14
Artificial Intelligence

Buyun Zhang **and** ***James P. Brody****

QUOTABLE QUOTES

"Nobody phrases it this way, but I think that artificial intelligence is almost a humanities discipline. It's really an attempt to understand human intelligence and human cognition."

Sebastian Thrun, German Computer Scientist, and CEO of Kitty Hawk Corporation (Marr 2017)

"Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, we will have multiplied the intelligence, the human biological machine intelligence of our civilization a billion-fold."

Ray Kurzweil, American Computer Scientist, and Inventor (Marr 2017)

"Some people worry that artificial intelligence will make us feel inferior, but then, anybody in his right mind should have an inferiority complex every time he looks at a flower."

Alan Kay, American Computer Scientist, and Fellow of American Academy of Arts and Sciences (Marr 2017)

"Anything that could give rise to smarter-than-human intelligence—in the form of artificial intelligence, brain-computer interfaces, or neuroscience-based human intelligence enhancement—wins hands down beyond contest as doing the most to change the world. Nothing else is even in the same league."

Eliezer Yudkowsky, American Researcher of Artificial Intelligence (Marr 2017)

"A year spent in artificial intelligence is enough to make one believe in God."

Alan Perlis, American Computer Scientist and Professor (Marr 2017)

Patient: "I'm having trouble breathing. I want my nurse."

Hospital Administrator: "We have something better than nurses, algorithms!"

Patient: "That sounds like a disease"

Hospital Administrator: "Algorithms are simple mathematical formulas that nobody understands. They tell us what disease you should have, based on what other patients have had."

Patient: "Okay that makes no sense. I'm not other patients, I'm me."

University of California, Irvine, Department of Biomedical Engineering, 3120 Natural Sciences II, Zot 2715, Irvine, CA 92697-2715.

* Corresponding author: jpbrody@uci.edu

Hospital Administrator: “Look, it’s not all about you. We’ve spent millions on algorithms, software, computers . . .”

Patient: “I need my nurse!”

National Nurses United Radio Ad. (National Nurses United, 2014)

“When you’re fundraising, it’s artificial intelligence. When you’re hiring, it’s machine learning. When you’re implementing, it’s linear regression. When you’re debugging, it’s printf()”

Baron Schwartz, Chief Technical Officer at VividCortex (B. Schwartz, 2017)

Learning Objectives

After completing this chapter, the reader should be able to:

- Understand the general definition of artificial intelligence (AI).
- Understand the relationship between Big Data and AI.
- Describe how classification algorithms are related to diagnosis of medical conditions.
- Read and understand a receiver operator characteristic curve.
- Understand the benefits and limits to characterizing a receiver operating characteristic (ROC) curve by its area under the curve (AUC).
- Understand the distinction between supervised and unsupervised learning.
- Give different examples of machine learning algorithms.
- Understand how ethical and regulatory issues effect the application of machine learning to medicine.

Introduction

Machine Learning vs. Artificial Intelligence

Artificial intelligence is a broad term that includes many different technologies and has a long history in medicine (Schwartz et al. 1987). In its broadest sense, artificial intelligence is defined as a machine that has thinking, reasoning, problem solving, and learning capabilities similar to humans.

Examples of artificial intelligence include knowledge bases and expert systems. Table 1 gives an example of an expert system for diagnosing congenital heart disease from 1968 (Gorry and Barnett 1968).

Table 1 demonstrates the early importance placed on natural language processing, the ability to extract information from written language. Natural language processing is still an area of active research with some success. However, today’s most successful artificial intelligence applications in medicine use machine learning.

The recent popularity of machine learning can be attributed to two major innovations, first the availability of extremely large digitized datasets and second, the development of computational techniques to control and reduce overfitting known as *regularization*.

These large digital datasets are often called Big Data. Big Data has led to a number of innovations. Large collections of text that have been translated into two or more languages led to machine translations that are almost as good as humans (Monroe 2017). Thousands of hours of television programs that include closed captions for the hearing impaired led to better voice recognition systems. Enormous sets of tagged images have vastly improved automated image recognition (Halevy et al. 2009).

Artificial intelligence advancements in a field follow the existence of Big Data in that field. A big impetus for the adoption of electronic health records in the United States was the passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act, part of the American Recovery and Reinvestment Act of 2009. We are beginning to see the applications of machine learning to medicine now.

Table 1. An artificial intelligence package for diagnosis, 1968.

```
PLEASE TELL ME ABOUT THE PATIENT
The patient here with me is 23 years old
PLEASE CONTINUE
He shows no evidence of cyanosis
OK PLEASE GO ON
There is no indication of the squatting symptom
ALL RIGHT.  GO ON PLEASE.
I'm afraid that's all I know about him at present.
CHECK THE HILAR VESSELS.  ARE THEY INCREASED OR DECREASED
They are increased
I SEE
LOOK FOR A LEFT SECOND SYSTOLIC MURMUR
First tell me what the three most probable diseases are at this time
2 WITH PROBABILITY 74
18 WITH PROBABILITY 7
30 WITH PROBABILITY 3
PLEASE CONTINUE
LOOK FOR A LEFT SECOND SYSTOLIC MURMUR
And the name of disease 18 is
PATENT DUCTUS ARTERIOSUS
ALL RIGHT. GO ON PLEASE
LOOK FOR A LEFT SECOND SYSTOLIC MURMUR
He has a left second midsystolic murmur
ALL RIGHT
IS THE AXIS IN THE ECG RIGHT, LEFT, OR NORMAL
It is the right axis
OK
I RECOMMEND A DIAGNOSIS OF DISEASE 2
What is your current view of the problem
MY CURRENT VIEW IS
     DISEASE    PROBABILITY
        2             93
        4              2
Name disease 4 for me
ATRIAL SEPTIC DEFECT WITH PULMONARY HYPERTENSION
I'm finished with this case. Thank you for your help
I HOPE THAT I HAVE BEEN OF SERVICE.  GOODBYE
```

Diagnosis, Prognosis in Medicine vs. Classification in Machine Learning

Diagnosis is a key part of medicine. A patient presents with several symptoms and laboratory tests. Based on these symptoms and tests, the physician is called on to classify the patient's condition into a disease. Once diagnosis is complete, treatment can begin.

Classification is a key part of machine learning. A dataset is characterized by a number of variables. Based upon these variables, the machine learning algorithm is called on to classify the dataset into a particular class. An example is optical character recognition. An image containing a single numeral can be digitized into a 20 × 20 array of pixels. The algorithm can classify these 400 pixel values into one of ten possible digits, 0–9.

Thus diagnosis in medicine is a natural target for the application of machine learning. Both medical diagnoses and machine learning classification share a set of terminology that describes the performance of diagnosis and classification tests.

Diagnosis or classification tests are judged by their ability to correctly predict both the correct answer and avoid the incorrect answer. For the simplest case, a binary classification, four rates are relevant: the true positive, false positive, true negative and false negative rates. The true positive and true negative rates are the correct answers, while the false negative and false positive errors are the incorrect answers.

Sensitivity versus Specificity

These four numbers can be combined into two: sensitivity and specificity. The sensitivity is defined as the probability of a positive test given that the sample is known to be positive. While the specificity is the probability of a negative test given that the sample is known to be negative.

Sensitivity and specificity depend on cutoff values. Different cutoff values give different sensitivity and specificity values. Ideally, a test will have both high sensitivity and high specificity, but a tradeoff exists between the two. Often a test will produce a numerical value. All values above a threshold have a positive test result and all values below are negative. The exact values of the sensitivity and specificity will depend on the threshold value. If one chooses a low threshold value, one gets a high true positive rate (high sensitivity), and a high false positive rate (low specificity). On the other hand, if one chooses a high threshold value, one gets a low true positive rate (low sensitivity) and a low false positive rate (high specificity).

Characterizing such a test is a very general problem first addressed in the field of signal detection theory (Peterson et al. 1954). The problem was posed this way: "Suppose an observer is given a voltage varying with time during a prescribed observation interval and is asked to decide whether the source is noise or is signal plus noise. What method should the observer use to make this decision and what receiver is a realization of that method?"

Petersen, Birdsall and Fox answered the question they posed. The best method to decide whether you have signal, or just noise is to set a threshold. If the voltage exceeds the threshold, then one can claim to have detected the signal. Of course, this alters the question to "how does one set a threshold". One sets the threshold based on the acceptable true positive rate and false positive rate. In the electrical engineers' formulation of the problem, the "test" was an electronic receiver that detected the voltage. Thus, the main task was to characterize their receiver's operating condition. They formulated a graphical expression of their receiver's performance that they named the receiver operating characteristic (ROC) curve. In our case, a better name might be the test characteristic curve, but the ROC nomenclature is firmly embedded in science.

ROC Curves and Area Under Curve to Quantify Quality of Test

The receiver operating characteristic curve expresses the full capabilities of a test. One often needs to answer, "How good is the test?" The naïve would expect an answer like "80% accurate." Where the naïve might define accuracy as "probability of a positive test given that the sample is known to be positive," which we have defined as sensitivity. Instead, we can first define two extreme answers to the question, "how good is the test?" We can have the answer, "it's a perfect test", and "it's a completely random test".

