Vietnam Journal of Science and Technology ***60*** *(6) (2022) 885-928*

doi:10.15625/2525-2518/17256

REVIEW

DEEP LEARNING - CANCER GENETICS AND APPLICATION OF DEEP LEARNING TO CANCER ONCOLOGY

Doan B. Hoang[1], Simon Hoang[2]

[1]*School of Electrical and Data Engineering, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia*

[2]*Sydney Local Health District, Sydney, NSW 2137, Australia*

[*]Emails: *Doan.Hoang@uts.edu.au*

Received: 4 July 2022; Accepted for publication: 26 August 2022

Abstract. Arguably the human body has been one of the most sophisticated systems we encounter but until now we are still far from understanding its complexity. We have been trying to replicate human intelligence by way of artificial intelligence but with limited success. We have discovered the molecular structure in terms of genetics, performed gene editing to change an organism's DNA and much more, but their translatability into the field of oncology has remained limited. Conventional machine learning methods achieved some degree of success in solving problems for which we do not have an explicit algorithm. However, they are basically shallow learning methods, not rich enough to discover and extract intricate features that represent patterns in the real environment. Deep learning has exceeded human performance in pattern recognition as well as strategic games and are powerful for dealing with many complex problems. High-throughput sequencing and microarray techniques have generated vast amounts of data and allowed the comprehensive study of gene expression in tumor cells. The application of deep learning with molecular data enables applications in oncology with information not available from clinical diagnosis. This paper provides an overview of the fundamental concepts of deep learning, some essential knowledge of cancer genetics, and a review of applications of deep learning to cancer oncology. Importantly, it provides an insightful knowledge of deep learning and an extensive discussion on its challenges. The ultimate purpose is to stimulate ideas and facilitate collaborations between cancer biologists and deep learning researchers to address challenging oncological problems using advanced deep learning technologies.

Keywords: deep learning, cancer genetics, cancer oncology, drug response prediction, deep learning applications.

Classification numbers: 4.7.4, 4.8.5

1. INTRODUCTION

Arguably the human body is an extremely complicated system which we are still far from completely understanding. We have been struggling to define and explain the working of the higher mental functions by way of artificial intelligence and have had limited success.

We have made breakthroughs in discovering the "imprint" DNA makes on the fundamental building blocks of the human body. We have performed genome editing to change an organism's DNA for many applications in health and medicine, but their translatability into the field of oncology has remained limited. We are still tentatively probing many fundamental issues and still merely exploring the scientific basis of the root cause of many problems including cancer [1, 2]. It is clear that when a system or a scenario is extremely complex and involves a combination of nonlinear and time varying relationships, mathematical frameworks are unable to capture the inherent complexity. Their simplified models with limited sets of parameters and preconceived assumptions are unable to offer solutions to many real-world problems such as preventing cancer or cybersecurity attacks. Turning to a more realistic scenario of building data-driven systems, conventional machine learning has achieved successes in solving some difficult problems. However, it relies on a limited set of crafted features to feed current machine leaning methods; therefore, the success of such an approach is expected only for cases where the data actually correlates with the crafted features. Furthermore, conventional machine learning methods are not able to handle problems with millions of parameters and vast amounts of data samples. Conventional machine learning methods are basically shallow learning methods that are not rich enough to formulate and extract a large number of complex features that represent objects or patterns in real environments.

Deep learning makes impressive progress towards human intelligence as it utilizes the process of "learning by examples." The AlphaGo computer program defeated the world boardgame Go champion, Lee Sedol, 4 to 1 in a five-game match in 2016 [3]. AlexNet image classification model based on a convolutional neural network (CNN), proposed by Krizhevsky [4], won the image classification competition of the image dataset ImageNet [5] in 2012. Many remarkable successes have also been achieved over the last decade. Yet, it is still puzzling how deep learning performed these feats. We continue testing the limits of deep learning by applying it to many unsolved problems and at the same time determining its reliability in critical situations. We define deep learning as follows, "*Deep learning is a subset of machine learning that employs a rich architectural neural network model with a large set of parameters. These parameters are learned through training with an extremely large amount of data to extract the features and their interrelationships to describe as closely as possible the system or the object represented by the data."*

The discovery of DNA and the genome as the building blocks of the human system has allowed us to decode many fundamental rules in the construction and development of a human body in terms of the DNA arrangement in each cell. With the completion of the Human Genome Project and the advances of technologies such as Next Generation Sequencing, a huge amount of genomic and molecular data has been generated and made available for cancer research and management. Cancer can be viewed as a "complex disease" or "numerous complex personalized diseases". Regardless of the availability of genomic data, cancer challenges us with a real-world problem for which we have yet to find widely applicable solutions. Nevertheless, deep learning has been applied to many cancer-related problems with excellent results which are being translated to clinical use.

Existing reviews either emphasize deep learning technologies and leave cancer biologists out of the equation or focus on molecular and cell biology of cancer without regard for deep learning practitioners who could apply their expertise to cancer research and translation of results to clinical usage. This paper focuses on deep learning methods, cancer genetics, and application of deep learning to cancer oncology. It will focus on providing cancer biologists with an essential understanding of deep learning methods and provide deep learning

researchers/practitioners with a crafted set of knowledge of cancer genetics and molecular biology, not covered by other reviews. This set of knowledge is essential not only for developing better cancer treatments but also for searching the patterns that trigger cancer. The paper discusses opportunities and challenges of deep learning and problems associated with its application to cancer. The prime purpose is to germinate ideas and facilitate the collaboration between cancer biologists and deep learning researchers to address challenging cancer issues using advanced deep learning technologies.

The remainder of the paper is organized as follows. Section 2 provides the fundamentals of deep learning and deep learning methods. Section 3 provides the essentials of human genomics that are relevant to cancer as well as a summarized background of molecular biology of cancer. Section 4 discusses "why deep learning", providing a detailed review of a drug response prediction application to demonstrate the research and development methodology, and identifies resources and datasets for deep learning applications. Section 5 reviews some recent applications of deep learning models to cancer oncology. Section 6 discusses issues and opportunities encountered in deep learning as well as its applications to cancer oncology. The conclusion summarizes the paper and discusses future directions for research.

2. DEEP LEARNING - ESSENTIALS AND INSIGHTS

Deep learning is a subset of machine learning but rather than building a mathematical model as in the case of conventional machine learning, it uses an artificial neural network as the model to extract features and make predictions from large raw datasets. This section aims to review the essentials of deep learning, its underlying features as well as the insights that enable superior performances in image processing, object recognition, board game competition against human champions, and other applications across diverse domains.

2.1. From Machine Learning to Deep Learning

For a computer to solve a problem, it requires an algorithm - a sequence of instructions mapping the input to the output. For example, when given a set of numbers, the computer only has to execute the instructions of the sorting algorithm to output the ordered list of numbers. However, we do not have an algorithm for many real-world problems. Differentiating spam emails from legitimate ones is such a problem. Given an email (a message) as input, the expected output should be either yes or no. Yet, we do not know how to devise an explicit set of instructions to directly decide whether the input email is spam. The difficulty is that what is considered spam is time varying and dependent on the specific context among individuals. However, if we have a lot of data, say, thousands of example messages, some are known to be spam and some are not, we can learn from the data what constitutes spam. Learning from examples is what we expect from a machine. In other words, machine learning is a process to extract automatically an implicit algorithm when we do not have an explicit algorithm but we do have lots of data [6]. When we are not able to identify or classify patterns buried deep in the data, machine learning helps detect and make prediction based on the data. To use machine learning to solve a problem for some set of data, an appropriate family of models is selected together with a set of selected parameters or features. The model is trained by solving an optimization problem that optimizes the selected parameters with respect to the selected loss function. In other words, a machine learning algorithm builds a mathematical model based on sample data (or training data) to make predictions or decisions without being explicitly

programmed to perform the task. Deep learning is a specialized form of machine learning with several differentiating features:

- The models used by deep learning and conventional machine learning are different. Deep learning uses artificial neural networks with multiple hidden layers as the model for training and learning, while conventional machine learning builds a mathematical model, based on the data.

- Deep learning methods allow a machine to be fed with a large volume of heterogeneous and high-dimensional raw data and to automatically discover the features needed for detection or classification. With conventional machine learning, machine learning experts in the problem domain have to extract manually a set of salient features from the raw data suitable for the learning system to detect or classify patterns in the input.

- In deep learning, most model parameters are learned not directly from the features of the training examples, but from the outputs of intermediate stages (or of the preceding layers of the neural network architecture) [7]. On the other hand, most machine learning algorithms are shallow in that they learn the parameters of the model directly from the features of the training examples. See **Box 1** for shallow learning vs deep learning.

2.2. Deep learning: architectures and learning methods

A deep learning method is characterized by the neural network architecture that it employs and the method it uses for training the parameters of the architecture to perform its intended function. The next section discusses various deep learning methods in terms of their architectures and learning methods.

2.2.1 Architectures

Deep learning is a form of machine learning where the model for learning is a multi-hidden-layer artificial neural network (ANN). An artificial neural network is inspired by neural networks in the brain [8]. To understand the deep learning architecture, let us briefly describe the basic working of a neuron and the composition of neurons in neural networks.

Neuron. Neurons are the specialized cells of the nervous system. Input to the nervous system is via sensory transducer neurons and output is through the triggering of muscle fiber contraction by motor neurons. Every neuron takes its inputs from, and sends its output to, groups of other neurons, transforming multiple inputs into a single output signal. A group of interconnected neurons is called a neural network [9].

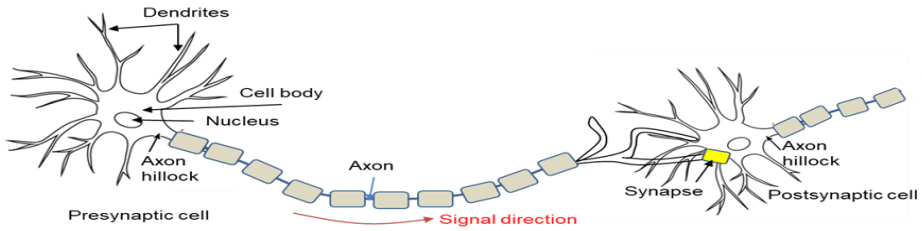


Figure 1. A neural cell in the nervous system.

A typical brain neuron (shown in Figure 1) has a cell body and a number of dendrites (highly branched extensions that receive signal from other neurons) and a single axon, an

extension that transmits signals to other neurons. Each branched end of an axon transmits information to another neuron at a junction called a synapse. At most synapses, chemical messengers called neurotransmitters pass information from the transmitting neuron to the receiving neuron. An input signal received by a neuron triggers a change in its membrane voltage called an action potential, which is a nerve pulse that carries information along the axon. A neuron can produce hundreds of nerve pulses per second and the frequency, with which a neuron generates pulses, can vary in response to input. The neuron's firing rate will therefore convey information about input signal strength [10].

Artificial Neuron. An artificial neuron (or a node) simplifies a neuron as follows. Inputs from impinging nodes are represented by $\{\mathbf{x}_i\}$, the node is represented as a processing unit that sums or integrates all its weighted inputs. The node processing unit (or the cell body) can be thought of as an integrator (summation) of the net input. The output of the node is often a nonlinear squashed value between 0 or 1 depending on the input sum. This transformation is called the activation function which may take several forms as dictated by practical training methods such as sigmoid and ReLU (Rectifier Linear Unit) - the two commonly used functions.

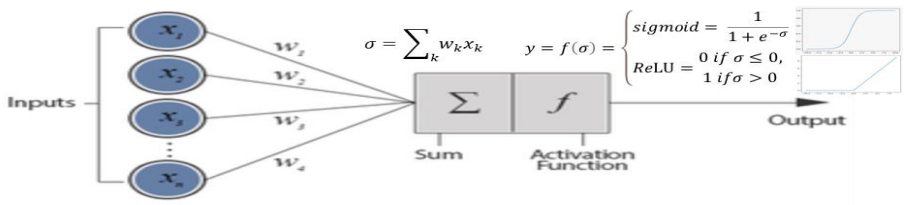


Figure 2. Operations of an artificial neuron.

The input into a typical node is a vector of elements $\{\mathbf{x}_k\}$. These elements are scaled by a vector of weights $\{\mathbf{w}_k\}$ to give a raw weighted sum strength of the node:

$$\sigma = \sum_k w_k x_k \tag{1}$$

The weighted sum is passed through an activation function f(σ) to yield the output *y*:

$$y = f(\sigma) = \begin{cases} sigmoid = \frac{1}{1+e^{-\sigma}}, values\ between\ 0\ and\ 1 \\ ReLU = \max(0, \sigma) \end{cases} \tag{2}$$

An activation function is selected for several reasons: (a) it provides a way to take into account a nonlinear relationship between the input and the output and (b) it is preferably a differentiable function so that the error between the desired output and its estimate can be optimized with a steepest gradient descent method (to be discussed later). The most common activation functions are the sigmoid function and the ReLU function as shown in (2). See **Box 2** for further explanation.

Neural Networks. The cerebral cortex of the brain is comprised of the grey matter (so-called because of the color) and the white matter. The grey matter forms the upper surface of the brain and below is the white matter, through which flow both axons of grey matter neurons (travelling to either other parts of the grey matter, or other parts of the brain, or the remainder of the body) and axons of neurons located in other parts of the brain (travelling to the grey matter). The cerebral cortex is a patchwork of distinct cortical areas, which are the functional modules of the cortex such as sensory function and association function, each dedicated to particular

information processing tasks. A cortical area is defined anatomically by the dense networks of axons that travel through the grey matter connecting neurons located within the same area. In contrast, connections between neurons located in different cortical areas are less numerous, and travel through the white matter. Cortical areas can also be distinguished through differences in the distribution and morphology of their neurons, as well as through differences in their patterns of intra-areal and inter-areal connectivity [9], [10]. Cortical areas are organized in a six-layered neural structure, with layer 1 at the surface of the cortex and layer 6 adjacent to the white matter. Each layer in a cortical area is a slab of cortex containing a common combination of neuron types, with each neuron of a particular type having a common pattern of synaptic input sources and axonal output targets. Each layer interacts in specific ways with both the layers in its own area and layers in other cortical areas.

Artificial neural networks. Inspired by human neural networks, an artificial neural network consists of groups of nodes structured in layers from left to right as shown in Figure 3. The input layer, the hidden layers, and the output layer. ANNs constructed in this way have been shown to be capable of performing tasks associated with learning and memory. Each artificial neural network (representing a cortical area) is dedicated to particular information processing tasks, for example learning to recognize an object from sensory neurons of the input layer. Figure 3 shows typical architectures of artificial neural networks: (A) no hidden layered network (only input and output layers), one hidden-layered network, and (B) multiple hidden-layered network. The inputs are presented as the input layer. Typically, once the inputs of a layer are weighted according to the signal strength and transformed through the activation function, the outputs of the layer become the inputs of the following layer. The last layer, the output layer of the neural network, provides the estimate or expected outcome, such as a classification or a regression or a feature extraction depending on the type of neural learning method.

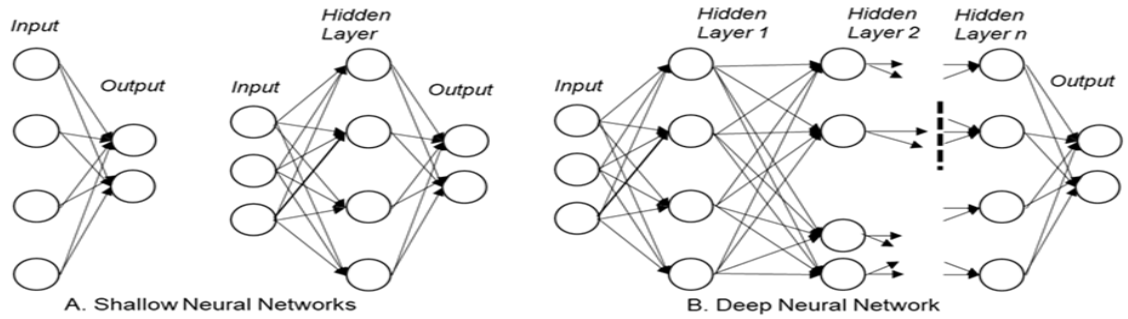


Figure 3. Artificial Neural Networks. (A) Shallow Networks. (B) Deep Networks.

