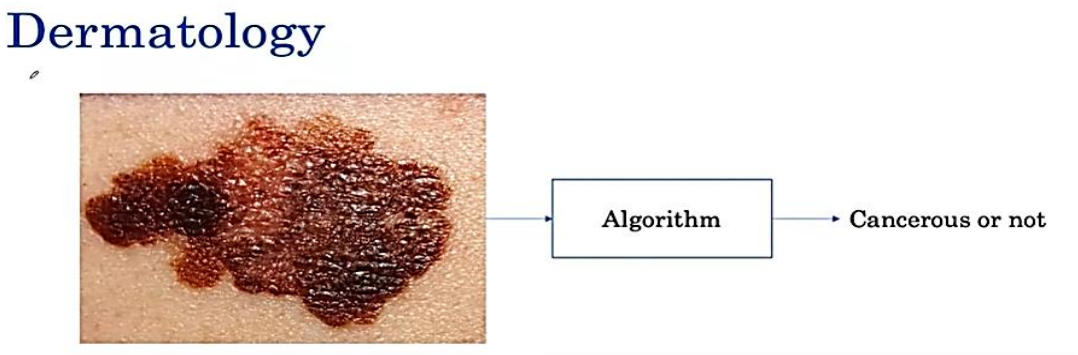
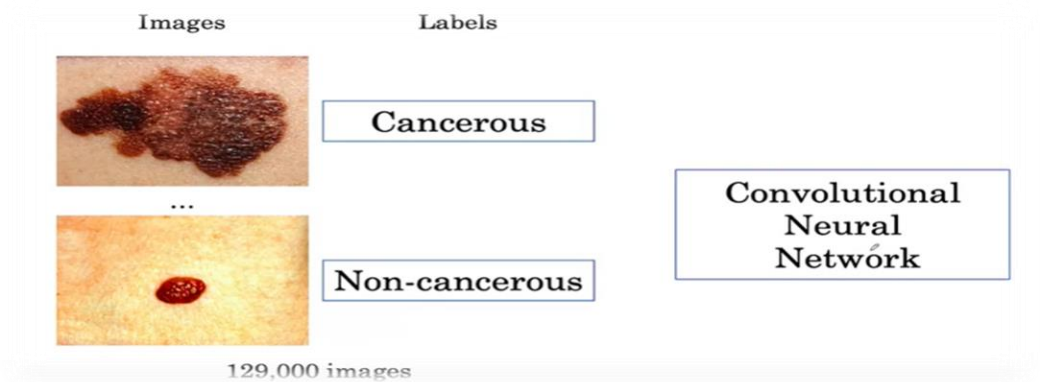
AI for Medicine Specialization
AI for Medical Diagnosis
Week 1
Medical Image Diagnosis

فى هذا الاسبوع سوف نغوص بإستقامة فى بناء موديل فى التعلم العميق لتصنيف للأشعة الصدرية العادية الكثير من الافكار سوف نتعلمها خلال هذا المثال (الأشعة الصدرية) القابل للتطبيق على نطاق واسع في العديد من الاختبارات التصوير الطبي
فى هذا الاسبوع سوف نبدأ فى النظر إلى ثلاثة أمثلة من ال medical diagnostic tasks حيث حقق التعليم العميق أداء مذهل. سوف نقفز إلى the training procedure لبناء (AI models for medical imaging)
فى النهاية سوف ننظر الى
the testing procedure for evaluating the performance of these models on real data

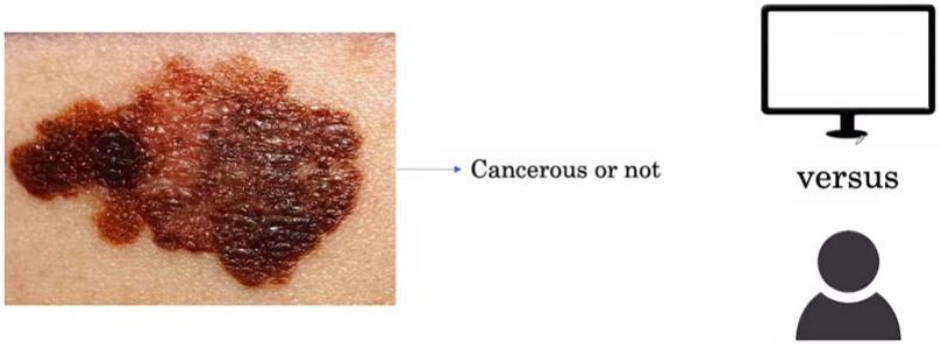
Our first example is in dermatology
طب الأمراض الجلدية



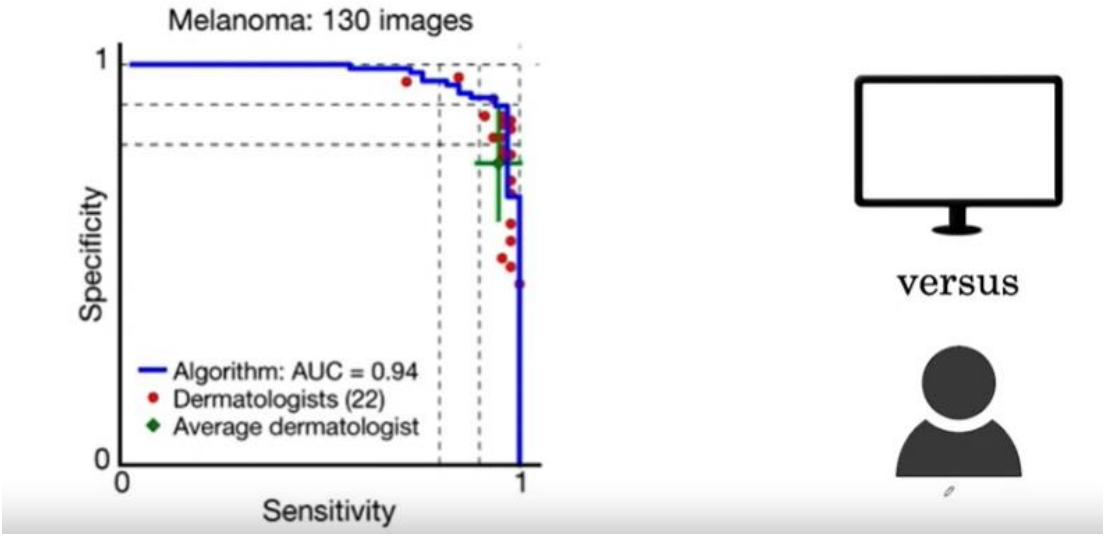
Dermatology هو فرع من الطب للتعامل مع الجلد وواحد من مهام ال dermatologists هو النظر الى المناطق المشبوهه فى الجلد لاتخاذ قرار اذا كانت A mole هوا cancerاو ليس
الكشف المبكر قد يكون لها تأثير كبير على نتائج سرطان الجلد
معدل البقاء على قيد الحياة لمدة خمس سنوات
من نوع واحد من سرطان الجلد يسقط بشكل ملحوظ إذا
اكتشفت في مراحل لاحقة
في هذه الدراسة، خوارزمية مدربه لتحديد ما إذا كانت المنطقة من الجلد تكون الأنسجة السرطانية أم لا



بإستخدام مئات من آلاف الصور المسماه كمدخل ف ال CNNتستطيع ان تتدرب على مثل هذه المهام فسوف نرى التدريب على مثل هذه الالجورزمات فى هذا الكورس



بمجرد ان الالجورزم تم تدريبه ف ال the algorithms predictions يستطيع ان يقيم ضد تنبؤات أخصائيو الأمراض الجلدية البشرية لمجموعه جديدة من الصور