A perfect test is one with 100% specificity and 100% sensitivity. It would have an ROC curve that looks like Fig. 1.

If the predictions are made at random, the specificity will be equal to the sensitivity for all thresholds, it looks close to Fig. 2.

The full range of possible threshold values and the associated true positive rate and false positive rate can only be expressed by a receiver operating characteristic curve (ROC). However, a summary of the test's receiver operating characteristic curve can be computed by taking the integral under the receiver operating characteristic curve. This quantity is widely known as the area under the receiver operating characteristic curve or area under the curve or AUC.

A perfect test has an AUC = 1.0. A completely random test has an AUC = 0.5. One can now answer the question, "how good is the test?" with a number. The test has an AUC = 0.7. A useful shorthand is to think of the AUC as a grade one might get in a class.

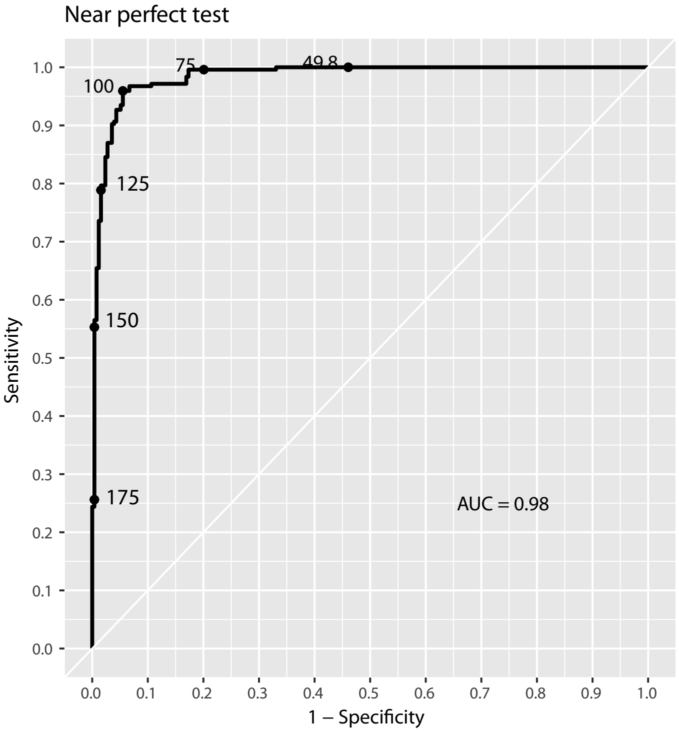

Fig. 1. A near perfect test has an AUC close to 1.0. In this figure, the threshold values are indicated on the curve.

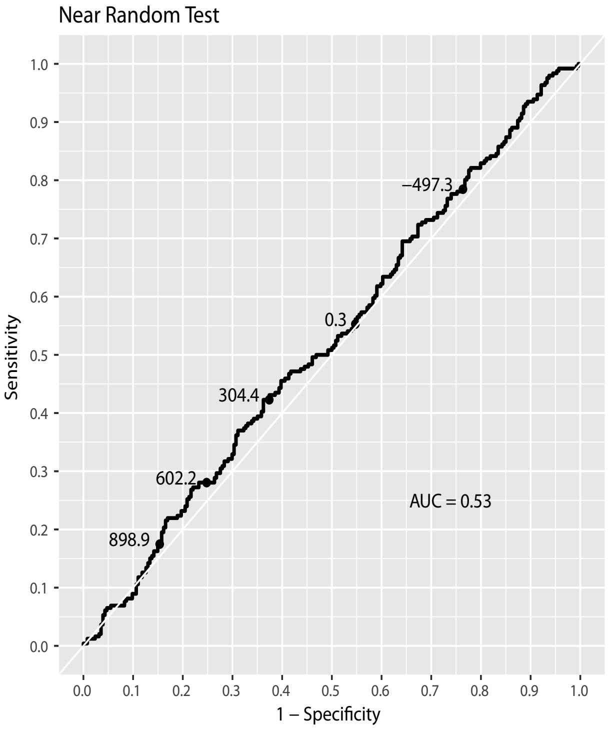

Fig. 2. This ROC curve represents a poor test, one that is just a bit better than guessing at random.

Table 2. AUC's can be interpreted as grades.

AUC	Grade
> 0.9	A
0.80–0.89	B
0.70–0.79	C
0.60–0.69	D
0.50–0.59	F

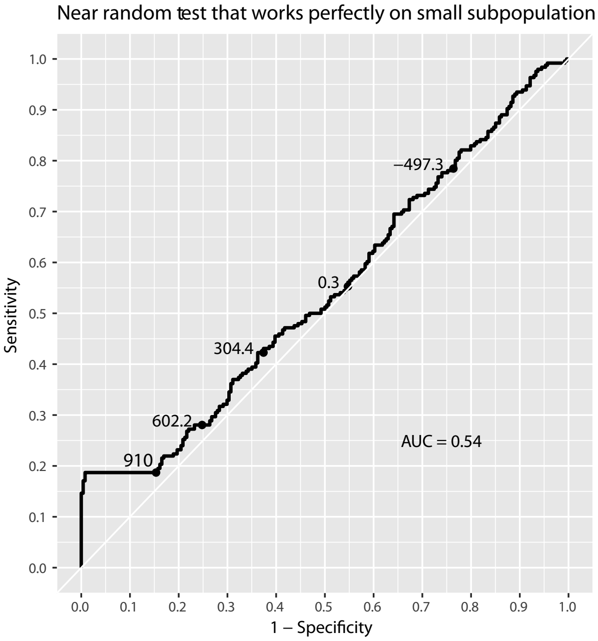

Fig. 3. This test has almost the same AUC as the one in Fig. 2, but it performs very well for a small subpopulation. This subpopulation performance is shown by the steep slope of the curve in the lower left. This figure demonstrates the importance of examining the entire AUC curve, and not just the summary statistic, AUC.

The AUC is only a summary of the test, and a test with a low AUC can be useful in a small subset of cases. Figure 3 shows an example from a BRCA-like test, where the test predicts outcome well for a small fraction of the population.

The AUC is widely used to quantify tests in machine learning, medicine, psychology and many other fields. In some fields, the AUC has other names including c-statistic, concordance statistic, and c-index.

One appealing feature of the ROC/AUC is that it is insensitive to class imbalance. Suppose a test set contains 90% of normal patients and 10% diseased patients. The machine learning task is to classify whether a particular patient is normal or diseased. An algorithm that simply guesses "normal" for all unknown patients will have an accuracy of 90%. The AUC, however, will be 0.50. The AUC is a better measure of the algorithms performance than the accuracy.

In cases of extreme class imbalance, screening for a rare cancer for instance, one often wants to identify the small number of patients most likely to be diagnosed with the rare cancer. In these cases, it is better practice to use a lift chart, which identifies what percentage of the target population (those with the rare cancer) can be identified with the smallest possible group. As an example, suppose the rare cancer occurs at a rate of 1 in 100,000. If we had an algorithm that could narrow identify a subset of the population in which the rare cancer occurs at a higher rate, say 1 in 10,000, that algorithm would have a lift of 10. The lift is computed as the ratio of the rate after prediction to the rate before prediction. An algorithm that provides no information (random) has a lift of 1.

Table 3. Several common medical tests and their published AUC values.

Test	AUC	Reference
PSA to detect prostate cancer	0.68	(Thompson et al. 2005)
Cardiac troponin to diagnose acute myocardial infarction	0.96	(Reichlin et al. 2009)
Cell free DNA test for Down's syndrome	0.999	(Norton et al. 2015)
HbA1c for diagnosing diabetes	0.958	(Tankova et al. 2012)
HEART score to predict major adverse cardiac events in 6 weeks from patients presenting with chest pain	0.86	(Poldervaart et al. 2017)
Circulating tumor cells to diagnose lung cancer	0.6	(Tanaka et al. 2009)

Artificial Intelligence Algorithms

Overfitting

In many cases, the primary objective of a machine learning algorithm is to predict a value or diagnose a condition based upon an input. Biomedical examples abound:

- Can one diagnose whether a patient has lung cancer based upon an immunofluorescent studies of cells captured from the blood?
- Can one diagnose whether a patient has diabetes by measuring glycated hemoglobin levels in the blood?
- Can one predict whether a patient will have a heart attack in the next month based upon a combination of EKG, age, body mass index, tobacco smoking status, family history, and measurements of troponin in the blood?

