Chapter

Applications of Machine Learning in Healthcare

Christopher Toh and James P. Brody

Abstract

Machine learning techniques in healthcare use the increasing amount of health data provided by the Internet of Things to improve patient outcomes. These techniques provide promising applications as well as significant challenges. The three main areas machine learning is applied to include medical imaging, natural language processing of medical documents, and genetic information. Many of these areas focus on diagnosis, detection, and prediction. A large infrastructure of medical devices currently generates data but a supporting infrastructure is oftentimes not in place to effectively utilize such data. The many different forms medical information exist in also creates some challenges in data formatting and can increase noise. We examine a brief history of machine learning, some basic knowledge regarding the techniques, and the current state of this technology in healthcare.

Keywords: machine learning, healthcare, big data, medicine, genetics, disease

1. Introduction

The advent of digital technologies in the healthcare field is characterized by continual challenges in both application and practicality. Unification of disparate health systems have been slow and the adoption of a fully integrated healthcare system in most parts of the world has not been accomplished. The inherent nature and complexity of human biology, as well as the variation between individual patients has consistently shown the importance of the human element in diagnosing and treating diseases. However, advances in digital technologies are no doubt becoming indispensable tools for healthcare professionals in providing the best care for patients.

The improvement of data technologies, including storage size, computational power, and data transfer speeds, has enabled the widespread adoption of machine learning in many fields—healthcare included. Due to the multivariate nature of providing quality healthcare to an individual, the recent trends in medicine have emphasized the need for a personalized medicine or "precision medicine" approach to healthcare. The goal of personalized medicine is to use large amounts of healthcare data to find, predict, and analyze diagnostic decisions, which physicians can in turn implement for each individual patient. Such data includes but is not limited to genetic or familial information, medical imaging data, drug combinations, population wide patient health outcomes, and natural language processing of existing medical documentation.

We will focus primarily on three of the largest applications of machine learning (ML) in the medical and biomedical fields. As a rapidly evolving field, there is a wide range of potential applications of machine learning in the healthcare field which may encompass auxiliary aspects of the field such as personnel management, insurance policies, regulatory affairs, and much more. As such, the topics covered in this chapter have been narrowed down to three common applications of machine learning.

The first is the use of machine learning in medical images such as magnetic resonance imaging (MRIs), computerized axial tomography (CAT) scans, ultrasound (US) imaging, and positron emission tomography (PET) scans. The result of these imaging modalities is a set or series of images which typically requires a radiologist to interpret and make a diagnosis. ML techniques have rapidly been advancing to predict and find images which may indicate a disease state or serious issue.

The second is natural language processing of medical documents. With the push towards electronic medical records (EMR) in many countries, the consensus from many healthcare professionals has been that the process is slow, tedious, and, in many cases, completely botched. This can sometimes lead to poorer overall healthcare for patients. One of the major challenges is the amount of physical medical records and documentation that already exists in many hospitals and clinics. Different formatting, hand-written notes, and a plethora of incomplete or non-centralized information has made the switch to adopting electronic medical records less than efficient.

The third machine learning application encompasses the use of human genetics to predict disease and find causes of disease. With the advent of next-generation sequencing (NGS) techniques and the explosion of genetic data including large databases of population-wide genetic information, the attempt to discern meaningful information of how genetics may affect human health is now at the forefront of many research endeavors. By understanding how complex diseases may manifest and how genetics may increase or decrease an individual person's risk can aid in preventative healthcare. This could provide physicians with more information on how to tailor a specific patients' care plan to reduce the risk of acquiring more complex diseases.

The common issue present in all three of these topics is how to translate health data acquired from the Internet of Things, into understandable, useful, trustworthy information for patients and clinician. How do we interpret hundreds of thousands of inputs and parameters from the data? How do we do this efficiently? What is the progress of addressing this problem currently?

2. Artificial intelligence and machine learning

Artificial intelligence (AI) has been intricately linked to the rise of modern-day computing machines. Machine learning has its roots and beginnings firmly planted in history. Alan Turing's work in cracking the German Enigma machine during World War II became the basis for much of modern computer science. The Turing Test, which aims to see if AI has become indistinguishable from human intelligence, is also named after him [1, 2].

At the height of the Second World War, the Allies had a significant logistical hurdle in the Atlantic. The United States and United Kingdom needed to set up secure shipping lines to move both armaments and troops to England in preparation for a mainland European invasion. However, the German U-boats were extremely effective at disrupting and sinking many of the ships traversing these shipping lanes [3]. As such, the Allies needed to intercept German communications to swing the

Battle of the Atlantic in their favor. The Germans encrypted their communications with The Enigma Machine, the most sophisticated encryption device of its time.

Turing and the rest of Bletchley Park were tasked with breaking the coded messages produced by The Enigma Machine and eventually produced The Bombe, a mechanical computing device which successfully decoded the cipher of The Enigma machine (**Figure 1**). Using the Bombe, they read the German orders sent to submarines and navigated their ships around these dangers. This was Turing's first intelligent machine. Alan Turing would later go on to describe the idea of a thinking machine which would eventually be called AI [4].

Machine learning is a subset of AI and the term was coined in the late 1950s by Arthur Samuel who published a paper on training computers to play checkers when he worked with IBM [5]. AI is best described as giving human-like intelligence to machines in a manner that directly mimics the decision making and processing of the human conscience. ML is the subset of AI that focuses on giving machines the ability to learn in an unaided manner without any human intervention.

By the late 1960s, researchers were already trying to teach computers to play basic games such as tic-tac-toe [6]. Eventually, the idea of neural networks, which were based on a theoretical model of human neuron connection and communication, was expanded into artificial neural networks (ANNs) [7, 8]. These foundational works laid dormant for many years due to the impracticality and poor performance of the systems created. Computing technology had not yet advanced enough to reduce the computational time to a practical level.

The modern computer era led to exponential increases in both computational power and data storage capacity. With the introduction of IBM's Deep Blue and Google's AlphaGo in recent decades, several leaps in AI have shown the capacity of

Figure 1.
Picture of the German Enigma machine which was used to code military communications. Taken from Wikimedia Commons.

AI to solve real world, complex problems [9, 10]. As such, the promise of machine learning has taken hold in almost every sector imaginable.

The widespread adoption of machine learning can be mostly attributed to the availability of extremely large datasets and the improvement of computational techniques, which reduce overfitting and improve the generalization of trained models. These two factors have been the driving force to the rapid popularization and adoption of machine learning in almost every field today. This coupled with the increasing prevalence of interconnected devices or the Internet of Things (IoT) has created a rich infrastructure upon which to build predictive and automated systems.

Machine learning is a primary method of understanding the massive influx of health data today. An infrastructure of systems to complement the increasing IoT infrastructure will undoubtedly rely heavily on these techniques. Many use cases have already show enormous promise. How do these techniques work and how do they give us insight into seemingly unconnected information?

2.1 Machine learning algorithms

Machine learning is broadly split into supervised and unsupervised learning. Algorithms falling under both categories implement mathematical models. Each algorithm aims to give computers the ability to learn how to perform certain tasks.

2.1.1 Supervised learning

Supervised learning typically employs training data known as labeled data. Training data has one or more inputs and has a "labeled" output. Models use these labeled results to assess themselves during training, with the goal of improving the prediction of new data (i.e., a set of test data) [11]. Typically, supervised learning models focus on classification and regression algorithms [12]. Classification problems are very common in medicine. In most clinical settings, diagnosing of a patient involves a doctor classifying the ailment given a certain set of symptoms. Regression problems tend to look at predicting numerical results like estimated length of stay in a hospital given a certain set of data like vital signs, medical history, and weight.

Common algorithms included in this supervised learning group are random forests (RF), decision trees (DT), Naïve Bayes models, linear and logistic regression, and support vector machines (SVM), though neural networks can also be trained through supervised learning [13]. Random forests are a form of decision trees but are an ensemble set of independently trained decision trees. The resulting predictions of the trees are typically averaged to get a better end result and prediction [14]. Each tree is built by using a random sample of the data with replacement and at each candidate split a random subset of features are also selected. This prevents each learner or tree from focusing too much on apparently predictive features of the training set which may not be predictive on new data. In other words, it increases generalization of the model. Random forests can have hundreds or even thousands of trees and work fairly well on noisy data [15]. The model created from aggregating results from multiple trees trained on the data will give a prediction that can be assessed using test data (**Figure 2**).

