well suited to analyzing multiple elements of a radiotherapy QA program. As a brief overview, the performance of linear accelerators (Linac) over time have been predicted by Li et al. [3]. More specifically, Carlson et al. used machine learning to predict the positional errors of the multi-leaf collimator (MLC) [4]. Valdes et al. were able to automatically detect problems with the Linac imaging system [5] and also continued to develop an ML method to predict IMRT QA passing rates [6, 7]. More broadly, El Naqa et al. developed an anomaly detection system to model QA errors and rare events in radiotherapy [8], and Ford et al. quantified the error detection efficiency of radiotherapy quality control checks [9]. These examples have demonstrated the ability of machine learning algorithms to identify QA cases that demand closer investigation as recommended by Task Group 100 [2]. This chapter will now discuss the application of machine learning to automated chart review, Linac QA, and IMRT QA in more detail.

12.2.1 Automated Chart Review

Medical physicists are tasked with verifying the integrity of each plan before it is delivered to the patient and through its course of treatment in a task commonly known as chart review. These reviews occur before the first fraction is delivered to the patient and then every week until the final treatment verification is performed. The AAPM Task Group 275 reviews this process and offers recommendations [10]. This task group report highlights different aspects of the plan that need to be checked before the plan is delivered: technical parameters and data transfer from the Treatment Planning System and the Treatment Console, calculation accuracy, image guidance, plan quality, and consideration of technical clinical factors. This check is highly important, and it constitutes the last safeguard against many mistreatments as 33% of near-miss incidents originate during the treatment preparation process [11]. However, due to the repetitive nature of the task and the overwhelming information needed to be processed by the reviewer, only 25–38% of all detectable errors are actually identified by the physicists [12, 13]. As such, the use of machine learning to help physicists perform this task is a natural and important application. Some examples include the use of clustering methods to detect outliers based on plan comparisons [14, 15] and Bayesian network architectures to detect those links in the workflow that are more likely to introduce errors [16].

12.2.2 Machine Learning Applied to Delivery Systems

This section will now focus on the application of machine learning for Linac QA. One element of the Linac that can have a major impact on the dose delivery is the multi-leaf collimator (MLC). Discrepancies in MLC positions were first

predicted using machine learning techniques by Carlson et al. [4] Leaf position and speed were calculated as predictive leaf motion parameters for the models in this study. The positional differences between the DICOM-RT files and the DynaLog files was used as the target to train ML models. The authors used three machine learning algorithms—linear regression, random forest, and a cubist model, with the cubist model performing best in terms of accuracy. In the future, these predictions could be incorporated into the treatment planning system to enable clinicians to visualize a more realistic dose distribution.

A machine learning model was developed by Chuang et al. to predict MLC discrepancies during delivery using prior trajectory log files [17]. Using a workflow that extracted discrepancies and mechanical parameters from these log files, the authors built multiple machine learning models including linear regression, decision tree, and ensemble methods to predict these discrepancies.

Separately to MLC discrepancies, the performance of the Linac over time has been predicted by Li and Chan by applying ANN time-series prediction modeling to longitudinal daily Linac QA data over a 5-year time period [3]. The network architecture was formed using a trial-and-error process which resulted in a set of one hidden layer, six hidden neurons, and two input delays. A benchmark autoregressive integrated moving average (ARIMA) model was used for comparison and was found to be less accurate than the new ANN time-series model.

More recently, QA results that were outside the suggested AAPM Task Group 142 tolerance limits were detected by Naqa et al. using a Support Vector Data Description approach on 119 EPID images from 8 Linacs [8]. Recently, equally important work on predicting different QA aspects of proton therapy is starting to emerge. In this case, usually predicting output from the proton machines is also intrinsically associated with patient-specific QA. Sun et al. developed a machine learning algorithm based on ensemble methods to predict monitor units for a compact proton machine. These models outperformed previously developed empirical approaches [18]. Additionally, Grewal et al. used a machine learning-based approach to predict outputs in uniform scanning proton therapy. The authors concluded that models that used machine learning outperformed previous models developed using empirical equations [19]. Table 12.1 lists recent studies applying ML methods to Linac QA that are discussed in this section.

12.2.3 Machine Learning Applied to IMRT QA

The section will focus on the application of machine learning to IMRT QA. The general workflow extracts features from each treatment plan and computes many complexity metrics which are associated with passing rates. Models are built using these features to predict passing rates for new plans.

The first such virtual IMRT QA model using machine learning was developed by Valdes et al. using 498 clinical IMRT plans from the University of Pennsylvania with associated QA passing rate results obtained using a MapCHECK (Sun Nuclear Corporation, Melbourne, FL) QA device [6]. A further data set of 203 clinical IMRT

Table 12.1 Summary of studies on machine QA using machine learning techniques

Group	QA source	Data set	ML model	Task
Carlson et al., *PMB, 2016* [4]	DICOM_RT, Dynalog files	74 VMAT plans	Regression, random Forest, cubist	MLC position errors detection
Li & Chan, *AMLS, 2017* [3]	Daily QA device	5-year daily QA data	ANN time-series, ARMA models	Symmetry prediction
Sun et al., Med Phys, 2018 [18]	Ion chamber	1754 proton fields	Random Forrest, XGboost, cubist	Output for a compact proton machine.
Naqa et al., *Med Phys, 2019* [8]	EPID	119 images from 8 Linacs	Support vector data description, clustering	Gantry sag, radiation field shift, MLC offset
Chuang et al., Med Phys, 2021 [17]	Trajectory log files	116 IMRT plans, 125 VMAT plans	Boosted tree outperformed LR	MLC discrepancies during delivery & feedback
Grewal et al., Med Phys, 2020 [19]	Ion chamber	4231 proton fields	Gaussian processes and shallow neural networks	Proton output and patient QA

plans was obtained from Memorial Sloan Kettering Cancer Center with associated QA passing rates acquired in this case using portal dosimetry [7]. All plans from both the data sets were planned in Eclipse (Varian Medical Systems, Palo Alto, CA). Parameters were automatically extracted for each IMRT beam from Eclipse using SQL queries. MLC positions and collimator rotations were extracted using scripts. Features were calculated for each beam by developing MATLAB functions (The MathWorks Inc., Natick, MA). A machine learning algorithm was trained to learn the relationship between the plan characteristics and the passing rates. The learning curve for the initial model demonstrated that approximately 200 composite plans are required for sufficient training. The learning curve for the portal dosimetry model showed that approximately 100 IMRT fields are sufficient for training a reliable model if the original model trained at the University of Pennsylvania is used as the starting point.

For the MapCHECK data, 78 features were extracted, and the most important features included the fraction of area delivered outside a circle with 20 cm radius (to capture symmetry disagreements), duty cycle, and fraction of opposed MLCs with aperture smaller than 5 mm (to quantify the effects of rounded leaves in the MLC). For the portal dosimetry data, an additional 10 features were calculated to account for the characteristics of portal dosimetry, and the important features included the CIAO area, the fraction of MLC leaves with gaps smaller than 20 or 5 mm, and the fraction of area receiving less than 50% of the total calibrated MUs.

There was a strong correlation between the MapCHECK measured QA passing rate and the predicted passing rate using the virtual IMRT model using data that had not previously been seen. All passing rate predictions were within a 3% error. Even

though passing rates are dependent on the site, different models were not built for the University of Pennsylvania and Memorial Sloan Kettering Cancer Center because, conditional on plan characteristics, this dependency disappears.

Implementation of virtual IMRT QA clinically requires the following workflow: (1) collect or extract IMRT QA data, (2) extract parameters of the IMRT fields from plans, (3) extract features and calculate complexity metrics, (4) apply a machine learning algorithm to build a predictive model. The most important features to predict passing rate should also be identified and can inform the physicists of possible failure modes that need to be addressed to tight up the QA program.

This process requires calculating features that correlate between the plan characteristics and passing rates. Interian et al., from the Valdes group, compared their own Poisson regression model using the same QA data to a model built using a Deep Neural Network capable of designing its own features [20]. Fluence maps for each plan were used as input to the convolution neural network (CNN), a specialized neural network for image analysis, and the models were trained using TensorFlow and Keras to predict QA passing rates. The CNN and virtual IMRT QA predictions were comparable even though the virtual IMRT QA model used features designed by physics experts. The authors concluded that CNNs with transfer learning can predict IMRT QA passing rates by automatically designing features from the fluence maps without human expert supervision.

In another study to predict IMRT QA gamma evaluation results, Tomori et al. applied deep learning methods to 60 IMRT QA plans [21]. Fifteen-layer convolutional neural networks were developed to learn the planar dose distributions from a QA phantom, and EBT3 film was used to measure the gamma passing rate. The volume of PTV, rectum, and overlapping region, and the monitor unit for each field were included as input to the model. The CNN was built using fivefold cross-validation to predict gamma passing rates at various criteria: 2%/2 mm, 3%/2 mm, 2%/3 mm, and 3%/3 mm. A linear correlation was found between measured and predicted values for all criteria which suggests that deep learning methods can predict gamma evaluation results for IMRT QA.

Deep learning can also be used to classify potential treatment delivery errors as demonstrated by Nyflot et al., who exported three sets of planar doses from QA plans corresponding to the error-free case, a random MLC error case, and a systematic MLC error case [22]. The plans were delivered to an EPID panel, and the EPID dosimetry software was used to perform gamma analysis. Two radiomic approaches (image features using a CNN and human-designed texture features) were used to identify metrics for input into four ML classifiers which were used to determine whether the images contained errors. The CNN performance was superior to the texture feature approach, and both the radiomic approaches were superior to using the gamma passing rate in their ability to predict clinically relevant errors.

Another approach to predicting results of VMAT QA measurements was used by Granville et al. which incorporated both treatment plan characteristics and Linac performance metrics [23]. This study used a support vector classifier (SVC) with the model output classes representing the median dose difference (±1%) between measured and expected dose distributions rather than passing rates.

During development of this model, unimportant features were removed using a recursive feature elimination (RFE) technique with cross-validation. The ten most important features for prediction consisted of five features representing treatment plan characteristics and five features representing Linac performance metrics. This model demonstrated the potential of using both machine and plan characteristics to predict QA results.

Delivery characteristics can also impact the dose accuracy of treatment plans. Li et al. extracted ten metrics from 344 QA plans and found that leaf speed is the most important factor affecting the accuracy of gynecologic, rectal, and head and neck plans [24]. They also found that the field complexity, small aperture score, and MU are the most important factors influencing the accuracy of prostate plans. Li et al. also explored the accuracy of VMAT QA result prediction using machine learning to build two prediction models: the classic Poisson regression model and a new Random Forest classification model [25]. The model performance was assessed under different gamma criteria and action limits with tenfold cross-validation on 255 VMAT plans. An independent validation set of 48 VMAT plans was used to validate these models without cross-validation. The accuracy of the prediction was greatly affected by the absolute value of the measured gamma passing rates and gamma criteria. Even though the passing rates for the majority of VMAT plans were accurately predicted by the regression model, the classification model had a much better sensitivity to accurately detect failed QA plans. Lam et al. applied tree-based machine learning models consisted of 1269 IMRT beams to improve the prediction accuracy of portal dosimetry at 2%/2 mm gamma criteria with a 5% threshold [26]. Later the same group of Li et al. [25] continued their work and improved on their model by using autoencoder-based classification-regression (ACLR) to generate GPR predictions for three different gamma criteria with 54 complexity metrics as input to the model [27]. This hybrid model was able to improve prediction accuracy over the classic Poisson Lasso regression model.

The validity of virtual IMRT QA was once again shown by Hirashima et al., who created a model to predict ArcCHECK measurements using plan complexity and dosiomic features as input to Gradient Boosting, considered the most accurate algorithm to date for the analysis of tabular data [28]. Table 12.2 summarizes the studies discussed in this section on virtual IMRT/VMAT QA in chronological order.

12.3 Future Directions

Radiation treatments are complex and require an extensive QA process to guarantee that they can be safely delivered to patients. With the increasing complexity of these techniques, the number of QA tasks that need to be implemented have grown exponentially. This has prompted the American Association of Physics in Medicine to propose, through Task Group 100, the prioritization of those tasks that are most important to guaranteeing the safe delivery of radiation therapy. Machine learning techniques, with their natural mechanism to process large amount of data, appear to be a valuable companion to facilitate QA tasks. Extensive demonstrations have been cited within this chapter exemplifying their use. They can be used to detect outliers

Table 12.2 Summary of studies on patient-specific QA using machine learning techniques

Group	TPS/ delivery	QA tool	Data source	ML model	Research highlight
Valdes et al. *Med Phys, 2016* [6]	Eclipse/ Varian	MapCHECK2	498 IMRT Plans	Poisson regression	Founding paper
Valdes et al. *JACMP, 2017* [7]	Eclipse/ Varian	Portal dosimetry	203 IMRT Beams	Poisson regression	Multi-sites validation
Interian et al. *Med Phys, 2018* [20]	Eclipse/ Varian	MapCHECK2	498 IMRT Plans	Convolutional neural network	Fluence map as input
Tomori et al. *Med Phys, 2018* [21]	iPlan/ Varian	EBT3 film	60 IMRT Plans	Convolutional neural network	Planar dose, volumes, MU
Nyflot et al. *Med Phys, 2019* [22]	Pinnacle/ Elekta	EPID	186 IMRT Beams	Convolutional neural network	Image, texture features
Granville et al. *PMB, 2019* [23]	Monaco/ Elekta	Delta4	1620 VMAT beams	Support vector classifier	1st VMAT & w/QC metrics
Li et al. *Red Journal, 2019* [24]	Eclipse/ Varian	MatriXX	248 VMAT beams	Poisson lasso & random forest	Specificity & sensitivity
Lam et al. *Med Phys, 2019* [26]	Eclipse/ Varian	Portal dosimetry	1269 IMRT beams	Ada-boost, RF, XGBoost	High accuracy at 2%/2 mm, 5% TH
Wang et al. *PMB, 2020* [27]	Eclipse/ Varian	MatriXX	400 VMAT beams	Hybrid model ACLR	High prediction accuracy
Hirashima et al. *RO, 2020* [28]	Acuros/ collapsed cone	ArcCHECK	1255 VMAT plans	XGBoost	Plan complexity & dosiomics

and alert physicists to take proactive actions and make informed decisions about those plans and the aspect of QA that require immediate attention. However, it is important to note that in the ever-changing world of radiotherapy, it is likely that the accuracy of these algorithms created to detect errors would decrease with time. As such, they should only be used to assist the physicists and facilitate the task and not as a replacement for human supervision. Additionally, the mechanisms and guidelines to perform QA on these algorithms are lacking and more work is needed to ensure a safe transition into ML-assisted QA programs.

Most machine learning application to QA has been to predict QA passing rates using a variety of methods. It is important to gain a full understanding of all contributing factors to delivery accuracy and QA failures in order to be able to implement a risk-based program as suggested in the AAPM TG-100 report. Future development could allow for the inclusion of QA predictions in the treatment planning system, so that the optimizer can ensure a passing QA result. This would allow

physicists to concentrate on running QA tests only on those plans with the lowest expected passing rates. A clinically implemented ability to predict QA results would have profound implications on the current treatment workflow and corresponding time to treatment for our patients.

Conflict of Interest The authors declare that they have read the chapter and there are no competing interests.

References

1. Feng M, Valdes G, Dixit N, Solberg TD. Machine learning in radiation oncology: opportunities, requirements, and needs. Front Oncol. 2018;8:110. https://doi.org/10.3389/fonc.2018.00110.
2. Huq MS, Fraass BA, Dunscombe PB, Gibbons JP Jr, Ibbott GS, Mundt AJ, et al. The report of task group 100 of the AAPM: application of risk analysis methods to radiation therapy quality management. Med Phys. 2016;43:4209–62. https://doi.org/10.1118/1.4947547.
3. Li Q, Chan MF. Predictive time series modeling using artificial neural networks for Linac beam symmetry—an empirical study. Ann N Y Acad Sci. 2017;1387(1):84–94. https://doi.org/10.1111/nyas.13215.
4. Carlson JN, Park JM, Park SY, et al. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. Phys Med Biol. 2016;61:2514. https://doi.org/10.1088/0031-9155/61/6/2514.
5. Valdes G, Morin O, Valenciaga Y, Kirby N, Pouliot J, Chuang C. Use of TrueBeam developer mode for imaging QA. J Appl Clin Med Phys. 2015;16:5363. https://doi.org/10.1120/jacmp.v16i4.5363.
6. Valdes G, Scheuermann R, Hung CY, et al. A mathematical framework for virtual IMRT QA using machine learning. Med Phys. 2016;43(7):4323–34. https://doi.org/10.1118/1.4953835.
7. Valdes G, Chan MF, Lim S, Scheuermann R, Deasy JO, Solberg TD. IMRT QA using machine learning: a multi-institutional validation. J Appl Clin Med Phys. 2017;18(5):278–84. https://doi.org/10.1002/acm2.12161.
8. Naqa IE, Irrer J, Ritter TA, et al. Machine learning for automated quality assurance in radiotherapy: a proof of principle using EPID data description. Med Phys. 2019;46(4):1914–21. https://doi.org/10.1002/mp.13433.
9. Ford EC, Terezakis S, Souranis A, Harris K, Gay H, Mutic S. Quality control quantification (QCQ): a tool to measure the value of quality control checks in radiation oncology. Int J Radiat Oncol Biol Phys. 2012;84:e263–9. https://doi.org/10.1016/j.ijrobp.2012.04.036.
10. Ford E, Conroy L, Dong L, et al. Strategies for effective physics plan and chart review in radiation therapy: report of AAPM task group 275. Med Phys. 2020;47(6):e236–72.
11. Novak A, Nyflot MJ, Ermoian RP, Jordan LE, et al. Targeting safety improvements through identification of incident origination and detection in a near-miss incident learning system. Med Phys. 2016;43(5):2053–62.
12. Ezzell G, Chera B, Dicker A, et al. Common error pathways seen in the RO-ILS data that demonstrate opportunities for improving treatment safety. Pract Rdiat Oncol. 2018;8:123–32.
13. Gopan O, Zeng J, Novak A, et al. The effectiveness of pretreatment physics plan review for detecting errors in radiation therapy. Med Phys. 2016;43:5181.
14. Azmandian F, Kaeli D, Dy JG, et al. Towards the development of an error checker for radiotherapy treatment plans: a preliminary study. Phy Med Biol. 2007;52:6511–24.
15. Furhang EE, Dolan J, Sillanpaa JK, et al. Automating the initial physics chart-checking process. J Appl Clin Med Phys. 2009;10:129–35.
16. Bojechko C, Philips M, Kalet A, et al. A quantification of the effectiveness of EPID dosimetry and software-based plan verification systems in detecting incidents in radiotherapy. Med Phys. 2015;42:5363.

17. Chuang K.-C, Giles W, Adamson J. A tool for patient-specific prediction of delivery discrepancies in machine parameters using trajectory log files. Med. Phys. 2021;48:978–90. https://doi.org/10.1002/mp.14670.
18. Sun B, Lam D, Yang D, et al. A machine learning approach to the accurate prediction of monitor units for a compact proton machine. Med Phys. 2018;45(5):2243–51.
19. Grewal HS, Chacko MS, Ahmad S, et al. Prediction of the output factor using machine and deep learning approach uniform scanning proton therapy. J Appl Clin Med Phys. 2020;21(7):128–34. https://doi.org/10.1002/acm2.12899.
20. Interian Y, Rideout V, Kearney VP, et al. Deep nets vs expert designed features in medical physics: an IMRT QA case study. Med Phys. 2018;45(6):2672–80. https://doi.org/10.1002/mp.12890.
21. Tomori S, Kadoya N, Takayama Y, et al. A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance. Med Phys. 2018;45(9):4055–65. https://doi.org/10.1002/mp.13112.
22. Nyflot MJ, Thammasorn P, Wooton LS, et al. Deep learning for patient-specific quality assurance: identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks. Med Phys. 2019;46(2):456–64. https://doi.org/10.1002/mp.13338.
23. Granville DA, Sutherland JG, Belec JG, et al. Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. Phys Med Biol. 2019;64:095017. https://doi.org/10.1088/1361-6560/ab142e.
24. Li J, Zhang X, Li J, et al. Impact of delivery characteristics on dose accuracy of volumetric modulated arc therapy for different treatment sites. J Radiat Res. 2019;60(5):603–11. https://doi.org/10.1093/jrr/rrz033.
25. Li J, Wang L, Zhang X, et al. Machine learning for patient-specific quality assurance of VMAT: prediction and classification accuracy. Int J Radiat Oncol Biol Phys. 2019;105(4):893–902. https://doi.org/10.1016/j.ijrobp.2019.07.049.
26. Lam D, Zhang X, Li H, et al. Predicting gamma passing rates for portal dosimetry-based IRMT QA using machine learning. Med Phys. 2019;46(10):4666–75.
27. Wang L, Li J, et al. Multi-task autoencoder based classification-regression (ACLR) model for patient-specific QA of VMAT. Phys Med Biol. 2020;65(23):235023.
28. Hirashima H, Iramina H, Mukumoto N, et al. Improvement of prediction and classification performance for gamma passing rate by using plan complexity and dosiomics features. Radiat Oncol. 2020;153:250–7.

Knowledge-Based Treatment Planning

13

Jiahan Zhang, Yaorong Ge, and Q. Jackie Wu

13.1 Introduction

Radiation therapy is a widely adopted and effective cancer treatment modality that leverages highly advanced and complex technologies. With the advent of intensity modulated radiation therapy (IMRT), physicians have a tremendous opportunity to maximize cancer control while minimizing toxicity to normal organs. However, achieving this inherently contradicting goal using IMRT requires extensive knowledge, experience, and time due to the complexity of technologies and the limitations in our understanding of patient conditions.

To tackle the challenges of radiation therapy, knowledge-based systems have been developed as early as 1980s to aid the design of radiation treatment plans [1, 2]. The knowledge-based systems reported during that period refer mainly to expert-based systems that aim to capture clinician knowledge and experience in terms of rules and algorithms. These rule-based approaches in recent years have led to a type of system that is commonly called "automatic (or automated) planning systems" (e.g., [3–5]). These systems aim to encode sophisticated planning knowledge that human planners have acquired through their planning experience into complex and often iterative algorithms to generate clinically optimal IMRT plans automatically. Note that these automatic planning systems and the earlier form of knowledge-based systems are not data-driven in the sense that their main algorithms are based on direct encoding of human knowledge and do not rely on predictive models that are based on a database of prior planning data.

J. Zhang · Q. J. Wu (✉)
Department of Radiation Oncology, Duke University Medical Center, Durham, NC, USA
e-mail: jackie.wu@duke.edu

Y. Ge
College of Computing and Informatics, The University of North Carolina at Charlotte, Charlotte, NC, USA

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_13

As IMRT experience and especially the carefully designed clinical plans are accumulated over the past two decades, a new set of data-driven methods have been developed in recent years with an aim to improve the quality and efficiency of IMRT planning by learning from the past high-quality clinical plans. The term "knowledge-based planning" or simply KBP has now frequently been used to refer to this specific class of data-driven approaches to IMRT planning. Some of this development has led to commercial products recently and allowed the investigation of KBP in numerous clinical applications. This has somewhat solidified the narrower definition of KBP that draws knowledge from a database of prior clinical plan data and assumes that other sources of knowledge, such as treatment trade-off and clinician experience, are embedded in the design of prior clinical plans. We note that recent studies [6] have strongly affirmed this assumption and highlight a potentially significant advantage of the data-drive KBP approach over automatic planning systems because sophisticated clinician experience such as trade-off decisions is difficult to encode into rules or algorithms.

At the center of the KBP technologies is the KBP models that predict the plan parameters (e.g., dose volume constraints, optimal beam angles, and fluence maps) that are best achievable for each patient. Since the KBP models are learned from high-quality prior plans created and delivered at a clinical entity (e.g., one cancer center or many cancer centers in a clinical trial), the predicted plan parameters reflect what are best achievable by the collective knowledge of the clinical entity and thus are in a sense optimal for the patients. Here lies a potential limitation of the KBP approach, that is, the quality of KBP may be limited by the quality of prior plans used to build the KBP models. This limitation has been addressed in a number of ways. To ensure that prior plans are of the highest quality before a KBP model is built, we can leverage the KBP technology to iteratively improve the overall quality of the prior plan database. Another solution is to focus on improving the overall plan quality of many centers with diverse resources by leveraging KBP models that are built on plans created by most experienced planners.

The plan parameters predicted by KBP models can be of many different types and forms. Due to relatively small training datasets, early models tend to predict summary dose parameters such as the dose volume histogram (DVH). Various approaches to predicting voxel level dose distribution [7, 8] have been attempted with limited success. With the advent of deep learning models in recent years, voxel level dose prediction has made significant progress in a number of cancer sites [9–11]. Moreover, direct prediction of fluence maps has also shown promising results [12–14]. Another important type of KBP models aims to predict the beam configurations. These models are essential for complex cancer cases where simple geometry with fixed and co-planar beam angles is not sufficient [15].

The development of KBP technologies faces many challenges. This is in part reflected in the fact that almost every modeling and machine learning technique has been applied to the development of various KBP models. One challenge that is becoming less prominent, but is still significant, is the lack of high-quality prior plan. This lack of data has required researchers to explore atlas-based models and other similarity-based machine learning approaches. It has introduced the need for

incremental learning as new plan data continue to accumulate and has also introduced the challenge for handling outliers because new plan data may come from very different type of patients and potentially with variable plan quality. Another major challenge in KBP modeling is the large number and complexity of the factors that are involved. Given the limited data samples, this challenge cannot be addressed simply by increasing the complexity of the models. Thus, careful design of model architecture that can tease out variations in the data samples [16–18] and proper design of features that can effectively extract predictive factors [19, 20] are often important tasks in developing successful KBP models. Finally, KBP is not just dose models or beam models. A successful KBP technology requires a complete and efficient planning workflow that seamlessly integrates multiple models, algorithms, and also planner input.

In the following sections, we will discuss some of the challenges of KBP development in more detail and introduce examples of solutions that effectively address these challenges.

13.2 Anatomical Feature-Based KBP Model

During IMRT treatment planning, the physician assigns organ-at-risk (OAR) dose constraints based on patient-specific anatomy and disease-specific considerations. The planner then sets optimization constraints in the TPS to achieve those constraints and make additional efforts to spare OARs as much as possible without compromising target coverage. This process is highly subjective and may introduce unnecessary plan quality variations. A data-driven approach has been developed to reliably predict the best achievable OAR dose sparing using knowledge embedded in previous treatment plans [19]. In this approach, we establish the correlations of OAR dose volume histograms (DVHs) and anatomy features for previously treated patients and use the correlation relationship to predict best achievable DVHs and guide treatment planning.

13.2.1 Distance to Target Histogram

DVHs are essentially one-dimensional representations of three-dimensional dose distributions. The reduced dimension makes DVH curves useful in interpreting OAR toxicity quantitatively. Similarly, to predict OAR DVHs, we need to first find low dimension representations of the spatial relationships between OAR voxels and the PTV. Considering that the primary predictor for dose distribution inside the patient is the distance from the PTV, we define the distance to target histogram (DTH) as a histogram of OAR volume fractions within certain distances from the PTV surface. The Euclidean form of the distance function r from an OAR voxel $v_{OAR}{}^{i}$ to the PTV surface, $r(v_{OAR}{}^{i}, PTV)$ is

$$r\left(v_{OAR}{}^{i}, PTV\right) = \min_{k}\left\{\left\|v_{OAR}{}^{i} - v_{PTV}{}^{k}\right\| \,\middle|\, v_{PTV}{}^{k} \in S_{PTV}\right\}.$$

Negative signs are assigned to the distance values for OAR voxels inside PTV. To improve the correlation of DTHs to DVHs, a modification to the Euclidean distance is made to account for the slower dose fall-off in the regions far away from the PTV. First, the voxel $v_{PTV}^{\min}$ on PTV surface which is closest to $v_{OAR}{}^{i}$ is located. The Euclidean distance $r(v_{OAR}{}^{i},PTV)$ is then reduced by a factor α ($\alpha < 1$) when the OAR voxel $v_{OAR}{}^{i}$ is outside a axial cutoff distance $r_{XY}{}^{cutoff}$ from $v_{PTV}^{\min}$,

$$r'\left(v_{OAR}{}^{i}, PTV\right) = \alpha\, r\left(v_{OAR}{}^{i}, PTV\right),$$

where $r'(v_{OAR}{}^{i}, PTV)$ is the modified distance.

13.2.2 Model Training and Validation

DTHs and DVHs are both continuous functions, and it is difficult to model DVHs directly as a function of DTHs. To effectively correlate the variations of DVHs to that of DTHs, principal component analysis (PCA) is applied to both type of histograms to reduce their dimensions. With PCA, much of the variability of the histograms can be explained by a small number of principle components. In particular, the first three components of the principal component scores (PCS) are selected as anatomical features. Additional anatomical factors, including OAR volume, PTV volume, fraction of OAR volume overlapping with PTV volume, and fraction of OAR volume outside the treatment fields, are combined with DTH PCS to form feature vectors.

One of the earliest KPB models studied at Duke University Medical Center used 88 prostate IMRT plans from the clinical database. The dose prescriptions for these patients are identical: the PTV enclosing the prostate and the seminal vesicles plus a standard margin of 5 mm is prescribed to 54 Gy, and the PTV enclosing the prostate is boosted to 76 Gy. These clinical plans were generated using the institutional prostate IMRT planning protocol with seven standard coplanar 15 MV beams at: 205°, 255°, 310°, 0°, 50°, 105°, and 55° gantry angles. Among these plans, 64 prostate patients are selected for training purpose and the remaining 24 prostate plans are reserved for validation.

The training workflow is summarized in Fig. 13.1. After data extraction and preprocessing, DVH PCS are fitted to anatomical features. To quantitatively establish the correlations, stepwise multiple regression method is utilized. The stepwise regression method selects the most significant feature to the model by the coefficient of partial determination, which measures the correlation between that factor and the DVH variation not explained by the factors already included in the model. The result of the training process is a KBP model consisted of regression coefficients and the DTH/DVH PCA basis vectors. The model can then be used to predict OAR DVHs for future patients.

Following the flowchart shown in Fig. 13.2, the DVHs of bladder and rectum for plans in the validation datasets are calculated by the regression model trained by the training dataset. These model-predicted DVHs are compared to their corresponding DVHs in the actual clinical plans to assess the effectiveness of the factors identified in the study. If the factors used in the trained regression model capture significant

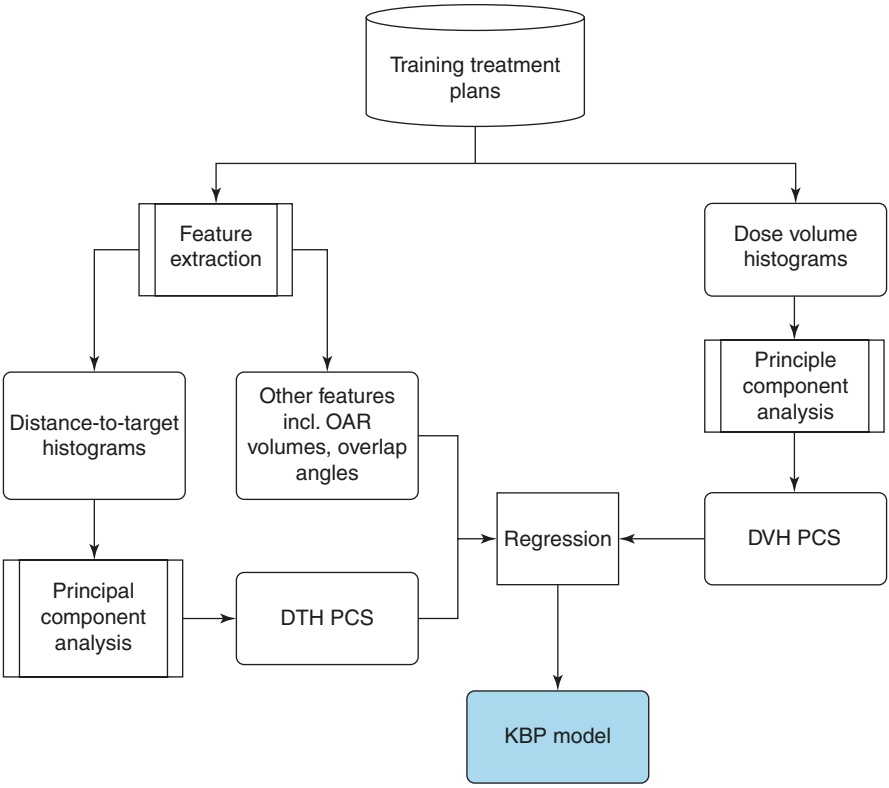


Fig. 13.1 The workflow of the KBP training process

portion of the interpatient OAR dose sparing variation, the model should be able to predict the DVHs in the validation datasets. The comparison of DVHs for a subset of the validation plans are shown in Fig. 13.3.

To quantify the level of agreement between the modeled DVHs and the actual plan DVHs, specific dose–volume parameters are analyzed. For the prostate validation plans, the volumes corresponding to 99%, 85%, and 50% of the prescribed dose in the modeled plans are compared with those values in the actual plans. For both the bladder and rectum, 17 out of 24 plans (71%) are within 6% error bound and 21 (85%) are within 10% error bound.

13.3 A Robust Ensemble Model with Outlier Filtering Mechanism

Forward selection, a type of stepwise regression, is used in the model discussed in Sect. 13.2. It finds the most significant features to add step by step, hence the name. The selected features are fitted to the data with ordinary least square, while the rest

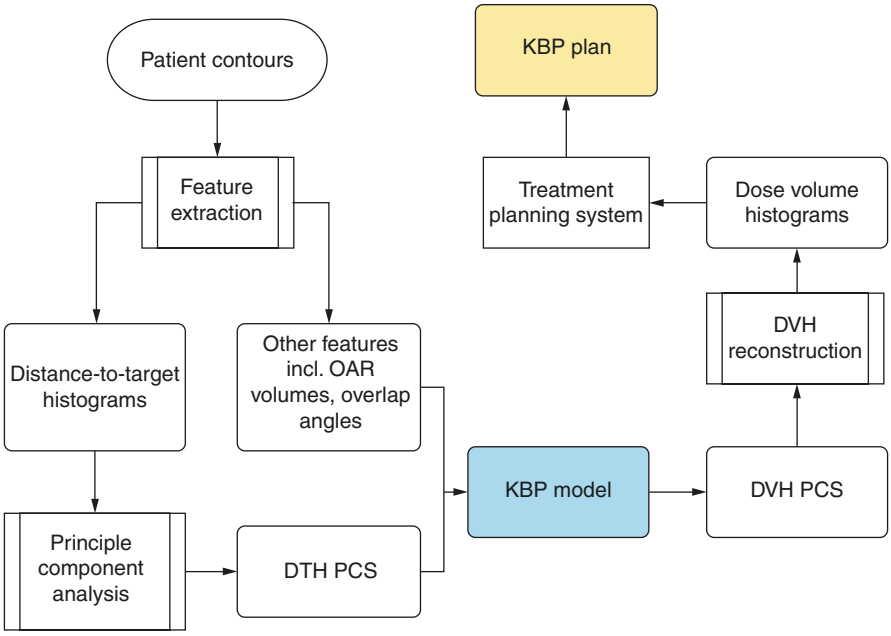


Fig. 13.2 The workflow of a KBP prediction

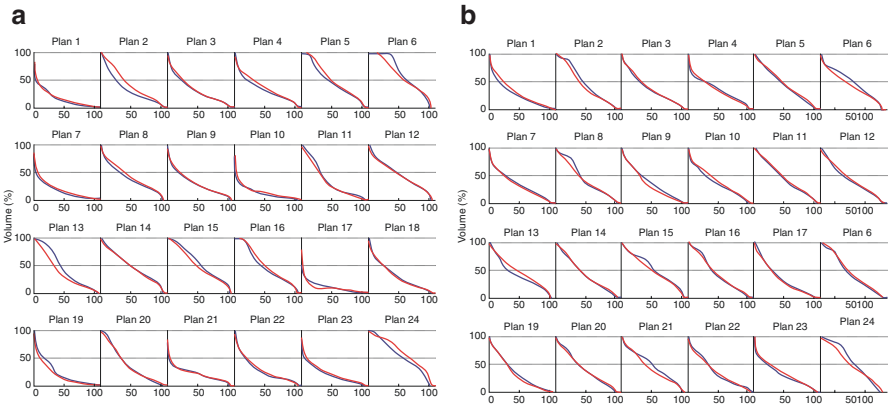


Fig. 13.3 Comparison of actual plan DVHs and the model predicted DVHs. The plans shown are a subset of the validation plans: (**a**) bladder and (**b**) rectum

of the features are discarded. There are some potential issues about this procedure, resulting in instabilities of the model training process. To address the model robustness issue and make KBP modeling more accessible to clinical environments, we developed an ensemble model [18] that takes advantage of various modeling methods and has a built-in outlier filter mechanism.

13.3.1 An Ensemble KBP Model

We first formulate the KBP model regression and describe four types of regression models, including ridge regression [21, 22], lasso [23], elastic net [24], and stepwise regression. These models also serve as base learners for the final ensemble model. The first three models share the same objective function

$$\beta = \arg\min\left\{\left\|Y - X\beta\right\|_2^2 + \varphi\left(\beta\right)\right\},$$

where $X \in \mathbb{R}^{N\times P}$ denotes P feature value from N training cases, $Y \in \mathbb{R}^N$ denotes OAR DVH principle component scores (PCS) of cases in the training set, and $\beta \in \mathbb{R}^P$ denotes regression coefficients corresponding to P anatomical features, such as PCS of distance-to-target histogram (DTH). The last term, known as the penalty term, balances the bias and variance of the trained model. The goal of KBP is to obtain regression coefficients β based on cases previously planned by experienced planners, and when a new case needs to be planned, the optimal OAR DVH can be calculated simply using the model predicted PCS of $X\beta$. In ridge regression, the penalty term $\varphi(\beta)$ is the square of ℓ 2-norm of the regression coefficients β; in lasso, the penalty term is the ℓ 1-norm of β; and in elastic net, the penalty term is simply a linear combination of ℓ 1-norm and ℓ 2-norm squared: $\varphi\left(\beta\right) = \lambda_1\left\|\beta\right\|_1 + \lambda_2\left\|\beta\right\|_2^2$.

In different settings, different regression models perform well, and none of these models consistently perform better than other models. Therefore, to improve model consistency, we present an ensemble model that combines the aforementioned individual models using a model stacking method. A previous study demonstrated that, even stacking ridge regression models alone with different penalty weight λ improved model generalization performance, and stacking models with different characteristics generated further improvement [25]. The ensemble approach is shown in Eqs. (13.1–13.3).

$$z_{kn} = \beta_k x_n, k = 1, K \tag{13.1}$$

$$\alpha_k^* = \underset{\alpha_k}{\operatorname{argmin}} \sum_{n=1}^{N}\left(y_n - \sum_{k=1}^{K}\alpha_k z_{kn}\right)^2, s.t. \forall \alpha_k \geq 0 \tag{13.2}$$

$$Y = \sum_{k=1}^{K}\alpha_k^*\beta_k X \tag{13.3}$$

First, individual models β_k, where $k \in [1, K]$ denotes individual model index, are trained separately on the training dataset repetitively with all the training data except for case n. Prediction of the in-training-set but out-of-model case z_{kn} is then generated. The process is repeated until all the models have covered all cases in the training set. Subsequently, the model weights α_k^* are optimized to minimize internal cross-validation error. A non-negative constraint is applied to prevent overfitting and increase the model interpretability. This step of optimization is done on the metadata, and the prediction results of each model for each case are

used to optimize the model weights. The individual models that perform well in the prediction task tend to get larger weightings. The K individual models $\boldsymbol{\beta}_k$ are combined and used for prediction of DVH PCS Y. The ensemble in this study consists of nine models, including stepwise, ridge, lasso, and elastic net with six different λ_2-to- λ_1 ratios. Figure 13.4 shows one example of the model weights from the individual models. This model is built using 50 prostate sequential boost cases. Y is the bladder DVH PCS1, and X consists of bladder anatomical features. It is apparent that regression coefficients differ from model to model, even though these are all variants of linear regression models. Note that model 1, stepwise regression, uses the least number of features, and model 2, ridge regression, evidently underfits.

13.3.2 Outlier Filtering

In previous studies, it has been pointed out that automatic outlier removal requires further investigation [26, 27]. To mitigate this, we incorporate a model-based automatic outlier removal routine in the ensemble model in order to further improve model robustness and address the volatile nature of clinical data. We utilize the cross-validation meta-data native to the proposed ensemble method to identify and remove impactful dosimetric and anatomical outliers.

13.3.2.1 Anatomical Outliers and Dosimetric Outliers

The first type of outliers is anatomical outliers. Anatomical outliers refer to cases with uncommon anatomical features relevant to DVH prediction, such as abnormal OAR sizes, unusual OAR volume distributions relative to PTV surface. Generally, anatomical outliers are more likely to deviate from the linear model, and when they do, the effect of these cases are generally larger than normal cases

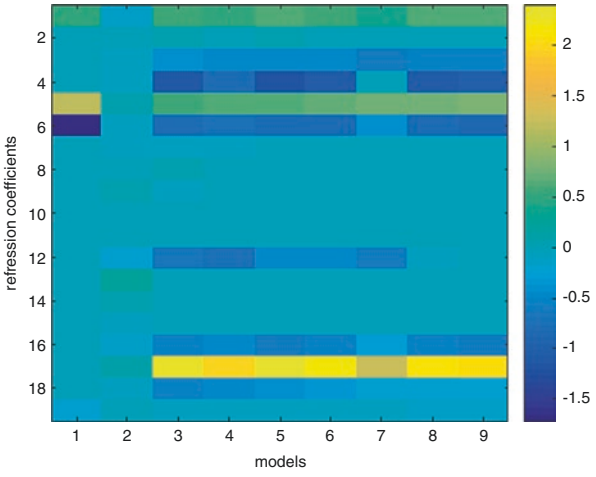


Fig. 13.4 Individual models trained on the same dataset. Vertical lines represent regression coefficients of individual models. Models 1–9 (from left to right) refer to stepwise regression, ridge regression, six elastic net regression models with various parameters, and lasso. Note that stepwise regression uses the least (four) features, and ridge regression uses all features but assigns small weights to the features

due to the quadratic data fidelity term of the regression model. Therefore, it is necessary to identify anatomical outlier cases that are detrimental to model building and remove those from the model before training. Other than anatomical outliers, there are cases that are detrimental to model building due to limited OAR sparing efforts and/or capabilities. These plans are categorized as dosimetric outliers. Dosimetric outliers include but are not limited to (1) treatment plans with inferior OAR sparing, (2) wrongly labeled data, such as 3D plans mixed in IMRT plans.

13.3.2.2 Prediction Performance Measure

Weighted root mean squared error (wRMSE) is defined to evaluate model prediction accuracy:

$$wRMSE = \sum_{i=1}^{N} w_i' \left(DVH_i - \widehat{DVH_i} \right)^2 .$$

wRMSE measures the overall deviation of predicted DVHs from ground truth DVHs, which are clinically planned. Weightings are introduced to emphasize higher dose regions of DVHs, which are generally considered to be of more clinical significance in OAR dose predictions. Here $w_i' = N w_i \Big/ \sum_{j=1}^{N} w_j$ denotes the normalized weighting factor for bin i of DVH curves.

13.3.2.3 Model-Based Case Filtering Method

To further improve the robustness of the ensemble model, cases with the highest s% median (of all individual models) internal cross-validation RMSE error are dropped from the training set. The percentage threshold s is selected to balance the trade-off between model robustness and accuracy. Empirically, we find that 10% is generally a good choice, even though the actual percentage of outlier cases is unknown and may differ from 10%. The complete workflow of the ensemble model with model-based case filtering is shown in Fig. 13.5. Similar to the model presented in Sect. 13.2, the end result of the training process is a set of regression coefficients, which can be used to predict OAR DVHs for future cases.

13.3.3 Retrospective Validation

In order to quantitatively evaluate the robustness of these regression methods in various challenging clinical environment, the KBP models were evaluated with limited training set size, training sets contaminated with anatomical outliers, and training sets contaminated with dosimetric outliers. In the outlier robustness tests, we purposefully mixed predefined outlier cases into the training set and validate the final model with normal cases. The rationale for adding outlier cases is to add controlled variation to the dataset and evaluate the robustness of the proposed model. Details regarding types of data used in the experiments are summarized in Table 13.1.

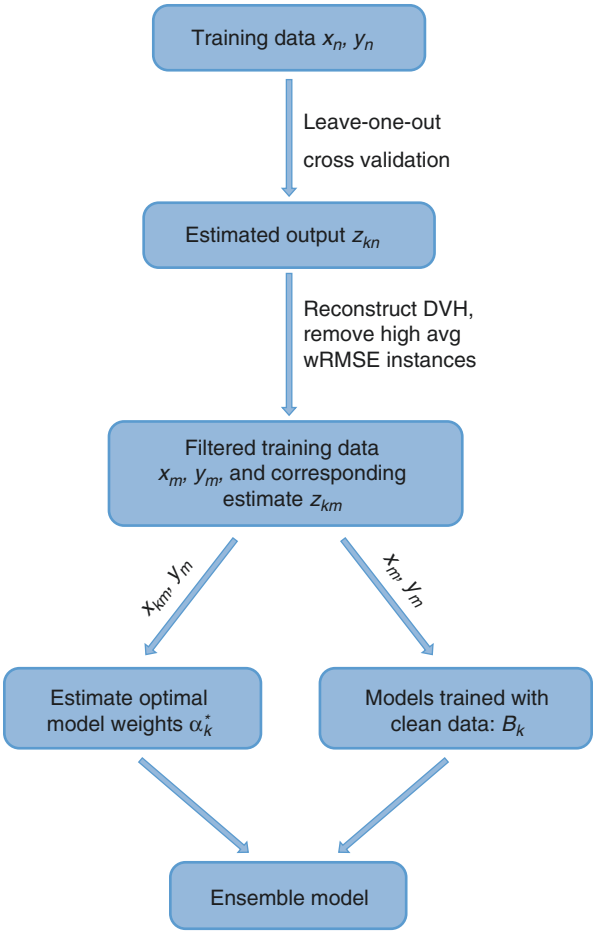


Fig. 13.5 The proposed ensemble learning workflow

Table 13.1 Summary of data used in the experiments

Dataset type	Training data	Validation data
Anatomical outliers	10 prostate cases treated with lymph nodes and 40 treated without lymph node	111 prostate cases treated without lymph node
Dosimetric outliers	40 prostate IMRT cases and 10 prostate conformal arc plans	110 prostate IMRT plans
Small training set	20 prostate IMRT cases	146 prostate IMRT cases

Figure 13.6 shows the prediction performance of individual models and the ensemble model, measured by DVH RMSE. The ensemble model consistently predicts better than, or similar to the best performing individual model in every challenging situation. It outperforms every individual model in at least one clinical scenario. With improved robustness, the proposed regression method potentially

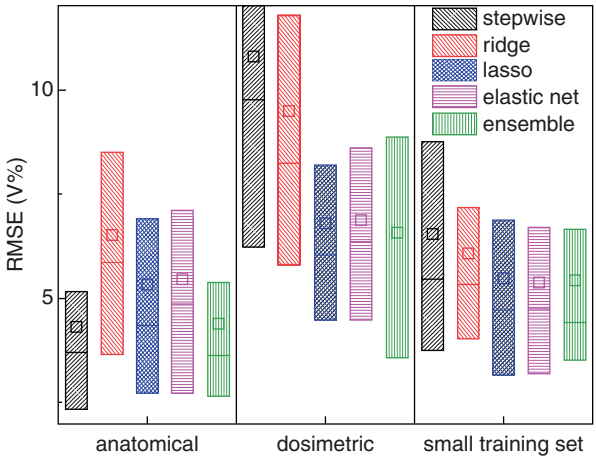


Fig. 13.6 Root mean-squared error (RMSE) for different models in different clinical scenarios. The three columns represent models trained with anatomical outliers, dosimetric outliers, and small training set, respectively (see Table 13.1 for dataset details). Squares denote mean values, and bars denote 25th, 50th, and 75th percentiles

enables end users to build site-specific, physician-specific, or even planner-specific models, without manually screening the training cases.

13.4 A KBP Model for Multiple-PTV Plans

The PCA-based DVH prediction model presented in previous sections only natively predicts DVHs in treatment plans with one PTV. The model can be modified to predict cases with multiple PTVs by using a simple feature summation method. However, the modification only works well under certain circumstances (e.g., similar prescribed dose ratio). In this section, we introduce a novel KBP model coupled with a generalized feature [20] to handle plans with multiple PTVs.

13.4.1 Generalized Distance to Target Histogram

We first define a multidimensional feature set, generalized distance-to-target histogram (gDTH), to represent the geometric variations of an OAR to multiple PTVs. In this section, we shall focus on the two-dimensional version of this concept but higher dimensional gDTHs can be developed similarly for cases with more than two PTVs. The elements of a gDTH matrix are defined as

$$G_{ij} = \frac{Number\ of\ voxels\ with\ d_1 < d(i), d_2 - d_1 < d(j)}{Number\ of\ voxels\ in\ the\ OAR},$$

where d_1 and d_2 denote distances from a voxel to primary and boost PTVs, respectively. G_{ij} is the fraction of the OAR volume with distances to the primary PTV surface smaller than $d(i)$, distances to boost PTV surface smaller than $d_1 + d(j)$. To generate a full gDTH for an OAR, we first calculate d_1 and d_2 for all voxels inside the OAR and then sort d_1 and $d_2 - d_1$ into discrete bins on a 2D map.

13.4.2 Modeling with a gDTH-Based Similarity Metric

To effectively use the gDTH feature, we incorporate a similarity metric that measures the geometrical similarities of OARs with respect to multiple PTVs. The first term is the Frobenius norm of the differences between the gDTHs of two cases. To account for prescription dose variations, we add a second term to represent the dose ratio similarity and define the similarity metric as:

$$\left\| gDTH_{target} - gDTH_{ref} \right\|_F^2 + \lambda \left(\frac{d_{target\ pri}}{d_{target\ bst}} - \frac{d_{ref\ pri}}{d_{ref\ bst}} \right)^2,$$

where $gDTH_{target}$ and $gDTH_{ref}$ denote the gDTH of the target plan and that of the plan being referenced from the database; λ is a balancing factor empirically tuned to match the mean values of the first term and the second term in the training dataset; and d denotes prescription dose.

Using this similarity metric, the k-nearest neighbors (kNN) search then selects a subset of training cases that resemble the validation case. T-distributed statistical neighboring embedding (t-SNE) [28] is used to visualize this high-dimensional feature space and to justify similarity metric measurements on the feature space distribution. T-SNE converts high-dimensional Euclidean distances to conditional probabilities and maps high dimension data to low dimension while preserving local structures of the datasets. A visualization of the proposed feature map of a dataset is shown in Fig. 13.7. Figure 13.7a shows a two-dimensional t-SNE map of the left parotid gDTHs of the 120-case training dataset in this study (the red and blue dots). Figure 13.7b is a validation case randomly picked to demonstrate the effectiveness of the proposed feature at differentiating cases with different OAR-PTV shape distributions. The blue dots on the map (Fig. 13.7a) are the cases selected by the similarity metrics to build the model to predict the parotid DVH of Fig. 13.7b (the validation case), while the red dots are the cases excluded from the modeling. Figure 13.7c-f further show the PTVs and left parotid anatomies of the selected (Fig. 13.7f) and unselected (Fig. 13.7c-e) cases, and their respective locations on the 2D t-SNE map are indicated by the arrows.

13.4.3 Data Augmentation

Head and neck treatment plans have high inter-patient spatial variability, considering that the boost PTVs (i.e., GTVs) from various sub-sites are in different regions and that the OARs also vary significantly from patient to patient. Therefore, to successfully train a reliable KBP model for head and neck treatments, many treatment plans are required. However, the treatment plans available for training purposes is limited. To make efficient use of the training cases and effectively increase the training dataset, we here present two data augmentation methods for our modeling process. Both methods utilize single-PTV cases and the primary plans of multiple-PTV cases as the means to synthesize gDTHs.

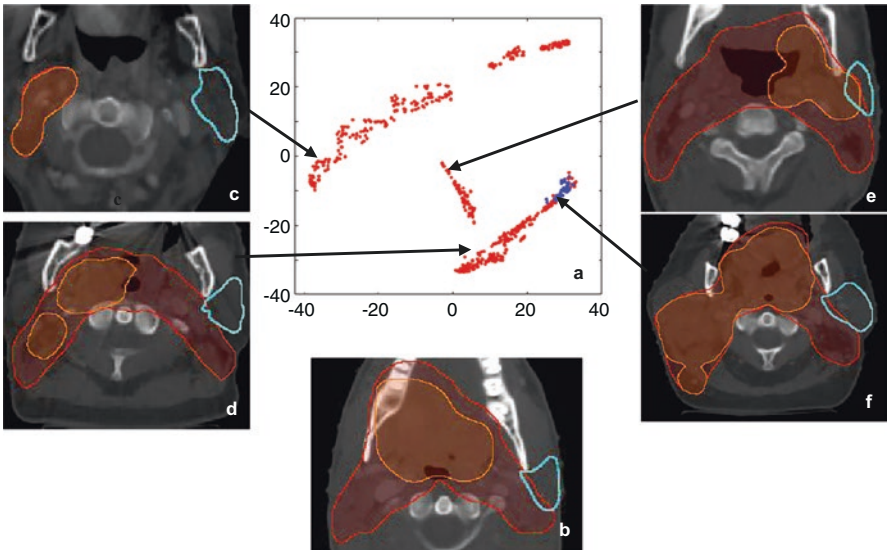


Fig. 13.7 A t-SNE visualization example: (**a**) A two-dimensional t-SNE map of the left parotid gDTHs; (**b**) The randomly selected validation case; (**c**-**f**) four example cases located in different regions of the feature map. In (**a**), blue dots mark the cases selected by the similarity metric for modeling the validation case in 3b, and red dots denote the rest of the dataset. (**c**), (**d**), (**e**) show three unselected cases, and the arrows indicate their locations on the t-SNE map, while (**f**) is one of the selected cases even though its PTVs (especially boost PTV) are significantly different to 3b in size and location. Both the x- and the y-axis of the t-SNE map are dimensionless and are of arbitrary units

The first data augmentation method originates from an intrinsic property of gDTH. The rightmost column of gDTH is equivalent to the DTH associated with primary PTV. In some cases, certain OARs are only affected by primary PTVs because the distance to boost PTVs is more than 5 cm larger than the distance to primary PTV. Such shape distribution is most common for brainstems and parotids. To simulate such cases, we scale OAR DVHs to various common clinical dose ratios—e.g., 44 Gy /70 Gy, 50 Gy/60 Gy, and generate zero filled gDTHs with only the rightmost columns remain unchanged from the original cases. We replicate the whole dataset in this fashion. By generating these additional cases, we effectively increase the number of training cases in which OAR DVHs are only affected by primary PTVs.

The treatment plans without boost PTVs can be utilized to simulate cases in which two PTVs have the same volume. For some cases, boost PTV surfaces overlap with primary PTV surfaces in the regions that are close to the OAR. We make the approximation that primary PTV plans can be scaled up to boost PTV dose level and can be treated as a plan with primary and boost PTVs of the same shape. For this type of augmented cases, gDTH can be generated by treating original primary PTV as both primary PTV and boost PTV.

13.4.4 Training and Validation

With IRB approval, 268 HN cases were retrieved for model training and validation. From these cases, 120 cases were randomly selected for model tuning (feature selection, augmentation method evaluation) to avoid positively biasing the results. The model performance was tested with the remaining 148 cases. To evaluate the effectiveness of the proposed data augmentation methods, we first map the training set to a two-dimensional PCA space. The first two principal component scores of the training set gDTHs are set as the X and Y axes, respectively (Fig. 13.8). As shown in the figure, the data augmentation procedure populates two opposite sides of the gDTH distribution, where data points are sparse. Therefore, when predicting a validation case's DVH with gDTHs near the edge of the map, augmented cases will be selected to build the model and help improve prediction accuracy and robustness.

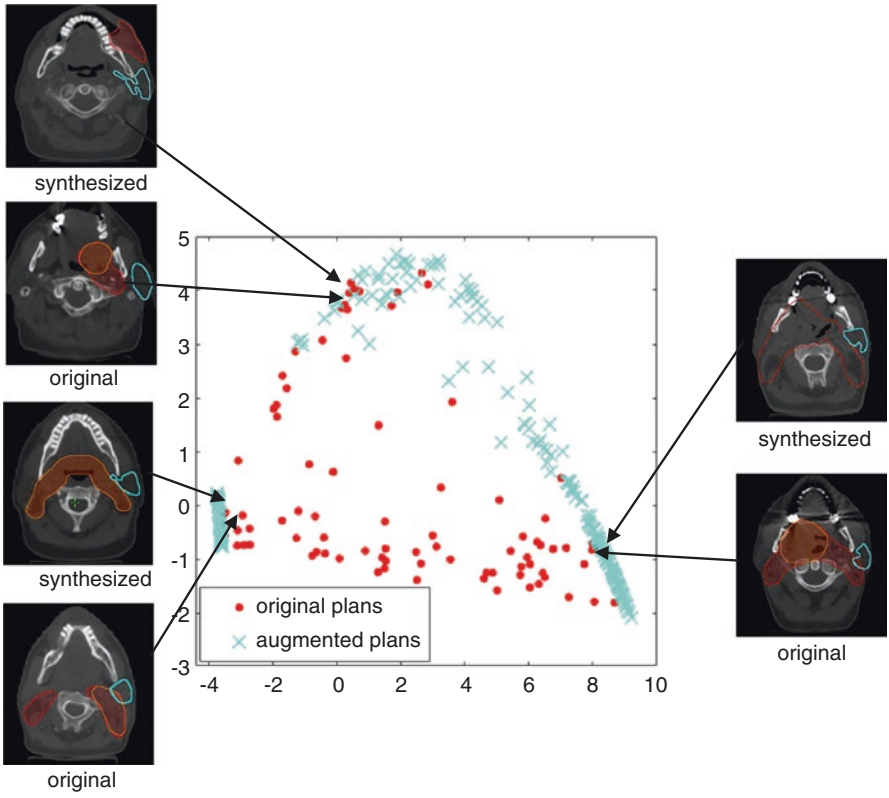


Fig. 13.8 The distribution of the first two principal component scores of gDTHs in the training dataset. Blue crosses represent augmented cases, and red dots represent the original training data. Also shown in the figure are three example pairs of structure sets that demonstrate the resemblances of the original and the synthesized cases. Left parotids, primary, and boost PTV structures are marked with cyan contours, red segments, and orange segments, respectively

Table 13.2 Model prediction accuracy comparison between the previous modeling process [19] and the proposed modeling process with and without data augmentation

	DVH RMSE (Vol %)		
	Previous model	Proposed model	Proposed model with data augmentation
Parotid	7.99 (0.36)	6.92 (0.30)	6.83 (0.28)
Brainstem	5.11 (0.39)	3.73 (0.23)	3.77 (0.26)
Cord	5.53 (0.27)	5.19 (0.25)[a]	4.92 (0.25)
Mandible	6.31 (0.23)	5.70 (0.21)	5.59 (0.22)
Larynx	9.32 (0.74)	8.46 (0.80)[a]	7.19 (0.35)
Oral cavity	8.23 (0.43)	7.58 (0.40)	7.33 (0.41)
Pharynx	7.63 (0.28)	7.04 (0.32)	6.65 (0.27)

[a]The improvement over the previous method is not statistically significant (paired-sample t-test, $p > 0.05$)

To quantitatively measure the improvements of the proposed modeling workflow over the previous process, we evaluate DVH prediction accuracy measured by the root-mean-squared error (RMSE). The model previously tuned with 120 HN cases is evaluated using a separate validation dataset consisting of 148 cases (all with 2 PTVs). Compared with the current state-of-the-art KBP model [19], this model results in significantly reduced prediction RMSE for brainstem ($p < 0.001$), mandible ($p = 0.004$), pharynx ($p = 0.034$), oral cavity ($p = 0.022$), and parotids ($p < 0.001$), but the improvements are not significant for cord ($p = 0.051$) and larynx ($p = 0.099$). When augmented cases are included in the training dataset, statistically significant improvements are observed for predicted DVHs of all OARs, including brainstem ($p < 0.001$), cord ($p < 0.001$), larynx ($p = 0.004$), mandible ($p < 0.001$), pharynx ($p = 0.001$), oral cavity ($p = 0.011$), and parotid ($p < 0.001$), as shown in Table 13.2. In particular, the DVH prediction accuracies are moderately improved when data augmentation is implemented, compared to the model without data augmentation. For some OARs with high DVH variances (e.g., larynx), the improvement is significant.

The multiple-PTV model workflow generates accurate and robust DVH. The KBP plans guided by the proposed model demonstrates that the improvement in the DVH prediction model can translate into better plan quality in knowledge-based planning. KBP with the proposed modeling method can potentially help planners to achieve higher and more consistent plan quality, compared with the current clinical planning process.

13.5 Head and Neck Trade-off KBP Model

In treatment sites such as head-and-neck (HN), one set of DVH predictions provided by the standard KBP models may not be sufficient for clinical planning due to the complexity of OAR trade-off relations and considerations. Physicians usually prescribe OAR dose constraints based on estimates of the best achievable DVHs with desired trade-offs, and planners have to interact with the treatment planning

system (TPS) and the physicians iteratively to achieve patient-specific optimal dose-sparing. This iterative process forgoes some of the time-saving advantages accomplished by the KBP. In this section, we present a preplanning trade-off estimation method [6] that seeks to support trade-off decision-making by modeling the clinically viable trade-off experience embedded in prior clinical plans.

13.5.1 Plan Trade-off Modeling

The workflow of the proposed method, as shown in Fig. 13.9, starts with building a localized KBP model. First, given a case to predict, a case reference set (CRS) consisted of N reference cases is built using the treatment plans from a reference database. The KBP model presented here is based on the KBP model discussed in Sect. 13.4. The generalized distance-to-target (gDTH) feature is modified to select similar plans in terms of all OARs. After the KBP model is built, it is applied to all the cases in the CRS to extract trade-off-related variations. The difference between the predicted and the actual DVH PCS (from clinical plans) for all OARs is estimated and formulated as $E \in R^{N \times MP}$, which is essentially the fitting residuals of N training cases in the CRS (N = 35 in this study), each with M OARs and P DVH PCS scores per OAR. We observe in the clinic that significant portions of the fitting errors are due to the trade-off relations. As a result, there is valuable information in the fitting residuals. The purpose of subtracting the predicted PCS is to remove the variations

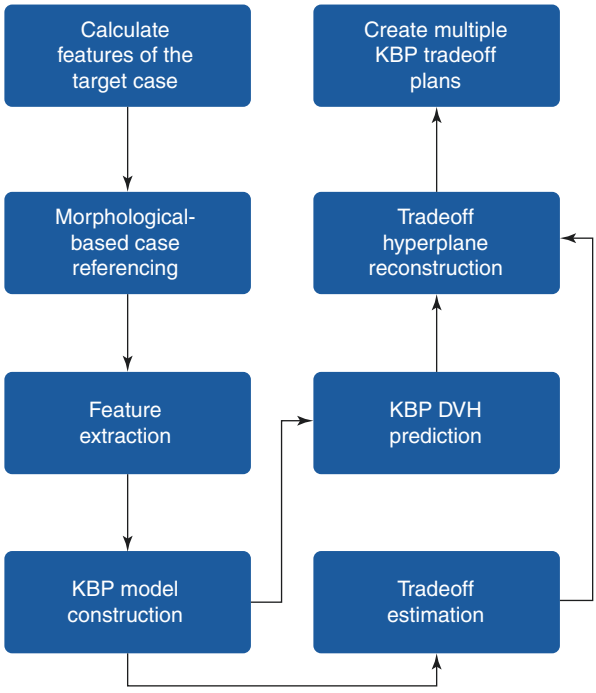


Fig. 13.9 The workflow of a knowledge-based trade-off model

linked to anatomical differences and extract only the discrepancy between the KBP model and the clinical plans that are likely caused by trade-off decisions. The matrix *E* of OAR DVH fitting residuals is subsequently processed by principal component analysis (PCA) to further reduce the dimensionality with the first three PCS taken as the principal trade-off directions. The trade-off directions effectively reveal the most prominent DVH variation patterns in the CRS after adjusting for anatomy variations. These patterns are representative of the commonly occurring trade-offs in the historical reference database. Since the CRS used as training cases in model building is selected based on morphological similarity to the current case, the extracted trade-off directions are thus specific to the current case. These directions form a trade-off hyperplane that provides effective guidance for a well-constrained and clinically viable trade-off subspace.

13.5.2 Trade-off Simulation and Validation

Figure 13.10 shows the RMSE between the DVHs predicted by the trade-off hyperplane and the corresponding DVHs realized by the auto-generated trade-off plans. These RMSE values represent the fidelity of the auto-generated trade-off plans compared with hyperplane model predictions and are evaluated against the KBP model baselines as well as the clinical-DVH-fitted results. RMSEs of the max-dose constrained OARs (cord, brainstem, mandible) have in general higher values compared to dose–volume constrained OARs (e.g., parotid). In clinical planning, when OARs are max-dose constrained, planners tend to place very low priorities on sparing these OARs at dose–volume points other than the D_{max}. This results in large DVH variations not attributed to anatomy differences. Therefore, these DVH curves are not predicted as accurately as DVHs of other OARs, such as parotid and oral cavity. Due to such variations, RMSE values, even for the KBP plans, are not expected to be zero. However, as it has been shown in the literature that KBP model predicted DVHs are indeed achievable [29], the non-zero RMSE of the KBP model predictions can establish the baseline for the hyperplane model

The RMSE of all 12 hyperplane-guided trade-off plans are not significantly different from the baseline RMSE ($p > 0.05$; paired t-test; $n = 30$ validation samples). These results suggest that all trade-off plans are as achievable as the KBP plans. Table 13.3 shows the averaged RMSE values for all OARs in all plans at various trade-off locations. The RMSE values for trade-off locations further away ($\pm 2\ \sigma$) appear to have higher RMSE values than other trade-off locations ($\pm\ \sigma$).