In the simplest case, one has an example set of results *y* and input values *x*. The goal is to identify the best function that will predict the value of *y* from the input values, *x*: $y = f(x)$ as shown in Fig. 4. This

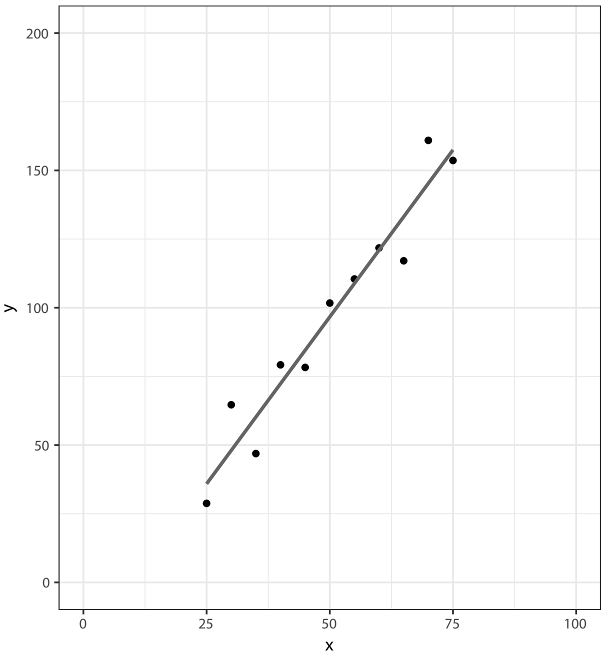

Fig. 4. The line represents the best prediction of y, given x. The points indicate observations of y, and the corresponding x value. The line is represented by two parameters a slope and an intercept.

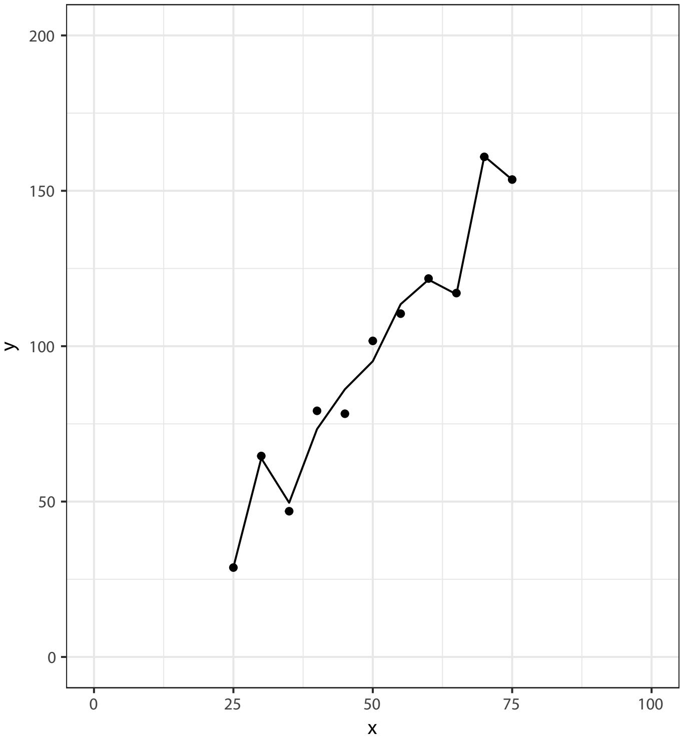

Fig. 5. This black curve represents a clear case of overfitting. The same data as shown in Fig. 4 are presented. In this case, the datapoints are fit with an 8th degree polynomial.

process is referred to as supervised learning. The challenge is to derive this function *f* without overfitting. An example of overfitting is shown in Fig. 5.

Every real set of data contains both signal and noise. The goal of machine learning is to build a model of the signal, while ignoring the noise. The problem is that the noise is often indistinguishable from the signal.

Many approaches to reducing overfitting exist. The first, and simplest, is to have some fundamental understanding of the relationship between *y* and *x*. For instance, if there is good reason to believe that *y* is linearly related to *x*, then the set of functions, *f*, should be limited to those in which *x* and *y* are linearly related: $y = \beta x + c$. If the relationship between *y* and *x* is complex and not well understood, then methods that are more complex are needed.

The main approaches to reduce overfitting in complex functions are known as regularization and drop out. To understand regularization better, we first need to state the cost function. A typical linear least square fit minimizes the cost function

$$C = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta x_i)^2,$$

This cost function is called the mean squared error. It takes the squared difference between the actual value, y_i and the predicted value β x. The problem can be posed as a minimization problem, where the goal is to minimize the cost function by adjusting β.

If multiple input variables, x_j, exist, instead of a single input variable, x, then a set of coefficients, β j is also needed. The cost function now looks like this, when there are a total of p input variables and n independent observations:

$$C = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2,$$

Minimizing this function will often lead to overfitting.

Overfitting can be reduced by adding a new term to the cost function that penalizes functions that are more complex. This addition to the cost function is known as regularization. The idea is that a simpler model is one that uses fewer of the x_j variables. The penalty is implemented by adding a term to the cost function that is proportional to the absolute value of the coefficient:

$$C = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

This addition to the cost equation is known as L1 or Lasso Regression. If, instead, the cost function includes a factor proportional to the square of the coefficient:

$$C = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2,$$

The process is known as L2 regularization.

In each case, the cost function is increased by the final term, which is proportional to the parameter λ.

Regularization Revolution

Dropout

Dropout is a revolutionary method to prevent overfitting (Srivastava et al. 2014). Dropout has primarily been applied to deep learning algorithms, but similar techniques are also used in other algorithms. The principle underlying the dropout mechanism is that by randomly dropping different connections within a network during the training process, the network becomes more robust and less susceptible to overfitting. The advent of dropout led to an immediate improvement in the performance of most deep learning algorithms.

Machine learning algorithms fall into two categories: supervised and unsupervised. Supervised algorithms require a batch of training data: a set of chest x-ray images from patients with pneumonia, for instance. The supervised algorithm uses the training data to build a model that can be applied to similar data (a chest x-ray) with an unknown diagnosis. An unsupervised algorithm is applied to a set of data to discover sub classifications. For instance, an analysis based on six variables from patients with adult onset diabetes found that these patients could be grouped into five different types. The different types had different risk of complications (Ahlqvist et al. 2018).

Supervised

The goal of supervised algorithms is to predict either a classification or a numerical result, called regression. Classification is ubiquitous in medicine. The patient presents with symptoms, the physician attempts to classify the patient, based upon symptoms into one of several possible diagnoses. Machine learning algorithms usually provide a measure of probability that a set of data belongs to one class or another. Regression is less common, but still useful. Regression can answer questions like, how many days is the patient expected to be in the hospital based on a set of data like age, initial diagnosis, vital signs, height, weight, days since last hospital stay, etc.

Most supervised algorithms are based upon one of two types of algorithms: decision trees and neural networks.

Decision Trees

A basic decision tree is shown in the figure. A patient has seven binary variables: A, B, C, D, E, F and G. Each variable has one of two values, indicated by the corresponding lower or upper case letter. In this particular decision tree, the results are independent of C, D, and E. The patient can be classified as either “cancer” or “normal” based upon the values of A, B, F, and G.

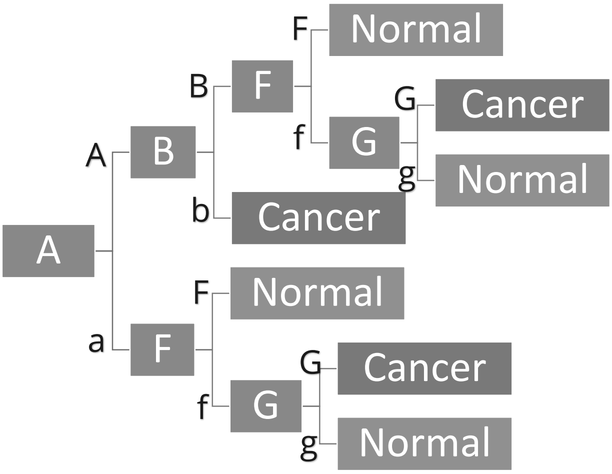

Fig. 6. This decision tree classifies a dataset as either normal or cancer.

This basic algorithm can be extended to cases that are more general. Instead of a definitive diagnosis for "cancer" or "normal" a probability is given. Continuous variables rather than binary variables can be used. The algorithm would find the optimal point at which the continuous variable be split into less than or greater than. Multiple decision trees can be built and an ensemble probability is computed by averaging the result of the many different decision trees. One can vary the number of different trees and the maximum depth of each tree. Specific popular algorithms based upon decision trees are XGBoost (Chen and Guestrin 2016), GBM (Friedman 2001, 2002), and LightGBM (Ke et al. 2017).

Neural Networks

Artificial neural networks are based upon biological neural networks found in the brain. An input layer, one or more hidden layers, and an output layer characterize an artificial neural network. Each layer is composed of multiple artificial neurons, the basic building block of the artificial neural network. An artificial neuron has multiple inputs and a single output. The artificial neuron takes the numbers provided on its input and applies some function to compute an output number. The output number of each neuron in a layer is then fed to neurons in the next layer. This process continues until the output layer.

An artificial neural network might diagnose pneumonia from chest x-rays. First, a network topography is identified. The input layer needs one neuron for each input variable. The output layer needs one neuron for each possible output. The geometry of the hidden layer varies, and identifying the best hidden layer geometry is one of the major challenges of building an effective neural network.