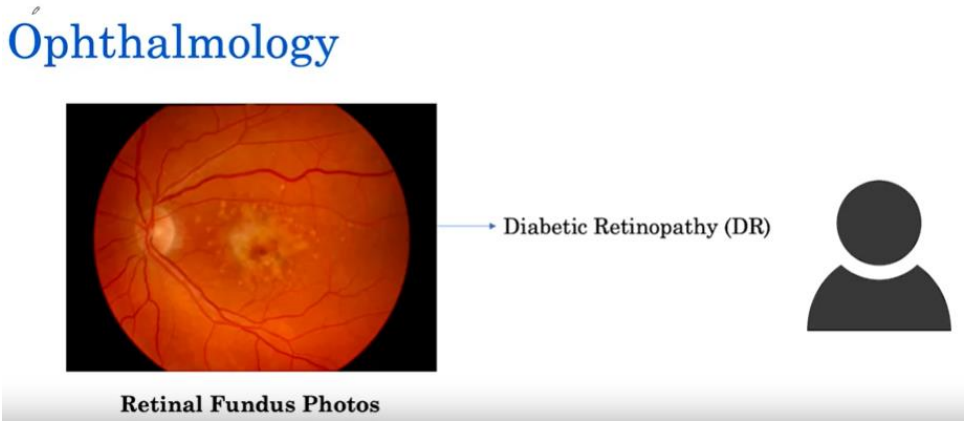


فى هذه الدراسة سوف نجد ان اداء الالجورزم زى بالظبط اداء ال dermatologists
لاتقلق بخصوص تفسير هذا الجراف فى الاسابيع القادمة من الكورس سوف نتتطلع اى تقييم الموديل باستخدان ال CURVE فالاستنتاج الرئيسى انك ستبقى قادرا على استخلاص من هذا الجراف ان الالجورزم يتنبأ بدقة بمقارنة بتنبا ال human dermatologists
للمزيد من القراءة يمكنك قراءة هذه الورقة البحثية
Dermatologist-level classification of skin cancer with deep neural networks
https://www.nature.com/articles/nature21056

انتهى الفيديو الاول
George Samuel

Lecture Eye Disease and Cancer Diagnosis

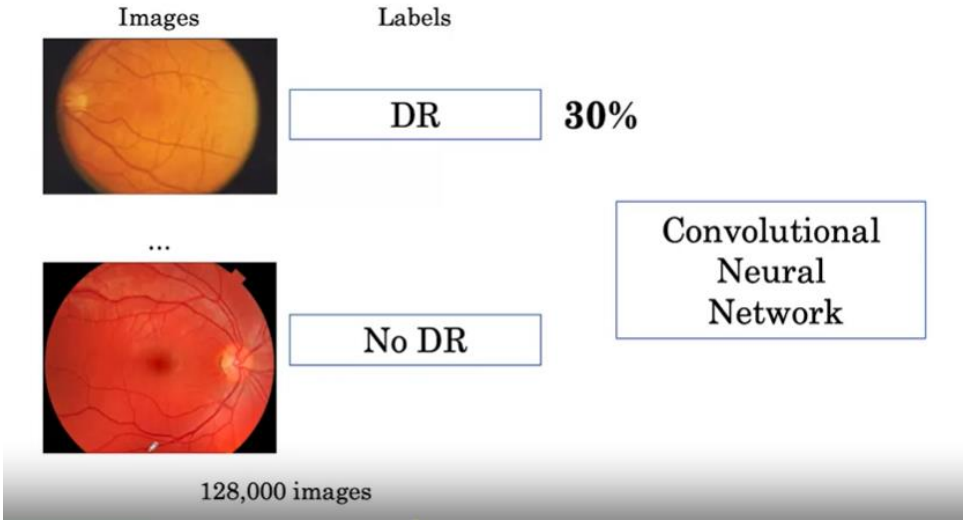
Our second example is an ophthalmology



ophthalmology التي تتعامل مع تشخيص وعلاج اضطرابات العين

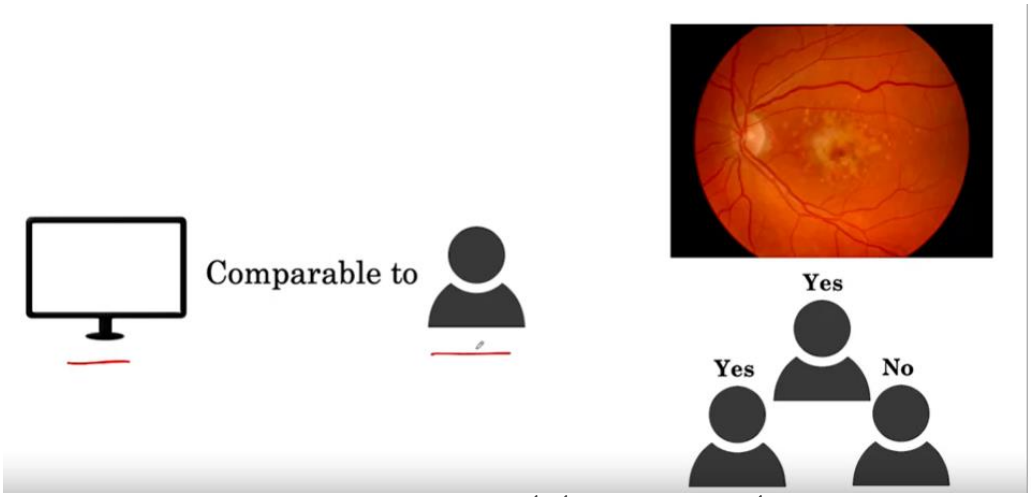
- دراسة واحدة معروفة في ٢٠١٦ بدات لصور في شبكية العين التي صورت الجزء الخلفي من العين للنظر هنا هو اعتلال الشبكية السكري وهو إلحاق الضرر شبكية العين الناجمة عن مرض السكري يعد من أهم أسباب العمى .

حاليا، كشف DR هو مضيعة للوقت و العملية يدوية تتطلب تدريب الاطباء لفحص هذه الصور .



فى هذه الدراسة تم تطوير هذا الألجورزم ليقرر اذا كان المرضى يعانون من اعتلال الشبكية السكري بالنظر الى الكثير من الصور فهذه الدراسة تستخدم أكثر من ١٢٨٠٠٠ صورة منها ٣٠ في المئة فقط كان اعتلال الشبكية السكري .

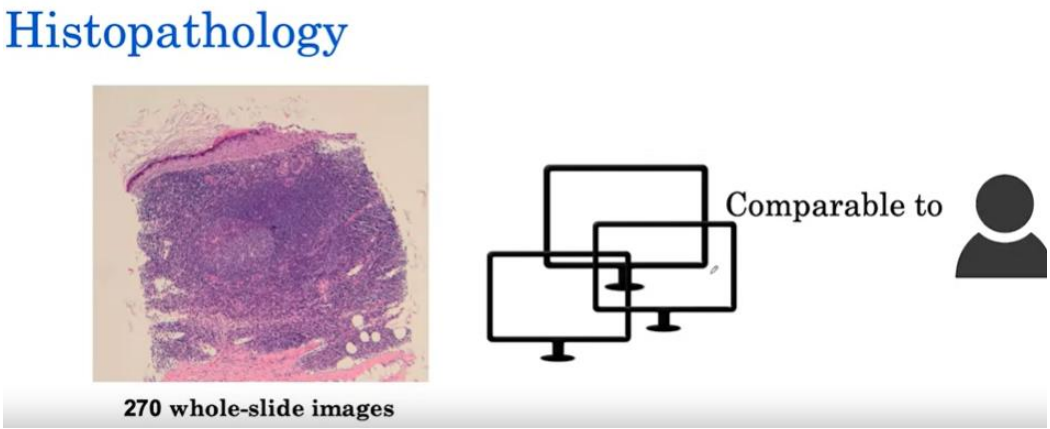
- دعونا نلقي نظرة على هذه البيانات مشكلة عدم الاتزان التى بارزه جدا في الطب والعديد من المجالات الأخرى باستخدام بيانات العالم الحقيقي .real-world data
- سوف نرى بعض الطرق التعامل مع هذا التحدي .



