BJR|Open https://doi.org/10.1259/bjro.20210072

Received: 12 November 2021 | Revised: 01 June 2022 | Accepted: 06 June 2022 | Published online: 22 June 2022

Cite this article as:
Mazaheri Y, Thakur SB, Bitencourt AGV, Lo Gullo R, Hötker AM, Bates DDB, et al. Evaluation of cancer outcome assessment using MRI: A review of deep-learning methods. *BJR Open* (2022) 10.1259/bjro.20210072.

REVIEW ARTICLE

Evaluation of cancer outcome assessment using MRI: A review of deep-learning methods

[1,2]YOUSEF MAZAHERI, PhD, [1,2]SUNITHA B. THAKUR, [3,4]ALMIR GV BITENCOURT, [2]ROBERTO LO GULLO, [5]ANDREAS M. HÖTKER, [2]DAVID D B BATES and [2]OGUZ AKIN

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, United States
[2]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, United States
[3]Department of Imaging, A.C.Camargo Cancer Center, São Paulo, Brazil
[4]Dasa, Sao Paulo, SP, Brazil
[5]Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, Zurich, Switzerland

Address correspondence to: Dr Yousef Mazaheri
E-mail: *mazahery@mskcc.org*

ABSTRACT

Accurate evaluation of tumor response to treatment is critical to allow personalized treatment regimens according to the predicted response and to support clinical trials investigating new therapeutic agents by providing them with an accurate response indicator. Recent advances in medical imaging, computer hardware, and machine-learning algorithms have resulted in the increased use of these tools in the field of medicine as a whole and specifically in cancer imaging for detection and characterization of malignant lesions, prognosis, and assessment of treatment response. Among the currently available imaging techniques, magnetic resonance imaging (MRI) plays an important role in the evaluation of treatment assessment of many cancers, given its superior soft-tissue contrast and its ability to allow multiplanar imaging and functional evaluation. In recent years, deep learning (DL) has become an active area of research, paving the way for computer-assisted clinical and radiological decision support. DL can uncover associations between imaging features that cannot be visually identified by the naked eye and pertinent clinical outcomes. The aim of this review is to highlight the use of DL in the evaluation of tumor response assessed on MRI. In this review, we will first provide an overview of common DL architectures used in medical imaging research in general. Then, we will review the studies to date that have applied DL to magnetic resonance imaging for the task of treatment response assessment. Finally, we will discuss the challenges and opportunities of using DL within the clinical workflow.

INTRODUCTION

In recent years, there has been a dramatic increase in research studies applying artificial intelligence (AI) approaches to a wide range of decision-making problems. In cancer imaging research, deep learning (DL) has shown promising performance in a wide range of tasks, including cancer screening, tumor characterization (which includes subtype classification), treatment response, and survival outcome assessment.

With respect to treatment response assessment, imaging has played an important role in this task for decades. The first attempt to establish standardized criteria for image-based treatment response assessment dates back to the 1970s, when the International Union Against Cancer and the World Health Organization (WHO) developed an evaluation scheme to classify treatment response of solid tumors based on bidimensional measurements of tumor size in the axial plane on imaging studies.[1] Since then, four categories have been used to classify image-based treatment response: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). In 2000,[2] the Response Evaluation Criteria in Solid Tumors (RECIST) was published, providing guidance on treatment response classification based mainly on changes in tumor size. These criteria recommended unidimensional instead of bidimensional measurements to quantify tumor burden and have since become the most used criteria for estimating solid tumor burden and determining clinical treatment response. However, RECIST is known to have certain intractable limitations, particularly when it comes to precision medicine approaches to cancer. For example, RECIST criteria are limited by large inter- and intraobserver variations, especially in tumors with irregular and complex shapes. Furthermore, when tumors are treated with targeted chemotherapy or immunotherapy, assessing tumor response based on changes in tumor size is likely inadequate. In such cases, tumor response would

be better reflected by morphologic, functional, and metabolic changes, such as residual cancer cell amount, extent of necrosis and fibrosis, or cystic degeneration. An alternative response criterion—the Choi criteria—was proposed for computed tomographic (CT) imaging and incorporates measurements of both tumor size and density.[3] Within the context of the Choi criteria, a patient is regarded as responding if CT images show a 10% reduction in tumor size or a 15% reduction in CT attenuation. Based on these criteria, Choi et al showed that response among patients with metastatic gastrointestinal stromal tumor showed significantly longer progression-free interval compared with nonresponses.

In 2009, RECIST 1.1,[4] a revised version of RECIST, was published. Additional response criteria have also been developed, including modified RECIST (mRECIST) for the evaluation of hepatocellular carcinoma (HCC) response to targeted therapy,[5] immune-related response criteria (irRC) for the assessment of response to immunotherapy,[6] and immune-related RECIST (irRECIST), which combines characteristics of irRC and RECIST.[7] The irRC are based on bidimensional measurements and new lesions are incorporated for the measurement of total measurable tumor burden.[6] irRECIST, reported by Nishino et al, requires only one-dimensional measurement and confirmation by two consecutive observations to judge complete response (CR), partial response (PR), or progressive disease (PD).[7]

DL methods have evolved since basic foundations were introduced in the 1940s, with methodologies advancing tremendously over the past decade. Qualitative and quantitative measurements on magnetic resonance imaging (MRI) data offer a promising technique for the assessment of treatment and survival outcomes. MRI and the numerous imaging sequences often acquired yield valuable information that can potentially serve as biomarkers for the assessment of treatment and survival outcomes. Suitable applications for DL methodology to MRI have the potential to enhance the prognosis and mortality assessments of cancer. DL methods promise to explore the complex relationship between MRI data and cancer outcomes.

The aim of this review is to highlight the use of DL in the evaluation of tumor response and survival outcome assessed on MRI. In this review, we will first provide an overview of common DL architectures used in general medical imaging research. Then, we will review published studies that have applied DL to MRI for the task of treatment response and survival outcome assessment in different types of cancer. Finally, we will discuss the challenges and opportunities of using DL within the clinical workflow.

Deep-learning architectures commonly used in medical imaging

DL is a machine-learning subset where features are learned directly from raw data rather than advanced specification.[8] DL techniques can be divided into unsupervised and supervised DL techniques. Supervised DL pre-specifies desired outputs with associated inputs while training the neural network algorithm. Accordingly, the algorithm is trained to learn relationships or transformations that allow it to predict expected outputs when given new inputs. By contrast, unsupervised DL finds relationships between variables in a given dataset without any labels. The algorithm discerns unlabeled data autonomously by relying on the extraction of inherent dominant features and patterns (Figure 1).