A method used to improve many supervised algorithms is known as gradient boosting. Taking decision trees as an example, the gradient boosting machine as it is commonly known, performs a similar ensemble training method as the random forest but with "weak learners." Instead of building the decision trees in parallel as in the random forest algorithm, the trees are built sequentially with the error of the previous tree being used to improve the next tree [16]. These trees are not nearly as deep as the random forest trees, which is why they are called "weak" (**Figure 3**).

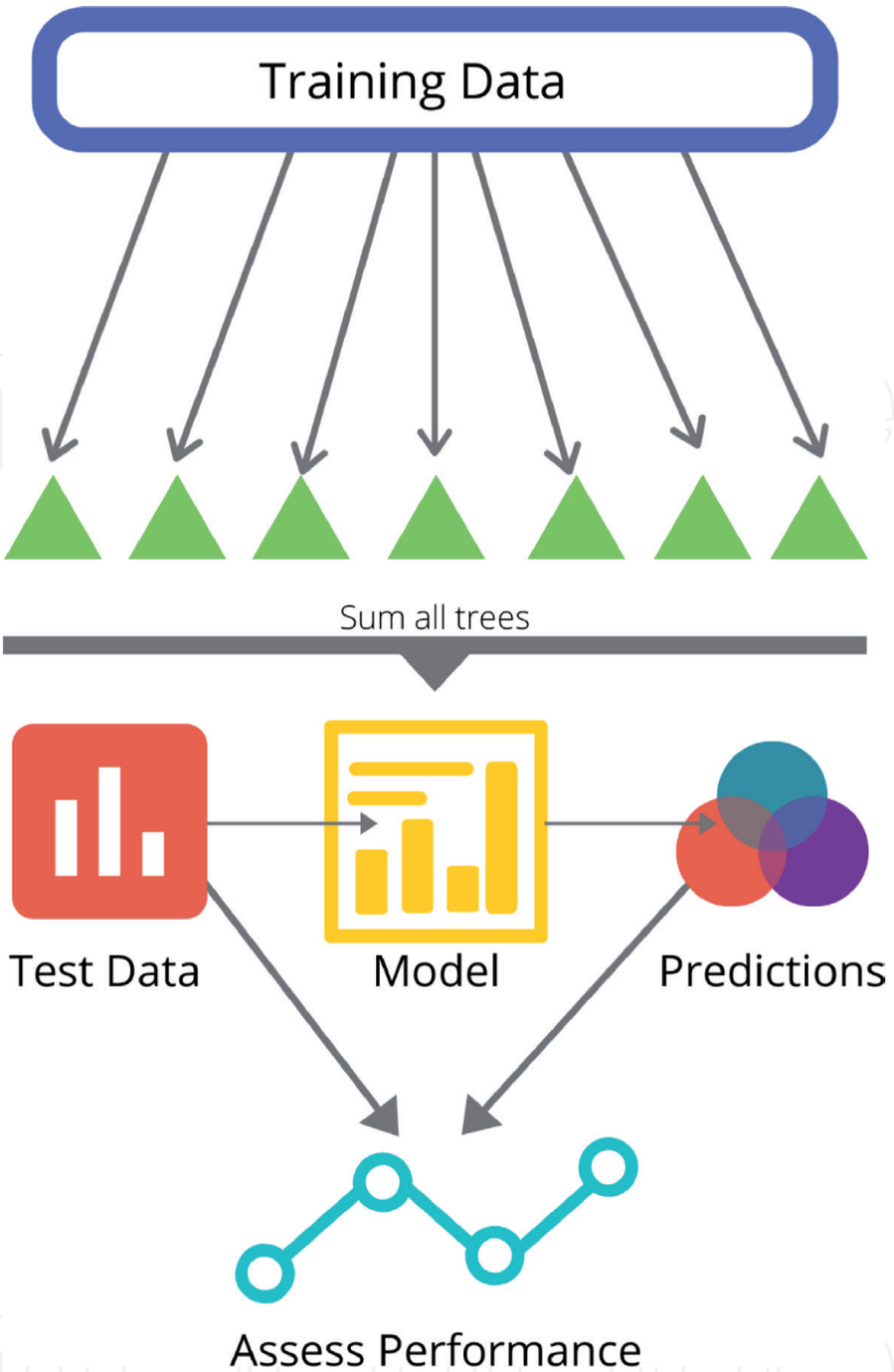

Figure 2.
Example of a workflow for training and assessing a random forest model. Each green triangle represents an independently trained tree from the training data. The prediction of each tree is summed and is represented as the model. Test data is then fed to the model, i.e., all the trees, and the resulting prediction is made. The prediction is then compared to the original test data to assess how the model performs.

Typically, better results can be achieved with gradient boosting, but tuning is much more difficult, and the risk of overfitting is higher. Gradient boosting works well with unbalanced data and training time is significantly faster due to the gradient descent nature of the algorithm [17, 18].

2.1.2 Unsupervised learning

Unsupervised machine learning uses unlabeled data to find patterns within the data itself [19]. These algorithms typically excel at clustering data into relevant groups, allowing for detection of latent characteristics which may not be

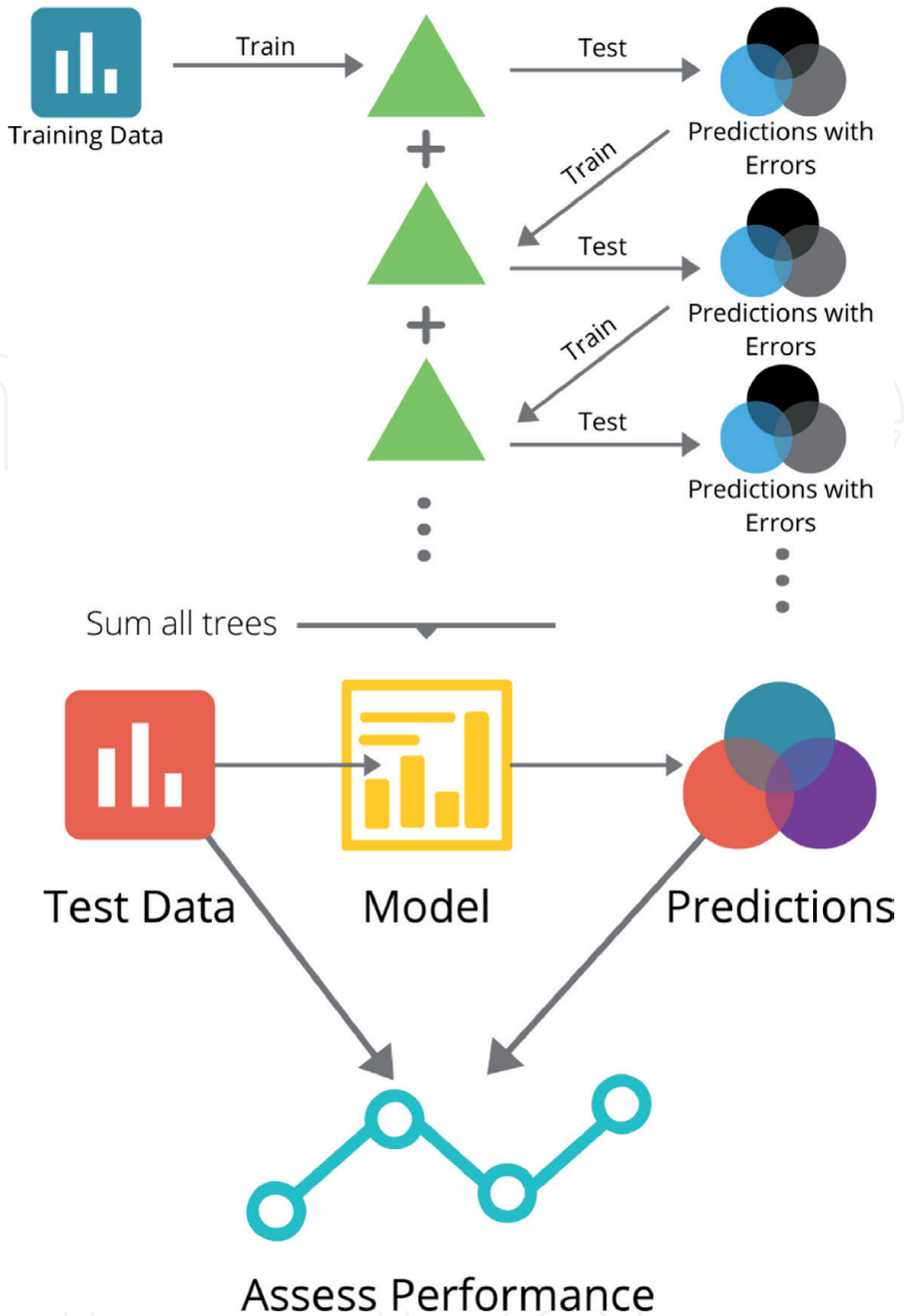

Figure 3.
Example of a simple workflow for training and assessing a gradient boosting machine model. Each green triangle represents a trained tree from the training data with the subsequent tree using the residuals or errors from the prior tree to improve its prediction. The prediction of each tree is summed and is represented as the model. Test data is then fed to the model, i.e., all the trees, and the resulting prediction is made. The prediction is then compared to the original test data to assess how the model performs.

immediately obvious. However, they are also more computationally intensive and require a larger amount of data to perform.