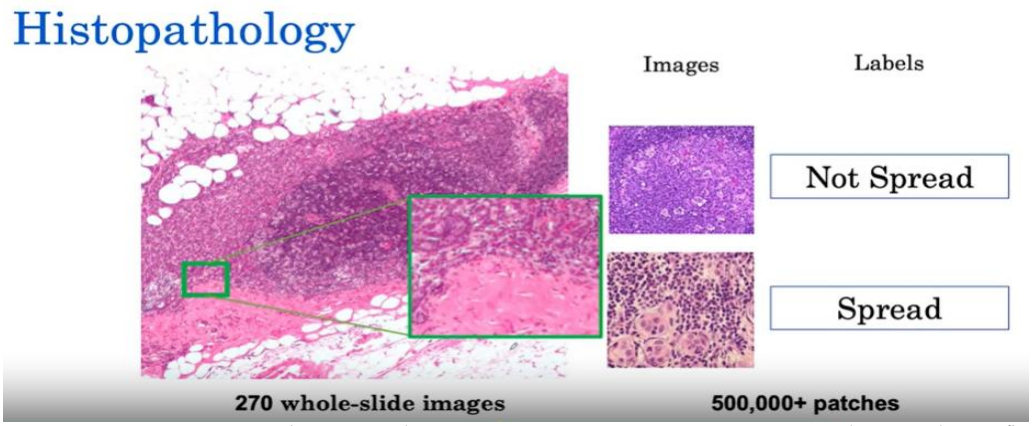
على غرار الدراسات السابقة أظهرت هذه الدراسة أن أداء الخوارزمية الناتجة كان مشابه لعلماء العيون . ففى هذه الدراسة majority vote لاكثر من طبيب عيون استخدموا مجموعه من ال reference standardاو ال ground troops اللى هيا مجموعه من الخبراء للتخمين الجيد والاجابة الصحيحة .
في وقت لاحق من هذا الأسبوع في الدورة ، ونحن سوف ننظر في كيفية أن ground-truth يمكن أن تحدد في مثل هذه الدراسات الطبية AI.

Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy
https://www.nature.com/articles/s41433-018-0269-y

Our third example is in histopathology,



المثال الثالث هو علم الانسجة تخصص طبي يتضمن فحص الأنسجة تحت المجهر
إحدى المهام لعلماء الأمراض ينظرون إلى الصور المجهرية الممسوحة للأنسجة المسماة صور الشريحة الكاملة ، ويحددون مدى انتشار ال cancer .
وهذا أمر مهم من أجل المساعدة في تخطيط العلاج، التنبؤ بمسار المرض وفرصة للتعافى .
فى دراسة سنة ٢٠١٧ تم استخدام ٢٧٠ صورة وال ذكاء الصناعى تم تطوير الجوؤزم وتم تقييمه ضد اطباء العيون ووجد ان الالجورزم افضل ف الاداء من اطباء العيون



الآن في التشريح، الصور بتكون كبيرة جدا ومانقدرش نغذى بيها الالجورزم بشكل مباشر بدون تقسيمها

الإعداد العام

من هذه الدراسات هو أنه بدلا من صور كبيرة ودجودة عالية ممكن استخلاص مجموعه من الباتشيس بتكبير عالى واستخدامها فى التدريب

وهذه الباتشيس تكون مسماه باسم اصلى خاص البصورة كلها وبعدين نغذى بها الالجورزم وبهذه الطريقة يستطيع الالجورزم ان يتعلم من مئات او الوف من الباتشيس وفى هذا الكورس سوف نطبق فكرة كيف نكسر الصور الكبيرة الى عده باتشات علشان ندرب بيها المودل لمهه مثل brain tumor segmentation

Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer.

https://www.ncbi.nlm.nih.gov/pubmed/30312179

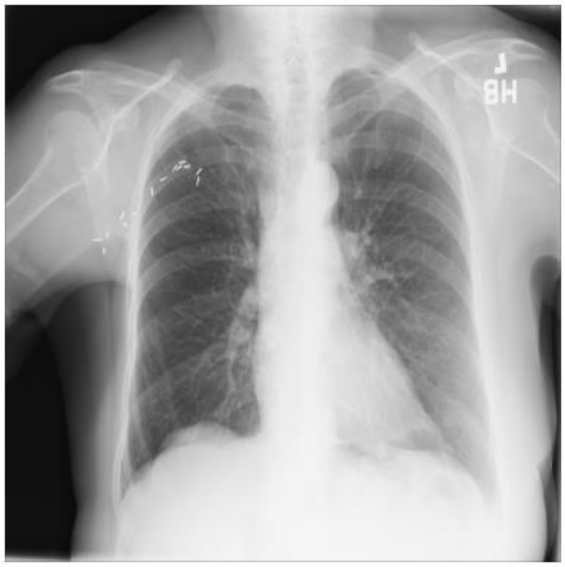
انتهى الفيديو الثانى

Gsamuel

Lab-1
Data Exploration & Image Pre-Processing

This lab will give you a little bit of practice loading and processing the chest x-ray dataset you'll be working with in this week's assignment. You aren't required to write any code in this notebook but feel free to explore, add or change things, and familiarize yourself with this exciting dataset.

This lab is ungraded, but you may find the examples here to be useful when you work on this week's graded assignment!

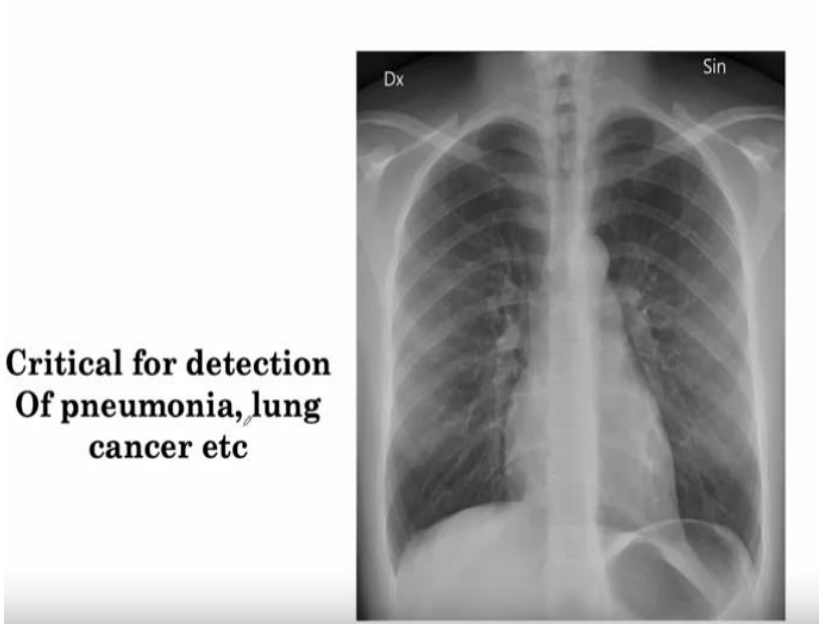


فى اول واجب فى هذا الكورس سوف نشتغل على صور الاكس راى المأخوذه من[public ChestX-ray8 dataset]

Building and Training a Model for Medical Diagnosis

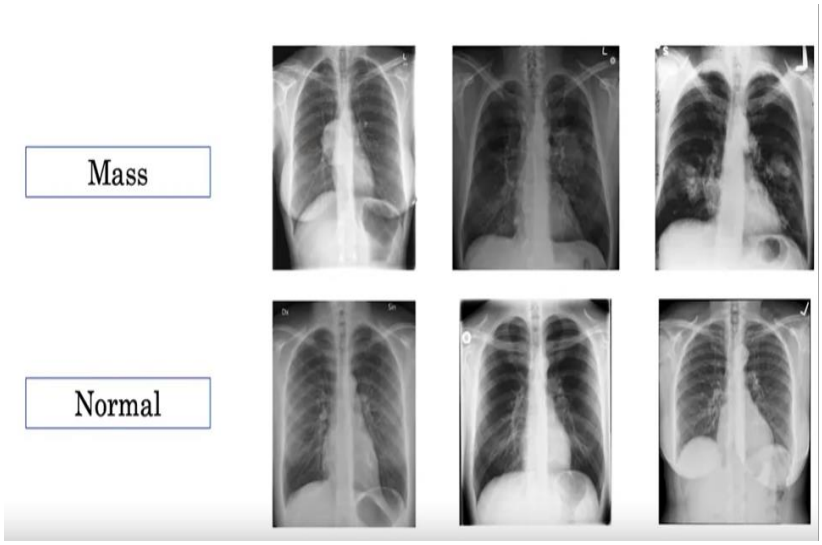
الآن بعد أن كنت قد رأيت بعض
تطبيق التعلم العميق في تصنيف الصور الطبية
دعونا نلقي نظرة فى كيفية بناء نموذج للتعلم العميق لصور طبية لمهمة تستخدم فى
using chest X-rays to detect multiple
diseases with a single model
نحن نسير خلال عملية التدريب لموديل الاكس راى وسوف نلقى نظرة على مفتاح التحدى التى سوف نواجهها فى هذه العملية واذاى هانقدر بشكل ناجح اننا نتغلب عليها

We'll start by looking at the task of chest X-ray interpretation.