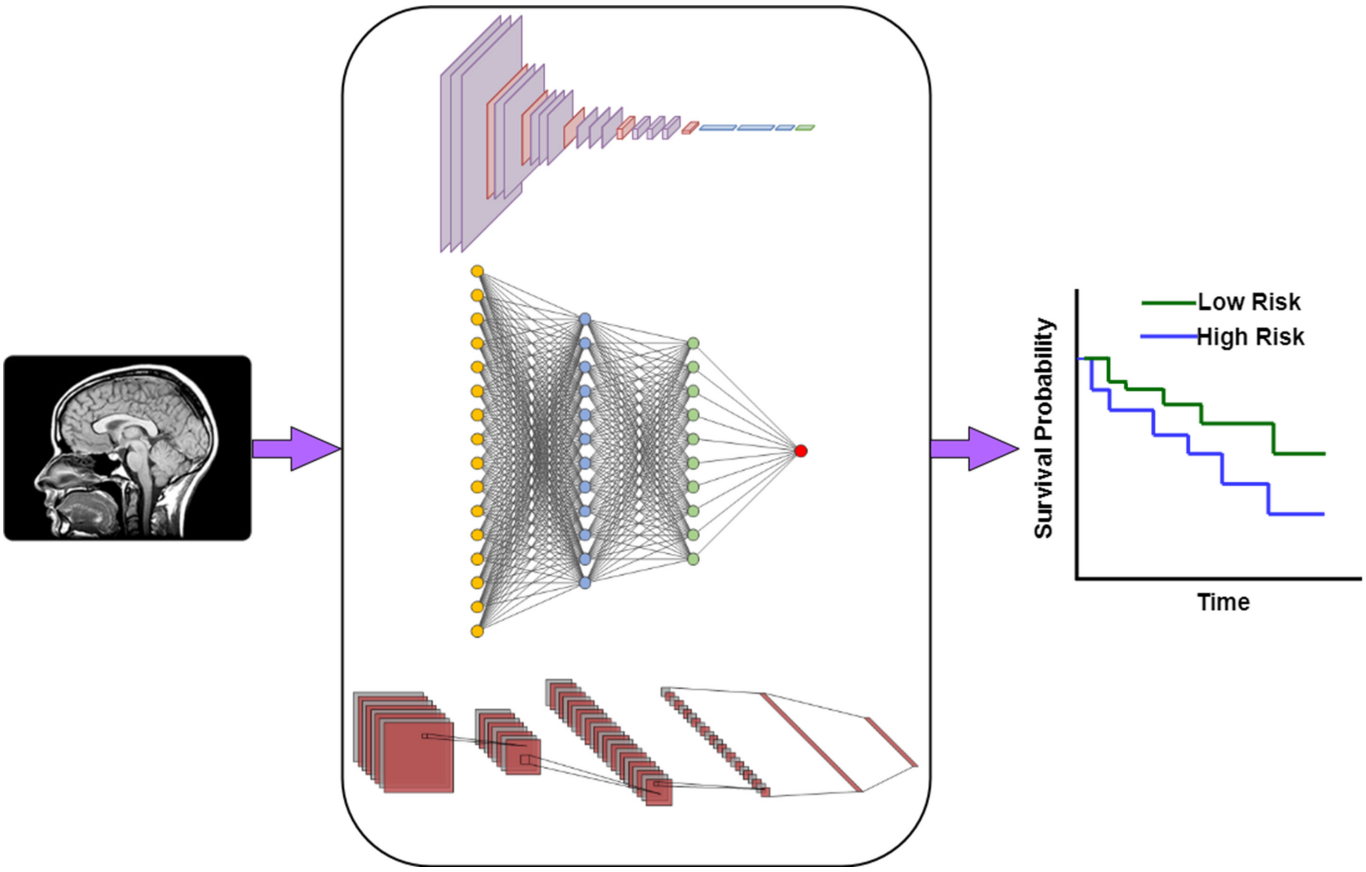
DL networks are characterized by hierarchical architectures consisting of multiple layers of non-linear information, whereby features in upper layers represent combinations of simpler features below. Neural networks use backpropagation as a learning algorithm to compute the gradient of the loss function for each weight in the network model. Subsequently, the gradient is used by an optimization algorithm to update model weights. In addition to calculating the gradient of a loss function with respect to variables of a model, a neural network model requires hyperparameter optimization or tuning of the learning algorithm. This task involves choosing a set of hyperparameters for a learning algorithm that yields an optimal model, or a model which minimizes a predefined loss function. Finally, cross-validation is often used to estimate the generalized performance of the model.

We will next review DL architectures commonly used in general medical imaging research. But it is important to first consider artificial neural networks (ANNs), the backbone of deep neural networks (DNNs). ANNs are inspired by the structure and function of the human brain. ANNs can be developed based on supervised, unsupervised, or semi-supervised learning. An ANN is composed of layers of connected nodes (also called artificial neurons), configured at multiple layers (depth) and in the order of hundreds to millions (Figure 2). The objective of this configuration is to maximize the correct output as compared with a reference value. This is accomplished on each forward propagation by calculating the error and adjusting the weightings on each node. The process is repeated at each iteration (epoch), until a more accurate solution is converged.

One of the first, simplest, and most widely used ANN in practical application is the feedforward neural network (FFNN). In FFNN, information flow is always in a single and forward direction from the input nodes, through any hidden nodes, and up to the output nodes. The objective is to learn the relationship between independent variables that are network inputs and the dependent variables that are assigned as network outputs.

The construction of an ANN involves training the network on a large dataset and subsequently validating the inferences of the network on a test set. Training and optimization are achieved through a loss index measuring algorithm-associated errors. Regularization refers to strategies employed to reduce the error of the test set at the expense of increasing training error. To tune an ANN, a loss index consisting of the sum of the error and regularization term is measured, and an optimization algorithm applied to adjust the weights and bias by backpropagating the errors from the output layers in the direction of the input layers. This iterative process is repeated until the loss index is minimized or until a predetermined value is reached. A key difference between ANN and DNNs is that DNNs entail a greater number

Figure 1. A general framework (based on deep learning algorithms) for the processing of images by classifying high and low risk of survival, thus assessing probability of treatment response. (A) Patient MRI images are input into a deep learning model for the purpose of training the model; (B) A deep-learning system developed and trained to characterize outcome assessment, such as survival probability; (C) The outcome of the deep-learning model is used to predict cancer outcome.



of hidden neurons, more layers, and innovative training paradigms to process larger amounts of data.

Convolutional neural network

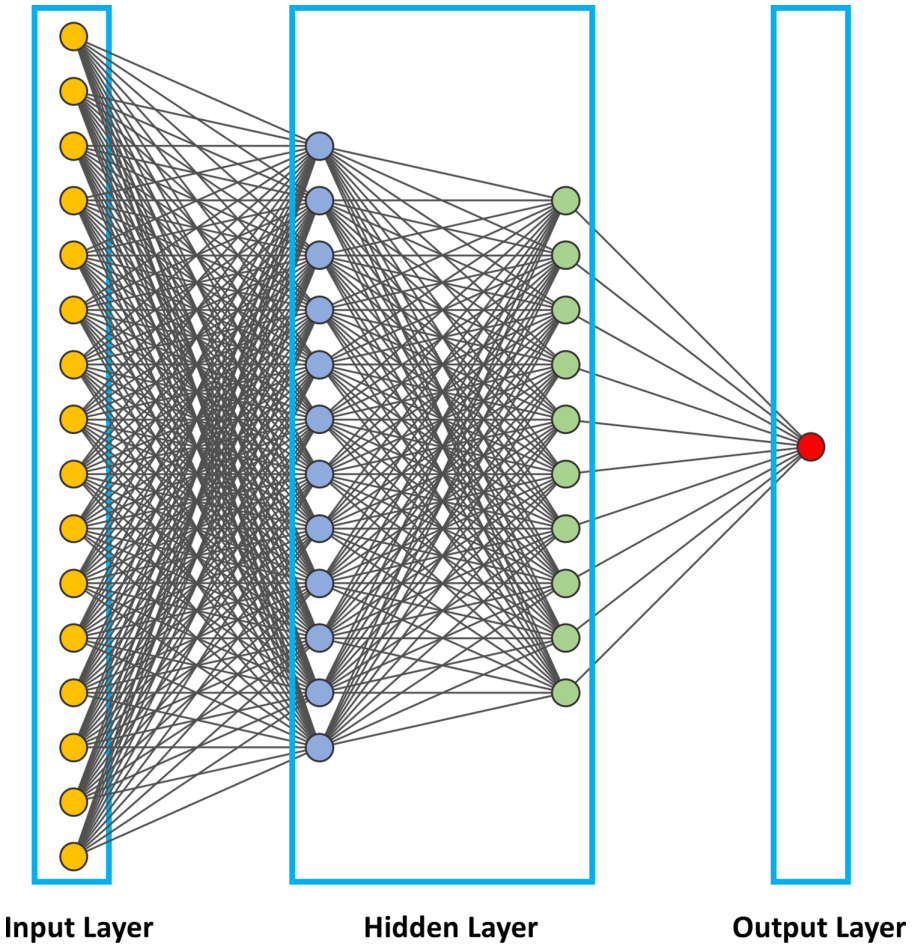
In 2012, Krizhevsky et al developed a convolutional neural network (ConvNet/CNN) that markedly improved image classification[9] (Figure 3). CNNs are the most widely used DL architectures for medical image analysis, having been developed for tasks including image recognition, image analysis, image segmentation, video analysis, and natural language processing. The best-known CNN architectures developed to date are ZFNet,[10] VGGNet,[11] GoogLeNet,[12] AlexNet,[13] and ResNet.[14]

CNNs are multilayered neural networks with three layer types: convolutional layers, pooling layers, and fully connected layers. CNNs are designed to extract features that capture the spatial and temporal patterns of the input images. Using convolutional and pooling layers, CNNs mimic the mathematical operations of convolution and pooling. The convolution layer constitutes the essential feature of CNNs and refers to the networks' operation based on a set of learnable filters to merge the input values and filter values onto the feature map. Pooling layers are used to reduce the dimensions of feature maps. The standard CNN employs a rectified linear unit (ReLU) as an activation function and a supplemental step to convolution. Another activation function which is very popular for neural networks is the sigmoid activation function, also called the logistic function. ReLU will give an output of zero for negative inputs but otherwise preserve the input. ReLU is the most used activation function in DL models due to its computational simplicity, representational sparsity, and linearity. As compared to the sigmoid activation function, ReLU are easier to train. The representational sparsity feature implies that the ReLU function, unlike the tanh and sigmoid activation functions, is capable of outputting a true zero value.