Next, all x-rays images are converted to a uniform image shape, say 1024 × 768. Each pixel of those images, 786,432 in total, corresponds to one neuron in the input layer. (Thus, this neural network would have an input layer with 786,432 input neurons, one or more hidden layers each with multiple neurons, and two neurons in the output layer.) The two neurons in the output layer correspond to the two states: "patient with pneumonia" and "patient without pneumonia".

The neural network is then "trained". A batch of chest x-rays previously identified as either from patients with pneumonia or from patients without pneumonia are identified. Training consists of applying the values of each pixel in the x-ray to the corresponding input neuron. Then the properties of each neuron in the hidden layer (known as weights) are adjusted to produce the appropriate output, a value of 1 in the appropriate output neuron and a value of 0 in the other neuron.

Training a large neural network can take significant computer time. If one had 10,000 chest x-rays from each class, each with 786,432 pixels, training could easily take weeks on a typical 2018 computer. Various methods are used to speed up training. The most common is to employ a graphical processing unit (GPU) in the computer. Another is specialized hardware: Google developed a Tensor Processing Unit to speed up training neural networks (Jouppi et al. 2017). Sometimes, the data is down sampled to reduce the training time.

Hyperparameters

Training a neural network requires not only determining the parameters (weights of the neurons), but also the hyperparameters. The hyperparameters are the properties of the network that are determined before the determination of the weights of the individual neurons. Hyperparameters include the structure of the network (how many hidden layers, how many neurons in each layer) and information on how the weights are to be determined during the training process (factors known as the learning rate, batch size, and momentum are examples).

Unsupervised

k means Clustering

The most common unsupervised machine learning algorithm is known as *k*-means clustering. The algorithm is rather old; it dates from at least the 1980's (Lloyd, 1982). The goal of this algorithm is to identify subgroups in the dataset. The number of subgroups must be pre-specified, and is known as *k*. With a complex dataset, the algorithm will find a close to ideal partition of the data into *k* different clusters. Usually, subgroups have some important characteristic that makes them useful with a common diagnosis into subgroups. A typical process is to divide a group of patients into sub-groups, where each sub-group has a different survival time for that disease.

One of the critiques of this process is that it will identify clusters even when none exists. Given a dataset with 1000 samples from a single subject, where the only difference between samples is noise from the measurement method, *k*-means will happily identify *k* different clusters in the data.

Hierarchical Clustering

Hierarchical clustering follows a different approach. This algorithm does not require that one specify several clusters beforehand. Instead, it computes a similarity score between different samples of the dataset. Pairs of samples that are most similar are then organized adjacent to one another in a dendrogram, see Fig. 7.

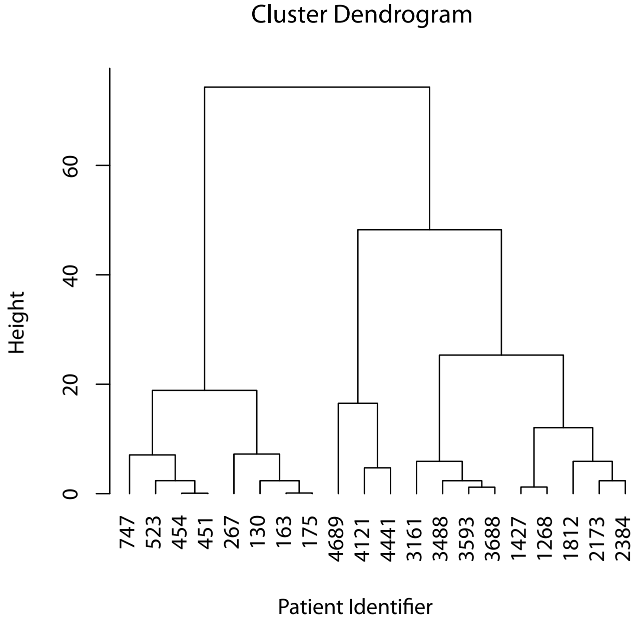

Fig. 7. This dendogram presents a hierarchical clustering of 20 patients. Each patient has a number of clinical variables associated with them. The patients are arranged such that the height of the line connecting them represents how difference in the patient's clinical variables.

Terminology: Machine Learning and Clinical

Supervised learning with hyperparameters requires that one be very careful to not over fit the data. Ideally, one has access to an unlimited dataset, or more data than one can easily handle. Often, however, only a limited set of data is available.

Given a limited dataset, best practices are to split the data into three parts known as the training data, the validation data, and the test data. A typical split is 70% training, 15% validation, and 15% test data.

The training data is used to determine the parameters of the initial model. The initial model is evaluated by applying it to the validation data. The results of the model application to validation data can be used to determine when the training data is being over fit and also to determine the hyperparameters of the model. The test data is only used to perform a final evaluation of the model.

One common pitfall is modifying the model after looking at how it performs on the test data. Any such modification is a form of fitting to the test data. Improvements gained through such modifications inevitably will not translate to the next batch of fresh data. A second pitfall is when the data is preprocessed/sub selected before the test data is extracted. This preprocessing can lead to information derived from the test data leaking into the training process.

The terminology for these sub-datasets: train, validate, test is commonly used by the machine learning field, but it is not universal. Clinical laboratory tests are developed on one dataset, and then validated on an independent dataset. Thus, clinical papers often refer to the final independent dataset as the validation step.

Applications

Applications of machine learning to medical data have grown rapidly in the past few years. We can organize these applications by different types of medical data. Examples of different data structures found in medical data include tabular data, medical images, time series, and natural language.

Tabular data refers to the unstructured data collected for each patient: sex, visit date, diagnosis, height weight. Although medical test results could be considered unstructured and included in tabular data, most medical tests might also vary over time and could also be considered time series: systolic and diastolic blood pressure, temperature, and many different blood tests (electrolytes, glucose, cholesterols) can vary on timescales as short as minutes or hours.

More traditional time series medical data include pulse rate, respiration, and electrocardiogram. The characteristic of time series data is that each data point includes both a time and a value. Different data points in the dataset are related to one another by the time. In contrast, tabular data refers to data in which the order is irrelevant and each data point is unrelated to another. One could switch the order of the data in tabular data with no effect, but not with a time series dataset.

Medical image data can be found in x-rays, ultrasound, magnetic resonance imaging, and computed tomography. Even plain old photography provides valid medical images and is often used by pathologists to record tissue samples.

Natural language is an important source of data, primarily from unstructured physician notes. These notes are still the most common and complete form of documentation in many electronic health records.

Supervised Learning

Deep Learning Image Analysis: Classification of Skin Cancer with a Neural Network

Melanoma, or skin cancer, starts as a discolored patch of skin. At this stage, it is easily removed and once completely removed will not recur at that location. Occasionally cells from this discolored patch of skin can travel, or metastasize, to other locations in the body. Common landing sites are the brain, liver, or lungs. The five-year survival rate of metastatic melanoma is less than 10%, while the survival 5 year survival rate for localized melanoma is over 90% (Jemal et al. 2017).

Diagnosing and removing melanoma in its earliest stages should lead to a significant reduction of metastatic melanoma and deaths due to melanoma. Early stage melanoma superficially appears similar to a mole, which is a general term for a harmless discoloration on the skin. Thus, differentiating early stage melanoma from a harmless mole, which is ubiquitous, is a regular task that dermatologists perform.

A group led by Sebastian Thrun developed a neural network to classify skin lesions as either benign or malignant (Esteva et al. 2017).

They started with a large dataset. Their training/validation dataset consisted of 129,450 images of skin lesions. Each image had been previously classified by a trained dermatologist. Some of the skin lesions depicted in the images had biopsies performed on them, providing an absolute truth to the classification problem.

They used a pre-trained convolutional neural network to build their classifier. Pre-trained neural networks use transfer learning (Quattoni et al. 2008). The disadvantage is that the supplied images must match the geometry of the pre-trained network expects. In this case, they started with a network called the Inception-V3 (Szegedy et al. 2015) network, which uses 299 × 299 pixel images. They pre-trained this network with 1.28 million images that contained 1000 different object categories, these categories are common descriptors and unrelated to skin or melanoma. Pre-training like this can substantially improve performance on a new unrelated dataset.

The pre-trained network was then trained with specific dermatology data. They split the 129,450 images into a train/validate dataset containing 127,463 images and a test dataset containing 1,942 test images. Each of the test images had a biopsy-confirmed classification.

The model, trained on the 127,463 images, was tasked with classifying the 1,942 test images. A subset of the 1,942 test images were also graded by at least 20 experienced dermatologists. The results of the model could then be compared to the dermatologists, as shown in Fig. 8.

In the specific case shown in the figure, for diagnosing melanoma, the machine learning algorithm had an AUC of 0.94.