Deep artificial neural network (DNN) architectures have an input layer, an output layer, and multiple hidden layers in between. The nodes in these different layers may be fully or partially connected. The input layer nodes receive input in the form of features, and thus typically the number of input layer nodes is equal to the number of features in the training data. The depth of a neural network (*we use the term neural network instead of artificial neural network from here on*) corresponds to the number of hidden layers, and the width to the maximum number of nodes in one of its layers. As shown in Figure 4, data is received at the input layer which transforms the data in a nonlinear way through the first hidden layers (l_1), which in turn transforms the output of the first hidden layer in a nonlinear way through the second hidden layer (l_2), and so on, before final outputs are determined at the output layer. Each node computes a weighted sum of its inputs and applies a nonlinear activation function to calculate its output $f(\sigma)$.

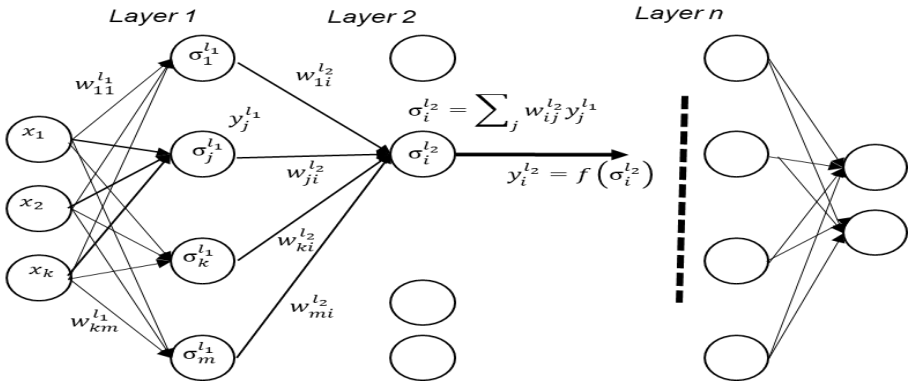


Figure 4. Deep Neural Network - Hidden Layer Processing.

The input into a typical neural network is a vector of elements $\{\mathbf{x}_k\}$. These are combined by a series of linear filters with weights $\{\mathbf{w}_k\}$ to give the inputs to the *hidden unit j* (eqn. 3)

$$\sigma_j^{l_1} = \sum_k w_{jk}^{l_1} x_k \tag{3}$$

This weighted sum of all the inputs to node *j* in the hidden layer is passed through an activation function f(σ) to give the output of *node j* (eqn. 4)

$$y_j^{l_1} = f(\sigma_j) = f\left(\sum_k w_{jk}^{l_1} x_k\right) \tag{4}$$

The activation function $f(\sigma_j)$ can be sigmoid or ReLU as seen above. The outputs from these hidden units of *layer 1* (l_1) then go through another layer of filters, produce a weighted sum at *node i of layer 2* (l_2)

$$\sigma_i^{l_2} = \sum_j w_{ij}^{l_2} y_j^{l_1} = \sum_j w_{ij}^{l_2} f\left(\sum_k w_{jk}^{l_1} x_k\right) \tag{5}$$

and be fed through another layer of activation functions to produce the outputs of layer 2. This process continues until the final outputs are produced at the output layer.

$$y_i^{l_2} = f(\sigma_i^{l_2}) = f\left[\sum_j w_{ij}^{l_2} f\left(\sum_k w_{jk}^{l_1} x_k\right)\right] \tag{6}$$

This set of operations is performed at every layer in the forward direction until the final outputs are obtained at the network output layer.

2.2.2. Learning

Neural Plasticity. In the human nervous system, it is known that the connection between an input neuron and its target neuron is adaptive in that the strength of the connection varies in accordance with the strength (or frequency) of signal sensed by the input neuron. This characterizes neural learning [10]. Learning is by repeatedly presenting the input pattern to optimize the intended performance measure, for example, minimize the difference between the output of the network and the intended output (called the loss function, in general) by varying the weights of the connections between neurons. In machine learning or deep learning, three major learning strategies include supervised learning, unsupervised learning, and reinforcement learning (Figure 5).

Box 1. Shallow versus Deep learning.

With no hidden layers, a neural network can classify only linearly separable problems. It has been shown that with one hidden layer a network can describe any continuous function with an "adequate" number of hidden nodes, often an extremely large number of nodes are needed in the hidden layer. With two hidden layers the network can describe any function at all [11], [12], [13]. It means that a deep learning network may include one or more hidden layers; however, for practical reasons, deep learning architectures imply two or more hidden layers. This differentiates deep learning from shallow learning where the architecture only has one or no hidden layer.

Supervised learning. It is a form of learning by examples whereby the algorithm (or the model) is trained to predict the desired output (label) from known input values. The function that measures the difference between the model estimate and the true label is called the loss function L(**w**) (or the error function E(**w**)). Learning is through training the network by repeatedly presenting the training examples (input-label samples) and letting the weights (the network free parameters) adapt in such a way to minimize the loss function L(**w**). This minimization is challenging, especially when the loss function is non-convex and high-dimensional. Backpropagation is the algorithm adopted for such a minimization in most neural architectures.

Backpropagation algorithm. As L(**w**) is a function of the weights between the input and the output, its rate of change is determined by its gradient relative to the input, provided the function is differentiable. Since neural networks are organised in a sequential chain of layers from the input layer through the hidden layers to the output layer, the gradient of the loss function can be computed via the chain rule for derivatives in a backward manner from the output to the input (layer *n* to layer *1*) as follows.

$$\frac{dL(W)}{dX} = \frac{dL(W)}{dY_{l_n}} \times \frac{dY_{l_n}}{dY_{l_{n-1}}} \times \times \times \frac{dY_{l_2}}{dY_{l_1}} \times \frac{dY_{l_1}}{dX} \quad (7)$$

During training, the predicted output (or the model estimate) is compared with the true label to compute a loss for the current set of model weights. The gradient of the loss function relative to the input is determined by computing the loss caused by the weights in the layer just before the output then the loss caused by the weights of the layer before that and the procedure is repeated in the backward direction until the input is reached. Hence the algorithm is called a backpropagation algorithm. The loss function L(**w**) and the weights are updated after each input-label sample. L(**w**) is optimized by following the steepest gradient descent (downhill slope) d**w** by learning rate α, a model parameter selected for the best performance.

Box 2. The Vanishing gradient problem.

The vanishing gradient problem arises whenever a network is trying to learn a model such as a multi-hidden-layered neural network but the gradients of a loss function become smaller and smaller through propagating backward layer by layer from the output to the input layer. The gradients may decrease to such a small value that adjustments to the weights make no difference to the optimization process and the learning process has to terminate as it never reaches a local or global minimum. Activation function such as sigmoid is prone to this vanishing problem as its gradients become small when the function operates near zero or one where the function is very flat. Surprisingly, it was discovered after a decade of experiments that the ReLU activation function solves this problem and allows faster learning in most deep learning cases. One issue is that ReLU is unbounded but this has not caused problems so far. For further exposure on ReLU and its variations, see [14 - 16].

The calculation using sigmoid is relatively complex and requires a long computation time, and the gradient may vanish during the process of back propagation. Although ReLU has its own issues, it requires little computational effort and has a faster convergence speed.

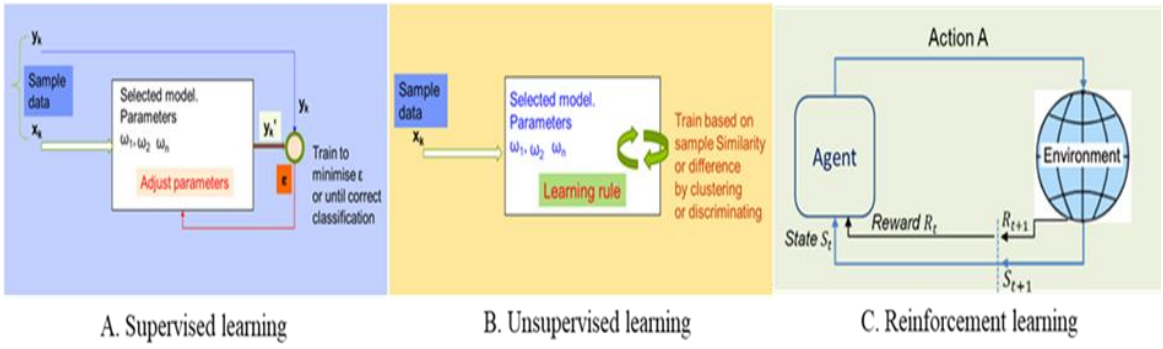


Figure 5. Learning methods: (A) Supervised, (B) Unsupervised, (C) Reinforcement.

Unsupervised learning. With this form of learning, only the input data but no label (or ground truth) is available to guide the learning process. A model is learnt by clustering groups of input samples based on their similarity so as to decide if they belong to the same class. Similarity is a measure of some property of the data such as the distance between a specific feature or a group of features of data samples. Similarity measures are often *preconceived* by the designer of the algorithm. For a true unsupervised learning, it is expected that an unsupervised learning method be able to discover features of the data itself and select the similarity measure by itself given the set of input data. This form of learning implies a degree of intelligence that humans possess as they do learn to perform tasks by themselves.

Reinforcement learning. Reinforcement learning (RL) is learning to map situations to actions through trial and error so as to maximize a numerical reward signal. In reinforcement learning, an agent is an entity that has explicit goals, can sense aspects of their environments, and can choose actions to influence their environments. RL agents learn through observing their environment, taking actions based on their observations, and assessing the utility of their behaviour through the incoming reward signal. RL agents learn by making observations of the state of the environment at each time step t. They use this information to select actions A to obtain a reward R_t. In a reinforcement learning system, a reward signal defines the goal of a reinforcement learning problem; a policy defines the learning agent's way of behaving at a given time; a value function specifies what is good in the long run. On each time step, the environment sends to the reinforcement learning agent a single number called the reward and the agent selects actions to influence the environment so as to accumulate and maximize the total amount of rewards [17].

2.3. Deep learning methods

In a previous section we describe a general, fully connected feedforward DNN. However, many deep learning methods have been developed for specific applications. These deep learning methods differ in the way nodes are arranged in the architecture *and* their specific learning methods, not just their architectures. In this section we describe the most useful deep learning methods including convolutional neural networks (CNNs), recurrent neural networks (RNNs),

Autoencoder-Decoder neural networks (AEs), and generative adversarial neural networks (GANs). The diagrams for these networks are shown in Figure 6.

Convolutional neural network [18] is one of the most popular types of deep neural networks because of their superior performance for spatial data. CNNs are inspired by the neural process of the human visual cortex and are designed to process multiple data types, such as language sequences or images. The CNN works by extracting features directly from the data without the need to define them. Convolutional neural networks (CNN) represent a specific type of feedforward DL model. CNNs learn hierarchical representations by detecting different features of the presented data using a large number of hidden layers. Every hidden layer increases the hierarchy and complexity of the learned features. For example, the first hidden layer could learn how to detect edges, the second may learn to detect other components and so on until the complete shape of objects can be presented to the classifier layer of the CNN which decides or recognizes the intended object.

The architecture of a typical CNN is structured as a series of stages. The basic structure of CNNs consists of three main types of hidden layers: (i) convolution layer, (ii) pooling layer, and (iii) fully connected layer. To deal with multidimensional data, the convolution layer of CNN has neurons arranged in three dimensions. The three-dimensional layers consist of several 2D filters. Each 2D filter is a window of neural nodes and *connects with only a local section* of a layer before it. The filter interacts with all the nodes in the previous layer through the convolution process by sliding the filter window over all sections of the previous layer. By computing convolutions between local sections and weight vectors (or filter weights), feature maps (local weighted sums obtained by sliding the filter over across all sections) are obtained at each convolution layer. This type of convolution operation allows CNNs to extract features that are highly correlated within and across sub-sections of data (Figure 6 and **Box 3**). The pooling layer of CNN reduces size of the layer by calculating mean, maximum, or other statistics of non-overlapping subsections in the feature maps. This type of non-overlapping subsampling not only reduces the size of the feature maps but also enables CNNs to merge local features to identify more complex features. The final layer of CNN consists of fully connected layers similar to traditional DNN to perform supervised classification or regression. The architecture of a typical CNN is structured as a series of stages. The earlier stages are composed of a series of convolutional-pooling combined layers. The final stage is usually composed of fully connected layers for classification or regression. Supervised learning is used for CNNs. Backpropagating gradients through a CNN is as simple as through a regular deep network, allowing all the weights in all the filter banks to be trained.

Box 3. Meaning of the Convolutional Operation

If one creates an event (an impulse) at a single point in time t_1 and applies it to any real system, its effect on the system will last or persist (with reduced effect) for a period afterward, counting from time t_1. At time t_2 (after t_1), if one then applies another impulse, the effect on the system will be the additive effects of both impulses at time t_2 onwards. In general, if an impulse is applied to the system at time t_n, the overall effect on the system at time t_n is the sum of the remaining effects of all other impulses that have been applied to the system (at t_1, t_2, …, t_n). Explicitly, the effects on the system at time t_n have n components: the remaining effect of the impulse input that was applied at time t_1 plus the remaining effect of the impulse input that was applied at time t_2 plus the remaining effect of the impulse input that was applied at time t_3 plus … plus the remaining effect of the impulse input that was applied at time t_{n-1} plus the effect of the impulse at time t_n. The time convolution effect is computed by shifting the time and adding the contributions from

all previous and current inputs at that timepoint. Time convolution is thus a computational method that takes into account the contribution of all inputs that affect the output at a specific timepoint.

Convolution in CNN mirrors the same concept but in the spatial domain (spatial convolution), rather than in the time domain. For a 2-D image, this is done by shifting a window from left to right and from top to bottom of an image to take into account the effect of all surrounding pixels over a particular spatial point in the image. Similar process applies for 3-D or higher dimension objects. This explains why CNNs perform well for spatial data.

In summary, the architecture and the learning/training of CNNs are designed to achieve the following properties. CNNs learn hierarchical representations from data in which higher-level features are obtained by composing lower-level ones. The CNN architecture reduces the number of model parameters compared to a fully connected network by applying convolutional operations to only small regions of the input space and by sharing parameters between regions. This allows more efficient training. Through convolution and pooling operations, CNNs achieve a high degree of invariance to location and translation, allowing superior object recognition.

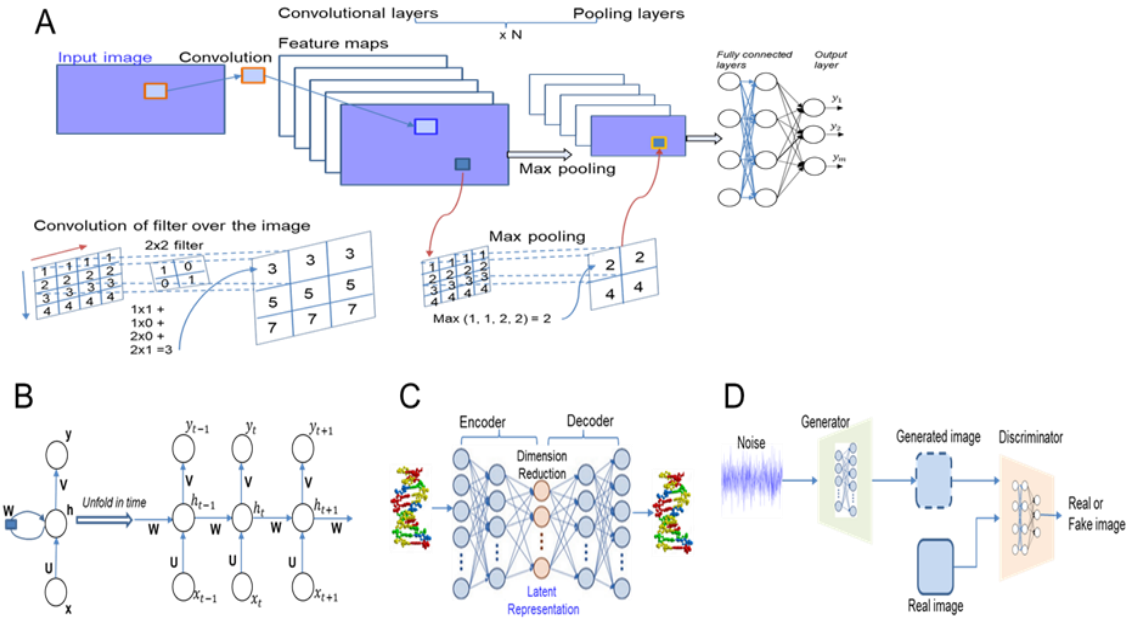


Figure 6. Diagrams of specific Multilayer Deep Learning Networks: (A) Convolution, (B) Recurrent, (C) Auto Encoder-Decoder, (D) Generative Adversarial Networks.