The most common and well-known algorithms are K-means clustering and deep learning, though deep learning can be used in a supervised manner [12, 20]. Such algorithms also perform association tasks which are similar to clustering. These algorithms are considered unsupervised because there is no human input as to what set of attributes the clusters will be centered on.

The typical k-means algorithm has several variations such as k-medians and k-medoids, however the principle is the same for each algorithm. The algorithm uses Euclidian distance to find the "nearest" center or mean for a cluster assuming there are k clusters. It then assigns the current data point to that cluster and then

recalculates the center for the cluster, updating it for the next data point [21]. The biggest drawback to this algorithm is that it must be initialized with an expected number of "means" or "centers." Improper selection of the *k* value can result in poor clustering.

Deep learning uses neural nets to perform predictions even on unlabeled data as well as classification techniques. Based off models of human neurons, perceptrons, as they are typically called, are organized into many networked layers making the network "deep" in nature [20]. Each perceptron has multiple inputs and a single output. They are organized into layers where the outputs of the previous layer serve as the inputs for the next layer. The input layer requires one perceptron per input variable and the subsequent layers are determined before training by a human (**Figure 4**). This is one of the difficulties and challenges in building an effective neural net. The computationally intensive nature of computing each perceptron for a large neural net can mean that training alone can take days to weeks for large data sets [22].

2.1.3 Hyperparameters

In machine learning, a model typically has a set of parameters as well as a set of hyperparameters. Parameters are variables about the model that can be changed during training. For example, parameters can be the values of the training data itself with each piece of data being different along one or several of the parameters. Whereas hyperparameters are typically set before training occurs and cannot change once learning begins. Hyperparameters typically are set to tune values like the model's learning speed and constrain the algorithm itself.

Different algorithms will have different sets of hyperparameters. For example, a common hyper parameter for artificial neural networks is the number of hidden layers. Additionally, a separate but related hyperparameter is the number of perceptrons in each hidden layer. Whereas a similar equivalent in decision trees would

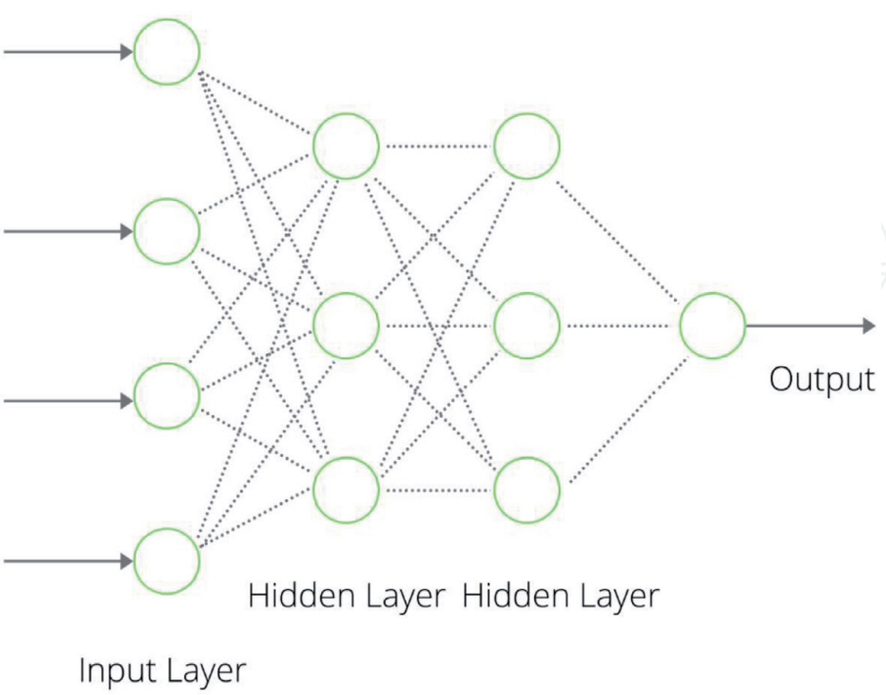

Figure 4.
Example of a simple neural net with two hidden layers of three perceptrons each. The number of inputs, number of hidden layers, and number of perceptrons in each layer can be changed. Additionally, the connections between layers and perceptrons can also be changed.

be the maximum number of leaves in a tree or the maximum depth for a tree. Other common hyperparameters include learning rate, batch size, dropout criterion, and stopping metric.

Properly selecting hyperparameters can significantly speed up the search for a proper generalized model without sacrificing performance. However, in many cases finding the proper set is more of an art than a science. Many researchers have attempted to make hyperparameter searching a more efficient and reproducible task [23–25]. Again, this process also highly depends on the algorithm, dataset, and problem you are trying to solve. A machine learning model can be tuned a nearly infinite amount of different ways to achieve better performance. Hyperparameters represent a way to reproduce results and also serve as a tool to properly validate models.

2.1.4 Algorithm principles

Considering the pace of research in the field, there are constant advances and improvements to many of these machine learning techniques, but the important thing to remember is that not all algorithms work for all use cases. Each algorithm has advantages and disadvantages. Certain data types may also affect the performance of individual algorithms and the time spent implementing such models will often be a result of testing different variations, parameters, and hyperparameters within these algorithms to achieve the best generalized performance.

2.2 Assessment of model performance

The goal of any machine learning algorithm is to utilize real data to create a model that performs the best on real-world scenarios, and that can be assessed in a quantitative, reproducible manner. Assessment of statistical models is a whole subfield in itself, but we will briefly discuss the basics, which are applicable for almost any machine learning algorithm you will come across.

2.2.1 Sensitivity vs. specificity

Sensitivity and specificity are two important metrics used in a statistical or machine learning model to assess if the model is performing successfully. As such, it is important to understand what each of these numbers tell us about what a trained model can do, and what the model cannot do.

Sensitivity is the probability that a positive result occurs given that the sample is indeed positive. Mathematically,

$$Sensitivity = \frac{(Number\ of\ True\ Positives)}{Number\ of\ True\ Positives + Number\ of\ False\ Negatives}$$

This is also sometimes referred to as the recall or hit rate, or just simply the true positive rate, and the sensitivity is equivalent to 1 – *False Negative Rate*.

Specificity is the probability of a negative result given that the sample is negative. Mathematically,

$$Specificity = \frac{Number\ of\ True\ Negatives}{Number\ of\ True\ Negatives + Number\ of\ False\ Positives}$$

This value is also referred to as the selectivity of the test. This is equivalent to 1 – *False Positive Rate*.

2.2.2 The receiver operator curve and area-under the curve

The standard metric for assessing the performance of machine learning models is known as the receiver operating characteristic (ROC). The ROC can be summarized by a number from 0 to 1, which is the measured area-under-the-ROC curve (AUC). The ROC curve is a 2D plot that measures the false positive rate vs. true positive rate. There are four numbers that are used to determine the effectiveness of a test: true positive rate, false positive rate, true negative rate, and false negative rate.

True positive and true negative are the correct answers to a test while false positive and false negative are incorrect answers to the test or model. These numbers can be condensed further into two numbers known as sensitivity and specificity. We have already discussed sensitivity and specificity but now we will discuss how they are used to create the ROC.

Ideally a test would have both high sensitivity and high specificity. However, there is a tradeoff, prioritizing one often leads to the detriment of the other. When setting the threshold low, one will receive a high true positive rate (high sensitivity) and a high false positive rate (low specificity). Conversely, setting the threshold high will result in a low true positive rate (low sensitivity) and a low false positive rate (high specificity).

The ROC and AUC metric is used to characterize most of the classification tasks many machine learning models are attempting to do; does this person have the disease or do they not? If a test has a high sensitivity and a high specificity it is considered a near perfect test and the AUC is close to 1 (**Figure 5**). If the test is random then the AUC is 0.5. The x-axis is typically the false positive rate (or 1 – specificity). Ideally, the false positive rate is as low as possible. The y-axis is typically the true positive rate (sensitivity). The sensitivity is what is usually maximized. On a typical curve, the midpoint of the curve is the most balanced trade-off between sensitivity and specificity though this is not always the case. The AUC value is a simpler, more generalized way, to assess the performance rather than the varying tradeoffs between sensitivity and specificity.

