Published in final edited form as:
Oncology. 2020 ; 98(6): 344–362. doi:10.1159/000493575.

Machine Learning and Imaging Informatics in Oncology

Huan-Hsin Tseng[*], **Lise Wei**[†], **Sunan Cui**[‡], **Yi Luo**[§], **Randall K. Ten Haken**[**], **Issam El Naqa**[††]
Department of Radiation Oncology, University of Michigan, Ann Arbor 519W William Street, Ann Arbor, MI 48103, USA

Abstract

In the era of personalized and precision medicine, informatics technologies utilizing machine learning (ML) and quantitative imaging are witnessing a rapidly increasing role in medicine in general and in oncology in particular. This expanding role ranges from computer-aided diagnosis to decision support of treatments with the potential to transform the current landscape of cancer management. In this review, we aim to provide an overview of ML methodologies and imaging informatics techniques and their recent application in modern oncology. We will review example applications of ML in oncology from the literature, identify current challenges, and highlight future potentials.

Keywords

Oncology; machine learning; imaging informatics

1. Introduction

MACHINE learning (ML) is an interdisciplinary field from artificial intelligence that draws upon advances in computer science, neuroscience, psychology and statistics for developing computer algorithms that can learn tasks from data, without being explicitly programmed for this purpose [1]. The application of ML is currently prevalent in a wide range of diverse fields (e.g., banking, sports, politics and advertising) producing reliable guidance to decision making and reducing manual labor [2]. A subarea of ML called deep learning, allowing abstract representation of data via deep neural networks (also known as multi-layer neural networks), has recently shown its potential in mimicking human cognition and challenging human intellectual abilities from video/board games to medicine. Recently, many high-profile companies have implemented machine learning techniques in their practice. For example, the Google Cloud [3] can convert audio to text and translate an arbitrary string into any supported language. Spotify music utilizes convolutional and recurrent neural networks

[*] thuanhsi@med.umich.edu. Phone : (734)936-9219
[†] liswei@umich.edu
[‡] sunan@umich.edu
[§] yiyiluo@med.umich.edu
[**] rth@med.umich.edu
[††] ielnaqa@med.umich.edu. Phone : (734)936-4290

Disclosure Statement
The authors have no conflicts of interest to declare.

for recognizing music genres and making recommendations for its users [4]. Rider sharing Apps like Uber and Lyft can predict rider demands using ML algorithms to minimize the waiting time of their customers [5].

In the field of medicine, information technology ventures are actively developing and seeking applications for their ML tools. For instance, Google DeepMind released mobile applications for diagnosis of eye disease, kidney injury and management of electronic patient records. In the field of oncology, there has been growing interest in applying machine learning for diagnosis, prognosis and treatment queries. For example, the IBM Watson for Oncology (WFO) system has demonstrated its effectiveness in making treatment recommendations for specific cancer patients. In breast cancer, WFO was able to learn an extensive corpus of medical journal, textbook and treatment guidelines information at Memorial Sloan Kettering Cancer Center (MSKCC) by natural language processing (NLP) to identify articles that are well-matched to the characteristics of specific patients. It also incorporated data from over 550 breast cancer cases in MSKCC, including variables like patient characteristics, comorbidities, functional status, tumor characteristics, stage, imaging and other laboratory findings [6]. WFO can further refine its analytical process according to feedback given by experts. The system can finally provide treatment planning recommendations (surgery, chemotherapy, immunotherapy, radiotherapy) and alternative options within each treatment plan (e.g., drugs or doses) to a specific patient. It has been tested at the Manipal Comprehensive Cancer Center, showing a high concordance (93%) with a multidisciplinary tumor board [7].

Algorithms based on deep learning, such as a convolutional neural network (CNN), have been applied in imaging diagnosis of a wide variety of cancers showing high accuracy comparable or superior to human experts. For instance, a CNN was able to distinguish between the most common and deadliest types of skin cancers by learning from a dataset consisting of 12,940 clinical images, outperforming two dermatologists on a subset of the same validation set [8]. In a challenge competition to develop automated solutions for detecting lymph node metastases in breast cancers from pathology images, the top-performing algorithm was also based on a deep learning CNN [9], which achieved a better diagnostic performance than a panel of 11 pathologists under a simulated exercise designed to mimic routine clinical workflow. Besides the high prediction accuracy offered by ML algorithms, they also enjoy high efficiency and can be cost effective. For instance, a well-trained CNN from previous clinical examples, can achieve accurate diagnosis in a fraction of a second at any time offering the possibility of a universal access to diagnostic care anywhere, anytime.

ML can be also an effective tool for molecular targeting by unveiling the complex relationships of underlying genetics and other biological information. A blood test called CancerSEEK [10] was reported to be able to detect eight common cancer types from very early stages of disease and localize the origin of cancer to a small number of anatomic sites by assessing the levels of circulating proteins and mutations in the DNA. This test can be applied in early detection of cancers and it can conceivably reduce deaths.

In the future, cognitive learning systems such as WFO may potentially offer physicians the necessary tools to tailor their treatments to an individual patient based on the synthesized knowledge by ML algorithms from the existing literature and/or interactive learning from clinical oncology experts. Due to the sophisticated technical advances and the tremendous growth of genetic and clinical data, it is becoming harder for busy physicians to stay current about every emerging new finding. On the other hand, ML-based systems can acquire such knowledge from a large volume of unstructured and structured datasets, aggregate and effectively present such synthesized knowledge to practicing physicians as a second opinion to aid and support their decision making and improve cancer patients' management.

In this article, we will review some of the basic concepts and methods commonly applied in ML. Then, we will present several examples of ML and imaging informatics applications in diagnosis, prognosis, and treatment of cancers. Finally, we will discuss current technical and administrative barriers for a more comprehensive and wider incorporation of ML techniques into clinical practice and offer some tentative recommendations to realize the tremendous potentials of ML for oncology and cancer patient's care.

2. What Is Machine Learning?

Machine learning is broadly referred to as computer algorithms that can provide computers with the ability to learn patterns from data or make predictions based on prior examples. Machine learning, a term first coined by Arthur Samuel, is considered as a major branch of Artificial Intelligence (AI) (see Fig. 1), as proposed by John McCarthy [11] and was defined as "involves machines that can perform tasks that are characteristic of human intelligence". Machine learning is generally designed to learn analytical patterns from data and making generalizations (predictions) based on its exposition to previous samples [12]. Thus, it is a field strongly tied to cognitive psychology, neuroscience, computational and statistical principles that also aim at data mining and performance predictions. To explain the concept of machine learning more concretely, the following provides more details.

2.1. Definition

A technical definition of machine learning is, quoted from [13], "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". Tom Mitchell further illustrated this definition by showing an example of playing checkers, where E=playing checkers, P=ability to win, and T=game rules. In another words, ML provide computers with the ability to performs tasks beyond what they were originally programmed for, i.e., they learn how to perform tasks in a more or less similar fashion to an autonomous human operator.

2.2. Categories

There are mainly three categories recognized in machine learning: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning requires a dataset containing input and output labels, which are the desired outputs or outcomes, so that a computer is trained by a labelled-dataset as if it were learning under the supervision of a teacher. Technically, supervised learning aims to find a mathematical function that can map input data pairs into output labels. On the other hand, unsupervised learning can operate on a dataset without given labels. In such a case a computer algorithm is tasked to figure out (intrinsic) structures within the data (e.g., Fig. 3), where these intrinsic structures could mean clusters or support regions of data.

Some typical supervised learning algorithms include (logistic, LASSO, Ridge, etc.) regressions, support vector machines (SVMs), random forests, neural networks (NNs), etc. Examples of unsupervised learning are principal component analysis (PCA) [14], Laplacian eigenmaps [15], t-SNE [16], p-SNE [17], and autoencoders [18], etc. Illustrations of a supervised learning and an unsupervised learning are given in Fig. 2(a) and Fig. 3(a), respectively. A clinical application by Dawson et al. [19] utilized an unsupervised PCA to indicate the presence of linear separability in xerostomia (dry mouth) data of patients at high or low risks post-radiotherapy exposure of the parotid gland, Fig. 3(b). Intuitively, supervised learning usually can perform more effectively in classifying data due to the additional guidance of known answers (labels) provided to it. Thus, in this respect, unsupervised learning is generally considered a harder computational problem, where cognitive learning is assumed to be implicit.

From a probabilistic perspective, supervised learning algorithms can also be categorized into *discriminant* or *generative* models. It is usually assumed that inputs (x) and their labels (y) in supervised classification arise from a joint probability $p(x,y)$. A discriminant classifier model, defined by the posterior probability $p(y \mid x)$, can be used to map inputs (x) to class labels (y) without necessarily knowing the underlying joint probability function. Whereas a generative classifier attempts to learn the architecture of such joint probability $p(x,y)$ first and then make their predictions by using Bayes rule to calculate conditional probabilities $p(y \mid x)$ and choosing the most likely label y [22]. The advantage of the generative approach is that we can use the algorithm to generate new synthetic data similar to the existing ones, while the discriminant algorithm generally offers better performance for classification tasks [16] [23].

If a given dataset contains temporal information, one may utilize dynamic machine learning algorithms that can take time information into account such as a Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) network [24]. In addition, classical Bayesian network techniques are able to perform both *dynamic* and *static* predictions, and recursive Bayesian methods can estimate an unknown probability density function recursively over time based on incoming information. If the variables are linear and normally distributed, the recursive Bayesian method become equivalent to the well-known Kalman filter widely used in control and signal processing applications [25] [26].

As for the third category, reinforcement learning (RL), it is designed to embody a software agent (which may represent a clinician in our case) to take actions when interacting with a given environment (e.g., clinical treatment). Usually there is a definite goal for the agent to reach, via a so-called a *reward function* (e.g., better treatment outcome). Winning a chess/GO game by an agent can be a goal of an RL algorithm in a board game, such as in the example AlphaGo of DeepMind [20]. It is worth noting that RL is a modern extension of classical statistical decision-making schemes known as Markov Decision Processes (MDPs), which originally appeared in the 1950s [21] and are currently empowered by advanced computing technologies.