Recurrent neural networks [19] are designed to model data that are sequential in nature, such as natural language or time series, and sequentially dependent data. RNNs have a basic structure with a cyclic connection. A node can receive as input not only the current data point but also the values of hidden units from previous time steps. RNNs, once unfolded in time, can be seen as deep feedforward networks with the same weights being shared among layers. The basic structure of RNNs is shown in Figure 6 with an input unit **x**, a hidden unit **h** and an output unit **y**. While processing data x_t at sequential step t, hidden layer weight parameters W and hidden layer activations h_{t-1} of previous step are used. The computed results propagate on to y_t and h_{t+1}. **U** are weight parameters for connection between the input node and hidden layer, whereas **V** are

weight parameters for connection between hidden layer node and output node y_t. The outputs of the hidden nodes at different discrete time steps can be considered as the outputs of different nodes in a deep multilayer network. *For RNN networks, the process of convolution applies naturally as an unrolled RNN is just a sequence of time events. RNNs, once unfolded in time, can be seen as deep feedforward networks. RNNs are thus appropriate for sequential data.* Supervised learning with a backpropagation algorithm is also used to train RNNs and it is called backpropagation through time (BPTT). Although the main purpose is to learn long-term dependencies, the BPTT method is more prone to the problem of vanishing gradients when the time sequence is long (i.e., large number of layers). Long short-term memory (LSTM) networks are specifically designed to handle this problem. LSTMs use special hidden units, called the memory cell, to maintain inputs for a long period. The cell has a connection to itself. At the next time step, it copies its own real-valued state and accumulates the external signal. This self-connection is gated by another unit that learns to decide when to clear the content of the memory. LSTM networks have subsequently proved to be effective for machine translation.

Autoencoders (AEs) are typically used for dimensionality reduction and feature representation learning before feeding other ML or DL methods for prediction. An AE is a NN that learns to reconstruct its input data by first reducing the dimension of the input and then reconstructing the original input data from the dimension-reduced input. AE neural networks have two components: the encoder and the decoder. The encoder transforms information from the input layer through a series of hidden layers into a hidden layer with fewer nodes at the end of the encoder, this encoder output layer represents the latent features of the original input. The decoder then takes the output layer of the encoder as its input and learns to reconstruct the original input of the AE. In the process of the encoding and the reconstruction, the AE maps the original input space to a specific feature space. A reduced dimension of the feature space is achieved by restricting the number of nodes in the hidden layers of the encoder. An autoencoder discovers latent features of the input data without relying on the labels associated with training instances of the input. In this sense, AE is an unsupervised deep learning neural network. To achieve this, the input sample acts both as the input to the encoder and as the label at the output of the decoder to calculate the loss function when the autoencoder learns to reconstruct its original input. The learning is by backpropagation through minimization of the errors between the output of the decoder and the original input to the autoencoder.

Generative Adversarial Networks (GANs) [20]**.** Architecturally, generative adversarial networks consist of two neural networks in adversarial roles: a generator that iteratively learns to generate more and more realistic samples and a discriminator that tries to identify whether the samples are model generated or real. This competition between the two subnetworks ultimately makes GANs capable of generating samples that are indistinguishable from the corresponding real-world samples. Explicitly, the generator is a deep neural network that takes as input, a random noise vector and transforms it into a model distribution and the discriminator is a deep neural network that acts like a classifier to distinguish between output data point (fake) and training data sample (real). With these settings the learnt weights of the generator are the parameters that represent the model distribution.

2.4. Deep learning training process

A dataset is a collection of examples or sample points. The training process consists of four steps: data preparation, training, testing, and validation

Data Preparation. Raw data has to be pre-processed or transformed into a dataset that is suitable for the intended deep learning method. This may involve ways to deal with missing data values, transforming categorical features to numerical features, or normalizing feature values, etc. The dataset obtained is often randomly divided into 3 distinct subsets: training set, validation set, and testing set. Nominally, 70 % of the dataset is used for training, 15 % for validation and 15 % for testing but these proportions can be changed depending on the size of the data [7]. The training set is used to train the learning model, the testing set is used to test the model on the data that is not in the training set, and the validation set is used to evaluate the hyperparameters or empirical parameters of the model. Hyperparameters are not the weights (i.e., free parameters) in training, they are chosen beforehand heuristically by the model designer, such as the learning rate, the number of hidden layers, the number of nodes in each layer, the size of the connection weights.

Training. Training is performed by repeatedly presenting data samples from the training dataset, allowing the network to adapt its weights across all layers until the loss function is reduced to an acceptable level. Once the procedure is completed, it is expected that the model will correctly predict the outcome when it is presented with a data sample from unseen datasets.

Testing. Testing is done by presenting samples from the testing set, which the model has not seen in the training set, to determine the accuracy of the model's prediction.

Validating. Validation is the process for selecting the best (deep) learning method by using the validation dataset to find best values for the hyperparameters for the (deep neural network) model. The validation set is distinct from the training dataset and the testing dataset and is used solely for performance validation. However, if the validation dataset is small and deemed inadequate to represent samples of each class, cross-validation may be used. *Cross-validation* [7] on the training dataset is used to simulate a validation set and its working can be described as follows. The training dataset is divided into *n* subsets of the same size. Each subset is called a fold. The hyperparameters are used as variables to evaluate a model. In an *n-fold cross-validation*, train the model F_i ($i = 1 \ldots n$) on old folds, except for the *i*th fold F_i. Once the best hyperparameter values are found, the whole training dataset is used to train the model using those best values and the final model is tested with the test dataset.

Underfit and Overfit problems. Often, we do not know the complexity of the data and hence the selected model (neural network architecture of deep learning) may be too complex or too simple for the data. *Underfitting* is when the model is too simple or the input features are inadequate to describe the complexity of the presented data. Underfit model is unable to correctly predict the labels of the training data. To fix the problem, either more data-correlated input features or a more complex model is needed. *Overfitting* is when the model is too complex for the presented data or the training dataset is too small to describe all the features of the data. Overfit models predict well with the training data but perform poorly for unseen data. The overfitting problem can be overcome by trying (i) simpler models with less parameters such as shallow neural networks or deep networks with smaller number of hidden layers, (ii) regularizing the model through an optimization process that constrains the hyperparameters associated with the complex model [7], or (iii) obtaining more training data.

2.5. Deep learning evaluation

In order to compare these different methods, multiple metrics are used to determine the effectiveness of a classification algorithm. These include accuracy, precision, recall, F-Score, and specificity, which are derived from the confusion matrix, and the area under the receiver

operating curve (AUC). The *confusion matrix* is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. In a binary classification problem, there are two classes. For example, the model predicts two classes: **A** and **not A**. The confusion matrix is shown in Table 1.

Table 1. Confusion Matrix.

Actual \ Predicted	A (predicted)	Not A (predicted)
A (actual)	TP (True Positive)	FN (False Negative)
not A (actual)	FP (False Positive)	TN (True Negative)

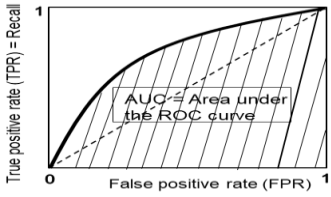
Other performance measures derived from the confusion matrix include precision, recall, accuracy, and specificity.

$$precision = \frac{TP}{TP + FP}; recall = \frac{TP}{TP + FN}; accuracy = \frac{TP + TN}{TP + TN + FP + FN};$$
$$specificity = \frac{TN}{TN + FP}; F\ Score = \frac{2TP}{2TP + FP + FN}$$

Figure 7. Formula for precision, recall, accuracy, specificity, and F-Score (or F1 Score).

Precision determines how many of the predictions are correct. *Recall* shows how many of the correct results are found. *Accuracy* determines how correct the values are predicted. *F-score* uses a combination of precision and recall to calculate a score that can be interpreted as an averaging of both scores. *Specificity* measures a true negative rate or specifies probability that an actual negative will test negative (Figure 7).

The *area under the ROC (receiver operating characteristic) curve* (AUC) is commonly used to assess the performance of classification models. The curve represents the true positive rate against the false positive rate of a classifier as shown in Figure 8. ROC curves can only be used to assess classifiers that return a probability of prediction. The higher the AUC, the better the classifier.



$$where\quad TPR = \frac{TP}{TP + FN} = \text{recall}$$
$$and\quad FPR = \frac{FP}{FP + TN}$$

Figure 8. The area under the ROC curve.

3. CANCER AND CANCER GENETICS

Tumors are formed by the abnormal growth of cells. The abnormal behaviour of these cells is caused by an alteration in their genome. The vast majority of genetic alterations found in cancers develop during the life of a patient in somatic cells [21].

In this section, we provide some introductory information about cancer, a fundamental understanding of human genetics in terms of genome, gene expression, DNA sequencing and various forms of mutations that may lead to cancer. We also focus on molecular and cell biology of cancer and the tumorigenesis processes. Given that deep learning technology is excellent and powerful in extracting latent features and patterns within data, we propose cancer genetics as a challenging area to which deep learning could be applied. There is a potential to not only develop better oncology but also prevent cancer as well as find its root cause which is still unknown [1, 2].

3.1. Cancer in a nutshell

Cancer is caused by unregulated cell growth which leads to the formation of a mass of cells known as a tumor. A benign tumor grows unregulated without tissue invasion. Malignant tumors invade adjacent tissues and metastasize or colonize other organs of the body. Specifically, cancer is caused by a progressive degradation of normal cell behaviour through gene mutations. Mutated genes may lose their normal function and promote out-of-control cell growth, resulting in tumors that can migrate to other parts of the body. Three classes of genes are found to interfere in the regulation and maintenance of the human genome and play a key role in the initiation and progression of cancer: oncogenes, tumor suppressor genes, and caretaker genes. Oncogenes, mutated from proto-oncogenes, interfere with the regulation and differentiation of cells, and cause out-of-control cell growth; mutated tumor suppressor genes are unable to perform their functions of inhibiting cell growth and/or promoting cell death, resulting in unregulated cell growth; and caretaker genes, when mutated, cannot repair DNA damage, allowing the damage to propagate and induce mutation of other genes, resulting in instability of the genome.

Tumor cells grow in a series of steps: the first step is hyperplasia: too many cells are produced from uncontrolled cell division. These cells appear normal, but genes have already mutated, resulting in some loss of growth control. The second step is dysplasia: further growth and abnormal changes to the cells. The third step is anaplastic: the cells become more abnormal, spread over a wider area of tissue and lose their original function. The tumor is still benign. The last step is metastasis: the cells in the tumor invade surrounding tissues, including spreading to other parts of the body through the bloodstream or the lymphoid system and often resulting in mortality. However, not all tumors progress to this point.

Most cancers arise as a consequence of genetic alterations in a single cell, but over time multiple genetic and epigenetic mutations occur in different cells within malignant tumors. This heterogeneity of tumors allows subsets of cells to be resistant to therapy. These cells survive and proliferate even if the majority of cells are killed [9].

3.2. Human genetics

To understand cancer and cancer mechanisms, we need to understand cells and their behavior in terms of genome, DNAs, chromosomes, and genes.

Genome. A genome is all of an organism's genetic material, coding or non-coding, contained in the nucleus and the mitochondria of every cell [22]. It is the complete set of instructions for building, running, and maintaining an organism, and passing on the organism's characteristics to the next generation. The genome is made of a chemical module called DNA. It contains genes, which are packaged in chromosomes. In other words, for an organism, every cell contains its genome, the genome contains chromosomes, chromosomes contain genes, and genes are made of DNA (Figure 9).

Chromosome. A chromosome is a package containing a portion of a genome, that is, it contains some of an organism's genes. Chromosomes are structured for division and only visible in the cell replication stage. Each *chromosome is a long DNA molecule* packed by a protein called histone to make it and all other chromosomes fit inside a cell's nucleus. Different chromosomes contain different genes. Each nucleus contains 46 chromosomes in the case of a human cell. Chromosomes come in pairs, one from the mother and one from the father. The members of a pair have the same size and shape (except the sex chromosomes), and they have the same banding patterns. In other words, each person possesses 23 such pairs with 22 autosomes (chromosomes pairs which are the same for males and females) and a pair of sex chromosomes, which differ between males and females. In humans, a female has two identical sex chromosomes (XX). A male has one sex chromosome that is like those of females, and one that is smaller and differently shaped (XY). *Mitochondrial chromosomes*. Mitochondria are tiny structures inside cells that synthesize molecules used for energy. Unlike other structures inside cells, each mitochondrion contains its own chromosome. This chromosome contains DNA (mitochondrial DNA) that codes for some, but not all, of the proteins that make up that mitochondrion.

Gene. A gene is the fundamental unit of heredity. Genes are found on chromosomes and each gene is made of a specific segment of the DNA molecule. Different genes with different genomic sequences and interaction with the environment determine the phenotypes (distinctive characteristics, or traits) of an organism. The genes on each chromosome are arranged in a particular sequence, and each gene has a particular location on the chromosome (called its locus). The gene locus also includes (non-coding) regions that often control the expression (transcription/interpretation of the code and production of the coded proteins) of the gene. Genes have chemical markers to indicate where transcription should begin and end. Humans have 20,687 protein-encoding genes [23]. Identification of cancer genes has led to a deep understanding of the tumorigenesis process and has resulted in many changes in cancer biology.

DNA. A DNA is a molecule made of two strands of sugar-phosphate (nucleotides), which runs in opposite directions, and linked by four special molecules called nucleobases or bases: adenine (A), thymine (T), guanine (G), and cytosine (C). The four bases will only pair off in the following combinations: A with T and G with C. The pairing is by hydrogen bond interactions that span the double-stranded helix. DNA can make a copy of itself through a replication process. The original double-helix DNA (the parent) is unwound (by breaking the hydrogen bonds) into two strands, each with their bases attached. Since each base only pairs with one of the other 3 bases, a complementary strand can be constructed on each of the parent stands. As a result, one DNA replicates into 2 identical DNAs.

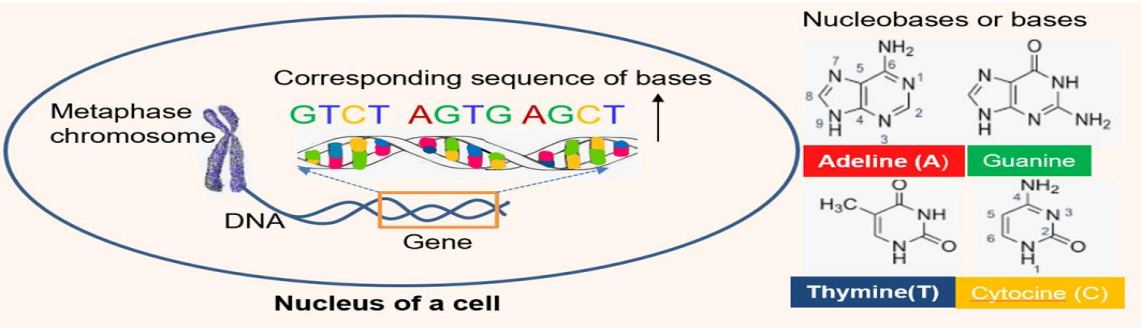


Figure 9. Chromosomes, Genes, DNA, and Nucleobases.

Cell. A cell is a base unit that contains its genome. The cell is the fundamental unit of living matter. The human body is made of over 200 types of cells. Cells are differentiated into different organ tissue types that perform specific functions for that organ. Cells not only perform a whole range of tasks required by the organ they belong to but they also contain all the DNA instructions for doing so. An organism is constructed from organ types. Structurally, all cells are bound by a plasma membrane. Contained inside the membrane is a fluid called *cytoplasm*, and within it, a nucleus and other components called organelles. Each component of a cell has a specific function. The nucleus contains the cell's genome and has its nuclear membrane with pores in it to allow small molecules to move between the nucleus and the cytoplasm. The organelles include ribosomes (for building proteins), endoplasmic reticulum (for transporting and storing molecules), Golgi apparatus (for modifying, packaging, and sorting molecules), vesicles (sacs for specialised processes), vacuoles (regions for storage and digestion), mitochondria (for producing energy used by the cell), and cytoskeleton (a network of protein filaments for maintaining the shape of the cell, anchoring components in place, and providing a basis for cell movement).

Cell cycle control and cell division process (Figure 10). Cells divide when body tissue is growing and when worn-out cells need to be replaced. The cell cycle is the highly regulated process by which a cell replicates itself, copies its genetic material (DNA) and divides into two identical daughter cells. Cell cycle control or the control of the cell division is singularly important, if functioning properly it prevents cell growth that leads to tumor. Cell cycle control is through G1/S, G2/M, and M checkpoints as explained in Figure 10.

The cell cycle consists of three phases: Interphase (G1 + S + G2), Mitotic phase, and Cytokinesis. Disruption of the cell cycle causes unregulated cell division that may promote excessive cell growth or even cancer [22, 10].