Further along in the network architecture, the pooling layers work to downsample the features in the convolved feature map, typically using max pooling, so that dominant features that are rotationally and positionally invariant are extracted. Finally, fully connected layers at network's end generate the required class prediction by taking the flattened matrix from the pooling layers as input.

The main advantage of CNNs is that it captures important image features (through a backpropagation algorithm) without any human supervision. Compared with alternative network designs like FFNN, CNNs capture the spatial dependencies in an image, hence better capturing its composition. The primary

Figure 2. Illustration of an artificial neural networks (ANNs), the backbone of deep neural networks (DNNs). In this figure, we show a fully connected neural network where all the nodes, or neurons, in one layer are connected to the neurons in the next layer. When the input increases, fully connected networks tend to be computationally expensive, resulting in poor scalability.



disadvantages of CNNs are that they require large training data, and that they do not encode the position and orientation of the object.

Many variants of the CNN architecture have been developed. For example, U-Net is a fully convolutional network developed by Ronneberger et al. in 2015 for medical image segmentation.[15]

Figure 3. An illustration of a simple convolutional neural network including convolutional, pooling, and fully connected layers. The two-dimensional input data undergoe multiple rounds of convolution and subsample layers. Feature extraction by filters are learned through back projection. The pooling operations, including max or mean, in a region are used to reduce the number of pixels in each layer of the network. Each operation increasingly extracts higher order discriminative features. Ultimately, the output layer is a class probability based on these higher order features.

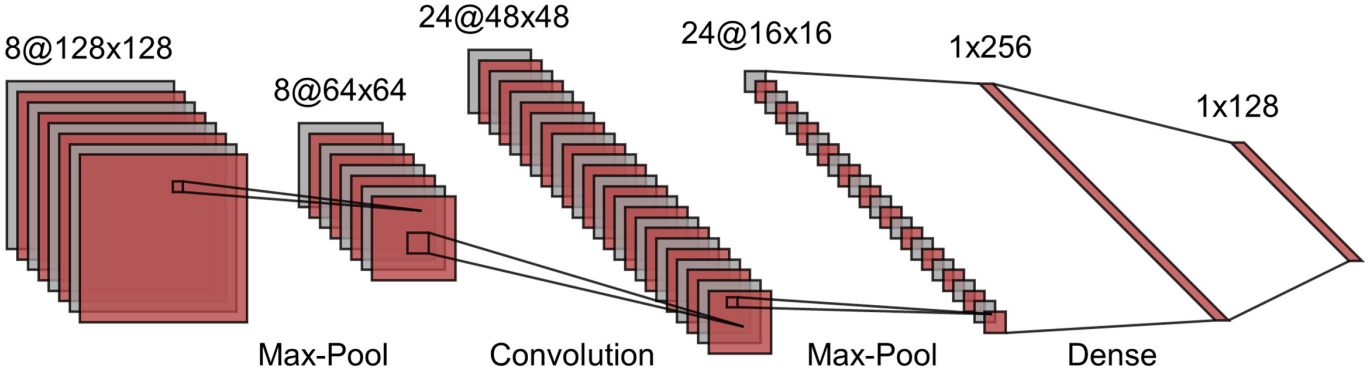
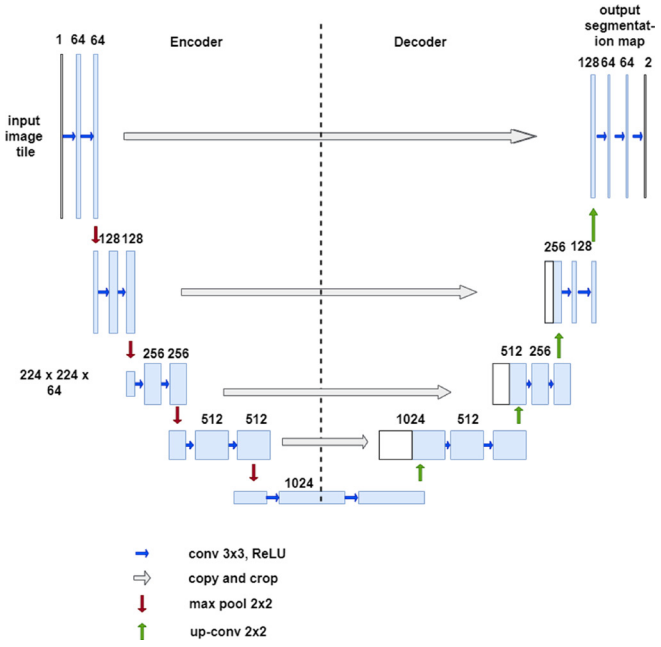


Figure 4. The U-net network structure has a deep-learning encoder-decoder architecture. The CNN is termed "U-net" due to the u-shaped structure. The network consists of encoder layers where there is first downsampling in the image size followed by upsampling in the expansive or decoder layer.



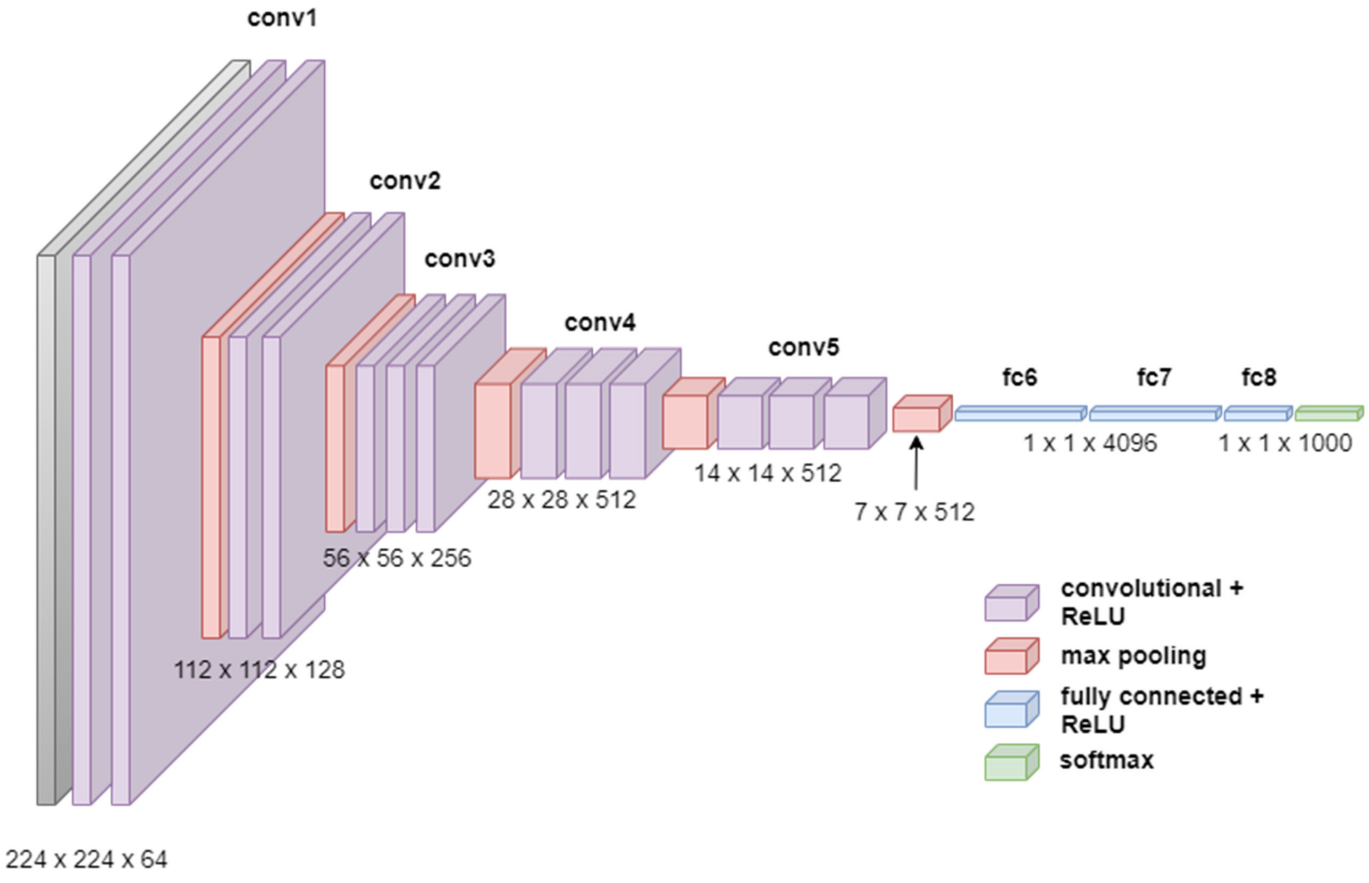
U-Net consists of a contracting path (also known as the encoder path) that downsamples the image into a feature map, followed by an expansion path (*i.e.,* decoder path) that upsamples the feature map to the target such as the output segmentation map. During downsampling, feature information is extracted while spatial information is reduced. During upsampling, feature and spatial information are combined through a sequence of up-convolutions and concatenations, generating high-resolution features. The resultant neural network yields more precise segmentations with fewer training images. The workflow for a U-Net network is illustrated in Figure 4.