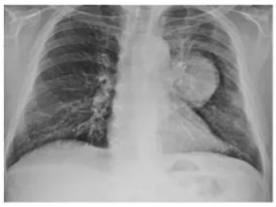
The chest X-ray is one of the most common diagnostic Imaging procedures in medicine with about 2 billion chest X-rays that are taken for a year. Chest X-ray interpretation is critical for the detection of many diseases , including pneumonia and lung cancer which affect millions of people worldwide each year.

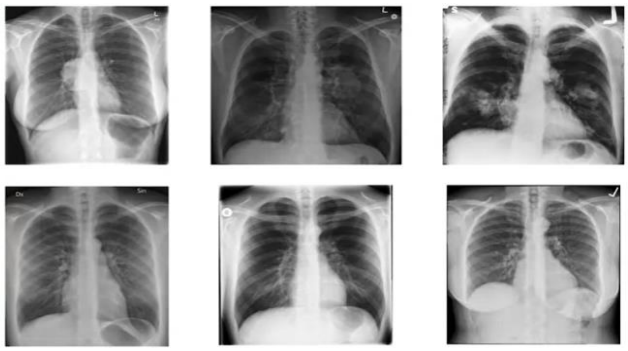
فال الاشعة الصدرية واحدة من أهم التشخيصات الشائعة لاجراءات الصور فى الطب حوالى ٢ بليون اشعة صدر تأخذ سنويا وتفسير الأشعة السينية للصدر أمر بالغ الأهمية لكشف العديد من الأمراض,بما في ذلك الالتهاب الرئوي وسرطان الرئة الذي يؤثر على الملايين من الناس في جميع أنحاء العالم كل عام



Now a radiologist who is trained in the interpretation of chest X-rays looks at the chest X-ray, looking at the lungs, the heart, and other regions to look for clues that might suggest if a patient has pneumonia or lung cancer or another condition.

الآن ، الأشعة السينية التي تم تدريبهم على تفسير الصدر بالأشعة السينية سوف نلاحظ الصدر بالأشعة السينية ، ومراقبة الرئتين والقلب وغيرها من المناطق للبحث عن أدلة على ما إذا كان المريض يعاني من الالتهاب الرئوي أو سرطان الرئة أو غيرها من الأمراض





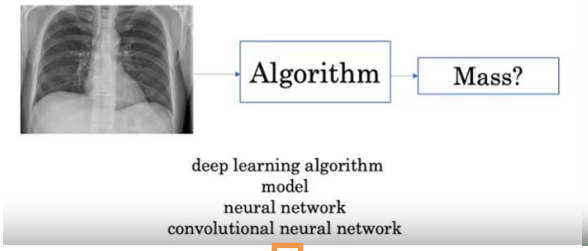
Let's look at one abnormality called a mass looks like. And I'm not going to first define what a mass is but let's look at three chest X-rays that contain a mass and three chest X-rays that are normal. I can then show you a new chest X-ray here and ask you to identify whether there is a mass. You might be able to correctly identify that this chest X-ray contains a mass

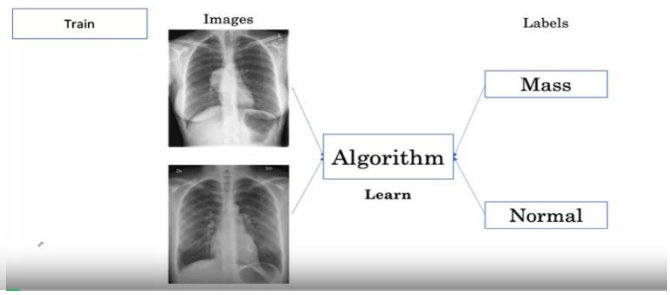
تعالوا نبص على صورة من ال abnormality ودى فيها mass بس دلوقتى مش هانعرف يعنى ايه ماس بس هانعرض ٣ صورة صدرية تحتوى على ماس و٣ آخرين طبعيين ممكن نعيطكم صورة جديدة للتعرف هل يوجد ماس ولا لأ فانت لاشك تستطيع التعرف على الصورة التى تحتوى على ماس

here's the mass that might look similar to things that you see in these images, but not similar to anything that you see in these images. The way you're learning is very similar to how we're going to teach an algorithm to detect mass. For our own reference, a mass is defined as a lesion or in other words damage of tissue seen on a chest X-ray as greater than 3 centimeters in diameter. Let's see how we can train our algorithm to identify masses

هنا يظهر الماس زى ماهوا باين ف الصور لكن مش زى باقى الصور ودى نفس الطريقة اللى هانعلم بها الموديل بناتعنا ازاى يكتشف الماس وطبقا للمراجع الطبية فالماس بيقى حجم قطره اكبر من ٣ سم تعالو بقى نشوف الموديل بتاعنا هايتعلم الماس ازاى

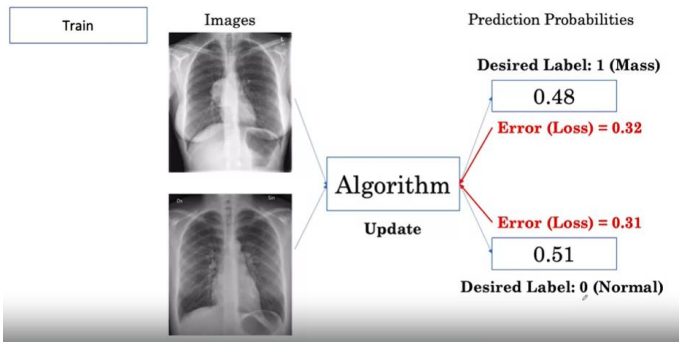
Training, prediction, and loss





اثناء التدريب الالجورزم بيعرض صور الاشعة الصدرية اللى اصلا معمول لها تسمية واللى يتحتوى على ماس او لاتحتوى فالالجورزم هايتعلم من الصور وال LABEL بتاعها وبعدين الخوارزمية تتعلم في نهاية المطاف أن تذهب من مدخل الأشعة السينية الصدر لإنتاج ما إذا كانت الأشعة السينية تحتوي على ماس. والالجورزم ده ليه عدخ اسماء ممكن تسمع عن deep learning algorithm او model او Neural networkاو CNN

During training, an algorithm is shown images of chest X-rays labeled with whether they contain a mass or not. The algorithm learns using these images and labels. The algorithm eventually learns to go from a chest X-ray input to produce the output of whether the X-ray contains mass. And this algorithm can go by different names. You may have heard of the terms deep learning algorithm or model or neural network or convolutional neural network.