In the *interphase*, most somatic cells have two gap phases: G1 precedes DNA replication (S phase) and G2 follows the S phase. In the gap phase G1, the cell synthetizes proteins and organelles (such as mitochondria and ribosomes), grows in size, and ensures that both daughter cells will inherit sufficient amounts (or number) of proteins/organelles. At the *G1/S checkpoint*, the cell checks if it satisfies conditions concerning daughter cells and if it is ready for DNA duplication before it can commit to divide by going to the S phase. If the cell does not pass all these conditions, it leaves the cell cycle, entering a resting G0 phase. In the *S phase*, the DNA is duplicated (the cell's chromosomes are duplicated into two sister chromatids), along with the centrosome (the center supporting cell separation).

In the gap phase G2, the cell continues to grow, organizes the microtubules to form a mitotic spindle structure that pulls the chromosomes apart and prepares for mitosis (Figure 10). The *G2/M checkpoint* ensures that the DNA was completely copied with no damage. If errors are detected, the cell will stay in G2 until repaired; otherwise, it undergoes a programmed cell death (apoptosis).

In the *mitotic phase* (M phase) the cell stops growing and divides into two daughter cells. Mitosis is highly regulated and is conventionally broken down into five stages: prophase, prometaphase, metaphase, anaphase, and telophase and cytokinesis [10, 22]. During the *prophase*, the DNA condenses into recognizable chromosomes, the migration of centrosomes to both poles of the cell and the mitotic spindle (a series of specially synthesized protein filaments anchored to opposite sides of the cell) begins to form. During the *prometaphase* the nucleus membrane breaks down. Each of the two chromatids of each chromosome now has a kinetochore (a specialized protein structure) at its center. During *metaphase*, all the chromosomes align in the middle and for each chromosome, the sister chromatids are attached (at the kinetochore) to

the mitotic spindle coming from opposite poles. During the *anaphase*, the sister chromatids are separated from each other and are pulled towards the opposite ends of the cell. During the *telophase and cytokinesis*, the mitotic spindle is broken down, two nuclei forms, one for each set of chromosomes, the nucleus membranes reform and the cell's contents are redistributed, and finally the cell membrane is physically separated to form two new cells (in cytokinesis). There is an *M or Anaphase checkpoint* to ensure the proper assembly of the mitotic spindle and its attachment to each chromosome.

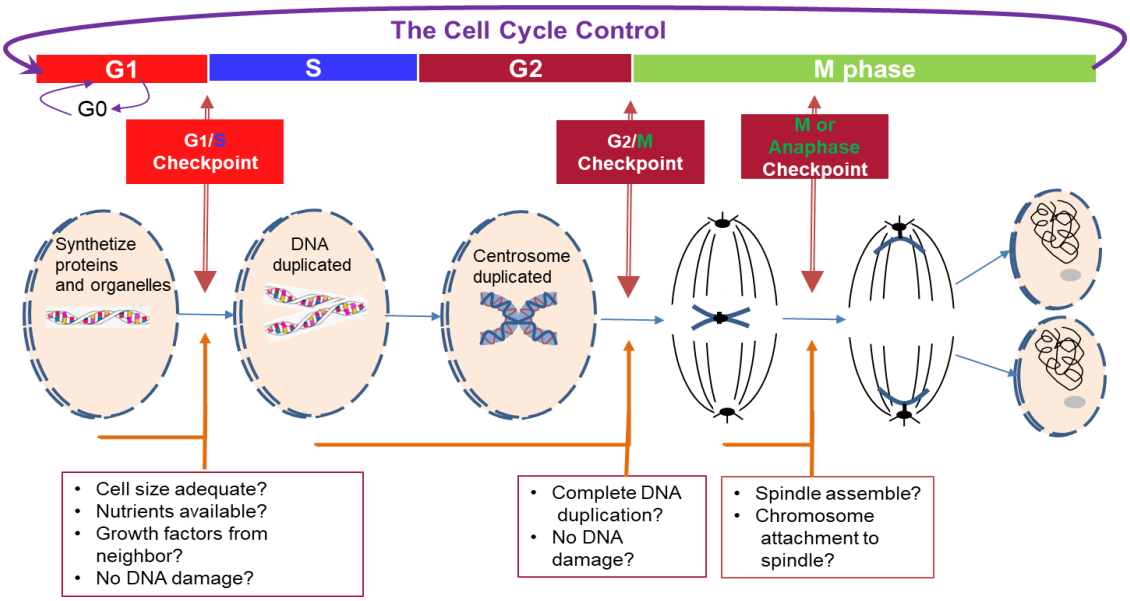


Figure 10. Cell cycle control and cell division process.

3.3. DNA sequencing

The fundamental unit of heredity, a gene, is a segment of the DNA molecule whose genetic instructions are coded in the order of A, T, G and C. Each gene has a unique ordered sequence of bases (see Figure 9).

Sequencing DNA means determining the order of the four chemical building blocks that make up the DNA molecule. That is to determine the order of DNA bases - the order of As, Ts, Gs, and Cs, which make up an organism's DNA. Sequencing is important as it tells us the genetic information that is carried in a particular DNA segment (or gene). For example, we can use sequence information to determine which stretches of DNA contain genes and which stretches carry regulatory instructions (turning genes on or off). Importantly, sequence data can highlight changes in a gene that may cause disease. In the DNA double helix, the four chemical bases always bond with the same partner (A to T, and G to C). This pairing is the basis for the mechanism by which DNA molecules are copied when cells divide, and the pairing also underlies the methods by which most DNA sequencing experiments are done. The human genome contains about 3 billion base pairs (6 billion bases) that provide the instructions for making and maintaining a human being. Each of us has the same set of DNA molecules (chromosomes) but we are different because our genes have different on-off patterns. Whether a gene is turned on (or expressed) depends partly on heredity and partly on the environment. If a

gene expresses, its DNA sequence ordering encodes information for producing some useful products for our body.

The Human Genome Project led to the completion of the DNA sequence for all human chromosomes in May 2006 [24]. The Next Generation Sequencing (NGS) and other sequencing technologies enable systematic mutational analysis of the cancer genome. Applications of sequencing technologies include mutational analysis, gene expression profiles for understanding cancer at the molecular level, and for providing prognosis information.

3.4. Gene expression

Gene expression is the process by which the instructions in our gene's DNA segment are converted and encode an RNA transcript, which is often translated into a functional product, such as a protein (Figure 11). RNA is a chemical similar in structure and properties to DNA, but it only has a single strand of bases and instead of the base thymine (T), RNA has a base called uracil (U). Gene expression is a tightly regulated process that allows a cell to respond to its changing environment. It controls when proteins are made and the amount of protein made. Key steps involved in gene expression include transcription, mRNA processing, translation, and posttranslational modifications such as phosphorylation, methylation, and acetylation [23].

Transcription. Transcription is when the DNA in a gene is copied to produce an RNA transcript called messenger RNA (mRNA). An enzyme called RNA polymerase uses available bases from the nucleus of the cell to form the mRNA. Along the mRNA, the DNA bases are arranged into codons, a triplet of bases that specifies a particular amino acid to be attached to a sequence of amino acids in a protein.

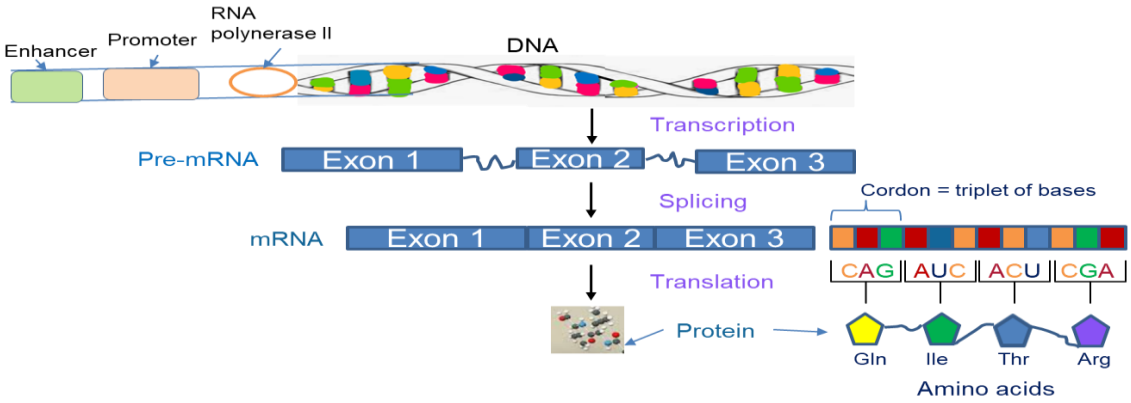


Figure 11. Gene expression process - Transcription, Splicing, and Translation to protein.

Translation. Translation occurs after the messenger RNA (mRNA) has carried the transcribed 'message' from the DNA to the ribosomes in the cell for making protein. The mRNA reads each codon at a time. The message carried by the mRNA is read by a carrier molecule called transfer RNA (tRNA). Each amino acid is attached specifically to its own tRNA molecule. When the mRNA sequence is read, each tRNA molecule delivers its amino acid to the ribosome and binds temporarily to the corresponding codon on the mRNA molecule. Once the tRNA is bound, it releases its amino acid and the adjacent amino acids all join together into a long chain called a polypeptide. This process continues until a protein is formed. By differential inclusion or exclusion of regions of pre-mRNA, a single gene can generate multiple spliced

messenger RNAs (mRNAs) called isoforms that generate multiple functional proteins (Figure 11). Exons refer to the portion of genes that are spliced together to form mRNA. Introns refer to the spacing regions between the exons that are spliced out of precursor RNAs during RNA processing.

3.5. Regulation of Gene Expression

The gene expression process is regulated by DNA-binding proteins (called transcription factors, TFs) that activate or repress transcription. A region near the transcription start site that contains binding sites for various transcription factors is called the promoter region. Changes in gene expression are also regulated by epigenetic mechanisms which modify gene expression without changes in the DNA sequence. Epigenetic mechanisms such as secondary modifications of DNA or histones (the proteins that packed and compressed DNA molecules) can result in the activation or silencing of gene expression or altering chromosomal loci [21], [23]. These mechanisms play a key role in genes involved in the formulation of tumors. Major epigenetic mechanisms include DNA *methylation, Histone methylation, and Histone acetylation.* DNA *methylation* involves the addition of a methyl group to cytosine of CpG dinucleotides in CpG islands in gene promoter regions. *Histone methylation* involves the addition of a methyl group to lysine residues in histone proteins. Methylations alter chromatin structure, making it either more open (allowing gene expression) or tightly packed (gene repression). *Histone acetylation* is another mechanism that results in an open chromatin configuration, which favors active transcription. Acetylation removes acetyl groups from histones.

Sequencing of tumour-derived RNA enables the identification of differentially expressed genes, gene fusions (a hybrid gene formed from two previously independent genes), small RNAs (short noncoding RNAs that can inhibit gene expression), aberrantly spliced isoforms (multiple proteins encode from a single gene due to variations in the splicing process of joining or skipping coding and noncoding portion of the pre-mRNA). Modifications of DNA or histones, and changes in chromatin structure can also be identified. Gene expression profiling and genome-wide sequencing approaches have allowed the understanding of cancer at the molecular level. It has been suggested that individualized knowledge of pathways or genes deregulated in a given tumor (personalized genomics) may provide a guide for therapeutic options on the tumor, thus leading to personalized therapy (also called precision medicine).

3.6. Mutations

A mutation can be defined as any change in the base sequence of DNA. Some mutations may be harmful, others may constitute an evolutionary advantage. Mutations occur when a sequence of coding DNA for gene expression is altered, resulting in a defective protein or the intended protein may not be produced at all. Mutations can involve the entire genome or structural alterations in chromosomes or individual genes. Sequencing technologies can reveal sequence mutations, small insertions and deletions, copy number alterations, structural rearrangements, and loss of heterozygosity in tumour DNA samples.

Mutations involving a single base-pair (or individual base substitutions) are referred to as point mutations. As each amino acid is specified by a codon consisting of three bases, if a single base is substituted the altered sequence may code for a different amino acid. If the DNA sequence change occurs in a coding region and alters an amino acid, it is a *missense mutation,* this may entail distinct phenotypes. However, many amino acids are encoded by more than one codon, if the point mutation does not alter the amino acid, it is a *silent mutation*. If the point mutation changes a protein's structure, it is a *nonsense mutation*. This type of mutation changes

a normal codon to an early stop codon, resulting in a shortened protein (i.e., UAC/UAU/UGC to stop codons UAA/UAG/UGA).

Polymorphisms are sequence variations that have a frequency of at least 1 %. Usually, they do not result in a perceptible phenotype. *Single-nucleotide polymorphisms (SNPs)* often consist of single base-pair substitutions that do not alter the protein coding sequence. Mutations may involve *insertions and deletions* of DNA sequences. The addition of a single base will lead to a frameshift and each subsequent codon is read wrongly *or the left-over bases cannot form a codon with just one or two bases.* Larger insertions or deletions may affect a portion of a gene or an entire gene as well as potentially causing a frameshift. *Unequal crossing-over mutation* is when mispairing of homologous sequences leads to unequal crossover, with gene duplication on one of the chromosomes and gene deletion on the other chromosome. *Errors in DNA repair are* mutations caused by defects in DNA repair when somatic cells divide. *Splicing mutations:* Mutations of sequences required for splicing may alter the protein product or the expression level of a gene.

Copy number variations. Copy number variations (CNVs) are relatively large genomic regions that have been duplicated or deleted on certain chromosomes. It has been estimated that as many as 1500 CNVs [9], scattered throughout the genome, are present in an individual. When comparing the genomes of two individuals, approximately 0.4 - 0.8 % of their genomes differ in terms of CNVs. Some CNVs have been associated with susceptibility or resistance to disease, and CNVs can be elevated in cancer cells.

Functional Consequences of Mutations. Functionally, mutations can be broadly classified as gain-of-function and loss-of-function mutations. Gain-of-function mutations typically confer an abnormal activity on a protein or produce a new trait (phenotypic alteration). An increase in gene expressions and hence gene products may also result in disease. Loss-of-function mutations prevent the normal gene product such as a protein from being produced or renders it inactive. A nonsense mutation is an example of a loss of function mutation that causes termination of the amino acid chain during gene translation. Mutations in introns (non-coding portions of an mRNA) or in exon (coding portion) junctions may produce splicing mutations. Mutations may also be found in the regulatory sequences of genes, resulting in reduced or enhanced gene transcription [9]. The advancement of techniques for genome-wide expression profiling and mutation analyses has provided a detailed picture of the molecular defects present in individual tumors.

3.7. Molecular and cell biology of cancer

In this section we study cancer from the molecular standpoint to understand the tumorigenesis process. In particular, we explore various strategies which tumor cells employ to sustain their growth and spread to other parts of the body. The aim is to appreciate various issues associated with the disease from its detection, diagnosis, to prognosis and treatment. Cancer is characterized by unregulated cell growth, avoidance of cell death, tissue invasion, tumor formation, and the ability to metastasize [25]. Although tumors have a clonal origin, they are not a homogeneous mass. Even identical cancer cells, sharing the same genome, express different profiles and exhibit multiple functional traits.

3.7.1. Cancer and mutation.

Cancer is caused by a stepwise degradation of normal cell behaviour by mutation and as a result, cancer cells grow out of control, leading to mortality if cancer spreads to other parts of the body. This implies that the mechanism that regulates the growth of cells fails due to mutation.

The two classes of genes responsible for the regulation/deregulation are oncogene and tumor suppressor genes. Proto-oncogenes promote growth when the body needs it, but when mutated they become **oncogenes** and allow unregulated cell growth. **Tumor suppressor genes** code for proteins to block cell growth, but when they are mutated, they become inactive and lose their designated function. **Activating** *oncogenes* or **inactivating** *tumor suppressor genes* are achieved by: (i) point mutations that may lead to hyperactive proteins versions or lead to truncated versions of tumor suppressor and rendering them inactive; (ii) gene amplifications that lead to protein overproduction; (iii) deletions that lead to a loss of function phenotype; (iv) chromosome rearrangements that generate fusion proteins that can hyper-activate the protein, or that relocate in the genome new regulatory units that lead to *overexpression of the oncogene* or that lead to a *reduction of expression of tumor suppressor genes* or its inactivation. Another class of genes that play a vital role in maintaining the stability of the genome and mitigating the accumulation of harmful mutations in other genes leading to the onset of cancer are **caretaker genes**. Their main role is to repair DNA damage. Damage can include chemical modification or loss of DNA bases, or single strand or double strand breaks. Each type of damage can lead to serious mutations during the critical phase of DNA replication and recombination. DNA damage can override cellular checkpoints and allow unchecked cell cycle progression [21]. Two of the best-studied caretaker genes are *mlh1* and *msh2*, which repair mismatch DNA bases. The tumor suppressor gene p53 also has caretaker capability. Mutations in these genes greatly increase the rate of point mutations in genes.