Another way to think of AUC is as a percentage the model can correctly identify and separate a positive result from a negative result. Given an unknown case, a model with an AUC of 0.75 has a 75% chance of correctly identifying whether the case is a positive case or a negative case. This number will quickly tell you the results of any model.

2.2.3 Overfitting

Overfitting is one of the main concerns when training any model [26]. Simply put, when training a model on a set of data, over-training the model will improve the performance of the model on that specific dataset but at the cost of losing generalization to other datasets. An overfitted model will not work when applied to new data it has never seen before. From a practical standpoint, such a model is not very useful in a real-world application.

When training any machine learning model, the ideal result is a *generalized* model. A generalized model works well on a variety of different cases and a variety of different datasets, especially data it has never seen before. As such, many researchers are hesitant to give too much credence to a model or method that utilizes a single dataset.

A variety of methods have been used to prevent models from overfitting and many of these are now encapsulated in the *hyperparameters* discussed earlier.

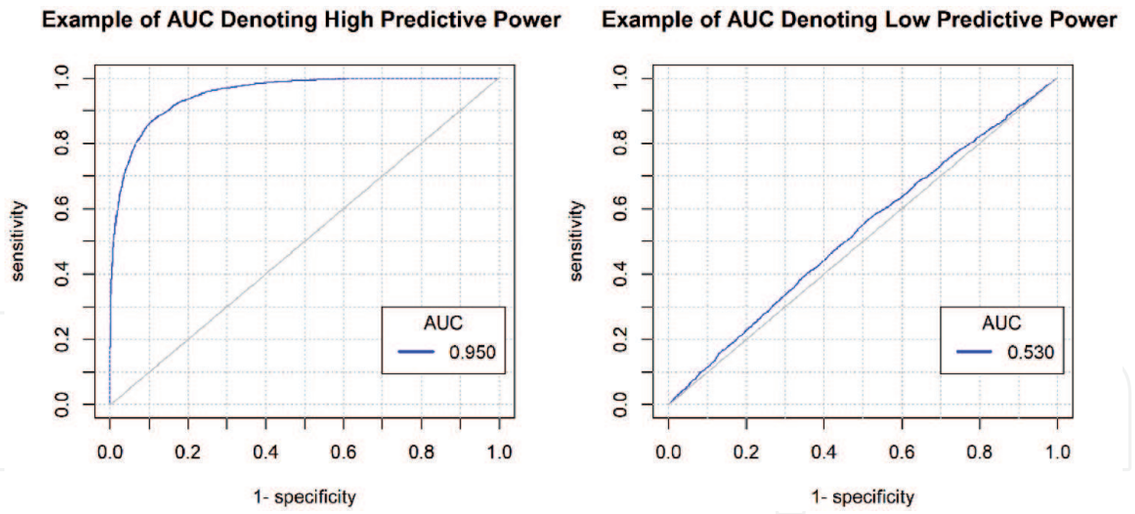

Figure 5.
Examples of an AUC denoting a model which has good predictive power (left) and an AUC denoting a model with poor or near random predictive power (right). 1 – Specificity is sometimes written as false positive rate (fpr) and sensitivity can be read as true positive rate (tpr).

The idea is to prevent the models from adapting too quickly to the dataset it is being trained on. This subset of methods is known as *regularization* [27].

One such method, used in neural nets, is called dropout. This method is widely used to prevent artificial neural nets from overfitting during classification tasks. The method is fairly simple. During the training process, random perceptrons and their corresponding connections are "dropped" from the network. These "thinned" networks have better performance compared to other regularization techniques on supervised learning tasks [28].

Often a method known as cross-validation is used to assess the performance and validate the *generalized* predictive ability of a model. The most common method for building machine learning models is the partitioning of the data set into roughly 80% for training and 20% for testing. This partition is typically less useful for linear models but splitting is more beneficial for complex models [29]. During cross-validation, this split is done in separate sections of the data to ensure proper coverage. For example, if a 10-fold cross-validation is performed, the first split in a data set with 100 observations could be a the first 80 for training and the last 20 for test, the second split could be the first 10 and last 10 for test and the middle 80 for training, etc. (**Figure 6**). This creates 10 models using the same algorithm just trained and tested on different portions of the same data. The average performance of these 10 models gives a good measurement of the generalized performance of the algorithm on that type of data.

2.3 Big data and the health information explosion

The healthcare sector has always had a very large amount of information, often times stored as physical documents in clinics, hospitals, regulatory agencies, and biomedical companies [30, 31]. With the push to electronic medical records (EMR), this information is rapidly being transformed into a form which can be leveraged by AI technologies. The estimated amount of healthcare data stored in 2011 was around 150 exabytes (1 EB = 10^{18} bytes), though that number is most likely exponentially larger almost a decade later [32, 33]. These large databases, when in a digitized form, are often known as Big Data.

However, such healthcare information is very different in both form and function. Visual data in the form of medical images is very different than familial history

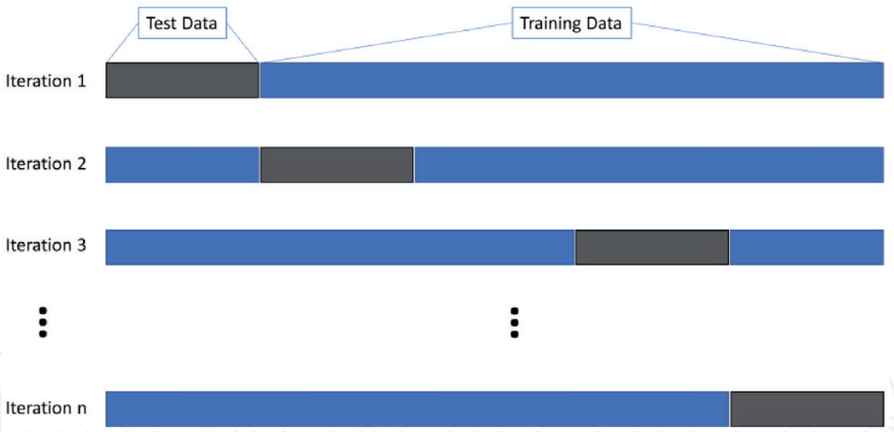

Figure 6.
Example of a set of cross validation splits. There are n splits for the number of iterations desired and the results of all iterations are averaged to assess the generalized performance of a model trained on a dataset.

which may be simple text-based information. Laboratory and clinical tests may be reported as numbers only, while health outcomes are often qualitative in nature and may be a simple yes or no entry in a spreadsheet. Insurance and administrative data is also indirectly linked to various information, such as patient outcomes, while information from sensor based technologies like EKGs, pulse oximeters, and EEG provide time-series data of vital signs [34].

Additionally, the genomic revolution has contributed enormously to the data explosion. Large-scale genetic databases such as the Cancer Genome Atlas (TCGA) and the UK Biobank include thousands of patients' genetic sequencing information along with various other health information such as disease state, age of diagnosis, time of death, and much more [35–38]. Copy number variation (CNV) data from the UK Biobank's roughly 500,000 patients, which does not even contain the raw sequence reads, is almost 2 Terabytes (TB) alone in flat text files. These genetic databases rely on an array of assays and sequencers spread across different hospitals

Database	Size of data	Number of participants	Status	Start date
The Cancer Genome Atlas	2.5 petabytes	11,300 [36, 41]	Completed	2005
The UK Biobank	~26 terabytes*,+	~488,377 [42]	Ongoing	2006
The European Prospective Investigation into Cancer and Nutrition (EPIC)	Unclear*	~521,000 [43]	Ongoing	1992
Estonian Genome Project	Unclear*	~52,000 [44]	Ongoing	2007
deCODE	Unclear*	~160,000 [45, 46]	Ongoing	1998
China Kadoorie Biobank	Unclear*	~510,000 [47, 48]	Ongoing	2004
Lifelines Cohort Study	Unclear*	~167,000 [49]	Ongoing	2006
All of Us (Precision Medicine Initiative)	Unclear*	Currently ~10,000, planned 1,000,000 [12, 50–52]	Ongoing	2015
FinnGen	Unclear*	Planned 500,000 [53]	Ongoing	2017

**Project is continuing to collect more data.*
+Number represents genetic data only. Project or study may also include unreported data including medical images and health records.