Recall that the ROC curve is determined by varying the threshold and plotting the resultant sensitivity/specificity that one would achieve with the threshold. Actual dermatologists do not have a specified threshold. Instead, they have an internal, hard to quantify, threshold. By measuring a particular dermatologist's true positive and false positive rate, one can put a single point on the ROC plot. Including several dozen dermatologists, results in an approximate curve on the ROC plot.

The conclusion of the paper is that the trained algorithm performed at level similar to trained dermatologists.

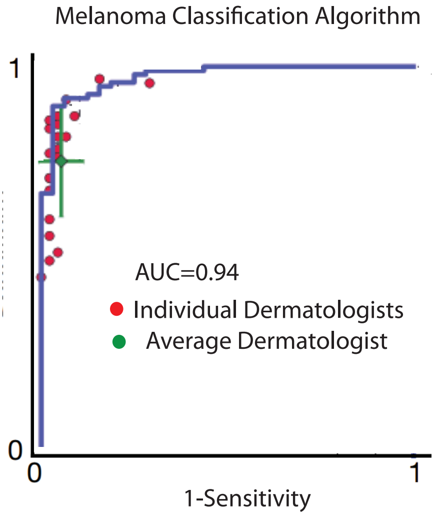

Fig. 8. The ROC curve for the melanoma classification algorithm, adapted from [21].

Deep Learning Image Analysis: Diagnosis of Diabetic Retinopathy from Fundus Photography

A leading cause of blindness in adults is diabetic retinopathy. Retinopathy occurs in untreated type 1 diabetes and can be reduced or delayed with tight control of glucose levels (Nathan 2014). Retinopathy is the degradation of small blood vessels in the retina. Using a special type of photography, called fundus photography, images of these blood vessels can be acquired, see Fig. 9. The degree of retinopathy can be quantified and tracked. At the earliest stages, retinopathy is easily diagnosed with fundus photography while it has no discernable effect on the patient's vision. At later stages, retinopathy can cause blurred vision, blank areas in the vision field and ultimately complete loss of vision.

Ophthalmologists typically grade the degree of retinopathy in fundus photographs. These ophthalmologists will have slightly different grades for identical photographs, known as inter-observer variability. The same grader shown the same fundus photograph at two different times will provide a slightly different grade, known as intra-observer variability (Abràmoff et al. 2010). Inter-observer and intra-observer variability hamper the measuring of small changes in retinopathy in patients. Thus, an ongoing goal is to have a consistent reliable method of grading fundus photography for the degree of retinopathy present.

This problem, the automated detection of micro vessel damage in fundus photographs, is one of ongoing interest to the biomedical imaging community. A challenge, known as the Retinopathy Online Challenge, was organized in 2009 to provide a uniform dataset and grading criteria (Niemeijer et al. 2010). The challenge provided 50 training images with the reference standard and another 50 test images where the reference standard was withheld. Competitors provided "answers" for the 50 test images and were graded based upon those answers. The best competitor algorithms did not reach the level of human experts.

In 2016, a team lead by Google employees developed a predictive model to analyze fundus photographs that does compare to the level of human experts (Gulshan et al. 2016). This work was noteworthy not only for its development of a deep learning algorithm, but also for its accumulation of a very large dataset of graded fundus photographs.

The Google group first built a dataset of 128,175 retinal photographs. Each was graded from 3 to 7 different experts. The experts consisted of a paid panel of 54 ophthalmologists who were licensed in the US or were senior residents. All grading occurred within an 8-month period in 2015.

The model building started with the Inception-v3 architecture neural network (Szegedy et al. 2015). Recall, this architecture requires that all input images are 299 × 299 pixels. This neural network was pre-trained with the ImageNet database, using transfer learning. The first step was to scale the fundus photographs such that the diameter of the dark circle was 299 pixels across. They split the data into 80% for training and 20% for validation (tuning hyperparameters). The training optimized the values of the 22 million parameters in the neural network

They had two separate independent test datasets consisting of just over 10,000 images. Seven or eight trained ophthalmologists graded each of these images for referable diabetic retinopathy. The grading resulted in a binary decision: the patient should be either "referred" or "not referred" for further evaluation. The trained neural network performed comparably to trained ophthalmologists.

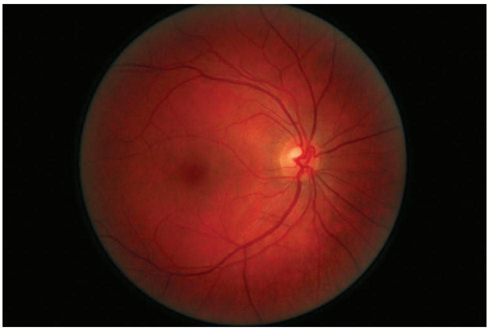

Fig. 9. A fundus photograph captures the microvasculature in the eye. Image by Danny Hope, CC by 2.0. (Image available at https://commons.wikimedia.org/wiki/File:Righ_eye_retina.jpg).

The results for a few other interesting experiments were also presented. One showed that the 128,175 initial retinal photographs in the training set might be too many. The results indicate that they could achieve similar results with half as many images. If they had used only 20,000 images, their results would be about 20% worse. A second experiment measured the effect of changing the number of experts who grade the training images. They found that increasing the average number of grades per image led to increases in the results right up to the limits of the data they had. This result implies that their neural network could be improved with more graders on the initial dataset, where the algorithm learns.

Deep Learning Image Analysis: Cardiovascular Risk from Retinal Images

Fundus photographs offer one of the easiest ways to image the microvasculature system. Many cardiovascular diseases stem from problems that occur with the blood vessels. Fundus photographs may offer a window into cardiovascular disease.

Several cardiovascular risk calculators exist. One of the first was the Framingham risk score (Lloyd-Jones et al. 2004, Pencina et al. 2009). The Framingham risk score provides a number that indicates the likelihood of the patient developing cardiovascular disease within the next ten years. The factors considered are age, sex, total blood cholesterol, history of cigarette smoking, HDL cholesterol levels in blood, and systolic blood pressure. Taken together, these factors can predict cardiovascular disease with an area under the receiver operator characteristic curve AUC of ROC of about 0.8 (Günaydın et al. 2016).

A group of researchers at Google built a cardiac risk predictor that uses only fundus photographs as input (Poplin et al. 2018). They reasoned that since much of cardiovascular disease manifests itself in microscopic blood vessels and fundus photographs offer the clearest, easily obtainable images of the microvasculature, fundus photographs might be able to predict cardiovascular disease.

They first acquired the training data, which they refer to as the development dataset. They used 48,101 patients from the UK Biobank project, which has retinal fundus images along with other health information. They combined the UK Biobank data with an additional 236,234 patients from EyePACS. The UK Biobank health information included sex, ethnicity, BMI, blood pressure, and tobacco smoking status, while the EyePACS patient information included age, sex, ethnicity, and HbA1c blood levels.

They also constructed a test dataset (which they refer to as the validation dataset, consistent with clinical terminology). The test dataset consisted of 12,026 patients from the UK Biobank dataset and 999 from the EyePACS data.

They trained a deep neural network model to predict clinical factors from the fundus images. The model is again based on the Inception-v3 architecture and was pre-trained with objects from the ImageNet dataset. In this case, the initial images were scaled to be 587 pixels square.

Separate models were trained to predict discrete factors (sex and smoking status) and continuous factors (age, BMI, blood pressure, and HbA1c blood test results).

Overall results were good. They found that they could predict age, blood pressure, and BMI from fundus photographs significantly better than just guessing. However, they could not predict glycosylated hemoglobin (HbA1c) blood levels better than guessing. Glycosylated hemoglobin measures long-term average glucose levels in the blood. High levels of glucose are an indication of diabetes, which is a major risk factor for cardiovascular disease.

Finally, they were able to predict major adverse cardiac events in the following five years with an AUC of 0.70. This AUC is comparable to state of the art risk prediction using clinical factors like age, sex, blood pressure, and BMI. However, it does not represent much of an advance in medicine, since those clinical factors are easily measured. While this application of machine learning is useful, it will not change the practice of medicine.

Image Analysis: Human Level Performance is Often the Goal, the Limitation is using Large Medical Image Sized Images

Images of interest to most consumers contain substantial information at large scales, but much less information on small scales. A face, for instance, is recognizable from its overall structure and shape. The information at smaller scales: wrinkles or blemishes for instance, do not contribute much to its recognizability. Thus many of these images can be downsized substantially without losing important information. Medical images, on the other hand, often contain the relevant features, a crack in a bone or a small lump in the breast, at the small scale.

When applying deep learning to medical images, the goal is often to meet the level of performance that the best-trained humans can provide. Several of the applications mentioned above meet that level, but do so with downsized images. Most medical images have resolution in the 1000s by 1000s pixel range, but this many pixels is often too large computationally for machine learning algorithms to handle in reasonable amounts of time. To make the problem tractable, images are down sampled to 299 × 299 pixels, or 587 × 587 pixels. One question is how does this down sampling effect the ability of the machine learning algorithm to diagnose as compared to human readers who diagnose using the full resolution of the original medical image.