Another CNN architecture is the VGG16, which was used to win ILSVR competition in 2014.[16] The VGG16 network improves upon AlexNet by replacing large kernel-sized filters with multiple 3 × 3 kernel-sized filters. The network applies the same kernel size of 3 × 3 filter throughout the feature extraction part and always uses the same padding and max pool layer of 2 × 2 filter of stride 2. This arrangement of convolution and max pool layers is consistently followed throughout the network (Figure 5).

Transfer learning

Often, training an entire network from scratch is impractical since this requires large training datasets. Should a large training dataset be unavailable, transfer learning can be employed. In transfer learning, information obtained through a pre-trained

Figure 5. The VGG-16 architecture. The VGG16 consists of 13 convolutional layers, five max-pooling layers, and three fully connected layers. Consequently, the number of tunable parameters is 16 (13 convolutional layers and three fully connected layers).



model using a large dataset (such as ImageNet) is transferred to a smaller dataset. For CNNs, one transfer learning strategy is to modify training in the convolutional layers, such that training occurs only during the last few convolutional layers to perform a prediction. This is based on the premise that the early convolutional layers extract low-level features that can be generalized across images, whereas the later convolutional layers are geared toward identifying high-level features within an image. Low-level features are local and include features such as edges and blobs. High-level features include objects, their states, and events in images, which are extracted using machine-learning techniques. Further strategies include fine-tuning all layers of the CNN by adjusting the weights of the pre-trained network or utilizing a pre-trained model that includes the CNN checkpoints and fine-tuning the network weights. Checkpoints allow pre-trained models to be used for inference without retraining. Alternatively, checkpoints allow model training to resume in case it was interrupted or for the purpose of model fine-tuning.

Recurrent Neural Network

Recurrent neural networks (RNNs) are employed to process sequential or time series data, whereby the nodes in RNNs are connected along the data sequence. RNNs are derived from transfer learning FFNNs. While FFNNs allow signals to travel in one direction from input to output only, RNNs allow information to cycle in loops allowing dependencies between data points. Consequently, RNNs possess internal state (memory), where they retain information about past inputs based on its weights and on input data, allowing them to harness past information to predict a later event (Figure 6). RNNs are commonly used for speech and language tasks, such as speech recognition and natural language processing. Of note, long-short-term memory (LSTM) networks are a subtype of RNN that extend the memory of RNNs.

Autoencoder and deep autoencoder

In the 1980s, Geoffrey Hinton designed the autoencoder (AE) to solve unsupervised learning problems. Autoencoders are a type of feedforward neural network for learning representation, in which the network receives the input and deconstructs it into an internal latent representation or code before reconstructing the input as closely to the original image as possible. Autoencoders consist of an input, an output, and multiple hidden layers (Figure 7). The training of the network can be unsupervised, with the goal of reconstruction error minimization: a measure of the differences between the original input and the reconstruction. Types of autoencoder include the multilayer autoencoder; the convolutional autoencoder, intended to reduce image noise or detect video anomalies; and the regularized autoencoder, intended to learn representations for subsequent classification tasks. Regular autoencoders have one layer between the input and output layer, whereas deep autoencoders have multiple hidden layers.

Generative adversarial network

Generative adversarial networks (GANs) are a branch of DL that I. Goodfellow introduced in 2014.[17] GANs have been successfully applied to unsupervised image translation, domain adaptation, image in-painting, and semi-supervised classification. They have also been studied for medical image synthesis.[18,19]

GANs entail simultaneous training of two adversarial models. The GAN architecture is composed of two networks, a Generator (G) and a Discriminator (D), which are trained in competition based on the two-person zero-sum game in game theory (one's win is another's loss). The Generator is responsible for generating data, and the Discriminator for estimating the probability that an image was drawn from the training data (is real) or produced

Figure 6. Illustration for the architecture of recurrent neural network (RNN). RNNs are a class of neural network commonly used for text and sequence data. They allow previous outputs to be used outputs to be used as inputs while having hidden states. An important class of RNNs are long-short-term memory (LSTM) which have feedback connections are often used for time series analysis.

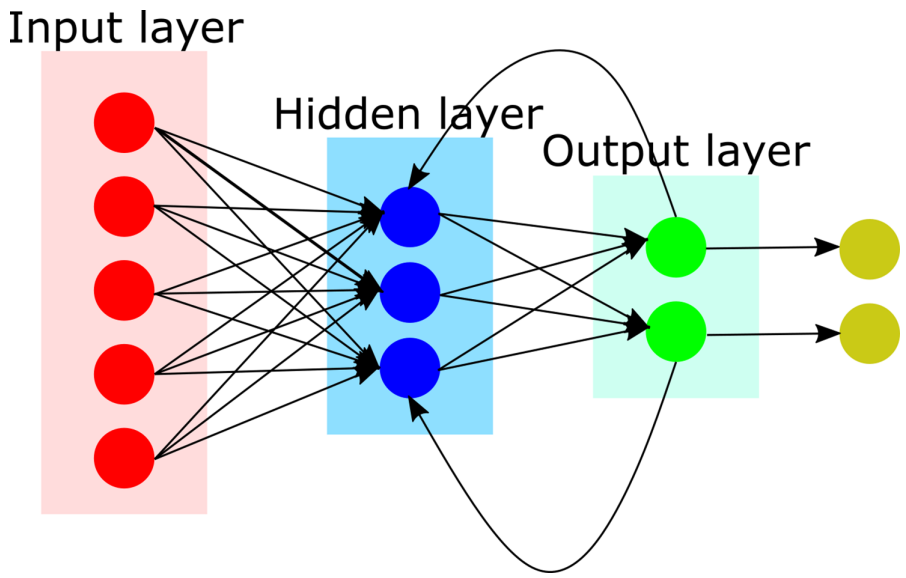
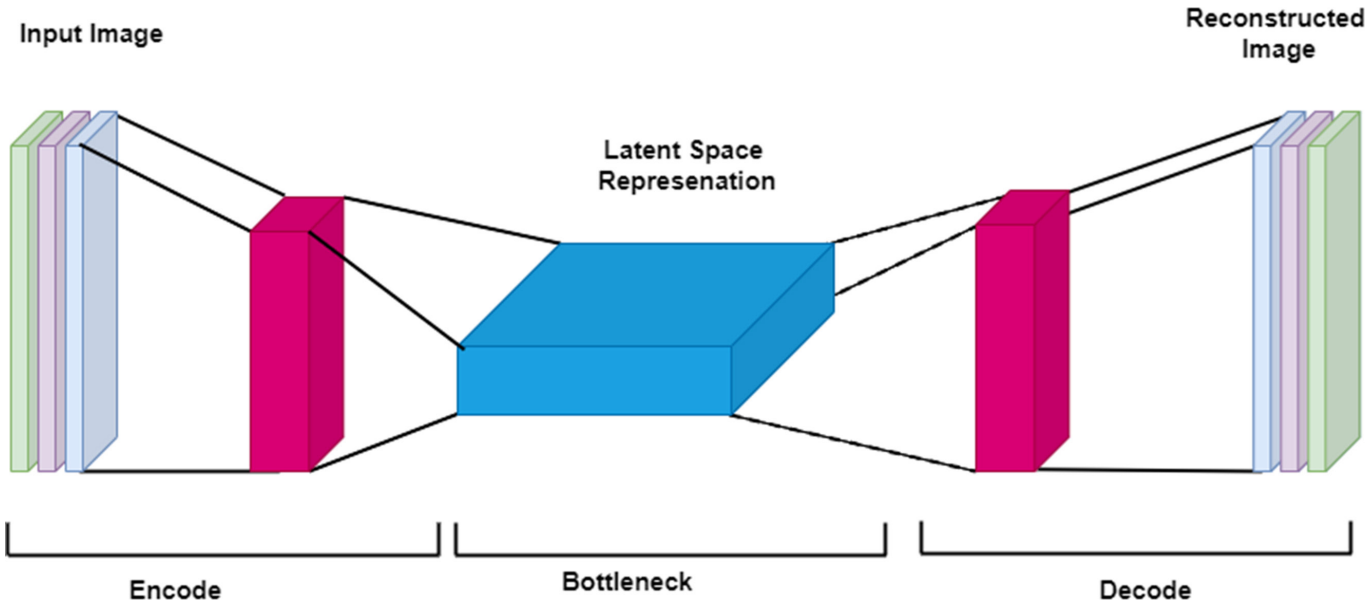