The trade-off hyperplane with three directions accounts for 68.9% ± 0.5% of the variances in the training plans, and 57.5% ± 3.0% in the validation plans. All 14 replaned cases match closely to the predicted hyperplane location (DVH RMSE<10%). The feature extraction, regression, and DVH prediction of the proposed model work similarly to the conventional model-based KBP. The endpoint of the workflow is an ensemble of best achievable plans with various trade-off preferences. Additionally, the workflow can provide clinicians real-time estimations of planning trade-offs,

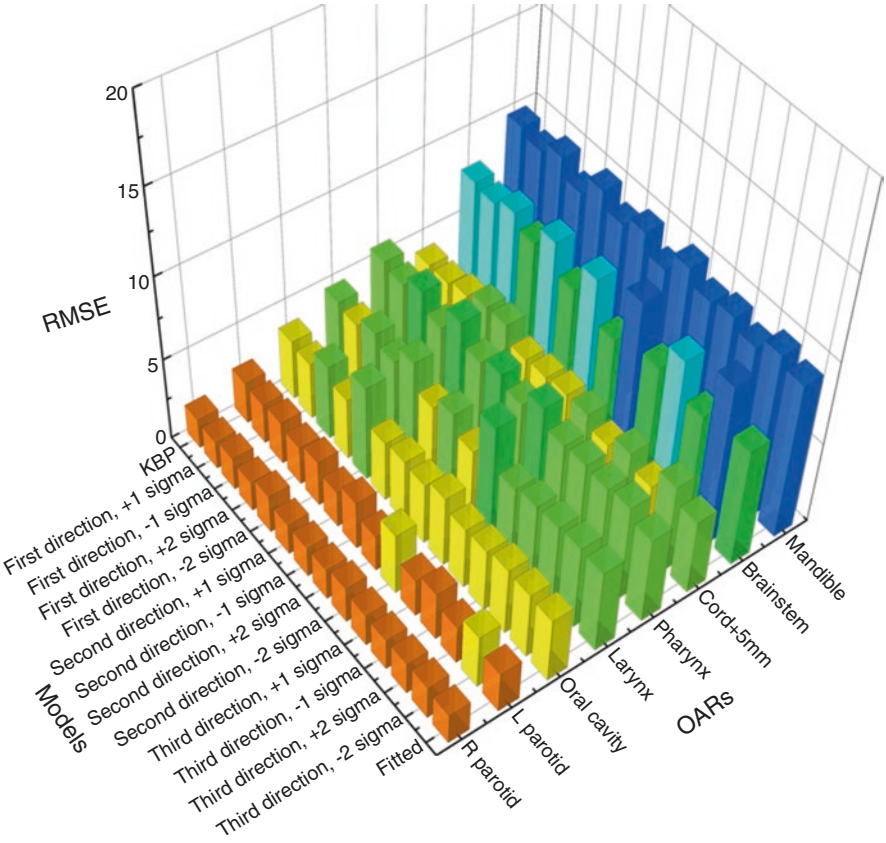


Fig. 13.10 RMSE values for model predicted DVHs measured against corresponding TPS plan DVHs. Small RMSE values indicate that hyperplane predicted DVHs closely resemble the realized DVHs of auto-generated plans and hence provide evidence that the hyperplane DVH predictions are highly achievable. Different colors denote different ranges of RMSE values

Table 13.3 Average RMSE value for different trade-off locations

Types of DVH guidance	Average RMSE (%)	Standard deviation (%)
KBP (no trade-off)	4.96	2.48
First trade-off direction, ± 1 σ	5.07	2.56
First trade-off direction, ± 2 σ	5.39	2.52
Second direction, ± 1 σ	5.05	2.50
Second direction, ± 2 σ	5.35	2.62
Third direction, ± 1 σ	5.08	2.46
Third direction, ± 2 σ	5.30	2.51
Clinical DVH fitted	5.02	2.33

thereby providing systematic guidance on the best achievable dosimetric parameters for informed decision-making.

13.6 A Complete Workflow for KBP Planning of Whole Breast Radiation Therapy

Whole breast radiation therapy (WBRT) is routinely used in post-operative settings to reduce the risk of locoregional recurrence. Describe tangential beam configuration. For WBRT, forward planning techniques such as field-in-field and fluence editing are the standard practice. In the fluence-editing method, the planner starts with an initial fluence and modifies the fluence iteratively. This technique requires hours of fine-tuning fluence manually, and the plan quality is highly dependent on the planner's experience. In this section, we present an automatic planning method for WBRT [13]. Different from the KBP models presented in earlier sections, the WBRT auto-planning method generates fluence maps directly without using the inverse optimization engine embedded in the TPS.

13.6.1 Digitally Reconstructed Radiograph (DRR)-Based Energy Selection

The first decision to make when planning a WBRT case is to select the appropriate beam energy. At our institution, two types of beams are typically used: (1) single energy (SE): 6X beams and (2) mixed energy (ME): 6X and 15X beams. The planner determines the beam energy configurations by examining the beam path length. To build an automated energy selection tool, prior clinical plans are used to build a binary decision model (a choice of single or mixed energy) and classify the query case for the energy selection. Digitally reconstructed radiographs (DRRs) are generated in each beam direction, and the gray-level histogram within the irradiation volume is calculated. Principal component analysis (PCA) is then performed on the gray-level histogram of each case in the training set (two beams combined) to reduce the data dimension. The first two principal component scores are subsequently used as features for the classification model. The energy decision boundary is then determined in the 2D feature space. It is worth noting that the idea behind this model is that the radiological path length in the beam direction can be effectively represented by the DRR gray values and captured by the PCA.

13.6.2 Anatomy-Driven Fluence Estimation

The second step of the workflow is to generate a fluence map to achieve the optimal dose distribution within the 3D target volume. The fluence map is generated by predicting the fluence intensity of each pixel on the fluence map. Ideally, the optimal dose distribution should cover the entire breast target with prescription dose

while minimizing the hot spot volume (105% of prescription dose). A machine learning algorithm has been developed to learn the correlation between anatomical features and the optimal fluence intensities. We utilize the random forest (RF) model to summarize the relationship between input features (shape-based features, including gray-level intensity, penetration depth in breast target, penetration depth in lung, etc.) and output variables (pixelwise fluence intensity). RF is a highly nonlinear model which initializes decision trees using randomly sampled data from a training dataset and generates a prediction by averaging the output from all trees. The RF model was trained using all 20 training plans with 150 trees. For query cases, the RF model predicted fluence intensity at the pixel level and the entire fluence map served as the fluence estimation for the corresponding beam. For ME cases, the entire predicted fluence map was divided into a low-energy (6X) and high-energy (15X) components. Low-energy and high-energy beams from the same side share same beam parameters such as gantry angle, collimator angle, and jaw sizes. The ratio of low-energy fluence intensity and high-energy fluence intensity for each pixel on the fluence map depends on the beamlet penetration depth, and this relationship was learned from the 10 training ME plans.

13.6.3 Patient-Specific Fluence Fine-Tuning

The fluence map generated from the RF model inherits the plan quality from the training cases. However, the physician may have a patient-specific requirement for the target coverage or a constraint for a high-dose volume or hot spot. The third step offers physician an opportunity to interactively fine-tune the 3D dose distribution. This is achieved by specifying the dose to be delivered to anchor points while balancing dose contribution from both beams. Dose anchor points are identified in two steps. First, they are identified on the iso-plane in the irradiated volume and later adjusted during the centrality correction step. Then, the centrality correction step actively balances the beamlet penetration depth inside the breast tissue from either side for each dose anchor point. Geometric and dosimetric parameters (penetration depth, dose at anchor point, etc.) of these dose anchor points are summarized from training plans to serve as baseline values, and these parameters can be further adjusted to provide specific coverage or dose reduction for any query patient.

13.6.4 Planning Validation

13.6.4.1 Data Selection

The previous sections already referred the number of cases for training, etc. A total of 40 institutional review board–approved WBRT plans from Duke University Medical Center were retrospectively studied. All plans were treated with 50 Gy in 25 fractions. Twenty plans, 10 with single energy (SE, 6MV) and 10 with mixed energy (ME, 6/15MV), were randomly selected to establish the optimization parameters of the proposed methodology. The remaining 20 plans (10 SE and 10 ME

plans) were reserved for validation. Among the 20 training cases, 12 are left breast cases. For validation cases, 9 out of 20 are left breast cases. SE plans use two 6MV beams, namely the medial beam and lateral beam, set by the attending physician to include the whole breast and skin flash. For ME plans, two high-energy beams (15MV) utilize the same beam setup and beam apertures as two low-energy beams (6MV) in the corresponding beam direction. Clinical plans of all 40 cases were manually generated in the Eclipse™ TPS by planners iteratively painting the fluence and calculating the dose distribution.

13.6.4.2 Model Training and Validation

The PCA analysis result is shown in Fig. 13.11. The DRR intensity histogram for each patient is shown in Fig. 13.11a. In Fig. 13.11b, red dots represent single energy cases and green dots represent mixed energy cases. Solid squares represent training cases while circles represent validation cases. PC1 = 0 served as a good classifier with an accuracy of 19/20 for the validation cohort, meaning the model suggested the same energy combination as the clinical plans.

13.6.4.3 Plan Quality Comparison

Dose distribution was qualitatively compared between clinical plans generated manually and automatically generated plans. Figure 13.12 shows the isodose distribution comparison for one large breast case (left three columns) and one small breast case (right three columns). Overall dose homogeneity was comparable between the clinical and auto-plans. The high-dose volume (105% Rx dose volume) was similar in location as well as volume between two plan groups.

Boxplots of dose–volume metrics are shown in Fig. 13.13. The median and interquartile range for each endpoint are comparable between two plan groups.

13.6.4.4 Plan Efficiency

The average optimization time for auto-plans is <20 s. Even as a stand-alone platform, the entire process including data transferring from and to the treatment

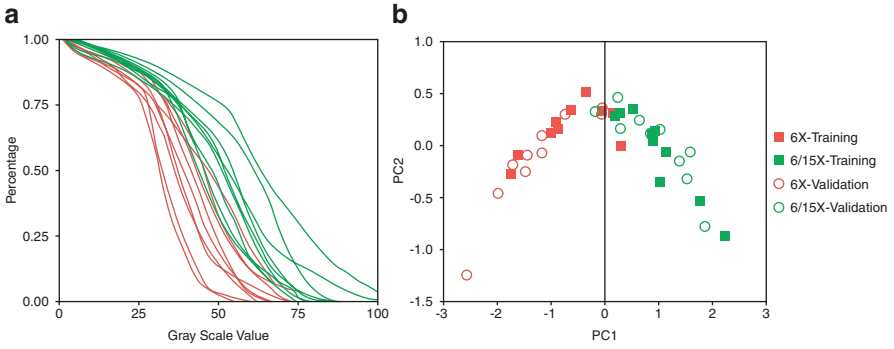


Fig. 13.11 (**a**) DRR intensity histogram for single energy cases (red) and mixed energy cases (green); (**b**) PC1 and PC2 score of single energy cases (red) and mixed energy cases (green) for training dataset (solid square) and validation dataset (circle), PC1 = 0 is shown as black line

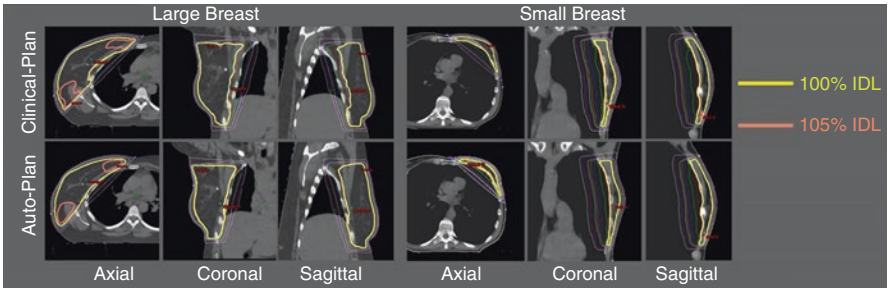


Fig. 13.12 Isodose comparison between ECOMP clinical plan (top row) and ECOMP auto-plan (bottom row) for one large breast patient (left three columns) and one small breast patient (right three columns)

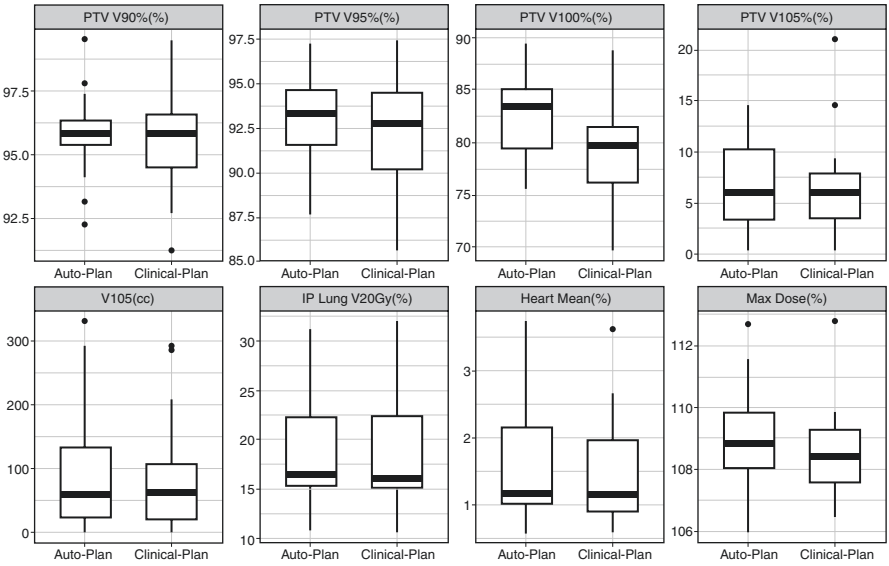


Fig. 13.13 Boxplot comparison of dose metrics between auto-plan (left) and clinical plan (right)

planning system can be accomplished within 5 min. This is substantially faster than the manual process which ranges from 30 min to 4 h in our clinic.

13.7 Beam Bouquet Knowledge Model for Lung IMRT Planning

Another important aspect of treatment planning is beam angle selection. For certain planning sites, such as lung, good beam angle selection decisions are critical in creating high-quality treatment plans. In clinical practice, beam angles are often selected based on the planner's experience and adjusted by trial-and-error to find an

optimal set of beam angles or a beam bouquet. In this section, we present a method of establishing a small set of standardized beam bouquets for lung IMRT planning [15]. The bouquets are determined by learning the patterns from the multi-dimensional beam configuration features of prior clinical plans using a cluster analysis method.

13.7.1 Dissimilarity Metric between Two Beam Bouquets

First, we define a dissimilarity measure between two beam bouquets. The dissimilarity measure is computed as the sum of angle separations between each pair of corresponding beams in the two bouquets. It takes into account the permutation of beams within each bouquet when comparing two beams. Specifically, a distance is first defined between two angles a and b,

$$\delta(a,b) = \min_{k\in Z} |a - b + 360k|,$$

where k can take any value in the integer set Z and the 360 k term accounts for the 360 degree modulo in the angle space. Then, the dissimilarity measure between two bouquets with the same number of beams $x_1 = (x_{11}, x_{21}, \cdots, x_{n1})$ and $x_2 = (x_{12}, x_{22}, \cdots, x_{n2})$ is defined as:

$$d^1(x_1, x_2) = \min_{\sigma\in\pi} \sum_{l=1}^{n} \delta\left(x^1_{\sigma(l)}, x^2_l\right),$$

where σ is any permutation π of the beam orders. In our dataset, the number of beams used in a plan ranges from 6 to 11, thus it is necessary to classify beam angle settings with different number of beams, by defining the dissimilarity measure between bouquets with different number of beams. If two bouquets x_1 and x_2 have different numbers of beams n_1 and n_2 and assume $n_1 > n_2$ without losing generality, we define the dissimilarity as the sum of two terms:

$$d(x_1, x_2) = \min_{x_1'\subseteq x_1} d^1(x_1', x_2) + \min_{x_2'\subseteq x_2} d^1(x_1 \setminus x_1', x_2'),$$

where x_1' is a subset of x_1 which has the same number of beams as x_2 and x_2' is a subset of x_2 with beam number $n_1 - n_2$. The first term compares n_2 beams in both x_1 and x_2 to calculate the distance, while the second term compares the remaining $n_1 - n_2$ beams in x_1 with x_2. This step ensures that every beam in the plan is taken into account when calculating the dissimilarity between two bouquets with different number of beams.

13.7.2 Establishing the Standardized Beam Bouquets

After the dissimilarity (or distance) is calculated between each pair of the beam bouquets, a k-medoids method is used to sort the beam angle configurations into clusters. A medoid is defined as the object in a cluster, with the average distance

to all the other objects in the same cluster (within-cluster distance) being the minimal. Thus, the medoid of a cluster is the most representative beam angle configuration of all cases within the cluster. The set of all medoids characterizes the major types of beam angle settings frequently used in clinical lung IMRT plans, and they are designated as the standardized beam bouquets. The medoid case that corresponds to a standardized beam bouquet is designated as the reference case of this bouquet.

The average silhouette width $\bar{s}$ is the average of the silhouette index s(i) over all the data points in the dataset. The silhouette index measures how close each point in one cluster is to the data points in the neighboring clusters. For a data point i in cluster A, let a(i) be the average distance of i to all other points in cluster A, d(i, C) be the average distance of i to all the data points in another cluster C. The silhouette index is defined as:

$$s(i) = 1 - \frac{a(i)}{\min\limits_{C \neq A} d(i,C)}$$

The silhouette index s(i) has a numerical value from +1 to −1. A large positive value indicates the point in one cluster is far from neighboring clusters, while negative value indicates the point may be assigned to a wrong cluster. A silhouette plot can be used to visualize how well separated the resulting clusters are. It plots the silhouette index s(i) for each data point as a horizontal bar. A wider silhouette plot indicates larger s(i) values. The beam settings in the training dataset were classified into 3–8 clusters by the k-medoids algorithm. The average silhouette width $\bar{s}$ for each classification result is plotted against the number of clusters in Fig. 13.14a. As shown in this figure, the classification with six clusters

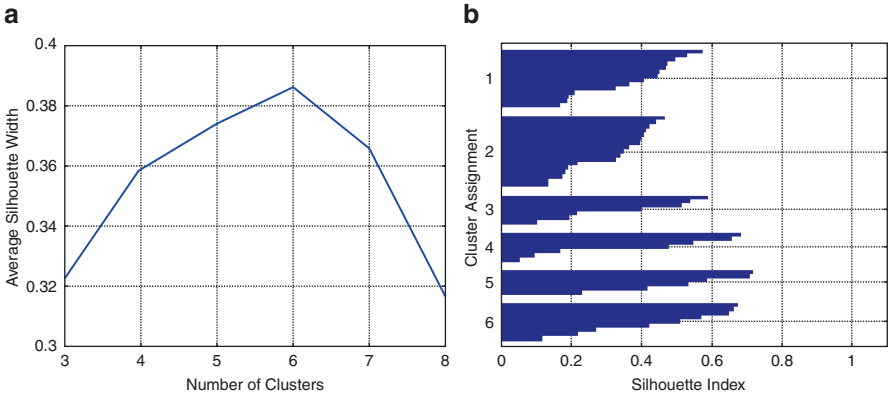


Fig. 13.14 (**a**) The average silhouette width for the classifications with each given number of clusters. (**b**) The silhouette plot for the classifications with six clusters. Each horizontal bar indicates the silhouette index for an object, which is grouped according to its cluster assignment. The width of the bar represents the silhouette index value. The vertical axis numbers indicate the indices of the clusters

has the largest average silhouette width value of 0.39, which suggests that six bouquets best represent the beam configuration patterns in the dataset. The silhouette plot with six clusters is shown in Fig. 13.14b. The all-positive silhouette index values indicate that the beam configurations align well within their assigned clusters.

The beam bouquets corresponding to the medoids of the six clusters are shown in Fig. 13.15. The number of beams in these bouquets ranges from 7 to 9, which reflects the number of beams used in the reference clinical plans. The representative axial CT image slices of these reference plans are also shown under each medoid in the figure. As shown, these beam configurations reflect the gross anatomical and tumor location characteristics.

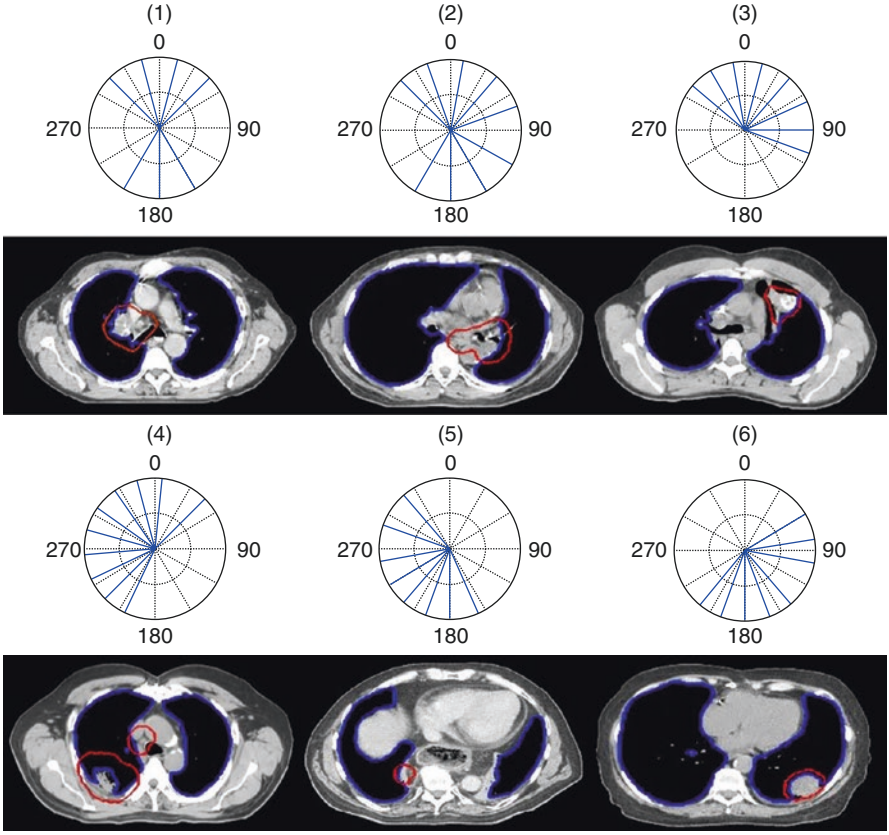


Fig. 13.15 The six beam bouquets are shown in polar coordinates using IEC beam angle convention at the first and third rows. The solid radial lines indicate the beam directions. The number inside the parenthesis on top of each bouquet labels the ID of the bouquet. The representative axial CT image slices of the reference cases corresponding to the medoids of the six clusters are shown under each medoid at the second and fourth rows. The PTV is denoted by the red contours and the lung by the blue contours

13.7.3 Validation with Clinical Cases

Sixty lung IMRT plans with prescription doses ranging from 45 Gy to 70 Gy were retrospectively studied under an IRB-approved research protocol. The plans have six to eleven co-planar beam angles, with an average beam number of eight. The dataset has a wide range of tumor size (from 12 to 4432 cm^3, mean 502 cm^3) and locations. The tumor locations in the dataset are distributed as follows: 26 cases in the right lung, 23 in the left lungs, 8 in the mediastinum, and 3 in the chest walls.

Twenty additional randomly selected lung cancer cases were re-planned to assess the validity of using the standardized beam bouquets. For each case, an experienced planner manually selected a standardized beam bouquet based on his/her judgment of the similarity between the tumor location and patient anatomical features of the case and those of the reference cases. The planner had no knowledge of the beam configurations used in the original clinical plans. After the beam bouquet was selected, inverse optimization was performed using the same dose objectives as in the clinical plans. The mean and standard deviations (SD) of the dosimetric parameters in plans using six beam bouquets and those in the clinical plans. They are compared by paired t-tests.

Table 13.4 lists the mean and standard deviations of the dosimetric parameters in the bouquets based and the clinical plans, as well as the paired t-test values. The lung V10Gy, the esophagus mean dose, cord D2%, and PTV dose homogeneity defined as D2%–D99% are statistically better in bouquet-based plans (p-value<0.05), but the improvements (<5%) were small and may not be clinically significant. Other dosimetric parameters are not statistically different.

13.8 Summary

In the modern era of radiation oncology, KBP is increasingly implemented to facilitate high quality and efficient treatment planning. KBP refers to the concept of modeling clinical planning knowledge embedded in previously treated cases and utilizing the knowledge to help generate new plans. KBP exists in various forms, ranging from forward planning to inverse planning. In this chapter, we have

Table 13.4 Dosimetric parameters in plans using six beam bouquets and those in clinical plans

OAR/PTV	Parameter	Bouquets plans	Clinical plans	p-value
Lung	V_{10Gy} (% OAR volume)	29.1 ± 11.7	32 ± 12.6	0.01
	V_{20Gy} (% OAR volume)	18.3 ± 8.1	18.9 ± 8.7	0.44
	Mean dose (% Dx)	18.8 ± 7.0	19.2 ± 7.0	0.28
Esophagus	Mean dose (% Dx)	32.0 ± 16.3	34.4 ± 17.9	0.01
Heart	V_{60Gy}(% OAR volume)	0.6 ± 1.1	1.2 ± 2.7	0.39
	Mean Dose (% Dx)	19.2 ± 16.5	19.4 ± 16.6	0.74
Spinal cord	$D_{2\%}$ (%Dx)	47.7 ± 18.8	52.0 ± 20.3	0.01
PTV	$D_{2\%}$–$D_{99\%}$ (% Dx)	17.1 ± 15.4	20.7 ± 12.2	0.03

discussed DVH-prediction models for inverse IMRT/VMAT planning, a breast WBRT auto-planning technique, and a lung beam angle selection model.

The key elements of successfully building a robust KBP model are data quality and data sufficiency. Considering the time-sensitive nature of the clinical environment and the large variations of patient population, there can be some plans in the clinical database that do not follow the general planning trends. These plans are considered outliers when training a KBP model. The effect of outliers of KBP modeling has been extensively studied. For instance, Delaney et al. have shown that dosimetric outliers have marginal effects on resulting plan quality. To directly address the effect of outliers, Sheng developed an outlier identification method [17]. Nevertheless, it is critical to examine a KBP model and properly validate it before clinical use to ensure the model indeed represents best achievable plan qualities from previous experience. Another important aspect of KBP modeling is the number of plans used to train a model. Boutilier et al. have shown that different models require different amount of cases [30]. Therefore, the number of cases for a robust model still requires careful consideration when designing a KBP model for clinical use.

Future research in KBP will likely focus on more sophisticated modeling methods and more complex planning scenarios. Both directions will be enabled by the development of larger database of high-quality clinical plans through integration efforts across consortium of institutions as well as accumulation of planning cases within individual institutions. Recent publications have shown promising results using complex nonlinear models such as convolutional neural networks to successfully predict voxel-level dose in some cancer sites. Work has also begun to handle more complex cancer targets, more complex trade-off decisions, as well as more complex treatment techniques. Beyond more complex and powerful models, the sophistication of modeling methods will also mean more advanced algorithms for learning, evolving, and integrating models. So far, data-driven KBP has focused on building models in a batch mode, that is, learning from static datasets. As these models mature and are deployed in clinical use, another important research question will address how these models can be improved as new clinical cases are accumulated and new treatment techniques are developed.

Acknowledgments This work is partially supported by an NIH grant (#R01CA201212) and a master research grant from Varian Medical Systems.

References

1. Kalet IJ, Paluszynski W. Knowledge-based computer systems for radiotherapy planning. Am J Clin Oncol. 1990;13(4):344–51.
2. Shwe MA, Tu SW, Fagan LM. Validating the knowledge base of a therapy planning system. Methods Inf Med. 1989;28(1):36–50.
3. Zhang X, et al. A methodology for automatic intensity-modulated radiation treatment planning for lung cancer. Phys Med Biol. 2011;56(13):3873.
4. Voet PW, et al. Toward fully automated multicriterial plan generation: a prospective clinical study. Int J Radiat Oncol Biol Phys. 2013;85(3):866–72.

5. Hazell I, et al. Automatic planning of head and neck treatment plans. J Appl Clin Med Phys. 2016;17(1):272–82.
6. Zhang J, et al. Knowledge-based tradeoff hyperplanes for head and neck treatment planning. Int J Radiat Oncol Biol Phys. 2020;106(5):1095–103.
7. McIntosh C, Purdie TG. Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. Phys Med Biol. 2016;62(2):415–31.
8. Shiraishi S, Moore KL. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy. Med Phys. 2016;43(1):378–87.
9. Zhang J, et al. Voxel-level radiotherapy dose prediction using densely connected network with dilated convolutions. In: Artificial intelligence in radiation therapy. Cham: Springer International Publishing; 2019.
10. Nguyen D, et al. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. Phys Med Biol. 2019;64(6):065020.
11. Lee MS, et al. Deep-dose: a voxel dose estimation method using deep convolutional neural network for personalized internal dosimetry. Sci Rep. 2019;9(1):10308.
12. Lee H, et al. Fluence-map generation for prostate intensity-modulated radiotherapy planning using a deep-neural-network. Sci Rep. 2019;9(1):15671.
13. Sheng Y, et al. Automatic planning of whole breast radiation therapy using machine learning models. Front Oncol. 2019;9:750.
14. Li, X., et al. Automatic IMRT planning via static field fluence prediction (AIP-SFFP): a deep learning method for real-time prostate treatment planning, in AAPM 61st Annual Meeting and Exhibition. St Antonio, Texas, USA; 2019.
15. Yuan L, et al. Standardized beam bouquets for lung IMRT planning. Phys Med Biol. 2015;60(5):1831–43.
16. Yuan L, et al. Incorporating single-side sparing in models for predicting parotid dose sparing in head and neck IMRT. Med Phys. 2014;41(2):021728.
17. Sheng Y, et al. Outlier identification in radiation therapy knowledge-based planning: a study of pelvic cases. Med Phys. 2017;44(11):5617–26.
18. Zhang J, et al. An ensemble approach to knowledge-based intensity-modulated radiation therapy planning. Front Oncol. 2018;8:57.
19. Yuan L, et al. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. Med Phys. 2012;39(11):6868–78.
20. Zhang J, et al. Modeling of multiple planning target volumes for head and neck treatments in knowledge-based treatment planning. Med Phys. 2019;46(9):3812–22.
21. Tikhonov AN. On the stability of inverse problems. Comptes Rendus De L Academie Des Sciences De L Urss. 1943;39:176–9.
22. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 2000;42(1):80–6.
23. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc Series B Methodol. 1996;58(1):267–88.
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B Stat Methodol. 2005;67:301–20.
25. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59.
26. Tol JP, et al. Evaluation of a knowledge-based planning solution for head and neck cancer. Int J Radiat Oncol Biol Phys. 2015;91(3):612–20.
27. Delaney AR, et al. Effect of dosimetric outliers on the performance of a commercial knowledge-based planning solution. Int J Radiat Oncol Biol Phys. 2016;94(3):469–77.
28. LJP M, Hinton GE. Visualizint high-dimensional data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
29. Wall PDH, Carver RL, Fontenot JD. Impact of database quality in knowledge-based treatment planning for prostate cancer. Pract Radiat Oncol. 2018;8(6):437–44.
30. Boutilier JJ, et al. Sample size requirements for knowledge-based treatment planning. Med Phys. 2016;43(3):1212–21.

14 Intelligent Respiratory Motion Management for Radiation Therapy Treatment

Martin J. Murphy

14.1 The Problem of Respiratory Movement During Radiotherapy

A number of treatment sites for external-beam radiation therapy, such as lung, breast, pancreatic, and liver cancers, move as the patient breathes, which compromises the precision of their irradiation. For the purpose of this chapter we will use lung tumors as the paradigm to represent this motion problem.

To achieve the best likelihood of effective beam coverage for a treatment target that moves during respiration, there are four basic approaches: (1) inhibit the movement via breathholding or physical restraints; (2) enlarge the therapy beam field so that the tumor never moves outside of it (the margin approach); (3) turn the beam on only when the tumor is at or near the beam isocenter (the gating approach); (4) move the beam or the patient synchronously with breathing so that the beam stays continuously aligned with the tumor (the tracking approach). In the tracking approach [45], the beam can be realigned by moving the linear accelerator (LINAC) itself [1, 9, 60, 61, 63], or shifting the multileaf collimator (MLC) aperture [5, 12, 20, 30, 39–42, 49, 55, 59, 68], or, in the case of a charged-particle beam, magnetically steering the beam [4]. Alternatively, the patient can be moved by shifting the couch, so that the tumor remains at a fixed beam isocenter [10, 38, 52, 54, 64, 65]. Gating and tracking are the two approaches that call for adapting to tumor motion in real time during treatment.

Some patients breathe regularly; some do not. Likewise, anatomical movement during breathing is sometimes simple; sometimes complex. Figures 14.1 and 14.2 illustrate two representative patients' breathing patterns, as measured by an optical marker placed on the chest. Figures 14.3 and 14.4 show how a sequence of

M. J. Murphy (✉)
Department of Radiation Oncology, Virginia Commonwealth University, Richmond, VA, USA

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_14

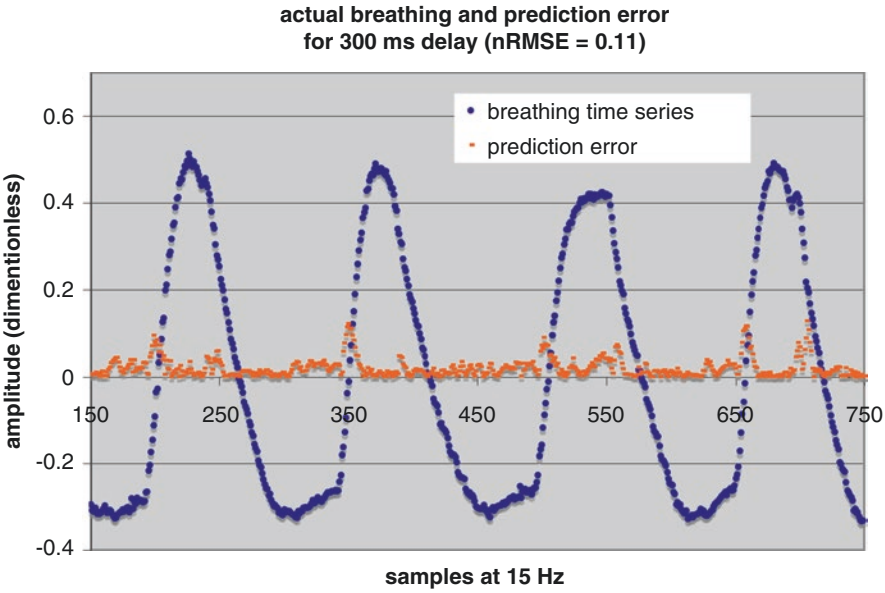


Fig. 14.1 An example of regular breathing

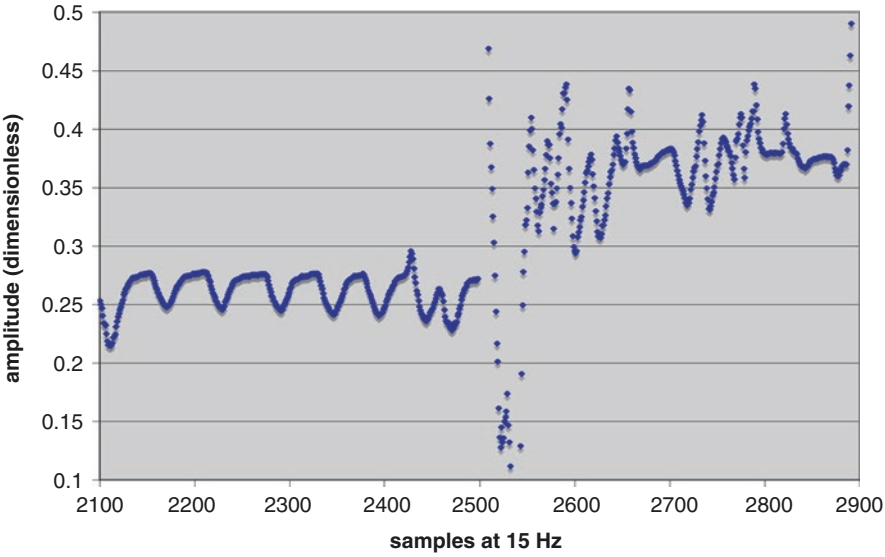


Fig. 14.2 An example of highly irregular breathing

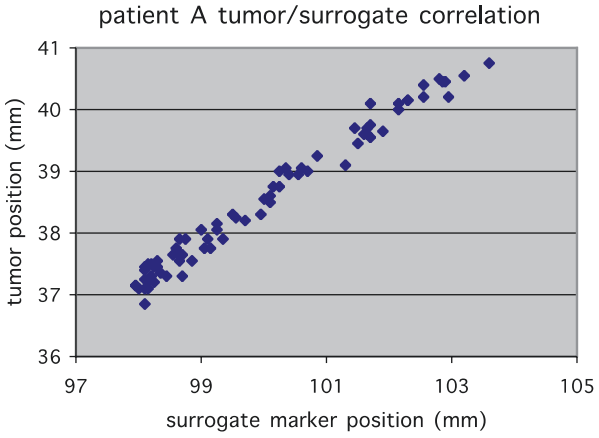


Fig. 14.3 A sequence of measurements of tumor position and chest marker position, showing a tight correlation over time [50]

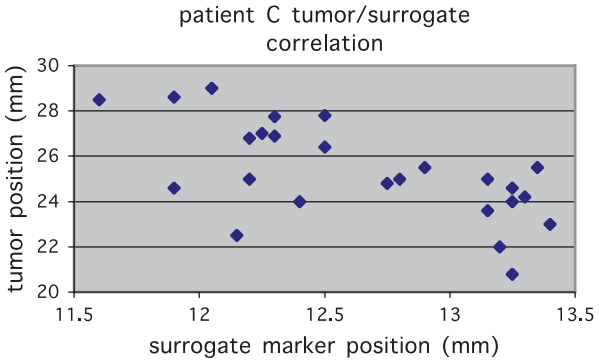


Fig. 14.4 An example of tumor and chest marker positions that do not maintain a tight correlation over time [50]

measurements of surface breathing movement (via the marker) can be related to the tumor's actual position, measured via x-ray fluoroscopic imaging. These four figures combine to demonstrate the complications presented by the breathing modeling and prediction problem. Although superficially regular (as in Fig. 14.1), normal breathing is not strictly periodic, but changes amplitude and period over time [34, 69]. In extreme cases, the breathing pattern can be highly irregular to the point of appearing chaotic (Fig. 14.2). The relationship between, e.g., tumor and chest movement can likewise range from stable, linear, and tightly correlated (Fig. 14.3) to unstable, nonlinear, or otherwise poorly correlated (Fig. 14.4) [50]. The tumor/surrogate correlation can vary over time (for example through changes in the relative amplitude and phase of the movements), so that a sequence of measurements of surrogate and tumor positions appear to be uncorrelated (as in Fig. 14.4). This greatly complicates any delivery compensation method (such as gating or tracking) that relies on respiratory surrogates to infer tumor movement.

These characteristics of breathing and tumor movement make it exceedingly difficult to devise an a priori bio-mechanical model of breathing that can accurately and continuously describe the movement of the anatomy and enable its prediction. The problem is instead a good candidate for a machine learning approach, using

algorithms that can learn to imitate and reproduce the movement patterns via training on examples of the patient's actual breathing. These models can be developed and applied during both the treatment planning and delivery processes.

The first step in respiration management occurs during the treatment planning stage, where the motion is first detected and characterized. Historically, this has been accomplished via 4DCT scanning. A 4DCT scan, however, typically has a coarse time resolution. Furthermore, respiratory movement can change from day to day, and even minute by minute, making any particular scan a potentially unrepresentative example. Daily rescanning to accommodate these variations becomes dose-intensive and requires replanning as well.

Lin et al. [35, 36] have developed and tested a super-learner model for motion prediction and management in the planning stage. Their model uses a collection of clinical and imaging features extracted from the treatment planning CT to train a respiration model that can anticipate actual tumor movement during treatment. They are primarily interested in predicting the likely direction and amplitude of tumor motion without recourse to a 4D CT. They propose that the model can be used to choose an appropriate motion compensation method during delivery (such as breathholding or gating) based on the range and other characteristics of the modeled tumor motion. Their super-learner is a hybrid algorithm incorporating Random Forest, adaptive neural network, and other machine learning components.

If the treatment planning process recommends that a dynamic compensation method (e.g., gating or tracking) should be used during treatment, then it is necessary to determine how the observable breathing signal that will be used during treatment is related to the actual tumor motion. This could involve observing the tumor itself via imaging (with or without implanted fiducials), the measurement of a surrogate breathing signal such as a chest marker, or both. If a surrogate such as a chest marker is to be used to infer the tumor position during treatment (as with the Varian RPM system, Varian Medical Systems, Palo Alto, CA), then Figs. 14.3 and 14.4 illustrate the potential difficulties. Sometimes the tumor and the surrogate movement have a nice, stable relationship that allows tumor prediction solely from the marker; sometimes the relationship is not simple or stable. In either case, a machine learning algorithm can be used to learn and follow the relationship, beginning with either 4DCT or fluoroscopic imaging during treatment planning, and then continuing during the dynamic control of treatment. In the following sections we will discuss dynamic respiratory compensation during treatment, with the understanding that the algorithms used for these processes are initialized during the treatment planning stage.

14.2 Dynamic Compensation Strategies during Delivery

There are two fundamental problems in adapting to tumor motion during treatment delivery: (1) determining the precise tumor position at any given time; (2) making a synchronized adaptive response to maintain beam/tumor alignment. Tumor position can either be measured directly via imaging or other detection methods, or it can be

inferred by measuring respiratory movement that is reliably correlated with the tumor movement and can act as a surrogate for it [2, 18, 24, 25, 29, 31, 50, 70, 75–77]. Here we are interested in learning and anticipating tumor position from some kind of surrogate respiratory signal.

The basic mechanisms for maintaining beam alignment with a moving tumor are illustrated schematically in Figs. 14.5 and 14.6. Figure 14.5 is an "open-loop" control architecture that is appropriate for either a gating or a beam tracking scheme. The tumor moves solely under the influence of patient movement (e.g., breathing). Respiration and/or tumor position sensors provide the input to the loop. The corrective signal propagates through various system components, each of which takes some time to react, resulting in a cumulative delay before the beam responds with the correction. Figure 14.6 is a "closed-loop" architecture in which the system's response combines with the patient's anatomical movement to influence the

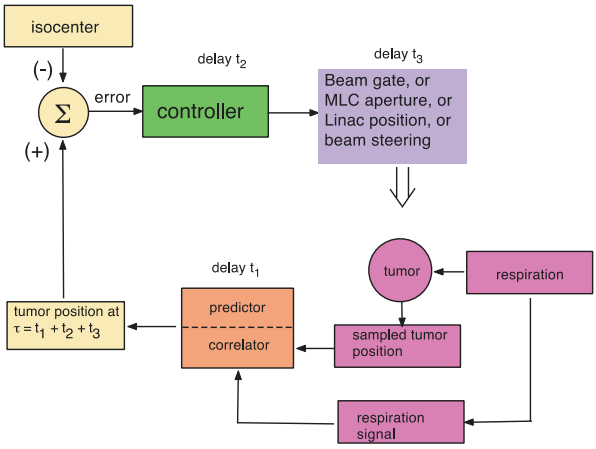


Fig. 14.5 An open control loop architecture for maintaining beam and tumor alignment

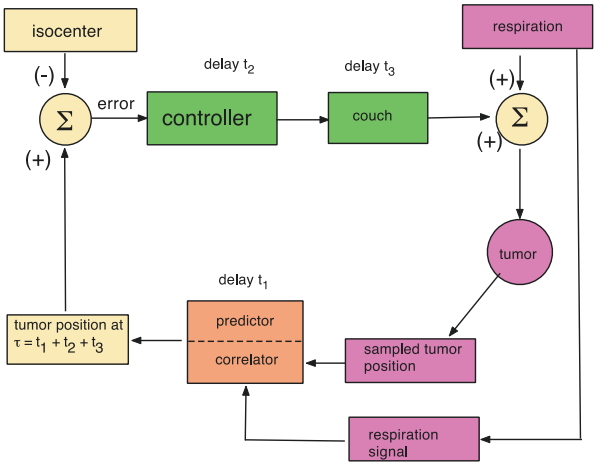


Fig. 14.6 A closed-loop beam alignment architecture

position of the target relative to the beam isocenter and thus the input to the loop. This is required for an adaptive system that moves the couch and patient relative to the beam as the tumor moves, so as to keep the tumor at a fixed position (set point) in space. In this case, respiration and couch shifts combine to move the tumor. In both architectures, the tumor position can be established either by following a surrogate breathing signal that correlates with tumor motion, or by directly observing the tumor's position, or both.

No adaptive response to movement can occur instantaneously, so it is necessary to compensate for delays between localization of the tumor and adjustment of the beam timing or alignment. This comes down to predicting the future tumor position (or its surrogate respiratory signal) by an amount equal to the response delay time so that the adaptation is synchronized to the tumor's actual position. These delays can range anywhere from 50 to 1000 ms, depending on the tracking/correction method [3, 9, 13–15, 26, 32, 58, 73]. Temporal prediction of the tumor's future position via machine learning algorithms has been the subject of a large and wide-ranging number of studies and continues to inspire new and more sophisticated approaches.

Intelligent delivery algorithms must be capable of continual adaptation to changes in the motion patterns, through methods of continuous retraining as the patient breathes. Many different tracking algorithms have been investigated (see, e.g., [11]). Among them, adaptive neural networks were found early on to be an effective machine learning approach to this problem [27]. They will therefore provide the entry point for the discussion of machine learning for dynamic treatment delivery.

The object of this discussion is the control loop element identified in Figs. 14.5 and 14.6 as the "correlator/predictor." This element receives as input some measurement of breathing and provides the anticipated position of the tumor as input to the beam or couch controller. To allow for control loop delays, the "correlator/predictor" must emulate the patient's breathing in order to predict the future respiratory signal and/or tumor position.

14.3 Using an Artificial Neural Network (ANN) to Model and Predict Breathing Motion

A machine/deep learning algorithm provides an effective means to perform the correlation/prediction function in both open-loop and closed-loop control systems. Its virtue lies in its trainability from real-world data, whereas the alternative would be to construct a detailed biomechanical model of respiration that would nevertheless need to be customized for individual patients.

An artificial neural network (ANN) is a trainable machine learning algorithm. One form that is very useful for predicting a signal amplitude has the basic architecture shown in Fig. 14.7. In this kind of application we have some measured signal $S(t)$ as input and a future instance of that signal $S(t + n)$ as the output target. The job of the ANN is to make an estimate $S'(t + n)$ of the future target signal from samples of the input signal. The input layer of the network is provided with discrete measurement samples from the past signal history, the hidden layers compute weighted

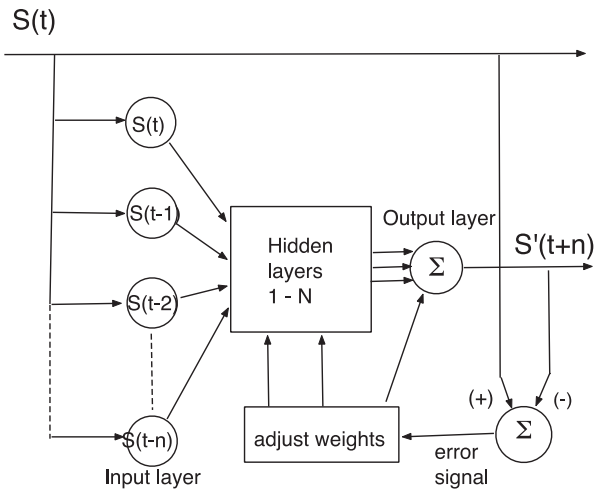


Fig. 14.7 An artificial neural network architecture to predict a signal amplitude *S(t)*

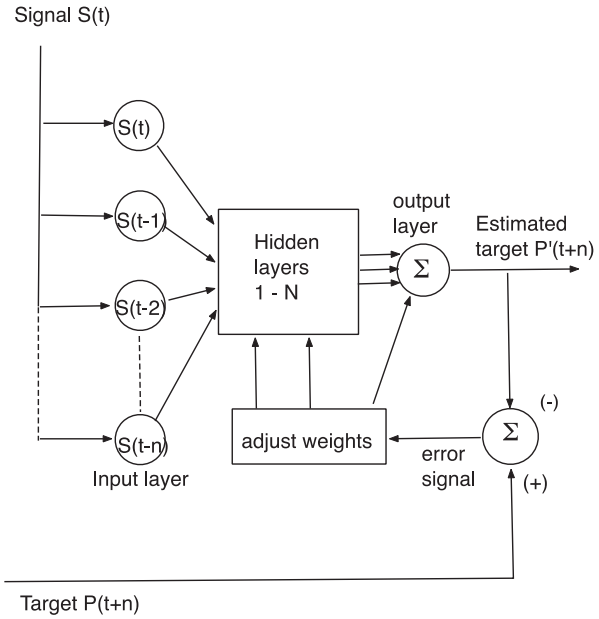


Fig. 14.8 An ANN configured to predict a different signal *P(t + n)* that is correlated with the input *S(t)*

combinations of the input data, and the output layer delivers an estimate of the target signal at a future time. In Fig. 14.7 the target signal is a future sample of the input signal, in which case the network is trained to imitate the input signal so that it can predict its future behavior. When the target signal finally arrives at time $t + n$, the prediction $S'(t + n)$ is compared to it, an error is computed, and this error is used to adjust the network weights so as to produce a more accurate prediction of the next sample. Figure 14.8 shows a configuration to use the input signal *S(t)* to predict a different signal *P(t)* that is correlated in some way with the input signal. In this case

the network is trained to predict the correlated target signal from the input. The target prediction might be for the present moment or some future time.

In our breathing prediction problem, we identify the input data with a sequence of discrete measurements of the patient's breathing. This could be as simple as the time history of the amplitude of a single breathing signal, such as a moving marker [44] or spirometer signal [25, 77], or it could comprise simultaneous measurements of multiple breathing signals [76]. If we are only interested in predicting breathing movement to compensate for a treatment system's lag time, then the target signal would be a future instance of the patient's measured breathing and the network's output would be an estimate of that future instance. If we are interested in deducing the tumor position from the measured breathing signal, then the input signal would be a breathing surrogate measurement and the target data would be a measurement of the tumor's spatial position at some particular time. It could be the tumor position at the present time, in which case the ANN makes a spatial correlation between the tumor and breathing motions, or it could be the future position of the tumor, in which case the network performs both a correlation and a temporal prediction to arrive at a good estimate of the tumor location.

14.4 Basic Neural Network Architecture for Correlation and Prediction

14.4.1 The Single Neuron, or Linear Filter

We can introduce the basic computational components of an artificial neural network for correlation and prediction by considering a simple network configured to predict the future amplitude of a single breathing signal sampled at discrete time intervals. It begins with a single neuron, as shown in Fig. 14.9. (This has historically been known as a linear perceptron.) The input is the amplitude history of the

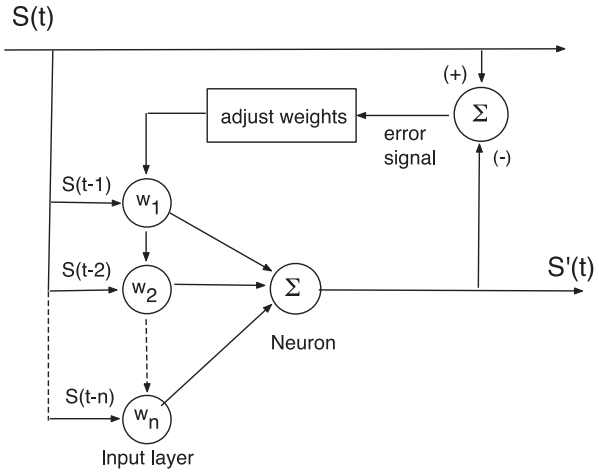


Fig. 14.9 A simple linear filter for prediction

measured signal *S(t),* sampled at *n* intervals of τ seconds. For breathing, which has a period of a few seconds for most people, τ might be on the order of 100 ms. We take the *N* most recent samples. Each sample is multiplied by a weight w_i and the *N* samples are summed:

$$S'(t) = {}^{N}\sum_{i=1} w_i S(t - i\tau) \tag{14.1}$$

If we stop here, we have a simple linear filter, where *S′(t)* is the filter's estimate of the signal amplitude at the present time, based on the previous N samples. *S′(t)* is compared to *S(t)* and the error is used to adjust the weights until the difference is minimized. If we want it to predict *S* at some future time $t + \Delta t$, rather than the present, we wait Δt seconds for the actual signal to arrive, compare it to *S′* to find the error, and adjust the weights accordingly.

The linear filter (i.e., a single neuron) in Fig. 14.9 and eq. 14.1 can do a reasonable job of predicting breathing, provided that the pattern isn't too changeable or irregular [46]. It provides a starting point to introduce several basic elements in the development of ANNs for prediction and correlation.

The weights are initially optimized in the training stage. For a basic signal prediction filter this typically consists of presenting the filter with pre-recorded signal histories that are representative of the signal that one ultimately wants to predict. For example, if one wants a filter customized to emulate and predict a particular patient's breathing, one begins by recording a segment of the patient's breathing signal. This is presented to the filter incrementally via a sliding window that is *N* samples wide, so that the filter gets a set of *N* samples up to a time *t* at the inputs, makes a prediction for $t + \Delta t$, compares the prediction to its target, which is the recorded signal at $t + \Delta t$, adjusts the weights, steps forward one sample, and repeats the process. This is an example of supervised sequential training. Sequential training has the advantage that, as the filter is presented with new breathing data that it hasn't seen before, it can continue the process, retraining continuously to adapt to new breathing patterns.

The initial training process must be done in such a way that it doesn't "see" future samples in the training stage before they would actually arrive in real time.

The simplest training algorithm for a linear neuron is the LMS (least mean square) method. Let $\boldsymbol{S}_i$ be the vector of *N* input samples from the *i'th* training signal history, let $\boldsymbol{W}_i$ be the vector of N weights assigned to the inputs, and let ε_i be the difference between the predicted and target signal sample. The updated weight vector is

$$\boldsymbol{W}_{i+1} = \boldsymbol{W}_i + \alpha \varepsilon_i \boldsymbol{S}_i \tag{14.2}$$

where α is a parameter that determines the speed of convergence. In the case of sequential training, each training signal history $\boldsymbol{S}_i$ is simply the previous signal history advanced by one sample.

There are numerous other algorithms to update the weights. For a more comprehensive review of ANN training algorithms, see for example [22], or [21], or another introductory textbook on neural networks.

14.4.2 The Basic Feed-Forward Artificial Neural Network for Prediction

Soon after the single-neuron perceptron was proposed as a primitive machine learning algorithm for pattern recognition, Minsky proved that it, and any linear combination of neurons performing the function of eq. 14.1, could only do linear discrimination and was incapable of performing even a simple exclusive-or function [43]. This led to the development of nonlinear networks of neurons for more complex pattern recognition and signal processing. Figure 14.10 is a schematic of the simplest nonlinear neural network—a feed-forward network with one hidden layer—for signal prediction. The inputs are distributed in parallel to two or more neurons like the one in Fig. 14.9 (the simple linear filter). These make up the "hidden" layer. (The layer is "hidden" because it can't be reached directly from the outside.) The output x of each neuron is passed through a nonlinear "activation" function (the sigmoid function in eq. 14.3 is the most commonly used), weighted, and summed with the others in the output neuron, which delivers the final signal estimate.

$$y = f(x) = 1/\left(1 + e^{-x}\right) \tag{14.3}$$

$$df \,/\, dx = y(1 - y) \tag{14.4}$$

The activation function must be nonlinear; otherwise the network is reducible to a single linear neuron and nothing is gained.

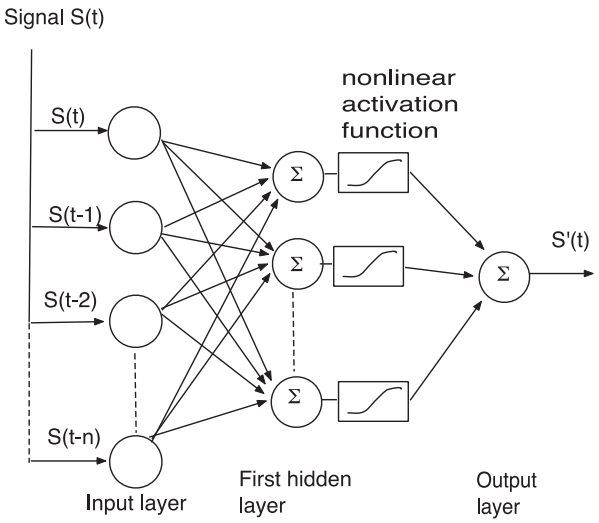


Fig. 14.10 A basic feed-forward network with one hidden layer of neurons and a single neuron in the output layer

14.4.3 Training the Feed-Forward Network

Each input to each neuron in the hidden and output layers in Fig. 14.10 has an independently variable weight. However, the weights in the hidden layer are "blocked" from the output signal error by the nonlinear activation function. This prevents a simple linear generalization of the LMS algorithm in eq. 14.2. The problem is solved by the method of error back propagation.

Although the basics of back propagation can be found in any textbook on neural networks (e.g., [21]), there is some advantage to providing them here, using the simple two-layer network in Fig. 14.10 as the architecture. Let layer 1 be the hidden layer and layer 2 be the output layer (in this case just one neuron.) Let the index i apply to the data samples and j to the number of neurons in layer 1 (and also the equal number of input weights to layer 2). Let $\boldsymbol{W}_{1,j}$ be the vector of weights for the *j'th* neuron in layer 1 (with components $w_{1,ji}$) and $\boldsymbol{W}_2$ be the weight vector for the output (layer 2) neuron (with components $w_{2,j}$). The outputs of the layer 1 neurons are $x_{1,j}$ before activation and $y_{1,j}$ after activation. The error in the predicted output signal is ε.

In the forward pass, the delta is calculated for layer 2:

$$\Delta_2 = \overline{\varepsilon}$$

In the backward pass, the deltas for layer 1 propagate through the derivative of the transfer function:

$$\Delta_{1,j} = \left[y_{1,j} \left(1 - y_{1,j}\right)\right]\left[\Delta_2 w_{2,j}\right].$$

The incremental changes to the weights in the two layers are then calculated (in this example via LMS):

$$\delta w_{2,j} = \alpha \Delta_2 \, y_{1,j}$$
$$\delta w_{1,ji} = \alpha \Delta_{1,j} \, S_i.$$

In addition to LMS, there are a number of other algorithms that can be used to update the weights [21, 22]. Regardless of which one is used, there are some general principles to be followed to get the best results. The first step is to initialize all of the weights. The usual practice is to choose them randomly, because this gets the neurons acting independently. However, there is always some chance that a random initialization will come up with an unfavorable filter that performs badly. This can be avoided by performing the random initialization and subsequent training multiple times, while testing each fully-trained filter on an independent validation signal. The set of weights that does the best job of predicting the validation signal becomes the optimal filter for application to test signals. The validation signal can be any part of the pre-recorded signal that wasn't used for training.

It is also generally the case that a single pass through the training data will not result in optimal convergence of the weights. It is therefore customary to run through the training data repeatedly, starting each subsequent training pass at the weights

from the prior training pass. Each pass is called an *epoch*. However, there is the risk of *overtraining* the filter after too many epochs. In this case the filter becomes completely optimized to emulate the training data, but cannot generalize effectively to signals it has not yet seen. This can be avoided by testing each epoch of trained filter on the validation data, and terminating the training when the filter's performance on the validation set is clearly worse than its performance on the training data.

The feed-forward breathing prediction network in Fig. 14.10 can be generalized to perform temporal prediction and position correlation by comparing its output to some measure of tumor position, as in Fig. 14.8.

14.4.4 The Recurrent Network

A static neural network is trained and then applied without modification to test data. An adaptive neural network retrains the neural weights as it acquires new data. A recurrent network is a closed-loop feedback architecture in which signals from the hidden and output layers are fed back to previous hidden layers and/or to the input layer. This architecture is inspired by the observation that the human brain is a recurrent network of neurons. Recurrent networks are sometimes referred to as dynamic networks because they adapt the neuron weights to ongoing experience through the feedback loop(s). The timing of the feedback determines the short- or long-term memory of the network. In Fig. 14.11, a simple recurrent network for prediction feeds the previous *m-1* predictions back to the input layer at each time step *S(t)* of the input signal. The output signals are held back by the prediction interval τ before they are supplied to the input, so that the error between *S(t)* and *S′(t)* can

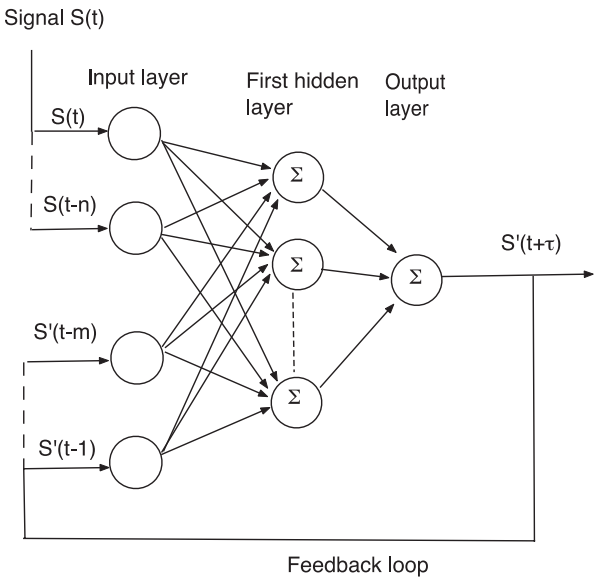


Fig. 14.11 A basic recurrent network

be computed and used to update the weights. The hoped-for advantage is that the raw input data from the (potentially noisy) measurements is supplemented by filtered data from the outputs that will smooth out the network's response. A recurrent network can be trained in the same way as a feed-forward network, e.g., via back propagation. Mafi and Moghadam [37] have investigated the use of recurrent/dynamic networks for breathing motion prediction.

14.5 Performance of Basic Neural Networks to Predict Tumor Motion

The problem of predicting breathing with an artificial neural network has been studied by a number of researchers (e.g., [8, 19, 27, 28, 33, 44, 47, 48, 56, 62, 67]).

14.5.1 Breathing Prediction Examples for a Simple Feed-Forward Network

A feed-forward network can have more than one hidden layer, each of which can have multiple neurons. It can also have more than one neuron in the output layer (cf Fig. 14.12). The output of each hidden neuron is passed through the activation function before it is summed by the neurons in the next layer. It has been found, however, that a feed-forward network with just one hidden layer of two neurons, and one output neuron, can predict breathing more or less as well as more complicated layered architectures [8, 27, 44]. We can therefore use such a simple network to learn

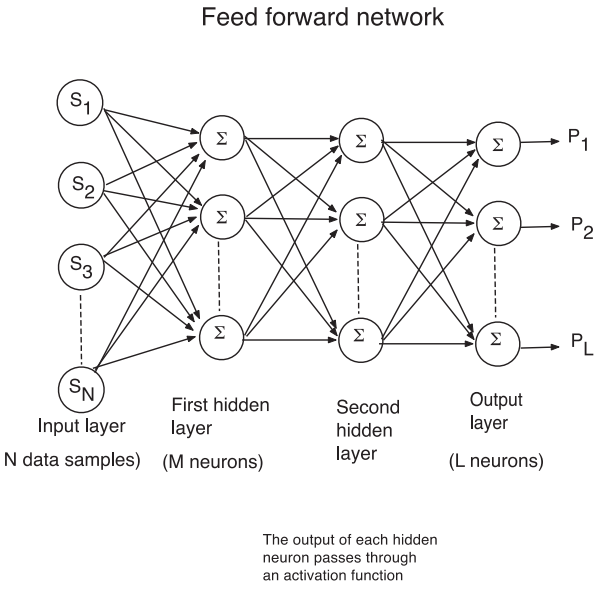


Fig. 14.12 A general feed-forward network with multiple layers containing multiple neurons

some important things about basic breathing prediction. The following examples of feed-forward ANN results for breathing prediction were all obtained with a single breathing amplitude (displacement of a chest marker) for the input signal, two neurons in the hidden layer, a sigmoid activation function, and one output neuron (for the future signal amplitude). After initial training via LMS, the network was updated (adapted) each time a new breathing data point became available. To quantify the accuracy of breathing prediction, the dimensionless quantity of normalized root mean square error (eq. 14.5) was used to compare the predicted (P_i) and actual (D_i) future amplitudes, for prediction horizons (i.e., lag times) ranging from 100 ms to 500 ms.

$$nRMSE = \left[\Sigma_i \left(D_i - P_i \right)^2 / \Sigma_i \left(D_i - \mu \right)^2 \right]^{1/2} \tag{14.5}$$

Here μ is the mean of all of the observations.

There are several parameters to determine when designing the feed-forward neural network prediction filter—the length (in seconds) of the input signal history and the number of samples in that history (i.e., the sampling rate), the number of training epochs, and the training rate $\propto$ in the LMS updating rule (eq. 14.2). Without going into detail about the testing of the network, which is reported in detail in [48], it suffices to say that the performance of the network in predicting a variety of different breathing examples was explored by varying each of these network parameters, to find the values that provided the best results. One obvious question to ask is whether a single network setup can do a reasonable job of predicting a wide range of breathing patterns, or if the filter setup needs to be optimized to each individual patient. To answer this question, the filter setup was first optimized for each patient breathing history, and its accuracy was noted. Then a globally optimal sampling length and rate, number of training epochs, and training rate was identified in the results and used to configure a standardized filter, which was then tested against all of the individual patient histories. Figure 14.13 shows the results [48].

Patients 1–14 were randomly selected from a cohort treated for lung cancer, and displayed a wide range of breathing patterns; patients 15–27 were healthy volunteers coached to regularize their breathing via audiovisual feedback [16,

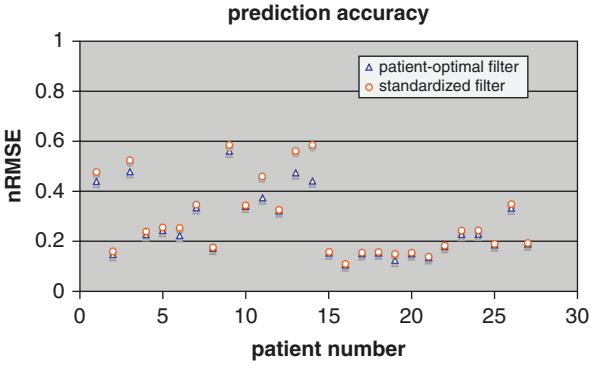


Fig. 14.13 Prediction accuracy of an ANN customized to each patient compared to the accuracy of an ANN with a fixed configuration [48]

17]. The standardized filter did essentially as well as the personalized filter for the healthy coached patients, and continued to do reasonably well even for the most erratic lung cancer patients. This offers encouragement that it is not always necessary to go through an involved filter optimization process for each patient.