Image Analysis: Breast Cancer Screening

Screening for breast cancer presents a good opportunity to test the effect of down sampling. First, breast cancer screening often involves multiple images. Each image is an x-ray taken at a different angle of the breast.

Currently, a radiologist will examine these different views and assign the patient into one of seven classes. These classes are defined by the American College of Radiology and known as the Breast Imaging Reporting and Data System (BI-RADS) (American College of Radiology. BI-RADS Committee 2013). BI-RADS has seven levels:

Table 4. The seven levels of BI RADS grading.

0	Incomplete –Need additional imaging evaluation
1	Negative
2	Benign
3	Probably Benign
4	Suspicious
5	Highly suggestive of malignancy
6	Known biopsy-proven malignancy

In a typical screening population about 10–15% of patients receive a score of 0, necessitating a second mammogram (Geras et al. 2017). Of these, less than 1% will ultimately lead to a diagnosis of cancer. Nevertheless, this second mammogram induces anxiety and significant medical costs (Tosteson et al. 2014). Reducing these incomplete mammograms is an important goal of breast screening research.

A group at New York University examined how well a deep learning network could assign mammography images to one of the BI-RADS categories (Geras et al. 2017). They specifically asked the question of how well the machine learning algorithm performed as a function of the image size.

They started by building a dataset of 201,698 breast screening exams from 129,208 different patients. All the patients were from the New York City metropolitan area. This dataset consisted of 886,437 different images. Each exam typically consisted of two images of each breast. One image provides the craniocaudally view, while the second provides the mediolateral oblique view. Each image was taken at a resolution of 2600 × 2000 pixels. The entire dataset consisted of about four terabytes. All the images had been previously assigned BI-RADS scores of 0, 1, or 2 from a radiologist.

They split the dataset into three parts: training, validation, and test datasets. Their dataset had about 22,000 exams with a BI-RADS score of 0, 75,000 exams with a BI-RADs score of 1 and 67,000 exams with a BI-RADS score of 2. Then they constructed a deep learning network to train a model on classifying the exams into one of the BI-RAD classes.

To evaluate the model, the treated the problem as three separate binary classification problems: BI-RADS 0 vs. BI-RADS 1 and BI-RADS 2, BI-RADS 1 vs. BI-RADS 0 and BI-RADS 2, and BI-RADS 2 vs. BI-RADS 0 and BI-RADS 1. They computed the area under the curve for the receiver operator characteristic curve (AUC) for each of the three problems, and then take the average of the three AUCs to arrive at a macro average AUC (macAUC). This value, macAUC, is the primary performance metric of their model.

The results show that downscaling the datasets images by a factor of 8 (325 × 250 images rather than the 2600 × 2000 images) reduces the macAUC from 0.73 to 0.68. This downscaling corresponds to a reduction of the total size of the images by a factor of 64. They measured the effect of decreasing the resolution at several intermediate values and found a monotonic increase in macAUC with the resolution of the images.

They also tested the effect of changing the size of the training data. As expected, the increasing the amount of training data leads to an increase in the macAUC. For instance, when using 100% of the training data, the macAUC was 0.73. If, however, one only used 10% of the training data, the macAUC dropped to 0.68.

One can compare the two effects to see that decreasing the resolution of the images had a larger effect on the model than decreasing the amount of training data. If one is constrained by the total size in bytes, it is better to reduce the number of training cases than to decrease the resolution of the individual images.

Unsupervised Clustering

A common observation in medicine is that patients with identical diagnoses can have vastly different outcomes. Perhaps some respond to a particular therapy, while others do not. Some survive for years, while others die within a short period. This observation suggests that subcategories exist for the given diagnosis.

Type 2 Diabetes

Diabetes is a good example of a diagnosis with heterogeneous outcomes. Diabetes is classified into two forms: type 1 and type 2. The outcome for patients with type 2 diabetes is particularly heterogeneous. Some patients develop severe forms of kidney disease, but have no vision problems. Other patients, who share the same type 2 diabetes diagnosis, develop vision problems, but have no kidney problems.

Identifying subgroups is a branch of machine learning known as unsupervised clustering. The most widely used algorithm for unsupervised clustering is known as k-means (Kanungo et al. 2002). The k means algorithm takes a list of patients, each of whom has several clinical variables related to the diagnosis, and groups them into k subgroups. The grouping is done in a way such that the patients in each group have similar clinical variables. This algorithm is always successful; it will find subgroups. The key question is whether the subgroups are useful: do the subgroups predict outcomes better than any other method?

A group from Sweden and Finland recently applied k-means clustering to type 2 diabetes (Ahlqvist et al. 2018). They suggested that a better classification of type 2 diabetes patients could identify individual patients with increased risk of specific complications and then allow customized treatments to prevent those complications.

The first step was constructing the dataset. The group used data on about 30,000 diabetes patients from five Scandinavian studies that had been running for several years. The largest study, called the ANDIS project ("All New Diabetics In Scania—ANDIS | Swedish National Data Service" n.d.), was responsible for over half the patients. The ANDIS project is a study based upon the National Diabetes Register in Sweden (Gudbjörnsdottir et al. 2003).

The second step was selecting the relevant clinical variables. Since they were using patient samples that had been previously collected, they were constrained to clinical variables that these studies had collected. They selected model variables that should be related to insulin production and demand. They ended up with body mass index, age of onset, calculated estimates of beta cell function and insulin resistance (based on c-peptide concentrations) and the presence or absence of GADA, an antibody indicative of autoimmune conditions.

The clusters were probably not just later stages of diabetes. They found similar clustering in both recently diagnosed patients and in patients who were diagnosed years before.

It is not clear if the clusters represent different forms of diabetes. Understanding how diabetes originates and progresses is of fundamental interest, and a better understanding of this process could lead to both treatments and preventative measures. This clustering does not necessarily point to different causes, but it does indicate different outcomes for each cluster.

Unsupervised Classification of Brain Tumors

Brain tumors are another example of a heterogeneous disease. Brain tumors are classified using histology. Major divisions of brain tumors like gliomas, astrocytomas, medulloblastomas, each have many well defined subtypes (Louis et al. 2016). However, diagnosis within subtypes is variable. For instance, different pathologists will classify gliomas differently. These different sub classifications result in a different course of treatment for the patients. This variability has an effect on both clinical trials, and on applying the results of clinical trials to other patients (van den Bent 2010). An accurate diagnosis of the specific type of brain cancer a patient has should result in better treatment of the patient.

One potential use for classification is then to accurately classify brain tumors. A group led by Stefan Pfister tackled this important problem (Capper et al. 2018). The group started with the WHO classification of central nervous system tumors (Louis et al. 2016). They collected a total of 2,801 samples, with at least eight cases for each group, and then analyzed each patient sample for a genome wide methylation profile (Capper et al. 2018). The methylation profile provides about 450,000 measurements of methylation at different locations across the genome.

Once they collected this data, they performed an unsupervised clustering. They ultimately ended up with 82 different central nervous system tumor classes. Each class was characterized by a distinct DNA methylation profile. Of these 82 classes of tumors, 29 mapped directly to a single entity in the WHO classification scheme (for instance, diffuse midline glioma), another 29 mapped to a subcategory of an entity in the WHO classification (for instance glioblastoma G34). The remaining 24 classes were more complicated: 19 of them could be associated with several WHO classes (a one to one mapping did not exist) and the remaining five were classes that simply had been defined by WHO.

Their unsupervised clustering used the random forest algorithm. This algorithm gave probability estimates that the tumor sample belongs to each class. They first used cross validation to establish the consistency of the algorithm. The cross validation established that the about 95% of the tumors were consistently classified. Most of the inconsistently classified tumors occurred within a small group of closely related classes that have no clinical difference.

Having established their classifier, they applied it to a new set of 1,155 central nervous system tumors. This test set consisted tumors extracted from both adult (71%) and children (29%) patients with 64 different histopathological classifications. This test set was enriched with rare cases and does not represent a typical population. First, about 4% of the samples had to be discarded because they could not obtain a methylation profile from the sample. Of the remaining samples, 88% of the samples matched to one of the classes they established with the training set. Of these, about three quarters agreed with the histopathology evaluation, but one quarter did not. These one quarter underwent a second histopathology examination. In 90% of these mismatches, the revised classification matched with the methylation profile.

The authors demonstrated that they could consistently classify methylation profiles from different brain tumors. This classification system is useful clinically. The authors set up a web site where a pathologist can upload the methylation profile of a tumor sample, and the web site will provide a report on the classification of the tumor ("MolecularNeuropathology" n.d.).

Natural Language: Reading Doctors Notes

Natural language processing extracts computer readable data from free flowing text. A perfectly working natural language processing system would find many uses in biomedicine. For instance, a project dubbed "Literome" attempts to extract molecular interactions that might be involved in complex disease by analyzing the scientific literature and extracting gene networks and genotype-phenotype associations (Poon et al. 2014).