Figure 7. Illustration of a basic autoencoder. An autoencoder is an unsupervised learning model assigned the task of transforming the input image into a latent or compressed representation by minimizing the reconstruction errors between input and reconstructed images of the network. An autoencoder performs two tasks. It first encodes an image, and subsequently it decodes it. Encoding an image in this context means that the autoencoder generates a compressed representation of the original image. Conversely, the decoder takes the output from the bottle neck (latent space representation) and attempts to recreate the input image. For the autoencoder to reconstruct an image, it will need to learn some latent representation of the image. Latent representation refers to a set of compressed features of the image which are learned by the network through an iterative process of training, and which are subsequently used to reconstruct the desired image.



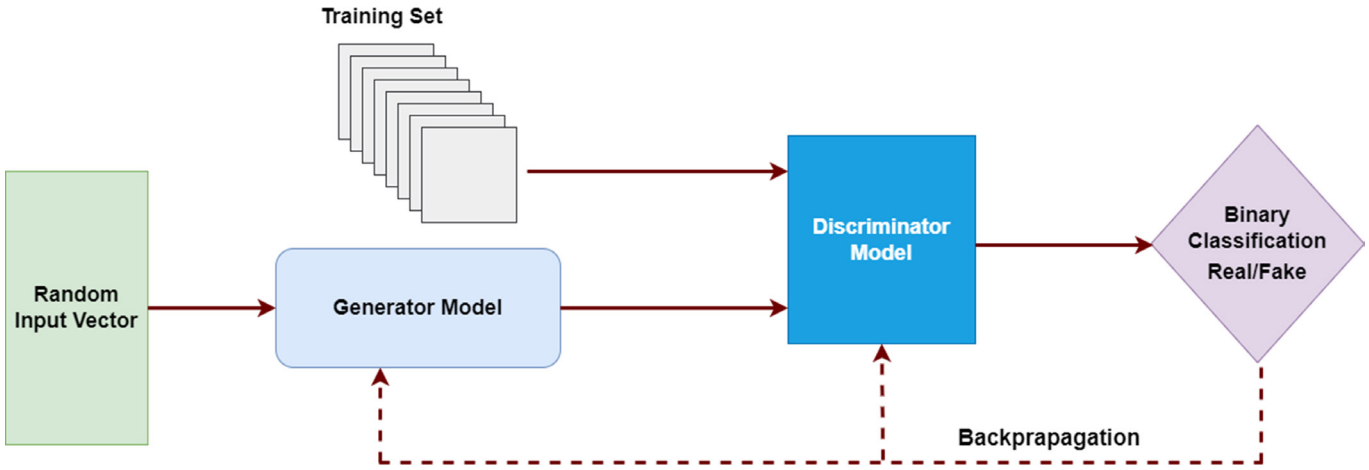
by the generator (is fake). The objective of these models is to learn the training data distribution and subsequently generate realistic data samples indistinguishable from the input data. They perform this task by minimizing the loss function through a second adversarial network. During training, the Generator increasingly improves image generation until the Discriminator can no longer distinguish between the real and fake data (Figure 8).

TREATMENT RESPONSE APPLICATIONS

Brain cancer

Nie et al[20] used multimodal images of 68 patients with high-grade gliomas to develop a 3D DL framework to predict the survival time (long *vs* short) of patients with high-grade glioma. An independent dataset of 25 patients was used to validate the model. Their approach consisted of a multichannel architecture of 3D CNNs to identify and extract high-level features from T_1-weighted MRI, resting-state functional MRI, and diffusion tensor MRI. Using extracted features as well as demographic and tumor-related features like gender, age at diagnosis, tumor

Figure 8. In Generative adversarial networks (GANs) consists of two models: the discriminator and the generator. GANs learn through deriving backpropagation signals through a competitive process involving a pair of networks.



location, tumor size, and WHO grade, a support vector machine was used to predict overall survival time. The combination of deeply learned as well as demographic and tumor-related features resulted in a classification accuracy of 90.66% with threefold cross-validation, and 90.46% with 10-fold cross-validation.

In another study predicting overall survival in patients with glioblastoma multiforme, a proposed radiomics model used deep features extracted from CNNs based on transfer learning and handcrafted features based on radiomics analysis.[21] The study consisted of 75 patients for training and an independent data set of 37 patients. Both handcrafted features (N = 1403) and deep features (N = 98304) were extracted from the preoperative multimodality MR images. After feature selection, a model was generated and a radiomics nomogram was constructed by combining signature and clinical risk factors. The radiomics signature outperformed traditional clinical risk factors such as age and the Karnofsky Performance Score for the prediction of overall survival (C-index = 0.710). The model combining the radiomics signature and traditional clinical risk factors further improved prediction performance (C-index = 0.739).

In a study by Kickingereder et al, a DL model using ANNs was developed for the quantitative assessment of tumor response.[22] Three datasets were used to train and test the model: Heidelberg training dataset (455 patients with brain tumors), Heidelberg test dataset (longitudinal dataset of 40 patients with data from 239 MRI scans), and EORTC-26101 test dataset (MRI scans from 532 patients obtained from 34 institutions). Using the Heidelberg training dataset, an ANN was developed for automated volumetric segmentation of contrast-enhanced tumors and non-enhancing T2-signal abnormalities on MRI. This ANN was derived from the authors' previously developed ANN, itself based on a U-Net architecture.[15] The newly developed ANN was asked to predict segmentation masks of contrast-enhanced tumors and non-enhancing abnormalities via an ANN ensemble model (five ANN models obtained from cross-validation of the Heidelberg training dataset) on the Heidelberg and EORTC-26101 test datasets. The tumor segmentation masks generated by the ANNs were shown to be highly accurate in comparison with a reference standard selected as the ground truth segmentation masks generated by a radiologist (median DICE coefficient = 0.89 for contrast-enhanced tumors and 0.93 for non-enhanced abnormalities in the Heidelberg test dataset; 0.91 for contrast-enhanced tumors and 0.94 for non-enhanced abnormalities in the EORTC-26101 test dataset). Moreover, the time to progression determined using ANN-based assessment of tumor response outperformed central RANO assessment for the prediction of overall survival in the EORTC-26101 test dataset (hazard ratios = 2.59 vs. 2.07; $p < 0.001$).

A study by Han et al combined hand-crafted radiomics and deep features generated by a pretrained CNN[23] from gadolinium-based contrast-enhanced T_1-weighted images of patients with high-grade gliomas from both their institution and from The Cancer Genome Atlas. Feature selection followed by Elastic Net-Cox modeling were performed to predict long- and short-term survivor groups. The model classified patients with high-grade gliomas into long- and short-term survivors (the log-rank test p value < 0.001 in patients from their institution, p = 0.014 in patients from The Cancer Genome Atlas, and p = 0.035 in all patients from both cohorts).