Table 1.
Overview of largest biobank databases as of 2019.

and research facilities around the globe, before being processed and transferred to their respective centralized storage databases [35, 39, 40].

The collection of biological data and creation of these databases show no evidence of slowing down. Many biobanks, databases which contain some form of biological samples such as blood or serum, contain thousands of participants and many have plans to collect hundreds of thousands of samples from patients (**Table 1**). Because many databases are growing so quickly it is unclear how much data resides in many of these databases. However, The Cancer Genome Atlas alone contains 2.5 petabytes (1 PB = 10^{15} bytes) of data and the UK Biobank contains 26 terabytes (1 TB = 10^{12} bytes) of just genetic information (UK Biobank also contains medical images such as brain scans which is not included in this table).

Implementing machine learning systems into a hospital with this complex information Web is usually slow, due to the abundance of caution needed to ensure patient health. Many physicians are also wary of adopting new systems that are unproven in a clinical setting due to the risk of litigation and potentially catastrophic consequences for their patients.

3. Machine learning of medical images

Modern medical images are digital in nature. To effectively utilize them in healthcare there are several challenges that must be overcome. Medical imaging describes a collection of techniques to create visual representations of interior portions of the human body for the purpose of diagnosis, analysis, and medical intervention. This is beneficial in avoiding or reducing the need for the older clinical standard of exploratory surgery. Since opening any portion of the human body through surgical means increasing the risk of infections, strokes, and other complications, medical imaging is now the preferred tool for initial diagnosis in the clinical setting.

The current clinical standard of assessing medical images is the use of trained physicians, pathologists, or radiologists who examine the images and determine the root cause of clinical ailments. This clinical standard is prone to human error and is also costly and expensive, often requiring years or decades of experience to achieve a level of understanding which can consistently assess these images. Considering that the demonstration of viable machine learning capabilities in the modern age was demonstrated by Andrew Ng using images pulled from YouTube videos, it is clear why medical images were one of the first areas addressed during the initial adoption of machine learning techniques in healthcare [54].

Accuracy of diagnosis is extremely important in the medical field as improper diagnosis could lead to severe consequences and results. If a surgery is performed where none was needed or a misdiagnosis leads to improper dosages of prescribed medication, the possibility of a fatal outcome increases. In the realm of image processing, most techniques rely fundamentally on deep learning (DL) and specifically in artificial neural networks (ANNs). Modern techniques utilize improvements to ANNs in the form of convolutional neural networks (CNNs) to boost performance when classifying images.

The majority of the current publications are using some form of CNNs when it comes to object detection in medical images [55]. Graphic-processing unit (GPU) acceleration has made the building of deep CNNs more efficient, however significant challenges in creating a competent model still exist. The biggest issue is the need for a large amount of annotated medical image data. The cost to aggregate and create such databases is often prohibitive since it requires trained physicians' time to annotate the images. Additionally, concerns involving patient privacy often hinders

the ability to make such databases open-source. Many studies only use around 100–1000 samples in training CNNs. This limited sample size increases the risk of overfitting and reduces the accuracy of the predictions [56].

Concerns regarding the implementation of machine learning into clinical diagnosis have been raised regarding proper validation of models [57]. The main fears entail properly scoping the intended goals of a machine learning model, reducing dimensionality of the data, and reproducibility of training such models on real-world and new clinical data. Validating results on other datasets can be difficult due to the lack of larger datasets for niche diseases, where the aggregation of this data can take more work than the actual training of the model. Medical imaging data is inherently more difficult to acquire and is more difficult to store and process. The infrastructure to handle the data has simply not kept up with the increase in the amount of data.

3.1 Lesion detection and computer automated detection

The most common use of current machine learning technologies in medicine is for computer automated detection (CAD) specifically in the detection of lesions such as those commonly found in mammograms, brain scans, and other body scans [58]. These methods use CNNs to arrive at the probability that a candidate lesion is in fact a lesion, often utilizing several 2D slices of 3D rotational scans of either CAT or MRI images.

Ultrasound images are also used in training and a variety of methods such as randomized rotation of the images or centering candidate lesions in the center of the image. Especially in mammography, CAD techniques have reached a level where they are used as a "second opinion" for most radiologists, greatly improving the accuracy of screenings without doubling the cost associated with using a human as the "second opinion" **Figure 7**.

CAD is also currently split into detection and diagnosis. This distinction is subtle but important. A lesion can be categorized as either benign or malignant, based off a physician's knowledge and assessment. However, the actual detection is a crucial first step in treating a patient.

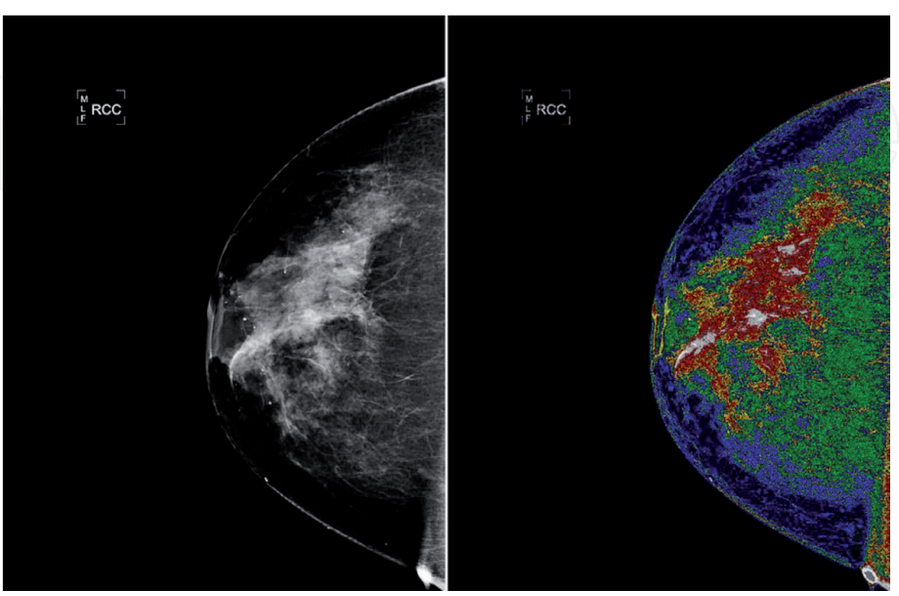

Figure 7.
Example of mammogram with the left image being that of a raw mammogram and the right hand being the image with the detection overlaid with the region of interest in white, using NASA software originally used to enhance earth science imagery. Taken from NASA press release, credited to Bartron Medical Imaging.

Computer aided detection is the actual recognition of potential lesions from a medical image. For example, detection and segmentation of glioblastoma is a difficult task, due to the invasive and widespread nature of these tumors. Unlike other brain tumors, they are not easily localized and assessing how treatments such as chemotherapy are performing is in itself a difficult task. Deep learning has aided in this by helping automate assessment of glioblastoma MRIs [59].

Computer aided diagnosis describes the probability a lesion is malignant in nature. These methods are primarily used to improve the accuracy of diagnosis and improve early diagnosis in the clinical setting. Again, these tasks have consistently been performed by machine learning especially in brain related applications, due to the difficult nature of assessing brain health. Additionally, diagnosis of Alzheimer's through medical imaging is a possible application for deep learning which is showing some promise [60, 61].

4. Natural language processing of medical documents and literature

Electronic medical records (EMR), the new standard in many hospitals, require complex digital infrastructure. Unification of health data in a formatted manner is a major goal as it should increase the efficiency of hospitals as well as improve patient health outcomes. However, a significant problem is the historical existing physical documentation. Transferring these existing documents into an electronic form is difficult and would be very tedious and expensive if people were hired to manually input such information into an electronic system.

One application of machine learning, which may aid in this problem, is natural language processing (NLP). By scanning these documents rapidly and integrating the resulting images into a database, these systems attempt to extract readable data from free text and incorporates image processing to identify key words and terms. Handwritten physician notes contain information such as patient complaints, the physicians own observations, and patient family history. This clinical information can be annotated. However, poorly worded or inaccurate writing by the physician can make it difficult to accurately assign this information to appropriate categories. Forms and documents that already have structure make for much easier language processing, though there is still the risk of missing data **Figure 8**.

Creating a system for improved clinical decision support (CDS) with old patient records is feasible. Any such system is structured to aid in clinical decision making for individual patients based on a database of computerized knowledge. Such a system could be envisioned as two-fold: 1. extracting facts about the patient from their medical record, either through written or typed physician notes or labs or dictation involving audio NLP, 2. Associating possible disease states based on extracted information from previous known cases or through literature search via NLP [62]. Integration of several specialized NLP systems is required for any true and practical implementation of such a CDS system.