2.3. Deep Learning

Recently, an interesting and powerful branch in ML called deep learning is demonstrating tremendous success in solving pattern recognition and computer vision problems compared to classical ML techniques. These learning algorithms are generally based on neural network (NN) architectures with more than 2 hidden layers (Fig. 2 (b)), and hence the qualifier "deep". Deep learning has empirically proven to be capable of efficient learning from complex tasks [27], which is essentially due to an inherent characteristic called the universal approximation property [28]. In CNNs, such a property can be interpreted as learning data representation in a hierarchal manner while optimizing prediction for the task at hand without the risk of overfitting [29], and avoiding the feature selection problem inherent in classical ML problems [27]. Some specialized (deep) NNs of a desired purpose are developed subsequently to perform different tasks such as Convolutional Neural Networks (CNNs), Fig. 4, for image recognition/classification and Recurrent Neural Networks (RNNs) for sequential learning such as text captions of images.

CNNs are said to be inspired by the work of Hubel and Wiesel on the animal visual cortex [30]. In practice, a CNN is one particular type of NN (Fig. 4) usually consisting of three parts: (1) a convolutional layer, (2) a pooling (down-sampling) layer and (3) one final fully-connected layer for classification purposes. The distinguished convolutional part is generally the most important piece (hence the name CNN) and it aims at learning and extracting features from the input data (e.g., edges, textures, etc., from images) [29]. The pooling layer serves as a data reduction mechanism and the last fully-connected layer makes final a judgment as to which class the input images may belong to.

A CNN can go deep by consecutively repeating the convolution and the pooling layers. With the ever-increasing computing power to deal with the growing size of data, deep learning can be efficiently trained and applied. As mentioned in [27], deep learning outperforms conventional shallow NNs and other ML algorithms by capturing complex structures in high-dimensional data. Each feature map via a convolutional kernel is considered a representation of higher or more abstract level [31]. Thus, representations of deeply transformed layers emphasize quantities that are crucial for computers to discriminate, such that they can be considered as latent (hidden) features in the data that are automatically learned by a CNN.

Certain CNN architectures have proven to be effective: LeNet is the first application of a CNN to read zip codes and digits by Yann LeCun in the 1990's [32]; AlexNet, the champion of the ImageNet pattern recognition challenge 2012, demonstrated surprising performance compared to other common methods and is considered a watershed moment in modern ML applications [33]. GoogLeNet, the laureate of ILSVRC 2014, introduced an inception module into the architecture to reduce the huge number of parameters by up to 12 times over that of AlexNet and found out an optimal local sparse structure [34]. Also, a VGGNet, introduced in 2014, used a 3 × 3 filters only stacked to increase the learning depth [35].

There are two other variants widely used in the literature, VGG16 and VGG19, of the original VGGNet, where the “16” and “19” refer to the layer depth, respectively. Notably, in general the depth of a layer could improve the classification accuracy, but this is not always the case. The learning accuracy empirically ceases to grow after a certain depth is reached. This is due to the vastly increasing amount of weights when adding more layers. In the case where using weights free of concern is a luxury, alternatives should be considered. Particularly, in the medical field where datasets are typically of small size, U-Net [36] may be a good choice for such scenarios.

The success of CNNs and RNNs owes to their design for taking related information into account, where a CNN considers multi-dimensional neighboring data (e.g., image pixels/ MRI voxels) and an RNN focuses on (1-dimensional) sequential (temporal) relations such as human voice recognition or texts. These features make them naturally suitable for decision support by incorporating both the spatial and the temporal information. Hence, making them strong candidates for aiding treatment planning and dose adaptation. Another possible way to achieve automatic decision support in radiotherapy, for instance, is via reinforcement learning, where desired benefits are maximally pursued by a software agent. It is the same principle that drove AlphaGo into winning the Chinese Go game. Thus, deep learning has the potential to optimize prediction of outcomes and identify optimal strategies for precision treatment in oncology. Later in Sec. III-B, we shall discuss how these technologies are currently applied to help extract useful cues from imaging data, as a valuable resource for precision oncology.

3. Machine Learning for Imaging Informatics in Oncology

Medical imaging has been widely applied clinically over the past decades. Recently, it is showing even more potential and utility due to the vast development of quantitative imaging techniques and recent breakthroughs in the machine learning community [31].

In general, there are two main types of imaging acquisitions: (1) anatomical imaging, including conventional X-ray, ultrasound, computed tomography (CT) and magnetic resonance imaging (MRI), etc.; (2) functional (molecular) imaging, including positron emission tomography (PET), single-photon emission computed tomography (SPECT) and diffusion-weighted MRI, etc. In order to combine the advantage of anatomical resolution and functional information of tissues, multimodality imaging techniques such as SPECT/CT, PET/CT, PET/MRI were also developed. With all these techniques available, images can provide valuable data encoded with patient individual information about tumor, tumor environment and genotype that can be data mined to help with the diagnosis, prognosis and prediction of oncology outcomes [37].

This raises the question: “how can we discover underlying biological relationships in these huge amounts of imaging data”? Beyond the already complex procedures clinicians use to read and interpret medical images for making routine decisions, they are also extracting other information; however, in a relatively qualitative way (for example, the boundary of the tumor, or its heterogeneity, etc.). Fortunately, with the development of advanced pattern recognition techniques and statistical learning tools, digital medical images can now be converted into mineable high-dimensional data via high-throughput extraction of quantitative features. This has shown great potential for precision medicine in oncology. The conversion of medical images into a large number of advanced features and the subsequent analysis that relates these features with biological endpoints and clinical outcomes give rise to the field of *Radiomics*.

3.1. Radiomics

Medical imaging plays an important role in oncological practice from diagnosis, staging of tumors, treatment guidance, evaluation of treatment outcomes, and follow-up of patients. Being a noninvasive method, images are able to provide both spatial and temporal information of the tumor. Extraction of quantitative features from medical images together with subsequently relating these features to biological endpoints and clinical outcomes is referred to as the field of Radiomics. “Radio” comes from radiology, which refers to radiology images, e.g. CT, MRI and PET. “-omics” stands for the technologies that aim at providing collective and quantitative features for the entire system and explore the underlying mechanisms. It is widely used in biology, such as in the study of genes (genomics), proteins (proteomics) and metabolites (metabolomics) [38]. The origin of radiomics is medical image analysis and understanding. The goal of radiomics is to take advantage of the digital data stored in those images to develop diagnostic, predictive, or prognostic radiomic models to help understand the underlying biological/clinical processes, support personalized clinical decisions and optimize individualized treatment planning. The core of radiomics is the extraction of quantitative features, with which we can apply all the advanced machine learning algorithms and build models to bridge between images and biological and clinical endpoints. A central hypothesis of radiomic analysis, is that the imaging features are able to capture distinct phenotypic differences, like genetics and proteomics patterns or other clinical outcomes so that we can infer these endpoints. This hypothesis has been proven recently by many researches. Segal et al. showed that the dynamic imaging traits (features) in CT systematically correlate with the global gene expression programs of primary human cancer [39]. Aerts et al. [40] found that a large number of radiomic features extracted from CT images have prognostic power in independent datasets of lung and head neck cancer patients. It is interesting to note that the study identified a general prognostic phenotype for both lung and head neck cancers. Vallieres et al. [41] extracted features from FDG- PET and CT images and performed risk assessment for locoregional recurrences (LR), distant metastases (DM) and overall survival (OS) in head and neck cancer. These studies assured the potential of radiomic features for analyzing the properties of specific tumors. A central component of all the examples stated above is the ability to obtain or infer the hidden information from the pixels (voxels) in the digital images. Generally, there are two main processes to help relate raw images to the endpoints: (1) feature extraction; (2) classification or regression using the extracted features. There are two general ways to extract useful features: (1a) hand-crafted techniques or directly using existing radiomic signatures; (1b) automatic learning of image representation by deep NNs (e.g. CNNs). For oncology classification problems, methods and tools for feature extraction via both conventional machine learning algorithms (e.g., SVMs, random forests) and newly emerging deep learning algorithms are growing at a rapid pace.

3.2. Hand-crafted feature extraction (conventional radiomics)

The regions of interest (ROIs) in cancer diagnosis/prognosis are generally spatially and temporally heterogeneous. It has been shown that radiomics features for capturing these intra-tumor heterogeneities were powerful in prognostic modeling of aggressive tumors. These changes in 4D (3D space + 1D time) space play an important role in analyzing and monitoring the disease status [40]. Thus, it is natural to divide radiomic features further into two types: spatial (static) and temporal (dynamic) [42]. Static features are based on intensity, shape, size (volume), texture and wavelet, while dynamic features are based on kinetic analysis using time-varying acquisition protocols, such as dynamic PET or MRI. Both of these features offer information on the tumor phenotype and its microenvironment (habitat). Examples of static features are: (a) morphological (shape descriptors), (b) first-order, and (c) second-order (texture) features. Texture features, specifically, can provide statistical interrelationships between voxels and capture special patterns in the ROIs to compensate for the loss of information from the first-order features due to the spatial information associated with the relative positions of the intensity levels of the voxels.

These features can then provide a quantitative representation that to some degree can mimic the features that clinicians may pay attention to, while also offer the potential of obtaining more information invisible to the human eye. After producing isotropic interpolated voxel sizes and discretize grey level images, features can be calculated from varying common texture matrices such as: the grey-level co-occurrence matrix (GLCM) [43], grey-level run-length matrix (GLRLM) [44], grey-level size zone matrix (GLSZM) [45], grey-level distance zone matrix (GLDZM) [46], neighborhood grey tone difference matrix (NGTDM) [47], and the neighboring grey level dependence matrix (NGLDM) [48]. These gray level matrices provide statistical methods to capture the spatial dependence of gray level intensities that constitute the textures of an image. For a more detailed introduction to radiomics analysis, one may refer to [49].

For time varying acquisition protocols, such as dynamic PET and MR, radiomic features are extracted based on kinetic analysis of the dynamic images. Compartment models are widely used for tracer transport, its binding rates and metabolism modeling. For example, FDG-PET imaging has shown great success in tumor detection and cancer staging using^{18}F labelled fluorodeoxyglucose (FDG) as the tracer of choice to visualize intra-tumoral glucose metabolism. Beyond these common radiomic features (mostly statistical texture features), one can also apply other advanced pattern recognition features like fractal features, which are based on the concept of fractional Brownian motion and represent the normalized average absolute intensity difference of pixel pairs on a surface at different scales (a); Scale Invariant Features, which are invariant to image spatial scaling and rotations, and are able to provide robust matching across a large range of affine transformation [50]; or Histograms of Oriented Gradients (HOG) features, which are obtained by counting occurrence of gradient orientation in localized parts of an image [51]. In addition, one may also develop their own new ad hoc features based on the understanding of the specific task at hand.