Breast cancer

In 2012, Hylton et al[24] reported that MRI outperformed clinical assessment in predicting pathologic complete response (pCR) to neoadjuvant chemotherapy (NAC), using MR images from 216 patients enrolled in the ACRIN 6657/1-SPY1 TRIAL. Tumor measurements on MRI were superior to clinical examination in predicting pCR to NAC at all timepoints. Particularly, tumor volume change at the second MRI examination obtained after one cycle of anthracycline-based treatment showed greatest predictive ability. This work motivated additional studies on the use of MRI to predict pCR, including those with DL tools.

Huynh et al[25] compared CNN-extracted features from DCE-MR images at different contrast timepoints to determine which timepoint would result in the best classifier for predicting response to NAC, finding that CNN-extracted features based on pre-contrast time points yielded the best classifier.

Several studies investigated deep learning applied to MRI to evaluate response to NAC, using publicly available MR images from the multiinstitutional I-SPY1 TRIAL. Ravichandran et al[26] applied a CNN to pre-treatment dynamic contrast-enhanced MR images from 166 patients with breast tumors of at least 3 cm in size, who received DCE-MRI imaging prior to treatment, had at least two post-contrast phases of DCE-MRI, and had undergone post-NAC surgery. The classifier to predict pCR based on CNN-extracted features from both pre- and post-contrast images achieved an accuracy of 82% in the testing set. The inclusion of HER2 status to the classifier improved the accuracy to 85%. Another study using MR images from the I-SPY TRIAL, Liu et al,[27] developed a CNN algorithm to predict pCR *vs* no-pCR response to NAC based on post-contrast images only, which yielded an accuracy of 72.5%. Due to the high computational burden associated with training customized CNNs, Comes et al[28] investigated a transfer learning approach, using the pre-trained CNN AlexNET (previously trained to extract both low-level features such as edge and dots and high-level features such as shapes and objects from a raw image), to evaluate the early efficacy of NAC before the completion of therapy. When optimized features extracted from pre- and early treatment exams were combined with clinical features such as ER, PgR, HER2 and molecular subtypes, the classifier achieved an accuracy of 91.4% on the subset of patients used for fine-tuning, and 92.3% on the independent database.

Single-institution studies have also shown that deep learning is promising to predict response to NAC. Ha et al[29] investigated a CNN to predict NAC response based on pre-treatment breast MRI for 141 patients with locally advanced breast cancer who had pre-treatment MRI followed by adriamycin/taxane-based NAC and surgical resection. Patients were divided into three groups based on NAC response: complete, partial, and no response/progression. Tumors underwent 3D segmentation on

the first post-contrast image. The CNN architecture consisted of ten convolutional layers, four max-pooling layers, and 50% dropout after a fully connected layer. The overall mean accuracy of the CNN was 88% (95% CI,±0.6%). In another study, the same group of authors[30] developed a CNN algorithm to predict post-NAC pCR of the axilla using breast MRI performed before NAC. The proposed CNN algorithm achieved an overall accuracy of 83%. El Adoui et al[31] evaluated a group of 42 breast cancer patients who had DCE-MR imaging before and after the first cycle of chemotherapy and developed a CNN that achieved an area under the receiver operating characteristic curve (AUC) of 0.91 and accuracy of 88% using both the pre- and post-treatment examinations without segmentation (multi-input CNN). Using single-input CNN of pre-treatment examinations only or post-treatment examinations only with or without segmentation achieved an AUC of 0.69–0.79 and accuracy of 68–80%.

In applying DL using both positron emission tomography/magnetic resonance imaging (PET/MRI) scans obtained before and after the first cycle of NAC in patients with advanced breast cancer, Choi et al[32] generated CNNs based on AlexNet that improved the classification of patients into pCR and non-pCR groups compared with the majority of conventional PET and MR imaging parameters.

Colorectal cancer metastases

Colorectal liver metastases (CRLM) are the third leading cause of cancer-related death in the US.[33] The assessment of treatment response at preoperative chemotherapy is crucial to inform therapeutic adjustments that maximize benefit. Zhu et al[34] applied DL to MR images to predict CRLM response to chemotherapy. The study included 101 patients in the training cohort, 54 patients in the testing cohort, and an additional 25 patients as an external validation cohort. The DL architecture was designed to import four inputs: pre- and post-treatment T_2-weighted image, and pre- and post-treatment apparent diffusion coefficient (ADC) images. The network was designed to extract features from the input data to distinguish pathology tumor regression grade (TRG) between the response and non-response group, as well as to distinguish survival outcomes after hepatectomy. Three models were developed: Model A (based on pre- and post-treatment MRI), Model B (based on pre-treatment MRI only), and Model C (based on post-treatment MRI only). The results of the DL algorithm were compared with RECIST to predict tumor response and determine survival outcome. The accuracy of Model A (accuracy of 87.5%) was significantly higher as compared with Models B and C (accuracy of 79.7 and 85.9%, respectively) and RECIST (accuracy of 57.8%). The p-values for comparison were as follows: 0.04 for comparison of Model A *vs* Model B, 0.04 for comparison of Model A *vs* Model C, and 0.03 for comparison of Model A *vs* RECIST.

Rectal cancer

The current standard-of-care treatment in patients with locally advanced rectal cancer (LARC) is neoadjuvant chemoradiation therapy (CRT) followed by total mesorectal excision (TME). Patients with pCR may be spared resection if followed with biopsy and MRI.[35] Assessment of response to chemoradiotherapy can impact treatment decision-making for these patients. The availability of additional treatment options or non-operative approaches is a motivating factor for the assessment of treatment response. MRI plays an important role in treatment response assessment after chemoradiotherapy. However, distinguishing between therapy-induced scarring and residual viable tumor on T_2-weighted sequences remains difficult.[36]

Recently, radiomics and DL methods have been used to predict pCR in patients with LARC. In a study, Shi et al[37] extracted radiomic features from pre-treatment MRI T1- and T_2-weighted images, axial DWI, and T_1-weighted DCE-MRI. They used a three-layer ANN to select parameters and build diagnostic radiomics models. Additionally, a CNN was developed with the image input a tight bounding box covering the tumor region of interest (ROI). Results showed that CNN based on pre-treatment and mid-radiation therapy MRI achieved an AUC of 0.83 for predicting pCR *vs* non-pCR, whereas the model combining ROI and radiomic features achieved an AUC of 0.80 based on pre-treatment images, 0.82 for mid-radiation therapy, and 0.86 for both pre-treatment and mid-radiation therapy images.

Radiomics methods provide a valuable mechanism for extraction of quantitative features from medical images. These can then be correlated with various biological features and clinical endpoints. Delta-radiomics is an emerging approach and an extension of radiomics based on the analysis of variations of radiomics features at different acquisition time points.[38] The points are typically pre- and post-treatment, with the objective of predicting response.[39] Delta-radiomic features have shown promise in predicting the response of colorectal liver cancer[40] and metastatic renal cell cancer[41] to chemotherapies, as well as the analysis of CT images to determine the treatment response of non-small cell lung cancer to radiation therapy.[42] One study, evaluating the ability of delta-radiomics to predict overall survival of patients with recurrent malignant gliomas who were treated with concurrent stereotactic radiosurgery and bevacizumab, indicated that delta-radiomic features potentially provided better treatment assessment than features extracted from a single time point.[43] While delta-radiomics is at an early stage, it has shown promising results in studies focusing on temporal changes of radiomic features in treatment response assessment. Delta-radiomics provides high-dimensional data, making machine-learning tools like DL suitable for feature analysis. An in-depth review providing detailed information on delta-radiomics is available elsewhere.[38]

In another study, Zhang et al[44] developed DL models to predict response based on diffusion kurtosis and T_2-weighted MRI from 383 patients with LARC who underwent baseline MRI prior to preoperative chemotherapy. The DL network architecture consisted of a multipath CNN with eight inputs comprising T_2-weighted imaging and diffusion kurtosis imaging pre- and post-treatment.[45] Three DL models were considered. The first was for pCR prediction and second for TRG (0 + 1) and TRG (2 + 3) classification. The third model was for T-downstage and non-T-downstage classification. The first model for pCR prediction achieved an AUC of 0.99, significantly better than the evaluation

by two radiologists (AUC of 0.66 for rater 1 and 0.72 for rater 2) ($p < 0.001$). The second and third models had AUC of 0.70 and 0.79, respectively. The DL model also served to reduce radiologist error rate; when radiologists were assisted by the DL model in predicting pCR, their AUC significantly improved to 0.82 for rater 1 and 0.83 for rater 2 ($p = 0.002$ and 0.01, respectively). However, the diagnostic performance of the DL models for classifying TRG and T stage downgrading did not exceed the two radiologist evaluations.