Likewise, compilation of the existing scientific research into central repositories is a difficult task. Sometimes physicians may be unaware of a promising new treatment just due to the difficulty of parsing the tidal wave of new papers. Scientific publications have always been widely dispersed across multiple journals and the modern-day information explosion has only exacerbated the issue. When it comes to compiling information such as results from genome-wide association studies (GWAS), the primary method has been a manual curation of the information by certain individuals within the scientific community: "librarians" so to speak.

Recently, a paper published in Nature Communications used machine learning systems to automatically compile GWAS information from open-access publications

Nursing Care Plan

CLIENT ID:
NAME:
D.O.B.:
DOCTOR:
PENSION:

LIFESTYLE ISSUES	GOAL OF CARE	CARE OR INTERVENTION REQUIRED Tick and/or Highlight Appropriate Response
Links to Assessment: Baseline Health Assessment *Communication assess (11-04)*		Hearing Hearing loss: Partial ☐ Profound ☐ Right ☐ Left ☐ Both ☐ Aids used: Behind the ear ☐ Inside the ear ☐ Right ☐ Left ☐ Both ☐ Storage details: ..... Assistance Required? ☐ Yes ☐No Are hearing aids wom ☐ Yes ☐No Battery type ☐ .....
Notes		

LIFESTYLE ISSUES	GOAL OF CARE	CARE OR INTERVENTION REQUIRED Tick and/or Highlight Appropriate Response
SKIN INTEGRITY: **Links to Assessment:** *Waterlow Risk Assessment (11-06)* *Wound Management-Complex (11-07a)* *Wound Management-Simple (11-07b)*	**RESIDETS SKIN INTEGRITY IS MAINTAINED AT OPTIMUM LEVEL**	☐ Bed/Chair bound: ..... ☐Repositioning frequency ☐2nd hourly ☐3rd hourly ☐4th hourly ☐Other ☐ Aid (in Bed): ..... ☐ Aid (on chair): ..... ☐ Creams: ..... ☐ Leg Protectors: ..... ☐ Bed Rail Protectors: Yes☐ No☐ ☐ Other: ..... Refer to Wound Management Charts as necessary

Name	**Designation**
Signature	**Date**
Notes	

Figure 8.
Example of a nursing care plan which represents a formatted health document. Most of these plans were filled out by hand and many hospitals have transitioned such forms to electronic records. However, older documents still need to be transferred to digital form. Taken from Wikipedia commons.

and extract GWAS associations into a database with the aim of helping curators. Though the results are somewhat inconsistent (60–80% recall and 78–94% precision) it represents one of the many ways NLP is being utilized to aid in medical discovery [63].

4.1 Examples of natural language processing in healthcare research

There are many exciting possibilities where NLP could be used to improve medicine and medical research. We will discuss a few interesting findings with similar approaches but different goals. This is by no means an expansive list but highlights the broad spectrum of possible machine learning applications.

In 2015, a research group published a paper reporting 100% accuracy of predicting onset of psychosis using recorded dialog of clinically high-risk youth. Each youth was interviewed over a period of 2.5 years every 3 months. Based on the

transcripts of these interviews, a machine learning algorithm was trained to predict whether a patient would develop psychosis. This was done using what is known as Latent Semantic Analysis to determine coherence of speech using NLP. The sample size for this study was rather small however (n = 34) [64].

Another study used NLP to identify cirrhosis patients and risk-stratify the patients. This study was able to correctly identify cirrhosis patients from electronic health records, ICD-9 code combinations, and radiological scans with a 95.71% sensitivity and 93.88% specificity [65]. This indicates that such a system could correctly identify cirrhosis patients based off existing medical data in most hospitals.

Yet another study used NLP to accurately identify reportable cancer cases for national cancer registries. This method analyzed pathology reports and diagnosis codes to identify patients with cancer patients using supervised machine learning. The accuracy was 0.872 with a precision of 0.843 and sensitivity of 0.848 [66]. The primary goal of this study was to automate the process of reporting cancer patients to the National Program of Cancer Registries in the United States.

These examples of NLP use in healthcare highlight the wide diversity of applications within medicine. Language is the primary means of communicating complex information, doctors' notes and annotated medical documents hold valuable insights in populations and individual patient health. The irregularity and variance of language and extraction of higher-level information into relevant subcategories makes analysis difficult. Machine learning is showing promising results in performing such complex analyses.

5. Machine learning in genetics for the prediction and understanding of complex diseases

Genetic information and technologies have exploded since 2008, creating difficult challenges in how to handle the exponentially increasing data. Advances in genetic sequencing speed, namely NGS technologies have exponentially increased the speed at which a whole human genome is sequenced, while also dramatically reducing costs. The human genome is a complex physical structure that encodes all the information of human development and characteristics. The genome is highly interconnected and deciphering most of these instructions is still a mystery to us. Variation of genomes between people also increases the complexity of understanding gene interactions.

Many health initiatives have focused on acquiring large sample sizes of human genomes to help identify statistically relevant trends among different populations of humans. However, the 23 chromosomes of the human genome contain around 20,000 genes which have been identified as the primary coding sequences for the proteins necessary in building the biological components of our cells [67]. This number is still a rough estimate and some estimates indicate that there may be as many as 25,000 genes or as few as 19,000 [68, 69]. A large swathe of genetic information that does not code for any proteins is not included in these estimates.

A growing body of literature indicates that certain sections of what has been colloquially called *genetic dark matter*, or *missing heritability*, exists [70–74]. These terms refer to the portions of DNA which have no apparent protein coding function, but may be relevant to the level of gene expression in a person's genetic code [75, 76]. Levels of gene expression may cause protein overload or deficiency, which can lead to a variety of health problems. Additionally, structural differences in the physical structure of how the DNA is bound into chromosomes and then subsequently unwrapped during both the duplication process and translation and transcription process, can also affect the level of gene expression.

For example, methylation or acetylation of the DNA backbone can make it more difficult (methylation) or easier (acetylation) to unravel the DNA strand during normal cell processes like replication or protein assembly. Evidence of multiple copies of the same gene have also been classified in what is described as copy number variations (CNV) which indicate duplication, triplication, and deletion events of certain areas of the genome in an individual. Understanding this highly interconnected and nonlinear relationship between all the different of the areas of the human genome is difficult.

With machine learning, scientists have begun to find patterns and trends which can be modeled in a more predictable manner. Utilizing the ever-growing amount of genetic data, machine learning has the potential of accurately predicting who is at risk of acquiring certain diseases such as cancers and Alzheimer's disease. Mental illnesses such as schizophrenia and bipolar disorder have also been known to run in families, indicating a possible genetic link.

5.1 Inherited vs. environmental risk

Disease risk can be broadly categorized into inherited risk and environmental risk. Inherited risk describes a person's disposition to acquiring complex diseases due to a trait which is genetically passed down from their predecessors. This includes genetic mutations contained within their germline DNA which may predispose them to cancers or other health conditions [77, 78].

Environmental risk describes somatic mutations, or mutations to a person's DNA due to something they have encountered in their environment. These mutations can still increase a person's risk of acquiring a disease but they do not affect the germline, and will not be passed on to their progeny and thus will not be inherited [79].

Inherited risk describes mutations that exist in the human germline and which will be passed onto the offspring through normal reproduction. Whereas, somatic mutations may affect organs or a set of cells, germline mutations exist in all the cells of the offspring. Many of these mutations may be passed through paternal lineage and there is some indication that certain individuals may have disease predisposition but which cannot be directly linked to familial history but could still be due to these hidden germline mutations [80–82].

Several different types of mutations may exist within a human genome. They are broadly categorized as single nucleotide polymorphisms (SNPs), structural variations or copy-number variations (CNVs), and epigenetic variations.

SNPs are a single or point mutation of one base pair in the human genome that occurs in at least 1% of the human population [83, 84]. These mutations are the most common source of genetic variation and can occur both within coding regions and outside of coding regions of the genome. SNPs contribute to vast differences even between relatives and can arise because of both inheritance and development in the womb. Within SNPs there are common and rare variants, with rare variants occurring less than 0.5% within the global sample [84].

Structural variations and specifically CNVs are deletions, insertions, duplications, and inversions of large regions of DNA. These structural differences are usually inherited and a typical human can have anywhere between 2100 and 2500 structural variations [84]. These variations were found to cover more of the human genome than SNPs alone [83].