The accuracy of any predictive filter can be expected to diminish if the breathing pattern changes over time, simply because the filter must retrain itself to adapt to the changes, and that takes time. This can be demonstrated by calculating a breathing regularity measure and then looking at prediction accuracy as a function of that measure.

For a finite length of continuous patient breathing signal $S(t)$, the autocorrelation coefficient $C(\tau)$ is defined as the cross-correlation integral of $S(t)$ with itself, at delay time τ:

$$C(\tau) = \int S(t) S(t-\tau) dt \tag{14.6}$$

For a stationary periodic signal the average value of $C(\tau)$ versus τ will be approximately constant, while for a non-stationary (time-changing) signal the average of $C(\tau)$ will become smaller with increasing τ. We can characterize the stability of the signal by the inverse of the rate at which the average correlation coefficient decays with τ. Call this the correlation decay time. To compute the decay time, a 60 s window was set at a point in the breathing time series and $C(\tau)$ was computed for $0 < \tau < 60$ s. The peak values of the positive half cycle of the autocorrelation coefficient were plotted in a semi-log scale as a function of τ. The inverse slope of the graph gave the decay time for that particular position of the breathing signal window. A rapidly changing signal will have a short decay time; a slowly changing signal will have a long decay time; a perfectly stationary signal will have an infinite decay time.

Figure 14.14 shows the prediction accuracy of the ANN filter as a function of the breathing signal's decay time (from [48]). As expected, rapidly evolving breathing patterns are harder to predict, no matter how well the filter is designed.

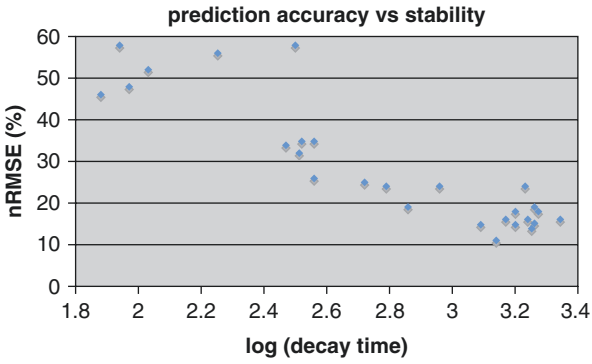


Fig. 14.14 The prediction accuracy of a neural network filter as a function of the stability of the breathing signal, as characterized by the decay time of its autocorrelation. Shorter decay times correspond to more rapidly changing breathing patterns (from [48])

14.6 Advanced Neural Network Architectures

14.6.1 Quadratic Neural Unit

In a basic feed-forward neural network perceptron, the neuron makes a linear summation of its weighted inputs. It uses a sigmoid activation function to pass the neuron output(s) of each layer to the next layer, where they are again combined in a linear summation. As noted earlier, without the nonlinear activation function connecting the layers, a multiple-layer perceptron reduces to a single layer perceptron. In a polynomial neural network, the neuron makes a higher-order polynomial combination of the inputs. For example, a quadratic neural unit computes:

$$y = \Sigma_{ij} w_{ij} x_i x_j$$

Bukovsky et al. [6] have tested a neural network with a quadratic neural unit for predicting breathing time series. The network was trained via a Levenberg–Marquardt algorithm adapted for the polynomial neurons. They found that it can be retrained in real time and achieved good accuracy out to a prediction horizon of one second.

14.6.2 Using a Kalman filter to Predict/Correct as Part of the Training Loop

Consider a system that is being observed via periodic data samples. Suppose each data sample fluctuates randomly due to the behavior of the system itself (plant noise) and uncertainty in the measurements (measurement noise). If the system's evolving state is governed by a linear function, then the best estimate of the next sample is provided by the Kalman filter predictor, which is a continuously updating algorithm that takes its present estimate of the system's state, makes a prediction of the next signal sample, combines it with the next available data measurement, and calculates a correction to update the state of the system, which is then recirculated via a prediction/correction loop. Such a filter continuously adapts to the evolution of the system.

A breathing signal has variability that can be divided between two sources—irregularity in the actual breathing (plant noise) and errors in the observations (measurement noise). This has inspired studies to predict breathing with a Kalman filter. However, the breathing system is nonlinear and consequently the Kalman filter must be generalized to an Extended Kalman Filter (EKF). The Extended Kalman Filter attempts to linearize the observations (typically via a Taylor expansion) so that the basic Kalman prediction/correction algorithm can be used. Unfortunately, this has proved problematic and the performance of an EKF for breathing has generally not been as favorable as other methods.

Looking back to Fig. 14.7, one sees that the weights are updated from the most recent error signal. These error signals also incorporate plant and measurement

noise, which suggests that an extended Kalman filter can be used to train an ANN [74]. In this application it would be used to calculate (predict) each successive update to the weights, and thus the state of the ANN, rather than model and update the breathing state itself. This could combine the advantages of both the ANN and the EKF. Figure 14.15 illustrates the strategy. This breathing prediction architecture has been studied by e.g., Lee et al. [33], who present the details for computing the EKF prediction/correction of the network weights.

A recurrent EKF-ANN with p outputs describes the system state with a vector of s neuron weights, which requires an error covariance matrix of size s^2 and computational complexity of order $O(ps^2)$. This can become demanding when there are a large number of inputs to the network. However, it is possible to decouple the individual weights in the EKF stage, so that the error covariance matrix becomes block-diagonal and the computational complexity is reduced to order $O(ps)$ [33, 53].

14.6.3 A Network with Multiple Breathing Signal Inputs

The discussion of ANNs for breathing prediction and correlation has so far used the simple case of a single one-dimensional breathing signal $S(t)$ supplied as input. In a clinical setting one can often have multiple sensors, each measuring up to three spatial degrees of freedom in movement. The CyberKnife (Accuray Incorporated, Sunnyvale CA) utilizes an array of optically tracked infrared emitters distributed on the patient's chest and abdomen to record breathing. The breathing data are correlated with periodic x-ray measurements of tumor position to provide a targeting signal to the linear accelerator, which makes compensating corrections to the

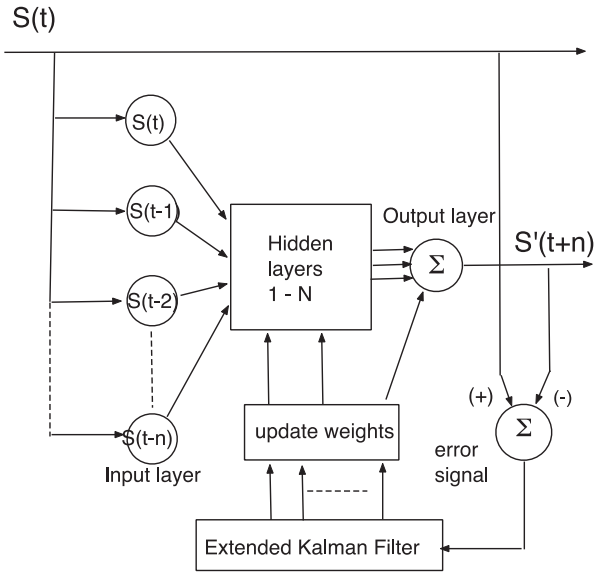


Fig. 14.15 A recurrent network employing an extended Kalman filter to compute the updates to the weights

treatment beam's direction [60, 61]. This has the advantage of multiple redundant measurements to reduce the influence of measurement noise and the capacity to determine during training whether the patient is a chest or abdominal breather.

The most basic generalization to m breathing signal sources is simply to make up an input layer that provides n taps of each signal, for a total of nm input nodes. However, for a breathing patient the m sets of input samples will be correlated with one another. In the EKF-ANN this correlation will be reflected in the error covariances. This can be dealt with by coupling the Kalman filters for each signal channel (while keeping the weights decoupled, as above). This has been studied by Lee et al. [33]. Alternatively, one can make a principal components analysis (PCA) of the signals to obtain an input vector of maximally uncorrelated data.

14.6.4 Deep Learning Neural Networks for Prediction

One of the limitations of conventional neural networks used for sequential pattern recognition is their tendency to forget older input data. This is partially mitigated in the short term in a recurrent neural network, but when a network is retrained in real time to adapt to a changing input, the older learning parameters are deliberately modified in response to the newer, different input signal. This has the advantage of keeping the network up to date, but limits the ability of the network to learn and remember its experiences with a large set of prior data. To mitigate this shortcoming, Long Short-Term Memory (LSTM) networks were developed [23].

In adaptive retraining for breathing prediction, the basic network architecture (the model) is fixed for a particular patient, but the learning parameters are modified in real time in order to adapt to changes in the patient's breathing. It has been found that, for a cohort of patients with regular breathing patterns, a standard network model could be used for all of the patients, but when patient breathing becomes irregular, it becomes increasingly favorable to tailor the network model to the individual patient. Tailoring the network model to each patient is a time-consuming process that must be accomplished before treatment begins. It would be highly desirable to have a standardized network architecture that has been developed and trained on a cohort of patients of widely variable breathing patterns. Such a network needs "memory" of each patient that it has seen in the past. Lin et al. [35, 36] have applied a LSTM network model to a cohort of patients whose individual breathing patterns were recorded with the Real-Time Position Measurement data (RPM, Varian Medical Systems, Palo Alto, CA). Wang et al. [72] similarly apply deep Long Short-Term Memory to the breathing prediction problem.

14.7 Support Vector Regression (SVR) as an Alternative to Neural Networks for Breathing Prediction

One of the drawbacks to a conventional feed-forward neural network for pattern recognition is its vulnerability to becoming trapped in local minima, leading it to miss the true loss function minimum. Feed-forward ANNs apply a hard decision boundary that can be distracted by outlier training data. This becomes more problematic as the multi-dimensional decision surface becomes fuzzy, with individual training examples overlapping so as to blur the true minimum. The concept of a support vector machine was developed by Vapnik [71] to overcome this problem. An SVM is a probabilistic learning algorithm that draws upon individual training examples in the same way as an ANN but uses a soft margin in the decision surface to allow the classifier to tolerate individual classification errors as long as they are less than some upper limit *epsilon*.

Minor variations of these algorithms are referred to as Support Vector Regression (SVR) [57]. Studies [7] have shown the SVR approach to achieve better time-series prediction accuracy when applied to breathing prediction for adaptive motion compensation.

14.8 Probabilistic Neural Networks

Another statistical learning algorithm is the so-called Probabilistic Neural Network (PNN) developed by Specht [66] in the 1960s at about the same time as Vapnik's early work on the SVM. In a PNN, the training data are smoothed using a Parzen window [51] with a Gaussian kernel to provide a functional representation of the multi-parameter training data set and then individual classification inputs are located in that multi-parameter function space and their classification is assigned a probability of being correct. This accommodates in a natural way both sparse training data sets and outliers in the data. Specht has shown that the PNN functions in the same way as a feed-forward neural network in which the sigmoid activation function has been replaced by an exponential function. These algorithms have been successfully applied to a variety of time-series prediction and forecasting problems, but at the time of writing there is no evidence of their application to respiratory time series.

14.9 Summary

The use of machine learning techniques to analyze and predict patient breathing and tumor motion patterns begins in the treatment planning stage, where they can be used to reduce the amount of CT imaging to characterize 4D movement. Adaptive breathing compensation during radiation therapy requires a means to compensate for inevitable lag times in the adaptation process by predicting tumor movement either directly from imaging data or indirectly from surrogate breathing data. Although breathing appears superficially regular in most individuals, it is actually

variable in period and amplitude. Furthermore, the relative movement of different parts of the anatomy under the influence of breathing can change over time, making it difficult to associate tumor movement with other surrogate movements. Machine learning algorithms offer an attractive way to emulate these complicated behaviors without recourse to biomechanical modeling. They are intrinsically capable of conforming to individual breathing patterns and adapting in real time to changes in breathing behavior.

The artificial neural network is a simple machine learning algorithm that has been shown to be effective at predicting breathing behavior. It offers a clear advantage over a basic linear adaptive filter without much additional computational burden [46]. More usefully, it has been found by numerous researchers that an acceptable level of prediction accuracy can be achieved with a very simple network architecture, and that adding feedback loops or more layers with more neurons often provides little or no further improvement [8, 27, 44]. Furthermore, it is not always necessary to customize the network architecture to each individual patient [48]. This is most clearly the case when steps are taken to regularize an individual's breathing through training and feedback [48]. Nevertheless, deep learning neural networks have been shown to be capable of generalized training from cohorts of patients that exhibit variable breathing characteristics [35, 36].

While the latencies of various motion-adaptive therapy devices can be (and have been) systematically reduced, so that temporal prediction becomes less important in a tumor tracking system, the problem of tracking the tumor's motion from surrogate breathing signals remains. This application of ANNs has not been studied as well as temporal prediction and invites further investigation.

The importance of intelligent breathing management during radiation therapy is evident in the large number of research studies devoted to it over the past twenty years. During that time, developers have progressed from simple linear filters to highly sophisticated machine learning and deep learning algorithms. The problem remains of significant interest to engineers and clinicians alike.

References

1. Adler JR, Murphy MJ, Chang S, Hancock S. Image-guided robotic radiosurgery. Neurosurgery. 1999;44:1299–306.
2. Ahn S, Yi B, Suh Y, Kim J, Lee S, Shin S, Choi E. A feasibility study on the prediction of tumour location in the lung from skin motion. Br J Radiol. 2004;77:588–96.
3. Bedford JL, Fast MF, Nill S, et al. Effect of MLC tracking latency on conformal volumetric modulated arc therapy (VMAT) plans in 4D stereotactic lung treatment. Radiother Oncol. 2015;117(3):491–5.
4. Bert C, Saito N, Schmidt A, Chaudhri N, Schardt D, Rietzel E. Target motion tracking with a scanned particle beam. Med Phys. 2007;34(12):4768–71.
5. Booth JT, Caillet V, Hardcastle N, et al. The first patient treatment of electromagnetic-guided real time adaptive radiotherapy using MLC tracking for lung SABR. Radiother Oncol. 2016;121(1):19–25.

6. Bukovsky I, Homma N, Ichiji K, Cejnek M, Slama M, Benes PM, Bila J. A fast neural network approach to predict lung tumor motion during respiration for radiation therapy applications. Biomed Res Int. 2015;
7. Choi SW, Chang Y, Kim N, Park SH, Song SY, Kang HS. Performance enhancement of respiratory tumor motion prediction using adaptive support vector regression: comparison with adaptive neural network method. Int J Imaging Systems and Technology. 2014;2(1):8–15.
8. Davuluri P, Hobson RS, Murphy MJ, Najarian K. Performance comparison of Volterra predictor and neural network for breathing prediction. In: First international conference on biosciences. Cancun, Mexico: IEEE; 2010. p. 6–10.
9. Depuydt T, Verellen D, Hass O, Gevaert T, Linthout N, Duchateau M, Tournel K, Reynders T, Leysen K, Hoogeman M, Storme G, De Ridder M. Geometric accuracy of a novel gimbals based radiation therapy tumor tracking system. Radiother Oncol. 2011;98(3):365–72.
10. Ehrbar S, Schmid S, Johl A, et al. Comparison of multi-leaf collimator tracking and treatment-couch tracking during stereotactic body radiation therapy of prostate cancer. Radiother Oncol. 2017;125(3):445–52.
11. Ernst F, Schweikard A. Robotic LINAC tracking based on correlation and prediction. In: Murphy MJ, editor. Motion adaptation in radiation therapy. New York: Taylor and Francis; 2012.
12. Falk M, Munck AF, Rosenchöld P, Keall P, Catell H. Cho B C, Poulson P, Povsner S, Sawant a, Zimmerman J and Korreman S, real-time dynamic MLC tracking for inversely optimised arc therapy. Radiother Oncol. 2010;94:218–23.
13. Fast MF, Nill S, Bedford JL, Oelfke U. Dynamic tumor tracking using the Elekta agility MLC. Med Phys. 2014;41(11):111719.
14. Fledelius W, Keall PJ, Cho B, et al. Tracking latency in image-based dynamic MLC tracking with direct image acquisition. Acta Oncol. 2011;50(6):952–9.
15. Freislederer P, Reiner M, Hoischen W, et al. Characteristics of gated treatment using an optical surface imaging and gating system on an Elekta linac. Radiation Oncology. 2015;10(1):68.
16. George R, Chung TD, Vedam SS, Ramakrishnan V, Mohan R, Weiss E, Keall PJ. Audio-visual biofeedback for respiratory-gated radiotherapy: impact of audio instruction and audio-visual biofeedback on respiratory-gated radiotherapy. Int J Radiat Oncol Biol Phys. 2006;65(3):924–33.
17. George R, Suh Y, Murphy M, Williamson J, Weiss E, Deall P. On the accuracy of a moving average algorithm for target tracking during radiation therapy treatment delivery. Med Phys. 2008;35(6):2356–65.
18. Gierga DP, Brewer J, Sharp GC, Betke M, Willett CG, Chen GTY. The correlation between internal and external markers for abdominal tumors: implications for respiratory gating. Int J Radiat Oncol Biol Phys. 2005;61(5):1551–8.
19. Goodband JH, Haas OCL, Mills JA. A comparison of neural network approaches for on-line prediction in IGRT. Med Phys. 2008;35(3):1113–22.
20. Hansen R, Ravkilde T, Worm ES, et al. Electromagnetic guided couch and multileaf collimator tracking on a TrueBeam accelerator. Med Phys. 2016;43(5):2387.
21. Haykin S. Neural networks and learning machines. 3rd ed. London: Pearson; 2009.
22. Haykin S. Kalman filtering and neural networks. New York: Wiley Interscience; 2001.
23. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80.
24. Hoisak JDP, Sixel KE, Tirona R, Cheung PCF, Pignol J-P. Correlation of lung tumor motion with external surrogate indicators of respiration. Int J Radiat Oncol Biol Phys. 2004;60(4):1298–306.
25. Hoisak JDP, Sixel KE, Tirona R, Cheung PCF, Pignol J-P. Prediction of lung tumour position based on spirometry and on abdominal displacement: accuracy and reproducibility. Radiother Oncol. 2006;78(3):339–46.
26. Hoogeman M, Prévost J-B, Nuyttens J, Pöll J, Levendag P, Heijmen B. Clinical accuracy of the respiratory tumor tracking system of the Cyberknife: assessment by analysis of log files. Int J Radiat Oncol Biol Phys. 2009;74:297–303.
27. Isaksson M, Jalden J, Murphy MJ. On using an adaptive neural network to predict lung tumor motion during respiration for radiotherapy applications. Med Phys. 2005;32(12):3801–9.

28. Kakar M, Mystrom H, Aarup LR, Nottrup TJ, Olsen DR. Respiratory motion prediction by using the adaptive neuro fuzzy inference system(ANFIS). Phys Med Biol. 2005;50:4721–8.
29. Kanoulas E, Aslam JA, Sharp GC, Berbeco RI, Nishioka S, Shirato H, Jiang SB. Derivation of the tumor position from external respiratory surrogates with periodical updating of the internal/external correlation. Phys Med Biol. 2007;52(17):5443–56.
30. Keall PJ, Cattell H, Pokhrel D, Dieterich S, Wong K, Murphy MJ, Vedam SS, Wijesooriya K, Mohan R. Geometric accuracy of a system for real time target tracking with a dynamic MLC. Int J Rad Onc Biol Phys. 2006;65(5):1579–84.
31. Koch N, Liu HH, Starkschall G, Jacobson M, Forster KM, Liao Z, Komaki R, Stevens CW. Evaluation of internal lung motion for respiratory-gated radiotherapy using MRI: part I–correlating internal lung motion with skin fiducial motion. Int J Radiat Oncol Biol Phys. 2004;60(5):1459–72.
32. Krauss A, Nill S, Oelfke U. The comparative performance of four respiratory motion predictors for real-time tumour tracking. Phys Med Biol. 2011;56:5303–17.
33. Lee SJ, Motai Y, Murphy M. Respiratory motion estimation with hybrid implementation of extended Kalman filter. IEEE Trans Ind Electron. 2012;59(11):4421–32.
34. Liang P, Pandit JJ, Robbins PA. Non-stationarity of breath-by-breath ventilation and approaches to modeling the phenomenon. In: Semple SJG, Adams L, Whipp BJ, editors. Modeling and control of ventilation. New York: Plenum; 1995. p. 117–21.
35. Lin H, Zou W, Li T, Feigenberg SJ, Teo B-K, Dong L. A super-learner model for tumor motion prediction and management in radiation therapy: development and feasibility evaluation. Nat Sci Rep. 2019a;14868
36. Lin H, Shi C, Wang B, Chan MF, Tang X, Ji W. Towards real-time respiratory motion prediction based on long short- term memory neural networks. Phys Med Biol. 2019b;64(8):1361–72.
37. Mafi M, Moghadam SM. Real-time prediction of tumor motion using a dynamic neural network. Med and Biol Engineering and Computing. 2020;58:129–39.
38. Malinowski K, D'Soua WD. Couch-based target alignment. In: Murphy MJ, editor. Motion adaptation in radiation therapy. New York: Taylor and Francis; 2012.
39. McQuaid D, Webb S. IMRT delivery to a moving target by dynamic MLC tracking: delivery for targets moving in two dimensions in the beam's-eye view. Phys Med Biol. 2006;51:4819–39.
40. McQuaid D, Webb S. Target-tracking deliveries using conventional multileaf collimators planned with 4D direct-aperture optimization. Phys Med Biol. 2008;53:4013–29.
41. McQuaid D, Partridge M, Symonds Tayler R, Evans PM, Webb S. Target-tracking deliveries on an Elekta linac: a feasibility study. Phys Med Biol. 2009;54:3563–78.
42. McQuaid D, Webb S. Fundamentals of tracking with a linac MLC. In: Murphy MJ, editor. Motion adaptation in radiation therapy. New York: Taylor and Francis; 2012.
43. Minsky M, Papert S. Perceptrons. Cambridge: MIT Press; 1969.
44. Murphy MJ, Jalden J, Isaksson M. Adaptive filtering to predict lung tumor breathing motion free breathing. Proceedings of the 16th International Congress on Computer-assisted Radiology and Surgery. Paris; 2002. P. 539–544.
45. Murphy MJ, Tracking moving organs in real time. In: Chen S and Bortfeld R, editors. Seminars in radiation oncology, vol.14 (1). 2004. p. 91–100.
46. Murphy MJ, Dieterich S. Comparative performance of linear and nonlinear neural networks to predict irregular breathing. Phys Med Biol. 2006;51:5903–14.
47. Murphy MJ. Using neural networks to predict breathing motion. In: Seventh International Congress on Machine Learning Applications. San Diego, CA: IEEE; 2008. p. 528–32.
48. Murphy MJ, Pokhrel D. Optimization and evaluation of an adaptive neural network filter to predict respiratory motion. Med Phys. 2009;36(1):40–7.
49. Neicu T, Shirato H, Seppenwoolde Y. Synchronized moving aperture radiation therapy (SMART); average tumour trajectory for lung patients. Phys Med Biol. 2003;48:587–98.
50. Ozhasoglu C, Murphy MJ. Issues in respiratory motion compensation during external-beam radiotherapy. Int J Radiat Oncol/Biol/Phys. 2002;52:1389–99.

51. Parzen E. On estimation of a probability density function and mode. Ann Math Stat. 1962;33:1065–76.
52. Podder TK, Buzurovic I, Galvin JM, Yu Y. Dynamics-based decentralized control of robotic couch and multi-leaf collimators for tracking tumor motion. IEEE Int Conf Robotics Automat. 2008;19(23):2496–502.
53. Puskorius GV, Feldkamp LA. Neurocontrol of nonlinear dynamical systems with Kalman filter trained recurrent networks. IEEE Trans Neural Netw. 1994;5(2):279–97.
54. Qiu P, D'Souza WD, McAvoy TJ, Liu KJR. Inferential modeling and predictive feedback control in real-time motion compensation using the treatment couch during radiotherapy. Phys Med Biol. 2007;52:5831–54.
55. Rangaraj D, Papiez L. Synchronized delivery of DMLC intensity modulated radiation therapy for stationary and moving targets. Med Phys. 2005;32:1802–17.
56. Ren Q, Nishioka S, Shirato H, Berbeco RI. Adaptive prediction of respiratory motion for motion compensation radiotherapy. Phys Med Biol. 2007;52(22):6651–61.
57. Riaz N, Shanker P, Wiersma R, Gudmundsson O, Mao W, Widrow B, Xing L. Predicting respiratory tumor motion with multi-dimensional adaptive filters and support vector regression. Phys Med Biol. 2009;54(19):5735–48.
58. Saito M, Sano N, Ueda K, et al. Technical note: evaluation of the latency and the beam characteristics of a respiratory gating system using an Elekta linear accelerator and a respiratory indicator device. Med Phys. 2018;45(1):74–80.
59. Sawant A, Venkat R, Srivastava V, Carlson D, Povzner S, Cattell H, Keall P. management of three-dimensional intrafraction motion through real-time DMLC tracking. Med Phys. 2008;35:2050–61.
60. Schweikard A, Glosser G, Bodduluri M, Murphy MJ, Adler JR. Robotic motion compensation for respiratory movement during radiosurgery. Computer-Aided Surgery. 2000;5:263–77.
61. Schweikard A, Shiomi H, Adler J. Respiration tracking in radiosurgery. Med Phys. 2004;31(1):2738–41.
62. Sharp GC, Jiang SB, Shimizu S, Shirato H. Prediction of respiratory tumour motion for real-time image-guided radiotherapy. Phys Med Biol. 2004;49(3):425–40.
63. Solberg TD, Medin PM, Ramirez E, Ding C, Foster RD, Yordy J. Commissioning and initial stereotactic ablative radiothereapy experience with Vero. J Appl Clin Med Phys. 2014;15(2):205–25.
64. D'Souza WD, Naqvi SA, Uu CX. Real-time intra-fraction motion tracking using the treatment couch: a feasibility study. Phys Med Biol. 2005;50:4021–33.
65. D'Souza WD, McAvoy TJ. An analysis of the treatment couch and control system dynamics for respiration-induced motion compensation. Med Phys. 2006;33(12):4701–9.
66. Specht DF. Probabilistic neural networks. Neural Netw. 1990;3:109–18.
67. Sun WZ. Jiang;MY, Ren L, dang J, you T, and Yin F-F, respiratory signal prediction based on adaptive boosting and multi-layer perceptron neural network. Phys Med Biol. 2018;62(17):6822–35.
68. Tacke MB, Nill S, Krauss A, Oelfke U. Real-time tumor tracking: automatic compensation of target motion using the Siemens 160 MLC. Med Phys. 2010;37:753–61.
69. Tobin MJ, Mador MJ, Guenther SM, Lodato RF, Sackner MA. Variability of resting respiratory drive and timing in healthy subject. J Appl Physiol. 1988;65:309–17.
70. Tsunashima Y, Sakae T, Shioyama Y, et al. Correlation between the respiratory waveform measured using a respiratory sensor and 3D tumor motion in gated radiotherapy. Int J Radiation Oncology Biol Phys. 2004;60(3):951–8.
71. Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.
72. Wang R, Liang X, Zhu X, et al. Feasibility of respiration prediction based on deep Bi-LSTM for real-time tumor tracking. IEEE Access; 2018.
73. Wiersma RD, McCabe BP, Belcher AH, Jensen PJ, Smith B, Aydogan B. Technical note: high temporal resolution characterization of gating response time. Med Phys. 2016;43(6):2802–6.

74. Williams RJ. Training recurrent networks using the extended Kalman filter. Int Joint Conf on Neural Networks. 1992;4:241–6.
75. Wu H, et al. Gating based on internal/external signals with dynamic correlation updates. Phys Med Biol. 2008;53(24):7137–50.
76. Yan H, Yin F-F, Zhu G-P, Ajlouni M, Kim JH. Adaptive prediction of internal target motion using external marker motion: a technical study. Phys Med Biol. 2006;51(1):31–44.
77. Zhang T, et al. Application of the spirometer in respiratory gated radiotherapy. Med Phys. 2003;30(12):3165–71.

Part IV

Machine Learning for Outcomes Modeling and Decision Support

Prediction of Oncology Treatment Outcomes

15

Sunan Cui and Issam El Naqa

15.1 Introduction

Outcome modeling plays an important role in oncology and treatment personalization. This includes understanding response to different therapeutic cancer agents (chemo, radiation, check point blockade, etc.), treatment adaptation, and designing of future clinical trials, which will be the subject of Chapters 18 and 19, respectively. Historically, the application of outcome models has accompanied oncology practices since its inception; however, 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 [1].

The notion of outcome modeling is motivated by the clinical need to personalize treatment to individual patient's cases. This concept originated from *Hippocrates of Kos*, the father of western medicine, 2500 years ago, who wrote: "different [drugs] to different patients, for the sweet ones do not benefit everyone, nor do the astringent ones, nor are all the patients able to drink the same things [2]." However, at the time of Hippocrates, physical and clinical exams were the only resources of information. This has drastically changed due to recent advances in quantitative multimodality imaging (e.g., radiomics) and high-throughput biotechnology (genomics, proteomics, transcriptomics, metabolomics, etc.), which are the subject of Chap. 16. Together these information would form the *Big data* or the *panOmics* of oncology as depicted in Fig. 15.1 [3].

Radiation oncology, as a cancer treatment modality, has been historically at the forefront of modeling responses to therapy; however, recent clinical trials

S. Cui (✉) · I. El Naqa
Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: sunan@umich.edu; ielnaqa@med.umich.edu

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_15

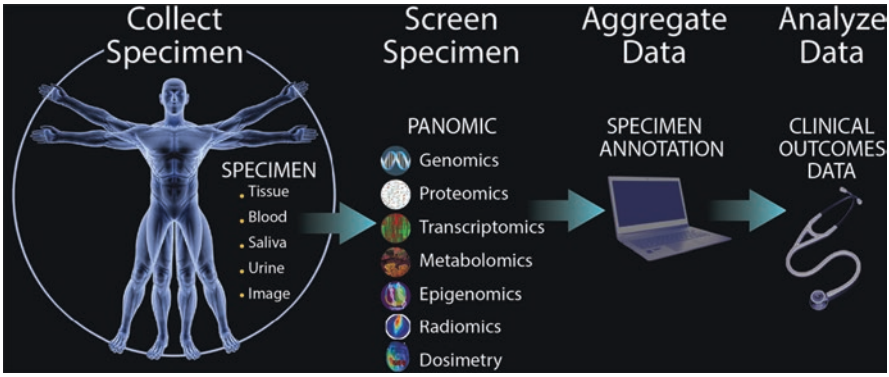


Fig. 15.1 The human body is a valuable resource for varying solid and fluid types of specimens, which can yield different -omics (genomics, proteomics, transcriptomics, metabolomics, radiomics) predictive biomarkers, in addition to dosimetric and clinical factors used in radiotherapy that would undergo major processes of annotation, curation, and preparation before being applied into outcome modeling of treatment outcomes (e.g., tumor response, side effects) [3]

examining treatment intensification in patients with locally advanced cancer have shown incremental improvements in local control and overall survival with several controversial results at instances [4] [5], in any case, radiation-induced toxicities remain major dose-limiting factors and likely culprit in such controversies [6–8]. Therefore, there is a need for studies directed toward predicting treatment benefit versus risk of failure. This is in addition to understanding how combining radiation with chemotherapy, surgery or most recently with immunotherapy can lead to improved outcome compared to either modality alone [9].

Clinically, the identification of such treatment predictors would allow for more individualization of cancer treatment plans. In other words, physicians may prescribe a more or less intense single or multimodality regimen(s) for an individual based on model predictions of local control benefit and toxicity risk. Such an individualized regimen 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 local control and side effects should be considered in the optimal treatment management design process.

In this chapter, we consider the subject of outcome modeling in radiation oncology as case study while highlighting similarities and differences when applied to other treatment modalities. We will provide an overview of the current status of data-driven outcome modeling techniques for predicting tumor response and side effects for patients who receive cancer treatment with special focus on the emerging role of machine and deep learning approaches to improve outcome modeling and response prediction. Then, we present examples of oncology data resources and its big data (panOmics) notion. Finally, we discuss the potentials and some of the challenging obstacles to applying bioinformatics and machine/deep learning strategies

to oncology outcome modeling. Interested reader in detailed treatment of this subject is recommended to check the dedicated textbook on this topic [1].

15.2 Outcome Modeling in Radiotherapy

Radiotherapy outcomes are usually characterized by two metrics: the tumor control probability (TCP) [10, 11] and the normal tissue complication probability (NTCP) [12] of surrounding normal tissues. TCP/NTCP models could be used during the consultation period as a guide for ranking treatment options [7, 8]. Alternatively, once a decision has been reached, these models could be included in an objective function, and the optimization problem driving the actual patient's treatment plan can be formulated in terms relevant to maximizing tumor eradication benefit and minimizing complication risk [9, 13, 14]. Traditional models of TCP/NTCP models and their variations use information only about the dose distribution and fractionation. However, it is well known that radiotherapy outcomes may also be affected by multiple clinical and biological prognostic factors such as stage, volume, tumor hypoxia [15, 16], etc. as depicted in Fig. 15.2. Therefore, recent years have witnessed the emergence of data-driven models utilizing informatics techniques, in

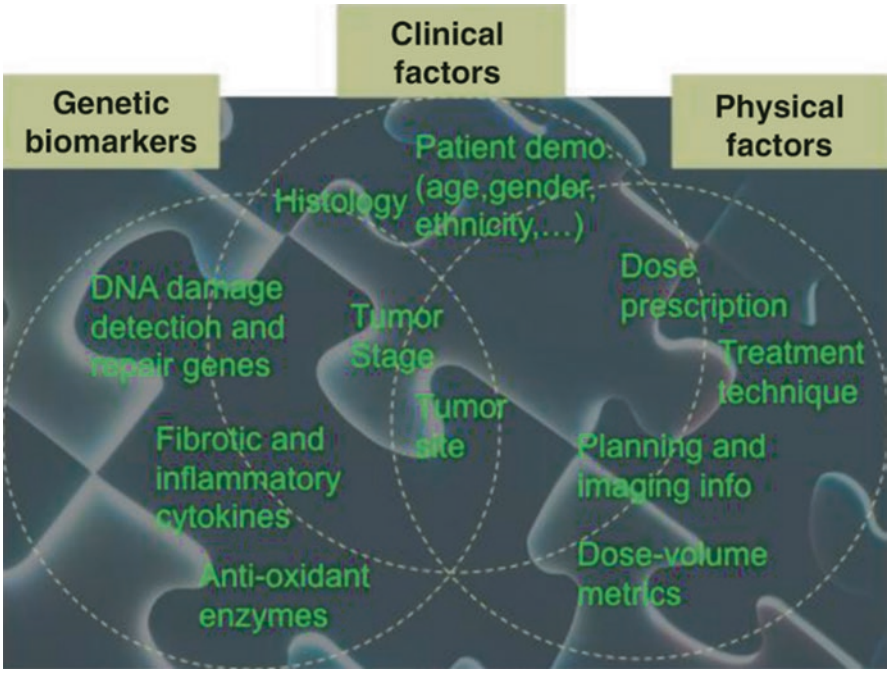


Fig. 15.2 Radiotherapy treatment involves complex interaction of physical, biological, and clinical factors. The successful informatics approach should be able to resolve this interaction "puzzle" in the observed treatment outcome (e.g., local control or toxicity) for each individual patient [24]

which dose-volume metrics are combined with other patient- or disease-based prognostic factors [17–23].

15.3 Data Resources

In oncology, there is a large pool of "big data" that comprise but are not limited to patient demographics, treatment prescription, 3D and 4D anatomical and functional disease longitudinal imaging features (radiomics), in addition to genomics and proteomics data derived from peripheral blood and tissue specimens (Fig. 15.1). In the case of radiotherapy, there is also an option of volumetric dosimetric data about radiation exposure to the tumor and surrounding tissues. Accordingly, this big data in oncology could be divided based on its nature into four categories: clinical, dosimetric, imaging, and biological. These four categories of radiotherapy big data are described in the following.

15.3.1 Clinical Data

Clinical data in oncology and particularly in chemoradiotherapy typically refers to cancer diagnostic information (e.g., site, histology, stage, grade, etc.), patient-related characteristics (e.g., age, gender, co-morbidities), and physiological metrics (e.g., pulmonary function measurements, heart/pulse rates, blood cell counts, body mass index (BMI)). Prior to the era of genetic profiling, these clinical variables were considered the only gold standard for clinical management and decision-making in oncology. From an informatics perspective, the mining of such data could be challenging particularly if the data is unstructured as typically the case; however, there are good opportunities for applying natural language processing (NLP) techniques to assist in the organization of such data [25].

15.3.2 Dosimetric Data

This type of data is related to the treatment planning process in radiotherapy or the chemical agent in chemotherapy. In radiotherapy, which involves radiation dose simulation using computed tomography imaging, specifically dose-volume metrics derived from dose-volume histograms (DVHs) graphs. Dose-volume metrics have been extensively studied [17–20] in the radiation oncology literature for outcome modeling [26, 27]. These metrics are extracted from the DVH such as volume receiving certain dose (Vx); minimum dose to x% volume (Dx); mean, maximum, and minimum dose; etc. More details are in our review chapter [28]. Moreover, we have developed a dedicated software tool called "Dose response explorer" (DREES) [29] for deriving these metrics and modeling of radiotherapy response.

There are different categories of chemical agents that aim on eradicating tumor cancers [30]. Among the most common ones are alkylating agents, which substitute

an alkyl groups (hydrocarbon) for hydrogen atom of organic compound including DNA (e.g., Temozolomide). There are also antibiotics (e.g., Doxorubicin, Blemoycin). Another common one is antimetabolites (e.g., Methotrexate, 5-Fluorouracil, Taxanes, vinca alkaloids). Other agents that do not fall into any of these classes include: Platinum compounds (Cisplatin) and topoisomerases (DNA winding enzymes) inhibitors. Recently, more signaling pathway targeted agents have been developed such anti-EGFR such as Cetuximab or Erbitux [31]. In addition to the agent type and dosage, the timing of the administration of the agent influences treatment response. Chemotherapy could be administrated after the completion of the local treatment such as radiation and is called adjuvant chemotherapy, before the local treatment and is called induction chemotherapy, or given during local treatment and is called concurrent chemotherapy. In particular, concurrent chemoradiation has been demonstrated to be effective in the treatment of several cancers, in which the chemotherapy agent can act as a radiosensitizer by aiding the destruction of radiation resistant clones or act systematically and potentially eradicate distant metastases [32].

15.3.3 Radiomics (Imaging Features)

Cancer patients are treated based on observational assessment from diagnostic imaging particularly computed tomography (CT) in combination with other clinical factors [33]. Information from multiple imaging modalities could be used to improve treatment monitoring and prognosis in different cancer types. For example, physiological information (tumor metabolism, proliferation, necrosis, hypoxic regions, etc.) can be collected directly from nuclear imaging modalities such as single-photon emission computed tomography (SPECT) and positron emission tomography (PET) or indirectly from magnetic resonance imaging (MRI) [34, 35]. The complementary nature of these different imaging modalities has led to efforts toward combining information to achieve better treatment outcomes. For instance, PET/CT has been utilized for staging, planning, and assessment of response to chemoradiation therapy [36, 37]. Similarly, MRI has been applied in tumor delineation and assessing toxicities in head and neck cancers [38, 39]. Moreover, quantitative information from hybrid-imaging modalities could be related to biological and clinical endpoints, a new emerging field referred to as “radiomics” [40, 41]. Potential of this new field to monitor and predict response to chemoradiotherapy has been demonstrated in esophageal [42], head and neck [43, 44], cervix [43, 45], lung [46] [47], and sarcoma [48] cancers, and more recently in the prediction of immunotherapy response [49, 50], in turn allowing for adapting and individualizing treatment [51].

15.3.4 Biological Markers

A biomarker is defined as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or

pharmacological responses to a therapeutic intervention" [52]. Biomarkers can be categorized based on the biochemical source of the marker into exogenous or endogenous.

Exogenous biomarkers are based on introducing a foreign substance into the patient's body such as those used in molecular imaging as discussed above. Conversely, endogenous biomarkers can further be classified as (1) "expression biomarkers," measuring changes in gene expression or protein levels or (2) "genetic biomarkers," based on variations, for tumors or normal tissues, in the underlying DNA genetic code. Measurements are typically based on tissue or fluid specimens, which are analyzed using molecular biology laboratory techniques [53]. Aggregation of large-scale genetic biomarkers has been the subject of large national efforts such as The Cancer Genome Atlas (TCGA) Data Portal, which provides a very useful platform for researchers to analyze datasets generated by TCGA. It contains clinical information, genomic characterization data, and high-level sequence analysis of the tumor genomes in different cancer types [54–57].

15.4 Database Technologies for Machine Learning in Oncology

Traditionally, relational database management systems (RDBMS) have been the technology of choice for storing and querying oncology information. RDBMS are based on organizing the data relation schema in a tabular format (sets of rows (tuples) and columns (attributes)) in accordance with Codd's 12 rules [58]. SQL (structured query language) is a fourth generational programming language that is used to process the data in an RDBMS. RDBMS and SQL have been the driving technology for Electronic Health Record (EHR) management software including that of oncology. Several governmental, commercial, and open-source resources for EHR exist. For instance, in the United States, more than 50% of patient records are stored in the Epic systems (Verona, WI), privately held software, which employs an object-oriented RDBMS. In addition, there are open-source EHR systems; however, they did not receive traction.

Recently, there has been resurgence in NoSQL (not only SQL) database technologies. NoSQL allows for a blend of structured and unstructured data with no commitment to a schema unless needed and enjoys a remarkable horizontal scalability for aggregating and querying massive datasets. Interestingly, the VistA EHR system developed by department of Veterans affairs in the 1960s is based on the MUMPS (Massachusetts General Hospital Utility Multi-Programming System), which is a key-value NoSQL database system. Today, the NoSQL open-source Hadoop architecture is considered the platform of choice for processing big data and potentially oncology data. The enabling technology that sprung its big data analytics potential is called MapReduce, which is a new parallel programming paradigm that involves two steps: a Map function for filtering and sorting and a Reduce function for grouping and aggregation of data. However, an issue that may impact NoSQL adoption in some instances is that the common so-called ACID (Atomicity,

Consistency, Isolation, Durability) properties of a reliable transactional processing may need to be compromised to achieve higher analytical performance. As a compromise, this created market for NewSQL that rely on storing large data in memory, which is advocated by M. Stonebraker (VoltDB, Inc., Bedford, MA).

15.5 Pan- Vs. P-OMICs

Due to advances imaging and biotechnology radiotherapy data has witnessed tremendous exponential growth in the past decade; however, number of cancer incidences has generally plateaued as depicted in Fig. 15.3. The fact that p (variables) $>> n$ (samples) constitute a serious challenge for class inference methods of statistical learning. This p-omics phenomenon may yield undesirable effects such as spurious correlations, echo chamber anomalies, Yule–Simpson reversal paradox, or misleading ghost analytics as discussed in the following.

15.5.1 Spurious Relationship

This pitfall commonly emerges in big data analysis when two variables have no true relationship; however, such one may be wrongly inferred due to confounding effects [59]. This is an important process when attempting to identify a biomarker of

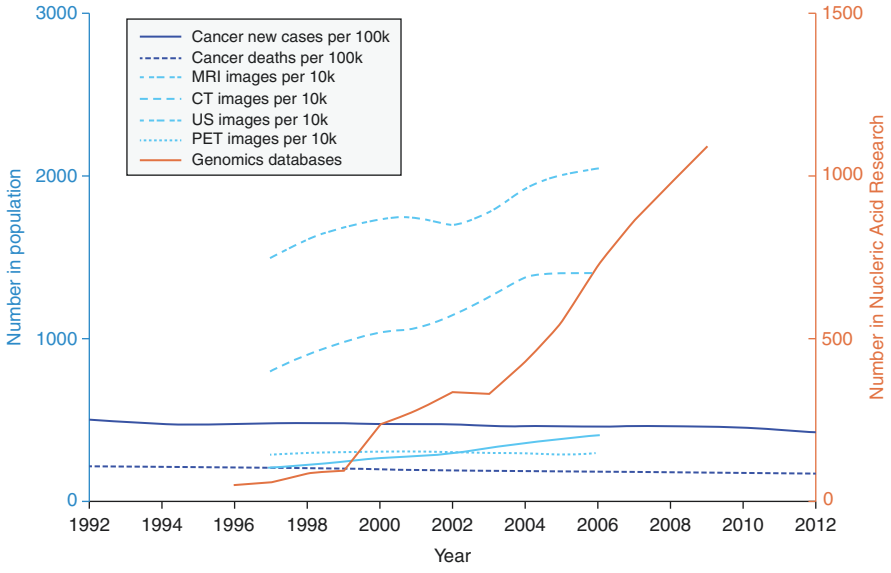


Fig. 15.3 The p-omics vs. pan-omics in cancer outcome modeling problem, where the number of variables grows rapidly in imaging and genetics while the number of samples has largely plateaued. The data is compiled querying cancer information from the SEER 2015 statistics, imaging information from Health Affairs 2008, and genomics from the Nucleic Acids Research archive

chemoradiation response, for instance. The famous example of such a case is the association of the number of ice cream sold and increased risk of drowning; the confounding effect or lurking explanatory variable is simply warm seasons. Understanding of the problem setup and possible prior knowledge of potential confounding effects is helpful in mitigating such effect.

15.5.2 Echo Chamber Effect

This happens when a relationship in the data is magnified by the data aggregation process itself in a cyclic manner [60]. This is typically a sampling problem (selection bias), in which the analyzed sample is not representative of the intended population. A common example is encountered in meta-analyses of previous oncology biomarker study findings, where negative results are typically less likely to be published [61].

15.5.3 Yule–Simpson Paradox

This is reverse effect of the echo chamber, where a true association is found in small datasets but lost or even reversed when larger data is aggregated. This paradox was reported by Simpson in the analysis of contingency tables in interpreting second-order interactions [62]. A lauded example in cancer research of this paradox is noted in the backlash generated by the paper by Tomasetti and Vogelstein that implied that variations in cancer risk are mainly explained by the number of stem cell divisions [63], a mere "bad luck" issue irrespective of the environment. This has been debated vigorously demonstrating selection bias and Yule–Simpson effects in the performed data analysis [64]. A common pitfall that could result in this paradox can happen when not accounting for known patient characteristics such as age and gender when conducting population studies. For instance, it is known that there exists a negative relationship between medicine dosages and recovery in both males and females; however, when the two group are combined together, a surprising positive relationship emerges as shown in Fig. 15.4 [65]. A possible remedy for this effect is using stratification by variables or more systematically performing unsupervised clustering to uncover such sub-population effects.

15.5.4 Ghost Analytics

This refers to erroneous (mis-) using of statistical tests or learning algorithms when analyzing large datasets. For instance, this problem arises when not accounting for assumptions embedded in a statistical test before applying it. A classical example is encountered when conducting multiple comparisons and reporting a "significant" p-value of the null hypothesis testing yielding misleading results by not adjusting the level of Type 1 error. Interestingly, when statistician R. Fisher introduced the

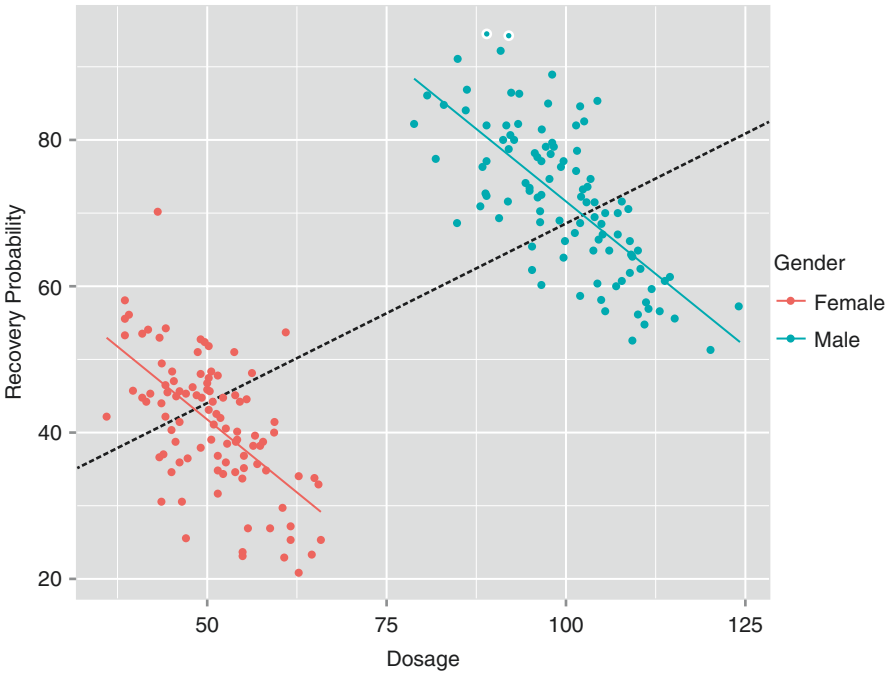


Fig. 15.4 Simpson paradox illustration where population and subgroups give contradictory results when analyzing the association between medical dosage and recovery in males and females (from [65])

notion of p-values in the 1920s, he did not intend to have it as a definitive test rather simply as an informal way to judge whether association evidence was worthy of a second look. Therefore, it is necessary to understand the assumptions made in a statistical test before attempting to apply it in order to achieve meaningful results.

15.6 Modeling Methods

Modeling techniques in oncology in general and radiation oncology in particular could be divided generally into bottom-up and top-down approaches as depicted in Fig. 15.5. The focus of this review will be on top-down approaches, while bottom-up methods are described for completeness and as a way to constrain the search space when conducting data mining exercises.

15.6.1 Bottom-up Approaches for Modeling Oncology Response

These approaches utilize first principles of physics, chemistry, and biology to model cellular damage temporally and spatially in response to treatment. Typically, they would apply advanced numerical methods such as Monte-Carlo (MC) techniques to

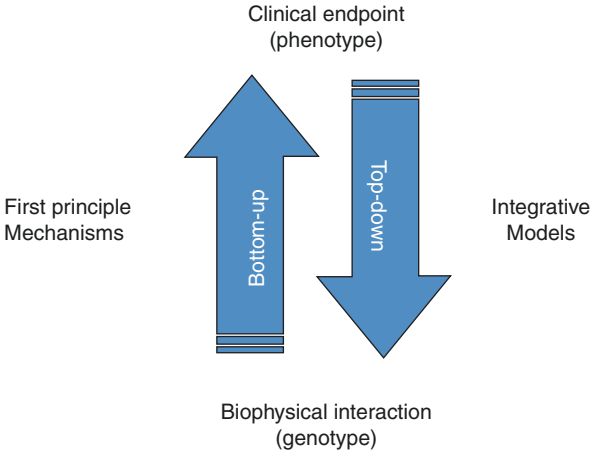


Fig. 15.5 Outcomes modeling schemes in oncology could be divided into: *top-down* (starting from the observed clinical outcome and attempting to identify the relevant variables that could explain the phenomena) or *bottom-up* (starting from basic principles to the observed clinical outcome in a multi-scale fashion)

estimate the molecular spectrum of damage in clustered and not-clustered DNA lesions ($\mathrm{Gbp}^{-1}\,\mathrm{Gy}^{-1}$) [66]. For instance in the case of radiotherapy, the temporal and spatial evolution of the effects from ionizing radiation can be divided into three phases: physical, chemical, and biological in a multi-scale fashion [67]. This information, however, could be used to guide incorporating prior knowledge or imposing constraints on a data mining approach narrowing its search space for optimal answers.

15.6.2 Top-Down Approaches for Modeling Oncology Response

These are typically phenomenological (non-mechanistic) models and depend on parameters available from the collected clinical, dosimetric and/or biological data [3]. In the context of data-driven and multi-variable modeling of outcomes, the observed treatment outcome is considered as the result of functional mapping of several input variables [68]. Mathematically, this is expressed as $f(\boldsymbol{x};\boldsymbol{w}^*) : X \rightarrow Y$, where $\boldsymbol{x} \in \mathbb{R}^N$ is composed of the input metrics (patient disease-specific prognostic factors, dosimetric metrics, or biological markers). The expression $y \in Y$ is the corresponding observed treatment outcome. The variable $\boldsymbol{w}^*$ includes the optimal parameters of the model $f(\bullet)$ obtained by the learning based on a designated objective function. Learning is defined in this context of outcome modeling as estimating dependencies from data [27]. Based on the human–machine interaction, there is two common types of learning: supervised and unsupervised. Supervised learning is used when the endpoints of the treatments such as tumor control or toxicity grades are known; these endpoints are provided by experienced oncologists following institutional or National Cancer Institute (NCI) criteria and it is the most commonly used learning method in outcomes modeling. Nevertheless, unsupervised methods such as clustering methods or principal component analysis (PCA) can be used to reduce the learning problem dimensionality through feature extraction, and to aid in

the visualization of multivariate data as well as in the selection of the optimal learning method parameters for supervised learning methods [69].

It is noted that the selection of the functional form of the model $f(\bullet)$ is closely related to the prior knowledge of the problem. In mechanistic models, the shape of the functional form is selected based on the clinical or biological process at hand; however, in data-driven models, the objective is usually to find a functional form that best fits the data [70]. Below we will highlight this approach using logistic regression and artificial intelligence methods in cases where the clinical endpoints are expressed as a binary dichotomy (failed/did not fail) as commonly practiced. However, the methods could be extended in cases with more than two classes or the endpoint is a continuous variable.

15.6.2.1 Logistic Regression

In oncology outcomes modeling, the response will usually follow an S-shaped curve. This suggests that models with sigmoidal shapes are the most appropriate to use [21–24, 71–74]. A commonly used sigmoidal form is the logistic regression model, which has nice numerical stability properties. The logistic model is primarily used in a binary classification, i.e., response $y = 0$ or $y = 1$. The probability of y equaling 1 can be written as, $p\left(y=1\right)=f\left(\boldsymbol{x};\boldsymbol{w}\right)=\frac{1}{1+e^{-w^T x}}$, where w are model parameters, whose optimal value $\boldsymbol{w}^*$ can be determined through maximum likelihood estimation [75]. Explanatory variables $\boldsymbol{x}$ containing patient-specific information may be chosen in a stepwise fashion to define the abscissa of the regression model $f(\bullet)$. Except for original patient-specific variables, users may determine whether interaction terms or higher order variables should be added. Penalty techniques based on ridge (L2-norm) or Lasso (L1-norm) methods could aid in the process by eliminating least relevant variables and imposing sparsity conditions [76]. An alternative solution to ameliorate this problem is offered by applying machine learning model selection methods based on cross-validation [75]. An example is provided below.

A Logistic Outcome Modeling Example

The logistic model is applied to predict local progression (LP) y (progression: $y = 1$, control: $y = 0$) in non-small-cell lung cancer (NSCLC) patients in TCGA-LUAD [77] and TCGA-LUSC [78] datasets. Only the 45 patient who received external beam radiotherapy as adjuvant therapy, with primary tumor as treated sites, and had complete dosimetric/ LP follow-up information were included in the analysis. Explanatory variables $\boldsymbol{x}$: $\boldsymbol{x} \in \mathbb{R}^7$ in this analysis include patients' gender, T stage, N stage, grouping stage, prior malignancy, tobacco history, and radiation total dose. Examples illustrating explanatory variables and responses in three randomly selected patients are shown in Table 15.1. The logistic model was implemented with python package scikit-learn [79]. In the code, $X : X \in \mathbb{R}^{S\times N}$ stands for explanatory variable $\boldsymbol{x}$ for S patients, $Y : Y \in \mathbb{R}^S$ stands for local progression for S patients. A summary of LP prediction results by our logistic model is presented together with ground truth (Fig. 15.6).

Table 15.1 TCGA patients' specific information that were used in LP prediction

	gender	T	N	stage	prior M	tobacco	dose	LP
1	M	T3	N0	T3	No	4	60	0
2	M	T1	N1	T1	No	3	64	0
3	F	T1	N2	T1	No	2	50	0

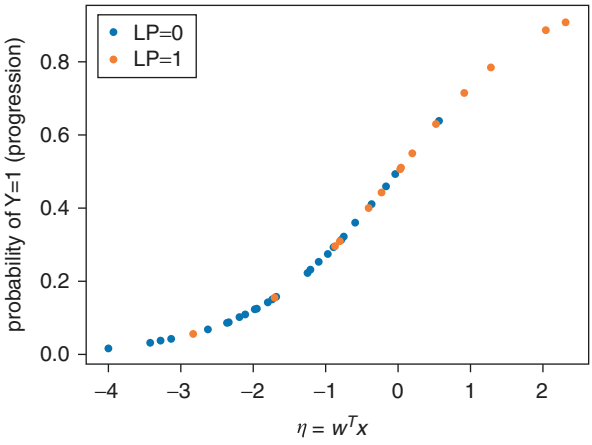


Fig. 15.6 LP Prediction results of 45 TCGA patients by logistic regression with ground truth shown

Code:

```
from sklearn.linear_model import LogisticRegression
# model building
clf = LogisticRegression(random_state=0, solver='lbfgs').fit(X, Y)
Pred=clf.predict(X)
Pred_prob=clf.predict_proba(X)
sc=clf.score(X, Y)
#estimated optimal parameters
w_est=clf.coef_.reshape(-1,1)
b_est=clf.intercept_
# the trained model
pred_fucntion=np.matmul(X,w_est)+b_est
```

15.6.2.2 Machine Learning Methods

Machine learning techniques are a class of artificial intelligence (e.g., neural networks, decision trees, support vector machines), which are able to emulate human intelligence by learning the surrounding environment from the given input data and can detect nonlinear complex patterns in such data. In particular, neural networks were extensively investigated to model post-radiation treatment outcomes for cases of lung injury [80, 81] and biochemical failure and rectal bleeding in prostate cancer [82, 83]. A rather more robust approach of machine learning methods is support vector machines (SVMs), which are universal constructive learning procedures based on the statistical learning theory [84]. For discrimination between patients who are at low risk versus patients who are at high risk of treatment, the main idea of SVM would be to separate these two classes with "hyper-planes" that maximize the margin between them in the nonlinear feature space defined by an implicit

kernel mapping [10, 11, 70]. However, these methods have been stigmatized as black boxes, hindering their application in practical clinical contexts.

In an effort, to alleviate the black box stigma of generic machine learning methods, more system-like approaches methods based on graphical approaches such as Bayesian networks (BNs) have been increasingly used in outcome modeling of cancer [85–87]. A BN provides graphical representation of the relationships between the variables represented as nodes in a directed acyclic graph (DAG), which encodes the presence and direction of relationship influence among the variables themselves and the clinical endpoint of interest. The relationship between parent and child nodes is modeled by conditional probabilities using Bayes chain rule. These methods are also robust for variable uncertainties and missing data, which would make them excellent candidates for clinical applications [88, 89].

For machine learning methods, a modeling example is provided below.

A Machine Learning Outcome Modeling Example

A multi-layer perceptron (MLP) is a fundamental neural network architecture and is the building block for deep learning applications [90]. It is composed of several layers, each with a number of nodes. Those nodes are interconnected in a feed-forward way that nodes have direct connection to the nodes in the subsequent layers. To calculate the nodes' value, a so-called activation function [91] is usually applied to a weighted sum of nodes in the previous layer to add non-linearity. Logistic function which is also known as sigmoid function is a common choice of activation. Specifically, logistic regression can be regarded an extreme case of MLP without any hidden layer.

An MLP with 2 hidden layer is applied to the same task of predicting LP in TCGA dataset. The dataset is randomly split into a training set ($\frac{1}{3}$) and a test set ($\frac{2}{3}$). All the explanatory variables were normalized to mean 0 and standard variation 1 to avoid numerical instability. The model was implemented with python package Pytorch [92]. Receiver operating characteristic (ROC) curves of LP prediction results are illustrated with calculated areas under ROC curve (AUC) for both training and test sets (Fig. 15.7).

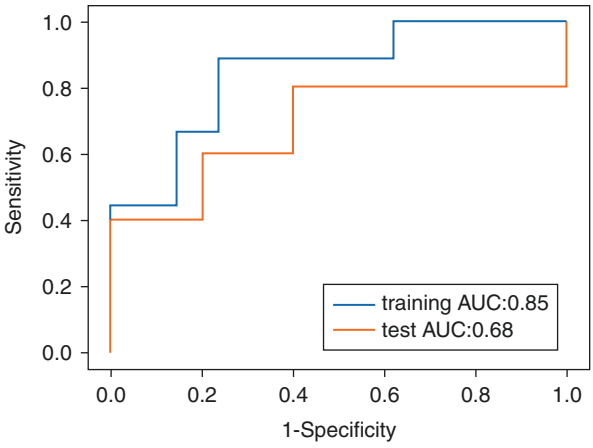


Fig. 15.7 ROC curves of LP Prediction results by MLP in 45 TCGA patients

Code:

```
import torch
#define the model
class MLP_NN(torch.nn.Module):
     def __init__(self, input_dim, hidden_dim_1,hidden_dim_2,
     output_dim):
          super(MLP_NN, self).__init__()
          self.L1=torch.nn.Linear(input_dim,hidden_dim_1)
          self.D1=torch.nn.Dropout(0.2)
          self.L2=torch.nn.Linear(hidden_dim_1,hidden_dim_2)
          self.D2=torch.nn.Dropout(0.2)
          self.L3=torch.nn.Linear(hidden_dim_2,output_dim)
     def forward(self,x):
          a1=torch.relu(self.L1(x))
          a1=self.D1(a1)
          a2=torch.relu(self.L2(a1))
          a2=self.D2(a2)
          outputs=torch.sigmoid(self.L3(a2))
          return outputs
#initialize the model
model=MLP_NN(input_dim,hidden_dim,hidden_dim,output_dim)
#use binary cross entropy as loss
criterion=torch.nn.BCELoss(reduction='mean')
#use Adam optimizer
optimizer=torch.optim.Adam(model.parameters(), lr=learning_rate)
for i in range(num_epoch):
     #training
     model.train()
     optimizer.zero_grad()
     y_pred=model(X_train_torch.float())
     #define the loss
     loss=criterion(y_pred,Y_train_torch.float())
     loss.backward()
     optimizer.step()
     # evaluate on test data
     model.eval()
     y_pred_test=model(X_test_torch.float())
loss_test=criterion(y_pred_test,Y_test_torch.float())
```

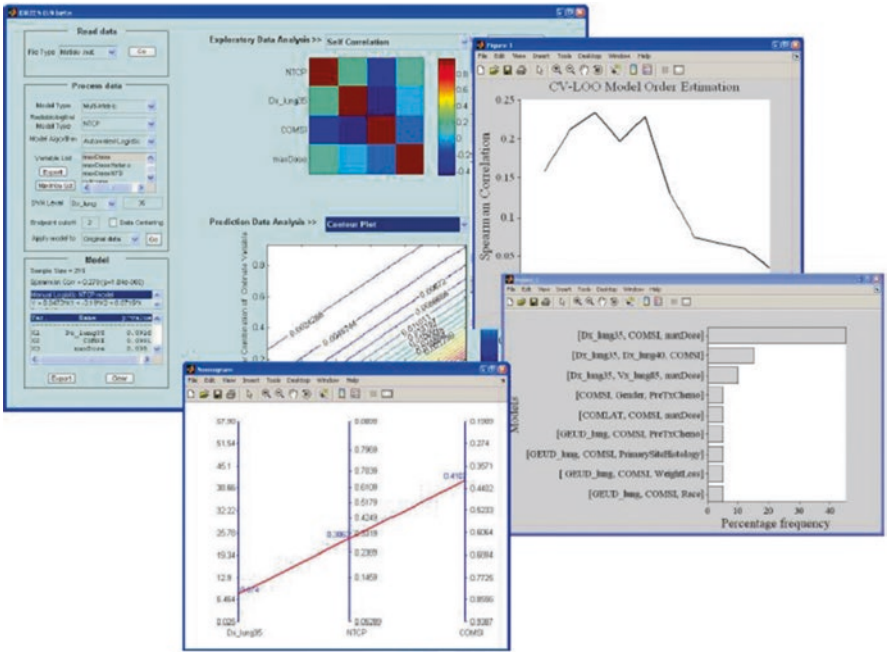


Fig. 15.8 DREES allows for TCP/NTCP analytical and multivariate modeling of outcomes data. The example is for lung injury. The components shown here are Main GUI, model order, and parameter selection by resampling methods, and a nomogram of outcome as function of mean dose and location

15.7 Software Tools for Outcome Modeling

Many of the TCP/NTCP outcome modeling methods require dedicated software tools for implementation. Examples of such software tools in the literature in the case of radiotherapy are BIOPLAN and DREES. BIOPLAN (BIOlogical evaluation of treatment PLANs) uses several analytical models for evaluation of radiotherapy treatment plans [93], while DREES is an open-source software package developed by our group for dose-response modeling using analytical and data-driven methods [94] presented in Fig. 15.8. It should be mentioned that several commercial treatment planning systems have currently incorporated different TCP/NTCP models, mainly analytical ones that could be used for ranking and biological optimization purposes. A discussion of these models and their quality assurance guidelines is provided in TG-166 [18]. In the general context of machine and deep learning applications, several open-source packages like Pytorch can be used as presented in the examples of Sect. 15.5.2. Further discussion of available software tools is provided in Chap. 7.

15.8 Discussion

In the era of personalized medicine in oncology, an multimodality and multidisciplinary approach that provides a unique combination of clinical, physical, technological, and biological data could be evaluated as an ideal case study for employing big data analytics to improve treatment effectiveness and outcomes in medicine. Oncology data is comprised of clinical patient characteristics, varying imaging acquisitions, laboratory and biochemical measurements, etc. carrying all the hallmarks of big data. It is believed that big data analytics hold great promise to improve safe treatment management and enable development of better clinical decision support systems for personalized medicine as lauded by the NIH Personalized Medicine Initiative (PMI). Furthermore, big data analytics has been highlighted in the American Society of Clinical Oncology (ASCO) progress report as one of the promising opportunity in the fight against cancer as envisioned in the development of its data aggregation portal known as CancerLinQ [95]. The same sentiment has been echoed in Radiation Oncology [96].

The path for data collection and aggregation in oncology has been traditionally implemented to develop a hypothesis based on a clinical or experimental observation then test this hypothesis in a controlled clinical trial institutionally, then multi-institutionally if it deemed promising. This path generally can account for about 3% of all patients' data which are generated and stored during regular clinical processes, with 97% of the data, termed "dark data" being not collected. However, the dark data are generally unstructured, untrusted, and fails to be useful for improving research, quality assessment, or clinical care. It is this "invisible data" that oncology Big Data initiatives such as CancerLinQ aim to bring to light. However, to make such data visible would require both cultural changes that would respect standardized lexicons and proper curation of this data on a routine basis. This would necessitate procedures that facilitate the data aggregation process, and local and national data champions within the oncology community. Moreover, making this data more visible would also need collaboration between all stakeholders to develop infrastructures and rigorous procedures to maintain its security and eliminate lingering patient privacy concerns.