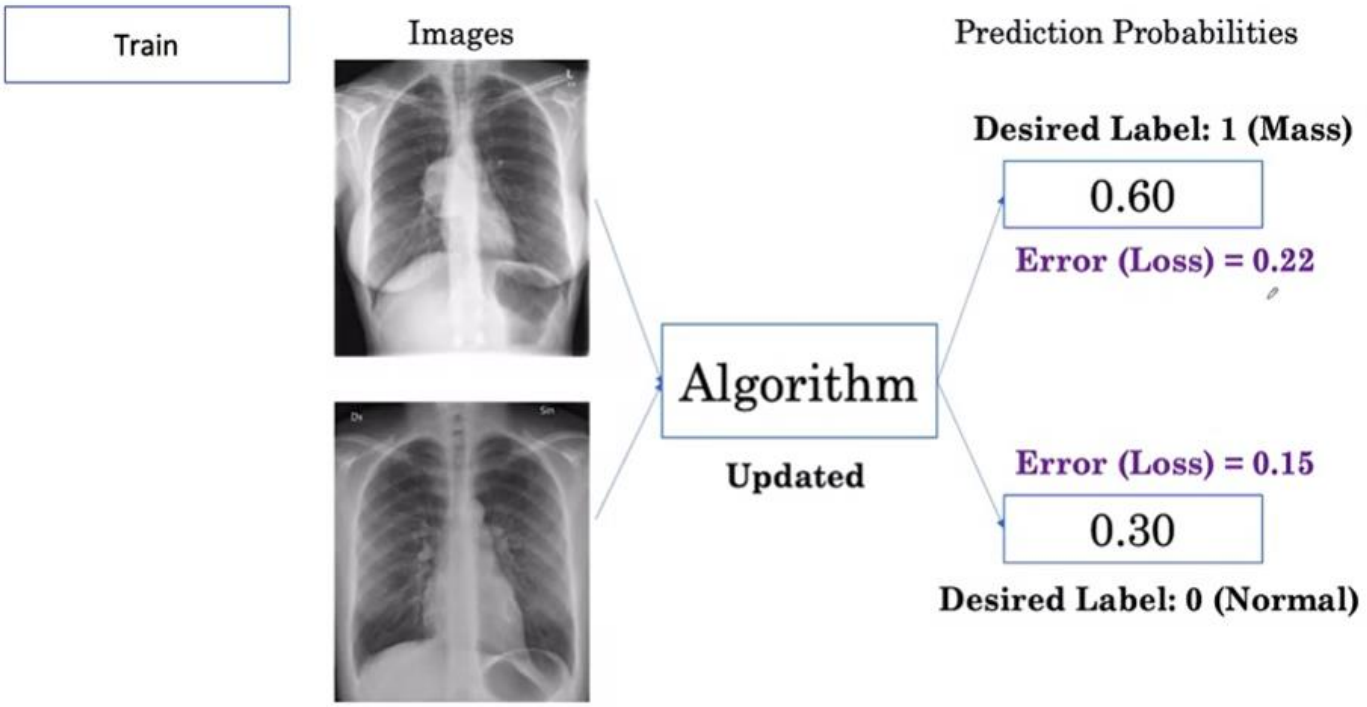


The algorithm produces an output in the form of scores, which are probabilities that the image contains a mass. So the probability that this image contains a mass is outputted to be 0.48, and the probability for this image is outputted to be 0.51. When training has not started, these scores, these probability outputs are not going to match the desired label.

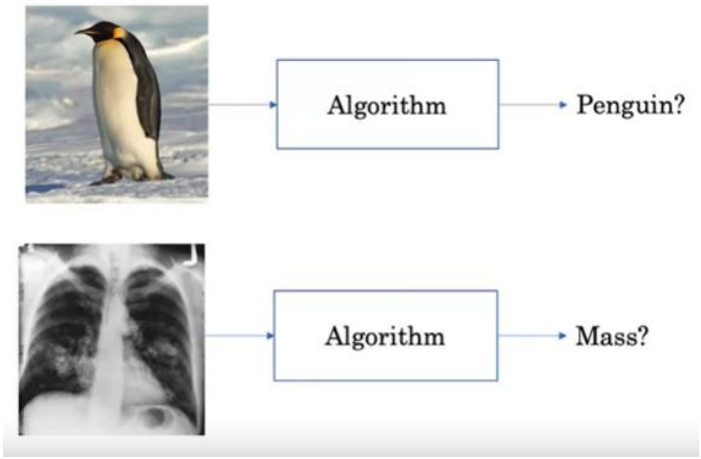
Let's say the desired label for mass is 1, and for normal is 0. And you can see that 0.48 is far off from 1 and 0.51 is far off from the desired label of 0. And we can measure this error by computing a loss function. A loss function measures the error between our output probability and the desired label. We'll look at how this loss is computed soon enough.

Then, a new set of images and desired labels is presented to the algorithm as it learns to produce scores that are closer to the desired labels over time. Notice how this output probability is getting closer to 1, and this output probability is getting closer to 0.

فالموديل هايتنج مخرجات بدرجة معينه يعنى فى شكل probabilities والاحتمال الاكبر هوا اللى هايعبر عن الصورة اذا كانت ماس ولا لأ فهنا فى الصورة احتمالية انه ماس بنسبة 0.48 واحتمالية ان الصورة تبقى 0.51 التدريب لم يبدأ بعدفهذه الاحتمالية لاتضاهى the desired label. فخلينا نقول ان the desired label للماس ب ١ وللحالة الطبيعية ب صفر ونقدر ساعتها نقول ان 0.48 بعيد جدا اننا نقول انها تنتمى لل ١ ونفس الكلام ال 0.51 بعيد جدا نها تبقى ٠ صفر فساعتها نقدر نقيس الخطأ اننا نحسب ال loss function وال loss function هيا اللى بتقيس الخطا مابين ال output probability and the desired label وهانعرف ازاى بتتحسب قريب جدا ... وبعد كده هايجى مجموعه صورة اخرى هتدخل للموديل والموديل هايتعلمها علشان يقترب من التسميات المطلوب بمرور الوقت



Training, prediction, and loss

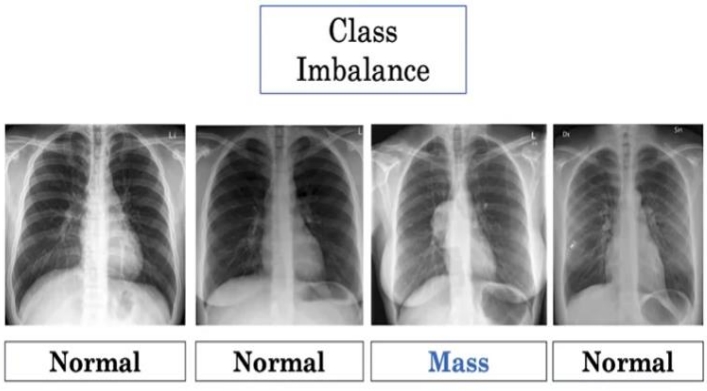


Typically, in the medical AI examples we've seen in previous videos, hundreds of thousands of images are show into the algorithm. This is a typical setup for image classification, which is a core task in the computer vision field where a natural image is input to an image classification algorithm, which says what is the object contained in the image. You may have seen deep learning algorithms that can do this.
Our example of chest X-ray classification is similar in many ways to the image classification setup. There a few additional challenges which make training medical image classification algorithms more challenging, which we'll cover next.