Epigenetic variation describes variations in the chemical tags attached to DNA or associated structures such as histones, which affects how genes are read and activated. Epigenetics includes DNA methylation and acetylation, histone modifications, and non-coding RNAs which all affect the degree to which a gene may be expressed [85]. As a newer field, it is unclear how much of these epigenetic

Type of cancer	Cases	Controls	AUC
Breast invasive carcinoma (men and women)	977	8821	0.81
Glioblastoma multiforme	484	9314	0.86
Ovarian serous cystadenocarcinoma	424	4268	0.89
Thymoma	111	9687	0.78
Uveal melanoma	80	9718	0.80

Table 2.
Sampling of performance of GBM models trained on data from the Cancer Genome Atlas.

variations are inherited from generation to generation, and how much is a result of environmental factors [86].

5.2 Prediction of cancers through germline copy number variations

One of the exciting methods we have discovered is the utilization of germline copy number variations in the prediction of different cancers. We have found that it is possible to use machine learning models, specifically gradient boosting machines (GBM), a form of decision trees (DT), to predict whether a person has a particular cancer. The models created were able to predict cancers such as ovarian cancer (OV) and glioblastoma multiforme with an AUC of 0.89 and 0.86 respectively [87], using copy number variation data taken from germline blood samples only. This result indicates that there is a significant inherited portion contributing to cancer risk in many, if not all cancers. Since these CNVs are also taken from germline DNA, the likelihood of continued inheritance to future generations is high.

This method does not look solely at SNPs as many previous methods rely on [88]. Most SNP data specifically looks at mutations within protein coding genes while ignoring the rest of the genome, whereas our method utilizes a whole genome approach by averaging the copy numbers of a person's entire genome as the basis for predicting cancer. Copy number variation accounts for a large amount of human genetic diversity and is functionally significant though the exact mechanisms are still unclear [77, 83].

These results demonstrate that almost all cancers have a component of predictability in germline CNVs which can be used to predict an individual's risk to acquiring that cancer **Table 2**. Experiments were performed on two independent databases: The Cancer Genome Atlas and the UK Biobank. The first database contains about 10,000 individuals and latter contains about 500,000 individuals.

Future studies may improve on the performance and the models could potentially be used as a tool to assess individual risk for diseases. Since the method can also be easily generalized to other diseases, we anticipate work to continue to encompass other potentially complex diseases which may have inherited components to them.

6. Conclusions

Application of digital technologies such as machine learning in the healthcare field is entering an exciting era. The collision of informatics, biology, engineering, chemistry, and computer science will rapidly accelerate our knowledge of both hereditary and environmental factors contributing to the onset of complex diseases. The potential of utilizing copy number variations in the prediction of cancer

diagnosis is exciting. Utilizing machine learning to create an interpretable method of understanding how the genomic landscape interlinks across genes to contribute to inherited cancer risk could potentially improve patient healthcare on an individual level.

Databases such as The Cancer Genome Atlas and UK Biobank are invaluable resources, providing high statistical power to scientific analysis. As other large-scale population data projects near completion in the coming decade, the methods laid on the foundation of The Cancer Genome Atlas and UK Biobank will continue to benefit and improve as sample sizes easily begin to move into the regime of millions of patients. Tracking populations around the world will truly aid in the goal of precision medicine.

Natural language processing will be essential in improving the practicality of translating scientific findings and results of other machine learning methods into a clinical setting. Multiple specialized systems will have to be integrated with each other to effectively extract the wealth of information into a format which can be utilized effectively by physicians and healthcare professionals.

Image analysis is becoming a staple in many diagnostic endeavors and will continue to improve the accuracy of radiological diagnosis. Detection of malignant masses and validation and verification of existing diagnosis has the potential to improve patient outcomes, while reducing errors. As a non-invasive method of looking inside the human body, any improvements in healthcare imaging will reduce the need for risky or ill-informed operations that could lead to other complications such as infections and blood clots.

The examples discussed in this chapter are some of the most promising works in applying machine learning in the healthcare field. Resolving big health data into a usable form will undoubtedly require machine learning techniques to improve. Infrastructure to support such learning techniques is currently not stable or standardized. Bringing such methods from concept to practical clinical use is contingent on both validation of these results and an appropriate infrastructure to support it.

A large variety of devices and storage methods will need to be unified and standardized to benefit from the increased data collection. Information about how human genetic variation can contribute to individual susceptibility allows patients and doctors to make early lifestyle changes in a preventative manner. Likewise, it can inform physicians of which types of prognostics and diagnostics would be the most relevant for a specific patient, saving both time and money, while improving patient outcomes in the long term. Just as AI started with Turing decoding the enigma machine, we are now going to use AI and machine learning to decode the secrets of the human body and genome.

Conflict of interest

The corresponding author is a distant relative of the editor of this book.

Notes/thanks/other declarations

The author would like to thank the University of California, Irvine for support during the writing of this chapter.

IntechOpen

Author details

Christopher Toh* and James P. Brody
University of California, Irvine, United States of America

*Address all correspondence to: tohc@uci.edu

IntechOpen

References

[1] Copeland J. The turing test. Minds and Machines. 2000;**10**(4):519-539

[2] French RM. The turing test: The first 50 years. Trends in Cognitive Sciences. 2000;**4**(3):115-122

[3] Edwards S. World war II at sea: A global history. The Journal of American History. 2019;**106**(1):237

[4] Turing AM. Computing machinery and intelligence. Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer. 2009. p. 23-65

[5] Samuel AL. Some studies in machine learning. IBM Journal of Research and Development. 1959;**3**(3):210-229

[6] Samuel AL. Programming computers to play games. Advances in Computers. 1960;**1**(C):165-192

[7] Fukushima K. Cognitron: A self-organizing multilayered neural network. Biological Cybernetics. 1975;**20**(3-4):121-136

[8] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics. 1980;**36**(4):193-202

[9] Huang K, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al. Pathogenic germline variants in 10,389 adult cancers. Cell. 2018

[10] Silver D, Hassabis D. AlphaGo: Mastering the ancient game of go with machine learning. Google Research Blog. 2016

[11] Brewka G. Artificial intelligence—a modern approach by Stuart Russell and Peter Norvig, Prentice Hall. Series in Artificial Intelligence, Englewood Cliffs, NJ. The Knowledge Engineering Review. 1996;**11**(1): 78-79

[12] Alpaydin E. Introduction to Machine Learning. London: The MIT Press. 2014;**3**:640

[13] Kotsiantis SB. Supervised machine learning: A review of classification techniques. Informatica (Ljubljana). 2007;**31**(3):249-268

[14] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer Series in Statistics. Switzerland: Springer. 2009;**2**(1):93-85

[15] Ho TK. A data complexity analysis of comparative advantages of decision forest constructors. Pattern Analysis & Applications. 2002;**5**(2):105-112

[16] Friedman JH. Stochastic gradient boosting. Computational Statistics and Data Analysis. 2002;**38**(4):367-378

[17] Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001;**29**(5):1189-1232

[18] Mason L, Baxter J, Bartlett P, Frean M. Boosting algorithms as gradient descent in function space. NIPS Conference Proceedings. 1999:512-518

[19] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews. Genetics. 2015;**16**(6):321-332

[20] Längkvist M, Karlsson L, Loutfi A. A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters. 2014;**42**(1):11-24

[21] Lloyd SP. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982;**28**(2):129-137

[22] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015;**61**(1):85-117

[23] Hazan E, Klivans A, Yuan Y. Hyperparameter optimization: A spectral approach. In: In: 6th International Conference on Learning Representations, ICLR 2018; Conference Track Proceedings. 2018

[24] Bardenet R, Brendel M, Kégl B, Sebag M. Collaborative hyperparameter tuning. In: 30th International Conference on Machine Learning; ICML. 2013. p. 2013

[25] Hutter F, Hoos H, Leyton-Brown K. An efficient approach for assessing hyperparameter importance. In: 31st International Conference on Machine Learning; ICML. 2014. p. 2014

[26] Hawkins DM. The problem of Overfitting. Journal of Chemical Information and Computer Sciences. 2004;**44**(1):1-12

[27] Ng A. Regularization—Solving the Problem of Overfitting. Coursera; 2011. Available from: https://www.coursera.org/learn/machine-learning/lecture/ACpTQ/the-problem-of-overfitting

[28] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research. 2014;**15**(56):1929-1958

[29] Picard RR, Cook RD. Cross-validation of regression models. Journal of the American Statistical Association. 1984;**79**(387):575-583