3.7.2. Mechanisms for increasing cell number.

The most important fact about cancer is that "*to form a tumor mass a cancer has to increase the number of cells*." The cancer cell achieves this goal through a number of mechanisms: Increasing cell division, stopping cell death, blocking cell differentiation, and becoming immortal.

To *increase cell division*, cells must receive proliferative signals to overcome a number of safeguards at the checkpoints in the cell cycle before the division can proceed. These safeguards include the CDK (cyclin-dependent kinase or enzyme) inhibitors or necessary cyclin proteins essential for cell division. Cancer cells possess mutations that break this cell cycle control, allowing cells to divide regardless of any safety checks [22].

Another way to increase cell numbers is *to stop cell death*. This occurs when a tumor suppressor gene is mutated or the cancer cell overexpresses anti-cell death proteins. For example, the p53 tumor suppressor, once mutated, can no longer trigger apoptosis (cell death) or some cancer cells overexpress *Bcl2*, an anti-apoptotic protein, which allows cells to avoid apoptosis.

Blocking differentiation is another strategy to block or change the differentiation program in order to increase the cell number. During embryonic development, the egg has to differentiate into all the different cells and tissues that form our organs in an organized manner. This regulated program is tightly controlled but if it is blocked cells proliferate.

Becoming *"immortal*" is another way cell numbers increase because "immortal" cells live indefinitely. Normal embryonic human cells canonly divide a finite number of times in culture (~ 60 times) and then stop when the length of the telomere (the end-parts of the chromosome) has reduced to zero. Each time a cell divides, the telomere loses a small amount of DNA and becomes shorter. This limit was named replicative senescence (RS). However, embryonic germ cells and most cell lines derived from tumors can divide indefinitely. Most tumor cells avoid this

shortening process and become immortal by overexpressing an enzyme called Telomerase Reverse Transcriptase for telomere synthesis.

Beyond the four mechanisms for increasing cell numbers, cancer cells employ additional mechanisms to feed, sustain, and spread themselves. These include developing their own metabolism mechanism and hijacking the host's mechanisms.

Cancer Metabolism. To sustain and increase the number of cells, energy and nutrients must be obtained to support growth. Cancer employs a special metabolism mechanism to sustain and promote tumor progression. Tumors metabolize glucose to lactate in aerobic conditions to produce new protein-biomass rather than more energy. The by-products of this metabolic mechanism lead to immune suppression and angiogenesis, promoting tumor progression.

Hijacking the Host. Tumor cells even hijack the host and deploy the host's various mechanisms to feed themselves, to evade the host's immune system, and to migrate and spread their tumors and hence promote cancer progression.

Immune Evasion. Tumor cells develop mechanisms to avoid detection by the host's immune system. They also hijack and corrupt the immune system to provide growth factors, inducing angiogenesis or helping tumor cells to metastasize.

Angiogenesis. To grow more than 1 mm in diameter [22], tumor cells have to recruit blood vessels to bring nutrients and oxygen to feed themselves. They establish mechanisms to induce this angiogenesis process.

Metastasis. Tumor cells exploit the embryonic strategies to migrate and invade other organs. The process is called metastasis and involves several steps: (i) invasion of surrounding tissues; (ii) intravasation into blood vessels; (iii) survival in the circulation; (iv) extravasation from the blood vessels; and (v) survival and proliferation at a secondary site.

Genomic Instability. Besides acquiring oncogenes and mutated tumor suppressor genes, tumor cells may also acquire mutations in caretaker genes and together the tumor cells accumulate more mutations. Consequently, the genome becomes unstable and generates sub-clonal diversity of the tumor environment.

Remark on cancer genetics. All cells are expected to obey a "cell cycle control" that regulates the process of cell division (Figure 10). The disruption of the cell cycle control process is the prime reason for unbounded cell growth leading to cancer. According to Schulz [21], "A cancer pathway is a cellular regulatory system whose activation or inactivation by a genetic or epigenetic mutation is essential for the development of at least one human cancer. Typically, cancer pathways become evident by alterations in different components of the same regulatory system in individual cases of one cancer type or in distinct cancers". With this knowledge of molecular and cell biology, it may be possible to identify pathways and their internal and external associated components, related to a particular cancer. If this is achieved cancer prevention or the stopping of cancer progression should be feasible. Extending the powerful mechanism of deep learning may be the right tool for tackling this problem.

4. DEEP LEARNING AS A VALUABLE TECHNOLOGY FOR CANCER

The section discusses the defining characteristics of deep learning that make it a valuable technology for addressing many issues related to cancer. An example is provided to illustrate the general method of how deep learning can be applied to oncology. This section also provides pointers to DL and data resources for readers who wish to engage in application implementation.

4.1. Deep learning as a problem solver for cancer oncology

Deep learning methods, such as those used by DeepMind's AlphaGo [26], [3] and object recognition [4] exceed human performance in visual tasks and are flexible and powerful analytical techniques for dealing with complex problems.

With today's high-throughput Next Generation Sequencing, diagnostic and imaging technologies we are overwhelmed with terabytes of multi-omics data of patient samples (i.e., genomics (whole genome data, single nucleotide polymorphism), gene expressions (mRNA, miRNA), proteomics and epigenetic (methylation and other chromosomal modifications)), millions of disease images (histopathology whole slide images, digitized film mammograms, expression data embedded into a 2-D images) and large number of clinical parameters. In these situations, human perception, statistical methods, and conventional machine learning approaches do not generally work. Deep learning leverages this explosion of data, since it is proven as an excellent method for processing vast numbers of images for identification and classification. It has demonstrated a powerful capability of extracting hierarchical presentation features of complex scenarios through processing of high dimensional and large amounts of data such as genomic data, gene expression data, mutation data, and cancer drug sensitivity data. Existing methods, however, fail on these complex problems in one or more aspects: i) they are not able to handle extremely large quantities of data, ii) they are not well-equipped to handle multi-label classification of data because information of classes may not be mutually exclusive, iii) they are not able to extract hierarchical features and building blocks required for recognition of complex and high dimensional objects, and iv) they rely on domain experts to extract features for classification and decision making.

For medical data, patients could have symptoms of multiple different diseases at the same time and it is important to develop tools that help to identify these multi-label classification problems early. Deep learning has excellent capability to integrate multi-omics data, clinical information and cancer images into its architectures for diagnosis and drug response prediction. Deep learning methods effectively handle different types of data formats that often coexist in healthcare applications.

Convolutional neural networks can extract hierarchical features through hidden layers constructively, especially for complex object recognition. It is clear that deep learning for pattern recognition can be applied to cancer detection and gene identification [27] with excellent performance as data for these applications can be mapped to image features. For example, a gene expression profile represents the state of a cell in the same way that patterns of pixels represent the content of an image.

Deep learning with recurrent neural networks is excellent for predicting sequential events such as natural language processing. This method clearly applies to cancer prognosis prediction [28] assignaling pathways in cancer biology present sequential events where a downstream event is related to upstream events.

Deep learning eliminates feature engineering that requires domain expertise to extract features for classification that is laborious and time-consuming. Deep learning can learn representative features automatically and directly from the raw input examples such as images of tumour tissue obtained from cancer patients, genomic data from DNA-sequencing and gene expression data from RNA-sequencing for cancer detection, classification, diagnostic, prognosis, and treatment purposes. Deep learning with autoencoders can extract features and reduce the dimension of data automatically without domain expert's input.

With deep learning, transfer learning can be used to avoid laborious and expensive training from scratch as a deep neural network trained on a large-scale dataset from a different domain can be used for pattern recognition in a new deep network application with fine tuning training on data pertaining to the new application. Furthermore, one advantage of using deep learning to train a model is its capability to continue training when more data is available.

As mentioned earlier, we are still far from finding out the root cause of many problems including cancer. Gene mutations may lead to tumors but may not be the cause of cancer [1] as the process from mutations to malignant tumors is not completely understood. Deep learning, with its excellent pattern recognition capability, presents itself as a powerful technology not only for oncology but also for detecting the causes or the triggering point to cancer.

4.2. An example of the application methodology

In this section, an application is described in detail to familiarize the readers with the methodology used by most applications of deep learning to issues related to cancer oncology. The selected application [29] is entitled "Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature". *We aim to preserve the spirit of the original paper by keeping close to the reported statements by the authors.*

Motivation for this work. This work takes genomics into consideration in developing their deep learning models becausethe majority of the deep learning-based drug development focuses on the prediction of drug-target interaction, based on molecular structures.

Aim. This study developed CDRscan, a deep learning model that predicts anticancer drug responsiveness based on large-scale drug screening assay data. Specifically, it predicts somatic mutation profile-based drug responsiveness by linking the tumour genomic fingerprint and its sensitivity to drugs.

Deep learning methods and model architecture. The application employs five CNN architectures to estimate the essential features of the genomic mutations of cancer cell lines and molecular drug features and combine them in a predictive model for predicting the half-maximal inhibitory concentration (IC_{50}) values of anticancer compounds from the genomic signature of tumour samples. All models generated predicted IC_{50} values across the 244 anticancer drugs for each cell line as a final output layer of the models. The average of the five values predicted by each model was then reported as the final outcome of CDRscan.

The datasets and data pre-processing. The datasets used to train CDRscan were extracted from COSMIC cell line project (CCLP) and GDSC databases. The CCLP contains various types of molecular profile data, including the whole exome sequencing data of 1,001 human cancer cell lines. The GDSC provides IC_{50} values from drug sensitivity assays for over 200,000 drug-cancer cell line pairs. In GDSC, the identical set of 1,001 cell lines characterised by CCLP was used and the IC_{50} values of 265 anticancer drugs were measured from the treatment of these cell lines.The datasets from the databases contain 686,312 mutation positions from 1,001 cell lines and 265 drugs, covering 30 cancer types as defined by The Cancer Genome Atlas (TCGA) studies. A subset of the data is selected that includes only gene mutations contained in Cancer Gene Census, which is a catalogue of 567 genes strongly associated with cancer pathology. The datasets also exclude the cancer types that have fewer than 10 different cell lines, drugs without PubChem Compound Identifier, and drugs with molecular weight greater than 1000 g/mol. The final datasets yielded a total of 152,594 instances which contained 787 cell lines across 25 TCGA cancer types, mutation information at 28,328 base positions in 567 genes, and IC_{50} measurements of cell line-drug treatment in 244 drugs. The input features of the entire instance were represented by 31,400 binary digits. Of these, 28,328 bits represented mutational status of

28,328 genomic positions in each of the 787 cell lines, while 3,072 bits encoded molecular profiles of the individual drugs [30].

Model training. 144,953 instances were randomly selected to train the models (95 % of the total 152,594 instances). To ensure that all 25 cancer types are represented equally in the training set, 95 % of the instances are randomly chosen from each cancer type. As a result, 25 subsets were created and subsequently compiled as a single training set. The remaining 5 % of the instances of individual cancer types were set aside to be used as test sets, both as 25 separate lists and as one consolidated list. To prevent overfitting, i) Three to four dropout layers were applied. In these layers, a subset of parameters (10 - 20 % of the total parameters) were randomly selected and ignored during training, making it less likely to overfit the training data, ii) Maxpooling layers were used to reduce dimensionality of the input, and iii) The performance score of CDRscan was measured by five-fold cross validation.

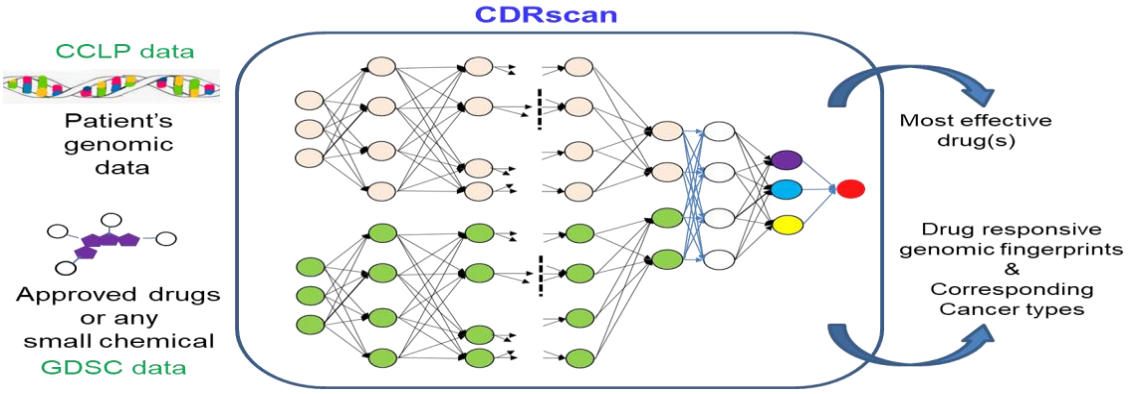


Figure 12. CDRscan architecture.

Performance evaluation. The experimentally obtained (observed) IC_{50} values and their counterparts predicted by CDRscan were plotted on a natural log scale. A coefficient of determination (R^2) [29] is used as the measure of prediction accuracy. In the drug-centric evaluation of CDRscan performance, AUC was computed for the compiled training set.

The observed and the predicted IC_{50} values showed a strong agreement with the mean coefficient of determination (R^2) value of 0.843, ranging from 0.838 to 0.853 across five models, confirming that the prediction was accurate in most instances. To further confirm the prediction accuracy of CDRscan, the area under the receiver operating characteristic curve (AUC) score of 0.98 was obtained for the test. The R^2 values of individual cell lines (785 in total) and drugs (244 in total) were assessed. Consistent with the high R^2 values, the predicted and observed IC_{50} values showed strong correlation across all cell lines. In the drug-centric correlation analysis, *dasatinib* (tyrosine kinase inhibitor drug) had the highest mean R^2 of 0.902 (n=288), and *bicalutamide* (androgen receptor inhibitor) had the lowest.

The authors compared the performance of CDRscan and a previously developed prediction model using the same databases [31]. It was confirmed that CDRscan exhibited significantly higher performance than the previous model ($R^2 = 0.843$ versus $R^2 = 0.72$).

Feasibility of drug repurposing using CDRscan. Thirty seven of the 102 approved anticancer drugs had the potential for new cancer type indications. In addition, 176 of 1,385 approved non-oncology drugs had the potential anticancer activities in addition to their original

drug indications. The number of approved oncology drugs with repurposing potential was 23. Nine of these 23 drugs had CDRscan-predicted anticancer activity against more than 90 % (23/25) of the total types, suggesting a universal antiproliferative/cytotoxic activity of the compounds.

From this review, it is clear that there are many important considerations for a successful application. The deep learning method chosen needs to be carefully selected to match the expected features of the problem. Not only the datasets but also the data pre-processing must be meticulously prepared to achieve a quality solution. The training, testing, and validating regimes have to be designed and executed methodologically to ensure reliable results regardless of the selected neural network architectures and the distribution of the chosen datasets. Furthermore, the performance analysis and evaluation must be extensive to demonstrate the quality of the results.

4.3. Resources for Deep Learning and Datasets

Deep learning resources. We recommend the following resources for deep learning. Zou [32] provides a simple explanation and lists many useful and practical resources for implementing deep learning applications. Zou also provides an interactive tutorial to build a convolutional neural network to discover DNA-binding genomic sequences that specifically bind to transcription factors. Goodfellow [33] provides an excellent resource for deep learning researchers. Sutton's book [17] is a comprehensive text in reinforcement learning. Burkov [34] provides a good resource for deep learning engineering. To get hands-on experience with practical deep learning, the book by Géron is recommended [35].

- Zou, J., Huss, M., Abid, A. *et al*. (2019). A primer on deep learning in genomics. Nat Genet 51, 12–18. https://doi.org/10.1038/s41588-018-0295-5
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. The MIT Press, Cambridge, MA, USA. ISBN: 978-0262035613
- Sutton, R. S. and Barto, A. G. (2018). Reinforcement Learning: An Introduction. Second edition, The MIT Press, Cambridge, Massachusetts, London, England.
- Burkov, A. (2020). Machine Learning Engineering. True Positive Inc. ISBN 978-1-7770054-5-0
- Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd edition, O'Reilly Media.

Clinical and molecular data resources. Most of the applications of deep learning to oncology use information from the following publicly available databases.