عادة ، في أمثلة الذكاء الإصطناعي الطبية التي شاهدناها في الفيديوهات السابقة ، مئات الآلاف من الصور تظهر في الخوارزمية.ود بشكل مثالى بيتم اعداده فى تصنيف الصور اللى بتكون مهمه اساسية فى نظام computer vision حيث تدخل الصور الطبيعية لل image classification algorithm علشان يقدر يحدد الاوبجيكت الموجود بالصورة وطبعا اكيد انك تعرضت لهذا النموذج من التعلم العميق لاكتشاف objects فى المثال اللى هنا xray classification فيه بعض التحديات البسيطة فهانتكلم عن ٣ تحديات لتدريب الموديل فى الصور الطبية the class imbalance challenge, the multitask challenge and the dataset size challenge ولكل تحدى سوف نغطى أكثر من تكنيك

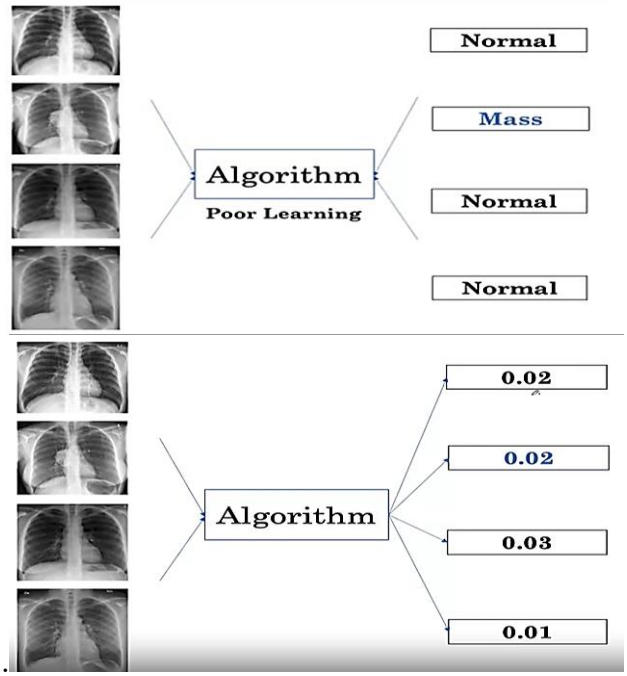
We'll talk about three key challenges for training algorithms on medical images; the class imbalance challenge, the multitask challenge and the dataset size challenge. For each challenge, we'll cover one or two techniques to tackle them

Let's start with the class imbalance challenge. So here's the challenge. There's not an equal number of examples of non-disease and disease in medical datasets. This is a reflection of the prevalence or the frequency of disease in the real-world, where we see that there are a lot more examples of normals than of mass, especially if we're looking at X-rays of a healthy population. In a medical dataset, you might see 100 times as many normal examples as mass examples

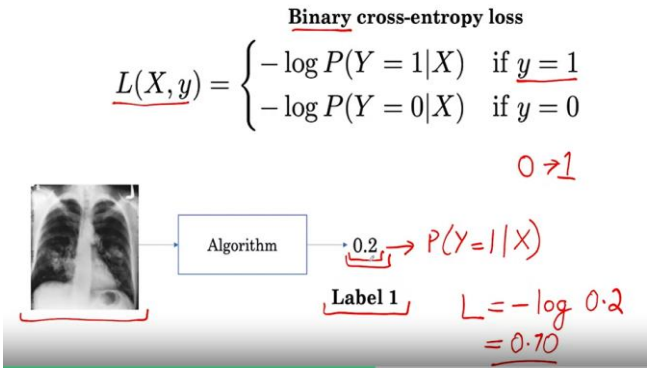


فال imbalance بيبقى عدد غير متساوى من الصور اللى فيها امراض والتى لا تشمل امراض فده وهيعكس مدى الانتشار او تكرار المرض فى الواقع لاننا سوف نرى امثلة كثيرة نورمال اكثر من اللى ظاهر فيها ماس ولا سيما إذا كنا نبحث فى الأشعة السينية على صحة السكان

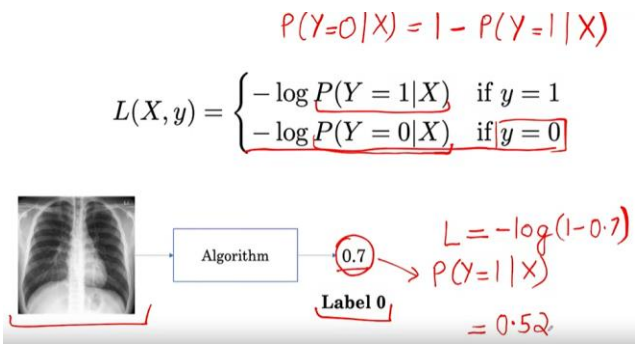
Binary Cross Entropy Loss Function



This creates a problem for the learning algorithm would seize mostly normal examples. This yields a model that starts to predict a very low probability of disease for everybody and won't be able to identify when an example has a disease. Let's see how we can trace this problem to the loss function that we use to train the algorithm. We'll also see how we can modify this loss function in the presence of imbalanced data. This loss over here is called the binary cross-entropy loss and this measures the performance of a classification model whose output is between zero and one Let's look at an example to see how this loss function evaluates. Here we have an example of a chest x-ray that contains a mass, so it gets labeled with one and the algorithm outputs a probability of 0.2. Now, the 0.2 here is the probability according to the algorithm of Y being equal to 1

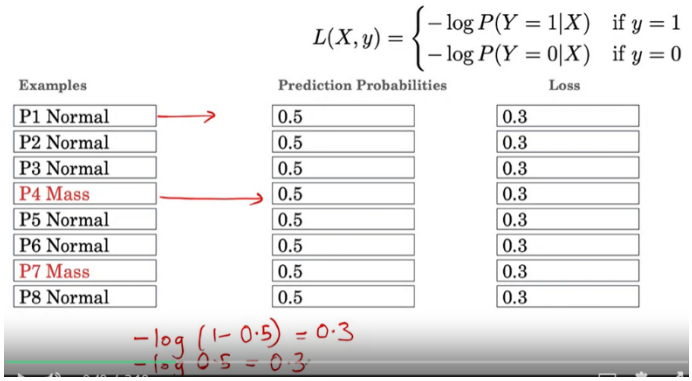


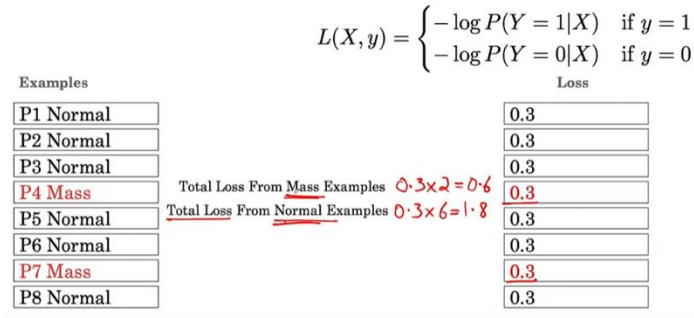
the probability that this example is a mass. So now, we can apply the loss function to compute
the loss on this example. Now notice that our label is one, so we're going to use the first term. Our loss is negative log and then we're going to take the algorithm output, 0.2. So this evaluates to 0.70. So this is the loss that the algorithm gets on this particular example. Let's look at another example. This time a non-mask example, which would have a label of zero. Our algorithm outputs
a probability of 0.7. Now, this time we're going to use this term of the loss rate here because Y equals 0. So now the loss is going to be negative log of the term PY equals 0 given X. We can get PY equals 0 given X using P of Y equals 1 given X. The way we can compute this quantity from that one is by recognizing that the probability that
an example is zero is 1 minus the probability that it's 1. An example as either mass or not. So the algorithm says 70 percent probability that something is mass, then there's 30 percent
probability it's not. So here we're going to
plug in 1 minus 0.7, that's going to
come out to 0.3 and this expression
evaluates to 0.52

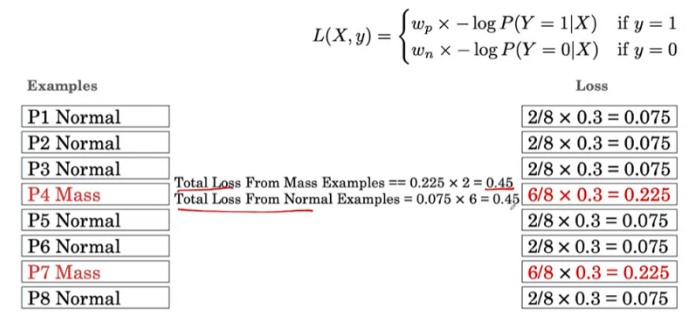


Impact of Class Imbalance on Loss Calculation

overall to the loss, as the normal examples. Let's pick six over eight as the weight we have on the mass examples, and two over eight as the weight we have on the normal examples. Then, you can see that if you sum up the total loss from the mass example, we get 0.45, and this is equal to the total loss from the normal examples here. In the general case, the weight we'll put on the positive class will be the number of negative examples over the total number of examples. In our case, this is six normal examples over a total examples. The weight we'll put on the negative class, will be the number of positive examples over the total number of examples, which is two over eight. With this setting of w_p and w_n, we can have over all of the examples for the loss contributions from the positive and the negative class to be the same. So this is the idea of modifying the loss using weights, in this method that's called the weighted loss, to tackle the class imbalance problem.