With the conversion of health records to electronic form, significant medical information is present in computer readable free form text. Physician notes and pathology reports are two particular areas that often contain text. Physician notes can contain everything from the patient's complaints, physician's observations, and patient's family history to the comments from the patient's primary caregiver. Physician notes can be inaccurate, incomplete, or poorly worded. These conditions make the natural language processing of physician notes challenging.

Pathology reports are more structured and focused than physician notes. These properties should make them more amenable to natural language processing. However, even these present challenges. One study of 76,000 breast pathology reports illustrates the problem (Buckley et al. 2012). They found 124 different ways of saying "invasive ductal carcinoma", even more troubling they found over 4000 different ways of saying "invasive ductal carcinoma was not present". Even with these challenges, a computerized system had 99.1% sensitivity at a level of 96.5% specificity compared to expert human coders.

One study of 500 colonoscopy reports provides a good test for a natural language processing system. The study, led by Thomas F. Imperiale, tested how well a natural language processing system could extract the highest level of pathology, and then the location of the most advanced lesion and the largest size and number of adenomas removed. This study found that the natural language processing system could identify the highest level of pathology correctly 98% of the time, but only could report the number of adenomas removed with 84% accuracy (Imler et al. 2013).

This 2013 study used the most popular open source natural language processing system, known as cTAKES (Savova et al. 2010). This software package was originally developed at the Mayo Clinic. Its name derives from the phrase "clinical Text Analysis and Knowledge Extraction System". It is now part of the Apache Software Foundation, which generally means that it is under active development and one can expect significant improvements over time. Performing the same test today should give substantially better results than the 2013 test.

Regulatory Issues

As the use of artificial intelligence and machine learning in medicine comes into widespread use, regulation should soon follow. The regulatory agencies are still creating a framework for dealing with these new medical tools. The International Medical Device Regulation Forum (IMDRF) is working on general principles that should be applicable to regulatory agencies around the world. IMDRF has representatives from Australia, Brazil, Canada, China, European Union, Japan, and the United States. In 2017, the IMDRF issued a working paper, "Software as a Medical Device (SaMD): Clinical Evaluation" (International Medical Device Regulation Forum 2017).

The US Food and Drug Agency (FDA) issued draft guidance in late 2017 on their approach for what they call "Clinical and Patient Decision Support Software" (Federal Drug Agency 2017). This guidance clarified what types of machine learning would and would not be subject to regulatory approval. The general principle that they follow is that software that makes a recommendation, but allows a physician to review the recommendation and the basis for the recommendation is exempt from FDA regulation. The FDA commissioner gave an example of software that knows current clinical guidelines and drug labelling that suggests ordering liver function tests first, when a physician attempts to prescribe statins for a patient (Gottlieb 2017).

However, the FDA would continue to regulate software that analyzes medical images, data from in vitro diagnostic tests, or signals from medical instrumentation like EKGs to make treatment recommendations. They view these types of software as an integral part of the medical device, which they already regulate. The FDA's justification is that if the software provides inaccurate information, significant patient damage could occur

Ethical Issues

The use of machine learning in medicine will bring new ethical issues to the forefront. Several ethical issues are predictable, have long been known (Schwartz et al. 1987), and should be addressed early on. Others will be more subtle.

Biases can be carried over from training. A machine learning algorithm is only as good as its training data. Since the majority of people in the United States are white, most medical data in the United States is collected from white people. However, the majority of genetic diversity in the United States comes from African Americans (Campbell and Tishkoff 2008, Gibbs et al. 2015). The dilemma then is how one determines the ethnic distribution of the training set. Should one aim to obtain a training set that benefits all people fairly well, or one that benefits one subgroup over another?

An example that we have worked on illustrates the problem. We have a machine learning algorithm that predicts a future diagnosis of ovarian cancer based on the patient's germ line DNA (Toh and Brody 2018). This algorithm was developed and trained with data from the Cancer Genome Atlas (TCGA) dataset (Bell et al. 2011). To test how the racial makeup of the training dataset effects predictions, we trained the model using a subset of the TCGA data solely from white patients. Then we applied the model to three different test sets each containing members labelled as "white", "Asian", or "black/African American". We found the AUC for predicting a future diagnosis of ovarian cancer to be 0.93 for the white test group, 0.98 for the Asian test group, and 0.70 for the black/African American test group. This general trend, where the AUC for the black/African American group is substantially lower, was also true for the other cancers tested: a form of brain cancer (glioblastoma multiforme), breast cancer, and colon cancer.

A second ethical issue is whether the machine learning algorithm should be designed to do what is best for the individual patient or for the medical system (Char et al. 2018). In our current system, physicians make medical decisions. This physician has an ethical obligation to treat patients to the best of their abilities. Medical systems, on the other hand, have no such obligation. A natural tension exists between physicians and medical systems, where the physician often takes the role of protecting the patient's best interest and the medical system tries to accommodate the physician, but must also deal with economic realities.

These issues might come up in a clinical decision support system. Suppose a private designer develops and markets a software package to review medical histories and recommend the next step in care. Should the private designer consider the effect of recommendations on the medical system's profit and quality metrics? Neither of these factors effect the patient's health, but both would be key selling points to the decision makers in the medical system who will authorize the purchase of such a software package (Char et al. 2018).

Conclusion

Today's machine learning applications in medicine are much more advanced than those from 50 years ago. Machine learning has made substantial progress on medical image analysis and the identification of subclasses of disease. The widespread adoption of electronic health records should lead to more innovations using this source of data. Regulatory agencies are developing new frameworks to ensure patient safety.

References

Abràmoff, M.D., J.M. Reinhardt, S.R. Russell, J.C. Folk, V.B. Mahajan, M. Niemeijer et al. 2010. Automated early detection of diabetic retinopathy. Ophthalmology 117(6): 1147–1154. doi:10.1016/j.ophtha.2010.03.046.

Ahlqvist, E., P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman et al. 2018. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The lancet. Diabetes & Endocrinology 6(5): 361–369. Elsevier. doi:10.1016/S2213-8587(18)30051-2.

All New Diabetics In Scania - ANDIS | Swedish National Data Service. n.d. Retrieved May 26, 2018, from https://snd.gu.se/en/catalogue/study/EXT0057.

American College of Radiology. BI-RADS Committee. 2013. ACR BI-RADS atlas: breast imaging reporting and data system. American College of Radiology.

Bell, D., A. Berchuck, M. Birrer, J. Chien, D.W. Cramer, F. Dao et al. 2011. Integrated genomic analyses of ovarian carcinoma. Nature 474(7353): 609–615. Nature Publishing Group. doi:10.1038/nature10166.

Bent, M.J. van den. 2010. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: A clinician's perspective. Acta Neuropathol. 120(3): 297–304. doi:10.1007/s00401-010-0725-7.

Buckley, J.M., S.B. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A.K. Belli et al. 2012. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J. Pathol. Inform. 3(1): 23. Medknow Publications and Media Pvt. Ltd. doi:10.4103/2153-3539.97788.

Campbell, M.C. and S.A. Tishkoff. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu. Rev. Genomics Hum. Genet 9: 403–33. NIH Public Access. doi:10.1146/annurev.genom.9.081307.164258.

Capper, D., D.T.W. Jones, M. Sill, V. Hovestadt, D. Schrimpf, D. Sturm et al. 2018. DNA methylation-based classification of central nervous system tumours. Nature 555(7697): 469–474. Nature Publishing Group. doi:10.1038/nature26000.

Char, D.S., N.H. Shah and D. Magnus. 2018. Implementing Machine Learning in Health Care—Addressing Ethical Challenges. N. Engl. J. Med. 378(11): 981–983. Massachusetts Medical Society. doi:10.1056/NEJMp1714229.

Chen, T. and C. Guestrin. 2016. XGBoost. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16, 785–794. New York, New York, USA: ACM Press. doi:10.1145/2939672.2939785.

Esteva, A., B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639): 115–118. Nature Publishing Group. doi:10.1038/nature21056.

Federal Drug Agency. 2017. Clinical and Patient Decision Support Draft Guidance for Industry and Food.

Friedman, J.H. 2001. Greedy function aproximation: A gradient boosting machine. Ann. Stat. 29(5): 1189–1232. doi:10.1214/aos/1013203451.

Friedman, J.H. and H., J. 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38(4): 367–378. Elsevier Science Publishers B. V. doi:10.1016/S0167-9473(01)00065-2.

Geras, K.J., S. Wolfson, Y. Shen, S.G. Kim, L. Moy and K. Cho. 2017. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks.

Gibbs, R.A., E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee et al. 2015. A global reference for human genetic variation. Nature 526(7571): 68–74. Nature Publishing Group. doi:10.1038/nature15393.

Gorry, G.A. and G.O. Barnett. 1968. Sequential Diagnosis by Computer. JAMA J. Am. Med. Assoc. 205(12): 849. American Medical Association. doi:10.1001/jama.1968.03140380053012.

Gottlieb, S. 2017. On advancing new digital health policies to encourage innovation, bring efficiency and modernization to regulation. FDA Statement.