New database technologies such as NoSQL or NewSQL whether terrestrial or in the cloud will yield better storage and query of oncology data while allowing application of more advanced analytics in real time as part of a clinical quality assurance or improvement program. The MapReduce framework allows embedding of machine learning algorithms as part of its architecture. This technology would work well with parallelizable algorithms. However, many oncology modeling schemes particularly ones that involve iterative or gradient descent optimization techniques do not lend themselves naturally to this framework. This would require further investigation to overcome this limitation and to exploit such technologies for oncology real-time analytics.

One of main challenge to big data analytics in oncology remains the inherent p-omics versus pan-omics problem. In the presented examples using primarily typical applications in oncology, we demonstrated different methods to mitigate this effect

such as using prior knowledge, information theory techniques, ensemble of machine learning, or different combinations of all these methods. Issues related to echo chamber or Yule–Simpson paradox need also to be carefully tested in the context of big data in oncology. However, the role of big data and its challenges is expected to grow as more current dark data being brought into light with many missing or poorly curated information and the pool of applications is ever expanding. This problem is further exacerbated when dealing with multiple clinical endpoints each may lead to different relationships with the input data. Moreover, despite decades of research many issues in dealing with multiple clinical or biological endpoints remain open [97]. The typical practice in oncology has been to optimize each point independently or to use heuristics to combine multiple endpoints in utility functions in order to account for competing risk effects and quantifying their subjective desirability [98]. Alternatively, such utilities could be presented as a multi-output system that would jointly optimize the prediction of the competing endpoints, of course, on the expense of increased sample size requirements posing further challenges to big data in oncology. Therefore, it is paramount to develop oncology-specific approaches that exploit bottom-up biological knowledge in cancer combined with advanced information theoretic and machine/deep learning methodologies to develop hybrid models that can mitigate current challenges of noisy analytic pitfalls and achieve the big data promise in cancer research.

15.9 Future Research Directions

The ability to maintain high-fidelity large-scale data for oncology studies remains a major challenge despite the high volume of clinical generated data on almost daily basis. It is paramount to implement FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for scientific data management to achieve the desired goals of precise outcome modeling, especially with machine/deep learning approaches [99]. In the case of radiation oncology, for instance, there have been several ongoing institutional and multi-institutional initiatives such as the RTOG, radiogenomics consortium, and EuroCAT to develop such infrastructure; however, there is plenty of work to be done to overcome issues related to, data sharing hurdles, patient confidentiality issues, lack of signaling pathways databases of radiation response, development of cost-effective multicenter communication systems that allows transmission, storage, and query of large datasets such as images, dosimetry, and biomarkers information. The use of NLP techniques is a promising approach in organizing unstructured clinical data. Dosimetry and imaging data can benefit from existing infrastructure for Picture Archiving and Communication Systems (PACS) or other medical image databases. Methods based on the new emerging field of systems radiobiology will continue to grow on a rapid pace, but they could also benefit immensely from the development of specialized radiation response signaling pathway databases analogous to the currently existing pharmacogenomics databases. Data sharing among different institutions is a major hurdle, which could be solved through cooperative groups or distributed databases by developing in a

cost-effective manner the necessary bioinformatics and communication infrastructure using open-access resources through partnership with industry.

Data quality is as important as data volume in big data analytics. Some important aspects of data quality in healthcare may include "completeness" which requires all data elements are present, "accuracy" which requires data are of the original source, "reliability" which requires data remain consistent, "legibility" which requires data (whether written, transcribed, or printed) should be readable, and "timeliness" which requires clinical information should be documented as event occurs or treatment is performed without much delay. Especially, efforts should be made to avoid biases in the data including age, gender and race, as these biases could potentially lead to AI algorithms that aggravate health care disparity, i.e., AI models resulting in misleading decision support for minority groups which are underrepresented. To ensure data quality, standards for healthcare documentation can be developed and implemented in radiation oncology. Routine monitoring can also be carried out to aid data quality improvement. In this process, quantitative metrics such as Shapley Value [100] should be developed and used as a guide to make evaluation and detect any abnormality that may occur.

To make data-driven outcome models real clinical tools, efforts should be made to improve interpretability of machine learning algorithms. Interpretability is particularly important as it can help act as fail-safe against scenario where algorithms may produce flawed results due to unforeseen bugs. Existing machine learning algorithms, specifically deep learning algorithms are known to suffer from a tradeoff between accuracy and interpretability. Hence, more work regarding interpreting and explaining machine learning algorithms' decisions [101] is expected. Specifically, human-in-to-loop (HITL) [102] concept which can guide to optimize 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 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 radiation oncology clinic.

15.10 Conclusion

Recent evolution in medical imaging and biotechnology has generated enormous amount of big data that spans clinical, dosimetric, imaging, and biological markers. This data provided new opportunities for reshaping our understanding of treatment response and outcome modeling. However, the complexity of this data and the variability of tumor and normal tissue responses would render the utilization of advanced bioinformatics and machine learning methods as indispensable tools for better delineation of treatment complex interaction mechanisms and basically a cornerstone to "making data dreams come true" [103]. However, it also posed new challenges for data aggregation, sharing, confidentiality, and quality. Moreover, oncology data and especially radiotherapy constitutes a unique interface between

physics and biology that can benefit from the general advances in biomedical informatics research such as systems biology and available web resources while still requiring the development of its own technologies to address specific issues related to this interface. Successful application and development of advanced data communication and bioinformatics tools for oncology big data so to speak is essential to better predicting treatment response to accompany other aforementioned technologies and usher significant progress toward the goal of personalized treatment planning and improving the quality of life for cancer patients. In the meantime, AI algorithms which balance the accuracy and interpretability are expected to be developed and serve as viable and reliable decision support tools in radiation oncology clinic.

References

1. El Naqa I. A guide to outcome modeling in radiotherapy and oncology: listening to the data. Boca Raton, FL: CRC Press. Taylor & Francis Group; 2018.
2. Sykiotis GP, Kalliolias GD, Papavassiliou AG. Pharmacogenetic principles in the Hippocratic writings. J Clin Pharmacol. 2005;45(11):1218–20.
3. El Naqa I, et al. Radiogenomics and radiotherapy response modeling. Phys Med Biol. 2017;62(16):R179–206.
4. Halperin EC, Brady LW. Perez and Brady's principles and practice of radiation oncology. 5th ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams&Wilkins; 2008.
5. Bradley JD, 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. Lancet Oncol. 2015;16(2):187–99.
6. Bentzen SM, et al. Quantitative analyses of Normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. Int J Radiat Oncol Biol Phys. 2010;76(3 Suppl):S3–9.
7. 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):S155–60.
8. Bradley JD, 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. Int J Radiat Oncol Biol Phys. 2017;99(2):S105.
9. Marciscano AE, et al. Immunomodulatory effects of stereotactic body radiation therapy: preclinical insights and clinical opportunities. Int J Radiat Oncol Biol Phys. 2019;12:360–1.
10. El Naqa I. Machine learning methods for predicting tumor response in lung cancer. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;2(2):173–81.
11. El Naqa I, et al. Datamining approaches for modeling tumor control probability. Acta Oncol. 2010;49(8):1363–73.
12. Deasy JO, El Naqa I. Image-based modeling of normal tissue complication probability for radiation therapy. Cancer Treat Res. 2008;139:215–56.
13. GG S. Basic clinical radiobiology. 3rd ed/. 2002, London/New York.
14. Armstrong K, et al. Individualized survival curves improve satisfaction with cancer risk management decisions in women with BRCA1/2 mutations. J Clin Oncol. 2005;23(36):9319–28.
15. Weinstein MC, et al. Modeling for health care and other policy decisions: uses, roles, and validity. Value in Health. 2001;4(5):348–61.
16. Moiseenko VK, 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.

17. Brahme A. Optimized radiation therapy based on radiobiological objectives. Semin Radiat Oncol. 1999;9(1):35–47.
18. 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):1386–409.
19. Choi N, et al. Predictive factors in radiotherapy for non-small cell lung cancer: present status. Lung Cancer. 2001;31(1):43–56.
20. Fu XL, et al. Study of prognostic predictors for non-small cell lung cancer. Lung Cancer. 1999;23(2):143–52.
21. Blanco AI, 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):1055–69.
22. 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):1106–13.
23. Marks LB. Dosimetric predictors of radiation-induced lung injury. Int J Radiat Oncol Biol Phys. 2002;54(2):313–5.
24. Hope AJ, 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.
25. Spencer SJ, et al. Bioinformatics methods for learning radiation-induced lung inflammation from heterogeneous retrospective and prospective data. J Biomed Biotechnol. 2009;2009:892863.
26. Härdle W. Applied multivariate statistical analysis. Berlin: Springer; 2003.
27. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations, Springer series in statistics, vol. xvi. New York: Springer; 2001. 533 p.
28. El Naqa I, Randall K. Dosimetric data in a guide to outcome modeling in radiotherapy and oncology: listening to the data. CRC Press. Taylor & Francis Group Boca Raton, FL; 2018.
29. El Naqa I, et al. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. Phys Med Biol. 2006;51(22):5719–35.
30. Chabner BA, Roberts TG. Chemotherapy and the war on cancer. Nat Rev Cancer. 2005;5(1):65–72.
31. Khoukaz T. Administration of anti-EGFR therapy: a practical review. Semin Oncol Nurs. 22: 20–27.
32. Seiwert TY, Salama JK, Vokes EE. The concurrent chemoradiation paradigm[mdash]general principles. Nat Clin Prac Oncol. 2007;4(2):86–100.
33. Wen PY, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. J Clin Oncol. 2010;28(11):1963–72.
34. Condeelis J, Weissleder R. In vivo imaging in cancer. Cold Spring Harb Perspect Biol. 2010;2(12):a003848.
35. Willmann JK, et al. Molecular imaging in drug development. Nat Rev Drug Discov. 2008;7(7):591–607.
36. Bussink J, et al. PET-CT for radiotherapy treatment planning and response monitoring in solid tumors. Nat Rev Clin Oncol. 2011;8(4):233–42.
37. Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging. 2010;37(11):2165–87.
38. Newbold K, et al. Advanced imaging applied to radiotherapy planning in head and neck cancer: a clinical review. Br J Radiol. 2006;79(943):554–61.
39. Piet, D., et al. Diffusion-weighted magnetic resonance imaging to evaluate major salivary gland function before and after radiotherapy. Int J Radiat Oncol Biol Phys 2008.
40. Lambin P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–6.
41. Kumar V, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30(9):1234–48.

42. Tixier F, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med. 2011;52(3):369–78.
43. El Naqa I, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recogn. 2009;42(6):1162–71.
44. Cheng NM, et al. Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma. J Nucl Med. 2013;54(10):1703–9.
45. Kidd EA, et al. FDG-PET-based prognostic nomograms for locally advanced cervical cancer. Gynecol Oncol. 2012;127(1):136–40.
46. Vaidya M, et al. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. Radiother Oncol. 2012;102(2):239–45.
47. Cook GJ, et al. Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? J Nucl Med. 2013;54(1):19–26.
48. Vallieres M, et al. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol. 2015;60(14):5471–96.
49. Sun R, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol. 2018;19(9):1180–91.
50. Trebeschi S, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Ann Oncol. 2019;30(6):998–1004.
51. El Naqa I, Ten Haken RK. Can radiomics personalise immunotherapy? Lancet Oncol. 2018;19(9):1138–9.
52. Group B.D.W. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69(3):89–95.
53. El Naqa I, et al. Biomarkers for early radiation response for adaptive radiation therapy. In: Li XA (ed) Adaptive radiation therapy. Taylor & Francis: Boca Baton, FL; 2011. p 53-68.
54. The Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576–82.
55. Abeshouse A, et al. The molecular taxonomy of primary prostate cancer. Cell. 163(4): 1011–1025.
56. The Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature, 2014;511(7511): 543–550.
57. Brennan CW, et al. The somatic genomic landscape of glioblastoma. Cell. 155(2): 462–477.
58. Codd EF. The relational model for database management: version 2. Reading, MA: Addison-Wesley; 1990. xxii, 538 p.
59. Pearl J, Glymour M, Jewell NP. Causal inference in statistics: a primer. Chichester: Wiley; 2016.
60. Lake P, Drake R. Information systems management in the big data era. New York: Springer; 2015.
61. Andre F, et al. Biomarker studies: a call for a comprehensive biomarker study registry. Nat Rev Clin Oncol. 2011;8(3):171–6.
62. Simpson EH. The interpretation of interaction in contingency tables. J Royal Stat Soc Series B (Methodological). 1951;13(2): 238–241.
63. Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. 2015;347(6217):78–81.
64. Albini A, et al. Strategies to Prevent "Bad Luck" in Cancer. J Natl Cancer Inst. 2015;107:10.
65. Kievit R, et al. Simpson's paradox in psychological science: a practical guide. Front Psychol. 2013;4
66. Nikjoo H, et al. Track-structure codes in radiation research. Radiat Meas. 2006;41(9–10):1052–74.

67. El Naqa I, Pater P, Seuntjens J. Monte Carlo role in radiobiological modelling of radiotherapy outcomes. Phys Med Biol. 2012;57(11):R75–97.
68. El Naqa I, et al. Multi-variable modeling of radiotherapy outcomes including dose-volume and clinical factors. Int J Radiat Oncol Biol Phys. 2006;64(4):1275–86.
69. El Naqa I, Li R, Murphy MJ. Machine learning in radiation oncology: theory and application. Geneva: Springer; 2015.
70. El Naqa I, et al. Predicting radiotherapy outcomes using statistical learning techniques. Phys Med Biol. 2009;54:S9–S30.
71. Tucker SL, et al. Dose-volume response analyses of late rectal bleeding after radiotherapy for prostate cancer. Int J Radiat Oncol Biol Phys. 2004;59(2):353–65.
72. Bradley JD, et al. A nomogram to predict radiation pneumonitis, derived from a combined analysis of RTOG 9311 and institutional data. Int J Radiat Oncol Biol Phys. 2007;69(4):985–92.
73. Huang EX, et al. Modeling the risk of radiation-induced acute esophagitis for combined Washington university and RTOG trial 93-11 lung cancer patients. Int J Radiat Oncol Biol Phys. 2011;12:60.
74. Huang EX, et al. Heart irradiation as a risk factor for radiation pneumonitis. Acta Oncol. 2011;50(1):51–60.
75. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Berlin: Springer; 2009.
76. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. Monographs on statistics and applied probability. Boca Raton: CRC Press/Taylor & Francis Group; 2015. xv, 351 pages.
77. Albertina B, Watson M, Holback C, Jarosz R, Kirk S, Lee Y, Lemmerman J. Radiology data from the cancer genome atlas lung adenocarcinoma [TCGA-LUAD] collection, T.C.I. Archive, Editor. 2015.
78. Kirk S, Lee Y, Kumar P, Filippini J, Albertina B, Watson M, Lemmerman J. Radiology data from the cancer genome atlas lung squamous cell carcinoma [TCGA-LUSC] collection. T.C.I. Archive., Editor. 2015.
79. Pedregosa AF, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011; 12: 2825–2830.
80. Munley MT, et al. A neural network to predict symptomatic lung injury. Phys Med Biol. 1999;44:2241–9.
81. Su M, et al. An artificial neural network for predicting the incidence of radiation pneumonitis. Med Phys. 2005;32(2):318–25.
82. Gulliford SL, et al. Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate. Radiother Oncol. 2004;71(1):3–12.
83. Tomatis S, et al. Late rectal bleeding after 3D-CRT for prostate cancer: development of a neural-network-based predictive model. Phys Med Biol. 2012;57(5):1399.
84. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
85. Oh JH, et al. A Bayesian network approach for modeling local failure in lung cancer. Phys Med Biol. 2011;56(6):1635–51.
86. Lee S, et al. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. Med Phys. 2015;42(5):2421–30.
87. Jayasurya K, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys. 2010;37(4):1401–7.
88. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Adaptive computation and machine learning. Cambridge, MA: MIT Press; 2009. xxi, 1231 p.
89. Sinoquet C, Mourad RL. Probabilistic graphical models for genetics, genomics, and postgenomics. First edition. ed. Oxford: Oxford University Press; 2014. xxvii, 449 pages, 4 unnumbered pages of plates.
90. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
91. Ding B, Qian H, Zhou J. Activation functions and their characteristics in deep neural networks. In 2018 Chinese Control And Decision Conference (CCDC);2018.

92. Paszke AaG, Sam and Massa, Francisco and Lerer, Adam and Bradbury, James and Chanan, Gregory and Killeen, Trevor and Lin, Zeming and Gimelshein, Natalia and Antiga, Luca and Desmaison, Alban and Kopf, Andreas and Yang, Edward and DeVito, Zachary and Raison, Martin and Tejani, Alykhan and Chilamkurthy, Sasank and Steiner, Benoit and Fang, Lu and Bai, Junjie and Chintala, Soumith, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in Advances in Neural Information Processing Systems 32, W.a.H. H., Larochelle and A., Beygelzimer and F, dAlcheBuc and E., Fox and R., Garnett, Editor. 2019, Curran Associates, Inc.
93. Eschrich S, et al. Systems biology modeling of the radiation sensitivity network: a biomarker discovery platform. Int J Radiat Oncol Biol Phys. 2009;75(2):497–505.
94. Shivade C, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014;21(2):221–30.
95. Dizon DS, et al. Clinical cancer advances 2016: Annual report on Progress against cancer from the American Society of Clinical Oncology. J Clin Oncol. 2015;
96. Benedict SH, El Naqa I, Klein EE. Introduction to Big Data in radiation oncology: exploring opportunities for research, quality assessment, and clinical care. Int J Radiat Oncol Biol Phys. 95(3): 871–872.
97. Gail M. A review and critique of some models used in competing risk analysis. Biometrics. 1975;31(1):209–22.
98. Murray TA, Thall PF, Yuan Y. Utility-based designs for randomized comparative trials with categorical outcomes. Stat Med. 2015;
99. Wilkinson MD, et al. The FAIR guiding principles for scientific data management and stewardship. Scientific Data. 2015;3(1):160018.
100. Ghorbani AZ. Data shapley: equitable valuation of data for machine learning. In Machine learning research 2019.
101. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018;73:1–15.
102. Zanzotto FM. Viewpoint: human-in-the-loop artificial intelligence. J Artif Intell Res. 2019;64:243–52.
103. Wang L, et al. An integrated approach for the analysis of biological pathways using mixed models. PLoS Genet. 2008;4(7):e1000115.

Radiomics and Radiogenomics

16

Ruijiang Li

16.1 Introduction

Radiographic imaging plays a critical role in radiation oncology, including radiation treatment planning and setup, evaluation of therapeutic response, and subsequent follow-up and disease monitoring [1–4]. In current clinical practice, the diagnostic interpretation of these images relies on visual assessment of few qualitative imaging traits. However, this approach does not allow a comprehensive characterization of the disease and has limited accuracy for prediction or assessment of treatment response and prognosis.

Radiomics is a powerful technique for discovering promising biomarkers by high-throughput quantitative analysis of medical images. This can be achieved either using a predefined set of manually handcrafted features, such as shape, histogram, and texture, or automated feature extraction by deep learning such as convolutional neural networks. By applying appropriate statistical or machine learning tools, predictive models can be developed to potentially improve the accuracy of outcome prediction. Radiogenomics is a closely related field that concerns the study of relationships between radiomic features at the tissue scale and underlying molecular features at the genomic, transcriptomic, or proteomic level.

Radiomics have been applied to all types of standard-of-care clinical images such as CT, MRI, and PET [5–9]. Many studies have identified novel imaging signatures that demonstrated improved diagnostic, prognostic, or predictive performance over currently used imaging metrics, while radiogenomics may allow identification of the underlying biological basis of these imaging phenotypes [10–17]. In the following sections, we will discuss the technical aspects, key findings,

R. Li (✉)
Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA, USA
e-mail: rli2@stanford.edu

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_16

and clinical applications of radiomics and radiogenomics with a focus on radiation oncology.

16.2 Technical Basis of Radiomics

Radiomics is multistep process that involves image acquisition, tumor segmentation, feature extraction, predictive modeling, and model validation. Figure 16.1 shows a general workflow of radiomics, which includes both manually handcrafted features and automated feature extraction by deep learning (discussed in detail in Sect. 16.4).

Various radiographic imaging modalities such as CT, MRI, and PET are used in radiation oncology practice to provide direct visualization and evaluation of the underlying anatomical or physiological properties of the tumor [18]. Because these standard-of-care images are normally acquired for every patient undergoing radiation treatment, they provide the necessary big imaging data for machine learning and modeling for outcome prediction.

For the radiomics approach with handcrafted features, segmentation of the region of interest—in most cases, the gross tumor, is required. For patients treated with radiotherapy, their primary tumors have already been manually delineated by radiation oncologists and are available from the treatment planning system. These preexisting contours can greatly facilitate retrospective radiomic analysis. However, there can be significant variations in tumor contours among different oncologists. To account for intra- and inter-rater variations, it is important to evaluate the robustness of image features and effect on downstream analysis by perturbing the tumor contours or using multiple delineations. Alternatively, tumor contoured can be defined more consistently using semiautomated segmentation algorithms with minimal human inputs [19–22].

Two types of handcrafted image features, semantic and agnostic, are often used to characterize the tumor phenotypes. Semantic features are defined based on existing radiology lexicon to qualitatively describe tumors and can be derived from the existing guidelines of specific imaging-reporting and data system by the American College of Radiology. On the other hand, agnostic features are computational

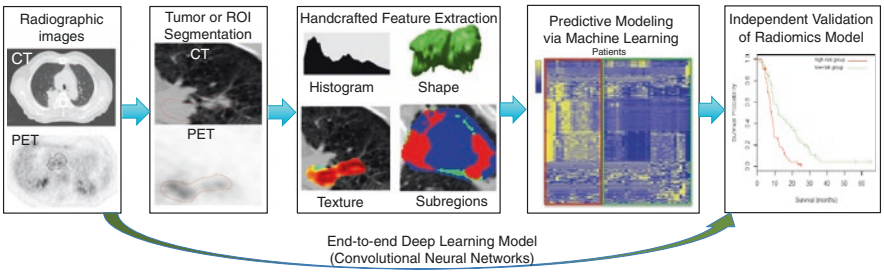


Fig. 16.1 Workflow of a typical radiomic study

metrics with predefined mathematical formulations. There are various types of agnostic image features that describe tumor shape, intensity, and texture to capture intratumoral heterogeneity. The details of agnostic features have been reviewed elsewhere [7, 23]. Commonly used radiomic features have been integrated into open-source software or commercial software platform. Among these [24, 25], Deasy and colleagues have provided an open-source platform, known as CERR [26] (http://www.cerr.info/), to prototype algorithms for radiomic analysis specifically for radiotherapy research.

Given the radiomics features for the tumor phenotypes, machine learning algorithms can be applied to discover and quantify their relations to relevant clinical endpoints or genomic traits. Supervised learning such as regression or classification methods are commonly used depending on the type of targeted variables being continuous values or class labels. Due to the large number of features compared with a relatively small number of samples, feature selection is an essential step to reduce the risk overfitting [27]. Image features that show minimal changes to tumor contour variations and minimal redundancy or overlap with other features may be preferentially selected. Various feature selection algorithms, including stepwise forward/backward selection and lasso, can be used to identify the most informative features to fit the prediction model. Cross validation is usually applied to minimize the potential feature selection bias. In addition to building predictive models with supervised learning algorithms, it is also possible to apply unsupervised clustering algorithms to the radiomic features in order to discover novel subtypes for a given disease [13, 14].

After initial discovery and training of promising signatures, any radiomics-based model should be tested in independent, preferably multiple external cohorts. The key for rigorous validation is that training and testing should be entirely separate and no information leakage should occur between the two procedures [28]. In addition, it is also important to evaluate the relationship between the newly proposed radiomics signatures and known clinical and pathologic factors. Those radiomic signatures that provide independent prediction power in a multivariable model are more likely to add clinical value for patient management.

16.3 Key Findings and Clinical Applications

In this section, we highlight some recent studies that may be potentially relevant for improving patient management in radiotherapy. Given the tremendous growth in radiomics research [5–8, 29–35], interested readers are referred to recent reviews.

Aerts and colleagues proposed a CT-based radiomics signature to predict overall survival in non-small cell lung cancer (NSCLC) patients treated with radiotherapy [36]. They extracted over 400 computational features from CT images to describe tumor intensity, shape, texture, and wavelet. Using 4 features selected from each category, they constructed a radiomic signature in a training cohort of over 400 patients and confirmed its prognostic value in two additional cohorts. Since publication of this landmark study, several groups have attempted to independently validate

this signature, which led to mixed results [37]. Potential concerns mostly regard reproducibility of the radiomics signature, and the fact that it is highly correlated to tumor size and volume [38], raising questions about its additive value beyond what's already known. The other issue is using overall survival as an endpoint, which is confounded by many non-disease related factors, and is less useful for guiding management.

In another radiomics study focusing on early stage NSCLC, Wu et al. investigated quantitative radiomic features of FDG-PET and CT for predicting distant metastasis after stereotactic ablative radiotherapy (SABR) [39]. Based on image features characterizing tumor morphology and intratumoral metabolic heterogeneity, they built a radiomic signature which significantly improved the prognostic value compared with conventional imaging metrics. Moreover, combining imaging and histologic information yielded further improvement in the accuracy of prediction for distant metastasis.

In addition to NSCLC, the potential of radiomics has been extensively investigated in head and neck cancer. El Naqa and colleagues studied FDG-PET/CT radiomics and combined them with clinical information to assess the risks of locoregional recurrence and distant metastasis in head and neck cancer patients [40]. The prognostic value of constructed prediction models was confirmed in an external cohort. In another recent study, Wu et al. analyzed quantitative image features of the primary tumor and involved lymph nodes defined at the baseline and midtreatment imaging characteristics of oropharyngeal cancer (OPC). They developed a machine learning model based on random survival forest and found that nodal imaging features were the most important features for predicting distant metastasis. The model provided independent prognostic information beyond established clinical factors including stage, smoking history, and HPV status. It further stratified patients within the subgroup of patients with HPV-positive OPC [41]. This study highlights the need to evaluate nodal imaging characteristics beyond primary tumor. The machine learning model has the potential to identify HPV-positive OPC patients who have a higher risk of distant relapse and should not be considered for treatment deintensification.

The majority of radiomic studies are focused on the analysis of the primary tumor as a whole. This bulk analysis approach implicitly assumes that the tumor is heterogeneous but well mixed, and the regional variations within a tumor are neglected. To address this issue, the concept of habitat imaging was proposed to capture spatial heterogeneity more explicitly at a regional level [8, 42]. Cao and colleagues proposed a clustering-based algorithm to identify the significant subvolumes for primary tumors from dynamic contrast-enhanced (DCE) MRI, which can predict local or regional failure in head and neck cancer [43]. Wu et al. developed a robust tumor partitioning method by a two-stage clustering procedure and identified three spatially distinct and phenotypically consistent subregions in lung tumors. One subregion associated with the most metabolically active, metabolically heterogeneous, and solid component of the tumor was defined as the "high-risk" subregion. The volume of high-risk intratumoral subregion predicted distant metastasis in patients with locally advanced NSCLC treated with radiation therapy [44].

The intratumor partitioning approach can be extended by combining with radiomic or texture analysis to allow more detailed and refined image phenotyping. Wu et al. [45] showed that the early change of texture features for the intratumoral subregion associated with fast contrast-agent washout at DCE MRI predicted pathological complete response to neoadjuvant chemotherapy in breast cancer. Cui et al. [46] performed radiomic analysis on tumor subregions and defined 120 multiregional image features on MRI in glioblastoma. A 5-feature radiomic signature was identified and independently validated in an external cohort to predict overall survival, which outperformed whole-tumor measurements. Stoyanova and colleagues investigated the association of MRI radiomic features with prostate cancer gene expression profiles from MRI-guided biopsy tissues [47].

These studies highlight the need for tumor partitioning to identify aggressive intratumoral subregions, which is applicable to many types of solid tumors. This may have significant implications for clinical oncology by identifying important tumor regions for biopsy. In addition, this is particularly relevant for radiotherapy treatment planning and adaptation, because high-risk subregions associated with the aggressive disease can then be targeted with a radiation boost to potentially improve local tumor control.

16.4 Emerging Paradigms: Deep Learning

The radiomics approach described above requires domain expertise to manually construct handcrafted images. Automated feature extraction using advanced deep learning is being increasingly used for radiomics. Deep learning as detailed in Chap. 4, is a machine learning technique that uses multiple processing layers and connections to learn complex relations between input data and desired output from a large set of labeled examples. Different from traditional machine learning techniques that require domain expertise to design features, deep learning directly deals with raw data and automatically develops its own representations needed for pattern recognition, thus eliminating the need to specify rules or features explicitly [48].

Recently, deep learning techniques have achieved promising and sometimes human expert-level performance in a variety of medical image interpretation tasks, such as detecting diabetic retinopathy in fundus photographs, classifying skin cancer in skin photographs, and detecting breast cancer lymph node metastasis in pathological images [49–53]. So far, the vast majority of deep learning applications have focused on disease detection and classification in a diagnostic setting [54, 55]. Several proof of concept studies suggest that deep learning could also be used to extract information from medical images for predicting survival outcomes [56–59]. At present, there are relatively few large cohort studies with independent validation of deep learning-based imaging signatures, although this likely will change soon.

In a recent study, Jiang et al. developed and independently validated a deep learning-based CT imaging signature to predict survival outcomes in a multicenter cohort study of over 1700 patients with gastric cancer [60]. They proposed a novel end-to-end deep neural network to predict a patient's risk of death based on CT

image, which was named "S-net" according to the shape of its architecture. Different from traditional deep CNN, S-net incorporates the concept of multi-level feature stream fusion. The rationale for this design is that both low-level features (e.g., local edges and textures) at shallow layers and high-level features (e.g., gross disease appearances) at deep layers contain useful information for survival analysis. This will allow the extraction and integration of comprehensive multi-scale image features for complex tumor phenotypes. The prognostic value of deep learning signature was independent of traditional clinicopathologic risk factors including TNM stage. By combining clinicopathologic and imaging predictors, they showed that an integrated nomogram had a much improved prognostic accuracy than did either alone. Importantly, the proposed imaging signature predicted benefit from adjuvant chemotherapy and could potentially guide personalized therapy in gastric cancer.

In another recent study, Abazeed and colleagues applied deep learning to analyze pre-treatment CT scans in a retrospective cohort study of 944 lung cancer patients treated with stereotactic body radiation therapy [61]. The single-institutional cohort was split into 849 patients in the training cohort and 95 patients in the independent validation cohort. The study integrated traditional radiomics features through multitask learning, by applying a time-based survival analysis, and incorporating new deep learning methods. They input pre-treatment CT images into a multi-task deep neural network and combined these data with clinical variables to derive *i*Gray, an individualized radiation dose that estimates the probability of treatment failure. Models that included deep learning model and clinical variables predicted treatment failures with a significant improvement in accuracy compared with traditional radiomics or clinical variables alone. This model could help identify tumors that are most resistant to radiation and may lead to personalized dosing of radiation therapy.

16.5 Radiogenomics: Integrating Imaging with Genomics

Radiogenomics builds upon radiomics, which integrates imaging and genomic data with the goal of gaining biological interpretation or improving patient stratification for precision medicine [10–15, 62–66]. Depending on the goal and study design, there are several major approaches for radiogenomic integration, which are depicted in Fig. 16.2.

The most commonly adopted radiogenomic study design is to find imaging correlates or surrogate of a specific genotype or molecular phenotype of the tumor. For instance, CT semantic and radiomic image features were found to be associated with *EGFR* or *KRAS* mutations in lung cancer [67–72]. MRI radiomic features were also correlated with intrinsic molecular subtypes or existing genomic assays in breast cancer [73–75]. Several studies have investigated the association between imaging features and tumor-infiltrating lymphocytes (TILs) [76–79]. For instance, Ferté et al. correlated both intra and peritumor radiomics features with the expression of CD8 in tumor core, and they showed the imaging signature may be useful in estimating CD8 cell count and predicting clinical outcomes of patients treated with immunotherapy [77].

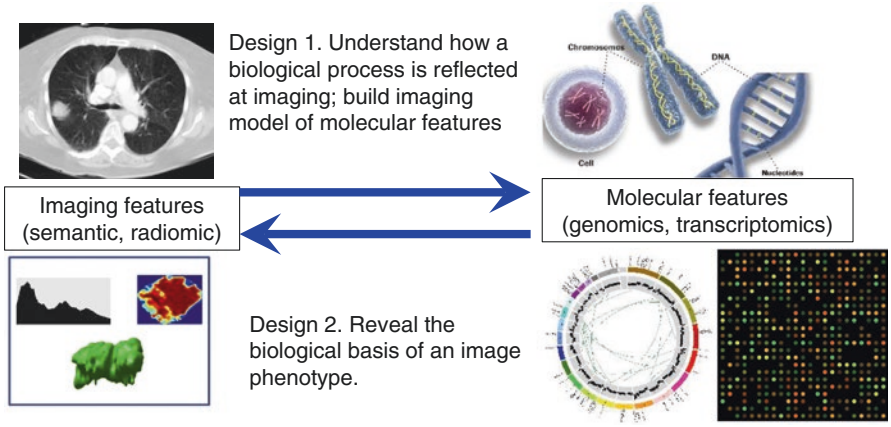


Fig. 16.2 Typical design of radiogenomic study

Radiogenomics can also be used create association maps between molecular features and a specific imaging phenotype to reveal its biological underpinnings. For example, tumors with higher maximum standardized uptake value from FDG-PET were demonstrated to be associated with the epithelial-mesenchymal transition in non–small cell lung cancer [80]. In another recent radiogenomic study, heterogeneous enhancing patterns of tumor-adjacent parenchyma from perfusion MRI were associated with the tumor necrosis signaling pathway and poor survival in breast cancer [15].

Beside radiogenomic association, one interesting area of investigation is to use unsupervised learning such as clustering algorithms to classify tumors into subtypes based on imaging phenotypes rather than molecular features. Itakura et al. identified novel glioblastoma subtypes based on MRI phenotypes that are associated with distinct molecular pathway activities [13]. Wu et al. [14] discovered and independently validated three breast cancer imaging subtypes, which were characterized by homogeneous intratumoral enhancement, minimal parenchymal enhancement, and prominent parenchymal enhancement. They were shown to be complementary to known intrinsic molecular subtypes. In a multi-cohort study of over 1000 patients, each of the 3 imaging subtypes was associated with distinct prognoses and dysregulated molecular pathways.

Another direction that is less investigated but particularly relevant for precision medicine is to leverage the complementary power of imaging and molecular data and integrate them into a unifying model to further improve prediction accuracy of clinical outcomes. Along this line, Cottereau et al. [81] showed that combination of molecular profile and metabolic tumor volume at FDG-PET imaging improved patient stratification for progression-free and overall survival in diffuse large B-cell lymphoma. Cui et al. [82] showed that integrating MGMT methylation status and volume of the high-risk subregion at multi-parametric MRI improved survival stratification in glioblastoma. Lee et al. developed a CT image-based prognostic signature and validated it in an external cohort of patients with stage I NSCLC [83]. Further, it was shown that a composite imaging and genomic signature improved

prognostic accuracy upon either one used alone. These studies provide the initial evidence that image-based biomarkers can provide additional information beyond molecular analysis alone and integrating both will provide more accurate individualized prediction of clinical outcomes.

16.6 Current Challenges and Potential Solutions

Despite the growing interest and promising findings in numerous studies, the progress for clinical translation of radiomics signatures has been slow. It is important to highlight some technical and practical challenges facing the field. These issues include standardization of image acquisition protocols and feature extraction, robustness and reproducibility of radiomic signatures, data sharing, and validation in multicenter cohorts.

16.6.1 Standardization and Quantitative Imaging

Clinical images are typically acquired with the goal of maximizing the contrast between normal and diseased tissues. There is often a lack of standardization of imaging protocols across institutions with different acquisition and reconstruction parameters, which can significantly hamper quantitative radiomic analysis. It is essential to standardize or harmonize the imaging data in multicenter validation studies. To overcome this issue, there have been several efforts that aim to standardize the imaging protocol by the quantitative imaging biomarkers alliance (QIBA) [84] and quantitative imaging network (QIN) [85] among others. In a retrospective analysis, several strategies have been proposed to harmonize imaging scans such that they are comparable across multiple cohorts. A common strategy is to derive the underlying physiological measures from the functional imaging. For instance, the perfusion maps can be computed from DCE MRI based on pharmacokinetic modeling [86]. Another practical strategy is to gauge the imaging values with the value of selected normal tissue region of interest as a baseline. For instance, on an individual basis, the average inter-quantile values of the background parenchyma can be used to normalize breast MRI scans [14]. In addition, the phantom study can be adopted to investigate the inter-scan and inter-vendor variability of the imaging-derived features [87, 88], which can provide useful insights into the uncertainties of quantitative imaging analysis.

16.6.2 Reproducibility and Need for Prospective Validation

Currently the biggest hurdle toward the clinical translation of radiomics is probably the lack of reproducibility. Several pitfalls can be implicated, including poor experimental design, multiple testing leading to false discovery and model overfitting, and unadjusted biases or confounding factors among others [89]. To enhance

reproducibility, a rational radiomic design should include proper imaging standardization, robustness test of radiomic features regarding segmentation variabilities, as well as rigorous model training and testing. Second, each radiomic analysis step should be well documented and original codes and data are easily accessible, allowing other investigators to replicate the results. Prior to clinical translation of any putative biomarkers, the most critical step is rigorous validation in prospective clinical trials.

16.6.3 Data and Software Sharing

Another challenge facing radiomics is the curation and sharing of image and relevant clinical data in large patient cohorts across multiple centers [85, 90]. It is important to match imaging with detailed clinicopathological and treatment information, as well as relevant clinical outcomes. There has been some progress toward data sharing under the initiative of the cancer imaging archive, where image and clinical data for various tumor sites are curated and shared publicly (http://www.cancerimagingarchive.net/). These cohorts are from single-institution or multicenter trails which should greatly facilitate discovery and validation of radiomic models. Nonetheless, compared with the abundant public gene expression data, the available imaging data are much less and continuing efforts should be spent to curate high-quality imaging datasets. Beyond technical challenges, there are also administrative and regulatory barriers that need to overcome in order to make large-scale data sharing feasible in the future. Due to the complexity of deep learning models and its black box nature, access and sharing of software codes will become essential to replicate and validate the models.

16.7 Conclusion and Future Outlook

In conclusion, radiomics and radiogenomics have shown a great promise for the discovery of novel imaging biomarkers with potential prognostic and predictive value. However, it should be noted that many studies so far are of hypothesis-generating nature, and few radiomic signatures have been prospectively validated in independent cohorts. Additionally, to be of practical value, any new candidate imaging biomarkers should provide complementary information to known clinical and pathologic factors. One critical and yet an under-explored area of investigation is how radiomics can be applied to serial imaging scans to better evaluate therapeutic response given the increasing availability of treatment regimens. Initial studies based on delta-radiomics are encouraging, but the optimum approach to characterizing longitudinal change remains to be defined.

Moving forward, advanced machine learning techniques, such as deep convolutional neural networks, are expected to play an increasingly important role in defining novel image-based prognostic and predictive models. In order for this approach to work, a sufficiently large dataset will be required to train a reliable model,

highlighting the need for curation of high-quality datasets and data sharing. Regardless of radiomics approaches with either handcrafted features or automated deep learning, prospective studies in multicenter clinical trials will be required to validate newly identified imaging biomarkers and truly establish the value of radiomics and radiogenomics in personalized radiation oncology.

Conflict of Interest None.

References

1. O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14(3):169–86.
2. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, et al. Quantitative imaging in cancer clinical trials. Clin Cancer Res. 2016;22(2):284–90.
3. Weissleder R, Schwaiger MC, Gambhir SS, Hricak H. Imaging approaches to optimize molecular therapies. Sci Transl Med. 2016;8(355):355ps16.
4. O'Connor JP, editor. Cancer heterogeneity and imaging. Semin Cell Dev Biol. 2017;64:48–57.
5. Aerts HJ. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncol. 2016;2(12):1636–42.
6. Sala E, Mema E, Himoto Y, Veeraraghavan H, Brenton J, Snyder A, et al. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. Clin Radiol. 2017;72(1):3–10.
7. Lambin P, Leijenaar RT, Deist TM, Peerlings J, de Jong EE, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14(12):749–62.
8. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2015;278(2):563–77.
9. Limkin E, Sun R, Dercle L, Zacharaki E, Robert C, Reuzé S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Ann Oncol. 2017;28(6):1191–206.
10. Diehn M, Nardini C, Wang DS, McGovern S, Jayaraman M, Liang Y, et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. Proc Natl Acad Sci U S A. 2008;105(13):5213–8.
11. Gevaert O, Xu JJ, Hoang CD, Leung AN, Xu Y, Quon A, et al. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data-methods and preliminary results. Radiology. 2012;264(2):387–96.
12. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat Biotechnol. 2007;25(6):675–80.
13. Itakura H, Achrol AS, Mitchell LA, Loya JJ, Liu T, Westbroek EM, et al. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. Sci Transl Med. 2015;7(303):303ra138.
14. Wu J, Cui Y, Sun X, Cao G, Li B, Ikeda DM, et al. Unsupervised clustering of quantitative image phenotypes reveals breast cancer subtypes with distinct prognoses and molecular pathways. Clin Cancer Res. 2017;23(13):3334–42.
15. Wu J, Li B, Sun X, Cao G, Rubin DL, Napel S, et al. Heterogeneous enhancement patterns of tumor-adjacent parenchyma at MR imaging are associated with dysregulated signaling pathways and poor survival in breast cancer. Radiology. 2017;285(2):401–13.
16. Colen R, Foster I, Gatenby R, Giger ME, Gillies R, Gutman D, et al. NCI workshop report: clinical and computational requirements for correlating imaging phenotypes with genomics signatures. Transl Oncol. 2014;7(5):556–69.

17. Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA, et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proc Natl Acad Sci. 2015;112(46):E6265–E73.
18. Timmerman R, Xing L. Image guided and adaptive radiation therapy. Baltimore: Lippincott Williams & Wilkins; 2009.
19. Parmar C, Velazquez ER, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One. 2014;9(7):e102107.
20. Arik SO, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham). 2017;4(1):014501.
21. Ibragimov B, Korez R, Likar B, Pernus F, Xing L, Vrtovec T. Segmentation of pathological structures by landmark-assisted deformable models. IEEE Trans Med Imaging. 2017;36(7):1457–69.
22. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys. 2017;44(2):547–57.
23. Larue RT, Defraene G, De Ruysscher D, Lambin P, Van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol. 2017;90(1070):20160665.
24. Echegaray S, Bakr S, Rubin DL, Napel S. Quantitative image feature engine (QIFE): an open-source, modular engine for 3D quantitative feature extraction from volumetric medical images. J Digit Imaging. 2017;31(4):403–14.
25. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. 2015;42(3):1341–53.
26. Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys. 2003;30(5):979–85.
27. Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng. 2014;40(1):16–28.
28. Subramanian J, Simon R. What should physicians look for in evaluating prognostic gene-expression signatures? Nat Rev Clin Oncol. 2010;7(6):327–34.
29. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. Phys Med. 2017;38:122–39.
30. Colen RR, Hassan I, Elshafeey N, Zinn PO. Shedding light on the 2016 World Health Organization classification of tumors of the central nervous system in the era of radiomics and radiogenomics. Magn Reson Imaging Clin. 2016;24(4):741–9.
31. Verma V, Simone CB, Krishnan S, Lin SH, Yang J, Hahn SM. The rise of radiomics and implications for oncologic management. J Natl Cancer Inst. 2017;109(7):djx055.
32. O'Connor JPB, Rose CJ, Waterton JC, Carano RAD, Parker GJM, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clin Cancer Res. 2015;21(2):249–57.
33. Mankoff DA, Farwell MD, Clark AS, Pryma DA. Making molecular imaging a clinical tool for precision oncology: a review. JAMA Oncol. 2017;3(5):695–701.
34. Zhang B, Tian J, Dong D, Gu D, Dong Y, Zhang L, et al. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clin Cancer Res. 2017;23(15):4259–69.
35. Huang Y-q, Liang C-h, He L, Tian J, Liang C-s, Chen X, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. J Clin Oncol. 2016;34(18):2157–64.
36. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
37. Ger RB, Zhou S, Elgohari B, Elhalawani H, Mackin DM, Meier JG, et al. Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT-and PET-imaged head and neck cancer patients. PLoS One. 2019;14(9):e0222509.

38. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol. 2019;130:2–9.
39. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo BW Jr, et al. Early-stage non–small cell lung cancer: quantitative imaging characteristics of 18F fluorodeoxyglucose PET/CT allow prediction of distant metastasis. Radiology. 2016;281(1):270–8.
40. Vallieres M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts H, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017;7(1):10116.
41. Wu J, Gensheimer MF, Zhang N, Han F, Liang R, Qian Y, et al. Integrating tumor and nodal imaging characteristics at baseline and mid-treatment computed tomography scans to predict distant metastasis in oropharyngeal cancer treated with concurrent chemoradiotherapy. Int J Radiat Oncol Biol Phys. 2019;104(4):942–52.
42. Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. Radiology. 2013;269(1):8–15.
43. Wang P, Popovtzer A, Eisbruch A, Cao Y. An approach to identify, from DCE MRI, significant subvolumes of tumors related to outcomes in advanced head-and-neck cancer. Med Phys. 2012;39(8):5277–85.
44. Wu J, Gensheimer MF, Dong X, Rubin DL, Napel S, Diehn M, et al. Robust intratumor partitioning to identify high-risk subregions in lung cancer: a pilot study. Int J Radiat Oncol Biol Phys. 2016;95(5):1504–12.
45. Wu J, Gong G, Cui Y, Li R. Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy. J Magn Reson Imaging. 2016;44(5):1107–15.
46. Cui Y, Tha KK, Terasaka S, Yamaguchi S, Wang J, Kudo K, et al. Prognostic imaging biomarkers in glioblastoma: development and independent validation on the basis of multiregion and quantitative analysis of MR images. Radiology. 2015;278(2):546–53.
47. Stoyanova R, Pollack A, Takhar M, Lynne C, Parra N, Lam LL, et al. Association of multiparametric MRI quantitative imaging features with prostate cancer gene expression in MRI-targeted prostate biopsies. Oncotarget. 2016;7(33):53362.
48. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9.
49. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10.
50. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
51. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318(22):2199–210.
52. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 2018;392(10162):2388–96.
53. Kickingereder P, Isensee F, Tursunova I, Petersen J, Neuberger U, Bonekamp D, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. Lancet Oncol. 2019;20(5):728–40.
54. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol. 2019;20(2):193–201.
55. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25(6):954–61.
56. Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLoS Med. 2019;16(1):e1002730.

57. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. PLos Med. 2018;15(11):e1002711.
58. Peng H, Dong D, Fang M, Li L, Tang LL, Chen L, et al. Prognostic value of deep learning PET/CT-based radiomics: potential role for future individual induction chemotherapy in advanced nasopharyngeal carcinoma. Clin Cancer Res. 2019;25(14):4271–9.
59. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velazquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A. 2018;115(13):E2970–E9.
60. Jiang Y, Jin C, Yu H, Wu J, Chen C, Yuan Q, et al. Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: a multicenter, retrospective study. Ann Surg. 2020.
61. Lou B, Doken S, Zhuang T, Wingerter D, Gidwani M, Mistry N, et al. An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction. Lancet Digit Health. 2019;1(3):e136–47.
62. Bakas S, Akbari H, Pisapia J, Martinez-Lage M, Rozycki M, Rathore S, et al. In vivo detection of EGFRvIII in glioblastoma via perfusion magnetic resonance imaging signature consistent with deep peritumoral infiltration: the φ index. Clin Cancer Res. 2017;23(16):4724–34.
63. Smits M, van den Bent MJ. Imaging correlates of adult glioma genotypes. Radiology. 2017;284(2):316–31.
64. Vargas HA, Huang EP, Lakhman Y, Ippolito JE, Bhosale P, Mellnick V, et al. Radiogenomics of high-grade serous ovarian cancer: multireader multi-institutional study from the Cancer Genome Atlas Ovarian Cancer Imaging Research Group. Radiology. 2017;285(2):482–92.
65. Lee J, Cui Y, Sun X, Li B, Wu J, Li D, et al. Prognostic value and molecular correlates of a CT image-based quantitative pleural contact index in early stage NSCLC. Eur Radiol. 2018;28(2):736–46.
66. Zhu Y, Li H, Guo W, Drukker K, Lan L, Giger ML, et al. Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. Sci Rep. 2015;5:17787.
67. Liu Y, Kim J, Balagurunathan Y, Li Q, Garcia AL, Stringfield O, et al. Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. Clin Lung Cancer. 2016;17(5):441–8.e6.
68. Rios Velazquez E, Parmar C, Liu Y, Coroller TP, Cruz G, Stringfield O, et al. Somatic mutations drive distinct imaging phenotypes in lung cancer. Cancer Res. 2017;77(14):3922–30.
69. Zhou M, Leung A, Echegaray S, Gentles A, Shrager JB, Jensen KC, et al. Non–small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. Radiology. 2017;286(1):307–15.
70. Yamamoto S, Korn RL, Oklu R, Migdal C, Gotway MB, Weiss GJ, et al. ALK molecular phenotype in non-small cell lung cancer: CT radiogenomic characterization. Radiology. 2014;272(2):568–76.
71. Rizzo S, Petrella F, Buscarino V, De Maria F, Raimondi S, Barberis M, et al. CT radiogenomic characterization of EGFR, K-RAS, and ALK mutations in non-small cell lung cancer. Eur Radiol. 2016;26(1):32–42.
72. Hasegawa M, Sakai F, Ishikawa R, Kimura F, Ishida H, Kobayashi K. CT features of epidermal growth factor receptor–mutated adenocarcinoma of the lung: comparison with nonmutated adenocarcinoma. J Thoracic Oncol. 2016;11(6):819–26.
73. Wu J, Sun X, Wang J, Cui Y, Kato F, Shirato H, et al. Identifying relations between imaging phenotypes and molecular subtypes of breast cancer: model discovery and external validation. J Magn Reson Imaging. 2017;46(4):1017–27.
74. Ashraf AB, Daye D, Gavenonis S, Mies C, Feldman M, Rosen M, et al. Identification of intrinsic imaging phenotypes for breast cancer tumors: preliminary associations with gene expression profiles. Radiology. 2014;272(2):374–84.

75. Li H, Zhu Y, Burnside ES, Drukker K, Hoadley KA, Fan C, et al. MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays. Radiology. 2016;281(2):382–91.
76. Braman N, Prasanna P, Whitney J, Singh S, Beig N, Etesami M, et al. Association of peritumoral radiomics with tumor biology and pathologic response to preoperative targeted therapy for HER2 (ERBB2)-positive breast cancer. JAMA Netw Open. 2019;2(4):e192561.
77. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol. 2018;19(9):1180–91.
78. Wu J, Li X, Teng X, Rubin DL, Napel S, Daniel BL, et al. Magnetic resonance imaging and molecular features associated with tumor-infiltrating lymphocytes in breast cancer. Breast Cancer Res. 2018;20(1):101.
79. Tang C, Hobbs B, Amer A, Li X, Behrens C, Canales JR, et al. Development of an immune-pathology informed radiomics model for non-small cell lung cancer. Sci Rep. 2018;8(1):1922.
80. Yamamoto S, Huang D, Du L, Korn RL, Jamshidi N, Burnette BL, et al. Radiogenomic analysis demonstrates associations between 18F-fluoro-2-deoxyglucose PET, prognosis, and epithelial-mesenchymal transition in non–small cell lung cancer. Radiology. 2016;280(1):261–70.
81. Cottereau AS, Lanic H, Mareschal S, Meignan M, Vera P, Tilly H, et al. Molecular profile and FDG-PET/CT total metabolic tumor volume improve risk classification at diagnosis for patients with diffuse large B-cell lymphoma. Clin Cancer Res. 2016;22(15):3801–9.
82. Cui Y, Ren S, Tha KK, Wu J, Shirato H, Li R. Volume of high-risk intratumoral subregions at multi-parametric MR imaging predicts overall survival and complements molecular analysis of glioblastoma. Eur Radiol. 2017;27(9):3583–92.
83. Lee J, Li B, Sun X, Cui Y, Wu J, Zhu H, et al. A quantitative CT imaging signature predicts survival and complements established prognosticators in stage I non-small cell lung cancer. Int J Radiat Oncol Biol Phys. 2018;102(4):1098–106.
84. Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. Radiology. 2011;258(3):906–14.
85. Kalpathy-Cramer J, Freymann JB, Kirby JS, Kinahan PE, Prior FW. Quantitative imaging network: data sharing and competitive AlgorithmValidation leveraging the cancer imaging archive. Transl Oncol. 2014;7(1):147–52.
86. Yankeelov TE, Gore JC. Dynamic contrast enhanced magnetic resonance imaging in oncology: theory, data acquisition, analysis, and examples. Curr Med Imaging Rev. 2007;3(2):91–107.
87. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring CT scanner variability of radiomics features. Invest Radiol. 2015;50(11):757.
88. Zhao B, Tan Y, Tsai W-Y, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep. 2016;6:23428.
89. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One. 2015;10(5):e0124165.
90. Roelofs E, Dekker A, Meldolesi E, van Stiphout RG, Valentini V, Lambin P. International data-sharing for radiotherapy research: an open-source based infrastructure for multicentric clinical data mining. Radiother Oncol. 2014;110(2):370–4.

Modelling of Radiotherapy Response (TCP/NTCP)

17

Sarah Gulliford and Issam El Naqa

17.1 Introduction

Recent years have witnessed tremendous technological advances in radiotherapy treatment planning, image guidance, and treatment delivery [1, 2]. Moreover, clinical trials examining treatment intensification in patients with locally advanced cancer have shown incremental improvements in local control and overall survival [3]. Radiotherapy outcomes are traditionally modelled using information about the dose distribution and the fractionation [4, 5] . However, it is well known that radiotherapy outcomes are multifactorial and may also be affected by multiple clinical and biological prognostic factors. For tumours these would include stage, volume, tumour hypoxia, etc. [6, 7]. The response of normal tissue incidentally and unavoidably irradiated during radiotherapy is the main factor limiting increase in prescription dose to the tumour. Optimising this trade-off, known as the therapeutic ratio, is the fundamental challenge in radiotherapy (Fig. 17.1).

The accurate prediction of both tumour response and corresponding risk of toxicity would provide patients and their treating clinicians with better tools for informed decision-making about expected benefits versus anticipated risks and higher likelihood of improved outcomes, in which machine learning (ML) methods are expected to play a prominent role (Fig. 17.2).

Typically, the 3D dose distribution to each delineated structure is characterised using a dose-volume histogram (DVH). A differential dose-volume histogram

S. Gulliford (✉)
University College London Hospitals NHS Foundation Trust, London, UK

University College London, London, UK
e-mail: s.gulliford@nhs.net

I. El Naqa
Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: ielnaqa@med.umich.edu

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_17

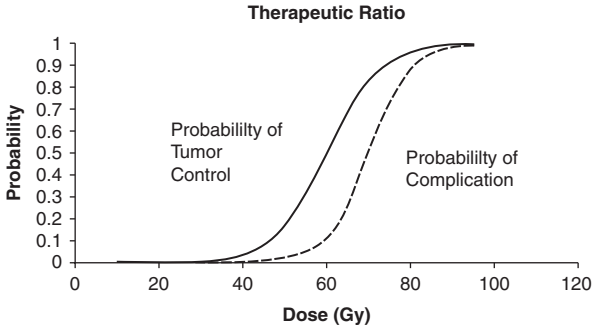


Fig. 17.1 As the dose delivered to a tumour increases, so does the probability of tumour control (TCP). However, the resultant increase in dose to surrounding healthy tissues increases the Normal tissue complication Probability (NTCP). Balancing TCP against NTCP is known as the Therapeutic Ratio and is a major goal for radiotherapy management

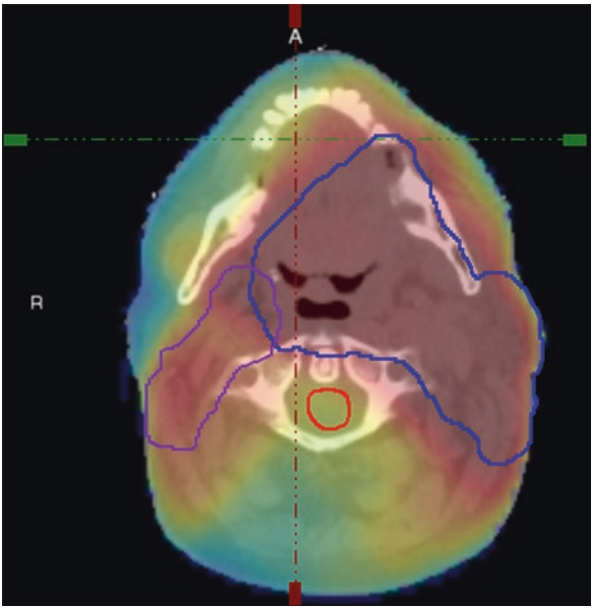


Fig. 17.2 An axial slice from a radiotherapy treatment plan of a patient treated for head and neck cancer. The Primary PTV and nodal volume are contoured. The colour wash indicates the dose distribution

reports the volume (absolute or relative) of a structure, which receives a specific dose (Fig. 17.3 top). Modern treatment planning systems usually calculate histograms with a bin width of <=0.1 Gy. More commonly histograms are displayed as cumulative dose-volume histograms where for each dose level, the volume of the organ or structure receiving at least that dose is reported (Fig. 17.3, bottom). These values are commonly reported as Vx where x is the relevant dose, e.g. V60 is the volume of a structure receiving at least 60 Gy or Dx where x is typically the relative volume, e.g. D90 is the minimum dose to 90% of a structure volume.

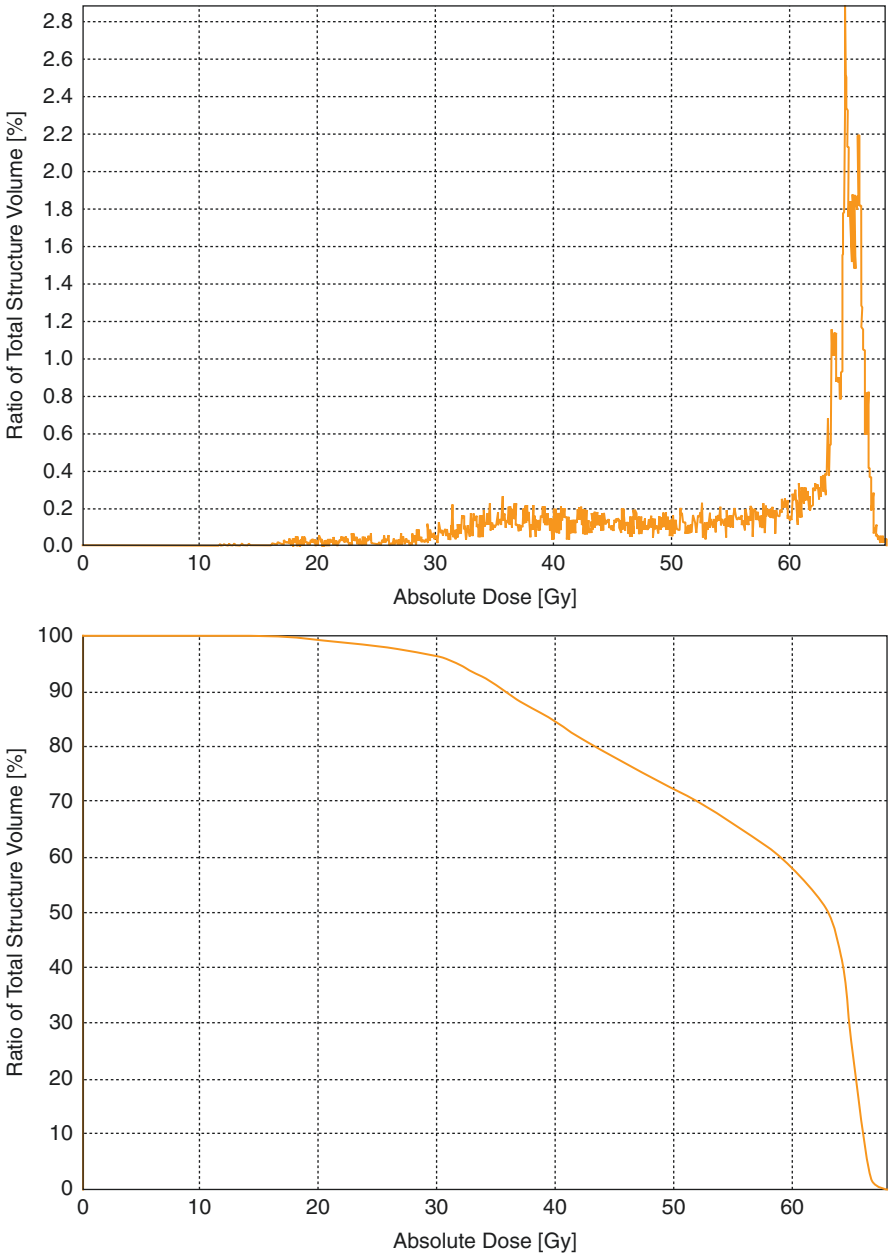


Fig. 17.3 Examples of differential and cumulative dose-volume histograms (DVH) for a normal tissue structure close to the tumour

For healthy normal tissues treatment factors will affect the dose distribution received but additional factors regarding comorbidities will also be important. Patient genetics have also been demonstrated to affect both TCP and NTCP. Recent years have witnessed the emergence of data-driven models incorporating advanced bioinformatics tools in which dose–volume metrics are mixed with other patient- or disease-based prognostic factors [8–16] in order to improve outcomes prediction [17] .

17.1.1 General Considerations

The use of machine learning (ML) is often favoured where the underlying relationship between the data and the endpoint is unknown and there is a need for future prospective evaluation of data. This is exactly the case for predicting radiotherapy response. Generally, the response of the tumour and organs at risk is related to increasing dose but will also depend on a multitude of other patient and treatment factors. For organs at risk particularly the dose distribution varies widely and is not well quantified. We prospectively evaluate every treatment plan going through the clinic and the development of knowledge-based tools to facilitate this process is highly desirable. The ability of a trained model to generalise to unseen data is imperative. Techniques to ensure this include statistical resampling methods such as cross-validation and bootstrapping, which reduce the dependency of a final model on a specific training dataset. The use of an independent (relevant) test set, to measure model performance, once the model has been finalised should also be regarded as standard practice and is highly recommended by the TRIPOD criteria (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) [18]. It is important to appreciate the extent to which the model can generalise. If a model is trained on data from a centre then a well-built model should be able to reflect the toxicity experience of that centre. However, it may not be able to predict toxicity for a similar cohort of patients from a neighbouring centre where subtle changes in treatment technique, toxicity reporting or patient demographic may render the model irrelevant.

Since the intention of radiotherapy is to keep the incidences of toxicity to a minimum the proportion of toxicity/no toxicity in the data set may be very unbalanced with only a small number of patients reporting toxicity. Whilst this is generally good news for the patient it is a challenge to model building. A number of approaches exist to try to account for this. First, the ratio of toxicity/nontoxicity cases should be standardised across training groups for example stratified cross-validation and in the independent test set. It is also possible to promote the number of cases within the dataset for the underrepresented class [19].

17.2 Tumour Control Probability

Tumour control is strictly defined by the probability of the extinction of clonogenic tumour cells at the end of treatment [20] . Several radiobiological models have been proposed in the literature to model TCP. The linear-quadratic model (LQ) is the

most frequently used model for including the effects of repair between treatment fractions. The LQ model is based on clonogenic cell survival curves and is parameterised by the radiosensitivity ratio (α/β). It is thought that it quantifies the effects of both unrepairable damage and repairable damage susceptible to misrepair after tumour sterilisation by radiation [21, 22]:

$$\mathrm{SF} = \exp\left(-\left(\left(\alpha + \beta * d\right) * D + ln2 * t / T_{pot}\right)\right) \tag{17.1}$$

where d is the fraction size, D is the total delivered dose, t is the difference between the total treatment time (T) and the lag period before accelerated clonogen repopulation begins (T_K), and T_{pot} is the potential doubling time of the cells. The ratio $\ln 2/T_{pot}$ is referred to as the repopulation parameter. Several variations of this model have been proposed including a Poisson-based [23] and a birth–death model [24]. Among the most commonly used LQ-based TCP models [25] is:

$$\mathrm{TCP} = \exp\left(-N \exp\left(-\left(\left(\alpha + \beta * d\right) * D + \ln 2 * t / T_{pot}\right)\right)\right) \tag{17.2}$$

A detailed review of analytical methods for TCP in radiation treatment has been recently published [26].

17.3 Machine Learning for TCP Modelling

Machine learning allows for exploiting nonlinear patterns in the data that may not be directly tractable from using analytical or phenomenological models. There are several steps into development of a TCP model using machine learning as shown in the examples below using dosimetric, clinical, imaging, and biological data in lung cancer.

17.4 Example 1: Dosimetric and Clinical Variables

17.4.1 Data Set

A set of 56 patients diagnosed with non-small cell lung cancer (NSCLC) and who have discrete primary lesions, complete dosimetric archives, and follow-up information for the endpoint of local control (22 locally failed cases) is used. The patients were treated with three-dimensional conformal radiation therapy (3D-CRT) with a median prescription dose of 70 Gy (60–84 Gy). The dose distributions were corrected for heterogeneity using Monte Carlo simulations [27]. The clinical data included age, gender, performance status, weight loss, smoking, histology, neoadjuvant and concurrent chemotherapy, stage, number of fractions, tumour elapsed time, tumour volume, and prescription dose. Treatment planning data were de-archived and potential dose–volume histogram (DVH) prognostic metrics were extracted using CERR [28]. These metrics included Vx (percentage volume receiving at least x Gy), where x was varied from 60 to 80 Gy in steps of 5 Gy, mean dose, minimum

and maximum doses, and centre of mass location in the craniocaudal (COMSI) and lateral (COMLAT) directions. This resulted in a set of 23 candidate variables to model TCP. The modelling process using nonlinear statistical learning starts by applying dimensionality reduction technique such as principal component analysis (PCA) to visualise the data in two-dimensional space and assess the separability of low-risk from high-risk patients. Separable cases could be modelled by linear kernels whilst non-separable cases are modelled by nonlinear kernels that allow for separability of the data but at the expense of increased dimensionality. This step could be preceded by a variable selection process and the generalisability of the model is evaluated using resampling techniques as discussed below [29].