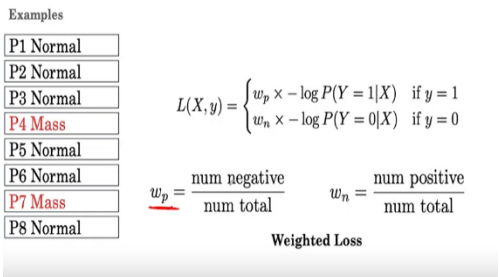






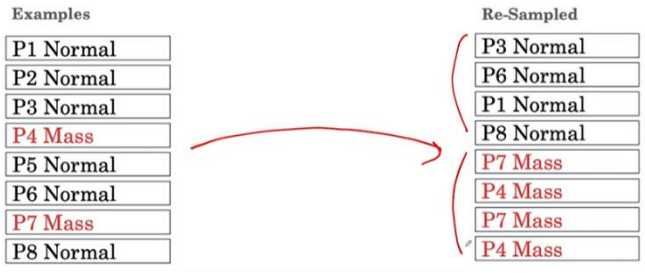
We've seen how the loss is applied to a single example. Let's see how it applies to a bunch of examples. Here we have six examples that are normal, and two examples that are mass. Note that P2, P3, P4 here, are the patient IDs. When the training hasn't started, let's say the algorithm produces an output probability of 0.5 for all of the examples, the loss can then be computed for each of the examples. For a normal example, we're going to use negative log of 1 minus 0.5, which is going to come out to 0.3. For a mass example, we're going to use negative log of 0.5, which is also going to come out to 0.3. The total contribution to the loss from the mass examples, comes out to 0.3 times 2 ,which is 0.6. While the total loss from the normal example, comes out to 0.3 times the 6 normal examples, which is 1.8. So notice how most of the contribution to the loss is coming from the normal examples, rather than from the mass examples. So the algorithm is optimizing its updates to get the normal examples, and not giving much relative weight to the mass examples. In practice, this doesn't produce a very good classifier. This is the class imbalance problem.

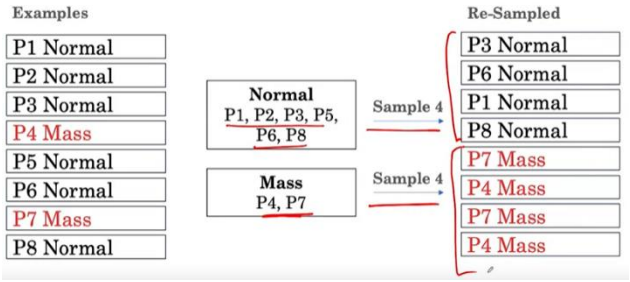
The solution to the class imbalance problem is to modify the loss function, to weight the normal and the mass classes differently. W_p will be the weight we assign to the positive or to the mass examples, and w_n to the negative or normal examples. Let's see what happens when we weight the positive examples more. We want to weight the mass examples more, such that they can have an equal contribution

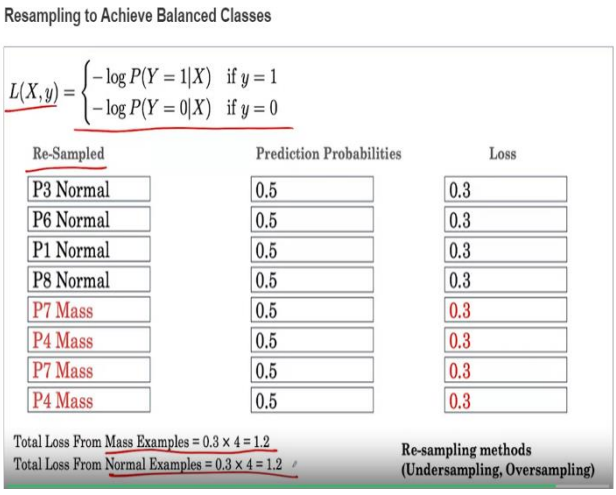


Resampling to Achieve Balanced Classes

Let's discuss another procedure that we can use to tackle the class imbalance problem, which is re-sampling. The basic idea here is to re-sample the dataset such that we have an equal number of normal and mass examples. Let's see how we can achieve this. First, we group the normal and the mass examples together. Notice that the normal group has six examples and the mass has two examples. Now from these groups, we will sample the images such that there's an equal number of positive and negative samples. We can do this by sampling half of the examples from the positive or mass class and half of the examples from the negative or normal class. Note that this means we may not be able to include all of the normal examples in our re-sample. Furthermore, we may have more than one copy of the mass examples in our re-sampled dataset. With this re-sampled dataset, if we now compute the loss in the same way that we did before, we can see that even without the weights, this is just our standard binary cross-entropy loss. We see there's an equal contribution to the loss from the mass examples and from the normal examples. There are many variations of this approach, like under-sampling the normal class or oversampling the mass class, these approaches fall under the category of re-sampling methods, which we can use to combat the data imbalance problem



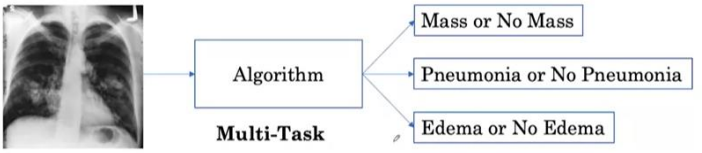




دعنا نناقش إجراء آخر الذي نحن يمكن أن نستعمل لمعالجة class imbalance problem ، الذي هوا re-sampling
الفكرة الأساسية لإعادة اخذ العينات ان احنا يبقى عندنا عدد متساوى من الاشعة السليمة و مثلها مثل الاشعة الغير سليمة

Multi-Task

Let's chat about another challenge that we encounter in the medical image classification setting, which is the multitask challenge. This far, we have looked at binary classification, where we care about classifying whether an example is a mass or not a mass. However, in the real-world, we care about classifying the presence or absence of many such diseases. Now, one simple way to do this, is to have models that each learn one of these tasks. However, maybe we can learn to do all of the tasks using one model. An advantage of this is that we can learn features that are common to identifying more than one disease, allowing us to use our existing data more efficiently. This is the setup of multitask learning. Let's look at how we can train the algorithm to learn all these tasks at the same time. So instead of the examples having one label, they now have one label for every disease in the example where zero denotes the absence of that disease and one denotes the presence of that disease. For the first one, we have an absence of mass, a presence of pneumonia and an absence of another disease, edema, which is excess fluid in the lungs. Instead of having one output from the model, the model now has three different outputs denoting the probability of the three different diseases. To train such an algorithm, we also need to make the modification to the loss function from the binary tasks to the multitask setting. Let's look at how we can do that