[30] Jothi N, Rashid NA, Husain W. Data mining in healthcare—A review. Procedia Computer Science. 2015;**72**(1):306-313

[31] Koh HC, Tan G. Data mining applications in healthcare. Journal of Healthcare Information Management. 2005;**19**(2):64-72

[32] Cottle M, Hoover W, Kanwal S, Kohn M, Strome T, Treister NW. Transforming Health Care through Big Data: Strategies for Leveraging Big Data in the Health Care Industry. Institute for Health Technology Transformation. New York, iHT2; 2013

[33] Wang Y, Kung LA, Byrd TA. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change. 2018;**126**(1):3-13

[34] Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Information Science and Systems. 2014;**2**(1):3

[35] The Cancer Genome Atlas Program—National Cancer Institute [Internet]. 2019. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[36] tcga-infographic-enlarge.__v100169753.png (1400×2580) [Internet]. 2019. Available from: https://www.cancer.gov/PublishedContent/Images/images/nci/organization/tcga/tcga-infographic-enlarge.__v100169753.png

[37] Peakman TC, Elliott P. The UK biobank sample handling and storage validation studies. International Journal of Epidemiology. 2008;**37**(2):234-244

[38] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine. 2015;**12**(3)

[39] Protocol for the UK Biobank—Wayback Machine [Internet]. 2019. Available from: https://web.archive.

org/web/20060214144838/http://www.ukbiobank.ac.uk/docs/draft_protocol.pdf

[40] Elliott P, Peakman TC. The UK biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. International Journal of Epidemiology. 2008;**37**(2):234-244

[41] Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, et al. Before and after: Comparison of legacy and harmonized TCGA genomic data commons data. Cell Systems. 2019;**9**(1):24-34.e10

[42] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK biobank resource with deep phenotyping and genomic data. Nature. 2018;**562**(7726):203-209

[43] Background—EPIC [Internet]. 2019. Available from: http://epic.iarc.fr/about/background.php

[44] Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, et al. Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. International Journal of Epidemiology. 2015;**44**(4):1137-1147

[45] Master Decoder: A Profile of Kári Stefánsson | The Scientist Magazine® [Internet]. 2019. Available from: https://www.the-scientist.com/profile/master-decoder--a-profile-of-kri-stefnsson-65517

[46] Gulcher J, Stefansson K. Population genomics: Laying the groundwork for genetic disease modeling and targeting. Clinical Chemistry and Laboratory Medicine. 1998;**36**(8):523-527

[47] China Kadoorie Biobank [Internet]. 2019. Available from: https://www.ckbiobank.org/site/

[48] Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK biobank: Opportunities for cardiovascular research. European Heart Journal. 2017;**44**:1158-1166

[49] Scholtens S, Smidt N, Swertz MA, Bakker SJL, Dotinga A, Vonk JM, et al. Cohort profile: LifeLines, a three-generation cohort study and biobank. International Journal of Epidemiology. 2015;**44**(4):1172-1180

[50] The Health 202: NIH wants 1 million Americans to contribute to new pool of gene data. The Washington Post [Internet]. 2019. Available from: https://www.washingtonpost.com/news/powerpost/paloma/the-health-202/2018/01/16/the-health-202-nih-wants-1-million-americans-to-contribute-to-new-pool-of-gene-data/5a5ba45a30fb0469e8840135/

[51] FACT SHEET: President Obama's Precision Medicine Initiative. whitehouse.gov [Internet]. 2019. Available from: https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

[52] Precision Medicine Initiative (PMI) Working Group. The precision medicine initiative cohort program—Building a research foundation for 21st century medicine. Precision Medicine Initiative Work Group Report to Advisory Committee to Director NIH; 2015

[53] FinnGen, a global research project focusing on genome data of 500,000 Finns, launched. EurekAlert! Science News [Internet]. 2019. Available from: https://www.eurekalert.org/pub_releases/2017-12/uoh-fag121917.php

[54] Le QV. Building high-level features using large scale unsupervised learning. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings. 2013

[55] Greenspan H, Van Ginneken B, Summers RM. Guest editorial deep

learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging. 2016;**35**(5):1153-1159

[56] Giger ML. Machine learning in medical imaging. Journal of the American College of Radiology. 2018;**15**(3):512-520

[57] Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research—Commentary. BioMedical Engineering Online. 2014. Online: Published 5 July 2014. Article number: 94

[58] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Zeitschrift fur Medizinische Physik. 2019;**29**(2):102-127

[59] Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. Medical Image Analysis. 2017;**35**(1):18-31

[60] Islam J, Zhang Y. Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. Brain Informatics. 2018;**5**(2)

[61] Liu M, Zhang J, Adeli E, Shen D. Landmark-based deep multi-instance learning for brain disease diagnosis. Medical Image Analysis. 2018;**43**(1):157-168

[62] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics. 2009;**42**(5):760-772

[63] Kuleshov V, Ding J, Vo C, Hancock B, Ratner A, Li Y, et al. A machine-compiled database of genome-wide association studies. Nature Communications. 2019;**10**(1):3341

[64] Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophrenia. 2015;**1**(1):15030

[65] Chang EK, Christine YY, Clarke R, Hackbarth A, Sanders T, Esrailian E, et al. Defining a patient population with cirrhosis: An automated algorithm with natural language processing. Journal of Clinical Gastroenterology. 2016;**50**(10):889-894

[66] Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. Journal of the American Medical Informatics Association. 2016;**23**(6):1077-1084

[67] Collins FS, Lander ES, Rogers J, Waterson RH. Finishing the euchromatic sequence of the human genome. Nature. 2004;**431**(7011):931-945

[68] Willyard C. Expanded human gene tally reignites debate. Nature. 2018;**558**. Online: Published 19 June 2018

[69] Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19000 human protein-coding genes. Human Molecular Genetics. 2014;**23**(22):5866-5878

[70] Galvan A, Ioannidis JPA, Dragani TA. Beyond genome-wide association studies: Genetic heterogeneity and individual predisposition to cancer. Trends in Genetics. 2010;**26**(3):132-141

[71] Insel TR. Brain somatic mutations: The dark matter of psychiatric genetics. Molecular Psychiatry. 2014;**19**(2):156-158

[72] Diederichs S, Bartsch L, Berkmann JC, Fröse K, Heitmann J, Hoppe C, et al. The dark matter of the cancer genome: Aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. EMBO Molecular Medicine. 2016;**8**(5):442-457

[73] Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews. Genetics. 2010;**11**(6):446-450

[74] Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America. 2012;**109**(4):1193-1198. DOI: 10.1073/pnas.1119675109

[75] Gibson G, Dworkin I. Uncovering cryptic genetic variation. Nature Reviews. Genetics. 2004;**5**(9):681-690

[76] Kiser DP, Rivero O, Lesch KP. Annual research review: The (epi) genetics of neurodevelopmental disorders in the era of whole-genome sequencing—Unveiling the dark matter. The Journal of Child Psychology and Psychiatry and Allied Disciplines. 2015;**56**(3):278-295

[77] Fanciulli M, Petretto E, Aitman TJ. Gene copy number variation and common human disease. Clinical Genetics. 2010;**77**(3):201-213

[78] Park RW, Kim TM, Kasif S, Park PJ. Identification of rare germline copy number variations over-represented in five human cancer types. Molecular Cancer. 2015;**14**(25):1

[79] Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nature Genetics. 2013;**45**(10):1134-1140

[80] Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. Trends in Genetics. 2013;**29**(10):575-584

[81] Kuusisto KM, Akinrinade O, Vihinen M, Kankuri-Tammilehto M, Laasanen SL, Schleutker J. Copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. PLoS One. 2013;**8**(8):e71802

[82] Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, et al. Germline mutations in predisposition genes in pediatric cancer. The New England Journal of Medicine. 2015;**373**(24):2336-2346

[83] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;**444**(7118):444-454

[84] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;**526**(7571):68-74

[85] Bredfeldt TG, Walker CL. Epigenetics. In: Comprehensive Toxicology. 2nd ed. 2010

[86] Heard E, Martienssen RA. Transgenerational epigenetic inheritance: Myths and mechanisms. Cell. 2014;**157**(1):95-109

[87] Toh C, Brody JP. Chromosomal scale length variation of germline DNA can predict individual cancer risk. bioRxiv. 2018;**10**(1101):303339. DOI: 10.1101/303339

[88] Lello L, Raben T, Yong SY, Tellier LC, Hsu SDH. Genomic prediction of complex disease risk. bioRxiv. 2018;**10**(1101):506600. DOI: 10.1101/506600