17.4.2 Data Exploration

In Fig. 17.4a, we show a correlation matrix representation of the selected candidate variables with clinical TCP and cross-correlations among themselves using Spearman's rank correlation coefficient (rs). Note that many DVH-based dosimetric variables are highly cross-correlated, which complicate the analysis of such data. In Fig. 17.4b, we summarise the PCA analysis of this data by projecting it into two-dimensional space for visualisation purposes. The plots show that two principal components are able to explain 70% of the data and reflect a relatively high overlap between patients with and without local control, indicating potential benefit from using nonlinear kernel methods.

17.4.3 Logistic Regression Modelling Example

The multimetric model building using logistic regression is performed using a two-step procedure to estimate model order and parameters. In each step, a sequential forward selection strategy is used to build the model by selecting the next candidate variable from the available pool (23 variables in our case) based on increased significance using Wald's statistics [30]. In Fig. 17.5a, we show the model order selection using the leave-one-out cross-validation (LOO-CV) procedure. It is noticed that a model order of two parameters provides the best predictive power with Spearman rank correction coefficient (rs = 0.4). In Fig. 17.5b, we show the optimal model parameters' selection frequency on bootstrap resampling (280 samples were generated in this case). A model consisting of GTV volume ($\beta = -0.029$, $p = 0.006$) and GTV V75 ($\beta = +2.24$, $p = 0.016$) had the highest selection frequency (45% of the time). The model suggests that increase in tumour volume would lead to failure, as one would expect due to increase in the number of clonogens in larger tumour volumes. The V75 metric is related to dose coverage of the tumour, where it is noticed that patients who had less than 20% of their tumour covered by 75 Gy were at higher risk of failure. However, a drawback of this logistic regression approach is that it does not automatically account for possible interactions between these metrics nor does it account for higher-order nonlinearities.

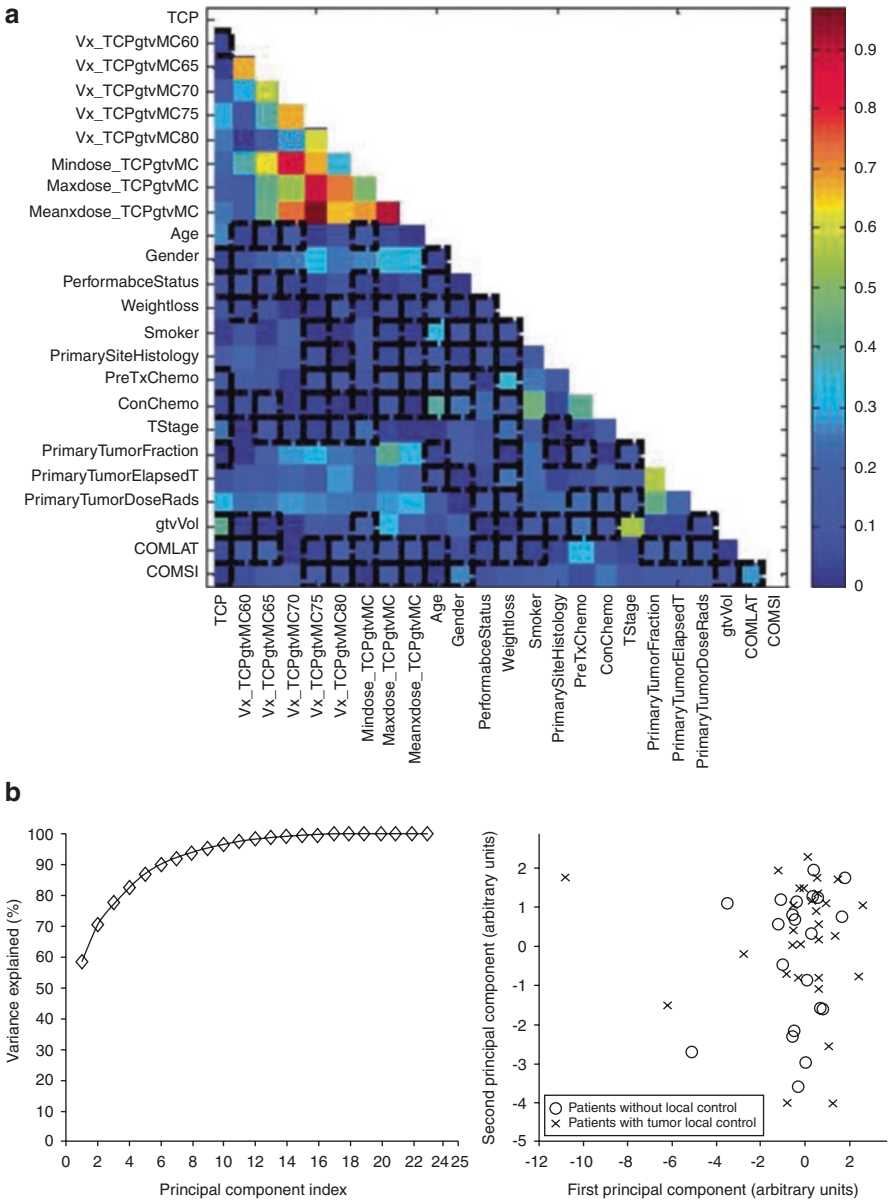


Fig. 17.4 (**a**) Correlation matrix showing the candidate variable correlations with TCP and among the other candidate variables. (**b**) Visualisation of higher dimensional data by principal component analysis (PCA). *Left* The variation explanation versus principal component (PC) index. *Right* The data projection into the first two principal component space. Note the cases overlap

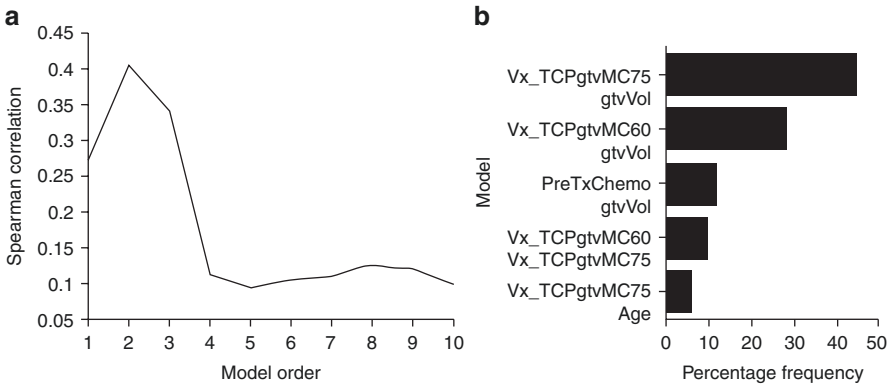


Fig. 17.5 TCP model building using logistic regression. (**a**) Model order selection using LOO-CV. (**b**) Model parameters estimation by frequency selection on bootstrap samples

17.4.4 Kernel-Based Modelling Example

To account for potential nonlinear interactions as revealed by the principal component analysis (PCA), we will apply kernel-based methods using support vector machines (SVM). Moreover, we will use the same variables selected by the logistic regression approach. We have demonstrated recently that such selection is more robust than other competitive techniques such as the recursive feature elimination (RFE) method used in microarray analysis. In this case, a vector of explored variables is generated by concatenation. The variables are normalised using the z-scoring approach to have a zero mean and unity variance [31]. We experimented with different kernel forms; best results are shown for the radial basis function (RBF) in Fig. 17.6a. The figure shows that the optimal kernel parameters are obtained with an RBF width $\sigma = 2$ and regularisation parameter $C = 10{,}000$. This resulted in a predictive power on LOO-CV rs = 0.68, which represents 70% improvement over the logistic regression analysis results. This improvement could be further explained by examining Fig. 17.6b, which shows how the RBF kernel tessellated the variable space nonlinearly into different regions of high and low risks of local failure. Four regions are shown in the figure representing high/low risks of local failure with high/low confidence levels, respectively. Note that cases falling within the classification margin have low confidence prediction power and represent intermediate-risk patients, i.e. patients with "border-like" characteristics that could belong to either risk group [29]. Klement et al. [19] describe using a SVM approach to predict TCP for early stage non-small cell lung cancers treated with stereotactic radiotherapy.

regions: (1) area of low-risk patients with high confidence prediction level, (2) area of low-risk patients with lower confidence prediction level, (3) area of high-risk patients with lower confidence prediction level, and (4) area of high-risk patients with high confidence prediction level. Note that patients within the "margin" (cases 2 and 3) represent intermediate-risk patients, which have border characteristics that could belong to either risk group

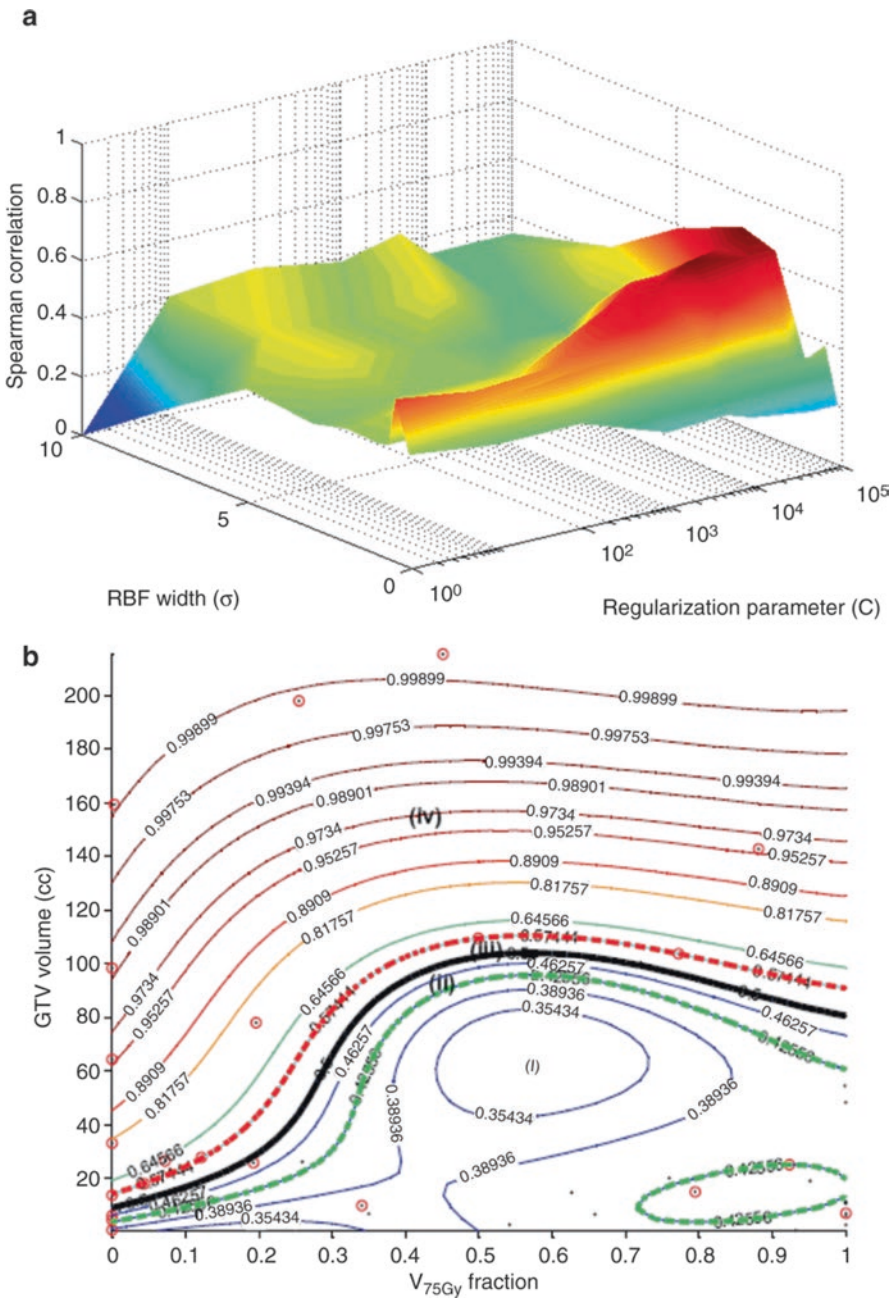


Fig. 17.6 Kernel-based modelling of TCP in lung cancer using the GTV volume and V75 with support vector machine (SVM) and a radial basis function (RBF) kernel. Scatter plot of patient data (*black dots*) being superimposed with failure cases represented with red circles. (**a**) Kernel parameter selection on LOO-CV with peak predictive power attained at $\sigma = 2$ and $C = 10{,}000$. (**b**) Plot of the kernel-based local failure (1-TCP) nonlinear prediction model with four different risk

(continued)

Forty-nine out of three hundred ninety-nine patients had a local failure after 6 months. Both under sampling and SMOTE (Synthetic Minority Over-sampling Technique) methods were used to account for the imbalance between classes. Only seven features were considered since dosimetric variables are known to be highly correlated. tenfold CV was employed, and variable selection was assessed using AUC. The final model was compared with a multivariate logistic regression and was found to have a higher area under the ROC curve AUC of 0.789 vs 0.696.

17.4.5 Comparison with Other Known Models

For comparison purposes with mechanistic TCP models, we chose the Poisson-based TCP model and the cell kill equivalent uniform dose (cEUD) model. The Poisson-based TCP parameters for NSCLC were selected according to Willner et al. work [32], in which the sensitivity to dose per fraction (α/β = 10 Gy), dose for 50% control rate (D50 = 74.5 Gy), and the slope of the sigmoid-shaped dose–response at D50 (γ_{50} = 3.4). The resulting correlation of this model was rs = 0.33. Using D50 = 84.5 and γ_{50} = 1.5 [33, 34] yielded an rs = 0.33 also. For the cEUD model, we selected the survival fraction at 2 Gy (SF2 = 0.56) according to Brodin et al. [35]. The resulting correlation in this case was rs = 0.17. A summary plot of the different methods predictions as a function of binned patients into equal groups is shown in Fig. 17.7. It is observed that the best performance was achieved by the nonlinear (SVM-RBF). This is particularly observed for predicting patients who are at high risk of local failure.

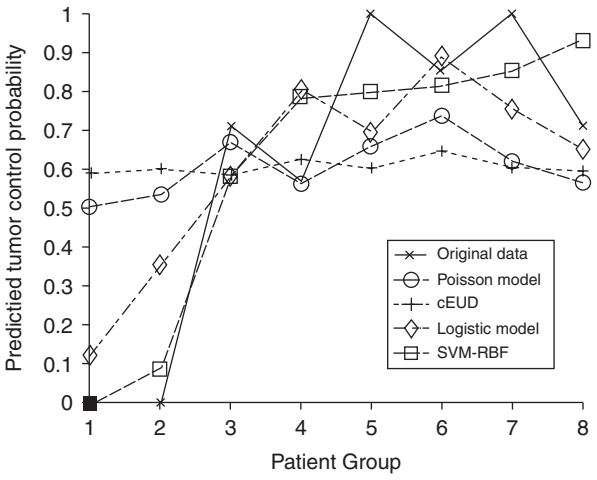


Fig. 17.7 A TCP comparison plot of different models as a function of patients being binned into equal groups using the model with highest predictive power (SVM-RBF). The SVM-RBF is compared to Poisson-based TCP, cEUD, and best two-parameter logistic model. It is noted that prediction of low-risk (high-control) patients is quite similar; however, the SVM-RBF provides a significant superior performance in predicting high-risk (low-control) patients

17.5 Use of Imaging Features

Pretreatment or posttreatment information from anatomical or functional/molecular imaging could be used to monitor and predict treatment outcomes in radiotherapy. For instance, changes in tumour volume on computed tomography (CT) have been used to predict radiotherapy response in NSCLC patients [36, 37]. On the other hand, functional/molecular imaging, in particular positron emission tomography (PET) with fluorodeoxyglucose (FDG), has received special attention as a potential prognostic factor for predicting radiotherapy efficacy [38–41]. For instance, high FDG-PET intensity has been shown to correlate with poor local control in lung cancer [42–45]. In our previous work, new features based on image morphology, intensity, and texture/roughness can provide a more complete characterisation of uptake heterogeneity [41]. Recently, we have shown that in addition to PET features, CT-derived features (from the gross target volume) may also improve prediction of local tumour response as shown in Fig. 17.8 [46].

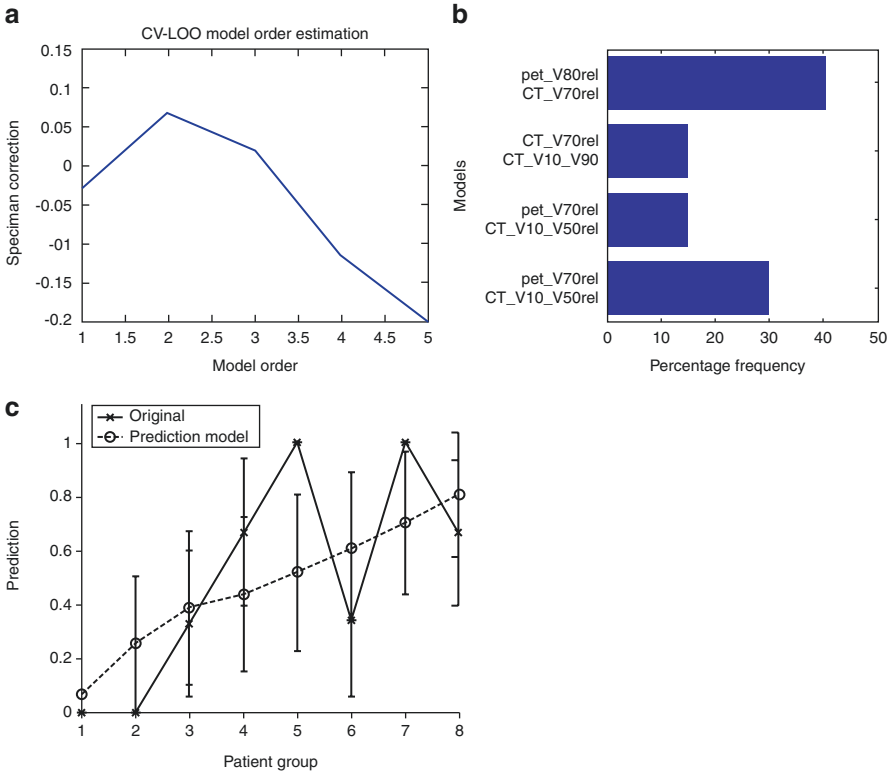
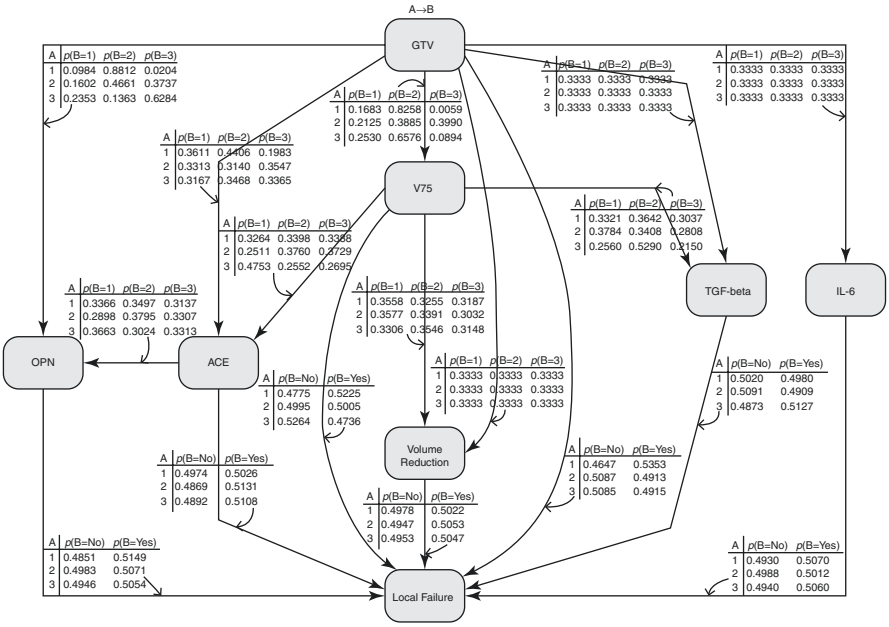


Fig. 17.8 Multimetric modelling of locoregional failure from PET/CT features. (**a**) Model order selection using leave-one-out cross-validation. (**b**) Most frequent model selection using bootstrap analysis. (**c**) Plot of locoregional failure probability as a function of patients binned into equal-size groups showing the model prediction and the original data

17.6 Use of Biological Markers

A biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention [47]. Biomarkers can be imaging biomarkers or measurements of gene expression or protein levels from tissue or fluid specimens. For instance, blood-based protein expression of hypoxia [48] and inflammation [49] were shown to be predictive of tumour response to radiotherapy. Therefore, we conducted a comparison study of physical factors, biological factors extracted from blood sera, and a combined model of local control in NSCLC patients. In order to account for the hierarchal relationship between the different variables, we utilised a graphical Bayesian network (BN) framework. A BN is a probabilistic graphical model of outcomes in which the variables (dosimetric, clinical, and biological) are presented as nodes in the graph and their conditional dependencies are represented by directed acyclic graph as shown in Fig. 17.9 [50].



	Volume reduction	IL-6	ACE	TGF-beta	OPN	GTV	V75
Low boundary	−0.39	1.64	47643	2450	12761	42.23	0
High boundary	−0.05	1.97	71805	28753	3E+07	138.11	0.16

Note: Thebinning task for each variable was performed in the following manner: values between mininnun value and low boundary were converted into 1; values between low boundary and high boudary were converted into 2; and values between high boundary and maximum value were converted into 3.

Fig. 17.9 *Top* Bayesian network with probability tables for combined biomarker proteins and physical variables for modelling local tumour control in NSCLC. *Bottom* The binning boundaries for each variable

17.7 NTCP Modelling

Although complimentary in approach, the complexity of predicting normal tissue response is a higher dimensional problem than predicting local control. The reasons for this are: (1) there are usually more than one organ at risk irradiated and protecting all of these structures requires compromise; (2) each structure responds differently to radiotherapy due to the type of cells and the structural and functional organisation of the tissue; and (3) the dose distributions to the surrounding normal tissues are inhomogeneous with gradients across the tissues commonly related to the proximity to the tumour (Fig. 17.2). This variability results in a large number of potential dose distributions to the structure. Consequently, the dose-volume relationship to toxicity is complex and not well understood.

Describing the dose distributions in order to model the response of the structure has been explored widely. The QUANTEC report published as a supplement in International Journal of Radiation Oncology Biology and Physics [51] provided a comprehensive report summarising the published data on the dose-volume response for 16 organs at risk whilst considering the limitations of the data and providing recommendations on how to improve future data collection and analysis. Commonly the dose measure is quantified as a metric such as maximum or mean dose or volume of the structure receiving a specified dose (Vx). Once developed and validated these metrics can be used prospectively as constraints during the treatment planning process. Each treatment plan is assessed prior to treatment in order to ensure safety and to evaluate the likely therapeutic success and risk of complication. In order to assess this risk, the concept of the NTCP has been developed. It is the probability that a given dose distribution to a defined tissue or structure will result in a quantifiable (unfavourable) response in the patient. The dose response of tumours to radiation is characterised using a sigmoidal response and this shape of response is translated as the basis for NTCP models. However, whereas in the case of a tumour where the dose is (ideally) homogenous, in the case of a normal tissue the dose distribution is ideally inhomogeneous with as much tissue as possible being spared. The result of this is the challenge of which metric to plot on the abscissa.

17.7.1 NTCP Models

A range of NTCP models have been developed, the most widely known and perhaps most regularly used is the Lyman Kutcher Burman (LKB) model. This model comprises an empirical model of dose response as a function of irradiated volume [52], the reduction of a dose-volume histogram to a single metric [53] and parameter fits for individual organs at risk [54] based on the tolerance doses summarising clinical knowledge by Emami et al. [55]. Originally the Lyman model was developed for particle therapy where dose distributions fall off steeply and essentially result in uniform dose D to a percentage of the organ with little dose to the remainder. The tolerance dose parameter TD50(1) or TD5(1) is the 50% or 5% probability of

experiencing toxicity where the whole structure is irradiated. The power law is employed to account for fractional irradiation.

$$NTCP = \frac{1}{\sqrt{2\prod}} \int_{-\infty}^{t} e^{-t^2/2} dt \tag{17.3}$$

where

$$t = \frac{D - TD_{50}(V)}{m * TD_{50}(V)} \tag{17.4}$$

$$TD_{50}(V) = TD_{50}(1) / V^n \tag{17.5}$$

$TD_{50}(V)$ is the tolerance dose for a partial volume V. The parameter m is the standard deviation of $TD_{50}(1)$ and n indicates the volume effect of the organ being assessed. $n = 0$ indicates a completely "serial" structure, where the maximum dose dominates outcome and $n = 1$ is a "parallel" structure where the mean dose is related to outcome.

17.7.2 Dosimetric Data Reduction-Summary Measure

In reality the dose distribution to an organ at risk is likely to be inhomogeneous. In this case a reduction is required to translate the inhomogeneous dose distribution to a single metric that results in the same radiation response as a corresponding homogenous dose distribution. The most commonly used metric is the generalised equivalent uniform dose [56]. Originally developed as the equivalent uniform dose to tumours [57] the concept was extended to include normal tissues. The formula is usually written as

$$EUD = \left(\sum_i V_i D_i^a \right)^{\frac{1}{a}} \tag{17.6}$$

where D_i is the dose in the ith bin of the DVH and V_i is the volume of tissue receiving dose D_i, a is the volume parameter and is equivalent to $1/n$.

Alternative models which consider the functional architecture of the organ/structure have also been employed [58, 59].

17.8 Machine Learning Approaches to NTCP Modelling

Machine learning brings a new toolbox to the challenges of predicting NTCP. The concept of allowing a nonlinear model to develop without an "a-priori" definition of the relationship between input variables and outcomes removes bias from our limited understanding of the response of normal tissues to radiation and enables us to uncover new information. Many of the considerations for predicting NTCP using

machine leaning are common to the different "flavours" of machine learning. As discussed, the data available includes dosimetric data, patient characteristics, previous health history, other current health conditions (comorbidities), systemic therapy (chemotherapy) and surgery. Little is known about the interaction between these different types of information and therefore the flexibility of being able to include variables without understanding higher-order interaction terms is a genuine advantage of machine learning. Many of the publications to date that predict NTCP from dosimetric variables present the data in the form Volume receiving (x) Gy or a reduction of the dose-volume histogram to EUD. The bins of the histograms for an individual patient are known to be highly correlated. Depending on the uniformity of the radiotherapy protocol for the cohort under observation, there is usually an inter-patient correlation to consider. Machine learning approaches are generally well placed to cope with such interactions.

17.8.1 Multivariable Logistic Regression

Conventionally models are obtained by fitting a sigmoidal shaped curve to a measure of dose to predict toxicity. This is achieved using data from retrospective cohorts of patients. Commonly, multivariate logistic regression [60] is performed where the model to predict probability of toxicity is comprised of coefficients describing the contribution of individual explanatory variables to the final model [61]. The outcome predicted by the model is compared to the known outcome and the error is minimised to find the optimal parameter fit. Statistical techniques of cross-validation and bootstrapping are employed to ensure generalisability of the models.

Logistic Regression assumes that the variables in the model are independent and uncorrelated. Since DVH data is neither of these, careful consideration is required on the use of logistic regression. As a result, dosimetric information is often limited to a single summary metric such as mean dose resulting in a compromise on the data included in the model. Dean et al. [62] described using penalised logistic regression to overcome the issues of correlated variables to predict acute dysphagia for 173 patients treated with radiotherapy for head and neck cancer. Fractional dose-volume histogram data for the pharyngeal mucosa was described in 20 cGy bins and augmented with clinical information such as chemotherapy and patient characteristics. Additionally, spatial descriptors of the dose distribution were also described using 3D moments and trained independently. Support vector machines (SVM) and random forests (RF) were also trained on both conventional and spatial dosimetric data. Model performance was addressed using AUC, Log loss and Brier Score. An independent dataset of 90 patients from University of Washington was available as an independent validation set. Overall the penalised logistic regression model using standard DVH metrics was not outperformed by any other model with an AUC of 0.76 on the internal validation and 0.82 on the external validation. Calibration curves were fitted to the models to present the predicted outcome compared to the known outcome. Curves can be recalibrated to improve the model fit further.

17.8.2 Feature Selection

Feature/variable selection can be regarded either as a pre-processing step or an integral part of model fitting. Where the existence or strength of correlation between individual features and toxicity is unknown a wide range of possibilities will need to be included in the original input data. It is important to also consider interactions between variables that may contribute to the predictive power of the model.

Advantages of pre-processing feature selection include reduction of model complexity, decrease in computational burden and improve generalisability of unseen data [63].

A wide range of methods for variable selection are available and a useful summary on this is found in [64]. Within the literature for predicting NTCP using machine learning undoubtedly one of the most popular is principal component analysis (PCA). Principal components are uncorrelated linear combinations of variables in a given dataset, which account for the variance in the input features in a dataset without reference to the corresponding outcome data, i.e. unsupervised learning. Ideally data with the same outcome class naturally cluster together and the clusters are separable from each other. PCA is particularly attractive feature for DVH-based analysis where variables are known to be highly correlated and has been coupled with conventional statistical models such as logistic regression as well as machine learning methodologies.

The reduction of dimensionality results in the ability to visualise high order data. One of the earliest studies using PCA to predict NTCP was published by Dawson et al. [65] who considered PCA for two different organs at risk. PCA was chosen in order to consider all the bins of a DVH without having to reduce to a single metric, such as mean dose, or summary metric such as EUD. The first cohort included 56 head and neck patients where data from the parotid glands was used to predict xerostomia (dryness of the mouth) 12 months after radiotherapy. The dosimetric data was characterised as a cumulative DVH with 1 Gy bins (84 bins in total). The first two principle components explained 94% of the variance in the DVH. When these were plotted against each other (Fig. 17.10) and labelled according to outcome class there was clear separation between the classes indicating that outcome classes were potentially linearly separable Logistic regression was applied to the first 3 principal components in addition to patient sex, age and diagnosis. Only the first principal component was significantly associated with toxicity.

In contrast to these clear-cut results the other cohort studied was 203 patients who received radiotherapy to either partial or whole liver. The first two principal components were plotted along with the Lyman NTCP model however no separation between clusters was observed. Despite this result the results of logistic regression including the first three principal components and relevant clinical factors demonstrated that only the first principal component was significantly associated with toxicity.

Following on from the work by Dawson, Bauer et al. [66] explored the use of PCA to quantify rectal bleeding in a cohort of prostate cancer patients treated with radiotherapy. The authors propose the use of a varimax rotation, an orthogonal

rotation applied to the subset of principal components that account for most of the variance in the data set. The varimax rotation maximises sparseness of the subset and only small regions of each mode (component) remain large allowing identification of specific regions of the DVH. However, the process re-introduces correlation which must be accounted for. A number of subsequent studies detailed the use of PCA to predict toxicity following prostate radiotherapy [67–69]. Sohn et al. [67] applied PCA to a cohort of 262 prostate cancer patients of which 50 patients reported late rectal bleeding CTCAE v. 3 ≥ G2. As with the previous study the bins of the cumulative DVH provided the input features however in this case the bin width was 0.1 Gy resulting in 850 variables. 96.1% of the variation was accounted for by the first three principal components.

A novel publication on the use of PCA in radiotherapy incorporates spatial information into the relationship between dosimetry and toxicity. Liang et al. [70] used (PCA) to identify patterns of irradiation of bone marrow in the pelvic region which were likely to increase acute haematological toxicity. White blood cell count nadir was used as an indicator for acute haematological toxicity in a cohort of 37 patients treated with chemo-radiotherapy for cervical cancer. The dose distribution for each patient was standardised by mapping each treatment planning CT, via deformable

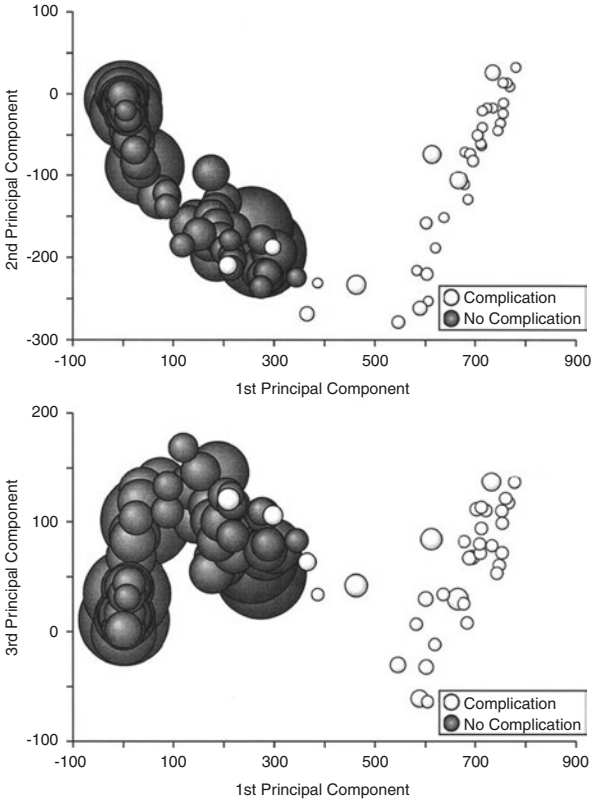


Fig. 17.10 Taken from Dawson et al. [65] Demonstrating linear separability of data describing xerostomia based on parotid gland dose distributions

registration, on to a pelvic bone template. The corresponding dose distributions were interpolated and mapped on to the template. The dose to each voxel in the standard image was calculated and considered as a predictor variable. The template ensured the same number of voxels for each patient and these voxels were sampled systematically, left-right, anterior-posterior and superior-inferior to form a row vector for each patient containing 44,146 elements. For each patient the same element referred to the same voxel. Clearly this data set would benefit from dimensionality reduction. As with some of the previous studies, since all of the variables were measured using the same scale (Gy), PCA was performed with the covariance matrix. Of the 36 non-zero eigenvalues with corresponding eigenvectors, 5 were statistically correlated with acute hematologic toxicity using univariate logistic regression. The results of the regression were used to test if the resultant dose space was related to toxicity. Acute haematological toxicity was defined by dichotomising the white blood cell nadir as <2000/μml for no toxicity (n = 23) vs. ≥2000/μml for toxicity (n = 14). Difference maps of the dose distribution were projected on to the pelvic bone template for those with/without the defined toxicity and compared with the voxels, which were shown to be statistically significant in the regression model. There was good agreement between the two assessments (Fig. 17.11). This mapping approach allowed the visualisation of important anatomical regions of active bone marrow which could be avoided using Intensity Modulated Radiotherapy (IMRT).

17.9 Classical Machine Learning Approaches

There are many flavours of machine learning as described in Chap. 3, however, most of the literature related to predicting NTCP is from the more established techniques. These can be mostly broadly separated in to supervised and unsupervised learning approaches. Conventionally, a model relates a number of variables to an outcome or classification, this is supervised learning. In contrast unsupervised learning finds patterns and groupings among the input variables only, these groupings should then naturally reflect the classification of the data. The following sections consider the use of supervised and unsupervised learning techniques for prediction of NTCP.

17.9.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are one of the classic machine learning approaches dating back to the seminal work of McCulloch & Pitts [71]. With the analogy of the way in which the human brain works it is tempting to think that the knowledge of an experienced clinician or medical physicist can be easily transferred. It has been a popular choice for applications relating to predicting the response of normal tissues to radiotherapy. One of the earliest papers was published by Munley et al. [72] who trained a feed forward, back-propagation, neural network to predict symptomatic lung injury following radiotherapy. Ninety-seven patients were included in the neural network of which 25 had a clinician assessed

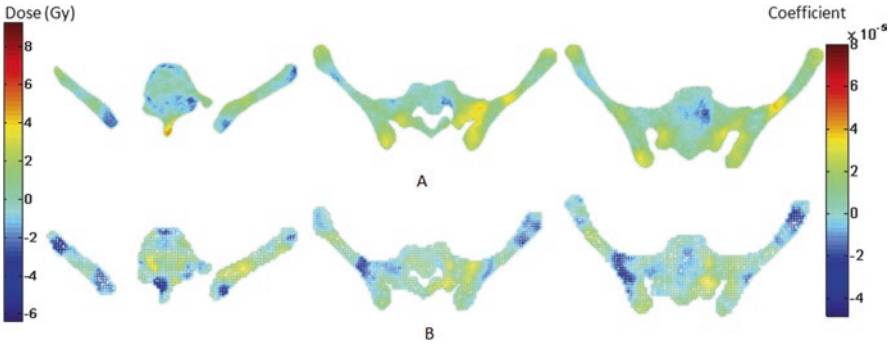


Fig. 17.11 Taken from Liang et al. [70] The top row indicates areas of pelvic bone marrow correlated to acute hematologic toxicity dichotomised as white blood cell nadir < or > /2000μml. The bottom row represents the regression coefficients produced after PCA

symptomatic lung injury. Patients from a number of tumour sites were included. Although 2/3 of patients were treated for lung tumours, the inclusion of other tumour sites increased the diversity of the dose distributions and confounding factors in the training cohort. The neural network had 29 inputs corresponding to pretreatment features Dosimetry which included dose-volume histogram reduction using both the Lyman [73] and Kutcher method [53], Volume of lung receiving 10 Gy (V10), V20, V30, V40, V50, V60, V70 and V80 and the full and effective dose to lungs and the lung volume. Each input was scaled 0–1. The architecture included 2–5 hidden nodes and a single output node each with a sigmoidal activation function. Training was performed using the leave-one-out approach where each patient case was taken out in turn and the neural network retrained. Training was terminated when the ROC analysis was maximised. The final result was an AUC of 0.833 ±0.04. This result was compared with Multivariate Logistic Regression which resulted in an AUC of 0.813 ± 0.064 and the dose-volume- histogram reduction method of Kutcher which yielded an AUC 0.521 ±0.08. The influence of each input variable was assessed by retraining the neural network with the leave-one-out approach applied to each variable and ranked by assessing the deterioration in AUC after a fixed number of iterations. The use of a leave-one-case-out approach to training the neural network is likely to result in overfitting but using a leave-one-input-out approach to investigate the contribution of individual features allowed useful insight into the prediction of toxicity.

In 2007, Chen et al. [74] reported results for a larger cohort of lung cancer patients from the same institution, Duke University Medical Centre, North Carolina. Radiation-induced pneumonitis (≥ Grade 2) was reported in 34 out of 235 patients, all of whom were treated using 3D conformal radiotherapy. ANNs were constructed using an algorithm that successively pruned and grew the input features and hidden nodes, using a training-validation cohort to assess improvement (or otherwise) of each successive iteration. To avoid local minima, weights and bias were trained from 5 randomised initial sets and the lowest error used overall. Weights were constrained to ensure reasonable responses between input variables and outcome. For

example, weights connecting dosimetric variables were constrained to have a positive value only. The authors acknowledged that this approach prohibits a complimentary subtractive effect between variables but suggest that this will safeguard against detrimental overfitting 93 potential input variables were available. Dosimetric information included V6 to V60 in 2 Gy increments and gEUD with a varying from 0.4 to 4 in increments of 0.1. The mean dose to the heart was also included. Since many of the dosimetric variables are highly correlated the training rules ensured that once a variable had been incorporated into the model no other highly correlated variables (>0.95) were eligible for inclusion in the model. The inclusion of non-dosimetric variables was justified by citing previous analysis of Normal Tissue response which was shown to be modified by interaction with chemotherapy [75] and age [76]. A wide range of non-dosimetric variables, similar to the previous publications were included covering patient demographics, treatment information and pre-radiotherapy assessment of lung function. A ten-fold cross-validation approach was used to ensure that the results were generalisable. Whilst a second approach using all patient data for training was developed for prospective testing. Leave-one-out analysis was used on this second architecture to assess the influence of individual chosen variables. Comparison of models was performed using ROC analysis. For the ANN trained using cross-validation the optimised architecture containing only dosimetric variables resulted in an ROC of 0.67 for the independent test. When non-dosimetric variables were added to the model construction this improved to 0.76. Each of the ANN developed using cross-validation contained different variables however the authors highlight that often highly correlated variables were represented in each model. The model trained for prospective testing included 6 variables, V16, gEUD a = 3.5, gEUD a = 1, forced expiration volume in 1 second (FEV1,) Carbon monoxide diffusion capacity in lung (DLCO%) (both of which were assessed prior to radiotherapy) and induction chemotherapy. All input features except FEV1 and induction chemo were shown to be individually statistically significant. It is clear from these results that different parts of the dose distribution were included in the final model despite dosimetric correlation being constrained. This result suggests that different parts of the dose distribution are important in predicting toxicity. We will consider this again with later publications.

To date we have considered neural networks where features from the dose distribution have been based on the cumulative dose distribution. The disadvantage of using dose-volume histograms is that all spatial information is discarded. It is known that each organ at risk has internal structure and function and that this is important for both damage and repair. There are a number of ways to incorporate spatial information into prediction of normal tissue toxicity. One example is the paper by Büttner et al. [77] where a dose surface map of the rectum was used as to provide the input features to an ensemble of neural networks which predicted rectal bleeding following prostate radiotherapy. A dose surface map is generated by unfolding the cylindrical structure of the rectum outlined in the treatment planning system. A number of unfolding methodologies have been suggested. In this study a slice wise method was chosen whereby the rectal contour outlined on each slice of

the treatment planning CT was virtually unfolded by cutting at the most posterior point. The maps were normalised on a slice-by-slice basis to produce maps as shown in Fig. 17.12. Since the dose in adjacent pixels is correlated four locally connected neural network architectures were constructed. The first connected a row of 3 neighbouring pixels to each node in the hidden layer with an overlap of 1 pixel. The second connected a 3x3 group of pixels to the first hidden layer where a group of 4x4 nodes was connected to a second hidden layer. In the third architecture a group of 3x3 pixels were connected to the first hidden layer. These nodes were connected to the second hidden layer row by row with no overlap. Finally, in the fourth architecture the input nodes were connected in the same way as the second architecture i.e. 3x3 group of pixels linked to the hidden nodes. The weights between each group were shared making the presumption that a global dose response could be modelled. In comparison a fully connected ANN using the dose surface histogram values, i.e. the area of the DSM receiving x Gy was constructed with 35 inputs characterising the dose between 5 and 73Gy.

An ensemble approach [78] was employed to train the ANN based classifier. An ensemble is a group of independently trained ANN each of which contributes to the output prediction. Ensembles should be less susceptible to overfitting and "choosing an unrepresentative" local minima. In this study an ensemble of 250 ANNs were constructed. Each ANN was trained using a different sample of cases from the training data with independent initialisation of the weights in each ANN. Since the incidence of rectal bleeding was relatively low (53/329 patients) 20% of the patients who did not report rectal bleeding and 75% of the patients who did report rectal

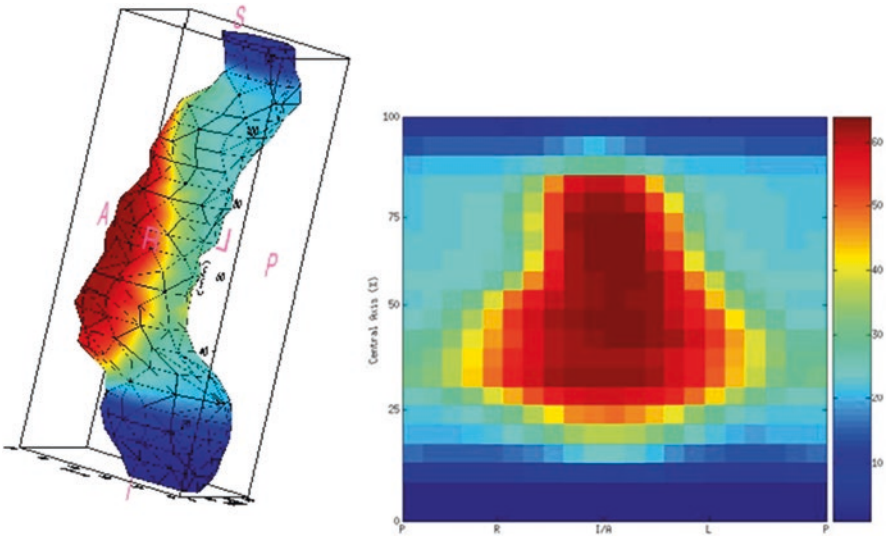


Fig. 17.12 Example dose distribution to the rectum shown as a mesh based on the contours delineated on the treatment planning CT and as a slicewise-unfolded, normalised Dose Surface Map (DSM)

bleeding were randomly chosen for each ANN. Expert ensembles were developed by sequentially adding ANN and evaluated using the area under the ROC curve for predictions on a subset of patients from the training set. If the AUC improved when predictions from the newest ANN were added, then the ANN was added to the ensemble. This process was repeated 3 times and ANN that were included in all 3 ensembles were incorporated into the expert ensemble. This whole process was repeated for each fold of the ten-fold cross-validation.

Architecture 2 was shown to produce the best predictive results with an AUC of 0.61 for all ANN and 0.64 for the expert ensemble this was compared to AUC of 0.59 for the dose surface histogram-based ANN. In order to assess the influence of the data partition resulting from cross-validation the cross-validation partitioning was repeated 100 times and the most promising locally connected architecture (2) retrained. The mean AUC was 0.65 ± 0.017.

Compared to other studies the AUC is relatively low. However, the improvement in the AUC when spatial information was incorporated suggests that using spatial information improves the input information and that overall shortcomings may well be due to a lack of non-dosimetric data or the fact that the radiotherapy dose distribution (either DVH or DSM) from the treatment planning scan is not representative of the actual dose distribution received by the patient over the course of the fractionated treatment.

17.9.2 Support Vector Machines (SVM)

Support Vector Machines are a class of machine learning that attempt to find a boundary plane that separates two classification outcomes in feature space. When the cases are linearly separable this is relatively straightforward, however, more often than not when considering prediction of normal tissue toxicity, the cases are not linearly separable. In this situation the variables can be transformed into a higher dimensional feature space where the cases may be separated by a hyper-plane. This is achieved using a nonlinear kernel such as a polynomial or radial basis function. Each data point represents a vector of the variables included in the model. The dual optimisation of separating the cases whilst improving fitting accuracy results in a balanced trade-off. This is computationally intensive to solve however it is possible to characterise the prediction function using only a subset of training data. The cases used to define the boundary between classes are known as support vectors. Unlike other approaches to machine learning SVM maximise the distance between the two classes rather than minimising the mean- square error and it is permissible for a defined number of cases to be on the "wrong side" of the boundary. The framework of a SVM implicitly includes higher-order interactions between variables without having to pre-define what they are.

In a publication complimentary to their work using neural networks (discussed in the previous section), Chen et al. describe using support vector machines to predict pneumonitis [79] on the same dataset reported for ANN [74]. A radial basis kernel function was chosen for the SVM in preference to a sigmoid or polynomial

kernel as the increase in free parameters might result in overfitting. SVM were constructed using only dosimetric variables and separately with all available variables. Parameter values C and σ were pre-determined using a grid search. Variable selection was performed using a similar approach to the ANN study whereby variables were added and substituted iteratively employing ten-fold cross-validation. Although each of the 10 results was independent there was a large crossover between the input variables selected. This level of consistency between folds is reassuring for generalisability. The AUC for the SVM including dosimetric and non-dosimetric variables was 0.76. A LOO out approach was employed to investigate the importance of individual variables in the SVMall model. The AUC was reduced by 0.19 with the exclusion of EUD and by 0.09 for induction chemotherapy. The importance of these two variables was consistent with results from the previous ANN study. However, the contribution of other variables demonstrates the risk of overfitting if techniques such as cross-validation are not employed.

El Naqa et al. describe the use of nonlinear kernel-based approaches for predicting Normal Tissue Toxicities [63] highlighting the challenges of mixed models built from different data types including dosimetric metrics, patient characteristics and disease/treatment based prognostic factors. They recommend the use of kernel-based methods, specifically support vector machines citing the following advantages over other machine learning approaches. Ability to adapt to artificial intelligence, ability to avoid excessive over-fitting, maintain computational efficiency of classical statistical methods and in summary state that SVM overcome the stigma of a black box due to rigorous mathematical foundations. Pre-processing of the data is achieved using PCA which also allows visualisation of the higher dimensional data. Examples from two clinical datasets were presented. The first was a small cohort of 55 head and neck cancer patients where a model is developed to predict xerostomia which results from a lack of salivary production following radiotherapy It was observed that the groups of patients with and without xerostomia were reasonably separated and it was subsequently demonstrated that a linear kernel produced a model which was not bettered by either Radial Basis function or Polynomial Kernel. The authors comment that this is "not the norm in radiotherapy" as exemplified by the second dataset presented. Data for 219 patients treated with radiotherapy for Non-small cell lung cancer (NSCLC) were used to predict Radiation Pneumonitis (RTOG Grade 3). Dosimetric characterisation of the dose to the lung was achieved using volume receiving x Gy (Vx). Vx with increments of 10 Gy from 10 to 80 Gy were included. Using these variables, it was demonstrated that the classes could not be separated using PCA. Using SVM it was demonstrated that an improvement in model performance was observed with increasing order of polynomial. A separate model was developed which included non-dosimetric variables including patient, disease and treatment variables. In addition, the dosimetric descriptors were expanded to include Dx (the volume of lung receiving a minimum dose x). In total 58 variables were included. The top 30 variables were selected using recursive feature elimination SVM. Variable pruning was used to account for multi-colinearity of correlated variables. The model resulted in a Matthew's correlation coefficient (MCC) of 0.22 and contained 6 variables. A further SVM was

developed using 3 variables discerned from a previous study using model order selection with resampling logistic regression. The resultant SVM with a Radial Basis function Kernel had an MCC of 0.34. The improvement in this value is attributed to the ability of a SVM to account for interactions between model variables.

In a subsequent, more comprehensive publication El Naqa et al. [80] expand on the data presented. Often in radiotherapy the incidence of complications can be quite low. Conventionally an SVM cost function treats the two potential classes equally however to account for the imbalance between classes; different weights can be assigned to samples in the two different classes with a higher penalty weight assigned to the underrepresented class.

As such the penalty term is expanded to

$$C^{+}\sum_{i=z+}\xi + C^{-}\sum_{i=z-}\xi \tag{17.7}$$

In addition to the datasets studied in the previous publication data predicting acute esophagitis in a cohort of 166 NSCLC patients was also presented. Finally, data from a multi-institutional RTOG study (9311) was used as an independent validation set to predict radiation Pneumonitis. As previously reported the best model to predict xerostomia was a linear classifier which yielded an MCC value of 0.64. The model to predict Esophagitis included concurrent chemotherapy and dosimetric information in the form Vx. No pre-model variable selection was performed. Optimal performance was achieved using a radial basis function with $\sigma = 2$ and $C = 100$ and yielded an MCC of 0.43.

The advantages of using an ensemble of Support vector machines is explored by Schiller et al. [81]. Using the Radiation Pneumonitis data from WUSTL the difference in AUC for differing size of Ensembles of SVM were compared using students' t-test. The results indicated that the AUC was statistically significantly improved for larger ensembles.

17.9.3 Unsupervised Learning SOM

Self-organising maps are an unsupervised form of artificial neural network. Unsupervised learning clusters similar data together based on the input features with no reference to corresponding output data. Similar to PCA, self-organising maps reduce the dimensionality of data. Proposed by Kohonen [82], self-organising maps are regularised grids of neurons which are trained by adapting weights. Each neuron contains information on the physical location and the weights which can be considered as typical values of the input features for that neuron. Neighbouring neurons will be more similar than distant nodes Once trained, subsequent cases are mapped on to the SOM by finding the neuron with the most similar weights. The weights can be initialised randomly however the process may be speeded up by performing PCA and using the first 2 principal components to initialise the weights Unlike PCA the use of self-organising maps to predict normal tissue complication probability is very sparse. The most prominent example is the study by Chen et al.

[83] which is complementary to their studies using ANN and SVM. As with previous studies, two models were developed SOM_{dose} which included dosimetric variables, and SOM_{all} which also incorporated the non-dosimetric variables such as chemotherapy status, tumour information and baseline lung function. Once the weights in an SOM are initialised each case is presented in turn to the map. Two parameters which steer the learning of the SOM are neighbourhood distance and learning rate. The neighbourhood distance defines the acceptable difference between the weights of an input and the weights associated with each neuron in order to decide if the patient case belongs to a particular node. In this study similarity was assessed using the Euclidean distance. The other parameter is the learning rate which in the context of SOM defines how much information from the input vector (i.e. how many of the input variables) are used in training. Once a case has been assigned to a neuron the associated weights are updated and the process repeated iteratively. In this study a ten-fold cross-validation approach was used. A map of 4x3 neurons was found to be optimal and input variables were included using trial and substitution. The resultant AUC was 0.67 for SOM_{dose} and 0.73 for SOM_{all} . The difference between the two AUC was shown to be statistically significant ($p < 0.05$). The influence of the cross-validation groups was tested by repeating the splitting of the data 200 times and retraining the SOMall model. Remarkably the AUC was 0.724 (SD = 0.017) suggesting a very consistent outcome. EUD with $a = 0.9.1$ and 1.1 chemotherapy, histology and tumour location were commonly selected variables. As mentioned previously EUD with a = 1 is mean dose which has been previously considered as being predictive of radiation pneumonitis. When this variable was removed from the model the decrease in AUC was shown to be statistically significant. The only other variable shown to produce a statistically significant decrease on exclusion was chemotherapy. These results are consistent with the two other publications by the same group.

17.9.4 Bayesian Networks

Bayesian Networks have become a popular statistical approach to challenging non-linear problems. Bayesian Networks are presented using directed acyclic graphs which summarise the joint probability distribution between a set of variables. The network is optimised by finding the conditional probabilities on each node which best represents the dataset. Oh et al. [84] describe using a Bayesian network to detect interaction of dose-volume-related parameters to predict radiation pneumonitis. The dataset comprised information on a cohort of 209 patients treated with radiotherapy for non-small cell lung cancer. Forty-eight of the patients were subsequently diagnosed with radiation pneumonitis. Input features included clinical features and dosimetric features characterised as Vx and Dx (minimum dose to the hottest x% volume). In all 160 features were available and the first step was to reduce the number of variables in the model. Information gain-based approach was employed for feature selection. Subsequently, the number of input features was reduced to 43. The Bayesian classifier assigns each case to the class with the highest

posterior probability, determined by Bayes Theorem. We have discussed previously that dose-volume data is highly correlated; however a naïve Bayesian classifier presumes that all features are mutually independent. Therefore, Oh et al. also implemented a tree augmented naïve Bayes classifier which allows connections between features, to overcome this challenge. Given the potential number of networks that may exist for a given dataset it is not feasible to find an exact solution and approximate solutions are usually employed. In this case both hill climbing and the K2 algorithm with random ordering were implemented with the maximum number of parents allowed on each node equal to three. The Bayesian networks were evaluated using the BDe score metric [85]. tenfold cross-validation was employed and each network assessed after 30 iterations. The performance of each network was assessed using Matthew's correlation coefficient. There was reasonable consistency between the different models with MCC between 0.25 and 0.3. Unexpectedly the tree augmented naïve Bayes classifier was reported to be inferior in predictive power to the naïve Bayesian classifier. One of the advantages of a Bayesian classifier approach is that it is inherently visual and therefore relationship between variables can be observed. In this study the dosimetric features relating to the heart and lung were shown to be clustered separately. Demonstrating that not only is there a relationship between heart and lung but also between the variables for each organ.

17.9.5 Decision Trees

Decision trees are constructed using recursive partitioning analysis which optimises successive dichotomisation of input variables resulting in a tree-like structure used for classification. Each tree is "grown" by starting at a root and splitting the training cases into two, maximally separated, classes. This branching continues until a terminal node (leaf) is reached. Each leaf has an associated probability of being assigned to a specific class. In the case of NTCP this is the probability of experiencing a defined toxicity. Once trained, prospective cases can be tested, by following the appropriate path along branches eventually ending at a leaf.

Das et al. [86] describe using decision trees to augment prediction of the classic Lyman NTCP [52] by producing a combined prediction as shown in Fig. 17.13. Using the same dataset as described previously by Chen et al. [74, 79, 83] decision tress with potential dosimetric and non-dosimetric factors were built using ten fold cross-validation with a balanced representation of cases experiencing radiation pneumonitis in each fold. The model was constructed using the AdaBoost algorithm which sequentially increases the number of weighted predictive units in the model. The first predictive unit contained only the Lyman model, subsequent predictive units contained both the Lyman model and a decision tree. The predictive error ε for each predictive unit was calculated as the sum of individual patient errors (deviation from binary outcome) multiplied by patient weights. The weight for the predictive unit and patient weights were updated and propagated to the next iteration. The success of the split was assessed using the Gini index split threshold criterion [78] which was expressed in this study as:

$$Ns\left(1-p_{inj,s-}^{2}p_{uninj,s}^{2}\right)-N_{L}\left(1-p_{inj,L-}^{2}p_{uninj,L}^{2}\right)-N_{R}\left(1-p_{inj,R-}^{2}p_{uninj,R}^{2}\right) \quad (17.8)$$

where S refers to the node being split, L and R refer to the left and right branches, N is the number of cases and p is the proportion of patients. The subscript *inj* refers to patients who reported radiation pneumonitis and *uninj* refers to patients who did not. The variables were ranked best to worst based on the Gini index. Only those variables with a Gini index >80% of the maximal Gini reduction were included in the model. Only three nodes were allowed on the decision tree in each predictive unit to avoid overfitting. Direction rules were implemented for a subset of variables to ensure that splits were logical for example as dose variables and disease stage were forced in a positive direction, i.e. higher value associated with increased risk of injury. AUC was used to assess the predictive accuracy of the model as successive predictive units were added. It was demonstrated that there was no further increase in AUC after 11 units. This model resulted in an AUC of 0.72 compared to predictions made solely using the Lyman NTCP model which yielded an AUC of 0.63.

A simplified model was constructed (Fig. 17.13) where the Lyman NTCP value was combined with the value on the appropriate terminal node to provide an overall predictive value. This simplified model was shown to have an AUC of 0.75 and included the use of induction chemotherapy, histology (squamous vs. other), gender and number of fraction per day. More recently Palma et al. [87] used recursive partitioning analysis to predict radiation pneumonitis on a cohort of patients identified from an international meta-analysis. Data from 836 patients who underwent concurrent chemo-radiation therapy for non-small cell lung cancer (NSCLC) from 12 different institutions in Europe, North America and Asia were collected. Patients were randomly assigned to either training or validation groups (2/3 vs. 1/3). Initially univariate logistic regression was used to identify input features that were predictive of radiation pneumonitis. These features were independently assessed using multivariate step-wise logistic regression and recursive partitioning analysis. The incidence of radiation pneumonitis was reported as 29.8% which was scored using a number of different scoring schemes where in each case Grade 2 or greater was counted as a radiation pneumonitis event.

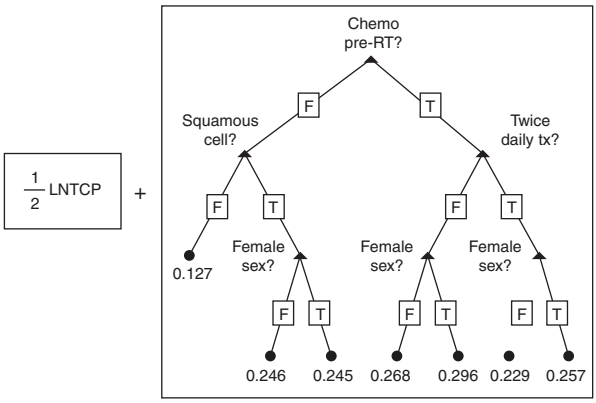


Fig. 17.13 Decision tree example taken from [86]

Chemotherapy regimen, Age > 65 years, V20, and mean lung dose were the variables used in the recursive partitioning model which defined 3 risk groups. A statistically significant difference between the risk of pneumonitis between the risk groups was observed for both the training and validation cohorts. The results of this study are strengthened by the inhomogeneity in the dataset although no quantification is made of predictive accuracy for comparison with other model-based studies.

Valdes et al. [88] also describe using decision trees to predict radiation pneumonitis. Univariate analysis was performed using the gini index. Multivariate decision trees were compared with RUSboost algorithm which is designed to overcome the challenges of class imbalance and also a random forest classifier (see below). The classification was evaluated using the F1 score.

17.9.6 Random Forests

Although decision trees are generally regarded as being highly interpretable, they are not necessarily the highest performing classifiers. This can be improved by creating an ensemble of decision trees known as a random forest. In this way the voting of individual trees is aggregated to make the final prediction. Dean et al. [89], in a study complimentary to the one described in the logistic regression section, describe the prediction acute mucocitis on a cohort of 183 head and neck cancer patients using random forests, support vector machines and logistic regression. Model performance was assessed using AUC and calibration curves. The random forest model produced an AUC of 0.71 which was similar to the SVC and penalised logistic regression but an improved calibration. Additionally, importance factors, derived from the Gini index provide information on the contribution of individual variables to the model. In this case the volume of oral cavity receiving intermediate and high doses was identified.

17.9.7 Hybrid Models and Comparative Studies

Each of the models here has shown strengths and weaknesses. None has been shown to be the perfect predictor. The question is whether an improvement can be made by combining predictions from different models to give "the best of both worlds". A useful illustration of this is the paper by Das et al. [90] who suggest that fusion of predictions from disparate models obtain a more realistic and robust estimate of the ground truth and that, where consensus exists between models this reinforces the predictions. The results of four previous studies discussed earlier in this chapter are combined to give a consensus prediction of the risk of radiation-induced pneumonitis using predictions from independently trained Decision Trees, Neural Network, Support Vector Machines and Self Organising Maps. Each model incorporated dosimetric and non-dosimetric features from the same pool of available input variables, individual reports [74, 79, 83] demonstrated that no two models chose the same set of variables. In this study the prediction of each model was averaged to generate an

analogue prediction value. One hundred random divisions of the data into ten-fold cross-validation was used to make predictions from each of the model types. These outcomes were converted to a binary value of 0 (no toxicity), and 1 (toxicity) prior to averaging to account for differences in scaling between the outputs of each type of classifier. These results were combined to produce an analogue prediction which was averaged over the 4 models. The resultant model was shown to have an AUC which converged at 0.79 when 10 randomly selected predictions were chosen for each model, this was an improvement on the results of each of the individual classifiers. The spearman correlation between any two of the predictions for each model was shown to be high ≥0.9 for all models except SVM whilst correlations between models was much lower. This emphasises the benefit of repeated cross-validation and the combination of different classifiers. The importance of individual input features was tested using reverse rank method whereby the patient predictions were ranked highest to lowest risk of pneumonitis based on the consensus prediction. The values of one input variable were then reverse so that the value for the top ranked patient was substituted with the bottom ranked patient and vice versa. The predictions were recalculated and the ranking recalculated. The spearman correlation coefficient was used to compare the pre and post switch rankings (which were resampled 10^5 times). A large negative coefficient would indicate a large impact on the predictions from the variable in question. As with previous publications highly correlated variables (Pearsons coefficient > 0.9) were excluded from being added to a model where another correlated feature was already present. Therefore, groups of dosimetric variables were grouped together. The largest negative coefficient was observed when 2 groups of dosimetric variables and induction chemotherapy were reversed. Female gender and squamous cell histology were also shown to be important. The dosimetric groups represented I EUD (a 0.5–1.2) & vol > 20—30Gy and II EUD (a 1.2–3). Subsequently the consensus variables were fitted to a logistic regression probability function. This translation of the consensus of machine learning into an easily interpretable model enables the transfer of learned knowledge in to the clinical context.

Nalbantov et al. [91] combined predictions from ten different models to predict radiation-induced acute dysphagia (swallowing difficulties). Each model was assigned equal voting rights and tested on a prospective cohort of patients. The results were compared to predictions made by physicians. All were given the same "input" information which included age, gender, WHO performance status, mean and maximum dose to the oesophagus, overall treatment time and concurrent/sequential chemotherapy. Predictions of acute dysphagia ≥G3 (CTCAE) [92] were made using Naïve Bayes, Bagging, Bayesian Networks, Boosting, Penalised Logistic Regression, Radial Basis Function network, Random Forest, Linear Support Vector Machine and LASSO and for a combined model with equal voting rights. The combined model resulted in a higher AUC (0.77) for the independent prospective validation cohort than for any of the individual models. The corresponding AUC for the physicians was 0.53.

Other studies have chosen not to create hybrid models but have made a direct comparison between Machine Learning approaches. Pella et al. [93] presented a

comparison between models based on ANN and SVM to predict acute toxicity for a cohort of 321 patients who received prostate radiotherapy. Both techniques were chosen for the flexibility that allows both dosimetric and clinical variables to be considered in the same model. The input features were selected by the authors based on clinical knowledge and appear to be limited compared to those in other studies we have considered. The dose distribution to the rectum was quantified by the dose received by 30% and 60% of the rectum (D30 and D60 respectively) and the absolute volume (cc) of rectum on the planning scan. The dose distribution to the bladder was described using only the dose received by 50% of the bladder and the absolute bladder volume (cc) from the treatment planning scan. Unusually, a single outcome of either GI or GU toxicity $\geq$ grade 2 was used, this choice was justified by the perceived low incidence of both GI (37%) and GU (11.5%) toxicity in the cohort. The Artificial Neural Network architecture was optimised using a genetic algorithm. The optimised ANN was reported to have two hidden layers with 47 neurons in the first hidden layer with a sigmoid activation function and 44 neurons also with a sigmoid activation function. A linear activation function was used in the output layer. The ROC for the optimised ANN was 0.697. In comparison the optimal SVM was found to have used a polynomial kernel of the ninth order which resulted in an AUC of 0.717. Both of these values related to a subset of 30 patients withheld from training. It should be noted that the optimisation of both ANN and SVM chose parameters that could lead to overfitting. An ANN with 13 inputs but nearly 100 hidden nodes is likely to be over-fitted as is an SVM using a ninth order polynomial. Since no cross-validation was employed it is impossible to infer how well these models would generalise. No statistical comparison was made between the AUC for the two techniques this may be again due to the singular nature of the result. Another study by Oh et al. [64] directly compares machine learning methods for outcome prediction for radiation pneumonitis. Comparison is made between both feature selection techniques and classification methods. The feature selection methods were SVM-Recursive Feature Elimination, Correlation based feature selection chi-square feature selection and information gain. Classifiers included SVM, Decision Tree, Random Forest and Naive Bayesian. Matthews's correlation Coefficient was employed to assess performance. Each method was tested on a cohort of 209 patients NSCLC patients from Washington University school of Medicine of whom 48 reported radiation pneumonitis, (which was also reported in the study of Bayesian Networks from the same group). Data included clinical variables including demographics and diseases stage and dosimetric variables quantified as Vx volume receiving x Gy and Dx dose received by x% of volume. Some input features were ranked highly by more than 1 feature selection approach but generally there was significant variability between feature selection methods. The feature selection was combined with each of the classification methods were starting with the highest rank variable models and subsequently increasing the number of variables. It was observed that SVM with a radial basis function or polynomial kernel function consistently resulted in the highest Matthews correlation coefficient values. Whilst caution is needed when comparing models since results may be data specific it is useful to consider the relative success of different approaches. Of note is the variability in

the results of the feature selection, It is not stated if any adjustment was made for correlated inputs which may have affected the results.