3.3. Machine-engineered feature extraction (deep radiomics)

An alternative approach for feature extraction is driven purely by the data itself using machine learning techniques such as CNNs and is usually referred to as “feature engineering”. Unlike obtaining hand-crafted features as mentioned above, CNNs are able to engineer important features considered critical for computers to learn characteristics from various image data automatically, although these latent features are not always human-recognizable. But owing to the powerful performance of CNNs, computer scientists have managed to unveil the feature maps a computer can recognize, making such features more interpretable [31].

The idea of CNNs has been applied to medical image processing as early as 1993, Zhang et al. used a shift-invariant NN to detect clustered microcalcifications in digital mammograms [37]. Sahiner et al. investigated the classification of ROIs on mammograms using a CNN with spatial domains and texture images [38]. Recently, a large number of inspiring works applying deep CNNs to medical image analysis have been presented. For example, Shin et al. used three CNN architectures, namely CifarNet, AlexNet and GoogLeNet with transfer learning for computer-aided detection (CAD) problems. They reported that the applications of CNN image features can be improved by either exploring the hand-crafted features or by transfer learning (i.e., using information from other domain such as natural images to inform the medical application at hand) [39]. Although CNN methods require little engineering by hand and learn their features automatically, they are also limited by the available data size. In the medical field, it is relatively difficult to collect such large amount of data comparable to that in other fields such as computer vision or board games. Another limitation is the lack of labelling for the data (clinical outcomes). Even if the former two issues are resolved, the features obtained from CNNs may be hard to interpret in the clinical sense, which may not be reassuring for medical practitioners and patient care. Transfer learning, data augmentation, Generative Adversarial Nets (GAN) [40], semi-supervised learning among others have been proposed to address the data limitation problem by providing additional data for training, while deep learning interpretation is still in its infancy.

3.4. Feature selection, model construction and validation

In imaging tasks for radiation oncology such as sementation or tumour contouring, supervised learning is typically useful. As mentioned earlier, in supervised learning one finds a “good fit” for the labeled data among several ML and statistical models, where such a “good fit” is usually determined by training error and internal/external validation performance. It can be challenging to find an accurate and stable model fitting data with a large number of features extracted from medical images, especially when the sample size (patients’ number) is much less than the data features. In these situations, overfitting and high variance are the major concerns due to the so-called the *curse of higher dimensionality.* One viable method is to trim the data with feature selection, namely selecting a subset of variables that are indicative for one’s classification purpose. Such feature selection of data may be useful when features are redundant, highly correlated, or sometimes irrelevant with respect to the classification task.

By performing feature selection, the variable space is reduced to mitigate the tendency of overfitting such that it helps building a more robust predicting model. Furthermore, computation cost and storage of data can be reduced. It may also be beneficial for interpreting and understanding the underlying data mechanism if selected structures properly reflect the classification purpose. Common feature selection methods are filtering methods and wrapper methods [52], where the former selects variables by ranking them with correlation coefficients and the latter searches for an optimal set of features by evaluating "usefulness" towards a given predictor using different combinations of features within a learning scheme.

After feature selection of data, one proceeds to build a classification model by evaluating the performance on training data and generalization error on independent test (validation) set. A complex model can fit the training data well (low bias), while its generalization performance to out-of-sample may be poor (high variance).

Such bias–variance tradeoff is due to complexity of a model. A fundamental quantity called *Vapnik–Chervonenkis dimension (VC-dimension)* characterizes such model complexity for a class of models, e.g., NNs, linear classifiers, and SVMs. It is known that linear classifiers on 2D plane have VC-dim=3, while NNs have VC-dim= $O(n \log n)$ where n is the total number of parameters (weights) in the network [53]. One then easily sees that NNs are equipped with stronger capacity to fit (training) data, which meets our empirical understanding. Therefore, the purpose of validation and model selection aims to choose a proper class of classifiers with suitable VC-dimension to characterize the data. In fact, Vapnik has proved a useful formula describing the relation between training/testing error and VC-dimension [54]:

$$\Pr\left(test\ error \le training\ error + \sqrt{\frac{\left[D\left(\log\left(\frac{2N}{D}\right) + 1\right) - \log\left(\frac{\eta}{4}\right)\right]}{N}}\right) \\ = 1 - \eta$$

With $0 \le \eta \le 1$ and $D \ll N$ when D is the VC-dimension and N is the size of training set. This equation exactly describes that when the probability of test error is likely to be larger than training error and by how much amount (the square root term) determined by D. In particular, when the VC-dimension is large, the test error is most likely to be larger than the training error, and hence meets our intuition of over-fitting. With the knowledge where overfitting/underfitting comes from, one is then dedicated to select a proper model describing data.

To perform model assessment and overcome model selection, two major methods of validation may be utilized: K-fold cross-validation (CV) and bootstrap method, where K-fold (K: a positive integer, e.g., $K = 5, 8, 10, \ldots$) CV randomly splits data K times into approximately K equal-sized and mutually-exclusive parts with one part reserved as validation data and the other $K - 1$ ones served as training set. Choice of K is arbitrary, however $K = 5, 10, N$ are commonly used with N as the total sample size. In general, large K leads to low bias and overestimates the true prediction error since the training sets will be approaching to the entire dataset but tends to give higher variance [55]. In an extreme case where $K = N$, called Leave-One-Out CV (LOOCV), is almost unbiased but will have higher variance [56] and thus it is suitable for smaller datasets. Other choices like $K = 5, 10$ will provide good compromise between bias and variance tradeoff [57]. Some studies further reported that 10-fold CV yielded the best results with experiments on real-world data of certain types [58].

Another validation method called bootstrapping has a basic idea of randomly drawing samples with replacement from the training data, called *bootstrap samples,* with each bootstrap sample having the same size as the original training set. To apply the bootstrap idea for assessing models, several statistical estimators have been proposed: in particular, *Leave-One-Out bootstrap, “.632 estimator”* [56] , *and “.632+ estimator”* [59]. are commonly used. Leave-One-Out bootstrap mimicking LOOCV keeps tracks of predictions from bootstrap samples not containing certain observation i, such that its error estimator can overcome overfitting problem when compared to pure bootstrap method, where the estimator is given as the following:

$$Err = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b \in C^{-i}} L(y_i, f^{*b}(x_i))$$

“.632 estimator” [56] further alleviate bias towards estimates of prediction error. Another technique “.632+ estimator” [59] further improves .632 estimator by considering the amount of overfitting. It is known that bootstrap can be shown to fail in certain exquisitely designed statistical examples as well as in a case that memorizer module is added [58]. However, such artificial counterexamples require more mathematical labor that is beyond the scope of this paper, therefore interested readers are encouraged to view the construction in [55]. In many occasions, both CV and bootstrap methods are shown to be valuable and have compatible results. As in [55], a comparison of CV and bootstrap is demonstrated for particular problems and fitting models. They found that either cross-validation or bootstrap yields a model fairly close to the best available. Therefore, one usually needs to determine which validation method actually provides the best description based on the field test with one’s data at hand.

As a final remark, model selection and feature selection should not exhaust all samples, since the features or model selected by exhausting all samples may derive more optimistic performance estimation. To avoid this issue, in small-sample-size problems, nested cross-validation techniques can be utilized to make full use of the data as well as to give an unbiased prediction estimation, where typically an outer loop is created for performance estimation, and an inner loop (in contrast to the outer loop) is established for searching optimal hyper-parameters, model training, or feature selection. In any case, the gold standard for validating a model is performing external validation on independent datasets. This is highlighted in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) recommendations [60].

4. Examples of Machine Learning Application in Oncology

In this section, we illustrate how machine learning can extract relevant imaging information for oncology applications. Three examples shall be investigated, where the first two compare the radiomics using classical analysis and the modern deep CNN method; the last one directly utilizes deep learning for identifying metastatic breast cancer motivated by a recent grand challenge [9] [61].

Radiomics signatures have been built with the power to detect tumor heterogeneity, predict outcomes, (e.g., recurrence, distant metastases, response to certain treatment), and conduct survival analysis of various cancers. We will review several results obtained with radiomics, using conventional machine learning methods (such as random forests, SVMs), and compare them with deep learning methods: e.g. CNNs, and present one example that fuses both conventional and deep models (i.e., deep radiomics).

Example 1. Conventional radiomics for predicting failure risks in head and neck cancers

Vallieres et al. [41] built a radiomics model based on pre-treatment FDG-PET and CT images to analyze the risk assessment of loco-regional recurrences (LR), distant metastases (DM) and overall survival (OS) in a head-and-neck cancer. 1,615 radiomic features from 300 patients from four institutions were extracted and divided into training and testing datasets. Considering the large number of features, feature set reduction was first applied by ranking the features based on an information gain criterion, where, the correlation of the feature with endpoints and features their inter-correlations were considered. The goal was to maximize the relevance of features with outcomes and minimize the redundancy. Subsequently, a forward feature selection method was applied to determine the final model. Feature selection, prediction performance estimation, choice of model complexity and final model computation processes were carried out using logistic regression (classification), Cox regression (survival analysis) and bootstrap resampling. Prediction models consisting of radiomic information only were first constructed for each of the three H&N outcomes. Meanwhile, clinical factors (e.g., tumor volume, age, T-stage) were analyzed by stratified random sub-sampling in the training set and resulted in three groups of clinical variables for each outcome. Final prediction models were constructed by combining the selected radiomic and clinical features via random forests. The performance of prediction models was estimated using receiver operating characteristic metrics (ROC), the concordance index (CI) and the p-value obtained from Kaplan-Meier analysis using the log-rank test between two risk groups (LR: AUC=0.69, CI=0.67; DM: AUC=0.86, CI=0.88). The analysis showed that radiomics can provide important prognostic information for the risk assessment of the three outcomes in head-and-neck cancer in this study.