- TCGA, The Cancer Genome Atlas (https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas), is a database that stores clinical and molecular data of over 11,000 tumor patients across 33 different tumor types, including genomic whole genome and/or exome sequencing, WGS/WES), transcriptomic (RNAseq, small RNAseq), epigenomic (methylation) and proteomic profiling (reverse-phase protein arrays, RPPAs) data [36], [37].
- CCLE, The Cancer Cell Line Encyclopedia project (https://sites.broadinstitute.org/ccle/) compiled genomic profiles of 947 human cancer cell lines, and pharmacologic profiles of 24 anticancer drugs across 479 cancer cell lines to benefit personalized medicine [38].
- GEO, Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), is a public searchable online data repository storing microarray and next-generation sequencing (NGS) data, as well as other high-throughput functional genomic datasets, such as genome methylation, chromatin structure, genomic mutation/copy number variation, protein profiling, and genome–protein interactions [39], [40].

- GTEx, Genotype-Tissue Expression (https://gtexportal.org/home/), contains whole-genome sequencing and RNA-sequencing profiles from ~960 post-mortem adult donors of many tissue samples that have tissue images stored in an image library for public access [41].
- GDSC, Genomics in Drug Sensitivity in Cancer (https://www.cancerrxgene.org), is a publicly available database providing experimentally measured drug sensitivities of 1,001 human cancer cells against 265 anticancer compounds.
- COSMIC, the Catalogue Of Somatic Mutations In Cancer (http://cancer.sanger.ac.uk/), is the world's largest and most comprehensive database of somatic mutations in human cancer from the Welcome Trust Sanger Institute.

5. REVIEW OF DEEP LEARNING APPLICATIONS TO DRUG RESPONSE PREDICTION, CANCER DIAGNOSIS, PROGNOSIS, TREATMENT, CANCER RISK PREDICTION

The formation process of a tumor (tumorigenesis process) is driven by mutations in tumor suppressors, oncogenes, caretaker genes and alterations in epigenetic regulation. The process brings about changes in gene expression. The knowledge allows the identification of differentially expressed genes and the understanding of the complex molecular mechanisms regulating normal and cancerous behaviors. Studies on molecular profiling of tumors have suggested methods for distinguishing tumors of various biological behaviors (molecular classification), elucidating pathways relevant to the development of tumors, and identifying molecular targets for the detection and therapy of cancer. High-throughput sequencing such as microarray techniques have enabled high-throughput DNA sequencing and RNA sequencing, and whole genome sequencing of cancer cells, leading to gene expression profiling and molecular profiling of tumors. Consequently, various omics datasets on mutation, gene expression, proteomic, and drug sensitivity are available. Combined with advances in cancer screening methods such as MRI, PET, etc., vast amounts of cancer images and traditional clinical diagnostic data are equally important and available for tackling cancer issues. Together with the advances of deep learning, these technologies have practical applications in oncology with information not available from traditional clinical tests alone. In this section, we provide a review of deep learning applications organized according to their purpose: i) drug response prediction, ii) classification, iii) diagnosis prediction, iv) prognosis prediction, v) treatment, and vi) cancer risk prediction. For each selected application, we highlight the problem leading to the development of the application, the relevant datasets, the deployed deep learning method, and its performance. Due to the limited space, we will only present substantial reviews of recent representative applications from 2016. However, for each category we will cover each application in some depth in order to gain an understanding of the issues underlying the application. We highlight the unique issue addressed by the application, the deep learning method employed for extracting features, and the significance of the specific data type of the selected datasets. We emphasize the importance of the preparation/pre-processing of data for training as well as the training, testing, and validating regimes. We leave out many applications that mainly use histology and image-based data modalities. Readers should refer to [42] for a survey on deep learning for image-based cancer detection and diagnosis and [43] for early cancer detection.

5.1. Deep learning for drug response prediction

Since most tumors are heterogeneous and molecular classification of cancer can reveal similarities in tumors of diverse tissue types, the precision medicine approach may be more effective than conventional therapeutic approaches when tumor resistance is an issue. This is particularly the case if tissue-specific mutations or pathways can be identified and targeted with a specific drug. Gene expression also offers the potential to predict drug sensitivities as well as provide prognostic information.

Due to essential differences between cell lines and tumors, to date the translation into predicting drug response in tumors remains challenging. Chiu *et al*. [44] tackled this problem by proposing DeepRD, *a deep learning model to predict drug response based on mutation and expression profiles of a cancer cell or a tumor.* The complete deep learning model is composed of 3 networks: i) a 4-layer mutation encoder, ii) a 4-layer expression encoder, and iii) a 5-layer prediction feedforward network (P). The two autoencoders are for feature extraction and dimension reduction of the mutation data and the expression data, the prediction network is for drug sensitivity prediction. The gene-level expression data of 935 cell lines of the Cancer Cell Line Encyclopedia (CCLE) from the CTD Data Portal [45] and 11,078 TCGA pan-cancer tumors from the UCSC TumorMap [46]. The drug response data of 990 CCLE cell lines to 265 anti-cancer drugs measured by the half maximal inhibitory concentration (IC_{50}) is from the GDSC Project [47]. In this study, the authors analyzed 622 cell lines with available expression, mutation, and drug sensitivity data (IC_{50}) and 9,059 tumors with expression and mutation profiles. Given a pair of mutation and expression profiles, the model predicts IC_{50} values of 265 drugs. They trained and tested the model on a dataset of 622 cancer cell lines and achieved an overall prediction performance of mean squared error at 1.96 (log-scale IC_{50} values). The performance of the model was evaluated using the testing samples for cell lines and then the final model was applied to predict drug response of TCGA tumors using tumor mutation data and gene expression data to determine the drug response (IC_{50}) of tumors. The performance was superior in prediction error or stability than two classical methods (linear regression and support vector machine) and four analog DNN models of DeepDRand built on individual types of input data. The model is then applied to predict the drug response of 9,059 tumors of 33 cancer types. Using per-cancer and pan-cancer settings, the model predicted both known and novel drug targets. The Chiu *et al*. analysis was comprehensive and novel resistance mechanisms and drug targets were identified.

In [48], the authors proposed DeepDSC, *a deep learning model, to predict drug sensitivity of cancer cell lines. An autoencoder is used to extract cell lines feature from gene expression data* in an unsupervised way. It also serves as a dimension reduction of the data from 20,000 to 500. This model is then integrated with the chemical features of compounds into a fully connected feedforward neural network to produce drug sensitivity output to given cell line-drug pairs. The expression profiles of cell lines and drug sensitivity data were collected from two public datasets, CCLE and GDSC. The compound chemical structure files were obtained from PubChem [49]. The *gene expression data* (CCLE) of cell lines contain the transcript level of about 20,000 genes, corresponding to a vector of the same length. The *drug response data* was extracted from CCLE dataset (504 cell lines) against 24 drugs. The final data matrix contains 491 cell lines and 23 drugs and 10,870 data points. For the GDSC dataset, the final data matrix contains 655 cell lines and 139 drugs and 73,075 data points. Using the GDSC dataset, DeepDSC was compared with three previous studies, using the same performance metrics, Root Mean Square Error (RMSE) and coefficient of determination R^2. Overall, DeepDSC outperformed state-of-the-art methods with the lowest prediction errors and high coefficient determination. DeepDSC had the best interpolation ability to fill in missing drug sensitivity values.

5.2. Deep learning for cancer classification

Different subclasses of cancer exist and may require different treatments. Molecular analyses have revealed that a cancer which appears uniform by morphological criteria can be further differentiated [21]. The authors of the study [50] address two challenges: reducing the dimensionality of the feature space in a way that ensures that sufficient information is retained to perform accurate classification while dealing with the problem of model overfitting due to limited number of training examples. The application proposed a general method for extracting relevant features from gene expression data, including for breast cancer, leukemia, colon cancer, prostate cancer, and ovarian cancer. The application provides cancer detection and cancer type analysis using the unsupervised learning autoencoder deep learning method. It uses a PCA-based method to reduce the dimension of the feature space and deploys an AE to learn an essential feature representation from unlabeled data. Reducing the dimension of the feature space is essential as the dimensionality of gene expression data is extremely high - in the order of 20,000 to 50,000. The AE is essential to capture the non-linearity of the relations between expressions of different genes using unlabeled data that may consist of data from different tumor cells but containing the same gene expression information. The output of the feature learning phases is fed into a softmax regression classifier for classification. The proposed algorithm is compared against two baselines: SVM with Gaussian kernel, and softmax regression algorithms. It was concluded that the proposed method outperforms the baseline algorithms which do not use unsupervised sparse features in terms of cancer classification accuracy. Importantly, the authors addressed the problem of limited data by allowing data from different cancers and other tissue samples to be used during feature learning independently of their applicability to the final classification task. They also demonstrated that deep learning has the ability to deal with gene expression data across different cancer types.

Among skin cancers, melanoma is one of most deadly, yet its identification is challenging due to high similarity between different skin lesions in terms of the color images of skin. The research work [51] addressed this challenging problem and proposed an automated skin lesion classification method to resolve this issue. A modified AlexNet was used for this purpose. AlexNet is a CNN with 5 convolutional layers and 3 fully connected layers that were used in the visual recognition of ImageNet. The last layer of AlexNet was replaced with a softmax classifier to classify three different skin lesions: melanoma, common nevus, and atypical nevus, instead of attempting to classify 1,000 classes as in the original ImageNet. The dataset for this application includes medical annotation of all the images namely medical segmentation of the lesion, clinical diagnosis and dermoscopic criteria (asymmetry, colors, and the presence of typical and atypical differential structures). The dataset consists of 200 RGB color images. It was divided into three classes: melanoma (40), common nevus (80), and atypical nevus (80). The *augmentation process was performed to overcome the lack of data*. To overcome the need for a huge number of labeled images to build a successful deep neural network, *transfer learning* and image augmentation are applied to a pre-trained AlexNet. The weights were updated using a stochastic gradient descent algorithm. The performance of the proposed method is compared with the existing methods (cited in the paper) that used the same dataset of skin lesions. Four performance measures, accuracy, sensitivity, specificity, and precision have been computed for the comparison; the achieved rates are 98.61 %, 98.33 %, 98.93 %, and 97.73 %, respectively. These results demonstrate that the proposed method outperforms the compared methods.

In [52] *DeepGene was proposed for somatic point mutation based cancer type classification*. The aim is to improve the classification performance and address the three obstacles in existing somatic point mutation based cancer classification (SMCC) studies: i) high

data sparsity where only a small discriminatory subset of genes is related to the cancer classification task but normal sequencing results include tens of thousands of genes, ii) a small sample size where even within the discriminatory subset, the majority of genes are not guaranteed to contain informative point mutations and often remain normal (i.e. zero values in the data), and iii) the application of simple linear classifiers where conventional machine learning is not effective since different genes related to specific types of cancer are generally correlated and have complex interactions. DeepGene is a deep neural network (DNN) based classifier, that consists of three stages: i) a clustered gene filtering is used to cluster the gene data using mutation occurrence frequency and filter out the majority of irrelevant genes, ii) an indexed sparsity reduction converts the gene data into indexes of its non-zero elements to suppress data sparsity, and iii) a DNN classifier that uses the filtered and indexed data for gene classification. The DeepGene dataset is a reformulated subset of The Cancer Genome Atlas (TCGA) dataset containing 12 selected types of cancer. The data is collected from the TCGA database with filter criteria IlluminaGA_DNASeq_Curated. Following this pre-processing a TCGA-DeepGene subset was obtained, where each sample (column) is assigned one of the labels {1, 2, …, 12} for each of the 12 types of cancer. A convolutional neural network with 4 hidden layers and 8,192 parameters per layer was used using the MATLAB MatConvNet toolbox. For comparison purposes, Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Naïve Bayes (NB) machine learning algorithms were selected. All methods use raw gene data as inputs. DeepGene outperforms these classifiers and achieves at least a 24 % performance improvement in terms of testing accuracy.

The conventional method used in standard practices of RNA-Seq analysis is to match the tumor samples to the normal samples, both being from the same tumor type. Using such methods fails to differentiate tumor types due to the lack of knowledge of other tumor types. The authors of [53] addressed this problem by proposing a new method to discover potential biomarkers for each tumor type. The authors used a convolutional neural network to classify tumor types based on the genomics data. *The method embedded the high dimensional RNA-Seq data into 2-D images* and used a convolutional neural network to do the classification of the 33 tumor types. By identifying significant genes and using the KEGG pathway analysis, the related genes in these pathways can then be viewed as tumor specific biomarkers for each tumor type. A convolutional neural network, consisting of three convolutional layers, and three fully connected layers, was used. The RNA-Seq gene expression data of 33 tumor types in Pan-Cancer Atlas was used. The data contains 10,267 tumor samples with respect to 20,531 genes. The high-dimension expression data (10381x1) was embedded into a 2-D image (102x102) to be suitable for the convolutional layers. With the trained neural network, heat-maps for all the classes were generated to identify high intensity genes that dominate the final classification. By functional analysis, it was validated that the top genes selected in this way are biologically meaningful for corresponding tumors. The proposed method achieved a tumor type classification accuracy of 95.59 %, higher than other works using the GA/KNN method on the same dataset.

5.3. Deep learning for cancer diagnosis

Although imaging and histopathology have become more sophisticated, they are limited by the size of a tumor. Developments in physics and information technology are revolutionizing imaging techniques, advances in computing enable localization of very small tumors by virtualizing tomography data [21]. Early detection of breast cancer on screening mammography is a challenging classification task because the tumors themselves occupy only a small portion of the image of the entire breast. A cancerous region of interest can be as small as 100 × 100 pixels

out of a full-field digital mammography (FFDM) image of 4000 × 3000 pixels. A deep learning algorithm was developed [54] for detecting breast cancer on screening mammograms using a training approach in which a model to classify local image patches is pre-trained using a fully annotated dataset with region of interest information. The patch classifier's weight parameters are then used to initialize the weight parameters of the whole image classifier, which can be further fine-tuned using datasets without region of interest annotations. The approach leverages training datasets with either complete clinical annotation or only the cancer status of the whole image. The 16-layer VGG (VGG16) and the 50-layer ResNet (ResNet50) CNN networks (VGG and ResNet are variants of the CNN architecture [55]) were used as patch classifiers. The total numbers of images in the training, validation and testing sets were: 1903, 199 and 376, respectively. The test set of digitized film mammograms from Digital Database for Screening Mammography (DDSM) showed the best result with AUC of 0.91 (sensitivity: 86.1 %, specificity: 80.1 %). The validation set of FFDM images from the INbreast database showed the best result with a per-image AUC of 0.98 (sensitivity: 86.7 %, specificity: 96.1 %). It was reported that convolutional network methods for classifying screening mammograms attained excellent performance in comparison with previous methods (cited in their paper). The authors demonstrated that a whole image classifier trained using their end-to-end approach on the DDSM mammograms can be transferred to INbreast FFDM images using only a subset of the INbreast data for fine-tuning and without further reliance on the availability of lesion annotations.

Much of the work on the identification of differentially expressed genes has focused on the most significant changes, and may not allow recognition of more subtle patterns in the data. As reported from [27], this issue was addressed through a deep learning approach for cancer detection and relevant gene identification. In particular, they aimed to i) use a deep learning architecture, an autoencoder to extract meaningful features from gene expression data, ii) evaluate the performance of the extracted representation through supervised classification models, and iii) discover highly relevant genes that could play critical roles and serve as clinical biomarkers for cancer diagnosis. The performance of the extracted features is evaluated through supervised classification models such as a shallow artificial neural network (ANN) and an SVM model, to verify the usefulness of these features in cancer detection. The weights of the model were used to extract genes for cancer prediction. An AE was selected with four layers of dimensions of 15,000, 10,000, 2,000, and 500. The encoded features of the AE were used as input features to the classification algorithms. *RNA-seq expression data* from The Cancer Genome Atlas (TCGA) database was analyzed for both tumor and healthy breast samples. These data consist of 1,097 breast cancer samples, and 113 healthy samples. The results and analysis illustrate that these extracted genes could be useful cancer biomarkers for the detection of breast cancer. However, for deep learning, the need for large data sets may not be available for cancer tissues. The authors stated that further analysis on the identified genes is needed since it can potentially improve methods for cancer diagnosis and treatment.

Gastric cancer (GC) is characterized as an aggressive malignancy which is difficult to detect at an early stage with no clear symptoms at onset. Malignancy development is a multistep process involving multiple genetic and epigenetic alterations leading to aberrant expression of key regulating factors. Manual pathological inspection of gastric slices is time-consuming and usually suffers from inter-observer variations. GastricNet, a deep learning-based framework, was proposed in [56] for automatic gastric cancer identification. The proposed network adopts different architectures for shallow and deep layers for better feature extraction. GastricNet adopts different architectures for its two modules: a CNN Multi-scale module (MSM) and a CNN Network in Network (NIN) module. The features extracted by the MSMs are concatenated

to form a composite feature for the NIN. A softmax classifier was used to determine the probability that the whole slice contains a tumor. The gastric dataset contains 560 gastric cancer slices and 140 normal slices. The cropping of original images of 2048x2048 to 224x224 patches generates 8,992 patches for gastric cancer and 14,000 patches for normal slices. The experimental results show that the proposed DL framework performs better than state-of-the-art networks and achieved an accuracy of 100 % for slice-based classification.