17.10 Deep Learning

Many of the papers described here were published in an era when machine learning was not in favour, in recent years this has changed dramatically and machine learning, particularly deep learning is now ubiquitous with modern life. Deep learning neural networks are now a commonly used tool and have their roots in the MLP. Chapter 4 describes this method in detail. The paper by Buettner described in the ANN section describes a handcrafted approach which is now automated in the concept of locally connected MLP. Wang et al. [94] describe using a deep perceptron network was also used to predict outcomes for lung cancer patient using varied data from a hospital electronic health record (EHR) systems. This ensemble network was trained using an automated approach to tune hyper parameters and a snapshot ensembles restarting strategy which takes snapshots of local minima to speed up training time for the ensemble. Multi objective optimisation utilising a pareto frontier was used for model selection. Twenty variables of different types were extracted from the EHR of 1007 patients treated with radiotherapy for lung cancer. Interestingly none of the characteristics related to the details of the radiotherapy treatment. Outcome at 1 year was predicted and compared to a classic support vector machine, a deep neural network (DNN) and a multi objective model. The multi-objective deep learning approach MoeDL was found to have the highest overall predictive metrics. This complex architecture was applied to a relatively simple dataset, deep learning is often associated with complex data such as images. Liang et al. [95] describe using a convolutional neural network approach to relate the 3D dose distribution to radiation-induced pneumonitis. The model is compared to results from a previous publication on the same dataset using a penalised logistic regression model with a more conventional characterisation of inputs of dose/volume parameters and dosiomics which are described in the next section.

17.11 Radiomics and Dosiomics

Radiomics is the term used to describe quantitative features extracted from a grey scale medical image such as CT in order to characterise spatial information, [96, 97]. These imaging biomarkers can be used for studies exploring classification and delineation of tumours and prediction of outcomes [98, 99]. A full exploration is beyond the scope of this chapter however the subsequent transfer of the concepts to 3D dose distributions have resulted in an interest in prediction of radiation response. The imaging biomarkers are intensity, texture and morphological descriptions of the voxels of the image such as the grey level run length matrix (GLRLM) which describes the number of runs of pixels of a particular grey scale value and defined length in the 3D image. This is one of the features used by Liang et al. in their

previous publication which uses penalised logistic regression to predict radiation pneumonitis. Gabrys et al. [100] have also detailed using Dosiomic features based on 3D moments to for multivariable NTCP models to predict xerostomia using a multitude of feature selection and classifier combinations. However, the highest AUC were obtained from univariate analysis of features related to the dose gradient across the parotid gland.

17.12 Radiogenomics

One of the many types of information which may influence the toxicity experienced by patients is the unique genetic code. In recent years there have been a number of genome wide association studies (GWAS) which have tried to identify genetic markers individual variations known as single nucleotide polymorphisms (SNPs) which are likely to indicate a difference in radiation response [101]. Since there are millions of potential SNPs in the genome extremely large datasets are preferable. However Cui et al. [102] demonstrated that combined careful consideration of feature selection and a-priori knowledge of likely candidates can be utilised to make prediction of radiation pneumonitis using a modest-sized datasets. The paper is a comprehensive exploration of feature selection, classical machine learning and deep learning approaches. Two hundred thirty features of interest were initially selected with a minority relating to dosimetry (5) and clinical factors (13) the remaining factors were measured cytokine levels, SNPs and microRNAs. Conventional feature selection approaches were compared with a variational autoencoder (VAE) approach which is an unsupervised deep learning concept whereby information is reduced and then expanded. The results were used as inputs to classic machine learning approaches of MLP, SVM & RF. Additionally, a joint VAE and MLP architecture was employed. Finally, a hybrid system which took the feature selection and VAE selection as input to a classic classifier was developed. Feature selection was explored with nested cross-validation and the top p% method which selected features which most frequently ranked highly. Overall the highest AUC (0.831) was obtained from the hybrid approach of feature selection using both weight pruning and latent variables which was then trained with an MLP. Of the 22 features included in the final model 7 were identified in every MLP model and included mean lung dose. This is an impressive result since it is known to predict for pneumonitis but was one of 230 input variables.

This field is advancing rapidly, and undoubtedly further publications will be available to explore the topic further [103, 104].

17.13 Challenges Modelling Radiotherapy Response

Despite many studies on large, high quality datasets, predicting NTCP remains a challenge. There are many potential reasons for this as discussed in the following.

1. In addition to the dosimetric response of normal tissues many other factors contribute to the incidence of toxicity, including patient characteristics, such as comorbidities or previous treatments which may modify the dose response and other treatments including chemotherapy which have the potential to cause side effects but may also affect the dose response of an organ [105].
2. Preliminary data is emerging to indicate that the response of normal tissues is partly determined by genetic susceptibilities. Genome wide association studies (GWAS) have so far shown inconsistent results when associations between toxicity and single nucleotide polymorphism (SNPs) have been investigated [106].
3. Currently the 3-D dose distribution to an organ is summarised and/or reduced to provide dosimetric information. However, this often results in the loss of spatial information. It is known that many organs contain substructure which is inherent to organ function. A classic example is the kidney [107] where dose specifically to the nephrons is known to be important.
4. Dosimetric data for an organ at risk relies on the contouring of the structure on the treatment planning system. Institutional protocols should be in place to ensure consistency of outlining. However, definitions may vary between institution, this is particularly important when applying a model to data from another institution [108].
5. WYSINWYG. What you see is not what you get. In addition to contouring consistency most NTCP studies use the treatment planning scan to define the organ at risk. Great care is taken at each fraction of radiotherapy to ensure that the treatment plan is reproduced and that the target is irradiated accordingly. However variation in normal tissues is not necessarily accounted for so unless an accumulated dose, based on daily imaging is constructed, there may well be a difference between the dosimetric data reported from the treatment plan and the actual dose to the normal tissue being modelled [109].

Awareness of these challenges and, where possible, incorporating them in to TCP/NTCP modelling will improve the robustness and generalisability of the resultant models.

17.14 Summary

This chapter has reviewed many of the studies which have implemented machine learning to further the knowledge of TCP and NTCP. Considering the total number of publications, machine learning has had a limited impact on the field. Here, we consider why this is the case and how that might be addressed. Machine learning, particularly artificial neural networks are traditionally regarded as being mystical black boxes where it is impossible to interpret the underlying model. Although it is challenging to interpret the weights of a black box it is not impossible, whilst other machine learning techniques, for example decision trees are considerably more transparent. There are a wide variety of machine learning techniques and deciding which one is appropriate can be daunting. The suite of

publications from Duke University [74, 79, 83, 90, 110] and comparative papers by Oh [111], Pella [93], Dean [62, 89] are insightful. It is not wise to necessarily take the AUC measure as the comparative standard between models as this may well be data specific. However, it is useful to consider the congruence of features selected by the final model. In some cases, combining different models improves predictive accuracy particularly where input features are potentially highly correlated. In this case an ensemble may facilitate similar information being used in slightly different forms. Alternatively, a hybrid approach can result in the best of all worlds. The flip side is that these models are inherently complex and may suffer from a lack of generalisability if not carefully trained. In addition, it may be more challenging to interpret the role of individual input features when many are distributed through the model. Many of the studies presented in this chapter have indicated that the results from machine learning were superior to standard techniques. This may be in part due to the flexible approach to combining different data types that are available. However only in rare cases does the AUC exceed 0.8. Although this is considered to be a very good result for both classic statistical and machine learning approaches in the medical arena ideally every patient would have a valid prediction. The reasons why we reach this glass ceiling are complex but essentially result from a failure to fully reflect the patient experience. No model can predict an outcome from data that is not provided as an input. The amount of data available for each patient is exploding as genetic information is incorporated into studies. In addition, the dose distribution to organs at risk is insufficiently characterised by DVH and steps to improve this by including spatial information will further increase the number of input features. Machine learning is a knowledge transfer tool allowing clinicians to present all of the data that they regard as relevant to a specific prediction situation. Clearly medical understanding evolves daily and therefore predictive models will need to continuously be updated to include this increased knowledge. Machine learning approaches are well equipped to deal with big data, and it is hoped that in the future the understanding of the response of normal tissues following cancer treatment including radiotherapy will be well understood and reliable knowledge-based models will be used as standard in the clinic.

17.15 Conclusions

Recent evolution in imaging and biotechnology has provided new opportunities for reshaping our understanding of radiotherapy response. However, the complexity of radiation-induced effects and the variability of tumour and normal tissue responses would render the utilisation of machine learning algorithms as indispensable tools for better delineation of these complex interaction mechanisms.

References

1. Bortfeld T, et al. Image-guided IMRT. Berlin: Springer-Verlag; 2006.
2. Webb S. The physics of three-dimensional radiation therapy : conformal radiotherapy, radiosurgery, and treatment planning, Series in medical physics, vol. xiv. Bristol: Institute of Physics Pub; 2001. 373 p.
3. Halperin EC, Perez CA, Brady LW. Perez and Brady's principles and practice of radiation oncology, vol. xxxii. 5th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. p. 2106.
4. Moissenko V, Deasy JO, Van Dyk J. Radiobiological modeling for treatment planning. In: Van Dyk J, editor. The modern technology of radiation oncology: a compendium for medical physicists and radiation oncologists. Madison: Medical Physics Publishing; 2005. p. 185–220.
5. El Naqa I. A guide to outcome modeling in radiotherapy and oncology : listening to the data, Series in medical physics and biomedical engineering, vol. xxviv. Boca Raton: CRC Press, Taylor & Francis Group; 2018. p. 367.
6. Choi N, et al. Predictive factors in radiotherapy for non-small cell lung cancer: present status. Lung Cancer. 2001;31(1):43–56.
7. Fu XL, et al. Study of prognostic predictors for non-small cell lung cancer. Lung Cancer. 1999;23(2):143–52.
8. Blanco AI, 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):1055–69.
9. 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):1106–13.
10. Marks LB. Dosimetric predictors of radiation-induced lung injury. Int J Radiat Oncol Biol Phys. 2002;54(2):313–6.
11. Hope AJ, et al. Clinical, dosimetric, and location-related factors to predict local control in non-small cell lung cancer. In: Astro 47th annual meeting. Denver: Denver, CO.; 2005.
12. Tucker SL, et al. Dose-volume response analyses of late rectal bleeding after radiotherapy for prostate cancer. International Journal of Radiation Oncology Biology Physics. 2004;59(2):353–65.
13. El Naqa I, et al. Multi-variable modeling of radiotherapy outcomes including dose-volume and clinical factors. Int J Radiat Oncol Biol Phys. 2006;64(4):1275–86.
14. Deasy JO, El Naqa I. Image-based modeling of normal tissue complication probability for radiation therapy. Cancer Treat Res. 2008;139:215–56.
15. Bentzen SM, et al. Quantitative analyses of Normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. Int J Radiat Oncol Biol Phys. 2010;76(3 Suppl):S3–9.
16. 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):S155–60.
17. El Naqa I. Machine learning methods for predicting tumor response in lung cancer. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;2(2):173–81.
18. Collins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. Ann Intern Med. 2015;162(1):55–63.
19. Klement RJ, et al. Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer. Int J Radiat Oncol Biol Phys. 2014;88(3):732–8.
20. Munro TR, Gilbert CW. The relation between tumour lethal doses and the Radiosensitivity of tumour cells. Br J Radiol. 1961;34(400):246–51.
21. Hall EJ, Giaccia AJ. Radiobiology for the radiologist, vol. ix. 6th ed. Philadelphia: Lippincott Williams & Wilkins; 2006. p. 546.

22. Joiner M, Kogel AVD. Basic clinical radiobiology, vol. vi. 4th ed. London: Hodder Arnold; 2009. p. 375.
23. Goitein M. Tumor control probability for an inhomogeneously irriadiated target volume. In: Zink S, editor. Evaluation of treatment planning for particle beam radiotherapy. Bethesda: National Cancer Institute; 1987.
24. Zaider M, Minerbo GN. Tumour control probability: a formulation applicable to any temporal protocol of dose delivery. Phys Med Biol. 2000;45(2):279–93.
25. Hall EJ. Radiobiology for the radiologist, vol. xii. 4th ed. Philadelphia: J.B. Lippincott; 1994. p. 478.
26. Zaider M, Hanin L. Tumor control probability in radiation treatment. Med Phys. 2011;38(2):574–83.
27. Lindsay PE, et al. Retrospective Monte Carlo dose calculations with limited beam weight information. Med Phys. 2007;34(1):334–46.
28. Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys. 2003;30:979–85.
29. El Naqa I, et al. Datamining approaches for modeling tumor control probability. Acta Oncol. 2009;49(8):1363–73.
30. El Naqa I, et al. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. Int J Radiat Oncol Biol Phys. 2006;64(4):1275–86.
31. Kennedy R, et al. Solving data mining problems through pattern recognition. Upper Saddle River: Prentice Hall; 1998.
32. Willner J, et al. Dose, volume, and tumor control prediction in primary radiotherapy of non-small-cell lung cancer. Int J Radiat Oncol Biol Phys. 2002;52(2):382–9.
33. Martel MK, et al. Estimation of tumor control probability model parameters from 3-D dose distributions of non-small cell lung cancer patients. Lung Cancer. 1999;24(1):31–7.
34. Mehta M, et al. A new approach to dose escalation in non-small-cell lung cancer. Int J Radiat Oncol Biol Phys. 2001;49(1):23–33.
35. Brodin O, Lennartsson L, Nilsson S. Single-dose and fractionated irradiation of four human lung cancer cell lines in vitro. Acta Oncol. 1991;30(8):967–74.
36. Seibert RM, et al. A model for predicting lung cancer response to therapy. Int J Radiat Oncol Biol Phys. 2007;67(2):601–9.
37. Ramsey CR, et al. A technique for adaptive image-guided helical tomotherapy for lung cancer. Int J Radiat Oncol Biol Phys. 2006;64(4):1237–44.
38. Borst GR, et al. Standardised FDG uptake: a prognostic factor for inoperable non-small cell lung cancer. Eur J Cancer. 2005;41(11):1533–41.
39. Levine EA, et al. Predictive value of 18-fluoro-deoxy-glucose-positron emission tomography (18F-FDG-PET) in the identification of responders to chemoradiation therapy for the treatment of locally advanced esophageal cancer. Ann Surg. 2006;243(4):472–8.
40. Ben-Haim S, Ell P. 18F-FDG PET and PET/CT in the evaluation of cancer treatment response. J Nucl Med. 2009;50(1):88–99.
41. El Naqa I, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recogn. 2009;42(6):1162–71.
42. Mac Manus MP, et al. Metabolic (FDG-PET) response after radical radiotherapy/chemoradiotherapy for non-small cell lung cancer correlates with patterns of failure. Lung Cancer. 2005;49(1):95–108.
43. Yamamoto Y, et al. Correlation of FDG-PET findings with histopathology in the assessment of response to induction chemoradiotherapy in non-small cell lung cancer. Eur J Nucl Med Mol Imaging. 2006;33(2):140–7.
44. Pieterman RM, et al. Preoperative staging of non-small-cell lung cancer with positron-emission tomography. N Engl J Med. 2000;343(4):254–61.
45. Wong CY, et al. Correlating metabolic and anatomic responses of primary lung cancers to radiotherapy by combined F-18 FDG PET-CT imaging. Radiat Oncol. 2007;2:17.
46. Vaidya M, et al. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. Radiother Oncol. 2012;102(2):239–45.

47. Group, B.D.W. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69(3):89–95.
48. Le Q-T, et al. An evaluation of tumor oxygenation and gene expression in patients with early stage non-small cell lung cancers. Clin Cancer Res. 2006;12(5):1507–14.
49. Rube CE, et al. Cytokine plasma levels: reliable predictors for radiation pneumonitis? PLoS One. 2008;3(8):e2898.
50. Oh JH, et al. A Bayesian network approach for modeling local failure in lung cancer. Phys Med Biol. 2011;56(6):1635–51.
51. Marks LB, Ten Haken RK, Martel MK. Guest editor's introduction to QUANTEC: a users guide. Int J Radiat Oncol Biol Phys. 2010;76(3 Suppl):S1–2.
52. Lyman JT. Complication probability as assessed from dose-volume histograms. RadiatResSuppl. 1985;8:S13–9.
53. Kutcher GJ, et al. Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations. IntJRadiatOncolBiolPhys. 1991;21(1):137–46.
54. Burman C, et al. Fitting of normal tissue tolerance data to an analytic function. Int J Radiat Oncol Biol Phys. 1991;21(1):123–35.
55. Emami B, et al. Tolerance of normal tissue to therapeutic irradiation. Int J Radiat Oncol Biol Phys. 1991;21(1):109–22.
56. Niemierko A. A generalized concept of equivalent uniform dose (EUD). Med Phys. 1999;26:1100.
57. Niemierko A. Reporting and analyzing dose distributions: a concept of equivalent uniform dose. Med Phys. 1997;24(1):103–10.
58. Källman P, Agren A, Brahme A. Tumour and normal tissue responses to fractionated non-uniform dose delivery. Int J Radiat Biol. 1992;62(2):249–62.
59. Jackson A, et al. Analysis of clinical complication data for radiation hepatitis using a parallel architecture model. Int J Radiat Oncol Biol Phys. 1995;31(4):883–91.
60. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression, vol. 398. Hoboken: Wiley; 2013.
61. El Naqa I, et al. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. Int J Rad Oncol Biol Phys. 2006;64(4):1275–86.
62. Dean J, et al. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. Clin Transl Radiat Oncol. 2017;8:27–39.
63. El Naqa I, Bradley JD, Deasy J. Nonlinear kernel-based approaches for predicting Normal tissue toxicities. In: Seventh international conference on machine learning and applications, proceedings; 2008. p. 539–44.
64. Oh JH, Al-Lozi R, El Naqa I. Application of machine learning techniques for prediction of radiation pneumonitis in lung cancer patients. In: Eighth international conference on machine learning and applications, proceedings; 2009. p. 478–83.
65. Dawson LA, et al. Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation. Int J Radiat Oncol Biol Phys. 2005;62(3):829–37.
66. Bauer JD, et al. Principal component, Varimax rotation and cost analysis of volume effects in rectal bleeding in patients treated with 3D-CRT for prostate cancer. Phys Med Biol. 2006;51(20):5105–23.
67. Sohn M, Alber M, Yan D. Principal component analysis-based pattern analysis of dose-volume histograms and influence on rectal toxicity. Int J Radiat Oncol Biol Phys. 2007;69(1):230–9.
68. Skala M, et al. Patient-assessed late toxicity rates and principal component analysis after image-guided radiation therapy for prostate cancer. Int J Radiat Oncol Biol Phys. 2007;68(3):690–8.
69. Vesprini D, et al. Role of principal component analysis in predicting toxicity in prostate cancer patients treated with hypofractionated intensity-modulated radiation therapy. Int J Radiat Oncol Biol Phys. 2011;81(4):e415–21.

70. Liang Y, et al. Impact of bone marrow radiation dose on acute hematologic toxicity in cervical cancer: principal component analysis on high dimensional data. Int J Radiat Oncol Biol Phys. 2010;78(3):912–9.
71. McCullough WS, Pitts W. A logical calculus of the ideas imminent in nervous activity. Bull Math Biol. 1943;52(1–2):99–115.
72. Munley MT, et al. A neural network to predict symptomatic lung injury. Phys Med Biol. 1999;44(9):2241–9.
73. Lyman JT, Wolbarst AB. Optimization of radiation therapy, III: a method of assessing complication probabilities from dose-volume histograms. Int J Radiat Oncol Biol Phys. 1987;13(1):103–9.
74. Chen SF, et al. A neural network model to predict lung radiation-induced pneumonitis. Med Phys. 2007;34(9):3420–7.
75. Mcdonald S, et al. Injury to the lung from cancer-therapy - clinical syndromes, measurable end-points, and potential scoring systems. International Journal of Radiation Oncology Biology Physics. 1995;31(5):1187–203.
76. Lind PA, et al. ROC curves and evaluation of radiation-induced pulmonary toxicity in breast cancer. International Journal of Radiation Oncology Biology Physics. 2006;64(3):765–70.
77. Buettner F, et al. Using dose-surface maps to predict radiation-induced rectal bleeding: a neural network approach. Phys Med Biol. 2009;54(17):5139–53.
78. Hastie TT, Friedman J, Tisbshirani R. The elements of statistical learning: data mining, inference and prediction. New York: Springer-Verlag; 2002.
79. Chen S, et al. Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. Med Phys. 2007;34(10):3808–14.
80. El Naqa I, et al. Predicting radiotherapy outcomes using statistical learning techniques. Phys Med Biol. 2009;54(18):S9–S30.
81. Schiller TW, et al. Improving clinical relevance in ensemble support vector machine models of radiation pneumonitis risk. In: Eighth international conference on machine learning and applications, proceedings; 2009. p. 498–503.
82. Kohonen T. Essentials of the self-organizing map. Neural Netw. 2013;37:52–65.
83. Chen SF, et al. Using patient data similarities to predict radiation pneumonitis via a self-organizing map. Phys Med Biol. 2008;53(1):203–16.
84. Oh JH, El Naqa I. Bayesian network learning for detecting reliable interactions of dose-volume related parameters in radiation pneumonitis. In: Eighth International Conference on Machine Learning and Applications, Proceedings; 2009. p. 484–8.
85. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks - the combination of knowledge and statistical-data. Mach Learn. 1995;20(3):197–243.
86. Das SK, et al. Predicting lung radiotherapy-induced pneumonitis using a model combining parametric Lyman probit with nonparametric decision trees. Int J Radiat Oncol Biol Phys. 2007;68(4):1212–21.
87. Palma DA, et al. Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: an international individual patient data meta-analysis. Int J Radiat Oncol Biol Phys. 2013;85(2):444–50.
88. Valdes G, et al. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. Phys Med Biol. 2016;61(16):6105–20.
89. Dean JA, et al. Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy. Radiother Oncol. 2016;120(1):21–7.
90. Das SK, et al. Combining multiple models to generate consensus: application to radiation-induced pneumonitis prediction. Med Phys. 2008;35(11):5098–109.
91. Nalbantov G, et al. Combining the predictions for radiation-induced dysphagia in lung cancer patients from multiple models improves the prognostic accuracy of each individual model. J Thorac Oncol. 2011;6(6 Supp 2):S549.

92. Trotti A, et al. CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment. Semin Radiat Oncol. 2003;13(3):176–81.
93. Pella A, et al. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. Med Phys. 2011;38(6):2859–67.
94. Wang R, et al. Multi-objective ensemble deep learning using electronic health records to predict outcomes after lung cancer radiotherapy. Phys Med Biol. 2019;64(24):245005. https://doi.org/10.1088/1361-6560/ab555e.
95. Liang B, et al. Prediction of radiation pneumonitis with dose distribution: a convolutional neural network (CNN) based model. Front Onco. 2020;9:1500. https://doi.org/10.3389/fonc.2019.01500. eCollection 2019.
96. Hatt M, et al. IBSI: an international community radiomics standardization initiative. J Nucl Med. 2017;59
97. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. Phys Med. 2017;38:122–39.
98. Wei L, et al. Machine learning for radiomics-based multimodality and multiparametric modeling. Q J Nucl Med Mol Imaging. 2019;63(4):323–38.
99. Nie K, et al. NCTN assessment on current applications of Radiomics in oncology. Int J Radiat Oncol Biol Phys. 2019;104(2):302–15.
100. Gabrys HS, et al. Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of xerostomia. Front Oncol. 2017;8:35.
101. El Naqa I, et al. Radiogenomics and radiotherapy response modeling. Phys Med Biol. 2017;62(16):R179–206.
102. Cui S, et al. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. Med Phys. 2019;46(5):2497–511.
103. Rosenstein BS, et al. Radiogenomics: radiobiology enters the era of big data and team science. Int J Radiat Oncol Biol Phys. 2014;89(4):709–13.
104. El Naqa I, Napel S, Zaidi H. Radiogenomics is the future of treatment response assessment in clinical oncology. Med Phys. 2017;45(10):4325–8.
105. Kasibhatla M, Kirkpatrick JP, Brizel DM. How much radiation is the chemotherapy worth in advanced head and neck cancer? Int J Radiat Oncol Biol Phys. 2007;68(5):1491–5.
106. Barnett GC, et al. Independent validation of genes and polymorphisms reported to be associated with radiation toxicity: a prospective analysis study. Lancet Oncol. 2012;13(1):65–77.
107. Dawson LA, et al. Radiation-associated kidney injury. Int J Radiat Oncol Biol Phys. 2010;76(3 Suppl):S108–15.
108. Groom N, et al. Is pre-trial quality assurance necessary? Experiences of the CONVERT phase III randomized trial for good performance status patients with limited-stage small-cell lung cancer. Br J Radiol. 2014;87(1037):20130653.
109. Jaffray DA, et al. Accurate accumulation of dose for improved understanding of radiation effects in normal tissue. Int J Radiat Oncol Biol Phys. 2010;76(3 Suppl):S135–9.
110. Das SK, et al. Decision fusion of machine learning models to predict radiotherapy-induced lung pneumonitis. In: Seventh international conference on machine learning and applications, proceedings; 2008. p. 545–50.
111. Adamina M, Tomlinson G, Guller U. Bayesian statistics in oncology a guide for the clinical investigator. Cancer. 2009;115(23):5371–81.

Smart Adaptive Treatment Strategies

18

Huan-Hsin Tseng, Randall K. Ten Haken, and Issam El Naqa

18.1 Introduction

Recent years have witnessed tremendous growth in cancer patient-specific information from multimodality imaging (Computer tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MR), ultrasound (US)) to biotechnology (genomics, transcriptomics, proteomics, etc.), ushering in a new era of Big Data in oncology. With the availability of the patient-specific data, such as clinical, treatment, imaging, molecular markers, before and/or during oncology courses, new opportunities are becoming available for personalized oncology and radiotherapy treatments [1, 2].

The synthesis of this information into actionable knowledge to improve patient outcomes is currently a major goal of modern oncology. Subsequently, adapted cancer treatments (ACTs) have emerged as an important framework that aims to develop personalized treatments by adjusting treatment prescription according to clinical, geometrical changes, and physiological parameters observed during an oncology treatment course. Our goal in this chapter is to explore in more details the processes involved in the ACT framework that would allow aggregating and analyzing relevant patient information in a systematic manner to achieve more accurate decision-making and optimize long-term outcomes. We will consider the special case of radiotherapy (RT) as an example of such applications.

H.-H. Tseng
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
e-mail: h.tseng@gapp.nthu.edu.tw

R. K. Ten Haken · I. El Naqa (✉)
Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, USA
e-mail: rth@med.umich.edu; ielnaqa@med.umich.edu

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_18

18.2 Adaptive Treatment in Radiotherapy

The notion of ACT extends the traditional concept of adapted radiotherapy (ART) [3, 4], which is primarily based on imaging information for guidance, into a more general ART framework that can receive and process all relevant patient-specific knowledge that can be useful for adaptive decision-making and personalization of treatment. For instance, during a course of RT, anatomical and biological changes occur to the tumor and surrounding normal tissue that should be accounted for and the plan for *treatment adapted* to achieve improved outcome. Schwartz et al. demonstrated in a prospective trial benefits of such adaptive approach in head and neck cancer [5]. This approach has been demonstrated in phase I/II trials in liver [6] and lung [7] cancers. However, adaptation in these studies is based on subjective assessment and application of short-term heuristics that do not take full advantage of intra-treatment/follow-up information for deciding the best adaptation action long term, leading to modest improvements and yielding in many cases disappointing suboptimal results [8].

The proposed ACT-ART framework can be thought of as being comprised of four stages, as depicted in Fig. 18.1. These stages include: (1) planning patients

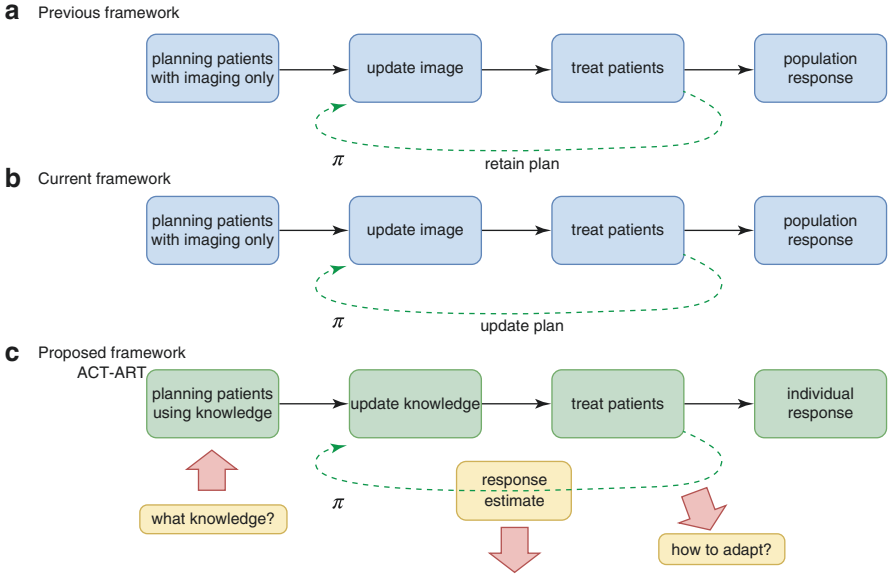


Fig. 18.1 Comparison of workflow of (**a**) nonadaptive RT, (**b**) current image-based ART, and (**c**) the proposed ACT-ART approach. The current ART (**b**) mostly relies on image guidance such as computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI). In ACT-ART, the planning patients' stage can utilize general knowledge about patient status (imaging + biological markers) as information for adapting treatment instead of using imaging only. Two major differences between previous/current RT and ACT-ART are that (1) knowledge is no longer restricted to imaging only and can include biological markers such as tumor genetics or blood-based inflammatory proteins (cytokines) to inform predictive modeling and decision-making; and (2) application process of machine learning for adapting a treatment plan π in ACT-ART. Adapted from [9]

using available knowledge, or pre- treatment modeling, (2) updating the prediction models with evolving knowledge through the course of therapy, or during-treatment modeling, (3) personalizing initial patient's treatments, and (4) adapting the initial treatment to individual's responses, where the two middle steps can be repeated at each radiation dose fraction (or few fractions) so that optimal treatment objectives are met and potentially long-term goals are optimized, i.e., long-term tumor control with limited side effects to surrounding normal tissues.

The first step in the implementation of an ACT-ART framework starts at the planning stage of patients by extending the current "image-only patients" into a more general preparation stage that can incorporate all relevant informatics signals for evaluating available treatment options, c.f., Fig. 18.1a, b. Thus, imaging information (CT/PET/MRI) is supplemented with biological markers (genomics, transcriptomics, proteomics, etc.) that can potentially aid the process of personalizing treatment to an individual patient's molecular characteristics and is not limited to imaging only as currently is the case as discussed later in the chapter. To develop an ACT framework, there are three essential questions pertaining to the successful development that need to be addressed [9]:

Q1 What knowledge should be synthesized for ACT planning?

Q2 How can we develop powerful predictive outcome modeling techniques based on such knowledge?

Q3 How can we use these models in a strategically optimal manner to adapt a patient's treatment plan?

The answer to these questions will be the subject of the subsequent sections of this chapter.

18.3 What Knowledge Is Needed for ACT?

There are four major types of oncology data that are potentially useful as part of the knowledge synthesis for ACT-ART: *clinical, treatment, imaging radiomics, and biological data*. To understand why and how they can be informative for assessing treatment outcomes, we provide a brief description about these four categories of data.

18.3.1 Clinical Data

Clinical data refers to cancer diagnostic characteristics (e.g., grade, stage, histology, site, etc.), physiological metrics (e.g., blood cell counts, heart/pulse rates, pulmonary measurements, etc.), and patient-related information (e.g., comorbidities, gender, age, etc.). Due to their nature, clinical data can usually be found in unstructured format

such that it the direct extraction of information can be challenging. Therefore, machine learning techniques for natural language processing could be useful for transforming such data into structured format (e.g., tabulated) before further processing [10].

18.3.2 Treatment Data

Treatment oncology data or dosimetric data in the context of RT for instance are informatic to the treatment planning process in RT, which includes simulated calculation of radiation dose using computed tomography (CT) imaging. In particular, dose–volume metrics obtained, for example, from dose volume histograms (DVHs) are extensively investigated for outcome modeling [11–15]. Useful metrics are typically the volume receiving greater than or equal to a certain dose (Vx), the minimum dose to the hottest $x\%$ of the volume (Dx), mean, maximum, minimum dose, etc. [16]. Notably, a dedicated software based on MATLAB called "DREES" can derive these metrics automatically and apply them in outcome prediction models of RT response [17].

18.3.3 Imaging Data

Quantitative imaging data or *radiomics* is a field of medical imaging study that aims to extract meaningful quantitative features from medical images and relate this information to clinical and biological endpoints. The most common imaging modality is CT, which has been considered the standard for treatment planning in oncology and RT specifically. Other imaging modalities include positron emission tomography (PET), and magnetic imaging resonance (MRI).

18.3.4 Biological Data

A biomarker can be defined [18] as *"a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention."* Measurements of biomarkers are typically based on tissue or fluid specimens, which are analyzed using molecular biology laboratory techniques [19] and have the following two categories according to their biochemical sources:

(a) *Exogenous biomarkers*: by injecting a foreign substance into patients such as that used in molecular imaging and are used in radiomics applications.
(b) *Endogenous biomarkers*: there exists two subclasses within this category:
    - *Expression biomarkers*: changes measured in protein levels or gene expression.
    - *Genetic biomarkers*: measuring variations between the underlying DNA genetic code and tumors or normal tissues.

18.4 How to Develop Outcome Models Using This Knowledge?

The modeling of treatment response is primary objective of modern oncology research as it is a key toward personalization of cancer care. In the following, we will provide a brief description of this subject in the context of RT, interested readers are referred to Chap. 15, which is dedicated to this subject or consult the literature for more details [20].

RT outcome models are typically expressed in terms of tumor control probability (TCP) and normal tissue complication probability (NTCP) [21, 22]. In principle, both TCP and NTCP may be evaluated using analytical and/or data-driven models. Though the former provides structural formulation, it can be incomplete and less accurate due to the complexity of radiobiological processes. On the other hand, data-driven models tend to learn empirically from the data observed, and thus they are capable of considering higher complexities and interactions of irradiation with the biological system. The trade-offs between analytical models and data-driven models can vary in terms of radiobiological understanding and prediction accuracy. Here, we will focus on machine learning methods only.

Traditionally, feed-forward neural networks were extensively investigated to model post-radiation treatment outcomes for cases of lung injury [23, 24] and biochemical failure and rectal bleeding in prostate cancer [25, 26]. Subsequently, support vector machines (SVMs), as universal constructive learning procedures based on the statistical learning theory [27] were utilized. For discrimination between patients who are at low risk versus patients who are at high risk of treatment, the main idea of a SVM would be to separate these two classes with "hyper-planes" that maximize the margin between them in the nonlinear feature space defined by an implicit kernel mapping [28–30]. However, these methods have been stigmatized as black boxes, hindering their application in practical clinical contexts. In an effort, to alleviate the black box stigma of generic machine learning methods and incorporate more system-like approaches methods based on graphical approaches such as Bayesian networks (BNs) have witnessed increased used in outcome modeling of cancer [31–36]. A BN provides graphical representation of the relationships between the variables represented as nodes in a directed acyclic graph (DAG), which encodes the presence and direction of relationship influence among the variables themselves and the clinical endpoint of interest. The relationship between parent and child nodes is modeled by conditional probabilities using Bayes chain rule. More recently, methods based on deep learning were adopted for outcome prediction [37, 38], with their distinct ability to learn from raw data. More details are discussed in Chap. 15.

18.5 How to Optimize Adaptation?

Machine learning offers a wide variety of tools for adapting decision-making. Specifically, dynamic machine learning algorithms such as reinforcement learning (RL) have been adopted in the design of adaptive clinical trials to estimate dynamic

treatment regimens [39]. For instance, the sequential multiple assignment randomized trial (SMART) has been applied for constructing adaptive interventions in different diseases related to drug abuse, HIV/AIDS, and mental illness with promising results [40]. However, this systematic approach has not been applied in (radiation) oncology. Berry argues that the potential benefit of adaptive designs is greatest in complicated settings exemplified by personalized medical research in oncology and suggests the use of Bayesian approaches to optimize sequential decision-making. However, the framework is yet to be defined in this case and optimality conditions still need to be proven for yielding the desired outcomes [41]. In radiation oncology, Kim et al. presented a simulation study based on Markov decision processes (MDPs), where they showed numerical examples of modifying dose fractionation schedules for adaptive RT applications [42]. Moreover, RL methods we also used to optimize the dose per fraction using different utility functions in cell culture experiments [43].

For modeling the ACT, we can use a reinforcement machine learning framework [44]. In this framework, the clinical (radiotherapy) environment is represented by a Markov decision process (MDP) as shown in Fig. 18.2. In the case of adaptive radiotherapy, at any time point (t) during the treatment adaptation there is a *state* s_t to describe a patient's biological status (e.g., tumor volume, tumor response (TCP) and toxicity risk of surrounding organs (NTCP)) and a reward associated to the state, $R_t = R_t(s_t)$, to evaluate the current patient status, and then an external action (a_t) (e.g., boosting the active part of the tumor, modifying the number of fractions, changing the fraction size, adding/removing chemotherapy agent, etc.) to be made by an agent (treating team). In the MDP setting, states and rewards are preordained pertaining to simulate a patient's biological configuration. An agent has no ability to directly change these factors, where it is only possible to influence a patient's state via agent's actions, so that the actions a_t are the only external variables to affect the MDP dynamics. The optimal decision of an action is based on maximizing the sum

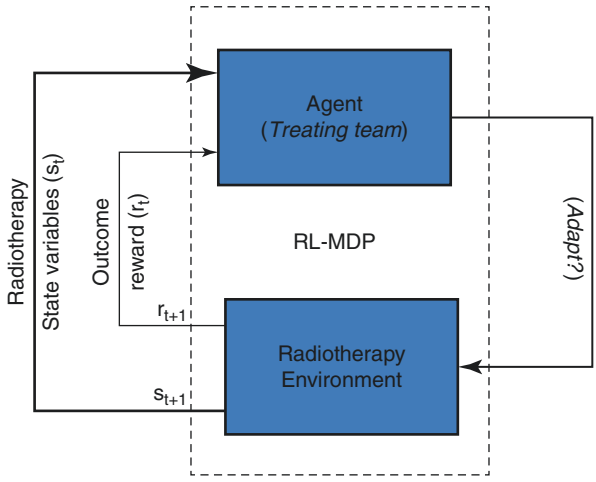


Fig. 18.2 Reinforcement learning representation of adaptive radiotherapy as a Markov decision process

of rewards in Eq. (18.1) such that a patient's tumor response (TCP) is highest and toxicity risk of surrounding organs (NTCP) is the lowest, for example.

Intuitively, one can imagine initially an agent has no knowledge of the relation between (s_t, R_t, a_t), much like new-born babies learning to interact with the world; it is through numerous trial-and-error process to understand the interactions between the three variables. This is where powerful statistical tools may help to clarify the relations and where machine learning involves. Most of the advanced deep reinforcement learning algorithms are dedicated to the interaction-learning process, e.g., Soft-Actor-Critic (SAC), Trust Region Policy Optimization (TRPO), Actor-Critic methods (A2C/A3C), etc.

The descriptions above can be described mathematically by an objective (of the MDP) to be maximized:

$$Q^{\pi}(s,a) = E_{P(s,a)}\left\{\sum_{k=0}^{\infty} \gamma^k R(s_k) | \pi, s, a\right\} \tag{18.1}$$

where $0 \le \gamma \le 1$ is a constant called *discounting rate* and $\pi : S \to A$ is called a *policy* (to be determined) that links actions a_t to states s_t, $\pi(s_t) = a_t$, Essentially, the policy function π is exactly the interaction an agent likes to figure out as mentioned above, where in medical physics term it corresponds to a given treatment adaptation policy desired.

Interestingly, taking $\gamma = 0$ in Eq. (18.1), $Q^{\pi}(s = s_0, a) = R(s_0)$, is equivalent to saying that an agent is extremely *myopic*, who only regards the current reward or short-term gains. In contrast when $\gamma = 1$ in Eq. (18.1) leads to an agent that looks into futuristic rewards or long-term gains. From the mathematical point of view, Eq. (18.1) as a definition is intuitive to comprehend but it is intractable for computational purposes due to the infinite sum involved. Therefore, one usually would convert Eq. (18.1) into a more computationally amicable form called *Bellman equation*:

$$Q_{i+1}(s,a) = E_{s' \sim P(s,a)}\left\{R(s,a) + \gamma \max_{a' \in A} Q_i(s', a')\right\} \tag{18.2}$$

such that $\lim_{i \to \infty} Q_i \to Q^*$, which is equivalent to Eq. (18.1). It can also be proved that the convergent point Q^* of the numerical iteration from the Bellman equation is optimal [27]. Once the optimal Q^* is derived from Eq. (18.2), one can subsequently solve for the corresponding policy function π defined by:

$$a^* \underline{\underline{\text{def}}}\ \pi(s) \underline{\underline{\text{def}}}\ \max_{a' \in A} Q^*(s, a'), \tag{18.3}$$

which completes the interaction between (s_t, R_t, a_t) as mentioned above. It is now clear that such link is achieved through the complex Q-function defined in Eq. (18.1). As such, the approach deriving a policy is called Q-learning [28], which is usually encountered in ACT applications.

In fact, there is another algorithm to derive a policy function π via the *policy-gradient* method [45]. However, the details are not in the scope of our discussion

here. A quick sketch is that a family of parametrizable functions $\pi_\theta(s_t) = a_t$ is used to directly approach the optimal policy π^* such that $\pi_\theta \to \pi^*$, see [45].

18.5.1 Classical MDP/RL Learning

To compute the Bellman equation, one needs to find a proper functional form to approximate each Q_i in Eq. (18.2). One traditional method assumes that it can be approached with a *linear* sum over a functional basis $\left\{f_j\right\}_{j=1}^{n}$,

$$Q_i\left(s,a\right) = \sum_{j=1}^{n} \alpha_{ij}\, f_j\left(s,a\right) \tag{18.4}$$

where the unknown parameters $\left\{\alpha_{ij}\right\}_{j=1}^{n}$ in Eq. (18.4) can be determined by usual linear regression. Note that Eq. (18.4) does not imply that Q_i is linear, as the Taylor expansion of $\sin x \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$ is a decent analog to Eq. (18.4), where none of $\{\sin x, x, x^3, x^5\}$ is a linear function. Due to the general observation that the optimal Q^* is in general highly nonlinear, such classical assumption suffers from two drawbacks: (a) how to choose a suitable functional basis $\left\{f_j\right\}_{j=1}^{n}$ beforehand for solving Q^*? and (b) how many functional members are to be used? (i.e., choice of n). In fact, any improper choice of (a) or (b) easily leads to overfitting or underfitting for the Q-learning. Nevertheless, this approach serves as a viable solution at least.

18.5.2 Deep MDP/RL Learning

The introduction of deep learning (Chap. 4) allows the predicament of the classical approach to extract relevant features from raw data to be evaded, since deep learning, utilizing deep neural networks (DNNs), provides a strong nonlinear approximator to almost every continuous function, which was rigorously proven in the Universal Approximation Theorem by Hornik [46].

To take advantage of deep learning, one simply approximates Q_i in the Bellman equation of Eq. (18.2) with a DNN Q_i^θ, i.e., $Q_i \leftarrow Q_i^\theta$ with θ denoting the neural weights. By this simple replacement, one avoids drawbacks (a), (b) of the classical method at once fundamentally, as DNN requires only little assumption on functions to be approximated. This DNN approximation of $Q_i \leftarrow Q_i^\theta$ is known as the Deep Q-Net (DQN) proposed by Google [47]. DQN is the cornerstone of Deep Reinforcement Learning (DRL).

Apart from the fact that DRL learns the environment interaction more efficiently, a tricky problem to be addressed in oncology or radiotherapy applications is the modeling of the transition probability. By definition it describes how two states transit under an action $s \xrightarrow{a} s'$, denoted by $s' \sim P(s,a)$, where in the course of a treatment this represents the transition of a patient's biological status affected by a given dosage. Therefore, from medical perspective it is crucial to grasp the knowledge of

transition probability modeling as it is eventually used in Eqs. (18.1 and 18.2) for Q-function computation.

18.6 ACT Example in Radiotherapy

In following we present an example of an ACT framework in the context of adaptive radiotherapy in non-small cell lung cancer (NSCLC). In a population of 42 patients who had inoperable or unresectable stage II to stage III NSCLC, the patients were enrolled in an adaptive RT-escalated dose study to improve local tumor control at 2-years. The adaptation was based on ^{18}F-deoxyglucose (FDG)-avid region detected by mid-treatment positron emission tomography (PET) as shown in Fig. 18.3 [48]. Conformal RT was individualized to a fixed risk of Radiation Pneumonitis (RP) (grade > 2) and adaptively escalated to the residual tumor defined on mid-treatment FDG-PET up to a total dose of 86 Gy in 30 daily fractions.

Figure 18.4 in [49] illustrates a detailed construction of a radiotherapy environment. The main idea was to use important features indicative of tumor Local Control

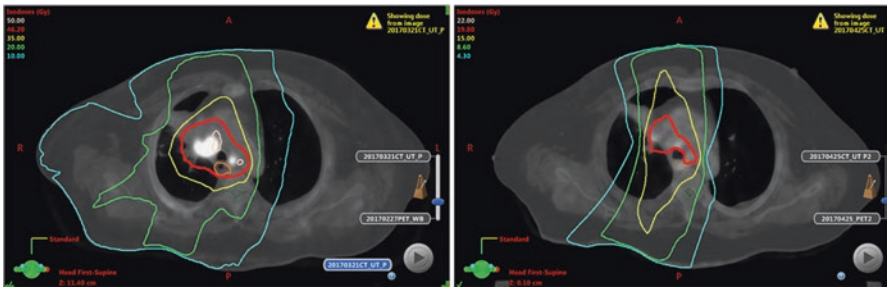


Fig. 18.3 Example of a NSCLC patient on an adaptive RT-escalated dose study to improve local tumor control by adaptation based on 18F-deoxyglucose (FDG)-avid region. [*Left]:* Pretreatment PET/CT plan showing original uptake. [*Right]:* mid-treatment PET/CT plan showing residual uptake for adaptation

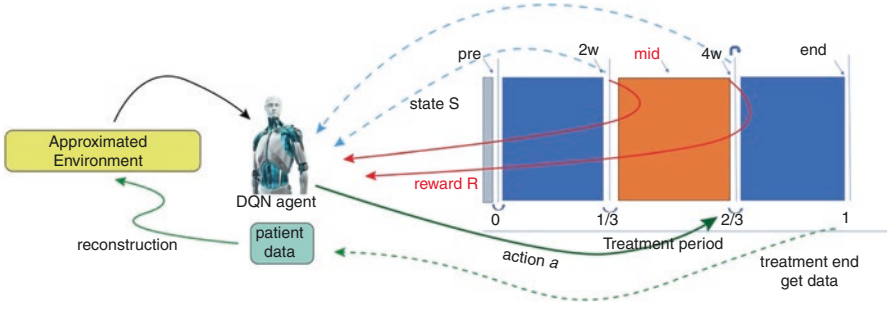


Fig. 18.4 Reinforcement learning for making decisions at two-third period of a treatment (right solid-green arrow). A first step in their framework is to learn transition functions from the historical data of two transitions recorded (RHS figure) so that the radiotherapy environment can be reconstructed (called approximated environment). With the transitions simulated, a DQN agent can then search for optimal dose at each stage [49]

(LC) and Radiation Pneumonitis (RP) after a complete treatment. In their study, factors of cytokines, SNPs, miRNA, and PET radiomics are considered informative as predicted by Bayesian Network (BN) analysis in [34]. They defined the BN selected features as states s_t with a properly defined reward R_t, combined with a set of DNNs to learn the transition probability $s' \sim P(s, a)$ from historical patient data.

Once a suitable radiotherapy environment is set up, Deep Reinforcement Learning can then be utilized to efficiently learn the MDP dynamical interaction, where in the study [49] they deployed DQN to optimize the total rewards for deriving optimal adaptation dose, i.e., an optimal policy $\pi^* : S \rightarrow A$.

18.7 Discussion and Recommendation

Smart adapted cancer treatments (ACTs) heavily rely on the proper integration of various components: information retrieval, multimodal data processing, statistical inference, and optimal control. Current machine learning techniques provide many opportunities to offer its help for improving ACT application. However, it is not yet entirely mature for us to wield these tools for clinical application. One major reason is due to the lack of comprehensive understanding of biological mechanisms of human body. Unlike healthcare, machine learning algorithms can be easily applied to industry with proper design and customization to complete desired tasks. Very often, one cannot be sure whether the data at hand is sufficient for prediction or desired treatment goals. Due to the uncertainty in many aspects, only when the algorithm is smartly implemented and correctly integrated then the adapted treatments can be executed successfully.

For example, it is noted in defining an MDP environment for an oncology treatment such as radiotherapy, there is no uniform way or standard procedure regarding the selection of indicative features. Although in the study [34], SNPs, miRNA, and PET radiomics, etc. MDP are adopted as states, other suitable or more effective choices may exist. Therefore, an MDP environment construction and setup requires intensive medical experience and background knowledge, and thus it serves a large field to be extensively explored.

On the other hand, to apply RL in oncology or radiotherapy effectively, there is also the art versus science of machine learning algorithm selection to be attended based on the nature of a task. Due to the variability of datasets and their intrinsic properties, it is unlikely to have a universal learning algorithm that fits all purposes. Particularly in the case of Deep Learning, a multilayered neural net generally serves as a black box. More attention to the analysis of medical inputs and outputs are required. Some interpretability tools developed such as LIME (Local Interpretable Model-Agnostic Explanations) [50] may be used to help provide better understanding of the performance of the machine learning algorithm.

18.8 Conclusions

An adapted cancer treatment is an approach to personalize medicine, where the characteristics and response of an individual are to be incorporated into the treatment prescription. An important factor to achieve these adaptations is the ability to improve prediction power, so that a treatment can be planned and adjusted beforehand. For a long time, it has been clinical team role to perform these tasks based on their professional medical training. With the maturity of statistical tools and computing hardware, machine learning methods has proven to be able to provide additional prediction power by mining hidden information and unnoticed patterns embedded in the numerical data. Among many possibilities, imaging data (PET/CT/MRI), radiomics, biological metrics, physiological information are known to provide valuable knowledge to be incorporated for adaptation; this is where the concept of knowledge-based response-adapted radiotherapy can reach its potentials.

Recent evolution in imaging and biotechnology is a major advantage to reshape the understanding of oncology response. However, to achieve the knowledge-based response-adapted radiotherapy goal one needs to be able to handle the complexity of treatment effects and consider the variability of tumor and normal tissue responses increase. With and after proper information fusion, one can then provide accurate and efficient methods for cancer treatment decision-making. One will rely on advanced algorithms for processing multidimensional data to achieve such goal, and the solution can be offered by current machine learning development. It is therefore the driving force why machine learning and deep learning have gradually become indispensable tools for better delineation of these complex interaction mechanisms such as the case in adapting treatment to patient's response.

Acknowledgments The authors would like to thank Dr. Michael Green and Dr. Benjamin Rosen for providing Fig. 18.3.

References

1. Benedict SH, El Naqa I, Klein EE. Introduction to big data in radiation oncology: Exploring opportunities for research, quality assessment, and clinical care. Int J Rad Oncol Biol Phys. 2016;95(3):871–2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360301615271969
2. El Naqa I. Perspectives on making big data analytics work for oncology. Methods. 2016;111:32–44. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1046202316302651
3. Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online adaptive radiation therapy. Int J Rad Oncol Biol Phys. 2017;99(4):994–1003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360301617308271
4. Xing L, Siebers J, Keall P. Computational challenges for image-guided radiation therapy: framework and current research. Semin Radiat Oncol. 2007;17(4):245–57. image-Guided Radiation Therapy. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053429607000616

5. Schwartz DL, Garden AS, Thomas J, Chen Y, Zhang Y, Lewin J, Chambers MS, Dong L. Adaptive radiotherapy for head-and-neck cancer: initial clinical outcomes from a prospective trial. Int J Rad Oncol Biol Phys. 2012;83(3):986–93. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360301611031713
6. Feng M, Suresh K, Schipper MJ, Bazzi L, Ben-Josef E, Matuszak MM, Parikh ND, Welling TH, Normolle D, Ten Haken RK, Lawrence TS. Individualized adaptive stereotactic body radiotherapy for liver tumors in patients at high risk for liver damage: a phase 2 clinical trial. JAMA Oncol. 2018;4(1):40–7.
7. Kong F-M, Hayman JA, Griffith KA, Kalemkerian GP, Arenberg D, Lyons S, Turrisi A, Lichter A, Fraass B, Eisbruch A, Lawrence TS, Haken RKT. Final toxicity results of a radiation-dose escalation study in patients with non-small-cell lung cancer (nsclc): Predictors for radiation pneumonitis and fibrosis. Int J Rad Oncol Biol Phys. 2006;65(4):1075–86. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360301606002495
8. Brouwer CL, Steenbakkers RJ, van der Schaaf A, Sopacua CT, van Dijk LV, Kierkels RG, Bijl HP, Burgerhof JG, Langendijk JA, Sijtsema NM. Selection of head and neck cancer patients for adaptive radiotherapy to decrease xerostomia. Radiother Oncol. 2016;120(1):36–40. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167814016311495
9. Tseng H-H, Luo Y, Ten Haken RK, El Naqa I. The role of machine learning in knowledge-based response-adapted radiotherapy. Front Oncol. 2018;8:266.
10. Wu JT, Dernoncourt F, Gehrmann S, Tyler PD, Moseley ET, Carlson ET, Grant DW, Li Y, Welt J, Celi LA. Behind the scenes: a medical natural language processing project. Int J Med Inform. 2018;112:68–73. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138650561730446X
11. Marks LB. Dosimetric predictors of radiation-induced lung injury. Int J Rad Oncol Biol Phys. 2002;54(2):313–6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360301602029280
12. Levegrun S, Jackson A, Zelefsky MJ, Skwarchuk MW, Venkatraman ES, Schlegel W, Fuks Z, Leibel SA, Ling C. Fitting tumor control probability models to biopsy outcome after three-dimensional conformal radiation therapy of prostate cancer: pitfalls in deducing radiobiologic parameters for tumors from clinical data. Int J Rad Oncol Biol Phys. 2001;51(4):1064–80. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S036030160101731X
13. Hope AJ, Lindsay PE, El Naqa I, Alaly JR, Vicic M, Bradley JD, Deasy JO. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. Int J Rad Oncol Biol Phys. 2006;65(1):112–24. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360301605030750
14. Bradley J, Deasy JO, Bentzen S, El Naqa I. Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma. Int J Rad Oncol Biol Phys. 2004;58(4):1106–13. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360301603020236
15. Blanco AI, Chao KC, El Naqa I, Franklin GE, Zakarian K, Vicic M, Deasy JO. Dose-volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy. Int J Rad Oncol Biol Phys. 2005;62(4):1055–69. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360301605000337
16. Deasy JO, Naqa IE. *Image-based modeling of Normal tissue complication probability for radiation therapy*. Boston: Springer; 2008. p. 211–52. https://doi.org/10.1007/978-0-387-36744-6_11.
17. El Naqa I, Suneja G, Lindsay PE, Hope AJ, Alaly JR, Vicic M, Bradley JD, Apte A, Deasy JO. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. Phys Med Biol. 2006;51(22):5719. [Online]. Available: http://stacks.iop.org/0031-9155/51/i=22/a=001
18. B. D. W. Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Therapeut. 2001;69(3):89–95. https://doi.org/10.1067/mcp.2001.113989.

19. El Naqa I, Craft J, Oh J, Deasy J. Biomarkers for early radiation response for adaptive radiation therapy. Adapt Rad Ther. 2011:53–68.
20. El Naqa I. A guide to outcome modeling in radiotherapy and oncology: listening to the data. Boca Raton: CRC Press; 2018.
21. Webb S. The physics of three dimensional radiation therapy: conformal radiotherapy, radiosurgery and treatment planning. Boca Raton: CRC Press; 1993.
22. Joiner MC, Van der Kogel A. Basic clinical radiobiology fourth edition. Boca Raton: CRC press; 2009.
23. Munley MT, Lo JY, Sibley GS, Bentel GC, Anscher MS, Marks LB. A neural network to predict symptomatic lung injury. Phys Med Biol. 1999;44(9):2241–9.
24. Su M, Miften M, Whiddon C, Sun X, Light K, Marks L. An artificial neural network for predicting the incidence of radiation pneumonitis. Med Phys. 2005;32(2):318–25.
25. Gulliford SL, Webb S, Rowbottom CG, Corne DW, Dearnaley DP. Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate. Radiother Oncol. 2004;71(1):3–12. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016781400300272X
26. Tomatis S, Rancati T, Fiorino C, Vavassori V, Fellin G, Cagna E, Mauro FA, Girelli G, Monti A, Baccolini M, Naldi G, Bianchi C, Menegotti L, Pasquino M, Stasi M, Valdagni R. Late rectal bleeding after 3d-CRT for prostate cancer: development of a neural-network-based predictive model. Phys Med Biol. 2012;57(5):1399–412.
27. Vapnik V, Vapnik V. Statistical learning theory, vol. 1. New York: wiley; 1998.
28. El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. Phys Med Biol. 2009;54(18):S9–S30.
29. El Naqa I, Deasy JO, Mu Y, Huang E, Hope AJ, Lindsay PE, Apte A, Alaly J, Bradley JD. Datamining approaches for modeling tumor control probability. Acta Oncol. 2010;49(8):1363–73.
30. El Naqa I. Machine learning methods for predicting tumor response in lung cancer. WIREs Data Min Knowled Discov. 2012;2(2):173–81.
31. Oh JH, Craft J, Lozi RA, Vaidya M, Meng Y, Deasy JO, Bradley JD, El Naqa I. A bayesian network approach for modeling local failure in lung cancer. Phys Med Biol. 2011;56(6):1635–51.
32. Lee S, Ybarra N, Jeyaseelan K, Faria S, Kopek N, Brisebois P, Bradley JD, Robinson C, Seuntjens J, El Naqa I. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. Med Phys. 2015;42(5):2421–30.
33. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, De Neve W, Lievens Y, Lambin P, Dekker ALAJ. Comparison of bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys. 2010;37(4):1401–7.
34. Luo Y, El Naqa I, McShan DL, Ray D, Lohse I, Matuszak MM, Owen D, Jolly S, Lawrence TS, Kong F-MS, Haken RKT. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via bayesian network analysis. Radiother Oncol. 2017;123(1):85–92. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167814017300634
35. Luo Y, McShan DL, Matuszak MM, Ray D, Lawrence TS, Jolly S, Kong F-M, Ten Haken RK, El Naqa I. A multi-objective bayesian networks approach for joint prediction of tumor local control and radiation pneumonitis in nonsmall-cell lung cancer (nsclc) for response-adapted radiotherapy. Med Phys. 2018;45(8):3980–95.
36. Luo Y, McShan D, Ray D, Matuszak M, Jolly S, Lawrence T, Kong F, Ten Haken R, El Naqa I. Development of a fully cross-validated bayesian network approach for local control prediction in lung cancer. IEEE Trans Rad Plasma Med Sci. 2018;3(2):232–41.
37. Cui S, Luo Y, Tseng H, Haken RKT, El Naqa I. Artificial neural network with composite architectures for prediction of local control in radiotherapy. IEEE Trans Rad Plasma Med Sci. 2018;3(2):242–9.

38. Cui S, Luo Y, Tseng H-H, Ten Haken RK, El Naqa I. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. Med Phys. 2018;46(5):2497–511.
39. Kosorok MR, Moodie EEM. Adaptive treatment strategies. In: Moodie EEM, Kosorok MR, editors. Practice. Philadelphia: Society for Industrial and Applied Mathematics; 2015.
40. Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy S. A "smart" design for building individualized treatment sequences. Annu Rev Clin Psychol. 2012;8(1):21–48.
41. Berry DA. Adaptive clinical trials in oncology. Nat Rev Clin Oncol. 2012;9(4):199.
42. Kim M, Ghate A, Phillips MH. A markov decision process approach to temporal modulation of dose fractions in radiation therapy planning. Phys Med Biol. 2009;54(14):4455–76.
43. Vincent RD, Pineau J, Ybarra N, El Naqa I. Chapter 16: Practical reinforcement learning in dynamic treatment regimes. In: *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. Philadelphia: Society for Industrial and Applied Mathematics. p. 263–96.
44. Andrew AM. Reinforcement learning: an introduction by richard s. Sutton and Andrew g. barto, adaptive computation and machine learning series, mit press (Bradford book), Cambridge, mass., 1998, xviii+ 322 pp, isbn 0-262-19398-1,(hardback,£ 31.95). Robotica. 1999;17(2):229–35.
45. Sutton RS, McAllester DA, Singh SP, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In: *Advances in neural information processing systems*; 2000. p. 1057–63.
46. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989;2(5):359–66. http://www.sciencedirect.com/science/article/pii/0893608089900208
47. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–33.
48. Kong F-M, Ten Haken RK, Schipper M, Frey KA, Hayman J, Gross M, Ramnath N, Hassan KA, Matuszak M, Ritter T, Bi N, Wang W, Orringer M, Cease KB, Lawrence TS, Kalemkerian GP. Effect of midtreatment PET/CT-adapted radiation therapy with concurrent chemotherapy in patients with locally advanced non-small-cell lung cancer: a phase 2 clinical trial. JAMA Oncol. 2017;3(10):1358–65.
49. Tseng H-H, Luo Y, Cui S, Chien J-T, Ten Haken RK, Naqa IE. Deep reinforcement learning for automated radiation adaptation in lung cancer. Med Phys. 2017;44(12):6690–705. https://doi.org/10.1002/mp.12625.
50. Ribeiro MT, Singh S, Guestrin C. â€œwhy should i trust you?: Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD â€™16. New York: Association for Computing Machinery; 2016. p. 1135–44. https://doi.org/10.1145/2939672.2939778.

Artificial Intelligence in Clinical Trials

19

Hina Saeed and Issam El Naqa

19.1 Introduction

The overall survival and quality of life for many cancer patients has improved dramatically due to the advancement in surgeries, devices, radiation techniques, systemic therapies including targeted agents, statistical tools, and evolution of clinical trials. It has been repeatedly estimated that 3–5% of adult cancer patients enroll in cancer clinical trials [1, 2]. This figure has now been estimated to be 8% [3]. Conversely, the vast majority of adult cancer patients (>95%) do not participate in clinical trials, even though 70% of Americans are estimated to be inclined or very willing to participate in clinical trials [4]. In order to improve the trajectory of oncology research, it is imperative to employ innovative ways to close the existing large gap between trial participation rates and the willingness of patients to participate as well as improve quality and patient safety through federal and international policies and ethics codes in clinical trials (Fig. 19.1).

19.1.1 Background on Clinical Trials in Oncology and Radiology

In order to maintain a quality environment for patient care in clinical trials, it is vital to understand the history of clinical trials, including successes, failures, and the risk for patient endangerment. Early experiments in oncology on human subjects can be traced to Rudolf Virchow's work in 1863 tracking cancer to its cellular origin by

H. Saeed (✉)
Department of Radiation Oncology, Medical College of Wisconsin, Milwaukee, WI, USA
e-mail: hisaeed@mcw.edu

I. El Naqa
Department of Machine Learning, Moffitt Cancer Center, Tampa, FL, USA
e-mail: issam.elnaqa@moffitt.org

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2_19

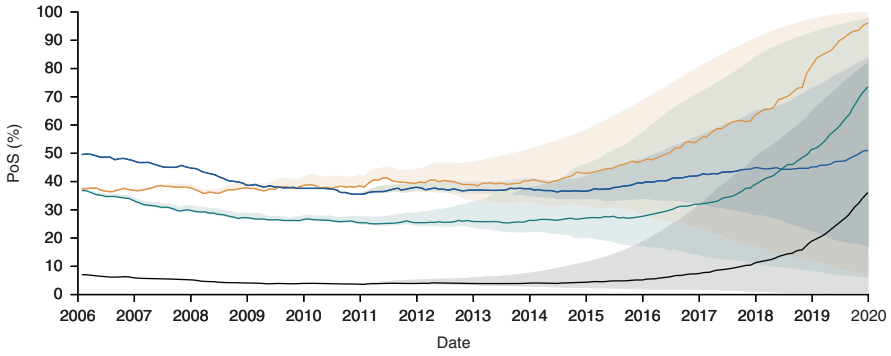


Fig. 19.1 State of clinical trials. Success rate declined for much of the period between 2008 and 2015, but since 2015, it seems trend has begun to reverse. Shaded areas represent the estimated probabilities of success (PoS) if all clinical trials yet to be completed failed (lower bound) or if all succeeded (upper bound) (right). *A* approval. PoS phase 1-2: blue. PoS phase 2-3: green. PoS phase 3-A: orange. PoS phase 1-A: black. (Data from Project ALPHA [5])

using a microscope [6]. A one-room laboratory in 1887 (later became National Institute of Health (NIH)) was established and National Cancer Institute (NCI) was formed in 1937 [6]. In the mid-1950s, NCI began to fund cooperative oncology groups in an effort to expand enrollment in clinical trials. The Clinical Trials Cooperative Group Program was originally composed of four pediatric and nine adult groups. The initial consolidation occurred in 2000 when the four pediatric groups became one group and the next one occurred in 2014, when the nine adult groups were merged into four adult groups. The Cancer Therapy Evaluation Program (CTEP), oversees the cooperative oncology groups [7].

By 1973, most oncology clinical trials were conducted at NCI-designated comprehensive cancer centers that received core grants from NCI to fund operations. According to NIH factsheet, approximately 80–85% of patients with cancer are seen and treated at community cancer centers or hospitals near their home communities with access to a wide array of clinical trial opportunities [3, 8]. In 2013, NCI Community Oncology Research Program (NCORP) was formed to bring state-of-the-art cancer prevention, control, treatment, and imaging clinical trials; cancer care delivery research; and disparities studies to individuals in their own communities. In 2004, the U.S. Food and Drug Administration (FDA) introduced the Critical Path Initiative with a goal of accelerating translation of basic research to safe and effective medicine and treatment options for patients [9]. This fostered identification of various biomarkers and other tools used to improve patient outcomes and survivorship rates and uncovered the potential to treat patients with targeted therapy, based on biomarkers and molecular abnormalities.