Example 2. Deep radiomics for breast cancer diagnosis

Antropova et al. [62] devised a method that extracted low- to mid- level features using a pretrained CNN and combined them with hand-crafted radiomic features for breast cancer diagnosis. Full-field digital mammography (FFDM), breast ultrasound, and dynamic contrast enhanced-MRI (DCE-MRI) images were used in this study. Three datasets were used separately to test the methodology. CNN features were extracted with VGG19 architecture, pretrained on ImageNet. There are five stacks in the architecture, with each stack containing two or four convolutional layers and a max pooling layer, followed by three fully connected layers. Images were duplicated to be input to the three-color channels for FFDM and ultrasound data extracted at precontrast and the first two postcontrast time points for the DCE-MRI dataset. CNN features were then obtained from each of the five max-pool layers. Five feature vectors were obtained by average-pooling along spatial dimensions. This method avoided the preprocessing step for images with varying sizes by preserving the spatial structure. Radiomic features describing lesion properties such as size, shape, texture, and morphology were also extracted from ROIs. A nonlinear SVM with a radial basis function (RBF) kernel was used to build models for both CNN and radiomic features. The two classifiers were fused by averaging the outputs to give the final model. From ROC analysis, they claimed that the fusion-based method, on all imaging modalities, performed significantly better than conventional radiomic models in the task of distinguishing benign and malignant lesions, with an AUC = 0.89 for DCE-MRI, AUC = 0.86 for FFDM and AUC = 0.90 for ultrasound. In summary, their analysis showed the feasibility of combining deep learning and conventional feature extraction methods for breast cancer diagnosis.

Example 3. A deep CNN for Detection of Lymph Node Metastases

An automated detection of cancer challenge, Camelyon16, was setup [9] to develop algorithms for lymph node detection of metastases in women with breast cancer, where a training dataset of 399 whole-slide images (WSIs) collected from two institutions, Radboud University Medical Center (RUMC) and University Medical Center Utrecht (UMCU), was provided. The competition was intended to motivate machine learning algorithms in the application of medical cancer imaging, where two independent tasks were asked to be evaluated. In task#1, the participants were asked to demonstrate the ability of localizing tumor; in task#2, the participants were asked to discriminate images with or without sentinel axillary lymph nodes (SLNs), i.e., an image classification problem. During the competition, a panel of 11 pathologists participated and independently reviewed the same dataset. Of all the 23 teams participating, the best results (algorithms) were derived by a joint team led by Harvard Medical School (HMS) and Massachusetts Institute of Technology (MIT), where Wang et al. [61] designed a composite image classifier using a deep CNN, GoogLeNet of 27 layers and in total more than 6 million parameters. Essentially, their neural networks were trained by millions of patches (out of the whole slide image, see Fig. 7(a)) for patch-level predictions to discriminate tumor patches from normal patches. These patch-bypatch predictions were then gathered to form a complete tumor probability heat map for one whole slide, as in Fig. 7(b). Furthermore, to decrease the computational time, several techniques were applied, including transforming images from RGB color into HSV color and identifying tissues via meaningless white background removed.

In fact, before Wang *et al.* presented their final winning model, several existing advanced architectures were tested for selecting strong candidates, such as GoogLeNet [34], AlexNet [63], VGG16 [64], and a face orientated deep network [65] giving the following testing performance:

	patch-wised classification accuracy
GoogLeNet	98.4 %
AlexNet	92.1 %
VGG16	97.9 %
FaceNet	96.8 %

Naturally, due to this experiment, GoogLeNet appeared to be the strongest candidate for their task, and thus were utilized as the winning model. By their delicate design of data preprocessing, model selection between several viable CNN architectures, and final data post-processing, their classifiers were able to achieve an AUC=0.925 for task#2 of whole slide image classification and a score of 0.7051 for the tumor localization of task#1, where a human pathologist independently cross-validated the image data, obtaining a whole slide image classification AUC=0.966 and a tumor localization score of 0.733. In terms of AUC score, it can be concluded that the designed CNN classifiers reached similar accuracy as a board-certified expert. Interestingly, they found that the errors made by the pathologist and the deep learning system were not strongly correlated. Therefore, by combining their deep learning classifiers with the human pathologist's diagnoses the pathologist's accuracy was increased to AUC=0.995, reducing in an approximately 85% human error rate. It was also noted in [9] that this was the first study to recognize that machine learning algorithms can rival human pathologists' performance. Another interesting observation was that deep learning-based algorithms significantly outperformed other conventional ML methods, where the top 19 out of all 23 teams utilized deep CNNs.

5. Discussion

Modern machine learning algorithms serve as powerful tools to improve medical practice by reducing human labor and possible errors. They can potentially improve a patient's diagnosis and treatment precision by complementing human perception. With the latest innovations in machine learning techniques, we are looking towards an exciting data revolution in the medical field in general and in oncology in particular. ML is expected to allow efficient utilization of resources and to save time and unnecessary medical expenses to patients, their doctors, and the society at large.

However, and before having a full-fledged embracement of this new digital revolution, one needs to validate whether these innovative technologies can be widely adopted in medicine. For example, it may still be unclear clinically how to decide which kind of features are better suited for solving a specific diagnostic or therapeutic task using a machine learning algorithm (e.g., hand-crafted features vs. machine-learned features). In the example of identifying metastatic breast cancer, researchers found that the recognition of deep learning systems and human pathologists can be complimentary, where the system helped reduce the pathologist error rate from over 3% to less than 1%. Therefore, at the current stage deep learning systems are more suited as a secondary opinion to aid in decision support or quality checks, rather than a stand-alone system.

One key factor for improving machine learning performance is the available quality and quantity of data. Before full development of AI (as in Fig. 1), computers are unlikely able to comprehend complex physical laws out of only a few examples, but rather are only capable of deducing empirical relations based on large observations by statistical inference algorithms. In the medical field where datasets tend to be small, partially observed or labeled, and sometimes noisy, the full optimization of machine learning can be a challenge. In the medical imaging scope, if a given dataset is too small and/or noisy, although hand-crafted features can be timeconsuming to generate, they can take advantage of prior domain knowledge and may outperform undertrained deep NN/CNN methods. However, there are a few steps that may mitigate the small data problem: (1) data preprocessing (e.g., PCA or autoencoders) for reducing the number of fitting parameters a priori; (2) data augmentation techniques such as transfer learning and GANs in attempt to overcome sample size limitations. In certain experiments such as image segmentation, diagnosis and endpoint prediction tasks [66] data augmentation techniques methods have demonstrated promising results; and (3) considering the combination of traditional features and CNNs features applied to images.

Gathering and sharing datasets across institutions certainly serves as another viable way to increase data size and improve machine learning utility. However, certain problems may also arise: e.g., how to train models from heterogenous datasets from different institutes that may also have varying data formats? The harmonization and standardization of naming conventions (abbreviations, code names, etc) alone can easily confuse computers and even lead to miscalculation by various feature definitions pertaining to one’s institute. Another important challenge is maintaining the confidentiality and the privacy of patient information in such data sharing processes where administrative (Institutional Review Boards), regulations, laws (e.g. HIPAA Privacy Rule) could be at risk of violation. For the purpose of privacy protection, a newly developing cryptography technology especially for datamining and statistical data queries called *differential privacy* [67] [68] can be applied to shared medical datasets, where the basic idea is the injection of proper noise level, hashing, or subsampling to scramble the original dataset from possible unwarranted probing. Large companies such as APPLE, GOOGLE, Facebook and Microsoft are applying and promoting such technologies for their customer privacy concerns [69]. Another approach is the utilization of distributed (rapid) learning presented in Eurocat [70] [71], where algorithms instead of data are shared across the different institutions. With all these exciting breakthroughs in machine learning and their potential in oncology, one still needs to be careful when wielding such methods and meticulously design the data validation experiments to avoid pitfalls of overfitting and mis-information [72].

6. Conclusion

The past few years have observed tremendous rise in machine learning applications to wide range of areas in oncology including: building predictive models of disease diagnosis, treatment response, and automation of workflow and decision support. But as methods and techniques in machine learning keep evolving, one can expect the role of machine learning to continue reshaping the field of oncology and cancer management. Machine learning is expected to alter the way patients receive treatments and doctors reach their clinical decisions. Diagnostics will be faster, cheaper, and more accurate than ever. However, to usher the advent of machine learning era, one has to be mindful of the characteristics and the limitations of this technology too. Machine learning methods require large amounts of data for their training and validation, which also beg the questions of computerized trust, data sharing and privacy concerns. With the pre-existing domain knowledge, the merits of man-crafted features standing for accumulative knowledge based on numerous observations should be incorporated and inherently infused with modern machine learning architectures. With the assistance of the state-of-the-art machine learning algorithms, imaging informatics holds the potential to provide better precision health care for cancer patients as well as revealing underlying biological patterns. The application of machine learning algorithms in the medical realm is promising, yet there remain many challenges before they can realize their potential into routine clinical oncology practice.

Acknowledgements

This work was supported in part by the National Institutes of Health P01 CA059827.

References

[1]. Mitchell TM et al., “Machine learning. wcb,” 1997.

[2]. Huang S-H and Pan Y-C, “Automated visual inspection in the semiconductor industry: A survey,” Computers in Industry, vol. 66, pp. 1 – 10, 2015 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166361514001845

[3]. LLC G. (2018) Google cloud speech-to-text enables developers to convert audio to text by applying powerful neural network models in an easy to use api. [Online]. Available: https://cloud.google.com/speech-to-text/

[4]. Ciocca S. (2017) How does spotify know you so well? [Online]. Available: https://medium.com/s/story/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe

[5]. Kooti F, Grbovic M, Aiello LM, Djuric N, Radosavljevic V, and Lerman K, “Analyzing uber’s ride-sharing economy,” in Proceedings of the 26th International Conference on World Wide Web Companion International World Wide Web Conferences Steering Committee, 2017, pp. 574–582.

[6]. Somashekhar SP, SepÃºlveda MJ, Puglielli S, Norden AD, Shortliffe EH, Rohit Kumar C, Rauthan A, Arun Kumar N, Patil P, Rhee K, and Ramya Y, “Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board,” Annals of Oncology, vol. 29, no. 2, pp. 418–423, 2018 [Online]. Available: [PubMed: 29324970]

[7]. ——, “Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board,” Annals of Oncology, vol. 29, no. 2, pp. 418–423, 2018 [Online]. Available: 10.1093/annonc/mdx781 [PubMed: 29324970]

[8]. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, and Thrun S, “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, pp. 115–118, Jan. 2017 [Online]. Available: 10.1038/-nature21056 [PubMed: 28117445]

[9]. E. B. B, V M, van Diest P J, and et al., “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” JAMA, vol. 318, no. 22, pp. 2199–2210, 2017 [Online]. Available: + 10.1001/-jama.2017.14585 [PubMed: 29234806]

[10]. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A, Hruban RH, Wolfgang CL, Goggins MG, Dal Molin M, Wang T-L, Roden R, Klein AP, Ptak J, Dobbyn L, Schaefer J, Silliman N, Popoli M, Vogelstein JT, Browne JD, Schoen RE, Brand RE, Tie J, Gibbs P, Wong H-L, Mansfield AS, Jen J, Hanash SM, Falconi M, Allen PJ, Zhou S, Bettegowda C, Diaz L, Tomasetti C, Kinzler KW, Vogelstein B, Lennon AM, and Papadopoulos N, “Detection and localization of surgically resectable cancers with a multi-analyte blood test,” Science, 2018 [Online]. Available: http://science.sciencemag.org/content/early/2018/01/17/science.aar3247

[11]. contributors W, “John McCarthy (computer scientist)—Wikipedia, the free encyclopedia,” 2018, [Online; accessed 16-March-2018]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=John_McCarthy_(computer_scientist)&oldid=823313442

[12]. Christopher MB, PATTERN RECOGNITION AND MACHINE LEARNING. Springer-Verlag New York, 2016.