5.4. Deep learning for cancer prognosis

A challenging question is whether novel deep learning methods could be used to directly learn the prognostically relevant features in microscopy images of the tumour, without prior identification of the known tissue entities, such as mitoses, infiltrating immune cells, or tumour budding. In [28], the authors aim to predict five-year disease specific survival of patients diagnosed with colorectal cancer (CRC) *directly* from digitized images of haematoxylin and eosin (H&E) stained diagnostic tissue samples. The authors trained a deep network *to directly predict patient outcome, without any intermediate tissue classification*. They *combined convolutional and recurrent neural network methods* into a model that is scalable to process images of different sizes in assessment of CRC tumor samples. Particularly, the authors applied *transfer learning* by utilizing a visual recognition model (VGG-16) to avoid training a convolutional neural network from scratch and showed that a deep convolutional neural network trained on a large-scale dataset from a different domain is also useful for pattern recognition in digitized images of CRC. A recurrent neural network (Long Short-Term Memory; LSTM) is then trained to read a sequence of VGG-16-produced features to predict five-year disease specific survival. The authors obtained images of H&E-stained TMA spots from 420 patients diagnosed with CRC together with follow-up time and outcome information for each of the patients as well as clinicopathological characteristics of the tissue samples. *The performance of the model is compared with the prognostic accuracy achieved by the visual assessment (tumour grading) performed by a skilled pathologist.* The results show that deep learning-based outcome prediction with only small tissue areas as input outperforms visual histological assessment performed by human experts on both TMA spot and whole-slide level in the stratification into low- and high-risk patients. The results suggest that state-of-the-art deep learning techniques can extract more prognostic information from the tissue morphology of colorectal cancer than an experienced human observer.

Estimating the future course of patients with cancer lesions is invaluable to physicians; however, current clinical methods fail to effectively use the vast amount of multimodal data that is available. To tackle this problem, The authors of [57] constructed *a multimodal neural network-based model to predict the survival of patients* (prognosis) for 20 different cancer types using clinical data, mRNA expression data, microRNA expression data and histopathology whole slide images (WSIs). The authors developed an unsupervised method to encode multimodal patient data into a common feature representation that is independent of data type or modality, compressing these four data modalities into a single feature vector for each patient. A dedicated deep learning CNN architecture is used for each data type. For the clinical data, fully connected layers with sigmoid activations are used to extract features. Highway networks [55] (variants of CNNs) are used for the gene and microRNA data. The CNN architecture (called SqueezeNet [58]) is used to extract features from whole slide image data. The authors used pancancer data to train these feature encodings and predict single cancer and pancancer overall survival, achieving an AUC of 0.78 overall. Their methods achieved comparable or better results

from previous research by resiliently handling incomplete data and predicting across 20 different cancer types.

5.5. Deep learning for cancer treatment

Medical imaging provides non-invasive means for tracking patients' tumor response and progression after treatment. However, quantitative assessment through manual measurements is tedious, time-consuming, and prone to interoperator variability, as visual evaluation can be non-objective and biased. In [59], the authors demonstrated the ability of deep learning networks to predict prognostic endpoints of patients treated with radiation therapy using serial CT images of patients with locally advanced non–small cell lung cancer (NSCLC). Models were developed using *transfer learning* of convolutional neural networks. The output of the pretrained network model was then input into *a recurrent neural network* for predictions of survival. The authors used two independent cohorts, dataset A and dataset B, consisting in a total of 268 patients with stage III NSCLC for this analysis. Dataset A contained 179 patients and was randomly split 2:1 into training (n = 107) and testing (n = 72). Dataset A included patients treated with chemotherapy and definitive radiation therapy and was used to train the combined CNN-RNN for predictions of survival. The test set from this cohort was used to assess performance and compared with the performance of radiographic and clinical features. Dataset B contains 89 patients treated with chemotherapy and surgery. This dataset was used as an additional test set to predict pathologic response, and the model predictions were compared with the change in volume. *Overall survival was assessed along with three other clinical endpoints for the definitive radiation therapy cohort: distant metastases, locoregional recurrence, and progression*. The results showed increases in performance of survival and prognosis prediction with incorporation of additional timepoints using CNN and RNN networks. It was demonstrated that *deep learning can integrate imaging scans at multiple timepoints to improve clinical outcome predictions*. Model performance was enhanced with each additional follow-up scan into the CNN model (e.g., 2-year overall survival: AUC = 0.74). The models stratified patients into low and high mortality risk groups, which were significantly associated with overall survival (HR = 6.16; 95 % confidence interval (CI), 2.17 - 17.44). The model also significantly predicted pathologic response in dataset B.

Cross-sectional X-ray imaging has become the standard for staging most solid organ malignancies. However, for some malignancies such as urinary bladder cancer, the ability to accurately assess the local extent of the disease and understand the response to systemic chemotherapy is limited with current imaging approaches. In [60], the authors explored the feasibility that radiomics-based predictive models, using pre- and post-treatment computed tomography images, might be able to distinguish between bladder cancers that have fully responded to chemotherapy and those that have not. The authors assessed three unique radiomics-based predictive models: a deep-learning convolution neural network, a deterministic radiomics feature-based approach and a bridging method between the two, for extracting radiomics features from the image patterns. They also compared the performance of the models, in predicting a complete response of bladder cancer to neoadjuvant chemotherapy, with that of expert physicians. The training data, based on a set of 82 patients with 87 bladder cancers who were evaluated with CT before and after the administration of neoadjuvant chemotherapy, was collected retrospectively. Data for an additional 41 patients with 43 cancers were collected as a test set. One set of chemotherapy regimens was used for the majority of these patients, while another set of regimens was used for other patients. Pathology obtained from the bladder at the time of surgery was used to determine the final cancer stage after chemotherapy and was used as

the reference standard to determine whether the patient had responded completely to treatment. The performances of all three methods are comparable to those of the radiologists. The study indicated that the computerized assessment using radiomics information from the pre- and post-treatment CT of bladder cancer patients has the potential to assist in assessment of treatment response.

5.6. Deep learning for cancer risk prediction

Mammographic density improves the accuracy of breast cancer risk models. However, the use of breast density is limited by subjective assessment, variation across radiologists, and restricted data. In [61], the study aims to develop a mammography-based DL breast cancer risk model that is more accurate than established clinical models. This retrospective study included 88,994 consecutive screening mammograms for 39,571 women. For each patient, all examinations were assigned to either training, validation, or test sets, resulting in 71,689, 8,554, and 8,751 examinations, respectively. Cancer outcomes were obtained through linkage to a regional tumor registry. By using risk factor information from patient questionnaires and electronic medical records review, three models were developed to assess breast cancer risk within 5 years: a risk-factor-based logistic regression model (RF-LR) that used traditional risk factors, a DL model (image-only DL) that used mammograms alone, and a hybrid DL model that used both traditional risk factors and mammograms. For the image-only DL model, the authors implemented a deep convolutional neural network (ResNet18 [62]). Given a 1664 x 2048 pixel view of a breast, the DL model was trained to predict whether breast cancer would develop within 5 years. Comparisons were made to an established breast cancer risk model, the Tyrer-Curzick (TC) tool that included breast density (TC is a tool used to calculate a woman's likelihood of developing breast cancer in 10 years). Model performance was compared by using AUCs with DeLong test (test to compare the difference between two AUCs). In summary, deep learning models that use full-field mammograms yield substantially improved risk discrimination compared with the Tyrer-Cuzick model. When the hybrid DL model was compared with breast density, it was found that patients with nondense breasts and model-assessed high risk had 3.9 times the cancer incidence of patients with dense breasts and model-assessed low risk. Overall, it was concluded that DL models can deduce informative indicators of risk contained in the mammograms not captured by traditional risk factors, hence DL models have the potential to replace conventional risk prediction models.

For medical data, patients could have symptoms of multiple different diseases at the same time and it is important to develop tools that help to identify problems early. This problem is challenging as it is difficult to infer information about classes that are not mutually exclusive. In the study described in [63], deep learning methods are used to tackle this challenging *multi-label classification of data problem* and predict chronic diseases for intelligent health risk prediction. In this study, hypertension, diabetes, and fatty liver are three chronic diseases that are analyzed to predict types of chronic diseases for a given patient. Overall, there are eight different diagnoses that can be given: one of the three diseases, two of the three diseases, all three diseases or no disease. The DNNs are used in this study. The number of hidden layers is experimented from 1 to 10. Physical examination records of 110,300 anonymous patients were used to predict diabetes, hypertension, fatty liver, a combination of these three chronic diseases, and the absence of disease. The dataset was split into training (90 %) and testing (10 %) sub-datasets. Deep Learning (DL) architectures were compared with standard and state-of-the-art multi-label classification methods: the decision tree C4.5, the Support Vector Machines (SVM), the Random Forest (RF), the KNN, and the Multilayer Perceptron (MLP). The results showed that DNNs give the highest accuracy among all six popular classifiers. The F-score of DNNs is

slightly lower (but compatible) than Random Forest and MLP classifiers but much higher than those of SVM andKNN. The authors concluded that deep learning architectures have the potential of inferring more information from physical examination data than common classification methods.

Remarks on reviewed applications. Most applications demonstrated that deep learning outperformed conventional machine learning, for this reason, we only provide observations that are common and that may help improve next generation applications. Most applications used simple mutation and expression data as input, few used proteomic data probably because i) proteomic data and its analysis are not readily available and ii) a gene's products are multiple and they come in varied shapes and sizes due to post translation as well as mutation. Autoencoders are often used in the first stage of the solution as the dimension of the feature space is extremely high in the order of 20,000 to 50,000. It is thus important to ensure that the integrity of the data is preserved in terms of the feature space. Omic-data is often very sparce as many features in the datasets have little to do with the latent features of the problem and do not correlate with the outcome, they need to be properly encoded and filtered out to save computational costs and produce more accurate results. For image-based applications, the use of "transfer learning" was prevalent for good reasons, since not only spatial and representational features enable excellent object recognition in deep learning but also the cost of training the DL architecture from scratch is substantially reduced. In general, omic-data seems abundant but in many cases the data is still deficient and/or imbalanced for classification as some portions of the data are too well represented and other portions underrepresented. Artificial but credible data may need to be generated to fill the gap. Multimodal input works well with deep learning as different types of input can be integrated easily and produce more accurate classification. It is still a challenge to infer information about classes that are not mutually exclusive. For example, patients could have symptoms of multiple different diseases at the same time and it is important to develop tools that help to identify problems early. Deep learning architectures can tackle this multi-label classification problem as they have the potential of inferring more information about the patterns of physical examination data than common classification methods.

6. CHALLENGES FOR DEEP LEARNING AND ITS APPLICATION TO CANCER ONCOLOGY

6.1. Challenges for Deep learning

The breakthrough in deep learning is the invention of the convolutional neural network which is inspired by the architecture of the human visual cortex. In the CNN method, undifferentiated features of the input data are extracted layer by layer to reveal the hierarchical knowledge of the patterns within and used in the final layer to produce the classification output. The method emulates the architecture and the learning scheme of the human visual cortex [18]. Humans learn to solve problems by decomposing them into a simpler and smaller set of problems and then finding solutions to these subproblems. Then the relevant parts of the solutions of the subproblems are selected and combined to provide the solution to the original problem. Polya [64] expressed this decomposition of the original problem and composition of the sub-solutions in a different way in his work on problem solving. Utilising just one specific learning scheme of the human visual cortex, deep learning has already achieved remarkable performances. The challenge is now to design more sophisticated/intelligent deep learning methods, inspired by other problem-solving strategies of the human brain. Or better still, to build

a true deep learning brain that has a repertoire of problem-solving skills and the intelligence to take appropriate decisions when facing a situation and adapting appropriate strategies for any specific environment.

Autoencoder deep learning methods have been used extensively to reduce the dimension and extract latent features of the input successfully; however, the principle behind the learning is very much supervised learning where the difference between the input and the network constructed input is used to minimize the loss function. The challenge is to discover other innovative unsupervised learning methods.

Reinforcement learning methods have had enormous successes in AlphaGo and other projects that pit deep learning machines against humans. This type of learning has an exciting potential for innovations and breakthroughs but it has not been thoroughly investigated. The challenge is to formulate deep learning and integrate it within a formal reinforcement learning framework to produce a higher level of intelligence.

In the meantime, due to the generality of problem decomposition and sub-solutions composition nature of deep learning, the potential of the current deep learning methods seems unlimited with many problems awaiting solutions. It will be sometime before application of the current deep learning methods is exhausted. The challenge is to formulate deep learning theoretically and quantify its capability.

On a more specific note, deep learning methods can be refined and improved in several ways. The model or the neural network architecture of deep learning is still ad hoc. The architecture is selected to solve a problem based purely on perception and experience such as CNNs for spatial data and RNN for sequential data. There is little consideration of the organization and complexity of the model - the depth of hidden layers, the size of a hidden layer, and the connectivity among nodes between layers to adequately solve the problem at hand. The challenge is to come up with a sound formulation and arrive at optimized deep learning models.

Overfitting is a manifestation of the mismatch problem between the order or the complexity of the architecture and the unknown dimension of the data. Research in matching the complexity of the problem and employable deep learning methods is necessary to obtain efficient and quality solutions [63]. Initial searches for new and efficient neural architectures are explored in [65]. The success hinges on some appropriately designed neural network architectures that encompasses a task-independent criterion of the quality of the representation that the network is required to learn [66].

Deep learning methods are still back boxes that learn by simple association. Other types of learning need to co-exist in future deep learning methods to deal with comprehensive sets of problems and data. An integrated AMI model was proposed in [66] that consists of a discriminative module, an associative module, and a feedback mechanism between them as seen in the human neural cortex. The discriminative compartment takes into account the need for analysing, discriminating, and clustering patterns (bottom up); the associative compartment provides the ability to associate, correlate, and make generalized decisions (top down); and the feedback between the compartments ensures the stability of the overall learning module and provides the dynamic balance between bottom-up and top-down learnings.

The backpropagation learning algorithm is the generally accepted algorithm used in current deep learning methods as it is efficient and yields desirable results. However, to achieve human-level learning, biologically plausible learning rules should be the subject of experimentation [67]. Bienenstock, Cooper, and Munro [68] introduced a learning rule that reflects the way neuron synaptic weights are changed and is stable without imposing external constraints on the

synapses as required in Hebbian or Von de Malsburg [69] learning rules. Deep learning methods are complex machines containing hundreds of millions of parameters, making training and regularization difficult. High capacity neural networks face serious challenges not only in designing appropriate architectures and learning strategies but also in validating reliable and superior performance in terms of generalizability to events not included in the training data for critical applications [70], [71].

6.2. Challenges for deep learning applications to cancer

Alterations of the genome that lead to uncontrolled cell growth manifest itself at various levels: DNA, gene expression, and protein. At the DNA level, a segment of DNA may be damaged or mutated; at the gene expression level, genes may be repressed, over-expressed, or expressed in the form of isoforms. At the gene product (i.e., protein) level, intended proteins may be overproduced or not produced at all. Unintended proteins may be produced because of mutations of the DNAs or post translation regulation damage. Current applications mostly utilized either simple mutation data or gene expression data and rarely proteomics data. Clearly, with multi-omics data deep learning methods will improve the accuracy of cancer diagnosis, prognosis, and treatment.

Deep learning accommodates multi-modal inputs that include clinical information to take into account of human expertise, images to take into account phenotype traits of a particular cancer, and proteomics data to take into account actionable elements that affect the tumorigenesis process. This allows deep learning to take human expertise into account.

Specific cancer instances are a particular manifestation of the general cancer pathway. However, "the designation 'cancer pathways' is in so far imprecise, as the same pathways also control the proliferation, differentiation, survival, and function of normal tissues. So, differences between normal and cancer cells are expected to be quantitative rather than qualitative" [21]. The challenge is to establish, quantify, and characterize clear pathways leading to a specific cancer and incorporate them in deep learning for more accurate solutions.

One issue is that current deep learning methods lack transparency and interpretability [72]. This presents not just a barrier for translating excellent results to clinical applications but also limits the ability to uncover causal and structural relationships common in biology. For lack of interpretability, several methods have been explored to determine the working of DNN architectures, including Taylor decomposition, layer-wise decomposition, and heat-maps for attributing the contribution of connection weights to the significance of the outcome [53].