Fraudulent claims of safety and efficacy related to drugs and devices were rampant in the United States in the late 1800s, resulting in serious injuries and deaths and prompting a series of actions that started with the Food and Drug Act in 1906. Other major events included Declaration of Helsinski (international ethical

guidelines) in 1964 [10], Belmont Report in 1979 that framed the concept of institutional review boards (IRBs), outlined protocol design criteria, and laid the recommendation for obtaining informed consent from all research subjects [11], Common Rule [12], and Health Insurance Portability and Account-ability Act (HIPAA) in 2003. Federal Wide Assurance for the Protection of Human Subjects (FWA) was passed in 2005 enforcing that all research involving nonexempt human study participants be subject to federal regulations and must be guided by ethical principles that includes the Belmont Report and Common Rule [13].

Lack of access to state-of-the-art healthcare; cultural or ethnic factors; economic status; language or literacy barriers; and long-standing fear, apprehension, and skepticism have been identified as obstacles to minority participation in clinical trials [12, 14]. Despite FDA mandated exclusion in 1977 of women in their childbearing years from participation in phase I clinical trials because of concerns about the potential teratogenic effects, in practice, the exclusion was extended to all women in all phases of clinical trials [15]. These policies severely limit knowledge about gender- and race-related differences in drug safety and efficacy [15–17], where members of racial and ethnic minorities, low-income individuals, and people who live in rural areas remain underrepresented [18]. The NIH Revitalization Act of 1993 mandated the inclusion of women and minorities in all NIH-sponsored clinical trials [19].

Since children represent a vulnerable population, special protections have been implemented to safeguard their treatment [20–22]. Additionally, children younger than 18 years old are now asked for their assent if they are mature enough to understand the expectations of the study trial. Although assent, unlike informed consent, is not required by law, many IRBs require it.

Cancer incidence and mortality rates are highest in elderly population. Despite FDA recommending guidelines for inclusion of older adults in clinical trials in 1989, they continue to be proportionally underrepresented in clinical trials [18]. Suggested reasons for underrepresentation include concerns about toxicities, the presence of comorbid conditions, perceived lack of benefits, advanced stage of disease at diagnosis, lack of awareness, quality-of-life concerns, and a variety of socioeconomic barriers [23–25]. The exclusion of the elderly not only limits the generalizability of results but can also be costly as elderly will not be treated effectively due to misconceptions about tolerance [23, 25, 26] as seen in the clinical example of Glioblastoma multiforme (GBM) [27]. Today, cooperative group trials, such as treatment (e.g., chemotherapy), quality-of-life trials, and registries, are designed specifically to include older adults.

In 2010, with the intent to improve access to care, the Patient Protection and Affordable Care Act (PPACA) led to the development of Patient-Centered Outcomes Research Institute (PCORI) and introduced requirements for insurance companies to provide coverage for routine costs associated with clinical trial participation [28]. With this in place, patients are able to choose the best option for treatment without worrying that certain tests or procedures will not be covered based on their participation in a clinical trial [29].

Due to improvements in policies, technology and clinical trials, the rate of cancer deaths began to decline by 1991 and continues to do so [30]. However, a lot still remains to be done. The focus today is not only the treatment and prevention of cancer but also symptom management and quality of life, genomics, personalized medicine, and biospecimens.

The number of clinical trials for patients with cancer dwarfs that of any other single disease, with cancer clinical trials comprising 22% in 2010 using ClinicalTrials.gov database to recently between 40% and 50% of all trials conducted in the United States [31, 32]. Oncology trials are predominantly early-phase studies that evaluate surrogate endpoints. They tend to be small, single arm, and open label. This orientation toward less robust design differs significantly from trials in other areas of medicine. Despite a wide variation in treatment options and survival between cancer types, the proportion of small, single-arm studies does not vary significantly between cancer types, and there is only moderate correlation between the number of trials for a given cancer type and relative incidence or mortality. Unfortunately, the high prevalence of small trials that lack comparator arms, rely on historical controls, and lack randomization limit the ability to assess the evidence supporting specific treatments through systematic reviews and comparative effectiveness research. Of note, this registry is not complete and suffers from a lack of standard ontology [33].

In order to effectively leverage limited resources, it is paramount to accurately characterize the current state of clinical research and available technology. Subsequent insights and metric development will allow us to monitor the activity and advance innovative approaches. The research community must take into account competing priorities, the importance of particular research questions, the urgency of disease, and the availability of trials across geographic regions and disadvantaged populations. It is essential to develop not only breakthrough treatments, but also improve the use of existing treatments. As nations such as the United Kingdom attempt to coordinate their approach to clinical research to align it better with public health priorities, we have an ongoing national debate in the United States regarding research priorities [34]. With the reorganization of the cooperative group system and implementation of PPACA, it is essential to prioritize research questions appropriately, understand the ideal mix of trials and be open to innovative trails that are more applicable in the modern era so as to optimize the generation of actionable evidence. Success from a financial and clinical research standpoint is going to become increasingly reliant on big data (including real-world data) dependent predictive analytics, real-time clinical decision support, precision medicine, and proactive population health management and these are driven largely by groundbreaking research in artificial intelligence (AI), which promises to transform the current clinical trial landscape.

19.1.2 Clinical Trials as the Gold Standard for Clinical Practice

Randomized controlled trials (RCTs) are the reference standard for driving clinical practice. RCTs measure the effectiveness of a new intervention or treatment [35]. In

order to provide a true, reliable assessment of effectiveness, RCTs need to be conducted prospectively and robustly. In a commentary describing six consecutive series of Phase III controlled trials, it was noted that when randomized care replaces random, opinion-based care, incremental progress can be anticipated in any specialty [36].

Although no study is likely on its own to prove causality, randomization reduces inherent bias and provides a rigorous tool to examine cause–effect relationships between an intervention and outcome. This is because the act of randomization in a large study balances participants' characteristics (both observed and unobserved) between the control and treatment groups, allowing attribution of any differences in outcome to the intervention [37].

All RCTs should be prospectively registered with a clinical trials database to avoid selective reporting [38]. When designing a clinical trial, the initial step is to carefully select the population, the interventions to be compared and pre-specified outcomes of interest. Once these are defined, the number of participants needed or the power calculation, the time scale of the study and the statistical and qualitative methods for analysis are determined. All RCTs should have appropriate ethical approvals and a trial protocol documenting full details of all trial processes. With appropriate ethical approvals in place, participants are then recruited, stratified (if needed) and randomly assigned (such as computer-generated randomization) to either the intervention or the comparator group. Following randomization, RCTs can be blinded if feasible so that the participants, doctors and nurses as well as researchers do not know what treatment each participant is receiving, further minimizing bias [39, 40].

RCTs can be analyzed by intention-to-treat analysis (ITT; subjects analyzed in the groups to which they were randomized), per protocol (only participants who completed the treatment originally allocated are analyzed), or other variations, with ITT often regarded least biased [41]. Adherence to the CONSORT (CONsolidated Standards of Reporting Trials) 2010 guideline enables readers to understand a trial's design, conduct, analysis and interpretation, and to assess the validity of its results. It contains a 25-item checklist and flow diagram template to improve the reporting in both groups of parallel RCT enabling [42].

RCTs can have their drawbacks, including their high cost in terms of time and money, problems with generalizability (participants that volunteer to participate might not be representative of the population being studied) and loss to follow up [43]. Besides careful conduction and interpretation of the study, potential conflict of interests and funding sources should be disclaimed [44]. Finally, the principle of equipoise should be met prior to and during the conduction of an RCT [45].

19.1.3 Why Do Clinical Trials Fail?

Clinical trials are prone to many opportunities for failure. The investments of resources, time, and funding grow with successive stages, from preclinical through phase III. Thus, the cost of a failed phase III trial is not just the cost associated with

the trial itself but the cost of all prior trials as well as the cost of lost time pursuing an opportunity to advance patient treatment and the efforts of the enrolled patients. It is important to maintain a philosophy of continual improvement with respect to clinical trials broadly and specifically with an aim toward optimizing every aspect of the research to justify merit in continuing to the next stage. Failures can arise due to multiple reasons—ranging from a lack of efficacy, issues with safety, a lack of funding to complete a trial, inability to maintain good manufacturing protocols, not following FDA guidance, or issues with patient recruitment, enrollment, and retention [46].

The primary source of trial failure has been and remains an inability to demonstrate efficacy. An assessment of 640 phase 3 trials with novel therapeutics found that 54% failed in clinical development, with 57% of those failing due to inadequate efficacy, 17% due to safety concerns and 22% failed due to commercial reasons. There are many reasons that potentially efficacious drugs can still fail to demonstrate efficacy, including a flawed study design, an inappropriate statistical endpoint, or an underpowered clinical trial which may result from insufficient enrollment and retention. Cancer drugs were significantly less likely to gain Food and Drug Administration (FDA) approval [47].

Clinical trials also fail with respect to safety. Safety is addressed in every clinical trial in every phase, but issues with safety may only become apparent with the larger populations associated with phase 3 studies, or at post-approval (phase IV) or post-market phase [47]. The appearance of rare toxic side effects (with the use of approved drugs) found as a result of spontaneous reporting and other unreliable detection methods has led to increased attention to phase IV research. This phase employs much more rigorous research methodology involving large datasets with obligatory and uniform reporting that can be queried in near real time to provide information on real-world efficacy and toxicity data. The need is further increased by the development of accelerated pathways to drug approval especially in areas without any effective treatments, such pathways lead to licensure without rigorous clinical efficacy data [48]. It is important also to recognize the desire for a sponsor to move a drug or device forward in the clinical trial process. Rushing studies into phase 3 after successful phase 2 trials may not provide time for reflection on how best to address safety in phase 3 and can be detrimental form a cost perspective as well [49].

Identifying safety issues is not always straightforward. People may have a greater propensity to present for care when they experience an adverse event that is of concern to them, and not necessarily when experiencing an adverse event of less concern to them but greater concern to the physician. This can influence which adverse events are reported, particularly if they are mild to moderate in severity. Reminding patients of the importance of reporting any adverse events and recording the patient-reported tolerability is recommended for improving the likelihood of detecting safety issues earlier rather than later [50].

Considering the huge cost involved, many trials (in phase 3, but also earlier) are underfunded, and may not have any reasonable opportunity to generate a positive outcome (even if protocols are amended, at additional cost) [51]. More generally,

particularly in the United States, the cost of complying with an increasing regulatory burden is also impactful, necessitating more staff, storage, and financial expenditure [52]. Underfunded trials are by definition more likely to be underpowered and thus, fail to demonstrate statistical significance at a predefined level of efficacy.

Such a premature discontinuation of trials for strategic reasons is ethically flawed as it deceives the patients, jeopardizes the patient-doctor relationship, and harms the medical community. Patients generally have an expectation that their participation in a trial will lead to an advancement of knowledge based on the trial's successful completion. Effort should be made to have patient representatives on steering committee and decrease or eliminate the involvement of the sponsor to limit the risk of premature discontinuation [53, 54].

In an ideal world, the inclusion/exclusion criteria should result in enrolling population that matches statistically the intended general patient population [55, 56]. However, researchers may have concerns such as the presence of multiple comorbidities, leading to additional risk of withdrawal and adverse events. They must consider the availability of competing therapies when designing a study. Many oncology studies have very specific inclusion/exclusion criteria based on prior treatment. Targeted treatments will exacerbate this issue as diagnostics screen out more individuals. Exclusion criteria are often presented without an explicit rationale including an attempt to exclude participants who may not show sufficient improvement against an endpoint, because their health is too good [57]. These factors can affect the duration and cost of a trial and eventually result in protocol amendment [58]. As many as 16% of protocol amendments are due to changes in inclusion/exclusion criteria, which can lead to differences in the patient populations before and after the amendment. Furthermore, across 3400 clinical trials, more than 40% had amended protocols prior to the first subject visit, delaying trials by 4 months [59, 60]. Some protocol amendments cannot be avoided; however, the potential for amendments can be reduced with better planning and anticipation of the consequences from design choices.

Despite the often seen patients' willingness to consent to participation in a clinical trial based on a belief that they might receive better treatment or the results of the trial can help others [4, 61–63], enrolling a sufficient number of subjects in a trial continues to be a long-standing problem [64, 65]. A UK study indicated that only 31% of the trials met enrollment goals [66]. Studies indicate that between 18% and 40% of centrally sponsored NCI trials fail to meet sufficient accrual goals to answer the study questions, with somewhere between several hundred to a thousand patients per year enrolled in these trials [67–69]. A broader study looking at cancer trials across all sponsors similarly found a rate of failure due to low accrual of 20%, with more than 6800 patients per year enrolled in these failing trials [70]. In addition, Campbell et al. reported that one-third of publicly funded trials required a time extension because they failed to meet initial recruitment goals [71]. Overall, trial participation rate averages 14.8% at academic centers and 6.3% at community centers. Over half of patients will not have a local trial available as a result of decisions about which trials an institution opens. Forty percent of patients with trials available (17% of total) will not be eligible to enroll on a trial due to eligibility requirements

established during the trial's design. Ultimately, 8% will enroll in a trial and 18% will not enroll. Multiple studies show that around 30% of eligible patients will not be asked to participate. Only a small fraction of patients overall ever has the opportunity to consent to a request to participate in a clinical trial, and when asked, over half typically agree [3, 69, 72].

Some studies offer remuneration to patients, generally to cover the patients' time and expenses but also in the hope that recruitment will be improved. While logic suggests that this might improve recruitment and patients sometimes report this as being important to them [73], evidence supporting this has been generally inconclusive [74–76]. However, Edwards et al. found that monetary incentives increased participant response to postal and electronic questionnaires [77]. Surveys show, however, that a high remuneration is often associated in patients' minds as being associated with a perception of higher risk in the trial and thus a reluctance to enroll [73, 78]. The effect and effectiveness of remuneration may depend on many factors and should remain an open area of research. Beyond remuneration, the additional costs associated with patient recruitment can be difficult to estimate and highly variable [79]. Marketing strategies can play an important role in the financial viability of some trials [80].

Healthcare providers can have a significant impact on patient recruitment and retention. Physicians are consistently rated as the most trusted source of information by patients [81]. Surveys consistently show that patients with cancer who have enrolled on trials first heard of a trial from their physician, and provider recommendation is a leading factor in enrolling on a trial [81, 82]. A site that has historically little focus on clinical trials or presents other non-scientific impediments may lead to low investigator enthusiasm and is associated with low recruitment [83]. Non-PIs have been shown to enroll fewer patients than PIs [84]. Issues can also arise when the investigator has competing trials. Studies have shown that the quality of the discussion around clinical trials as a treatment option is highly variable; however, training can improve this conversation [85, 86].

Specialized, dedicated, in-house research personnel, have been shown to significantly increase site enrollment [87–89], and providers have reported lack of staffing as a leading barrier to enrolling patients in cancer clinical trials [90, 91]. Recruitment and retention can suffer when patients perceive support staff to be unavailable or uninterested, or if they have to interface routinely with new staff or a lack of prioritizing the clinical trial over day-to-day operations [72, 92]. Given these challenges, retention of quality research staff is imperative. Frequent turnover of staff can lead to greater numbers of inexperienced study coordinators, which can impact the data quality and timeliness of completing a trial [93]. Increasing job satisfaction and incentivizing staff (providing funds for enrolling patients) has been shown to improve patient recruitment [72, 94]. Using nurses instead of surgeons to perform recruitment has not evidenced any difference in outcomes; however, cost savings have been realized [95, 96] which may be important in supporting recruitment and retention, or other aspects of the clinical trial, indirectly.

Study centers with a track record of successful performance are historically more likely to meet enrollment targets [59]. Smaller sites, or those with few or no clinical trials, can still help patients consider and find clinical trials, but these sites typically lack the staff and infrastructure to do so [52, 97–99]. Several factors most associated with above-average recruitment rate: implementation of a systematic pre-screening of patient for trial matching, engagement of other staff, time from ethics approval to first recruit, and the provision of a dedicated trial coordinator [100, 101]. Proper site selection in terms of having a nearby larger population pool and minimizing total time investment including travel time [102] has been correlated positively with the likelihood of meeting recruitment targets patient recruitment and retention is affected negatively when patients are concerned about being assigned to a control group rather than receiving active study drug [103–105]. Part of this effect may be due to patients having poor knowledge about placebos or what specific treatment is given in the control group [106]. For patients with poor prognoses, the concern may center around not having effective treatment at all. Other concerns the patients might have is fear of side effects [90], logistical challenges in terms of time and travel [90, 107–109] especially if elderly [110] and any associated costs such as lost work or medical expenses [104, 108, 111, 112]. The cost of participating under these circumstances may biase participation to those in higher socioeconomic levels [25, 113–115]. Studies also show that the financial impact of some trials can adversely affect patient adherence as well as retention [116, 117]. Sometimes patients are not presented with a clear rationale for why their participation is important and receive minimal feedback. Education may also help patients feel more inclined to participate in clinical research [118]. Encouraging patient trust in the clinical trial process may be expected to lead to better participation [119]. Ineffective scheduling and waiting time have been associated negatively with patient satisfaction [120, 121]. Sources of stress such as those associated with long waiting time [122] or being on-site without reciprocal empathy would be intuitively associated with lower retention [123].

Informed consent is often lengthy with complex language and concepts and is often not understood by patients [123–127]. Scientific literacy in the general population can be limited, leading to difficulty understanding information associated with a clinical trial [128, 129]. Individual site informed consent processes vary, with research showing that more interactive processes involving videos, questions and answers, and more human interaction can result in greater understanding and improve patient satisfaction [130–133].

Decreased retention may underpower a trial. This may prompt a sponsor to adapt by expanding the number of sites (costly protocol amendments and delays), increasing the allocated funds to meet minimum enrollment or even close the trial. By consequence this sometimes necessitates eliminating certain planned tests in order to reallocate available funds. In turn, certain endpoints may have an insufficient sample size to detect an important result [83]. This in turn, affects the trial ethics as patients know that their results will not be likely to contribute to a statistically significant outcome [134, 135].

19.2 Types of Clinical Trial Design

In oncology, patients are generally classified by their primary cancer and stage, and randomized controlled trials are conducted to create standard therapies. Historically, cytotoxic agents have been developed based on this perspective. However, research and development in the past two decades enabled cancer cell growth and progression to be defined at cellular and molecular levels, and the presence or absence of molecular markers or genetic mutations enabled detailed classification of particular tumor types into several subtypes. Similarly, there were developments in the chemotherapeutic drugs, shifting from treatments centered on cytotoxic agents to those using molecularly targeted agents, which act selectively on cancer cells. Recently, there is active research on immune checkpoint inhibitors [136], molecularly targeted agents that target specific molecular markers such as *EGFR* gene mutation-positive inoperable, recurrent, or metastatic non-small cell lung cancer [137], as well as *ALK* fusion gene-positive non-small cell lung cancer [138].

With the aforementioned changes, biomarker-based clinical trial designs started to rise in popularity over the past two decades. Trials based on biomarker-strategy design such as ERCC1 [139] entailed randomly assigning patients to an experimental treatment arm that uses the biomarker to determine therapy versus standard therapy or to a control arm that uses standard therapy. Trial designs such as "enrichment designs" or "targeted designs" included patient population with a single molecular marker for which a drug's effects can be expected in a specific tumor type. The biomarker is evaluated on all patients, but random assignment is restricted to patients with specific biomarker values. The established molecular marker could be evaluated with a diagnostic tool and would have a strongly correlation with the efficacy of the investigational drug. Additionally, in marker-negative cases, the drug is expected to have no efficacy from a biologic standpoint. Examples include clinical trials involving trastuzumab [140, 141] and CALGB 10603 [142]. If the molecular marker is not established as a reliable marker, the use of a marker-stratified design may be considered. In this design, patients were assigned to arms by molecular marker positivity or negativity and were randomized within each arm. Clinical trials that used the marker-stratified design include the INTEREST [143] and MARVEL [144, 145] trials. After this type of design was introduced, sequential subgroup-specific, marker sequential test (MaST) and fallback designs were proposed as extensions of the marker-stratified design [146]; this eventually led to the proposal of clinical trials that use the master protocol design.

With the advent of next-generation sequencing and comprehensive genomic profiling in oncology, there has been significant enthusiasm to pursue the concept of personalized or precision medicine with an aim to use tailored therapies to target specific genetic changes that cause the tumor to develop [147]. However, it is unrealistic to conduct phase I–III trials with adequate power according to each subpopulation based on patient molecular subtypes [148–152]. Common protocols that assess the combination of several molecular markers and their targeted therapies by means of multiple sub-studies for single and/or multiple tumor types are required.

These protocols are called "master protocols," and are drawing attention as a next-generation clinical trial design.

A master protocol is a comprehensive protocol created for evaluating multiple hypotheses of sub-studies that are concurrently conducted. It comprises of different sub-protocols of multiple concurrently operating sub-studies (Fig. 19.2), where the sub-studies are commonly conducted on populations based on specific tumor types, histologic types, and/or molecular markers. Master protocol trials can be exploratory or confirmatory. Exploratory master protocol trials are often composed of multiple single-arm sub-studies, and confirmatory master protocol trials are composed of multiple randomized sub-studies. For either trial type, the design and statistical considerations are commonly standardized between all sub-studies [153–155].

A master protocol trial uses a common centralized system for patient selection, logistics, templates, and data management. In order to collect standardized data, histologic and hematologic specimens of enrolled patients are measured and analyzed using a common basic system (e.g., next-generating sequencing and immunohistochemistry) to collect coherent molecular marker data. Patients can participate in sub-studies for which they meet eligibility criteria based on their molecular marker data. Thus, enrolling in a master protocol trial increases the chance of trial participation for which the patients can expect optimal therapeutic effects. Importantly, even if there are no sub-studies that a given patient can participate in, they will be followed-up, and can be placed on a waiting list until an appropriate sub-study is started. Furthermore, natural history data from a waiting list can be used as controls in evaluating the efficacy of an investigational drug in a single-arm sub-study [151]. On the other hand, the challenges associated with master protocol

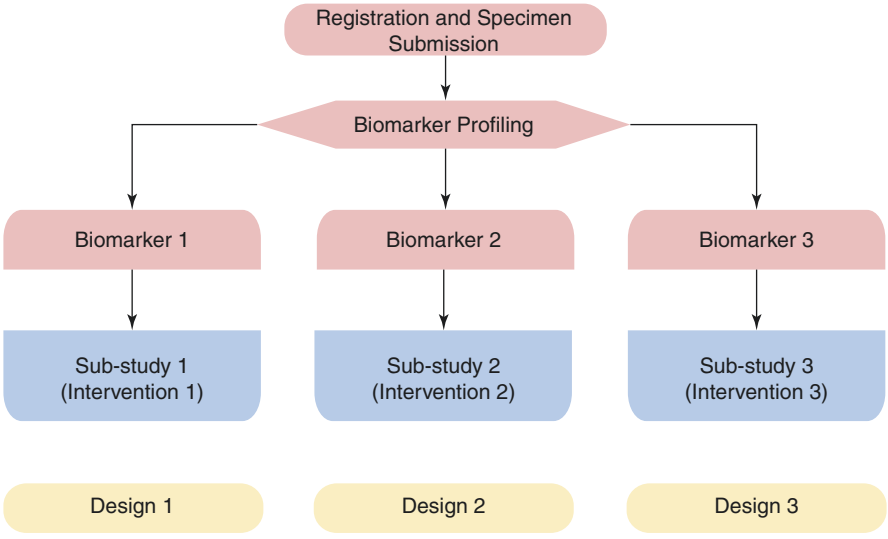


Fig. 19.2 Master protocol schema

trials, include the fact that several small sub-studies are being conducted in parallel, which may increase the rate of false positive findings.

A master protocol can alleviate some of the modern era challenges faced by clinical trials [152, 156]. First, classic trial design paradigms are challenged by inter-patient and intra-patient heterogeneity, as they are unable to test targeted therapeutics against low frequency genomic "oncogenic driver" aberrations with adequate power. These low frequencies of any given molecular driver further exacerbate clinical trial accrual difficulties. Identical tumor types can exhibit different responses to treatments depending on patient characteristics or disease stage. Even within the same patient, intratumor heterogeneity or variations in stromal tissue can generate a different treatment response [157, 158]. A master protocol enables efficient enrollment of rare fraction patients so that centralized patient management, based on a common protocol, promotes the acceleration of clinical development and trial data from multiple sub-studies can be comprehensively used to evaluate inter- and intra-patient heterogeneity. Second, findings on specific signal pathways strongly associated with driver gene mutations (oncogene driver and oncogene addiction) and cancer cell growth and progression can be obtained [159–161]. Third, combining two or more targeted therapies makes it possible to expand the genetic mutations being studied [153, 162, 163].

A master protocol trial is often classified into basket, umbrella, and platform trials based on characteristics of the study population (e.g., disease, histologic type, molecular marker) and on both the type and number of study therapies (Table 19.1, Fig. 19.3). The trial definitions are not standardized and have some overlap [151, 152, 156, 164–166].

Basket trial in oncology are often conducted as single-arm, phase II, usually non-randomized trials with the purpose of evaluating proof-of-concept (POC) in an early stage of development. It examines the therapeutic effects of molecularly targeted agents for several tumor types that may have a common single molecular marker, or genetic mutation, by tumor type and/or across tumor types. Each arm is a separate "basket" that assigns small cohorts of patients and focuses on testing one treatment against a specific target, regardless of cancer types [152]. The term "basket" refers to the fusion of potentially different cancers (according to the common classification by the body organ where they begin [166] or by their histological type of origin [167, 168] into one similar disease at the molecular level. The absence of a control group is a limitation in evaluating therapeutic effect; thus, it is desired to collect control data. An example of a basket trial is phase II NCI-MATCH (Molecular Analysis for Therapy Choice, NCT02465060) trial launched in 2015 by the US National Cancer Institute [169].

Table 19.1 Types of master protocol trials

Trial type	Definition
Basket	Evaluates one targeted therapy on multiple diseases or multiple disease subtypes
Umbrella	Evaluate multiple targeted therapies for one disease or several diseases
Platform	Evaluate several targeted therapies for one disease perpetually, and further accept additions or exclusions of new therapies or patient populations during the trial

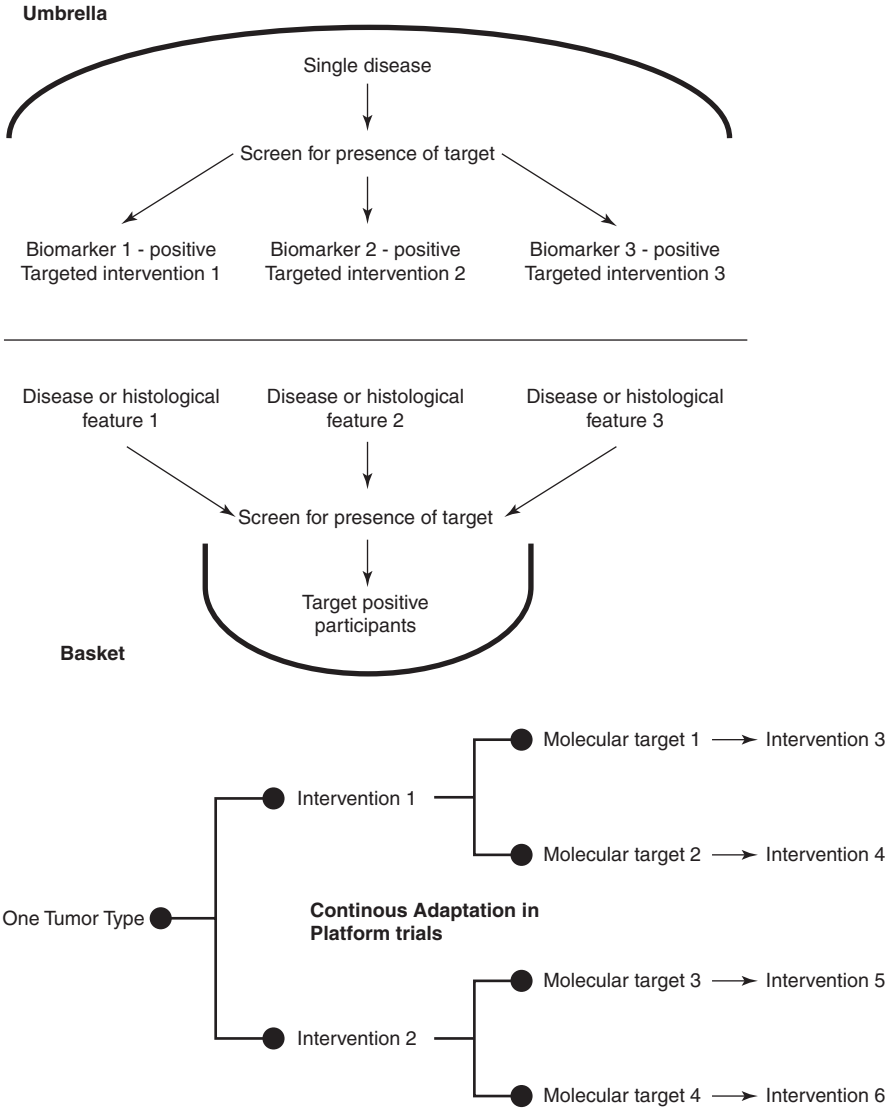


Fig. 19.3 Basket, Umbrella, and Platform designs

Basket trials are characterized by the comprehensive execution of single-arm trials with a small number of patients, which enables efficient patient enrollment for rare cancers or rare fractions. Generally, the number of participants in individual sub-studies are between 20 and 50, and hypotheses that can demonstrate statistical significance are made only when there is major therapeutic efficacy; therefore, a basket trial is considered a "signal-finding" trial. For sub-study designs, two-stage or multistage designs may be used [156].

Notably, basket trials are assumed to allow a fairly accurate prediction of whether a tumor with particular molecular characteristics will respond to a targeted therapy; furthermore, such response to a targeted therapy is established irrespective of the histologic type of the tumor. A number of studies have attempted to evaluate the general policy of using tumor genomics to select treatments for individual patients. Two meta-analyses of non-randomized phase II basket studies concluded that trials with a "personalized" strategy resulted in a higher proportion of patients achieving responses and longer PFS and overall survivals [170, 171]. The SHIVA result highlights the fact that using tumor genomics to guide therapy is not always useful [172]. Ignoring the possibility that activity of the drug can depend on the histology of the tumor as well as the genomic alteration can have implications as suggested by vemurafenib trial for patients harboring V600E BRAF mutation and the drug was active in NSCLC and several other histologies, but not in colorectal cancer [173]. Consequently, in designing basket trials, great care must be taken in defining the genomic alterations to be included and the matching strategy to be used [174, 175].

In order to counter the aforementioned issue, a drug regimen can be evaluated separately for each histology or each histology group by increasing the sample size to conduct a separate two-stage phase II design for each histology. However, this can become inefficient if the drug is found to be active or inactive in all the histology groups. A suggested approach that improves efficiency is the parallel evaluation of each histology by assessing the homogeneity of the response rates for the histology groups during an interim analysis. If the response rates seem heterogeneous, then a two-stage phase II design is conducted separately for each histology. If the response rates seem homogeneous, then a single two-stage phase II design is conducted with the histologies pooled [164].

Basket trials often explore a range of potential off-label uses for a drug approved in one histology for patients with a genomic abnormality. For some of these rare histologies, the response rate may be sufficiently large that if the responses are durable the evidence from the basket trial may be sufficient for extension of the approved indication. A suggested approach involves pooling the rare histologies that are found sensitive to the drug in the basket trial in a phase III licensing trial in which the test drug is randomized against best available treatment controls. This would allow extending indications to several rare histologies simultaneously [176].

Patients enrolled in each sub-study in a basket trial are often composed of a heterogeneous group in terms of tumor type, histologic type or patient characteristics. Therefore, as it is difficult to evaluate time-to-event endpoints (e.g., progression-free survival or overall survival). Thus, primary endpoints are often response rates, which are less sensitive to the effects of population heterogeneity [176].

Umbrella trials in oncology evaluate multiple targeted therapies that correspond to different molecular markers or genetic mutations within a particular tumor type [166]. Sub-studies may be single arm, phase II, or phase II/III trials that are randomized and compared to placebo or a standard therapy. The term "umbrella" refers to separation of one alleged cancer into many sub-cancers depending on their molecular features. There is also a "default arm" which assigns patients without a specific marker to receive standard treatment. Umbrella trials have in common a system that

unifies molecular profiles of patient specimens for evaluation. An example of an ongoing trial is Lung-MAP: S1400 Phase II/III Biomarker-Driven Master Protocol for Second Line Therapy of Squamous Cell Lung Cancer (NCT02154490) study sponsored by Southwest Oncology Group (SWOG) [177].

While basket trials are generally single-arm sub-studies that are exploratory in nature, umbrella trials are often single-arm or randomized sub-studies that are confirmatory. Therefore, a randomized sub-study with appropriate eligibility criteria by tumor type and/or stage can generate confirmatory evidence related to the targeted therapy for the tumor type under study. Sometimes, patient enrollment may be slowed in the case of compartmentalization by molecular markers for examining rare cancers or rare fractions. In addition, umbrella trials are normally large-scale and long-term master protocol trials, but when the standard therapy changes during that period, the clinical significance of comparing to a control group of patients undergoing standard therapy is lost [148, 163].

Platform trials evaluate several targeted therapies for one disease perpetually, and further accept additions or exclusions of new therapies or patient populations during the trial. Basket and umbrella trials could also be considered platform trials, if they permit the addition or exclusion of new treatments during the trial. In a platform trial, interim analyses evaluate the efficacy or futility of each targeted therapy, and their results are used to exclude certain targeted therapies or to add new ones. Futility is often evaluated by the Bayesian method [154, 178]. Since sub-studies by molecular markers are not mutually independent trials, the efficacy of the targeted therapy of each sub-study can be estimated by a Bayesian hierarchical model [152, 154, 178]. As such, platform trials permit relatively flexible addition or exclusion of treatment methods or patient populations, thereby enabling an efficient transition to a confirmatory trial. Examples include BATTLE trials in NSCLC [179, 180] and I-SPY II trial in breast cancer [181, 182]. Challenges of the platform trial include its large-scale, long-term nature, associated high costs of management and execution of the trial, and the need to build organizations or frameworks that can operate these trials perpetually. Platform trials also run the risks of bias and prognostic imbalances as a result of adaptive randomization. Consequently, apparently positive findings from these analyses should not be accepted as correct until confirmed by an independent phase III trial.

In recent years, there have been drugs for which efficacy was not observed in a broad study population, but they did demonstrate marked efficacy in specific patients. These patients are called “exceptional responders” [183], and new initiatives are in place to elucidate their molecular profiles. Master protocol trials can also be used to identify exceptional responders and are anticipated to become one of the standard clinical trial designs to promote individualized medical care.

Although the benefit, in terms of time and resources, of using a master protocol, cannot be denied, it come at the cost of increased up-front planning and coordination to bring a larger number of parties into agreement on trial design, execution, and governance than a stand-alone trial requires. The complexity and real-time decision-making further result in the need for more up-front planning. The need for coordination amongst multiple stakeholder, appropriate infrastructure, and complex

trial design elements can extend the start-up time for a master protocol considerably, as compared with that for a single-purpose trial.

With the approval of new interventions, the master protocol should allow for coordinated design adjustments. Usually, single-purpose trials tend to adapt better to it. The adaptations might result in the need for a temporary halt in recruitment, statistical design modifications to accommodate changing comparators, or ethical issues related to patient's perception of being on an inferior arm [151].

While designing these trials, it is important to consider any associated ethical issues [184]. The scientific validity is vital for an ethically sound study. Focusing only on molecular therapy targeting single mutation without considering the complexity of tumor biology or heterogeneity, may introduce bias. Due to the innovative trial designs, patients with rare malignancies have the opportunity to be enrolled and benefit from the trial, but due to insufficient accrual, the trial may generate clinically insignificant findings [184]. Inadequate sample size in study arms and the use of surrogate endpoints may result in an approval of an intervention without confirmed efficacy [185]. Hence, analysis methods should be aligned with the research objectives during the planning stage to minimize the chances of coming to an erroneous conclusion. Moreover, complexity, limited quality and availability of tumor samples may not only cause bias and unreliability, but also can potentially harm patients by assigning them to an inappropriate therapy arm. Other threats include publication bias and lack of explanation for a protocol modification [165].

An ethical requirement of conducting clinical trials is a favorable risk–benefit ratio based on non-maleficence and beneficence and serves to protect the participants against exploitation [184]. Novel clinical trials can gain important knowledge, which can be used in future trials to develop effective therapies. However, they offer limited direct benefits to patients. The excessive use of phrases like "personalized medicine" or "precision oncology" as opposed to "genome-based therapy" during informed consent can result in misleading information or therapeutic misconception. This can falsely indicate that the trial's goal is to provide personalized care with regard to the patient's best interest and direct therapeutic benefit rather than gathering data for contributing to scientific knowledge [165]. All potential participants must wait about 2 weeks for the results of the genetic screening, which may be stressful and produce anxiety. Moreover, surrogate endpoints do not necessarily translate to patient-centered outcomes [185, 186]. The enrollment of patients whose tumors harbor multiple mutations in treatments matching a single mutation may be controversial from an ethical standpoint [165]. Moreover, the recruitment of thousands of participants generates a huge amount of data that must not only be rapidly processed, but also reliably and safely stored, so that undesirable people have no access to it [165].

19.2.1 Adaptive Clinical Trials

The size and expense of phase III clinical trials in modern oncology continue to increase, but the success rate remains unacceptably low—only 34% of phase III

oncology drug trials with results announced from 2003 to 2010 achieved statistical significance in their primary end points [187]. The challenges associated with the modern era demand appropriate adaptation.

An adaptive design is one in which the accumulating data are used to modify the trial's course. Adaptive designs are ideal for addressing many questions at once. For instance, a single trial might identify the appropriate patient population, dose and regimen, and therapeutic combinations, and then switch seamlessly into a phase II/III confirmatory trial. Adaptive designs rely on information, including from patients who have not achieved the trial's primary end point. Longitudinal models of biomarkers (serum, molecular and imaging based) can be adaptively validated for long-term primary end point prediction. Adaptive trial designs can make development more informative by addressing whether a drug is safe and effective while showing how it should be delivered and to whom and shorten the duration of the development [188]. Examples of adaptive clinical trials include BATTLE-2, I-SPY 2 and FOCUS4 [189–191].

Both the Bayesian perspective and the more-traditional frequentist perspective can be used in statistical design of adaptive clinical trials. The Bayesian perspective facilitates building an efficient and accurate trial, including using longitudinal information adaptively. To take advantage of this attribute, researchers who favor the frequentist approach can build an adaptive design using the Bayesian perspective and then find its frequentist operating characteristics using computer simulations and then playing with the design to achieve more-desirable operating characteristics [192–196].

Predictive probability calculations based on interim monitoring can be used as part of an adaptive design and can indicate the superiority or futility of a treatment arm [144, 145, 197–199]. CALGB 49907 is an example of a trial that used predictive probability to adapt sample size [200]. Once interim monitoring commences, there is very little cost in having frequent monitoring from then on [201], so one can analyze the data more often and stop earlier (when appropriate). It accelerates public dissemination of important study results and protects patients on trials from ineffective treatments. Adaptive design allows for seamless phase I–II and phase II–III trials, increasing the overall efficiency of the process [202]. An application of interim monitoring is in the use of phase II/III designs, which can be very fast and effective (especially in the setting of multiple experimental treatments and a reliable intermediate end point) but do have the cost of having to commit earlier to the phase III question than if separate phase II and phase III trials were performed and loss of flexibility of being able to modify the phase III trial design based on the results of the phase II trial [203, 204].

The logistics of adaptive planning can be more expensive than traditional trials due to inherent complicated design [205]. The crucial outcome data that drive the adaptive aspects of the design must be deposited into a central database while patients are being accrued and followed. The increased use of real-time electronic data entry, processing, and analysis should allow for more frequent interim analyses, leading to quicker decisions. This database must be connected to the software that determines treatment assignments or other adaptive aspects of the trial. Adaptive

methods that require complex statistical modeling that is neither transparent nor reproducible should be avoided. The adaptive elements of stopping treatment arms based on unfavorable or very favorable interim monitoring results and adding treatment arms when they become relevant further increase the efficiency of the master protocol trial design. With adaptive planning, the aim is to improve on our gold standard randomized controlled trials in the modern oncology era.

19.3 Artificial Intelligence and Clinical Trial Design

19.3.1 Need for Artificial Intelligence in Clinical Trial Design

It is recognized that the success rate of clinical trials is low. An analysis of clinical trial data from January 2000 up to April 2019 estimated that less than 12% of drug-development programs ended in success [206]. As described earlier, majority of the failures of clinical trials can be traced back to lack of efficacy or safety of an intervention, flawed study design, lack of funds, retention or recruitment issues. In order to improve the success rate of clinical trial in the modern era of personalized medicine, it is imperative to understand the traditional barriers as well as the ones arising due to design innovations. Nonetheless, interventions to overcome these barriers is key to a successful and meaningful trial outcome.

The recruitment process is often the biggest barrier in clinical research and can be time-consuming and expensive. According to a 2016 study, 18% of cancer trials that launched between 2000 and 2011 as part of the US National Cancer Institute's National Clinical Trials Network failed to find even half the number of patients they were seeking after 3 or more years of trying, or had closed entirely after signing up only a few volunteers [69]. Haddad et al. found an estimated 20% of people with cancer are eligible to participate in such trials, but fewer than 5% do [207].

Every clinical trial poses individual requirements on participating patients with regards to availability (trial is available to a patient 44% of the time), eligibility (27% patients are eligible), physician motivation (70% of the patients will be asked to participate), and empowerment to enroll (8% of the patients will enroll) [3]. Eligible and suitable patients might not be properly incentivized to participate, and, even if they are, they might not be aware of a matching trial or find the recruitment process too complex and cumbersome to navigate. Moving enough patients under these tight recruitment timelines constitutes a major challenge and is in fact the number one cause for trial delays: 86% of all trials do not meet enrolment timelines, and close to one-third of all Phase III trials fail owing to enrolment problems [208]. This illustrates one of the most severe shortcomings of state-of-the-art clinical trial design: those trials with the highest patient demand suffer most from inefficient patient recruitment techniques.

A concept important to grasp is that clinical trials are usually not designed to demonstrate the effectiveness of a treatment in a random sample of the general population, but instead aim to prospectively select a subset of the population in which the effect of the intervention, if there is one, can more readily be

demonstrated. In such trial enrichment designs, inclusion of unsuitable patients will automatically decrease the observed efficacy of the drug being tested. Recruiting a high number of suitable patients does not guarantee success of a trial, but enrolling unsuitable patients increases the likelihood of its failure.

Another fundamental concept to increase the generalizability of a trial is to ensure diversity in the trial participants. In 2014, 86% of clinical trial participants worldwide were white people [209]. And a 2019 study found that 79% of genomic data comes from people of European descent [210], even though they only comprise 16% of the world's population.

Patients are often encouraged to search for a trial themselves. Typically, people rely on their doctors to inform them about suitable studies. Some patients search the website ClinicalTrials.gov, which lists more than 300,000 studies that are being conducted in the United States and 209 other countries. The complexity of trial eligibility criteria in terms of number and use of highly technical terms generally makes it intimidating for a patient. Patients often find it challenging to comprehend and assess their own eligibility. Manually extracting meaningful information from this large and unstructured data-source is a significant task that imposes a heavy processing burden on doctors and patients alike.

With the increasing generation and availability of medical big data to researchers including electronic health records, "omics" and wearables devices, it is difficult for oncologists to process all the available information that could influence outcome prediction and decision-making with traditional analytics. Success in clinical trials is going to become increasingly reliant on sophisticated machine learning (ML) algorithms to handle big data dependent predictive analytics and precision medicine. These algorithms have the potential to save billions of dollars, speed up medical advances and expand access to experimental treatments. AI- and ML-driven systems can help to improve patient cohort composition and provide assistance with patient recruitment.

Moving beyond patient cohort selection and recruitment, other important factors are a lack of technical infrastructure to cope with the complexity of running a trial. It is imperative that patients stay in the trial, adhere to trial procedures and rules throughout the trial, and that all data-points for monitoring the impact of the tested intervention are collected efficiently and reliably. Only 15% of clinical trials do not experience patient dropout, and the average dropout rate across clinical trials is 30%. A linear increase of the non-adherence rate in a trial leads to an exponential increase in additional patients required to maintain the statistical power of the outcomes [208]. These additional recruiting efforts lead to trial delays and substantial additional costs. For example, a study in which 20% of the patients are non-adherent means an additional 50% of patients need to be recruited. Similarly, 50% non-adherence rate leads to an increase in recruitment of an additional 200% of patient.

Improved patient monitoring and coaching methods during ongoing trials can be used to lower the adherence burden, make data point detection more efficient, and thus reduce dropout and non-adherence rates. To comply with adherence criteria, patients are required to keep detailed records of their medication intake and of a variety of other data-points related to their bodily functions, response to medication,

and daily protocols. This can be an overwhelming and cumbersome task, leading to on average 40% of patients becoming non-adherent after 150 days into a clinical trial [211].

Every clinical trial follows a protocol that describes its design in detail. This protocol involves tremendous resources including prior studies, and regulatory information. Any problems that arise during the trial and that require amendments to the protocol can lead to months of delays and add hundreds of thousands of dollars to the cost. A faster speed in designing a trial, writing and protocol and subsequent amendments can be achieved by the use of artificial intelligence.

The incredible need for AI to overcome the aforementioned issues is indisputable in the world of clinical trials. With many fields under its umbrella, AI is poised to overcome not only the historical roadblocks but also the challenges posed by the continuous medical innovations.

19.3.2 The Multiple Roles of Artificial Intelligence (AI) in Clinical Trial Design

The term "artificial intelligence" (AI) was coined by John McCarthy and colleagues at the Dartmouth Summer Research Project (Dartmouth College, Hanover, 1956) [212]. The use of AI in medicine dates back to the early 1970s when expert systems such as MYCIN were first introduced to provide rule-based diagnostic decision support [213]. However, several radiological applications in medical imaging preceded MYCIN [212]. Early medical AI systems relied heavily on medical domain experts to train computers by encoding clinical knowledge as logic rules for specific clinical scenarios. This not only made the system labor-intensive and time-consuming but also less flexible and more difficult to update [214]. More advanced ML systems that are capable of training themselves to learn these rules by identifying and weighing relevant features from data such as unstructured text, medical images, and EHRs emerged in the 1990s and 2000s, but were relatively slow to be adopted by the clinicians, largely because of the lack of widely available data and the fact that the early methods required intense feature-engineering efforts involving serious commitments from medical domain experts [215].

Several factors have changed this situation dramatically. The field of AI has seen multiple transformations recently, enabled by hardware improvements and availability of very large training datasets [216, 217]. Medical data in digital form is now widely available due to technological advancement. Public policy efforts such as the Public Health Monitoring and Promoting Interoperability System that strive to achieve a critical national goal of meaningful use of interoperable electronic health records throughout the United States healthcare delivery system will further contribute to enhance the use of AI [218].

Recent years have witnessed a surge in efforts as well as early proof-of-concept successes of AI in medicine, starting from medical imaging including clinical photographs, digital pathology, radiographic images [217, 219–227] to AI in clinical outcome prediction [228–237] to AI in translational oncology [238–244] and AI in

clinical decision-making [245–247]. AI techniques have advanced to a level of maturity that allows them to be employed under real-life conditions to assist human decision-makers.

The time is ripe for AI to contribute to clinical trials with a great potential to transform key steps of clinical trial design from study preparation to execution toward improving trial success rates, with a positive impact on not only time and money but also optimizing and advancing the medical and patient care.

Performing a requisite literature review for related studies remains a labor-intensive task requiring personnel with specific knowledge who can interpret the framework, criteria, and results of prior clinical trials. AI can help to overcome these shortcomings of current clinical trial design. ML, and particularly deep learning (DL) with its convoluted neural networks (CNNs) are able to automatically find patterns of meaning in large datasets such as text, speech, or images. Other AI branches, such as natural language processing (NLP) can understand and correlate content in written or spoken language such as searching through biomedical literature [248], human–machine interfaces (HMIs) allow information exchange between computers and humans and reasoning techniques convert the content into actionable recommendations for the human decision-maker [249]. These capabilities can be used for correlating large and diverse datasets such as electronic health records (EHRs), medical literature, and trial databases for enhanced trial designs.

Every clinical trial follows a detailed protocol that describes exactly how the study will be run. Any problems that arise during the trial and that require amendments to the protocol can lead to months of delays and add hundreds of thousands of dollars to the cost. AI can support a faster and more efficient process for a needed amendment as well as protocol development. A study designer must think through the implications of different inclusion/exclusion criteria (as well as objectives and endpoints) and the effects they will have on recruitment, enrollment, retention, and ultimately time and cost to completion [250–253]. The budget of a trial is limited, and therefore various trade-offs need to be considered, including not only the speed of enrollment, but the likelihood of meeting the enrollment goal. By choosing a cheaper but less expensive and more remote study center, cost can be lowered at the expense of slow recruitment. This may necessitate spending more on additional study centers, which come with additional costs of evaluating, training, protocol amendments, and trial execution. Quantifying these trade-offs can assist with making better decisions.

Fundamental to a trial design are the concepts related to perfect cohort composition, effective patient recruitment, and efficient patient monitoring. These are dependent on patient-related features such suitability, eligibility, decision-making power, and motivation, as well as trial features including datapoint monitoring, endpoint detection, compliance, and patient retention. Clinical trial enrichment and biomarker verification helps to reduce population heterogeneity by electronic phenotyping and improves prognostic and predictive enrichment. This augments the suitability of patients for the trial. Clinical trial matching is aided by automatic eligibility assessment, simplified trial description and automatic trial recommendation. Automatic event logging, encouraging compliance and datapoint monitoring,

promotes the use of a record keeping disease and study protocol diary. To improve patient retention, dropouts can be forecasted and interventions such as patient coaching are implemented to prevent it [254].

The above functionalities are enabled through individual combinations of the three main AI technologies: machine/deep learning, reasoning, and HMI, which analyzes specific sets of data sources such as electronic records (EMR), omics, internet of things (IoT), wearables, clinical trial databases, trial announcements, social media, medical literature, speech and video. Collectively, the AI is expected to improve the study outcomes related to optimized cohort composition, improved trial efficiency and success probability, decreased cost, increased retention and compliance.

In order to decrease the cost, while still aiming to optimize a clinical trial patient cohort composition, a realistic approach in the absence of a comprehensive omics and list of biomarkers, is to apply complex analytic methods to combine -omic data with EMR and other patient data, present in different location and in variable formats, to develop efficiently measured biomarkers reflective of meaningful endpoints. Branches of AI such as NLP and computer vision algorithms such as optical character recognition (OCR) can serve to automate the reading and accumulation of this evidence. With the use of prognostic and predictive features of AI including ML, DL and reasoning, a more extensive but still applicable discovery of correlations desired for optimization of clinical trials has been reported [255].

A wide variety of ML methods have recently shown substantial improvement in handling complex real-world situations to assist in electronic phenotyping and thus reducing population heterogeneity [256]. ML methods, including computational modeling, can be employed to improve the prognostic and predictive enrichment of patients in clinical trials, as evidenced by their growing utility in approximating key biomarkers [257, 258] and application in model-based clinical trial designs based on complex disease processes [259–261]. The overall productivity of oncology trials as measured by the success rates relative to trial complexity and duration, can be improved by 104% and 71% by the availability of pools of pre-screened patients and biomarker tests, respectively, by 2023 [262].

Several AI techniques can assist with clinical trial matching by automatically assessing the patient's eligibility, making trial recommendations and allowing the recruiting site to be cognizant of the patient. NLP [46, 263], reasoning techniques and ML/DL permit the systems to learn and improve on the quality of their analytic output based on an adapted underlying algorithm [216, 249]. AI-based clinical trial matching systems can perform an automatic analysis of EMR and clinical trial databases to find matches between specific patients and recruiting trials [263]. Additionally, NLP and OCR can mine publicly available online content such as, trial announcements and social media to automatically identify potential patient matches with relevant trials. Technology, such as patient matching/eligibility algorithms built into electronic medical record systems, has the potential to reduce the human workload associated with identifying eligible patient [264–266] but even with technology, staff time is required to find and enroll patients in clinical trials.

Patients can be provided with AI tools to ease the complexity associated with the process of finding a trial. This motivates the patients, keeps them informed of

pertinent trials and allows involvement with clinicians for further assessment. Such digital enrollment with AI tools enables the casting of a wider net and results in substantial improvement in more diverse recruitment. Patients prefer simplified informed consent forms and testing showed no lower level of patient comprehension of the details of the clinical trial when using a simplified form [267]. Similarly, patient education leaflets should be easily understandable. AI tools can play a role in ensuring that these items remain straightforward as well as transparent and at the appropriate education grade level. AI-based sentiment analysis can play a role in developing patient materials to provide a more compassionate tone and evoke greater patient trust [268–270] in addition to maintaining an appropriate reading level.

An open-source web tool called Criteria2Query uses AI to convert the trial eligibility criteria into standardized database query format that enabling researchers and clinicians to search databases without needing to know a database query language [271]. Software developed by Deep 6 AI, an AI-based trials recruitment company has also improvement in recruitment. Another open-source web tool, called DQueST helps the patients to make sense of eligibility criteria by reading trials on ClinicalTrials.gov and then generates plain-English questions such as "What is your BMI?" to assess users' eligibility. An initial evaluation showed that after 50 questions, the tool could filter out 60–80% of trials that the user was not eligible for, with an accuracy of a little more than 60% [272]. A digital-health company, Antidote, has developed a tool that helps people to search for trials. Similarly, IBM's Watson for Clinical Trial Matching system, in a pilot study, increased the average monthly enrolment for breast cancer trials by 80% [207]. Trials.ai, describes its AI tool as a data-driven guide to designing better trials protocols. It uses AI techniques to collect and analyze publicly available data as well as certain owned private data. The company's software can then help determine how aspects of the customer's proposed trial, such as the strictness of its eligibility criteria, might affect outcomes such as cost, length or participant retention.

With the rise of commercially available wearable sensors with medical-grade health-sensing capabilities and video monitoring as part of trial design, continuous real-time patient data can be automatically monitored, logged and reviewed. This decreases the burden on patients and serves to increase compliance. ML/DL models can then be used to analyze such data to identify relevant events and endpoints reliably and efficiently without manual patient initiated self-monitoring processes. DL algorithms coupled with at the point of sensing, ultra-low-power consumption mobile processors can be used for analyzing **time-series data** from wearable sensors [273, 274]. Wearables measuring biometric parameters using customized mobile processors and "cognitive sensing" DL models allow not only storage and transmission but also, analysis of information by filtering raw data in real-time to automatically log disease and outcome diaries, extracting actionable material, and providing patient with adaptive personalized feedback and support and thus ensuring compliance [275, 276].

With ML algorithms-based pattern recognition and segmentation techniques on medical images (from, e.g., retinal scans, pathology slides and body surfaces, bones and internal organs), faster diagnoses and tracking of disease progression [227, 277–279] is enabled. This increases trial efficiency.

Companies are developing ML models to predict which patients are at risk of dropping out of clinical trials to prevent threats to trial validity. ML algorithms based on continuous data monitoring may be used to assess patient behavior and presence/absence of adherence with the trial protocol and use the information to predict the risk of dropout for a specific patient. A timely intervention at the earliest warning signs for non-adherence can prompt the deployment of effective patient coaching techniques, coping mechanisms, trial approved modifications and remedies for any toxicities before they lead to a dropout [280]. Notably, these wearables and associated AI tools tend to be highly disease specific. The combination of these wearables with DL algorithms on smartphone platforms opens the door for wide array of opportunities to be explored. A study found that use of advanced technology-enabled non-invasive diagnostic screening (TES) using low-cost smartphones and other point-of-care medical sensors versus conventional vital signs examination can synergistically support population stratification and personalized screening [281]. This can provide support by increasing the trial efficiency.

Continuous real-time monitoring of patients and disease progression will be further permitted with advancement in healthcare-related **Internet of things** (IoT) [282]. It is imperative that such sensitive patient data and all the analysis resulting from it should be stored securely. With standardization and interoperability, the medical devices focusing on IoT can act as efficient cognitive sensors for a clinical trial. It is conceivable to see blockchain technology being used in this regard to provide secure immutable exchange of data.

During a trial follow up, artificial intelligence has the potential to reduce patient time investment regardless of constraints on study site location [283, 284]. In particular, evolutionary computer simulations algorithms can assign the most appropriate study center for each prospective patient in a trial based on patient and study center availability [285]. There is also the opportunity to match staff with patients so that patients tend to see familiar faces at each visit or alter staff based on patient desires [286]. Scheduling software can search for opportunities to reschedule patients adaptively when openings develop, making the most efficient use of the clinical trial's time. Similarly, it can be used to minimize other conflicts that a patient may have.

AI and ML approaches have a great potential in improving adaptive trial designs such as basket, umbrella and platform. These tools can facilitate enrolling patients in a trial, using large datasets to profile them using omics or biomarkers and then using real-world data (RWD) for matching them to profile-dependent interventions. Additionally, EMR data, omics and RWD such as patient-reported concerns, can be explored with ML techniques to create a more complete picture for intervention and biomarker discovery.

In master protocols and adaptive design, an AI-based framework for making and implementing decisions about which treatments to study, which to discontinue, and which to advance for further study or for regulatory submission typically involves the development of statistical models and algorithms as well as procedures to ensure the rapid flow of information among the involved parties (e.g., steering committee, sponsors, and data monitoring committee). Information from RWD-based simulations can be used to model the impact of different study eligibility criteria, the

timing of endpoint assessments, and study timelines. Medical product developers are using this to support and develop clinical trial designs and observational studies. Study site selection, eligible patient identification, and simulation of study control arms, with a potential for replacing the need to randomize a patient to a control arm in some scenarios, are some uses of RWD. ML algorithms can be employed to track patients longitudinally and identify clinically meaningful patterns from hospital EMR "data lakes". With RWD and longitudinal tracking of patients, certain traditionally conducted late-phase trials could be reduced or eliminated altogether [287].

AI occupies a significant place in the design of clinical trials and can be used in numerous ways to promote efficient and thus, successful trial completion. With innovations in healthcare and technology, rise in big data and shift toward personalized medicine, the likely role of AI will continue to expand and support advances in clinical research as summarized in Table 19.2.

Table 19.2 Opportunities for AI in clinical trials

Factor	Opportunity	Role for Artificial Intelligence
Poor study design	More complete review of literature, EMR, publicly available databases, social media	NLP and ML (structured and unstructured) of available literature, finding similar trials, trials addressing similar issues, or trials addressing different issues utilizing similar techniques, summarized for the study designer.
	Appropriate outcomes/ endpoints	NLP and ML of available literature, showing endpoints/measures/biomarkers used in other similar studies.
	Appropriate eligibility criteria	NLP and ML assessment of similar published trials to determine eligibility and suitability of inclusion and exclusion criteria.
	Appropriate statistical analysis	ML of available literature, summarizing statistical methods and designating these methods with successful or failed outcomes.
	Appropriate sample size	NLP/ML/DL to predict sample size, estimate patient dropout rates and pre-trial simulation of critical sample size.
	Reducing likelihood of amendments	ML/DL to present designer with pertinent information to consider and potential amendments.
	Inconsistencies in protocol	NLP/ML use in tabular and modular format to check time and events schedule against text, as well as summary of changes for any amendments.
	Optimal arms/ interventions for adaptive therapy	ML/DL to simulate the various scenarios with pertinent parameters. Assist in personalized treatment
	Regulatory/ethics review	ML to ensure regulatory and ethical compliance, especially with adaptive therapy and changes in marketplace
Poor site selection	Effective measurement of trade-offs for each site	ML/DL modeling to assess trade-offs: site history, staff support and experience, investigator enthusiasm, available population pool, patient burden and site-associated cost.

(continued)

Table 19.2 (continued)

Factor	Opportunity	Role for Artificial Intelligence
Poor recruitment	Improved use of funds	Maximize cost effectiveness by targeted communication to meet patient profile, including sentiment analysis.
	Ensuring appropriate eligibility criteria	NLP/ML on prior publications to identify suitable and eligible criteria, and also criteria associated with other trial failures.
	Optimizing cohort composition	ML/DL to search through EMR and omics databases to enrich patient cohort from prognostic and predictive standpoint
	Facilitating locating eligible patients	Database coordination, prompting investigators and patients when appropriate trials are available for specific patients. Assist in trial matching.
	Reducing dropouts	NLP/ML to profile patients based on prior data on who is more likely to complete a trial, reducing dropouts, predicting patient burden
	Informed consent	NLP and ML for sentiment analysis and allow for easily understandable consent forms and information leaflets
	Representative population	ML to ensure adequate diversity including racial, ethnic, sexual and gender biases
Poor patient retention	Transport and time investment	Adaptive patient scheduling. Incorporate patient profiles to tailor site assignment/schedules to patient constraints where possible.
	Financial burden	Systematic review of all patient costs to identify opportunities to minimize impacts.
	Safety	Automated review of contraindicated prior and concomitant medications and procedures.
	Increase likelihood of feeling respected	Sentiment analysis and other AI tools applied to all documents provided to patients. Training of interacting staff for personalized interactions.
	Compliance and adherence	Real-time monitoring to ensure mild toxicities are being addressed and subtle struggles experienced by the patient in keeping up with trial requirements. Tailored messaging and coaching to at-risk participants to increase likelihood of retention.
Poor trial execution	Automating reporting of events	Automated prompting of events for patients and staff, more objective monitoring, decreased patient burden, detects adherence, prompts for required reporting, including protocol deviations and adverse events.
	Preparing data and reporting for write-up	ML for cleaning data for periodic reporting.
	Lack of general awareness	Situation awareness provided to investigator/study coordinator monitoring study progress, patient progress, indicating any adaptive interventions if needed.
Overall	Multiple factor analysis to improve trade-offs based on budget and other constraints	Multicriteria AI-based decision-making to quantify trade-offs in order to achieve the implementation of a successful trial design.

19.3.3 Challenges for Artificial Intelligence in Clinical Trial Design

Despite the significance of AI/ML to contribute and accelerate clinical research, it has not been widely adopted yet. The limited successes of AI have been attributed to multiple things. An important factor is insufficient time lapse from an earlier era of lack of relevant technologies and AI models that were not able to generalize to more complex and realistic medical data sets [288, 289].

The efficiency of every AI-based study design application is directly dependent on the quality and amount of data, and hence faces the same fundamental challenges of electronic medical record (EMR) data interoperability, data privacy, security and integrity. Data lakes specifically designed for research need to be created that can handle large clinical, pathologic, imaging, omics and therapy data sets. This, in turn, requires money for infrastructure and personnel such as data scientists [290].

Furthermore, the digitalization and accessibility of EMR data is a bug hurdle for AI algorithms. Lack of regulatory frameworks on standardization of data collection causes EMR formats to differ widely, to be incompatible with each other or not digital at all, and to reside in a decentralized ecosystem with multiple sources without established secure data exchange for AI analysis. Similarly, clinical trial matching is based on AI algorithms. Any added future functionality and improved performance predominantly will depend on the quality and amount of data which are accessible for analytical model development and pilot study field validation work. Care must be taken to reduce overfitting of ML models as a result of class-imbalance in the training data.

Some AI processes such as NLP algorithms can be very beneficial for healthcare providers and researchers. However, they still require labor and time intensive manual annotation of data. The text in clinical documents is often free flowing and unstructured, and valuable information might only be implicit, requiring some background knowledge or context to understand. Medical providers might refer to the same thing in different ways. An NLP algorithm can be trained to spot all such synonyms by exposure to sample medical records that have been annotated by researchers. The algorithm can then apply that knowledge to interpret unannotated records.

The majority of currently implemented AI approaches rely upon supervised learning techniques that are in turn dependent on large well annotated datasets [212]. Thus, clinicians and researchers face the daunting task of annotating these large data sets. The process of curation and annotation can be helped through mining of record using natural language processing (NLP), semi- or weak-supervised learning to help reduce the number of cases that must be manually annotated.

In addition to curation of data, there is a need to rigorously curate or quality assurance (QA) of the data analytic pipelines developed to train, test and validate an AI/ML/DL algorithm. Changes in versions of software, hardware platforms and open-source libraries can all affect the reproducibility and predicted output. The lack of a shared framework, standardization and interoperability for evaluating AI tools is an issue. Due to copyright issues, it is often difficult to compare and assess such technologies in a standardized way. Many AI approaches related to trial design

criteria, assessment of trial efficiency, searching through literature and databases, patient matching, and outcome monitoring are not standardized and interoperable and thus, cannot be systematically validated.

With the rise of wearables and Internet of things (IoT), medical-grade devices with advanced analytics capabilities for continuous real-time monitoring of patients, disease progression and cognitive sensing, will continue to grow [282]. There is valid emphasis on much needed standardization and interoperability protocols, while at the same time, promoting sensitive data exchange with appropriate integrity and security. Blockchain technology could potentially be an answer for secure, immutable data transfers. Furthermore, a collaboration of regulatory, academic, medical, and industrial institutions, have started to produce standardization frameworks and best practice recommendations for incorporating wearable technology into clinical trial design.

During these early days of implementing AI testing and adoption, the explainability of AI models or how the classifier reached its conclusion is of utmost importance, particularly, if the conclusion is discrepant with the physician's own opinion. Unfortunately, many DL algorithms suffer from the black box stigma [291]. Several approaches ranging from proxy models, attention maps, disentangled representation or learning with known operators have been surfacing to address the issue of uninterpretable AI paradigm [292–296]. There is a need for developing new tools to quantify uncertainties [297] to understand feature attributions [298] or to investigate model's behavior.

In both EMR mining and clinical trial matching, the legal aspects of data privacy and security as well as a sufficient degree of transparency of AI models need to be addressed to ensure that AI-based systems are operable, ethical and have regulatory approvals.

Furthermore, data provenance and governance policies that address concerns of institutional review boards and the broader ethical issues around the use of large patient data sets for research and sharing of large amounts of data with for profit companies also need to be developed. There is a lack of guidance from a legal perspective surrounding physician liability and medical AI use. New guidelines on assigning liability for injuries that may arise from interaction between algorithms and practitioners (including appropriate use of the AI system by the human) is needed [299, 300].