[13]. Michalski RS, Carbonell JG, and Mitchell TM, Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

[14]. Jolliffe IT, “Principal component analysis and factor analysis,” in Principal component analysis. Springer, 1986, pp. 115–128.

[15]. Belkin M and Niyogi P, “Laplacian eigenmaps for dimensionality reduction and data representation,” Neural Computation, vol. 15, no. 6, pp. 1373–1396, 2003 [Online]. Available: 10.1162/089976603321780317

[16]. Maaten L. v. d. and Hinton G, “Visualizing data using t-SNE,” Journal of machine learning research, vol. 9, no. Nov, pp. 2579–2605, 2008.

[17]. Tseng HH, Naqa IE, and Chien JT, “Power-law stochastic neighbor embedding,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3 2017, pp. 2347–2351.

[18]. Baldi P, “Autoencoders, unsupervised learning, and deep architectures,” in Proceedings of ICML Workshop on Unsupervised and Transfer Learning, 2012, pp. 37–49.

[19]. Dawson LA, Biersack M, Lockwood G, Eisbruch A, Lawrence TS, and Ten Haken RK, “Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation,” International Journal of Radiation Oncology Biology Physics, vol. 62, no. 3, pp. 829–837, 2005.

[20]. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot Met al., “Mastering the game of go with deep neural networks and tree search,” nature, vol. 529, no. 7587, pp. 484–489, 2016. [PubMed: 26819042]

[21]. Bellman R, “A markovian decision process,” Indiana Univ. Math. J, vol. 6, pp. 679–684, 1957.

[22]. Ng AY and Jordan MI, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in Advances in Neural Information Processing Systems 14, Dietterich TG, Becker S, and Ghahramani Z, Eds. Press MIT, 2002, pp. 841–848. [Online]. Available: http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers\-a-comparison-of-logistic-regression-and-naive-bayes.pdf

[23]. Andel PMGNAI, Jirí, “engEstimating the dimension of a linear model,” engKybernetika, vol. 17, no. 6, pp. 514–525, 1981 [Online]. Available: http://eudml.org/doc/28796

[24]. Hochreiter S and Schmidhuber J, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997 [Online]. Available: 10.1162/neco.1997.9.8.1735 [PubMed: 9377276]

[25]. Sarkka S, Solin A, and Hartikainen J, “Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering,” IEEE Signal Processing Magazine, vol. 30, no. 4, pp. 51–61, 7 2013.

[26]. Madhavan P, Systems Analytics: Adaptive Machine Learning Workbook, 1st ed. USA: CreateSpace Independent Publishing Platform, 2016.

[27]. LeCun Y, Bengio Y, and Hinton G, “Deep learning,” nature, vol. 521, no. 7553, p. 436, 2015. [PubMed: 26017442]

[28]. Hornik K, “Approximation capabilities of multilayer feedforward networks,” Neural Networks, vol. 4, no. 2, pp. 251 – 257, 1991 [Online]. Available: http://www.sciencedirect.com/science/article/pii/089360809190009T

[29]. Zeiler MD and Fergus R, “Visualizing and understanding convolutional networks,” in Computer Vision – ECCV 2014, Fleet D, Pajdla T, Schiele B, and Tuytelaars T, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.

[30]. Hubei DH and Wiesel TN, “Receptive fields and functional architecture of monkey striate cortex,” The Journal of Physiology, vol. 195, no. 1, pp. 215–243. [Online]. Available: https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1968.sp008455

[31]. Zeiler MD and Fergus R, "Visualizing and understanding convolutional networks," in Computer Vision – ECCV 2014, Fleet D, Pajdla T, Schiele B, and Tuytelaars T, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.

[32]. Lecun Y, Bottou L, Bengio Y, and Haffner P, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 11 1998.

[33]. Krizhevsky A, Sutskever I, and Hinton GE, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, Pereira F, Burges CJC, Bottou L, and Weinberger KQ, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[34]. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, and Rabinovich A, "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6 2015, pp. 1–9.

[35]. Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014 [Online]. Available: http://arxiv.org/abs/1409.1556

[36]. Ronneberger O, Fischer P, and Brox T, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Navab N, Hornegger J, Wells WM, and Frangi AF, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[37]. Naqa I, *A Guide to Outcome Modeling in Radiotherapy and Oncology: Listening to the Data,* ser Series in Medical Physics and Biomedical Engineering. CRC Press, Taylor & Francis Group, 2018.

[38]. Gillies RJ, Kinahan PE, and Hricak H, "Radiomics: Images are more than pictures, they are data," Radiology, vol. 278, no. 2, pp. 563–577, 2016, pMID: [Online]. Available: 10.1148/radiol.2015151169 [PubMed: 26579733]

[39]. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, chen X, Chan BK, Matcuk GR, Barry CT, Chang HY et al., "Decoding global gene expression programs in liver cancer by noninvasive imaging," Nature biotechnology, vol. 25, no. 6, p. 675, 2007.

[40]. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," Nature communications, vol. 5, p. 4006, 2014.

[41]. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJ, Khaouam N, Nguyen-Tan PF, Wang C-S, Sultanem K et al., "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," Scientific reports, vol. 7, no. 1, p. 10117, 2017. [PubMed: 28860628]

[42]. Naqa IE, "The role of quantitative pet in predicting cancer treatment outcomes," Clinical and Translational Imaging, vol. 2, no. 4, pp. 305–320, 8 2014.

[43]. Haralick RM, Shanmugam K, and Dinstein I, "Textural features for image classification," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610–621, 11 1973.

[44]. Galloway MM, "Texture analysis using gray level run lengths," Computer Graphics and Image Processing, vol. 4, no. 2, pp. 172 – 179, 1975 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0146664X75800086

[45]. Thibault G, FERTIL B, Navarro C, Pereira S, LÃ©vy N, SEQUEIRA J, and MARI J-L, "Texture indexes and gray level size zone matrix application to cell nuclei classification," 11 2009.

[46]. Thibault G, Angulo J, and Meyer F, "Advanced statistical matrices for texture characterization: Application to cell classification," IEEE Transactions on Biomedical Engineering, vol. 61, no. 3, pp. 630–637, 3 2014. [PubMed: 24108747]

[47]. Amadasun M and King R, "Textural features corresponding to textural properties," IEEE Transactions on Systems, Man, and Cybernetics, vol. 19, no. 5, pp. 1264–1274, 9 1989.

[48]. Sun C and Wee WG, "Neighboring gray level dependence matrix for texture classification," Computer Vision, Graphics, and Image Processing, vol. 23, no. 3, pp. 341 – 352, 1983 [Online]. Available: http://www.sciencedirect.com/science/article/pii/-0734189X83900324

[49]. Zwanenburg A, Leger S, Vallières M, and Löck S, "Image biomarker standardisation initiative - feature definitions," CoRR, vol. abs/1612.07003, 2016 [Online]. Available: http://arxiv.org/abs/1612.07003

[50]. Lowe DG, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 11 2004 [Online]. Available: 10.1023/B:VISI.0000029664.99615.94

[51]. Freeman WT and Roth M, "Orientation histograms for hand gesture recognition," MERL - Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, Tech. Rep. TR94-03, Dec. 1994 [Online]. Available: http://www.merl.com/publications/TR94-03/

[52]. Guyon I and Elisseeff A, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. Mar, pp. 1157–1182, 2003.

[53]. Baum EB and Haussler D, "What size net gives valid generalization?" in Advances in neural information processing systems, 1989, pp. 81–90.

[54]. Vapnik V, The Nature of Statistical Learning Theory. Springer New York, 2013 [Online]. Available: https://books.google.com/-books?id=EoDSBwAAQBAJ

[55]. Hastie T, Tibshirani R, and Friedman J, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, ser. Springer Series in Statistics. Springer New York, 2013 [Online]. Available: https://books.google.com/books?id=yPfZBwAAQBAJ

[56]. Efron B, "Estimating the error rate of a prediction rule: Improvement on cross-validation," Journal of the American Statistical Association, vol. 78, no. 382, pp. 316–331, 1983 [Online]. Available: http://www.jstor.org/stable/2288636

[57]. Breiman L and Spector P, "Submodel selection and evaluation in regression. the x-random case," International Statistical Review / Revue Internationale de Statistique, vol. 60, no. 3, pp. 291–319, 1992 [Online]. Available: http://www.jstor.org/stable/1403680

[58]. Kohavi R, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence - *Volume* 2, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143. [Online]. Available: http://dl.acm.org/citation.cfm?id=1643031.1643047

[59]. Efron B and Tibshirani R, "Improvements on cross-validation: The .632+ bootstrap method," Journal of the American Statistical Association, vol. 92, no. 438, pp. 548–560, 1997 [Online]. Available: http://www.jstor.org/stable/2965703

[60]. GS C, JB R, DG A, and KM M, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement," Annals of Internal Medicine, vol. 162, no. 1, pp. 55–63, 2015 [Online]. Available: + [PubMed: 25560714]

[61]. Wang D, Khosla A, Gargeya R, Irshad H, and Beck AH, "Deep learning for identifying metastatic breast cancer," arXiv preprint arXiv: 1606.05718, 2016.

[62]. Antropova N, Huynh BQ, and Giger ML, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," Medical Physics, vol. 44, no. 10, pp. 5162–5171. [Online]. Available: https://-aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12453

[63]. Krizhevsky A, Sutskever I, and Hinton GE, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[64]. Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv: 1409.1556, 2014.