Gudbjörnsdottir, S., J. Cederholm, P.M. Nilsson, B. Eliasson and Steering Committee of the Swedish National Diabetes Register. 2003. The National Diabetes Register in Sweden: an implementation of the St. Vincent Declaration for Quality Improvement in Diabetes Care. 26(4): 1270–6. American Diabetes Association. doi:10.2337/DIACARE.26.4.1270.

Gulshan, V., L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy et al. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 316(22): 2402. American Medical Association. doi:10.1001/jama.2016.17216.

Günaydın, Z.Y., A. Karagöz, O. Bektaş, A. Kaya, T. Kırış, G. Erdoğan et al. 2016. Comparison of the Framingham risk and SCORE models in predicting the presence and severity of coronary artery disease considering SYNTAX score. Anatol. J. Cardiol. 16(6): 412–8. Turkish Society of Cardiology. doi:10.5152/AnatolJCardiol.2015.6317.

Halevy, A., P. Norvig and F. Pereira. 2009. The Unreasonable Effectiveness of Data. IEEE Intell. Syst. 24(2): 8–12. doi:10.1109/MIS.2009.36.

Imler, T.D., J. Morea, C. Kahi and T.F. Imperiale. 2013. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clin. Gastroenterol. Hepatol. 11(6): 689–694. Elsevier. doi:10.1016/j.cgh.2012.11.035.

International Medical Device Regulation Forum. 2017. Software as a Medical Device (SaMD): Clinical Evaluation.

Jemal, A., E.M. Ward, C.J. Johnson, K.A. Cronin, J. Ma, A.B. Ryerson et al. 2017. Annual Report to the Nation on the Status of Cancer, 1975–2014, Featuring Survival. JNCI J. Natl. Cancer Inst. 109(9). Oxford University Press. doi:10.1093/jnci/djx030.

Jouppi, N.P., A. Borchers, R. Boyle, P. Cantin, C. Chao, C. Clark and et al. 2017. In-Datacenter Performance Analysis of a Tensor Processing Unit. Proc. 44th Annu. Int. Symp. Comput. Archit. - ISCA '17, Vol. 45, 1–12. New York, New York, USA: ACM Press. doi:10.1145/3079856.3080246.

Kanungo, T., D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman and A.Y. Wu. 2002. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. 24(7): 881–892. doi:10.1109/TPAMI.2002.1017616.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma et al. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

Lloyd-Jones, D.M., P.W. Wilson, M.G. Larson, A. Beiser, E.P. Leip, D'Agostino et al. 2004. Framingham risk score and prediction of lifetime risk for coronary heart disease. Am. J. Cardiol. 94(1): 20–24. Excerpta Medica. doi:10.1016/J.AMJCARD.2004.03.023.

Lloyd, S. 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28(2): 129–137. doi:10.1109/TIT.1982.1056489.

Louis, D.N., A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W.K. Cavenee et al. 2016. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 131(6): 803–820. doi:10.1007/s00401-016-1545-1.

Marr, B. 2017. 28 Best Quotes About Artificial Intelligence. Forbes. Retrieved August 30, 2018, from https://www.forbes.com/sites/bernardmarr/2017/07/25/28-best-quotes-about-artificial-intelligence.

MolecularNeuropathology. (n.d.). Retrieved May 31, 2018, from https://www.molecularneuropathology.org/mnp.

Monroe, D. 2017. Deep learning takes on translation. Commun. ACM 60(6): 12–14. doi:10.1145/3077229.

Nathan, D.M. 2014. The diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: Overview. Diabetes Care 37(1): 9–16. American Diabetes Association. doi:10.2337/dc13-2112.

National Nurses United. 2014. Insist on a Registered Nurse. Retrieved from https://soundcloud.com/national-nurses-united/radio-ad-algorithms.

Niemeijer, M., B. van Ginneken, M.J. Cree, A. Mizutani, G. Quellec, C.I. Sanchez et al. 2010. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans. Med. Imaging 29(1): 185–195. doi:10.1109/TMI.2009.2033909.

Norton, M.E., B. Jacobsson, G.K. Swamy, L.C. Laurent, A.C. Ranzini, H. Brar et al. 2015. Cell-free DNA Analysis for Noninvasive Examination of Trisomy. N. Engl. J. Med. 372(17): 1589–1597. Massachusetts Medical Society. doi:10.1056/NEJMoa1407349.

Pencina, M.J., R.B. D'Agostino, M.G. Larson, J.M. Massaro and R.S. Vasan. 2009. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation 119(24): 3078–84. doi:10.1161/CIRCULATIONAHA.108.816694.

Peterson, W., T. Birdsall and W. Fox. 1954. The theory of signal detectability. Trans. IRE Prof. Gr. Inf. Theory 4(4): 171–212. doi:10.1109/TIT.1954.1057460.

Poldervaart, J.M., M. Langedijk, B.E. Backus, I.M.C. Dekker, A.J. Six, P.A. Doevendans et al. 2017. Comparison of the GRACE, HEART and TIMI score to predict major adverse cardiac events in chest pain patients at the emergency department. Int. J. Cardiol. 227: 656–661. Elsevier. doi:10.1016/J.IJCARD.2016.10.080.

Poon, H., C. Quirk, C. DeZiel and D. Heckerman. 2014. Literome: PubMed-scale genomic knowledge base in the cloud. Bioinformatics 30(19): 2840–2842. doi:10.1093/bioinformatics/btu383.

Poplin, R., A.V. Varadarajan, K. Blumer, Y. Liu, M.V. McConnell, G.S. Corrado et al. 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. 2(3): 158–164. Nature Publishing Group. doi:10.1038/s41551-018-0195-0.

Quattoni, A., M. Collins and T. Darrell. 2008. Transfer learning for image classification with sparse prototype representations. 2008 IEEE Conf. Comput. Vis. Pattern Recognit., 1–8. IEEE. doi:10.1109/CVPR.2008.4587637.

Reichlin, T., W. Hochholzer, S. Bassetti, S. Steuer, C. Stelzig, S. Hartwiger et al. 2009. Early Diagnosis of Myocardial Infarction with Sensitive Cardiac Troponin Assays. N. Engl. J. Med. 361(9): 858–867. Massachusetts Medical Society. doi:10.1056/NEJMoa0900428.

Savova, G.K., J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler et al. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Informatics Assoc. 17(5): 507–513. Oxford University Press. doi:10.1136/jamia.2009.001560.

Schwartz, B. 2017. Tweet. Retrieved August 30, 2018, from https://twitter.com/xaprb/status/930674776317849600.

Schwartz, W.B., R.S. Patil and P. Szolovits. 1987. Artificial Intelligence in Medicine. N. Engl. J. Med. 316(11): 685–688. Massachusetts Medical Society. doi:10.1056/NEJM198703123161109.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15: 1929–1958.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. IEEE Conf. Comput. Vis. Pattern Recognit. 2818–2826.

Tanaka, F., K. Yoneda, N. Kondo, M. Hashimoto, T. Takuwa, S. Matsumoto et al. 2009. Circulating tumor cell as a diagnostic marker in primary lung cancer. Clin. Cancer Res. 15(22): 6980–6. American Association for Cancer Research. doi:10.1158/1078-0432.CCR-09-1095.

Tankova, T., N. Chakarova, L. Dakovska and I. Atanassova. 2012. Assessment of HbA1c as a diagnostic tool in diabetes and prediabetes. Acta Diabetol. 49(5): 371–378. doi:10.1007/s00592-011-0334-5.

Thompson, I.M., D.P. Ankerst, C. Chi, M.S. Lucia, P.J. Goodman, J.J. Crowley et al. 2005. Operating Characteristics of Prostate-Specific Antigen in Men With an Initial PSA Level of 3.0 ng/mL or Lower. JAMA 294(1): 66. American Medical Association. doi:10.1001/jama.294.1.66.

Toh, C. and J.P. Brody. 2018. Analysis of copy number variation from germline DNA can predict individual cancer risk. bioRxiv 303339. Cold Spring Harbor Laboratory. doi:10.1101/303339.

Tosteson, A.N.A., D.G. Fryback, C.S. Hammond, L.G. Hanna, M.R. Grove, M. Brown et al. 2014. Consequences of False-Positive Screening Mammograms. JAMA Intern. Med. 174(6): 954. doi:10.1001/jamainternmed.2014.981.

QUESTIONS

1. What types of medical problems can be tackled by machine learning? What types of problems cannot?
2. When is the area under the curve (AUC) an inappropriate metric for quantifying the performance of a classification algorithm?
3. What types of biases will a machine-learning algorithm contain?
4. When is a machine learning algorithm subject to FDA regulatory approval?

PROBLEMS

Each student will suggest a real life dataset that machine learning could be applied to. The dataset should represent a single medical condition or disease. The dataset is easiest to imagine as a table. Each row represents a different patient. The first column should indicate whether the patient is positive/negative for the medical condition. The other columns are variables specific to the patient. Examples of variables are blood pressure, resting pulse, height, weight, results from any blood test, etc.