As compared to radiomics analysis, applying DL methodology to evaluate tumor response to treatment using MRI offer several key advantages. First, DL approaches typically do not require precise tumor delineation. Second, they often outperform radiomic feature analysis. Third, they automatically learn and hierarchically organize task-adaptive image features. The extracted features might not be visually identifiable but reflect associations between the classifier and images, providing tremendous potential in clinical decision-making.

Challenges and opportunities of Deep-learning

DL has both numerous advantages over traditional machine learning and tremendous potential to transform MRI-based evaluation of tumor treatment response. CNN, a popular DL architecture, allows the network to independently learn by performing prediction tasks, such as identification of useful regions or extraction of salient features from those regions, without the need for human intervention.[46,47] CNN provides a general-purpose learning procedure for an end-to-end image analysis workflow. CNNs learn specific patterns of their given task from the images themselves instead of relying on preprocessing steps, 'handcrafted' features, or subsequent model building. The objective is for the network to automatically extract relevant features from images, resulting in easy clinical application. However, some challenges warrant consideration when constructing a network that incorporates DL into clinical decision-making. We present these below.

Data availability and annotation

A key challenge is the availability of data, specifically medical images related to the clinical task at hand. The lack of sufficient data for training DL models in medical image analysis can limit the ability of deep neural networks to perform adequately. This problem is further exasperated by the time-consuming, expensive, and error-prone process of medical imaging annotation. One common solution is to transfer learning from pretrained models, *e.g.*, ImageNet. However, this approach could be ineffective in many instances due to differences in learned features between natural and medical images. To overcome the challenges associated with transfer learning, several novel approaches have been proposed.[48–50] Alzubaidi et al proposed training the DL model on large, unlabeled medical image datasets. This knowledge is then transferred to train the DL model on the small amount of labeled medical images.[50]

Overfitting and class imbalance

Given the large number of parameters that need to be optimized, a major concern in DL is insufficient disease representation, which may result in overfitting or class imbalance. The design and evaluation of DL networks should consider the risks associated with overtraining and overfitting of a particular network, which can lead to poor performance on data that has not been used for training purposes. A reliable network must incorporate sufficient instances of disease and/or rare diseases that might not be fully reflected within the network architecture and could therefore lead to reduced performance.[51] Several solutions have been proposed to address the class imbalance problem. These include training the network with random undersampling, or removing some observations of the majority class; random oversampling, or higher sampling of the minority classes; Synthetic Minority Oversampling Technique (SMOTE)[52] ; the NearMiss family of methods,[53] which is an undersampling technique; and penalizing learning algorithms, which is a cost-sensitive training; among others.

Data bias

At least seven types of data bias have been identified in machine learning literature, including sample bias, exclusion bias, measurement bias, recall bias, racial bias, and association bias. A complete survey of bias and fairness in machine learning is beyond the scope of this paper but the reader is referred to[54] for further details. Tools exist to address these issues, such as dividing datasets to train the model and including datasets from multiple testing centers. Further data availability and the sharing of MR images across institutions would mitigate concerns regarding generalizability, as well as enhance confidence in the reliability of DL methods for clinical use.

Interpretability: the 'black-box' approach

Another important challenge associated with DL models is the 'black-box' approach, which focuses primarily on optimizing outcome performance. It provides limited insight into internal structures or features of the models that lead to treatment decisions based on given model inputs. This limitation effectively diminishes the confidence required for such implementations to be broadly accepted within a clinical setting. To address this well-recognized challenge, several investigators have advocated for approaches based on interpretable models from the beginning[55] 155. At present, there is no consensus on the proposed approach. Divergence exists among researchers who highlight interpretability of the models.[55] For example, while Alex John London advocates optimal performance and predictive power as the primary basis for model evaluation,[56] others prefer models that are highly transparent. They refer to these as 'explainable medicine' and require causality.[57] This is an active area of research that will have a significant impact on the trajectory of the field.

Regulatory approval, ethical challenges, and reimbursement

Several key obstacles need to be overcome for DL methods to be widely accepted in a clinical setting. These include regulatory approval, which requires FDA approval in the USA and a separate approval process within the European Union. Further, DL implementation in clinical practice requires that legal and ethical issues of liability be resolved ahead of time. Finally, there must be a mechanism in which radiology AI can be reimbursed for usage. The current state of affairs is carefully reviewed by Chen et al..[58]

ACKNOWLEDGMENT

We thank Joanne Chin and Cecile C. Berberat for valuable assistance in the preparation and submission of this manuscript.

REFERENCES

1. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer* 1981; **47**: 207–14. https://doi.org/10.1002/1097-0142(19810101)47:1<207::aid-cncr2820470134>3.0.co;2-6
2. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. european organization for research and treatment of cancer, national cancer institute of the united states, national cancer institute of canada. *J Natl Cancer Inst* 2000; **92**: 205–16. https://doi.org/10.1093/jnci/92.3.205
3. Choi H, Charnsangavej C, Faria SC, Macapinlac HA, Burgess MA, Patel SR, et al. Correlation of computed tomography and positron emission tomography in patients with metastatic gastrointestinal stromal tumor treated at a single institution with imatinib mesylate: proposal of new computed tomography response criteria. *J Clin Oncol* 2007; **25**: 1753–59. https://doi.org/10.1200/JCO.2006.07.3049
4. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009; **45**: 228–47. https://doi.org/10.1016/j.ejca.2008.10.026
5. Llovet JM, Di Bisceglie AM, Bruix J, Kramer BS, Lencioni R, Zhu AX, et al. Design and endpoints of clinical trials in hepatocellular carcinoma. *J Natl Cancer Inst* 2008; **100**: 698–711. https://doi.org/10.1093/jnci/djn134
6. Wolchok JD, Hoos A, O'Day S, Weber JS, Hamid O, Lebbé C, et al. Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. *Clin Cancer Res* 2009; **15**: 7412–20. https://doi.org/10.1158/1078-0432.CCR-09-1624
7. Nishino M, Giobbie-Hurder A, Gargano M, Suda M, Ramaiya NH, Hodi FS. Developing a common language for tumor response to immunotherapy: immune-related response criteria using unidimensional measurements. *Clin Cancer Res* 2013; **19**: 3936–43. https://doi.org/10.1158/1078-0432.CCR-13-0895
8. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**: 436–44. https://doi.org/10.1038/nature14539
9. . A Krizhevsky IS, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in neural information processing systems. 2012;25:1097-105.
10. Zeiler RF. (n.d.). Visualizing and understanding convolutional networks. *European Conference on Computer*; 818–33.
11. . K Simonyan AZ. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations2015.
12. C. Szegedy WL, Yangqing Jia PS, Reed S, Anguelov D, Erhan D, Vanhoucke V, et al. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition. ; 2015. pp. 1–9.
13. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017; **60**: 84–90. https://doi.org/10.1145/3065386
14. . He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770-78.
15. . Ronneberger O FP, Brox T, editor. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention; 2015.
16. . Simonyan K, Zisserman A. ery Deep Convolutional Networks for Large-Scale Image Recognition. The 3rd International Conference on Learning Representations (ICLR2015)2015.
17. Goodfellow AJ, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, et al. (n.d.). Generative adversarial nets. *Advances in Neural Information Processing*.
18. . Wang Y, Yu B, Wang L, Zu C, Lalush DS, Lin W, et al. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. NeuroImage. 2018;174:550-62.
19. . Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst CK. Generating synthetic CTs from magnetic resonance images using generative adversarial networks. Medical physics. 2018.
20. Nie D, Lu J, Zhang H, Adeli E, Wang J, Yu Z, et al. Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Sci Rep* 2019; **9**(1): 1103.
21. Lao J, Chen Y, Li Z-C, Li Q, Zhang J, Liu J, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci Rep* 2017; **7**(1): 10353. https://doi.org/10.1038/s41598-017-10649-8
22. 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**: S1470-2045(19)30098-1: 728–40: . https://doi.org/10.1016/S1470-2045(19)30098-1
23. Han W, Qin L, Bay C, Chen X, Yu K-H, Miskin N, et al. Deep transfer learning and radiomics feature prediction of survival of patients with high-grade gliomas. *AJNR Am J Neuroradiol* 2020; **41**: 40–48. https://doi.org/10.3174/ajnr.A6365
24. Hylton NM, Blume JD, Bernreuter WK, Pisano ED, Rosen MA, Morris EA, et al. Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy--results from ACRIN 6657/I-SPY TRIAL. *Radiology* 2012; **263**: 663–72. https://doi.org/10.1148/radiol.12110748
25. HB Q, NGML A, eds. Comparison of breast DCE-MRI contrast time points for predicting response to neoadjuvant chemotherapy using deep convolutional neural network features with transfer learning. *Medical Imaging* 2017.
26. Ravichandran K, Braman N, Janowczyk A, Madabhushi A, Mori K, Petrick N. A deep learning classifier for prediction of pathological complete response to neoadjuvant chemotherapy from baseline breast DCE-MRI. In: *Computer-Aided Diagnosis*. ; 2018. https://doi.org/10.1117/12.2294056
27. Liu MZ, Mutasa S, Chang P, Siddique M, Jambawalikar S, Ha R. A novel CNN algorithm for pathological complete response prediction using an I-SPY TRIAL breast MRI database. *Magn Reson Imaging* 2020; **73**: S0730-725X(20)30273-3: 148–51: . https://doi.org/10.1016/j.mri.2020.08.021
28. Comes MC, Fanizzi A, Bove S, Didonna V, Diotaiuti S, La Forgia D, et al. Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-mris. *Sci Rep* 2021; **11**(1): 14123. https://doi.org/10.1038/s41598-021-93592-z
29. Ha R, Chin C, Karcich J, Liu MZ, Chang P, Mutasa S, et al. Prior to initiation of chemotherapy, can we predict breast tumor response? deep learning convolutional neural networks approach using a breast MRI tumor dataset. *J Digit Imaging* 2019; **32**: 693–701. https://doi.org/10.1007/s10278-018-0144-1
30. Ha R, Chang P, Karcich J, Mutasa S, Van Sant EP, Connolly E, et al. Predicting post neoadjuvant axillary response using a novel convolutional neural network algorithm. *Ann Surg Oncol* 2018; **25**: 3037–43. https://doi.org/10.1245/s10434-018-6613-4
31. El Adoui M, Drisis S, Benjelloun M. Multi-input deep learning architecture