[65]. Wang D, Otto C, and Jain AK, "Face search at scale: 80 million gallery," arXiv preprint arXiv:1507.07242, 2015.

[66]. Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q, and Shen D, "EnglishMedical image synthesis with context-aware generative adversarial networks," in EnglishMedical Image Computing and Computer Assisted Intervention âˆ' MICCAI 2017 – 20th International Conference, Proceedings, ser. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10435 LNCS. Springer Verlag, 2017, pp. 417–425.

[67]. Dwork C, Roth A et al., "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.

[68]. Vu D and Slavkovic A, “Differential privacy for clinical trial data: Preliminary evaluations,” in 2009 IEEE International Conference on Data Mining Workshops, 12 2009, pp. 138–143.

[69]. contributors W, “Differential privacy — wikipedia, the free encyclopedia,” 2018, [Online; accessed 4-April-2018]. [Online], Available: https://en.wikipedia.org/w/index.php?title=Differential_privacy&oldid=833357507

[70]. Lambin P, Zindler J, Vanneste B, van de Voorde L, Jacobs M, Eekers D, Peerlings J, Reymen B, Larue RTHM, Deist TM, de Jong EEC, Even AJG, Berlanga AJ, Roelofs E, Cheng Q, Carvalho S, Leijenaar RTH, Zegers CML, van Limbergen E, Berbee M, van Elmpt W, Oberije C, Houben R, Dekker A, Boersma L, Verhaegen F, Bosmans G, Hoebers F, Smits K, and Walsh S, “Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine,” Acta Oncologica, vol. 54, no. 9, pp. 1289–1300, 2015, pMID: [Online]. Available: 10.3109/0284186X.2015.1062136 [PubMed: 26395528]

[71]. Lambin P, Zindler J, Vanneste BG, Voorde LVD, Eekers D, Compter I, Panth KM, Peerlings J, Larue RT, Deist TM, Jochems A, Lustberg T, van Soest J, de Jong EE, Even AJ, Reymen B, Rekers N, van Gisbergen M, Roelofs E, Carvalho S, Leijenaar RT, Zegers CM, Jacobs M, van Timmeren J, Brouwers P, Lal JA, Dubois L, Yaromina A, Limbergen EJV, Berbee M, van Elmpt W, Oberije C, Ramaekers B, Dekker A, Boersma LJ, Hoebers F, Smits KM, Berlanga AJ, and Walsh S, “Decision support systems for personalized and participative radiation oncology,” *Advanced Drug Delivery Reviews*, vol. 109, pp. 131 – 153, 2017, radiotherapy for cancer: present and future. [Online]. Available: http://www.sciencedirect.com/science/-article/pii/S0169409X16300084

[72]. Naqa Issam El, Ruan Dan, Valdes Gilmer, Dekker Andre, Todd McNutt Yaorong Ge, Wu Jackie, Jung Hun Oh Maria Thor, Smith Wade, Rao Arvind, Fuller Clifton, Xiao Ying, Manion Frank, Schipper Matthew, Mayo Charles, Moran Jean, Ten Haken Randall, “Machine learning and modeling: Data, validation, communication challenges (in press),” Medical Physics, vol. 2018.

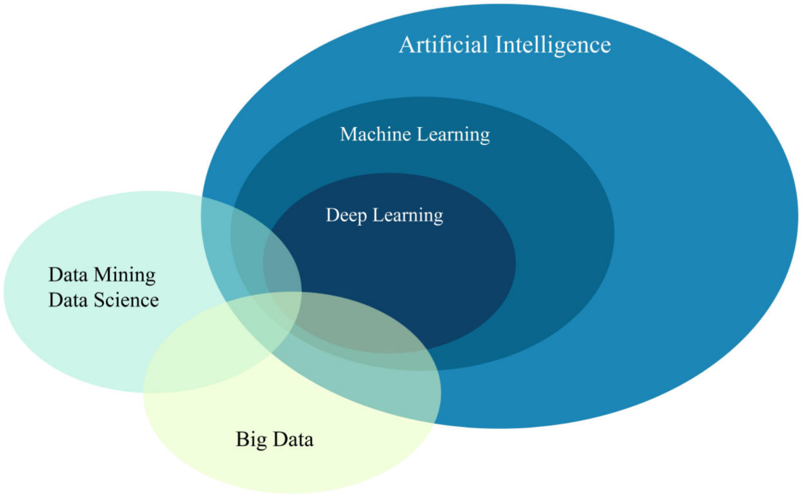


Fig. 1.

A schematic of the relation between AI, ML, Deep Learning, Big Data, and Data Science. It is noted that machine learning is a computational branch from AI that aims to provide computers with ability to perform tasks beyond their original programming such as data mining and big data analytics.

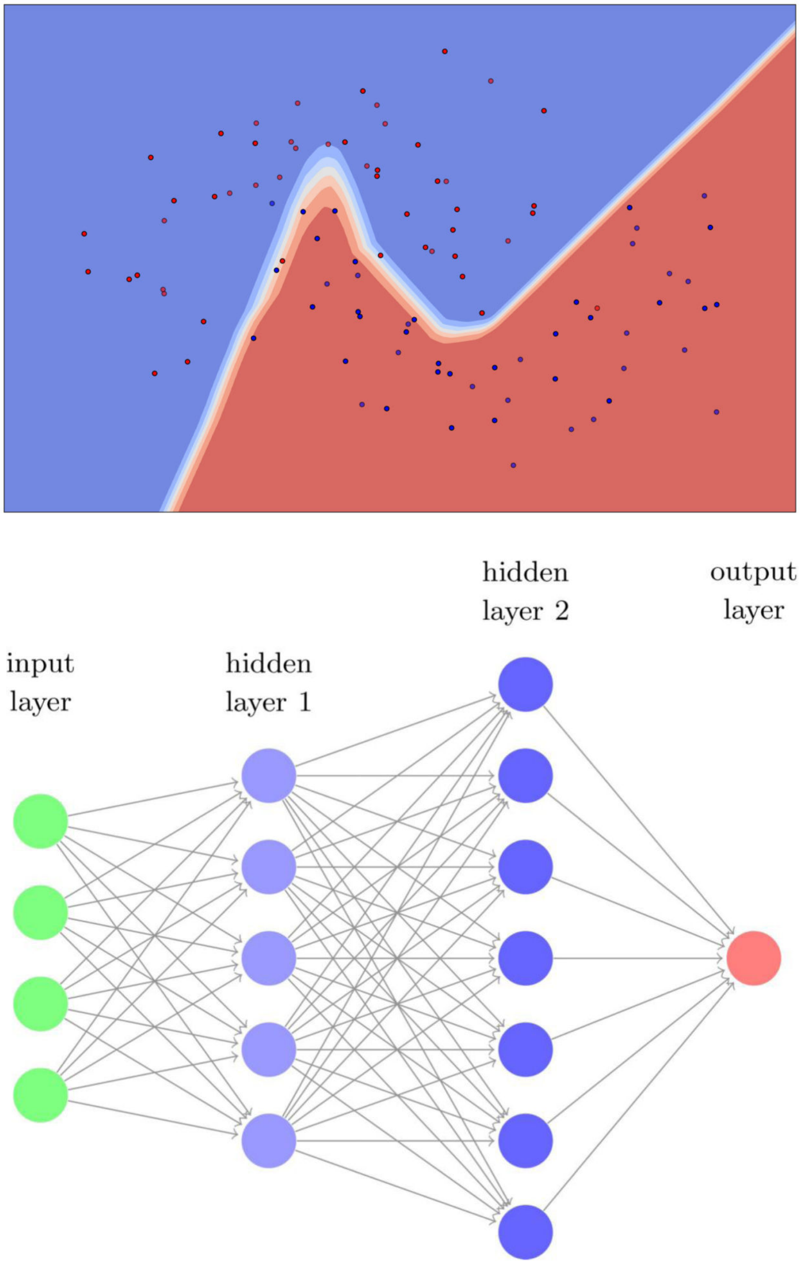
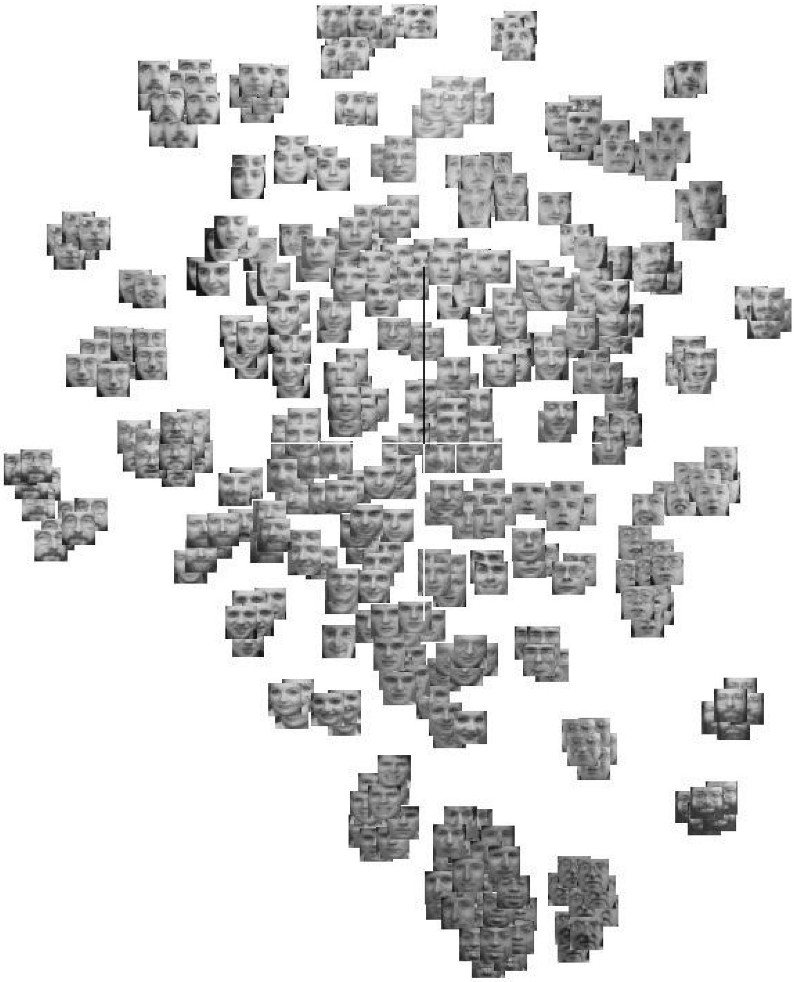


Fig. 2.