19.3.4 Example Application of Artificial Intelligence in Trial Design (SMART)

The management of a chronic health disorder can be optimized through a sequential, individualized approach whereby treatment is adapted and readapted over time in response to the specific needs and evolving status of the individual [301]. A high level of individual heterogeneity in response to treatment exists in the world of oncology. This necessitates the use of adaptive trials with sequential decision-making to optimize and personalize interventions and outcomes.

In many oncologic scenarios, there is a need for personalization and a quest to discover appropriate intervention. Furthermore, compliance or adherence to treatment can vary. It is not feasible due to multiple factors such as cost, duration, toxicity or lack of efficacy to launch a clinical trial with complete disregard of tumor heterogeneity or lack of compliance. Adaptive interventions provide a way to implement sequential strategies such as continue, step-up, switch, or step-down resulting in personalized sequences of treatment.

The Sequential Multiple Assignment Randomized Trial (SMART)—a multistage randomized trial designs—was devised with the aim to build optimal adaptive interventions by providing individuals with treatments designed to be efficacious and feasible without being onerous. Despite increasing popularity due to their real-world clinical appeal, individualized approach, feasibility and conformity with modern era research aimed at developing high-quality adaptive interventions, SMARTs remain relatively uncommon [302]. Previously, SMART has been used in oncology to develop medication algorithms to treat prostate cancer [303, 304].

Here, we briefly present some highlights from a scenario examined by Almirall et al. regarding the development and application of adaptive interventions—formalize individualized sequences of treatment—for optimizing weight loss among adult individuals who are overweight.

An appropriate adaptive intervention must consider variations in treatment response due to underlying tumor heterogeneity and patient compliance. The approach to craft an adaptive intervention workflow comprises of the following steps:

1. Constructing an individualized plan based on biomarker status, baseline tailoring variables and available treatment options
2. Identify individuals showing early signs of nonresponse
3. Intermediate tailoring variables
4. Adapt subsequent treatment to these nonresponding individuals

Often, a wide variety of critical questions must be answered before developing a high-quality adaptive intervention. Yet, at times, there is often insufficient empirical evidence or theoretical basis to answer these questions with accuracy and reliability. SMART uses experimental design principles to supplement the use of theoretical models and expert clinical consensus to obtain answers to many of the challenging critical questions around building adaptive interventions [305–308].

Typically, SMART trials have main effect aims (main comparison), embedded adaptive intervention aims (interactive effect) and optimization aims (more deeply tailored). Methods from reinforcement learning such as Q-learning have been employed [307, 309]. Adaptive intervention can span across the acute and maintenance phase continuum.

SMART designs differ significantly from standard randomized clinical trials (RCT) in terms of their central aim. Whereas the central aim of a SMART is to construct a high-quality optimal adaptive intervention based on data, the central aim of an RCT is to evaluate the effectiveness of an already-developed intervention

versus a suitable control. The end result of a SMART is a proposal for an optimal adaptive intervention. Following the development of an optimized adaptive intervention using data arising from a SMART, an investigator may choose to evaluate the optimized adaptive intervention versus a suitable control using a subsequent RCT.

AI-based approaches can be employed for the development of SMART design, gathering of data to create an adaptive intervention, match eligible and suitable patient, monitor outcome, improve retention, data analysis and design re-adaptation strategies.

19.4 Discussion and Recommendations

The evolution of clinical trials has resulted in compelling progress in the prevention and treatment of many diseases, including cancer. The focus today also includes patients' symptom management, quality of life, omics and personalized medicine. Advances in medicine, improved surgical techniques, the development of new drugs and devices, the application of statistical techniques to research studies, recognition of the need for regulation, and the development of ethical codes impact the conduction of trials, both in the United States and internationally. Developments have been made to include more of the disparate populations and increase outreach activities within communities. Patients are more educated about their diagnosis and potential clinical trials and frequently seek out participation in research activities.

A randomized controlled trial is the gold standard for establishing standard treatments, by evaluating the safety and efficacy of a therapy through assessment of a pre-established statistical hypothesis. The unacceptably high failure rate of clinical trials necessitates a fundamental transformation of the underlying clinical and innovation model of the research industry to enable a needed paradigm shift to allow for a new feasible trajectory of progress and success. The rising costs of operating clinical trials have limited the feasibility of conducting randomized controlled trials for all important clinical hypotheses. It is crucial to prioritize research questions appropriately and to understand the ideal mix of trials so as to amplify the generation of actionable evidence. Accurately characterizing the state of clinical research and reasons for failure is the first step toward an effective leveraging of limited resources and to alter the current course to ultimately result in more successful trials.

Multiple factors including, trial design, cost of trial, lack of treatment efficacy, safety concerns, ineffective site selection, eligibility criteria, patient recruitment and retention, patient burden, poor outcome monitoring, and inefficient data analysis, can impact the success of a trial Each of the facets of protocol design, execution, and successive trial planning offers opportunities for trading off different concerns, as well as simply making inappropriate judgments leading to poor outcomes. Formulating a list of these factors to consider when designing and executing a clinical trial can provide a foundation for better outcomes. However, not all factors are equally important. For instance, there is a direct trade-off between the speed of enrollment and the cost of executing a trial A well-structured mathematical

framework for trading off degrees of achievement in various parameters can offer a quantitative measure for comparing alternative choices [310].

To stay on par with rapid development of molecular sciences and technology innovations, new experimental trial designs and methods of data analytics have been developed to efficiently evaluate single or multiple hypotheses. In oncology, new clinical trials that use basket, umbrella, platform, or other master protocols are expected to increase due to the focus on omics. Master protocol trials can be used to identify responders to a specific intervention and are anticipated to become one of the standard clinical trial designs to promote personalized treatment. Increased planning efforts and coordination to satisfy the objectives of different stakeholders is required. Improvement in data quality, trial efficiency due to innovative design and required coordination due to changing marketplace innovative design necessitates an AI-based infrastructure. Currently, there is ongoing work in improving the ethical and statistical design of these master protocol trials.

Adaptive trial design methods reflect real oncologic scenarios and have the potential to deliver better treatment to patients. It can efficiently address multiple questions at the same time including early assessment of superiority or futility of a trial/group sequential design, adaptive randomization or dosing, adding or dropping arms, and changing accrual rates/sample size re-estimation. With its adaptable approach, it allows seamless phase I–II and phase II–III trials with a positive impact on both time and resources. The potential benefits of adaptive designs are greatest in complicated settings exemplified by personalized medical research. It employs modeling longitudinal information from individual patients, kept in a central database connected to appropriate software, in order to predict outcomes. Thus, the logistics of adaptive trials are more complicated than the logistics for traditional trials. From the statistical point of view, a Bayesian statistical approach facilitates building complicated but maximally informative clinical trials. Regulatory guidelines can not only prevent possible misuse and/or abuse of adaptive design methods in clinical trials, but also maintain the validity and integrity of the trial.

Merging the concepts of adaptive design with platform protocols has led to adaptive platform trials (APTs)—perpetually testing alternative care strategies for a particular disease and efficiently using information generated during trial conduct to alter subsequent operations in a pre-specified way (Fig. 19.4). This design can play a promising role in patient-centered precision medicine. AI methods will undoubtedly play a role from trial pre-design (assessment of protocol feasibility, analysis of structured and unstructured data from pervious trials and scientific literature, data mining EMRs and publicly available content), iterative trial design (patient cohort optimization, study arms, within-trial adaptations, specification of parameters to be varied in simulation (underlying frequencies of subtypes, event rates and accrual rates; size of absolute and relative treatment effects, sample size, endpoint completion, failure vs. success designation), trial start-up (registration, site selection, systematic flexible protocol, statistical plan including different simulated trajectories, documentation of design process (including software for model, algorithm and simulations), ethical review of the risk–benefit ratio of various designs), trial execution (patient recruitment and retention, assessment of site performance, ongoing

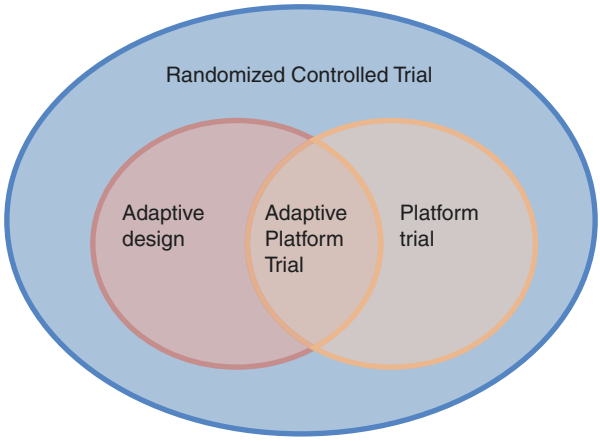


Fig. 19.4 Adaptive Platform trial design

biomarker analysis, secure and interoperable data sharing, real-time monitoring and coaching of patients through smartphone and video tools, predicting compliance issues and dropouts, alerts for missed appointments) and trial reporting (detailed accounting and reporting of patients, periodic analysis of entire APT, data cleaning by ML) [311].

Over the past decade, AI tools have progressed to a level that allows them to be implemented in practical life to complement human decision-making capabilities in the medical field in general and oncology in particular. AI has the potential to transform key steps of clinical trial design from study preparation to execution toward improving trial success rates, thus lowering the associated cost. However, the field of healthcare and clinical research has been relatively slow to adapt to this rapid pace of innovation despite the dire need to implement techniques to augment clinical trial success and the unprecedented explosion of big data. Tight regulations and policies contribute to the rigidity of the system. Innovative changes challenging the established practices need to be implemented cautiously and systematically. Furthermore, the added cost of investing in such technology, at least, in the up-front setting, and in an already underfunded clinical trials area, seems to discourage this trend as well.

There is a need for clinicians, researchers, biomedical engineers, and data scientists to work together and collaborate to assess the current state of AI, evaluate the new technology's added value in the current setting and pool their resources to apply the AI tools to promote maximum achievable benefit leading to the clinical trial's overall efficiency. The AI applications need to be validated in a repeatable and reproducible way. Any ethical, transparency and explainability concerns should be addressed for it to be accepted in real-world applications. The combined efforts of healthcare, AI industry and regulatory bodies toward the ideals of data integrity, provenance, secure data exchange and interoperability are ongoing. To gain and maintain the trust of researchers and clinicians, it is imperative to establish high quality and rigor from the beginning. It is important to note that the measurable impact of

AI tools on the reshaping and efficiency of the clinical trials system is not going to be instantaneous and will not show up in the statistics until after a 5–10-year delay.

Key recommendations to encourage widespread adoption of AI processes in clinical trials are outlined below:

1. Collect, extract, and organize large sets of global omics data, past clinical studies, journal articles, and real-world data, to advance patient selection and eligibility, trial matching, patient burden and visit management, site performance, compliance, retention statistics, adverse event detection and outcome prediction.
2. Shared protocols and algorithm repository to provide a greater insight and transparency
3. Encourage secure exchange of datasets to allow the data to be collectively used to train and test AI and ML models.
4. Encourage secure exchange of encrypted AI and ML algorithms.

19.5 Conclusions

High failure rates of clinical trials contribute substantially to the inefficiency of the clinical trial process. With AI and ML tools and related personnel increasing at a rapid pace, the trickling of these processes to assist with innovative clinical trial design, big data collation, data analysis, patient monitoring and outcome modeling will unquestionably increase. AI, including imaging informatics, is expected to develop decision-support systems for precision medicine and personalized healthcare. However, one has to be cognizant of the limitations of this disruptive technology as well. AI methods require large amounts of data for their training and validation, which also beg the questions of trust, secure data sharing, and privacy concerns. Focus on interoperability, transparency and secure algorithm exchange will assist in validation and promotion of trust amongst the stakeholders. Aside from media hype, the cautious deployment of the state-of-the-art AI/ML technologies, hold tremendous potential to revolutionize clinical trial designs and deliver intended results (Box 19.1).

Box 19.1 Different Terminology Used in AI

Artificial Intelligence (AI): The ability of a software or machine to perform tasks commonly associated with intelligent beings. The term is usually referred to systems endowed with intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience.

Brain–machine interface (BMI): A direct communication pathway between an enhanced or wired brain and an external device. Also referred to

as a brain–computer interface (BCI), a mind–machine interface (MMI), or a direct neural interface (DNI).

Human–machine interface (HMI): A user interface or dashboard allowing direct communication between a human and a device.

Machine learning (ML): The ability to learn, improve and predict from data using algorithms and mathematical modeling without being explicitly programmed to perform the task.

Deep learning (DL): A type of machine learning that uses layered artificial neural networks to progressively extract higher level features from large amounts of raw data.

Raw Data: Unprocessed data such as EMR.

Classification: Machine learning to separate data into categories such as biomarker positive vs. negative.

Regression: Machine learning to predict a continuous numerical output from data.

Dimensionality Reduction: Process in machine learning to reduce the number of random/complex variables under consideration by obtaining a set of principal/simplified variables.

Clustering: Machine learning to separate a group of abstract data into distinct classes or bins such that the objects in the same class are more similar to each other than those in other classes.

Supervised Learning: Machine learning to learn patterns from labeled training data.

Unsupervised Learning: Machine learning to learn patterns from unlabeled data, allowing the model to discover information on its own.

Association rule mining: Rule-based machine learning algorithms for discovering and extracting interesting relations between uncategorized variables in large databases.

Deep reinforcement learning (DRL): Machine learning that focuses on deep learning (DL) and reinforcement learning (RL) to create efficient algorithms that can take actions in an environment so as to maximize some notion of cumulative reward.

Natural language processing (NLP): A subfield of AI, broadly defined by the automatic manipulation of natural language such as text and speech, by software.

Optical character recognition (OCR): A subfield of AI, defined by pattern recognition aimed at the electronic conversion of images of typed, handwritten, or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo, or from subtitle text superimposed on an image.

References

1. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. JAMA. 2004;291(22):2720–6.
2. Fouad MN, Lee JY, Catalano PJ, Vogt TM, Zafar SY, West DW, Simon C, Klabunde CE, Kahn KL, Weeks JC, Kiefe CI. Enrollment of patients with lung and colorectal cancers onto clinical trials. J Oncol Pract. 2013;9(2):e40–7.
3. American Cancer Society Cancer Action Network. Barriers to patient enrollment in therapeutic clinical trials for caner – a landscape report. 2018. https://www.fightcancer.org/sites/default/files/National%20Documents/Clinical-Trials-Landscape-Report.pdf. Accessed 29 Mar 2020.
4. Unger JM, Cook E, Tai E, Bleyer A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. Am Soc Clin Oncol Educ B. 2016;35:185–98. https://doi.org/10.1200/EDBK_156686.
5. Estimates of Clinical Trial Probabilities of Success (PoS). Project ALPHA. Updated 7 Jan 2020. 2019. https://projectalpha.mit.edu/pos/. Accessed 15 Apr 2020.
6. DeVita VT, Rosenberg SA. Two hundred years of cancer research. N Engl J Med. 2012;366:2207–14. https://doi.org/10.1056/NEJMra1204479.
7. Cheson BD. Cancer clinical trials: clinical trials programs. Semin Oncol Nurs. 1991;7:235–42.
8. National Cancer Institute. NCI Community Cancer Centers program pilot: 2007–2010. 2014. http://ncccp.cancer.gov/Media/FactSheet.htm.
9. Kaitlin KI, DiMasi JA. Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000–2009. Clin Pharmacol Ther. 2011;89:183–8.
10. World Medical Association. Declaration of Helsinki. BMJ. 1996;313:1448–9. https://doi.org/10.1136/bmj.313.7070.1448a.
11. Jenkins J, Hubbard S. History of clinical trials. Semin Oncol Nurs. 1991;7:228–34.
12. National Cancer Institute. Cancer clinical trials: the in-depth pro-gram. NIH Publication No. 05-5051. Bethesda, MD: NCI; 2005.
13. U.S. Department of Health and Human Services. Federal policy for the protection of human subjects. 2014. http://www.hhs.gov/ohrp/humansubjects/commonrule/index.html.
14. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. Am J Public Health. 2014;104:e16–28.
15. McCarthy CR. Historical background of clinical trials involving women and minorities. Acad Med. 1994;69:695–8.
16. Allen M. The dilemma for women of color in clinical trials. J Am Med Wom Assoc. 1994;49:105–9.
17. Merkatz RB, Junod SW. Historical background of changes in FDA policy on the study and evaluation of drugs in women. Acad Med. 1994;69:703–7.
18. EDICT. The EDICT project: policy recommendations to eliminate disparities in clinical trials. Houston, TX: EDICT Project; 2008.
19. Pinn VW. The role of the NIH's Office of Research on Women's Health. Acad Med. 1994;69:698–702.
20. Burns JP. Research in children. Crit Care Med. 2003;31:S131–6.
21. Hirtz DG, Fitzsimmons LG. Regulatory and ethical issues in the conduct of clinical research involving children. Curr Opin Pediatr. 2002;14:669–75.
22. Sparks J. Timeline of laws related to the protection of human subjects. 2002. http://history.nih.gov/about/timelines_laws_human.html.
23. Hutchins LF, Unger JM, Crowley JJ, Coltman CA, Albain KS. Underrepresentation of patients 65 years of age or older in cancer-treatment trials. N Engl J Med. 1999;341:2061–7.
24. Lewis JH, Kilgore ML, Goldman DP, Trimble EL, Kaplan R, Montello MJ, Housman MG, Escarce JJ. Participation of patients 65 years of age or older in cancer clinical trials. J Clin Oncol. 2003;21:1383–9. https://doi.org/10.1200/JCO.2003.08.010.

25. Talarico L, Chen G, Pazdur R. Enrollment of elderly patients in clinical trials for cancer drug registration: a 7-year experience by the U.S. Food and Drug Administration. J Clin Oncol. 2004;22:4626–31. https://doi.org/10.1200/JCO.2004.02.175.
26. Herrera AP, Snipes SA, King DW, Torres-Vigil I, Goldberg DS, Weinberg AD. Disparate inclusion of older adults in clinical trials: priorities and opportunities for policy and practice changes. Am J Public Health. 2010;100(Suppl. 1):S105–12.
27. Perry JR, Laperriere N, O'Callaghan CJ, Brandes AA, Menten J, Phillips C, Fay M, Nishikawa R, Cairncross JG, Roa W, et al. Trial Investigators. Short-course radiation plus temozolomide in elderly patients with glioblastoma. N Engl J Med. 2017;376:1027–37. https://doi.org/10.1056/NEJMoa1611977.
28. Phillips C. Insurance coverage expanding for cancer clinical trials. NCI Cancer Bull. 2010;7(10) http://www.cancer.gov/ncicancerbulletin/051810/page5
29. Repucci N. A step-by-step checklist for conducting a clinical trial Medicare coverage analysis. In: Medical Research Law and Policy report; 2012. p. 1–9. http://www.dentons.com/en/insights/articles/2012/october/4/a-stepbystep-checklist-for-conducting-a-clinical-trial-medicare-coverage-analysis.
30. NCI. Annual report to the nation: cancer death rates continue to decline. Bethesda, MD: NCI; 2020.
31. Pharmaceutical Research and Manufacturers of America. Biopharmaceutical industry-sponsored clinical trials: impact on state economies. 2015. http://phrma-docs.phrma.org/sites/default/files/pdf/biopharmaceutical-industry-sponsored-clinical-trials-impact-on-state-economies.pdf.
32. Krall RL. State of the controlled clinical trial enterprise in the United States. Clin Pharmacol Ther. 2011;89(2):225–8. https://doi.org/10.1038/clpt.2010.292.
33. Hirsch BR, Califf RM, Cheng SK, Tasneem A, Horton J, Chiswell K, Schulman KA, Dilts DM, Abernethy AP. Characteristics of oncology clinical trials: insights from a systematic analysis of ClinicalTrials.gov. JAMA Intern Med. 2013;173(11):972–9. https://doi.org/10.1001/jamainternmed.2013.627.
34. Stead M, Cameron D, Lester N, et al. National Cancer Research Networks across the UK. Strengthening clinical cancer research in the United Kingdom. Br J Cancer. 2011;104(10):1529–153421364584.
35. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: study design: randomised controlled trials. BJOG. 2018;125(13):1716. https://doi.org/10.1111/1471-0528.15199.
36. Djulbegovic B, Kumar A, Glasziou P, et al. Trial unpredictability yields predictable therapy gains. Nature. 2013;500:395–6. https://doi.org/10.1038/500395a.
37. Tarnow-Mordi W, Cruz M, Morris JM, Mol BW. RCT evidence should drive clinical practice: a day without randomisation is a day without progress. BJOG. 2017;124:613. https://doi.org/10.1111/1471-0528.14468.
38. Prior M, Hibberd R, Asemota N, Thornton JG. Inadvertent P-hacking among trials and systematic reviews of the effect of progestogens in pregnancy? A systematic review and meta-analysis. BJOG. 2017;124:1008–15.
39. Bonnie S, Martin R. Understanding controlled trials: why are randomised controlled trials important? BMJ. 1998;316:201.
40. Lane S. The best evidence comes from the right study design, not just randomised trials. BJOG. 2018;125:1504. https://doi.org/10.1111/1471-0528.15197.
41. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: intention-to-treat versus per-protocol analysis. Perspect Clin Res. 2016;7(3):144–6. https://doi.org/10.4103/2229-3485.184823.
42. Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med. 2010;8:18.
43. Sanson-Fisher RW, Bonevski B, Green LW, D'Este C. Limitations of the randomized controlled trial in evaluating population-based health interventions. Am J Prev Med. 2007;33(2):155–61. https://doi.org/10.1016/j.amepre.2007.04.007.

44. Gasparyan AY, Ayvazyan L, Akazhanov NA, et al. Conflicts of interest in biomedical publications: considerations for authors, peer reviewers, and editors. Croat Med J. 2013;54:600–8.
45. Spieth PM, Kubasch AS, Penzlin AI, Illigens BM, Barlinn K, Siepmann T. Randomized controlled trials - a matter of design. Neuropsychiatr Dis Treat. 2016;12:1341–9. https://doi.org/10.2147/NDT.S101938.
46. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun. 2018;11:156–64. https://doi.org/10.1016/j.conctc.2018.08.001.
47. Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of investigational drugs in late-stage clinical development and publication of trial results. JAMA Intern Med. 2016;176:1826–33.
48. Crowther M. Phase 4 research: what happens when the rubber meets the road? ASH Educ B. 2013;2013:15–1.
49. Shanley A. Preventing phase III failures. Pharm Technol. 2016;2016:24–7.
50. Henon C, Lissa D, Paoletti X, Thibault C, Le Tourneau C, Lanoy E, Hollebecque A, Massard C, Sorea J-C, Postel-Vinay S. Patient-reported tolerability of adverse events in phase 1 trials. ESMO Open. 2017;2:e000148.
51. Institute of Medicine. A national cancer clinical trials system for the 21st century. Washington, DC: National Academies Press; 2010. https://doi.org/10.17226/12879.
52. Getz KA. Characterizing the real cost of site regulatory compliance. Appl Clin Trials. 2015;
53. Lievre M, Menard J, Bruckert E, Cogneau J, Delahaye F, Giral P, Leitersdorf E, Luc G, Masana L, Moulin P, Passa P, Pouchain D, Siest G. Premature discontinuation of clinical trials for reasons not related to efficacy, safety, or feasibility. BMJ. 2001;322:603–6.
54. Williams RJ, Tse T, DePiazza K, Zarin D. Terminated trials in the ClinicalTrials.gov results database: evaluation of availability of primary outcome data and reasons for termination. PLoS One. 2015;10:e0127242.
55. Heneghan C, Goldacre B, Mahtani KR. Why clinical trial outcomes fail to translate into benefits for patients. Trials. 2017;18:122.
56. Verster J, van de Loo AJ, Roehrs T, Roth T. Are clinical trial participants representative for patients with insomnia? Sleep. 2017;40:A148.
57. Schmidt AF, Groenwold RHH, van Delden JJM, van der Does Y, Klungel OH, Roes KCB, Hoes AW, van der Graaf R. Justification of exclusion criteria was underreported in a review of cardiovascular trials. J Clin Epidemiol. 2014;67:635–44.
58. Babbs CF. Choosing inclusion criteria that minimize the time and cost of clinical trials. World J Methodol. 2014;4:109–22.
59. Getz KA, Zuckerman R, Cropp AB, Hindle AL, Krauss R, Kaitlin KI. Measuring the incidence, causes, and repercussions of protocol amendments. Drug Inf J. 2011;45:265–75.
60. Lösch C, Neuhäuser M. The statistical analysis of a clinical trial when a protocol amendment changed the inclusion criteria. BMC Med Res Methodol. 2008;8:16.
61. Daugherty C, Ratain MJ, Grochowski E, Stocking C, Kodish E, Mick R, Siegler M. Perceptions of cancer patients and their physicians involved in phase 1 trials. J Clin Oncol. 1995;13:1062–72.
62. Godskesen T, Hansson MG, Nygren P, Nordin K, Kihlbom U. Hope for a cure and altruism are the main motives behind participation in phase 3 clinical trials. Eur J Cancer Care. 2015;24:133.
63. Moorcraft SY, Marriott C, Peckitt C, Cunningham D, Chau I, Starling N, Watkins D, Rao S. Patients' willingness to participate in clinical trials and views on aspects of cancer research: results of a prospective patient survey. Trials. 2016;17:17.
64. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, Elbourne DR, Francis D, Garcia J, Roberts I, Snowdon C. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials. 2006;7:9.
65. Prescott RJ, Counsell CE, Gillespie WJ, Grant AM, Russell IT, Kiauka S, Colthart IR, Ross S, Shepherd SM, Russell D. Factors that limit the quality, number and progress of randomised controlled trials. Health Technol Assess. 1999;3:1143.

66. Bower P, Wallace P, Ward E, Graffy J, Miller J, Delany B, Kinmonth AL. Improving recruitment to health research in primary care. Fam Pract. 2009;26:391–7.
67. Cheng S, Dietrich M, Finnigan S, et al. A sense of urgency: evaluating the link between clinical trial development time and the accrual performance of CTEP-sponsored studies. J Clin Oncol. 2009;27(18S):CRA6509. https://doi.org/10.1200/jco.2009.27.18s.cra6509.
68. Korn EL, Freidlin B, Mooney M, Abrams JS. Accrual experience of National Cancer Institute Cooperative Group phase III trials activated from 2000 to 2007. J Clin Oncol. 2010;28(35):5197–201. https://doi.org/10.1200/JCO.2010.31.5382.
69. Bennette CS, Ramsey SD, McDermott CL, Carlson JJ, Basu A, Veenstra DL. Predicting low accrual in the National Cancer Institute's Cooperative Group clinical trials. J Natl Cancer Inst. 2016;108(2):djv324. https://doi.org/10.1093/jnci/djv324.
70. Stensland KD, McBride RB, Latif A, Wisnivesky J, Hendricks R, Roper N, Boffetta P, Hall SJ, Oh WK, Galsky MD. Adult cancer clinical trials that fail to complete: an epidemic? J Natl Cancer Inst. 2014;106:dju299. https://doi.org/10.1093/jnci/dju229.
71. Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, Grant A. Recruitment to randomised trials: strategies for trial enrollment and participation study: the STEPS study. Health Technol Assess. 2007;11:105. iii–ix.
72. Dickson SA, Logan J, Hagen S, Stark D, Glazener C, McDonald AM, McPherson G. Reflecting on the methodological challenges of recruiting to a United Kingdom-wide, multi-centre, randomised controlled trial in gynaecology outpatient settings. Trials. 2013;14:389.
73. Slomka J, McCurdy S, Ratliff E, Timpson P, Williams M. Perceptions of financial payment for research participation among African-American drug users in HIV studies. J Gen Intern Med. 2007;10:1403–9.
74. Bryant J, Powell J. Payment to healthcare professionals for patient recruitment to trials: a systematic review. BMJ. 2005;331:1377.
75. De Wit NJ, Quartero AO, Zuithoff AP, Numans ME. Participation and successful patient recruitment in primary care. J Fam Pract. 2001;50:97681.
76. Pearl A, Wright S, Gamble G, Doughty R, Sharpe N. Randomised trials in general practice: a New Zealand experience in recruitment. N Z Med J. 2003;116:6817.
77. Edwards PJ, Roberts I, Clarke MJ, Diguiseppi C, Wentz R, Kwan I, Cooper R, Felix LM, Pratap S. Methods to increase response to postal and electronic questionnaires. Cochrane Database Syst Rev. 2009;(3):M000008.
78. Cryder C, London AJ, Volpp K, Lowenstein G. Informative inducement: study payment as a signal of risk. Soc Sci Med. 2010;70:455–64.
79. Chin Feman SP, Nguyen LT, Quilty MT, Kerr CE, Nam BH, Conboy LA, Singer JP, Park M, Lembo A, Kaptchuk TJ, Davis RB. Effectiveness of recruitment in clinical trials: an analysis of methods used in a trial for irritable bowel syndrome patients. Contemp Clin Trials. 2008;29:241–51.
80. Okuyemi KS, Cox LS, Nollen NL, Snow TM, Kaur H, Choi W, Nazir N, Mayo MS, Ahluwalia JS. Baseline characteristics and recruitment strategies in a randomized clinical trial of African-American light smokers. Am J Health Promot. 2007;21:183–91.
81. Comis RL, Miller JD, Colaizzi DD, Kimmel LG. Physician-related factors involved in patient decisions to enroll onto cancer clinical trials. J Oncol Pract. 2009;5(2):50–6. https://doi.org/10.1200/JOP.0922001.
82. Memorial Sloan Kettering Cancer Center. National clinical trials survey findings overview. New York, NY: Memorial Sloan Kettering Cancer Center; 2016.
83. Institute of Medicine. Public engagement and clinical trials: new models and disruptive technologies. Workshop summary. Washington, DC: The National Academies Press; 2012.
84. Mannel RS, Walker JL, Gould N, et al. Impact of individual physicians on enrollment of patients into clinical trials. Am J Clin Oncol. 2003;26(2):171–3. https://doi.org/10.1097/01.COC.0000017798.43288.7C.
85. Ulrich CM, James JL, Walker EM, et al. RTOG physician and research associate attitudes, beliefs and practices regarding clinical trials: implications for improving patient recruitment. Contemp Clin Trials. 2010;31(3):221–8. https://doi.org/10.1016/j.cct.2010.03.002.

86. Parreco LK, DeJoice RW, Massett HA, Padberg RM, Thakkar SS. Power of an effective clinical conversation: improving accrual onto clinical trials. J Oncol Pract. 2012;8(5):282–6. https://doi.org/10.1200/JOP.2011.000478.
87. Porter M, Ramaswamy B, Beisler K, et al. A comprehensive program for the enhancement of accrual to clinical trials. Ann Surg Oncol. 2016;23(7):2146–52. https://doi.org/10.1245/s10434-016-5091-9.
88. Copur MS, Ramaekers R, Gönen M, et al. Impact of the National Cancer Institute Community Cancer Centers program on clinical trial and related activities at a community cancer center in rural Nebraska. J Oncol Pract. 2016;12(1):67–8. https://doi.org/10.1200/JOP.2015.005736. e44–51.
89. Saphner T, Thompson MA, Planton S, et al. Insights from building a new National Cancer Institute Community Oncology research program site. WMJ. 2016;115(4):191–5. http://www.ncbi.nlm.nih.gov/pubmed/29099156. Accessed 19 Feb 2018.
90. Meropol NJ, Buzaglo JS, Millard J, et al. Barriers to clinical trial participation as perceived by oncologists and patients. J Natl Compr Cancer Netw. 2007;5(8):655–64. http://www.ncbi.nlm.nih.gov/pubmed/17927923. Accessed 4 Oct 2017.
91. Holcombe RF, Hollinger KJ. Mission-focused, productivity-based model for sustainable support of academic hematology/oncology faculty and divisions. J Oncol Pract. 2010;6(2):74–9. https://doi.org/10.1200/JOP.091075.
92. Thoma A, Farrokhyar F, McKnight L, Bhandari M. How to optimize patient recruitment. Can J Surg. 2010;53:205–10.
93. Raikar S. The impact of study coordinators effectiveness on trial site efficiency: evidence from a pilot study. 2016
94. Speicher LA, Fromell G, Avery S, et al. The critical need for academic health centers to assess the training, support, and career development requirements of clinical research coordinators: recommendations from the clinical and translational science award research coordinator taskforce*. Clin Transl Sci. 2012;5(6):470–5. https://doi.org/10.1111/j.1752-8062.2012.00423.x.
95. Donovan JL, Peters TJ, Noble S, Powell P, Gillatt D, Oliver SE, Lane JA, Neal DE, Hamdy FC. Who can best recruit to randomized trials? J Clin Epidemiol. 2003;56:605–9.
96. Fletcher B, Georghe A, Moore D, Wilson S, Damery S. Improving the recruitment activity of clinicians in randomised controlled trials: a systematic review. BMJ Open. 2012;2:e000496.
97. Jones RH, White H, Velazquez EJ, Shaw LK, Pietrobon R, Panza JA, Bonow RO, Spoko G, O'Connor CM, Rouleau J-L. STICH (surgical treatment for ischemic heart failure) trial enrollment. J Am Coll Cardiol. 2010;56(6):490–8.
98. Lincoff AM, Tardif JC, Neal B, Nicholls SJ, Ryden L, Schwartz GG, Malmberg K, Buse JB, Henry RR, Wedel H, Wichert A, Cannata R, Grobbee DE. Evaluation of the dual peroxisome proliferator-activated receptor α/γ agonist aleglitazar to reduce cardiovascular events in patients with acute coronary syndrome and type 2 diabetes mellitus: rationale and design of the AleCardio trial. Am Heart J. 2013;166:429–434.e1.
99. Schroen AT, Petroni GR, Gray HWR, Cronin W, Sargent DJ, Benedetti J, Wickerham DL, Djubegovic B, Slingluff CL. Preliminary evaluation of factors associated with premature trial closure and feasibility of accrual benchmarks in phase III oncology trials. Clin Trials. 2010;7:312–21.
100. Levett KM, Roberts CL, Simpson JM, Morris JM. Site-specific predictors of successful recruitment to a perinatal clinical trial. Clin Trials. 2014;11:584–9.
101. Van den Bor RM, Grobbee DE, Oosterman BJ, Vaessen PWJ, Roes KCB. Predicting enrollment performance of investigational centers in phase III multi-center clinical trials. Contemp Clin Trials Commun. 2017;7:208–16.
102. Yen W. How long and how far do adults travel for primary care? Washington State Health Services research project. Research brief no. 70. Washington, DC: Washington State Health Services; 2013.
103. Unger JM, Hershman DL, Albain KS, et al. Patient income level and cancer clinical trial participation. J Clin Oncol. 2013;31(5):536–42. https://doi.org/10.1200/JCO.2012.45.4553.

104. Zaleta AK, Miller MF, Johnson J, McManus S, Buzaglo JS. Perceptions of cancer clinical trials among racial and ethnic minority cancer survivors. In: American Psychological Association Annual Convention. Washington, DC: APA; 2017.
105. Javid SH, Unger JM, Gralow JR, et al. A prospective analysis of the influence of older age on physician and patient decision-making when considering enrollment in breast cancer clinical trials (SWOG S0316). Oncologist. 2012;17(9):1180–90. https://doi.org/10.1634/theoncologist.2011-0220.
106. Hughes J, Greville-Harris M, Graham CA, Lewith G, White P, Bishop FL. What trial participants need to be told about placebo effects to give informed consent: a survey to establish existing knowledge among patients with back pain. J Med Ethics. 2017;43:867–70.
107. Chang B-H, Hendricks AM, Slawsky MT, Locastro JS. Patient recruitment to a randomized clinical behavioral therapy for chronic heart failure. BMC Med Res Methodol. 2004;4:8.
108. Nipp RD, Powell E, Chabner B, Moy B. Recognizing the financial burden of cancer patients in clinical trials. Oncology. 2015;20:572–5.
109. Majhail NS, Rizzo JD, Hahn T, Lee SJ, McCarthy PL, Ammi M, Denzen E, Drexler R, Flesch S, James H, Omondi N, Pedersen TL, Murphy E, Pederson K. Pilot study of patient and caregiver out-of-pocket costs of allogenic hematopoietic cell transplantation. Bone Marrow Transplant. 2013;28:865–71.
110. McNeely EA, Clements SD. Recruitment and retention of the older adult into research studies. J Neurosurg Nurs. 1994;26:57–61.
111. Ulrich CM, Knafl KA, Ratcliff SJ, Richmond TS, Grady C, Miller-Davis C, Wallen GR. Developing a model of the benefits and burdens of research participation in cancer clinical trials. AJOB Prim Res. 2012;3:10–23.
112. Stump TK, Eghan N, Engleston BL, Hamilton O, Pirollo M, Schwartz JS, Armstrong K, Beck JR, Meropol NJ, Wong Y-N. Cost concerns of patients with cancer. J Oncol Pract. 2013;9:251–7.
113. Baquet CR, Elison GL, Mishra SI. Analysis of Maryland cancer patient participation in national cancer institute-supported cancer treatment clinical trials. J Clin Oncol. 2008;26:3380–6.
114. Sateren WB, Trimble EL, Abrams J, Brawley O, Breen N, Ford L, McCade M, Kaplan R, Smith M, Ungerleider R, Christian MC. How sociodemographics, presence of oncology specialists, and hospital cancer programs affect accrual to cancer treatment trials. J Clin Oncol. 2002;20:2109–17.
115. Townsley CA, Selby R, Siu LL. Systematic review of barriers to the recruitment of older patients with cancer onto clinical trials. J Clin Oncol. 2005;33:3112–24.
116. Zafar SY, Peppercorn JM, Schrag D, Taylor DH, Goetzinger AM, Zhong X, Abernethy AP. The financial toxicity of cancer treatment. A pilot study assessing out-of-pocket expenses and the insured cancer patient's experience. Oncology. 2013;18:381–90.
117. Bernard DS, Farr SL, Fang Z. National estimates of out-of-pocket health care expenditure burdens among nonelderly adults with cancer: 2001 to 2008. J Clin Oncol. 2011;29:2821–6.
118. Stiles C, Johnson L, Whyte D, Nergaard TH, Gardner J, Wu J. Does increased patient awareness improve accrual into cancer-related clinical trials? Cancer Nurs. 2011;34(5):E13–9. https://doi.org/10.1097/NCC.0b013e31820254db.
119. Ward PR, Rokkas P, Cenko C, Pulvirenti M, Dean N, Carney AS, Meyer S. "Waiting for" and "waiting in" public and private hospitals: a qualitative study of patient trust in South Australia. BMC Health Serv Res. 2017;17:333.
120. Dansky KH, Miles J. Patient satisfaction with ambulatory healthcare services: waiting time and filling time. Hosp Health Serv Adm. 1997;42:165–77.
121. Oermann MH. Effects of educational intervention in waiting room on patient satisfaction. J Ambul Care Manag. 2003;26:150–8.
122. Larson G. Vitals. 9th annual vitals wait time report. 2018.
123. Lopienski K. Retention in clinical trials -- keeping patients on protocols. 2015. https://forteresearch.com/news/infographic/infographic-retention-in-clinical-trials-keeping-patients-on-protocols/.

124. Schumacher A, Sikov WM, Quesenberry MI, et al. Informed consent in oncology clinical trials: a Brown University Oncology Research Group prospective cross-sectional pilot study. PLoS One. 2017;12(2):1–14. https://doi.org/10.1371/journal.pone.0172957.
125. Paasche-Orlow MK, Taylor HA, Brancati FL. Readability standards for informed-consent forms compared with actual readability. NEJM. 2003;348:721–6.
126. Beardsley E, Jefford M, Mileshkin L. Longer consent forms for clinical trials compromise patient understanding: so why are they lengthening? J Clin Oncol. 2007;25(9):2005–6. https://doi.org/10.1200/JCO.2006.10.3341.
127. Hadden K, Holland J, James L. Understanding the subject – plain language IRB informed consents for research. In: PRIM&R Adv Ethical Res Conf (Poster) San Antonio, Abstract 26; 2017.
128. Bostock S, Steptoe A. Association between low function health literacy and mortality in older adults: longitudinal cohort study. BMJ. 2012;344:e1602.
129. Krieger JL, Neil JM, Strekalova YA, Sarge MA. Linguistic strategies for improving informed consent in clinical trials among low health literacy patient. J Natl Cancer Inst. 2017;109:djw233.
130. Institute of Medicine. Informed consent and health literacy (workshop summary). Washington, DC: The National Academies Press; 2015. https://doi.org/10.17226/19019.
131. Hoffner B, Bauer-Wu S, Hitchcock-Bryan S, Powell M, Wolanski A, Joffe S. Entering a clinical trial: is it right for you? Cancer. 2012;118(7):1877–83. https://doi.org/10.1002/cncr.26438.
132. Sood A, Prasad K, Chhatwani L, Shinozaki E, Cha SS, Loehrer LL, Wahner-Roedler DL. Patients' attitudes and preferences about participation and recruitment strategies in clinical trials. Mayo Clin Proc. 2009;84:243–7.
133. Cartmell KB, Bonilha HS, Matson T, Bryant DC, Zapka J, Bentz TA, Ford ME, Hughest-Halberg C, Simpson KN, Alberg AJ. Patient participation in cancer clinical trials: a pilot test of lay navigation. Contemp Clin Trials Commun. 2016;3:86–93.
134. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. J Am Med Assoc. 2002;288:358–62.
135. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. Clin Trials. 2015;12:77–83.
136. Robert C. Pembrolizumab versus ipilimumab in advanced melanoma. N Engl J Med. 2015;372:2521–32.
137. Zhou C. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicenter, open-label, randomized, phase 3 study. Lancet Oncol. 2011;12:735–42.
138. Shaw AT. Ceritinib in ALK-rearranged non-small-cell lung cancer. N Engl J Med. 2014;370:1189–97.
139. Cobo M, Isla D, Massuti B, et al. Customizing cisplatin based on quantitative excision repair cross-complementing 1 mRNA expression: a phase III trial in non-small-cell lung cancer. J Clin Oncol. 2007;25(19):2747–54.
140. Slamon DJ. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344:783–92.
141. Perez EA. Four-year follow-up of trastuzumab plus adjuvant chemotherapy for operable human epidermal growth factor receptor 2-positive breast cancer: joint analysis of data from NCCTG N9831 and NSABP B-31. J Clin Oncol. 2011;25:3366–73.
142. Gilliland DG, Griffin JD. The roles of FLT3 in hematopoiesis and leukemia. Blood. 2002;100(5):1532–42.
143. Kim ES. Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomized phase III trial. Lancet. 2008;372:1809–18.
144. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: design issues. J Natl Cancer Inst. 2010a;102:152–60.

145. Freidlin B, Korn EL, Gray R. A general inefficacy interim monitoring rule for randomized clinical trials. Clin Trials. 2010b;7:197–208.
146. Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. Nat Rev Clin Oncol. 2014;11:81–90.
147. Idikio HA. Human cancer classification: a systems biology-based model integrating morphology, Cancer stem cells, proteomics, and genomics. J Cancer. 2011;2:107–15. https://doi.org/10.7150/jca.2.107.
148. Redman MW, Allegra CJ. The master protocol concept. Semin Oncol. 2015;42:723–30.
149. Berry DA. The Brave New World of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research. Mol Oncol. 2015;9:951–9.
150. Malik SM. Consensus report of a joint NCI thoracic malignancy steering committee: FDA workshop on strategies for integrating biomarkers into clinical development of new therapies for lung cancer leading to the inception of 'master protocols' in lung cancer. J Thorac Oncol. 2014;9:1443–8.
151. Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. N Engl J Med. 2017;377:62–70.
152. Renfro LA, Mandrekar SJ. Definitions and statistical properties of master protocols for personalized medicine in oncology. J Biopharm Stat. 2018;28(2):217–28. https://doi.org/10.1080/10543406.2017.1372778.
153. Redig AJ, Jänne PA. Basket trials and the evolution of clinical trial design in an era of genomic medicine. J Clin Oncol. 2015;33:975–7.
154. Saville BR, Berry SM. Efficiencies of platform clinical trials: a vision of the future. Clin Trials. 2016;13:358–66.
155. Hyman DM, et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. N Engl J Med. 2015;373:726–36.
156. Hirakawa A, Asano J, Sato H, Teramukai S. Master protocol trials in oncology: review and new trial designs. Contemp Clin Trials Commun. 2018;12:1–8. https://doi.org/10.1016/j.conctc.2018.08.009.
157. Catennaci DVT. Next generation clinical trials: novel strategies to address the challenge of tumor molecular heterogeneity. Mol Oncol. 2015;9:967–96.
158. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013;501(7467):355–64. https://doi.org/10.1038/nature12627.
159. Weinstein IB. Cancer. Addiction to oncogenes: the achilles heal of cancer. Science. 2002;297:63–4.
160. Weinstein IB, Joe A. Oncogene addiction. Cancer Res. 2008;68:3077–80.
161. Vogelstein B. Cancer genome landscapes. Science. 2013;339:1546–58.
162. Galbraith S. The changing world of oncology drug development—a global pharmaceutical company's perspective. Chin Clin Oncol. 2014;3:2.
163. Menis J, Hasan B, Besse B. New clinical research strategies in thoracic oncology: clinical trial design, adaptive, basket and umbrella trials, new endpoints and new evaluations of response. Eur Respir Rev. 2014;23:367–78.
164. Cunanan KM. Basket trials in oncology: a trade-off between complexity and efficiency. J Clin Oncol. 2017;35:271–3.
165. Strzebonska K, Waligora M. Umbrella and basket trials in oncology: ethical challenges. BMC Med Ethics. 2019;20(1):58. https://doi.org/10.1186/s12910-019-0395-5.
166. West HJ. Novel precision medicine trial designs. JAMA Oncol. 2017;3(3):423. https://doi.org/10.1001/jamaoncol.2016.5299.
167. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, et al. International classification of diseases for oncology (ICD-O). 3rd ed. Geneva: WHO; 2013. 1st Rev.
168. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 Cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44. https://doi.org/10.1016/j.cell.2014.06.049.

169. National Cancer Institute. NCI-MATCH trial (molecular analysis for therapy choice). Bethesda, MD: NCI; 2015. https://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/nci-match.
170. Schwaederle M, et al. Association of biomarker-based treatment strategies with response rates and progression-free survival in refractory malignant neoplasms: a meta-analysis. JAMA Oncol. 2016;2:1452–9.
171. Janku F, Berry DA, Gong J, Parsons HA, Stewart DJ, Kurzrock R. Outcomes of phase II clinical trials with single-agent therapies in advanced/metastatic non-small cell lung cancer published between 2000 and 2009. Clin Cancer Res. 2012;18:6356–63.
172. Le Tourneau C, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol. 2015;16:1324–34.
173. Hyman DM. Precision medicine at Memorial Sloan Kettering Cancer Center: clinical next-generation sequencing enabling next-generation targeted therapy trials. Drug Discov Today. 2015;20:1422–8.
174. Andre F, Mardis E, Salm M, Soria JC, Siu LL, Swanton C. Prioritizing targets for precision cancer medicine. Ann Oncol. 2014;25:2295–303.
175. Carr TH, et al. Defining actionable mutations for oncology therapeutic development. Nat Rev Cancer. 2016;16:319–29.
176. Beckman RA, Antonijevic Z, Kalamegham R, Chen C. Adaptive design for a confirmatory basket trial in multiple tumor types based on a putative predictive biomarker. Clin Pharmacol Ther. 2016;100:617–25.
177. Steuer CE, Papadimitrakopoulou V, Herbst RS, Redman MW, Hirsch FR, Mack PC, Ramalingam SS, Gandara DR. Clin Pharmacol Ther. 2015;97(5):488–91.
178. Hobbs BP, Chen N, Lee JJ. Controlled multi-arm platform design using predictive probability. Stat Methods Med Res. 2016;27:65–78.
179. Kim ES, et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov. 2011;1:44–53.
180. Liu S, Lee JJ. An overview of the design and conduct of the BATTLE trials. Chin Clin Oncol. 2015;4:33.
181. Rugo HS, et al. Adaptive randomization of veliparib-carboplatin treatment in breast cancer. N Engl J Med. 2016;375:23–34.
182. Park JW, et al. Adaptive randomization of neratinib in early breast cancer. N Engl J Med. 2016;375:11–22.
183. Takebe N, McShane L, Conley B. Biomarkers: exceptional responders-discovering predictive biomarkers. Nat Rev Clin Oncol. 2015;12:132–4.
184. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA. 2000;283:2701–11. https://doi.org/10.1001/jama.283.20.2701.
185. Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? BMC Med. 2017;15(1):134. https://doi.org/10.1186/s12916-017-0902-9.
186. Haslam A, Hey SP, Gill J, Prasad V. A systematic review of trial-level meta-analyses measuring the strength of association between surrogate end-points and overall survival in oncology. Eur J Cancer. 2019;106:196–211. https://doi.org/10.1016/j.ejca.2018.11.012.
187. Bio/BioMedTracker. Clinical trial success rates study. 2011. http://insidebioia.files.wordpress.com/2011/02/bio-ceo-biomedtracker-bio-study-handout-final-2-15-2011.pdf.
188. Berry DA. Adaptive clinical trials in oncology. Nat Rev Clin Oncol. 2012;9(4):199–207. https://doi.org/10.1038/nrclinonc.2011.165.
189. Papadimitrakopoulou V, Lee JJ, Wistuba II, et al. The BATTLE-2 study: a biomarker-integrated targeted therapy study in previously treated patients with advanced non-small-cell lung cancer. J Clin Oncol. 2016;3430:3638–47.
190. Barker AD, Sigman CC, Kelloff GJ, et al. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin Pharm Ther. 2009;86:97–100.

191. Kaplan R, Maughan T, Crook A, et al. Evaluating many treatments and biomarkers in oncology: a new design. J Clin Oncol. 2013;31:4562–8.
192. Berry DA. Bayesian clinical trials. Nat Rev Drug Discov. 2006;5:27–36.
193. Berry DA. Chapter 35. In: Hong WK, et al., editors. Holland-Frei cancer medicine. 8th ed. Shelton, CT: People's Medical Publishing House; 2010. p. 446–63.
194. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. Clin Trials. 2005;2:295–300.
195. Berry DA. Statistics: a Bayesian perspective. Belmont: Duxbury Press; 1996.
196. Berry SM, Carlin BP, Lee JJ, Muller P. Bayesian adaptive methods for clinical trials. New York, NY: CRC Press; 2010.
197. Ellenberg SS, Eisenberger MA. An efficient design for phase III studies of combination chemotherapies. Cancer Treat Rep. 1985;69:1147–52.
198. Wieand S, Schroeder G, O'Fallon JR. Stopping when the experimental regimen does not appear to help. Stat Med. 1994;13:1453–8.
199. Zhang Q, Freidlin B, Korn EL, Halabi S, Mandrekar S, Dignam J. Comparison of futility monitoring guidelines using completed phase III oncology trials. Clin Trials. 2016;14:48.
200. Muss HB, et al. Adjuvant chemotherapy in older women with early-stage breast cancer. N Engl J Med. 2009;360:2055–65.
201. Freidlin B, Korn EL, George SL. Data monitoring committees and interim monitoring guidelines. Control Clin Trials. 1999;20:395–407.
202. Bretz F, Schmidli H, Konig F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biom J. 2006;48:623–34.
203. US Food and Drug Administration. Draft guidance for industry—adaptive design clinical trials for drugs and biologics. Rockville, MD: U.S. Department of Health and Human Services; 2010.
204. Cuffe RL, Lawrence D, Stone A, Vandemeulebroecke M. When is a seamless study desirable? Case studies from different pharmaceutical sponsors. Pharm Stat. 2014;13:229–37.
205. Gaydos B, et al. Good practices for adaptive clinical trials in pharmaceutical product development. Drug Inf J. 2009;43:539–56.
206. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273–86. https://doi.org/10.1093/biostatistics/kxx069.
207. Haddad TC, et al. J Clin Oncol. 2018;36(Suppl):Abstr. 6550.
208. CLINPAL. Recruitment infographic. Engaging patients in clinical research. Stirling: CLINPAL; 2019. https://www.clinpal.com/blog/recruitment-infographic/. Accessed 5 Mar 2020.
209. Knepper TC, McLeod HL. Nature. 2018;557:157–9.
210. Martin AR, et al. Nat Genet. 2019;51:584–91.
211. Blaschke TF, et al. Adherence to medications: insights arising from studies on the unreliable link between prescribed and actual drug dosing histories. Annu Rev Pharmacol Toxicol. 2012;52:275–301.
212. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol. 2020;93:20190855.
213. Clancey WJ, Shortliffe EH. Readings in medical artificial intelligence: the first decade. Boston, MA: Addison-Wesley Longman; 1984.
214. McCauley N, Ala M. The use of expert systems in the healthcare industry. Inf Manag. 1992;22:227.
215. Niu F, et al. HOGWILD!: a lock-free approach to parallelizing stochastic gradient descent. arXiv. 2011;
216. LeCun Y, et al. Deep learning. Nature. 2015;521:436–44.
217. Wang P, Xiao X, Brown JRG, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng. 2018;2:741.
218. Center for Disease Control. Meaningful use. Atlanta, GA: CDC; 2020. https://www-cdc-gov.proxy.lib.mcw.edu/ehrmeaningfuluse/introduction.html. Accessed 28 Mar 2020.

219. Chang K, Bai HX, Zhou H, et al. Residual convolutional neural network for determination of IDH status in low- and high-grade gliomas from MR imaging. Clin Cancer Res. 2018;24:1073–81.
220. Ribli D, Horváth A, Unger Z, et al. Detecting and classifying lesions in mammograms with deep learning. Sci Rep. 2018;8:4165.
221. Lu Y, Yu Q, Gao Y, et al. Identification of metastatic lymph nodes in MR imaging with faster region-based convolutional neural networks. Cancer Res. 2018;78:5135–43.
222. Kann BH, Aneja S, Loganadane GV, et al. Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks. Sci Rep. 2018;8:14036.
223. Nikolov S, Blackwell S, Mendes R, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv. 2018. https://arxiv.org/abs/1809.04430. Accessed 8 Nov 2018.
224. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24:1559–67.
225. Liao S, Gao Y, Oto A, Shen D. Representation learning: a unified deep learning framework for automatic prostate MR segmentation. Med Image Comput Comput Assist Interv. 2013;16:254–61.
226. Fehr D, Veeraraghavan H, Wibmer A, et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proc Natl Acad Sci U S A. 2015;112:E6265–73.
227. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
228. Lao J, Chen Y, Li ZC, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Sci Rep. 2017;7:10353.
229. Yousefi S, Amrollahi F, Amgad M, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep. 2017;7:11707.
230. Zhen X, Chen J, Zhong Z, et al. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. Phys Med Biol. 2017;62:8246–63.
231. Bibault JE, Giraud P, Durdux C, et al. Deep learning and radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Sci Rep. 2018;8:12611.
232. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
233. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res. 2018;24:1248–59.
234. Sun R, Limkin EJ, Vakalopoulou M, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol. 2018;19:1180–91.
235. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6:26094.
236. Pella A, Cambria R, Riboldi M, et al. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. Med Phys. 2011;38:2859–67.
237. Carrara M, Massari E, Cicchetti A, et al. Development of a ready-to-use graphical tool based on artificial neural network classification: application for the prediction of late fecal incontinence after prostate cancer radiation therapy. Int J Radiat Oncol. 2018;102:1533.
238. Feng Q, Dueva E, Cherkasov A, Ester M. PADME: a deep learning-based framework for drug-target interaction prediction. arXiv. 2018. http://arxiv.org/abs/1807.09741. Accessed 14 Feb 2019.
239. Preuer K, Lewis RPI, Hochreiter S, et al. DeepSynergy: predicting anti-cancer drug synergy with deep learning. Bioinformatics. 2018;34:1538–46.
240. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34:i457–66.

241. Han Y, Kim D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. BMC Bioinformatics. 2017;18:585.
242. Eulenberg P, Köhler N, Blasi T, et al. Reconstructing cell cycle and disease progression using deep learning. Nat Commun. 2017;8:463.
243. Aliper A, Plis S, Artemov A, et al. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. Mol Pharm. 2016;13:2524–30.
244. Menden MP, Iorio F, Garnett M, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS One. 2013;8:e61318.
245. Sennaar K. AI and machine learning for clinical trials: examining 3 current applications. Boston, MA: Emerj - Artificial Intelligence Research and Insight; 2013. https://emerj.com/ai-sector-overviews/ai-machine-learning-clinical-trials-examining-x-current-applications/. Accessed 28 Mar 2020.
246. Somashekhar SP, Sepúlveda MJ, Puglielli S, et al. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol. 2018;29:418–23.
247. Liu C, Liu X, Wu F, et al. Using artificial intelligence (Watson for Oncology) for treatment recommendations amongst Chinese patients with lung cancer: feasibility study. J Med Internet Res. 2018;20:e11087.
248. Mohan S, et al. Deep learning for biomedical information retrieval: learning textual relevance from click logs. In: Proc. BioNLP 2017 Workshop, Association for Computational Linguistics; 2017. p. 222–31.
249. LeCun Y. The power and limits of deep learning: in his IRI Medal address, Yann LeCun maps the development of machine learning techniques and suggests what the future may hold. Res Technol Manag. 2018;61:22–7.
250. Alam H, Hartono R, Kumar A, Rahman F, Tarnikova Y, Wilcox C. Web page summarization for handheld devices: a natural language approach. In: Proc 7th Int Conf Doc Anal Recog. Edinburgh: IEEE Computer Society; 2013. p. 1153–7.
251. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, Crockett SD, Gourevitch RA, Dean KM, Mehrotra A. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. JAMIA. 2017;24:986–91.
252. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42:760–72.
253. Hripcsak G, Austin JHM, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology. 2002;224:157–63.
254. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci. 2019;40(8):577–91.
255. Chen Y, et al. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. Clin Ther. 2016;38:688–701.
256. Fanda JM, et al. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci. 2018;1:53–68.
257. Goudey B, et al. A blood-based signature of cerebrospinal fluid Aβ1–42 status. Sci Rep. 2019;9:4163.
258. Palmqvist S, et al. Accurate risk estimation of β-amyloid positivity to identify prodromal Alzheimer's disease: cross-validation study of practical algorithms. Alzheimers Dement. 2019;15:194–204.
259. Romero K, et al. The future is now: model-based clinical trial design for Alzheimer's disease. Clin Pharmacol Ther. 2015;97:210–4.
260. Sun Z, et al. A data driven method for generating robust symptom onset indicators in disease registry data. AMIA Annu Symp Proc. 2017;2017:1635–44.
261. Che C, et al. An RNN architecture with dynamic temporal matching for personalized predictions of Parkinson's disease. In: Proceedings of the 2017 SIAM International Conference on Data Mining. Philadelphia, PA: SIAM; 2017. p. 198–206.

262. IQVIA. Global oncology trends. Durham, NC: IQVIA; 2019. https://www.iqvia.com/insights/the-iqvia-institute/reports/global-oncology-trends-2019. Accessed 10 Apr 2020.
263. Young T, et al. Recent trends in deep learning based natural language processing. IEEE Comput Intell Mag. 2018;13:55–75.
263. Helgeson J, Rammage M, Urman A, Roebuck MC, Coverdill S, Pomerleau K, Dankwa-Mullan I, Liu L-I, Sweetman RW, Chau Q, Williamson MP, Vinegra M, Haddad TC, and Goetz MP. Clinical performance pilot using cognitive computing for clinical trial matching at Mayo Clinic. J Clin Oncol. 2018;36:15_suppl, e18598-e18598.
264. Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, Sacchi L. Secondary use and analysis of big data collected for patient care. Yearb Med Inform. 2017;26(1):28–37. https://doi.org/10.15265/IY-2017-008.
265. Penberthy L, Brown R, Puma F, Dahman B. Automated matching software for clinical trials eligibility: measuring efficiency and flexibility. Contemp Clin Trials. 2010;31(3):207–17. https://doi.org/10.1016/j.cct.2010.03.005.
266. Embi PJ, Jain A, Clark J, Bizjack S, Hornung R, Harris CM. Effect of a clinical trial alert system on physician participation in trial recruitment. Arch Intern Med. 2005;165(19):2272. https://doi.org/10.1001/archinte.165.19.2272.
267. Davis TC, Holcombe RF, Berkel HJ, Pramanik S, Divers SG. Informed consent for clinical trials: a comparative study of standard versus simplified forms. J Natl Cancer Inst. 1998;90:668–74.
268. Fan W, Gordon MD. The power of social media analysis. Commun ACM. 2014;57:74–81.
269. Neiger BL, Thackeray R, Van Wagenen SA. Use of social media in health promotion. Health Promot Pract. 2012;13:159–64.
270. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. Comput Ling. 2011;37:267–307.
271. Yuan C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc. 2019;26:294–305.
272. Liu C, et al. DQueST: dynamic questionnaire for search of clinical trials. J Am Med Inform Assoc. 2019;26:1333. https://doi.org/10.1093/jamia/ocz121.
273. Labovitz DL, et al. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. Stroke. 2017;48:1416–9.
274. Roy S, et al. Machine learning for seizure type classification: setting the benchmark. arXiv. 2019. https://arxiv.org/abs/1902.01012.
275. Harrer S. Measuring life: sensors and analytics for precision medicine. In: van den Driesche S, editor. Bio-MEMS and medical microdevices II. Bellingham, WA: SPIE; 2015. 51802-1-951802-5.
276. Yetisen AK, et al. Wearables in medicine. Adv Mater. 2018;30:1706910. 1–26.
277. Gulshan V, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
278. Rodriguez-Ruiz A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst. 2019;111:916.
279. Mahapatra D, et al. Deformable medical image registration using generative adversarial networks. Biomed Imaging. 2018:1449–53.
280. Yauney G, Shah P. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. PMLR. 2018;85:161–226.
281. Shah P, et al. Technology-enabled examinations of cardiac rhythm, optic nerve, oral health, tympanic membrane, gait and coordination evaluated jointly with routine health screenings: an observational study at the 2015 Kumbh Mela in India. BMJ Open. 2018;8:e018774.
282. M. Gabrani, et al. When data meets disease head-on: new trends in treating and managing epilepsy. 2018. https://researcher.watson.ibm.com/researcher/files/au1-sharrer/EpilepsyWP_Nov2018.pdf.
283. Berg BP, Denton BT, Erdogan SA, Rohleder T, Huschka T. Optimal booking and scheduling in outpatient procedure centers. Comput Oper Res. 2014;50:24–37.

284. Chien C-F, Tseng F-P, Chen C-H. An evolutionary approach to rehabilitation scheduling. A case study. Euro J Oper Res. 2008;189:1234–53.
285. Keller JM, Liu D, Fogel DB. Fundamentals of computational intelligence. New York, NY: Wiley; 2016.
286. Burke EK, Causmaecker DP, Berghe GV, Van Landeghem H. The state of the art of nurse rostering. J Sched. 2004;7:441–99.
287. FDA. Framework for FDA's real-world evidence program. 2018. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.
288. Marcus G. Deep learning: a critical appraisal. arXiv. 2018. https://ui.adsabs.harvard.edu//#abs/2018arXiv180100631M.
289. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol. 1996;49:1225–31.
290. Bycroft C, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.
291. Watson DS, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019;l:886.
292. Philbrick KA, Yoshida K, Inoue D, Akkus Z, Kline TL, Weston AD, et al. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. Am J Roentgenol. 2018;211:1184–93. https://doi.org/10.2214/AJR.18.20331.
293. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. Radiology. 2019;290:514–22. https://doi.org/10.1148/radiol.2018180887.
294. Luna JM, Gennatas ED, Ungar LH, Eaton E, Diffenderfer ES, Jensen ST, et al. Building more accurate decision trees with the additive tree. Proc Natl Acad Sci U S A. 2019;116:19887–93. https://doi.org/10.1073/pnas.1816748116.
295. Nazmul Haque K, Latif S, Rana R. Disentangled representation learning with information maximizing Autoencoder. arXiv. 2019;
296. Maier AK, Syben C, Stimpel B, Würfl T, Hoffmann M, Schebesch F, et al. Learning with known operators reduces maximum training error bounds. Nat Mach Intell. 2019;1:373–80. https://doi.org/10.1038/s42256-019-0077-5.
297. Eaton-Rosen Z, et al. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. Cham: Springer International Publishing; 2018.
298. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. arXiv. 2017;
299. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA. 2019;322:1765. https://doi.org/10.1001/jama.2019.15064.
300. Formation. E.G.o.L.a.N.T.N.T. Liability for artificial intelligence and other emerging digital technologies; 2019.
301. Lavori PW, Dawson R. Adaptive treatment strategies in chronic disease. Annu Rev Med. 2008;59:443–53.
302. Almirall D, Nahum-Shani I, Sherwood NE, Murphy SA. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. Transl Behav Med. 2014;4(3):260–74. https://doi.org/10.1007/s13142-014-0265-0.
303. Thall P, Logothetis C, Pagliaro L, et al. Adaptive therapy for androgen-independent prostate cancer: a randomized selection trial of four regimens. J Natl Cancer Inst. 2007;99(21):1613–22. https://doi.org/10.1093/jnci/djm189.
304. Almirall D, Lizotte D, Murphy SA. SMART design issues and the consideration of opposing outcomes. J Am Stat Assoc. 2012;107(498):509–12. https://doi.org/10.1080/01621459.2012.665615.
305. Lavori PW, Dawson R. Dynamic treatment regimes: practical design considerations. Clin Trials. 2004;1:9–20. https://doi.org/10.1191/1740774S04cn002oa.
306. Murphy SA. An experimental design for the development of adaptive treatment strategies. Stat Med. 2005;24:1455–81. https://doi.org/10.1002/sim.2022.
307. Nahum-Shani I, Qian M, Almirall D, et al. Q-learning: a data analysis method for constructing adaptive interventions. Psychol Methods. 2013;17(4):478–94.

308. Lei H, Nahum-Shani I, Lynch K, et al. A "SMART" design for building individualized treatment sequences. Annu Rev Clin Psychol. 2012;8:21–48. https://doi.org/10.1146/annurev-clinpsy-032511-143152.
309. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. Stat Med. 2009;28(26):3294–315. https://doi.org/10.1002/sim.3720.
310. Michalewicz Z, Fogel DB. How to solve it: modern heuristics. 2nd ed. Berlin: Springer; 2004. p. 444–9.
311. Adaptive Platform Trials Coalition. Adaptive platform trials: definition, design, conduct and reporting considerations. Nat Rev Drug Discov. 2019;18(10):797–807. https://doi.org/10.1038/s41573-019-0034-3. Published correction appears in Nat Rev Drug Discov. 2019.
312. Thomas DW, Burns J, Audette J, Carrol A, Dow-Hygelund C, Hay M. Clinical development success rates 2006–2015. San Diego, CA: Biomedtracker; 2016.
313. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014;32:40–51.

Index

I. El Naqa, M. J. Murphy (eds.), *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, https://doi.org/10.1007/978-3-030-83047-2