It has been claimed that deep learning is unsuitable for some critical medical applications because it can experience catastrophic forgetting when its learned memories (weight parameters of the neural networks) suddenly collapse. An alternative model, the Adaptive Resonance Theory for both biological and artificial intelligence has been proposed in [73].

Despite the enormous amount of cancer genomic data, as well as drug sensitivity data and the availability of numerous learning methods, going from simplest mutation profiles to selecting the most effective cancer drugs remains a challenge. This is partly due to the lack of labelling data and partly due to the complexity of tumor resistance, and possibly due to our less than complete knowledge of cancer pathways.

Providing reliable labels for training is challenging. For example, most of the established tissue entity labels include errors due to the subjective nature of visual interpretation by the human observer [28]. More effort has to be dedicated to improving both the quantity and quality of labelling/annotation of medical data.

Overfitting and underfitting present real problems in all deep learning applications as the designer does not have a complete knowledge of the complexity of the data and the selected model may be too complex or too simple for the data. Lack of data is also a common overfit problem.

It has been demonstrated that "the millions of adjustable parameters make deep neural networks capable of performing perfectly in training sets even when the target outputs are randomly generated and, therefore, utterly meaningless" [74]. This occurs when the data is inadequate to represent the complete distribution of the input or it may be skewed towards a particular set of training samples. Strategies and methods are needed to ensure relevant critical information is represented in the training as well as the validation datasets if the deep learning method is suitable for the intended medical/cancer applications [75]. A remedy is to fully integrate medical images, clinical data as well as phenotypically rich data, and omics data in multimodal learning to realize meaningful results.

7. CONCLUSION

The ultimate objective of this paper is to stimulate ideas and facilitate collaboration between cancer and deep learning researchers to address challenging oncological problems using advanced deep learning technologies. To this end, we have introduced the fundamentals of deep learning in terms of their architectural models and insights into various learning methods to cancer biologists. We have presented the essentials of cancer molecular biology to deep learning practitioners. We have reviewed a number of recent applications of deep learning to cancer diagnosis, prognosis, treatment, drug response, and cancer risk prediction to demonstrate how deep learning methods were selected and applied. We have discussed extensively the challenges of deep learning and its application to cancer as well as indicating possible directions for deep learning and cancer research. Clearly, to design better diagnosis, prognosis, and treatment of cancer, collaboration between cancer biologists and deep learning practitioners is essential.

New technologies such as single-cell sequencing, spatial transcriptomics and multiplexed imaging will enrich available datasets with new dimensions that improve the performance of deep learning methods in cancer research and clinical application. Recent technological advances have initiated a new era of personalized or precision medicine through data-driven assessment of diseases by combining deep learning and biomedical science. Deep learning with molecular diagnostic data will be helpful for the determination of tumor stage, such as early detection of tumor cells in the blood or identifying populations at risk of cancer which is often not possible by traditional methods [21]. Ultimately, for deep learning to be accepted in routine patient care, collaboration between experts in both the oncology and deep learning fields is imperative for clinical validation of interpretable deep learning methods [72]. Finally, cancer development is a gradual process that dynamically involves numerous events and actors that interplay through complex patterns. Deep learning is the most powerful technology developed so far to detect intricate latent features and patterns, both spatially and sequentially. Deep learning may prove to be a valuable tool for preventing cancer and also for determining the cause of cancer. One would also expect deep learning to play a role in pinpointing the pathway for cancer development or precisely identifying the triggering points along the tumorigenesis process.

Acknowledgements. The authors express their thanks to Adjunct Professor Michael Quigley for his thoughtful comments and meticulous editing of the final version of the paper.

Declaration of competing interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Adjiri A. - Mutations May Not Be the Cause of Cancer, Oncology and Therapy **5** (1) (2017) 85-101.
2. Motofei I. G. - Biology of cancer; from cellular and molecular mechanisms to developmental processes and adaptation, Seminars in Cancer Biology, 2021, doi: https://doi.org/10.1016/j.semcancer.2021.10.003.
3. DeepMind - "AlphaGo." https://www.deepmind.com/research/highlightedresearch/alphago (accessed June 13, 2022).
4. Krizhevsky A., Sutskever I., and Hinton G. - ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems **25** (2012) 1097-1105.
5. Russakovsky O., *et al.* - ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput Vis **115** (2015) 211-252. doi: https://doi.org/10.1007/s11263-015-0816-y.
6. Alpaydin E. - Introduction to Machine Learning, 3 Ed. MIT Press, 2014.
7. Burkov A. - The Hundred-Page Machine Learning Book, Kindle Ed. 2019.
8. McCulloch W. S. and Pitts W. - A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics **5** (1943) 115-133. doi:https://doi.org/10.1007/BF02478259
9. Morin P. J., Trent J. M., Collins F. S., and Vogelstein B. - Cancer Genetics, in Harrisson's Principles of Internal Medicine, D. L. Kasper, A. S. Fauci, S. L. Hauser, D. L. Longo, and J. L. Jameson Eds., 19 edS.: McGrawHill Education, 2015.
10. Campbell M. A., *et al*. - Biology, Pearson Education Australia, 2009.
11. Kolmogorov A. N. - On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition, Dokl. Akad. Nauk SSSR.**115** (5) (1957) 953-956.
12. Cybenko G. - Approximation by superpositions of a sigmoidal function, Math. Control Signal Systems **2** (1989) 303-314. doi: https://doi.org/10.1007/BF02551274
13. Gershenfeld N. - The nature of Mathematical Modeling, Cambridge University Press, 2002.
14. He K., Zhang X., Ren S., and Sun J. - Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, doi:https://doi.org/10.48550/arXiv.1502.01852.
15. Xu J., Li Z., Du B., Zhang M., and Liu J. - Reluplex made more practical: Leaky ReLU, presented at the 2020 IEEE Symposium on Computers and Communications (ISCC), 2020.
16. Clevert A., Unterthiner T., and Hochreiter S. - Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), 2016. [Online]. Available: arXiv:1511.07289v5.

17. Sutton R. S. and Barto A. G. - Reinforcement learning: An introduction, MIT press Cambridge, 1998.
18. LeCun Y., Bengio Y., and Hinton G. - Deep learning, Nature **521** (2015) 436-444. doi: https://doi.org/10.1038/nature14539
19. Graves A., Mohamed A. R., and Hinton G. - Speech recognition with deep recurrent neural networks, presented at the 2013 IEEE international conference on acoustics, speech and signal processing, 2013.
20. Murugan P. - Facial information recovery from heavily damaged images using generative adversarial network-part 1, [Online]. Available: arXiv preprint arXiv:180808867
21. Schulz W. A. - Molecular Biology of Human Cancers - An Advanced Student's Textbook, Springer, 2007.
22. Fior R. and Zilhão R. (Eds.) - Molecular and Cell Biology of Cancer - When Cells Break the Rules and Hijack Their Own Planet, Springer, 2019.
23. Jameson J. L. and Kopp P. - Principles of Human Genetics, in Harrisson's Principles of Internal Medicine, D. L. Kasper, A. S. Fauci, S. L. Hauser, D. L. Longo, and J. L. Jameson (Eds.): McGrawHill Education, 2015, ch. 82.
24. The Human Genome Completed [Online] Available: https://www.nature.com/news/2006/ 060515/full/news060515-12.html
25. Hanahan D. and Weinberg R. A. - Hallmarks of cancer: the next generation, Cell **144** (5) (2011) 646-674. doi: https://doi.org/10.1016/j.cell.2011.02.013.
26. Silver D. *et al.* - Mastering the game of Go with deep neural networks and tree search, Nature **529** (2016) 484-489. doi: https://doi.org/10.1038/nature16961.
27. Danaee P., Ghaeini R., and Hendrix D. - A deep learning approach for cancer detection and relevant gene identification, Pac Symp Biocomput **22** (2017) 219-229. doi:10.1142/ 9789813207813_0022.
28. Bychkov D., *et al.* - Deep learning based tissue analysis predicts outcome in colorectal cancer, Scientific Reports **8** (2018) Art no. 3395, doi: https://doi.org/10.1038/s41598-018-21758-3.
29. Chang Y., *et al.* - Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature, Sci. Rep. **8** (2018) Art no. 8857, doi: https://doi.org/10.1038/s41598-018-27214-6
30. Yap C. W. - PaDEL-descriptor: an open source software to calculate molecular descriptors and fngerprints, J. Comput. Chem. **32** (2011) 1466-1474.
31. Menden M. P., *et al.* - Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties, PLoS ONE **8** (4) (2013). doi:https://doi.org/ 10.1371/journal.pone.0061318.
32. Zou J., Huss M., A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti - A primer on deep learning in genomics, Nature Genetics **51** (2019) 12-18. doi:https://doi.org/10.1038/ s41588-018-0295-5.
33. Goodfellow I., Bengio Y., and Courville A. - Deep Learning, Cambridge, MA, USA: The MIT Press, 2016.
34. Burkov A. - Machine Learning Engineering, True Positive Inc., 2020.

35. Géron A. - Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2 Ed. O'Reilly Media, 2019.
36. Weinstein J., Collisson E., *et al.* - The Cancer Genome Atlas Pan-Cancer analysis project, Nature Genetics **45** (2013) 1113-1120. doi: https://doi.org/10.1038/ng.2764.
37. Tomczak K., Czerwińska P., and Wiznerowicz M. - Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," Contemp Oncol (Pozn) **19** (1A) (2015) A68-77. doi: 10.5114/wo.2014.47136.
38. Barretina J., Caponigro G., Stransky N., *et al.* - The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, Nature **483** (2012) 603-607. doi:https://doi.org/10.1038/nature11003.
39. Clough E. and Barrett T. - The Gene Expression Omnibus database, Methods in Molecular Biology **1418** (2016) 93-110. doi: doi:10.1007/978-1-4939-3578-9_5.
40. Edgar R., Domrachev M., Lash A. E. - Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Research **30** (2002) 207-210. doi:https://doi.org/10.1093/nar/30.1.207.
41. Lonsdale J., Thomas J., Salvatore M., *et al.* - The Genotype-Tissue Expression (GTEx) project, Nature Genetics **45** (2013) 580-585. doi: https://doi.org/10.1038/ng.2653.
42. Hu Z., Tang J., Wang Z., Zhang K., Zhang L., and Sun Q. - Deep learning for image-based cancer detection and diagnosis - A survey, Pattern Recognition **83** (2018) 134-149. doi: https://doi.org/10.1016/j.patcog.2018.05.014.
43. Khanam N. and Kumar R. - Recent Applications of Artificial Intelligence in Early Cancer Detection, Curr. Med. Chem. (2022). doi:10.2174/0929867329666220222154733.
44. Chiu Y. C., *et al.* - Predicting and characterizing a cancer dependency map of tumors with deep learning, Sci. Adv. **7** (34) (2021). doi:10.1126/sciadv.abh1275.
45. CTD Data Portal. https://ocg.cancer.gov/programs/ctd2/data-portal (accessed.
46. Newton Y., *et al.* - TumorMap: exploring the molecular similarities of Cancer samples in an interactive portal, Cancer Res. **77** (2) (2017) 111-114.
47. Iorio F., *et al.* - A landscape of Pharmacogenomic interactions in Cancer, Cell **166** (3) (2016) 740-754.
48. Li M., *et al.* - DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines, IEEE/ACM Trans Comput Biol Bioinform **18** (2) (2021) 575-582. doi:10.1109/TCBB.2019.2919581.
49. Bolton E. E., Wang Y., Thiessen P. A., and Bryant S. H. - PubChem: Integrated Platform of Small Molecules and Biological Activities, Annual Reports in Computational Chemistry **4** (2008) 217-241.
50. Fakoor R., Ladhak F., Nazi A., and Huber M. - Using deep learning to enhance cancer diagnosis and classication, Presented at the Proceedings of the 30 th International Conference on Machine Learning, Atlanta, Georgia, 2013.
51. Hosny K. M., Kassem M. A., and Foaud M. M. - Skin Cancer Classification using Deep Learning and Transfer Learning, Presented at the 9th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, 2018.

52. Yuan Y., *et al.* - DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations, BMC Bioinformatics **17** (2016) Art no. 476, doi:https://doi.org/10.1186/s12859-016-1334-9.
53. Lyu B. and Haque A. - Deep Learning Based Tumor Type Classification Using Gene Expression Data, Presented at the BCB '18: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018.
54. Shen L., Margolies L. R., Rothstein J. H., Fluder E., McBride R., and Sieh W. - Deep Learning to Improve Breast Cancer Detection on Screening Mammography, Sci. Rep. **9** (2019). doi: https://doi.org/10.1038/s41598-019-48995-4.
55. Khan A., Sohail A., Zahoora U., and Qureshi A. S. - A survey of the recent architectures of deep convolutional neural networks, Artif Intell Rev. **53** (2020) 5455-5516. doi:https://doi.org/10.1007/s10462-020-09825-6.
56. Li Y., Li X., Xie X., and Shen L. - Deep learning based gastric cancer identification, Presented at the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018.
57. Cheerla A. and Gevaert O. - Deep learning with multimodal representation for pancancer prognosis prediction, Bioinformatics **35** (14) (2019) 446-454. doi:https://doi.org/10.1093/ bioinformatics/btz342
58. Hu J., Shen L., Albanie S., Sun G., and W. E. - Squeeze-and-Excitation Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence **42** (8) (2020) 2011-2023. doi:10.1109/TPAMI.2019.2913372.
59. Xu Y., *et al*. - Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging, Clin Cancer Res. **25** (11) (2019) 3266-3275. doi:10.1158/1078-0432.CCR-18-2495.
60. Cha K. H., *et al.* - Bladder Cancer Treatment Response Assessment in CT using Radiomics with Deep-Learning, Scientific Reports **7** (2017) Art no. 8738. doi:https://doi.org/10.1038/s41598-017-09315-w
61. Yala A., Lehman C., Schuster T., Portnoi T., and Barzilay R. A. -A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction, Radiology **292** (1) (2019) 60-66. doi: 10.1148/radiol.2019182716.
62. He K., Zhang X., Ren S., and Sun J. - Deep residual learning for image recognition, Presented at the The IEEE Conference on Computer Vision and Pattern Recognition, 2016.
63. Maxwell A., *et al.* - Deep learning architectures for multi-label classification of intelligent health risk prediction, BMC Bioinformatics **18** (2017) Art no. 523. doi:https://doi.org/10.1186/s12859-017-1898-z.
64. Polya G. - How to solve it - A new aspect of Mathematical method, 2 Ed. Princeton University Press, 1973.
65. Liu H., Simonyan K., Vinyals O., Fernando C., and Kavukcuoglu K. - Hierarchical Representations for Efficient Architecture Search, 2018, doi:https://doi.org/10.48550/arXiv.1711.00436

66. Hoang D. B. and James M. R. - Stability and discriminative properties of the AMI model, Presented at the Proceedings of International Conference on Neural Networks (ICNN'97), 1997.
67. Hoang D. B. and James M. R. (Eds.) - AMI: A model of intelligence (PRICAI'96: Topics in Artificial Intelligence. Lecture Notes in Computer Science. Berlin: Springer, 1996.
68. Bienenstock E. L., Cooper L. N., and Munro P. W. - Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex, The Journal of Neuroscience **2** (1) (1982) 32-48.
69. Von der Malsburg C. - Self-organization of orientation sensitive cells in the striate cortex, Kybernetic **14** (1973) 85-420.
70. Liu Y., Chen P. H. C., Krause J., and Peng L. - How to read articles that use machine learning: users' guides to the medical literature, JAMA **322** (2019) 1806-1816.
71. Ransohoff D. F. - Bias as a threat to the validity of cancer molecular-marker research, Nat. Rev. Cancer **5** (2005) 142-149.
72. Tran K. A., Kondrashova O., Bradley A., Williams E. D., Pearson J. V., and Waddell N. - Deep learning in cancer diagnosis, prognosis and treatment selection, Genome Med. **13** (2021) Art no. 152. doi: https://doi.org/10.1186/s13073-021-00968-x
73. Grossberg S. - The resonant brain: How attentive conscious seeing regulates action sequences that interact with attentive cognitive learning, recognition, and prediction, Atten Percept Psychophys **81** (2019) 2237-2264. doi:https://doi.org/10.3758/s13414-019-01789-2
74. Zhang C., Bengio S., Hardt M., Recht B., and Vinyals O. - Understanding deep learning requires rethinking generalization, Presented at the Proc. Int. Conf. Learn. Represent, 2017. [Online]. Available: https://arxiv.org/abs/1611.03530.
75. Kleppe A., Skrede O. J., De Raedt S., Liestol K., Kerr D. J., and Danielsen H. E. - Designing deep learning studies in cancer diagnostics, Nat. Rev. Cancer **21** (2021) 199-211. doi: https://doi.org/10.1038/s41568-020-00327-9