(a) (Left) An illustration of supervised learning using neural networks (right figure) classifying synthetic data of binary labels (blue and red scatter dots), where the nonlinear decision boundary is shown in white. (b) A multi-layer (deep) neural network with two hidden layers. The so-called deep learning usually refers to learning algorithms heavily relying on such computational units.



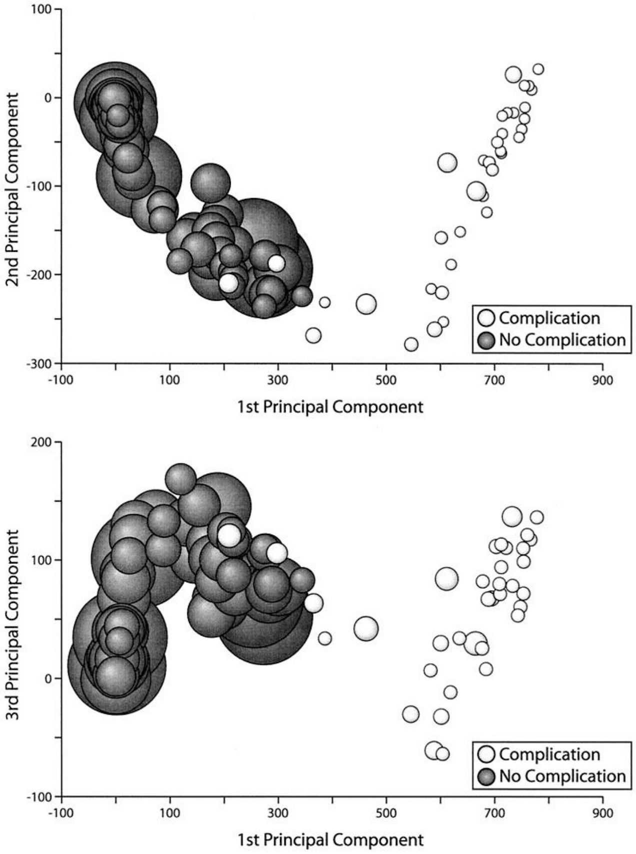


Fig. 3.

(a) [Left] An illustration of an unsupervised learning using p-SNE with open image database Olivetti faces, where similar images (same person) are clustered automatically without providing any identity information. (b) [Right (reprint permission granted)] Dawson et al. [13] demonstrated that PCA can be used to observe clinical data structure. In this case the data describing the xerostomia occurrences due to parotid gland dose distributions is linearly separable.

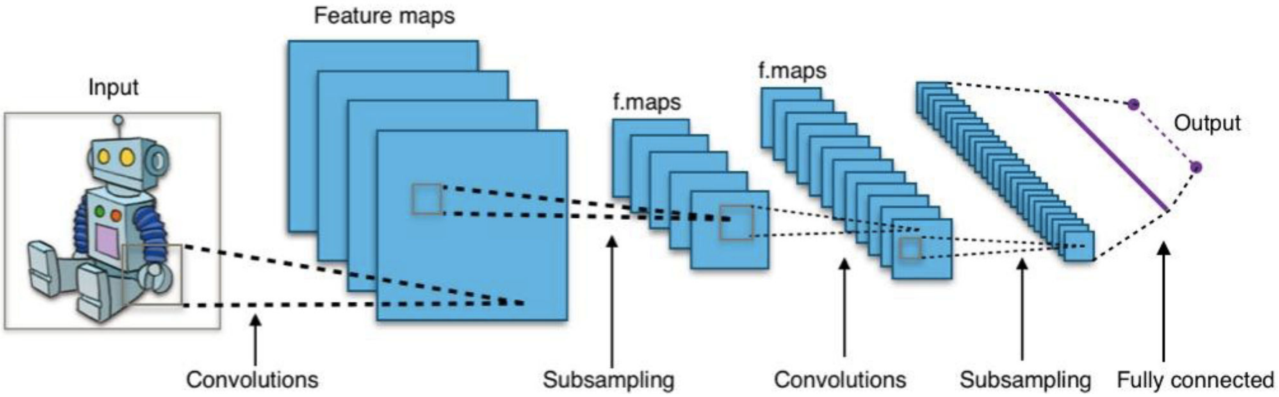
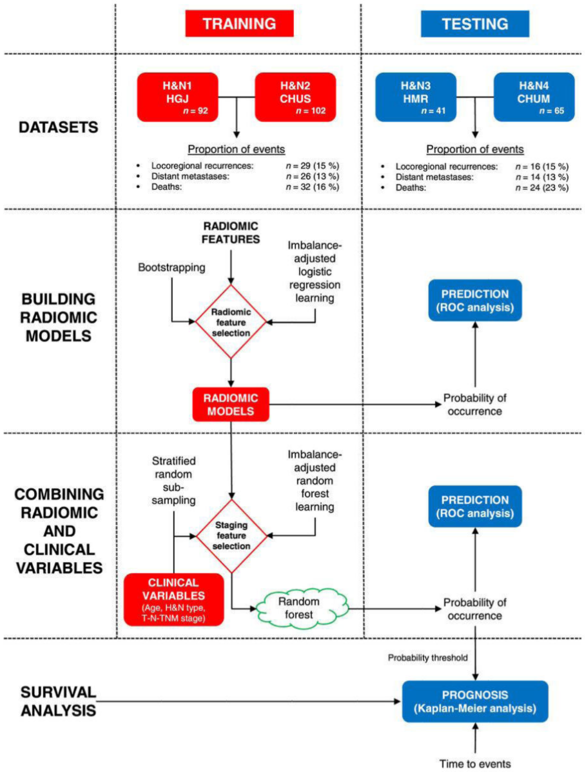


Fig. 4.

The structure of an CNN, usually consisting of three distinct layers: the convolution layer, the pooling layer, and a final fully-connected layer (Fig. 2(b)), where the convolution layer and pooling (subsampling) layer may be connected several times before a final fully-connected layer is encountered. An image mapped by a convolution layer is called a feature map, which triggers attention of many computer scientists. Figure created by Aphex34 distributed under a CC BY-SA 4.0 license (from Wikimedia Commons).


TRAINING
TESTING
DATASETS
H&N1 HGJ n = 92
H&N2 CHUS n = 102
H&N3 HMR n = 41
H&N4 CHUM n = 65
Proportion of events
Locoregional recurrences: n = 29 (15 %)
Distant metastases: n = 26 (13 %)
Deaths: n = 32 (16 %)
Proportion of events
Locoregional recurrences: n = 16 (15 %)
Distant metastases: n = 14 (13 %)
Deaths: n = 24 (23 %)
BUILDING RADIOMIC MODELS
RADIOMIC FEATURES
Bootstrapping
Imbalance-adjusted logistic regression learning
Radiomic feature selection
RADIOMIC MODELS
PREDICTION (ROC analysis)
Probability of occurrence
COMBINING RADIOMIC AND CLINICAL VARIABLES
Stratified random sub-sampling
Imbalance-adjusted random forest learning
Staging feature selection
CLINICAL VARIABLES (Age, H&N type, T-N-TNM stage)
Random forest
PREDICTION (ROC analysis)
Probability of occurrence
Probability threshold
SURVIVAL ANALYSIS
PROGNOSIS (Kaplan-Meier analysis)
Time to events