for predicting breast tumor response to chemotherapy using quantitative MR images. *Int J Comput Assist Radiol Surg* 2020; **15**: 1491–1500. https://doi.org/10.1007/s11548-020-02209-9

32. Choi JH, Kim H-A, Kim W, Lim I, Lee I, Byun BH, et al. Early prediction of neoadjuvant chemotherapy response for advanced breast cancer using PET/MRI image deep learning. *Sci Rep* 2020; **10**(1): 21149. https://doi.org/10.1038/s41598-020-77875-5
33. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015; **136**: E359-86. https://doi.org/10.1002/ijc.29210
34. Zhu H-B, Xu D, Ye M, Sun L, Zhang X-Y, Li X-T, et al. Deep learning-assisted magnetic resonance imaging prediction of tumor response to chemotherapy in patients with colorectal liver metastases. *Int J Cancer* 2021; **148**: 1717–30. https://doi.org/10.1002/ijc.33427
35. Hötker AM, Garcia-Aguilar J, Gollub MJ. Multiparametric MRI of rectal cancer in the assessment of response to therapy: a systematic review. *Dis Colon Rectum* 2014; **57**: 790–99. https://doi.org/10.1097/DCR.0000000000000127
36. Barbaro B, Fiorucci C, Tebala C, Valentini V, Gambacorta MA, Vecchio FM, et al. Locally advanced rectal cancer: MR imaging in prediction of response after preoperative chemotherapy and radiation therapy. *Radiology* 2009; **250**: 730–39. https://doi.org/10.1148/radiol.2503080310
37. Shi L, Zhang Y, Nie K, Sun X, Niu T, Yue N, et al. Machine learning for prediction of chemoradiation therapy response in rectal cancer using pre-treatment and mid-radiation multi-parametric MRI. *Magn Reson Imaging* 2019; **61**: S0730-725X(19)30145-6: 33–40: . https://doi.org/10.1016/j.mri.2019.05.003
38. Nardone V, Reginelli A, Grassi R, Boldrini L, Vacca G, D'Ippolito E, et al. Delta radiomics: a systematic review. *Radiol Med* 2021; **126**: 1571–83. https://doi.org/10.1007/s11547-021-01436-7
39. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017; **14**: 749–62. https://doi.org/10.1038/nrclinonc.2017.141
40. Rao S-X, Lambregts DM, Schnerr RS, Beckers RC, Maas M, Albarello F, et al. CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy? *United European Gastroenterol J* 2016; **4**: 257–63. https://doi.org/10.1177/2050640615601603
41. Goh V, Ganeshan B, Nathan P, Juttla JK, Vinayan A, Miles KA. Assessment of response to tyrosine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker. *Radiology* 2011; **261**: 165–71. https://doi.org/10.1148/radiol.11110264
42. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Sci Rep* 2017; **7**(1): 588. https://doi.org/10.1038/s41598-017-00665-z
43. Chang Y, Lafata K, Sun W, Wang C, Chang Z, Kirkpatrick JP, et al. An investigation of machine learning methods in delta-radiomics feature analysis. *PLoS One* 2019; **14**(12): e0226348. https://doi.org/10.1371/journal.pone.0226348
44. Zhang X-Y, Wang L, Zhu H-T, Li Z-W, Ye M, Li X-T, et al. Predicting rectal cancer response to neoadjuvant chemoradiotherapy using deep learning of diffusion kurtosis MRI. *Radiology* 2020; **296**: 56–64. https://doi.org/10.1148/radiol.2020190936
45. Bates DDB, Mazaheri Y, Lobaugh S, Golia Pernicka JS, Paroder V, Shia J, et al. Evaluation of diffusion kurtosis and diffusivity from baseline staging MRI as predictive biomarkers for response to neoadjuvant chemoradiation in locally advanced rectal cancer. *Abdom Radiol (NY)* 2019; **44**: 3701–8. https://doi.org/10.1007/s00261-019-02073-5
46. Napel S, Mu W, Jardim-Perassi BV, Aerts HJWL, Gillies RJ. Quantitative imaging of cancer in the postgenomic era: radio(geno)mics, deep learning, and habitats. *Cancer* 2018; **124**: 4633–49. https://doi.org/10.1002/cncr.31630
47. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI. *Radiology* 2019; **290**: 290–97. https://doi.org/10.1148/radiol.2018181352
48. Raghu MZ, Kleinberg J, Bengio ST. (n.d.). Transfusion: understanding transfer learning for medical imaging.
49. Alzubaidi L, Fadhel MA, Al-Shamma O, Zhang J, Santamaría J, Duan Y, et al. Towards A better understanding of transfer learning for medical imaging: A case study. *Applied Sciences* 2020; **10**: 4523. https://doi.org/10.3390/app10134523
50. Alzubaidi L, Al-Amidie M, Al-Asadi A, Humaidi AJ, Al-Shamma O, Fadhel MA, et al. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* 2021; **13**: 1590. https://doi.org/10.3390/cancers13071590
51. BudaM, Mazurowski MA. (n.d.). A systematic study of the class imbalance problem in convolutional neural networks.
52. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Jair* 2002; **16**: 321–57. https://doi.org/10.1613/jair.953
53 . Mani I ZI, editor. knn approach to unbalanced data distributions: a case study involving information extraction. Proceedings of workshop on learning from imbalanced datasets; 2003.
54. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv* 2021; **54**: 1–35. https://doi.org/10.1145/3457607
55 . Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.
56. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep* 2019; **49**: 15–21. https://doi.org/10.1002/hast.973
57. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; **9**(4): e1312. https://doi.org/10.1002/widm.1312
58. Chen MM, Golding LP, Nicola GN. Who will pay for AI? *Radiol Artif Intell* 2021; **3**: e210030. https://doi.org/10.1148/ryai.2021210030