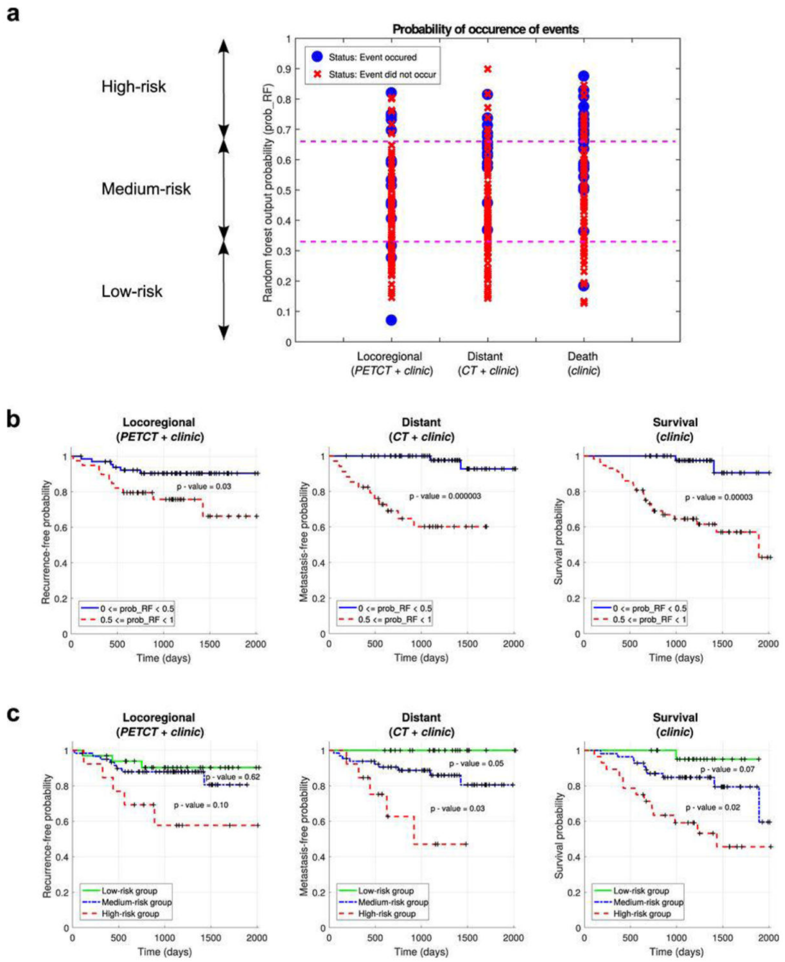
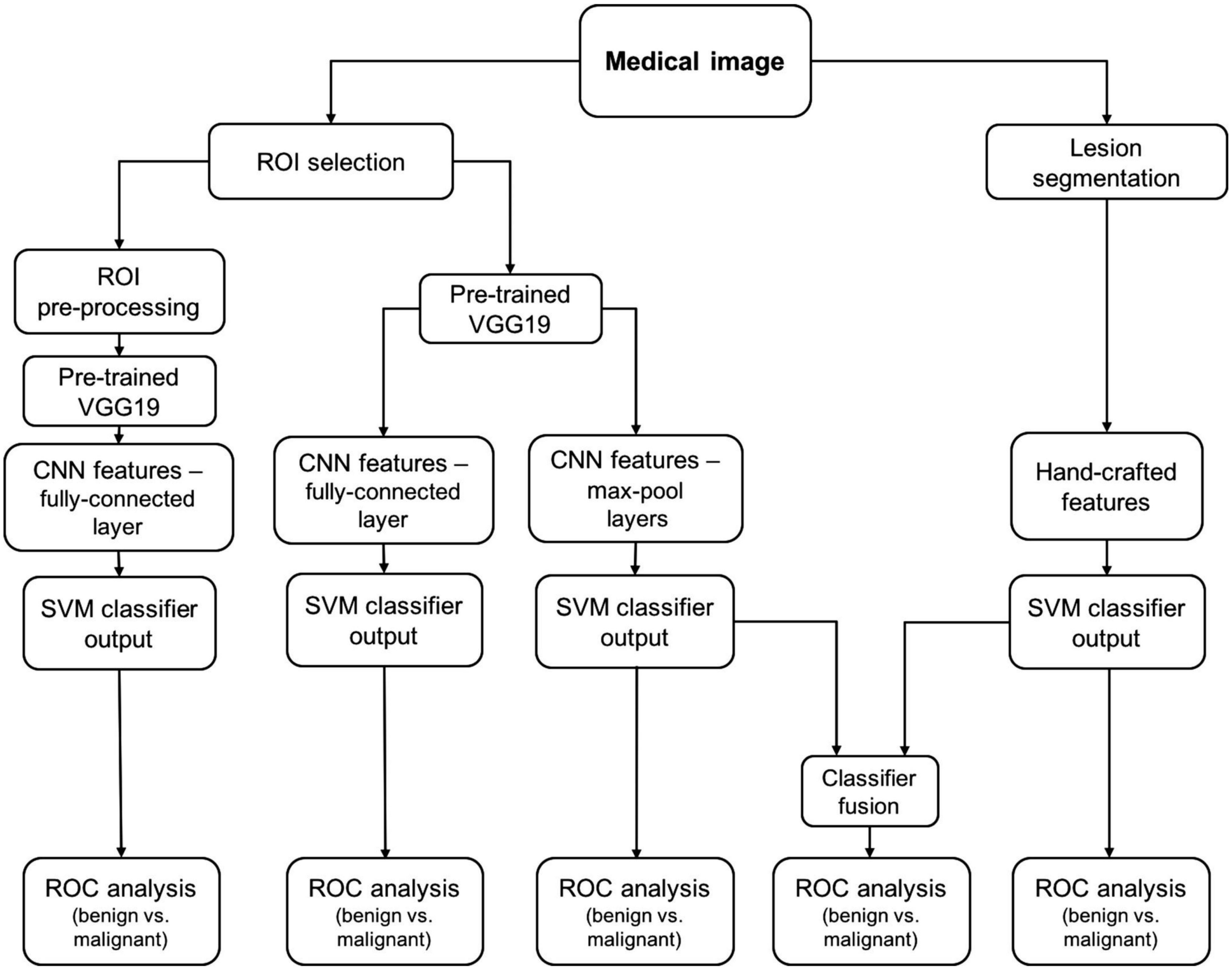


Fig. 5.

(a)[Left] The workflow of the model built by Vallieres et al. [41] The best combinations of radiomic features were selected in the training set, where these radiomic features were then combined with selected clinical variables in the training set. Independent prediction analysis was later performed in the testing set for all classifiers fully constructed in the training set. (b)[Right] Risk assessment of tumor outcomes in [41]. (1) Probability of occurrence of events for each patient of the testing set. The output probability of occurrence of events of random forests allows for risk stratification. (2) Kaplan-Meier curves of the testing set using a risk stratification into two groups as defined by a random forest output probability threshold of 0.5. All curves show significant prognostic performance. (3) Kaplan-Meier curves of the testing set using a risk stratification into three groups as defined by random forest output probability thresholds of 1/3 and 2/3.


Medical image
ROI selection
Lesion segmentation
ROI pre-processing
Pre-trained VGG19
Pre-trained VGG19
CNN features – fully-connected layer
CNN features – fully-connected layer
CNN features – max-pool layers
Hand-crafted features
SVM classifier output
SVM classifier output
SVM classifier output
SVM classifier output
Classifier fusion
ROC analysis (benign vs. malignant)
ROC analysis (benign vs. malignant)
ROC analysis (benign vs. malignant)
ROC analysis (benign vs. malignant)
ROC analysis (benign vs. malignant)

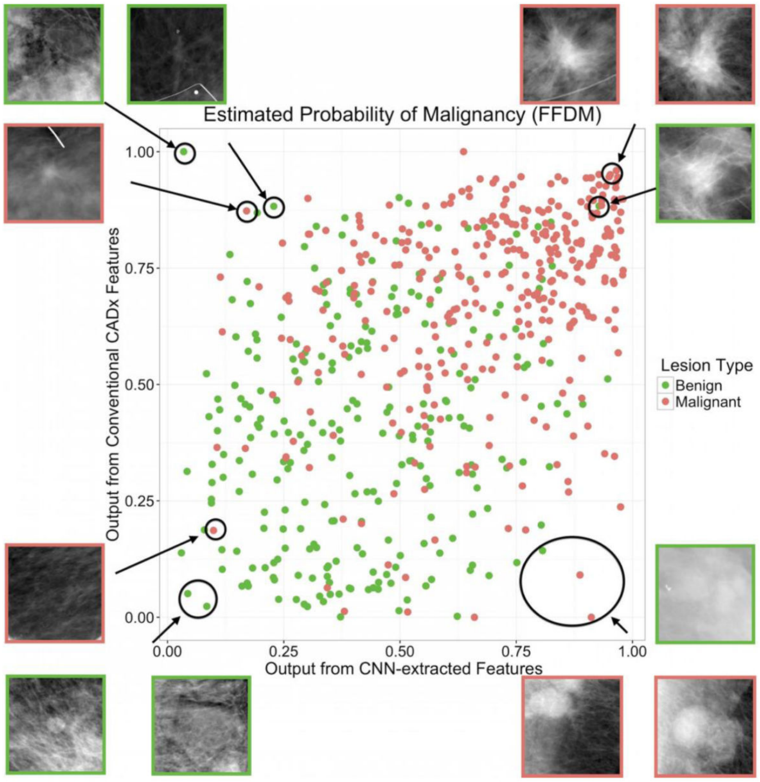


Fig. 6.

Lesion classification pipeline based on diagnostic images. Two types of features are extracted from a medical image: (a) CNN features with pretrained CNN and (b) handcrafted features with conventional CADx. High and low-level features extracted by pretrained CNN are evaluated in terms of their classification performance and preprocessing requirements. Furthermore, the classifier outputs from the pooled CNN features and the handcrafted features are fused in the evaluation of a combination of the two types of features. [permissions required!!]

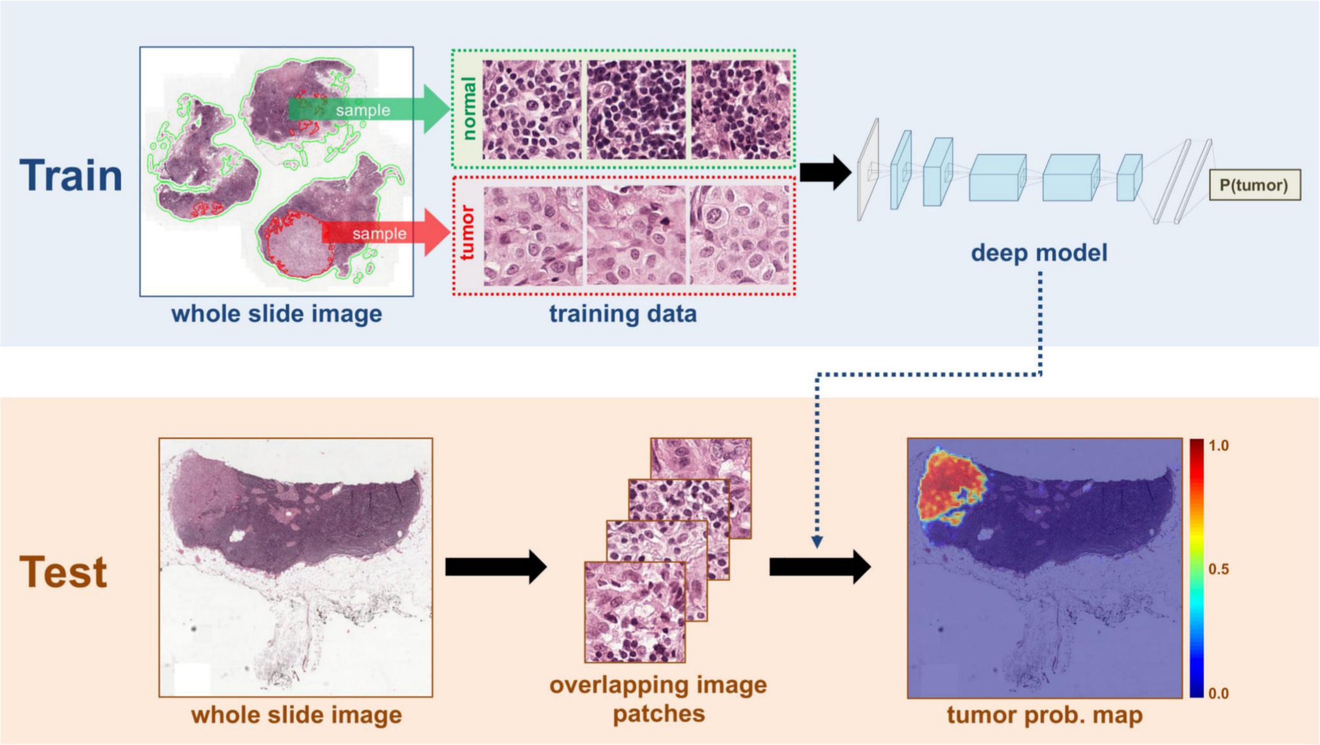


Fig. 7.

One proposed framework for cancer metastases detection by Wang et al. [61] who won the first prize in Camelyon16 cancer detection competition [9]. The model was based on deep CNNs, GoogLeNet of 27 layers.