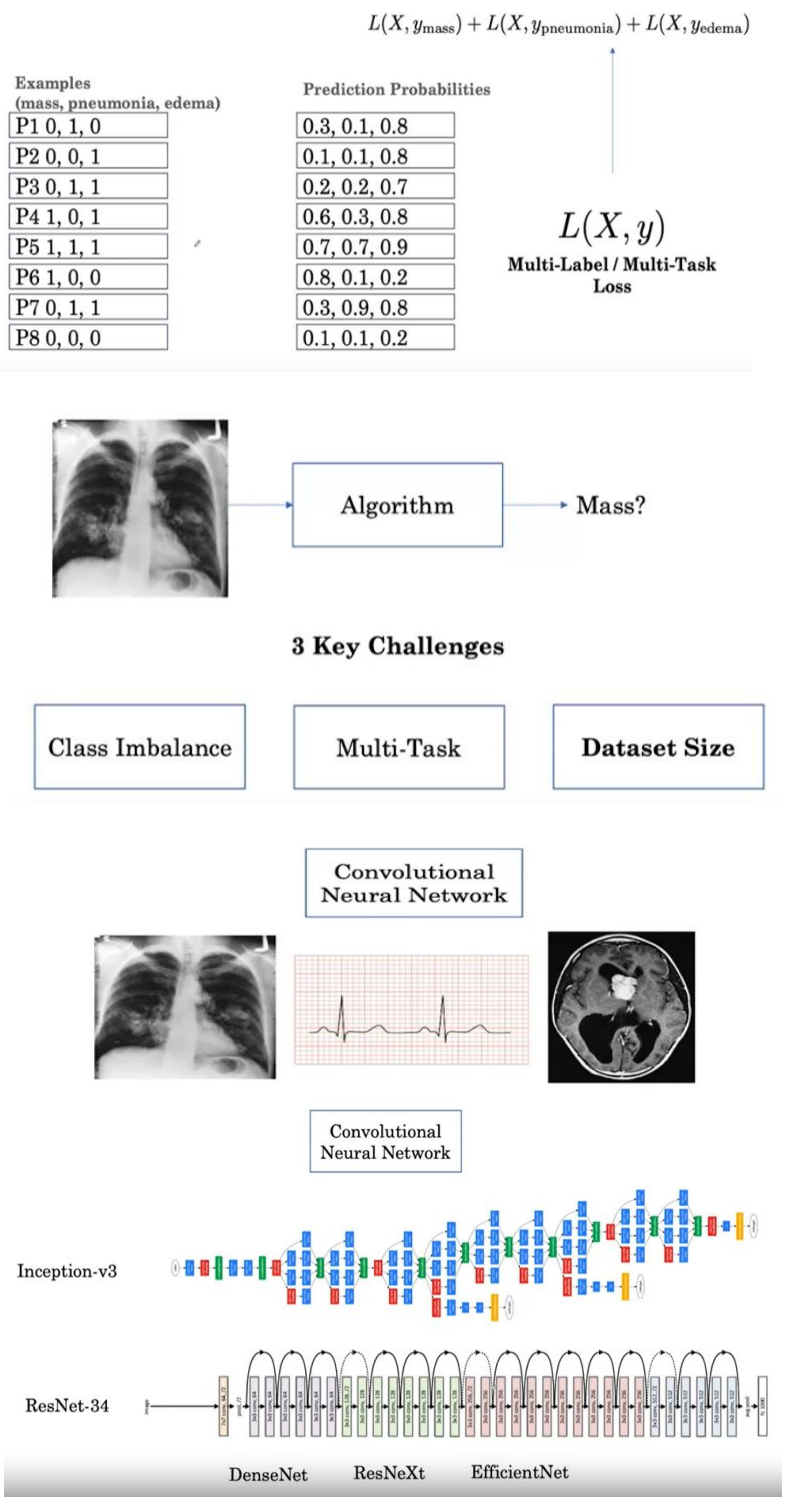
Examples (mass, pneumonia, edema)	Prediction Probabilities
P1 0, 1, 0	0.3, 0.1, 0.8
P2 0, 0, 1	0.1, 0.1, 0.8
P3 0, 1, 1	0.2, 0.2, 0.7
P4 1, 0, 1	0.6, 0.3, 0.8
P5 1, 1, 1	0.7, 0.7, 0.9
P6 1, 0, 0	0.8, 0.1, 0.2
P7 0, 1, 1	0.3, 0.9, 0.8
P8 0, 0, 0	0.1, 0.1, 0.2

$$L(X, y)$$
?

.

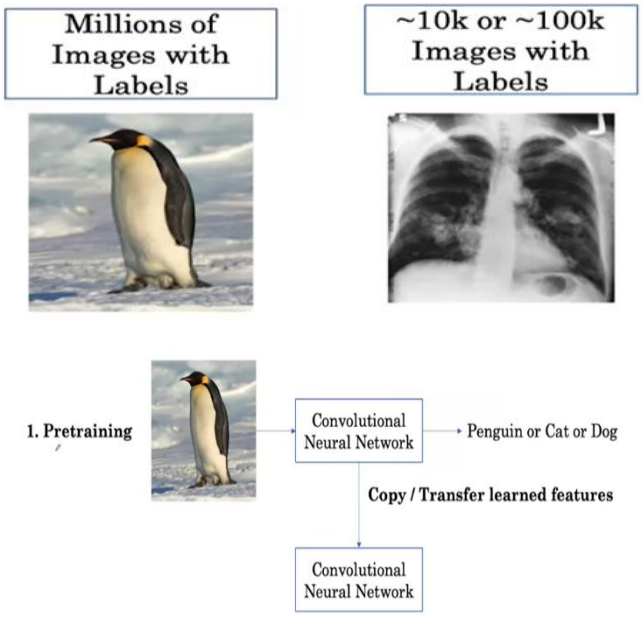
Multi-task Loss, Dataset size, and CNN Architectures

We modify the loss function such that we look at the error associated with each disease. We can represent our new loss as the sum of the losses over the multiple diseases. This is called the multi-label loss or the multi-task loss. In this case, here's the loss that we get for the eight examples. This first term over here represents the loss associated with the mass class using this prediction probability and using this label. Similarly, here we have the loss associated with the pneumonia class coming with this model output and this label. So we add up the three losses given to us by the individual loss function components. One final consideration is how we can account for class imbalance in the multitask setting. Once again, we can apply the weighted loss that we have covered earlier. This time, we not only have a weight associated with just the positive and the negative labels, but it's for the positive label associated with that particular class and the negative label associated with that particular tasks such that for mass, there will be a different way to the positive class than to pneumonia or to edema. This covers their solution to the second challenge of multitask learning. Let's look at the third challenge which is the data set size challenge. For many medical imaging problems, the architecture of choice is the convolutional neural network, also called a ConvNet or CNN. These are designed to process 2D images like x-rays. But variants of these are also well suited to medical signal processing or 3D medical images like CT scans, which we will look at in a future week. Several convolutional neural network architectures, such as Inception ,ResNet, DenseNet, ResNeXt and EfficientNets have been proposed and are widely popular in image classification. These architectures are composed of various building blocks. In medical problems, the standard is to try out multiple models on the desired tasks and see which ones work best.

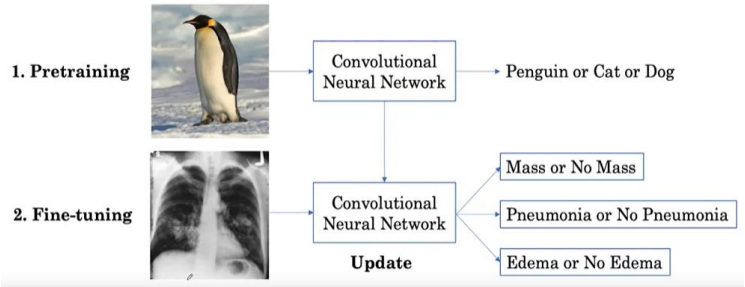


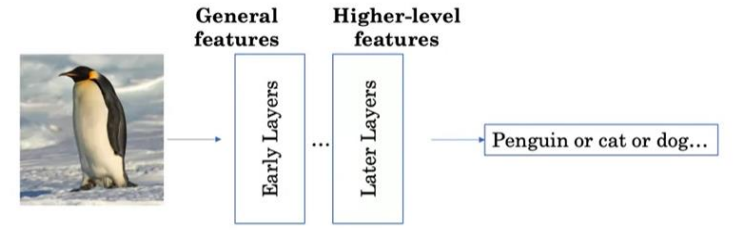
Working with a Small Training Set

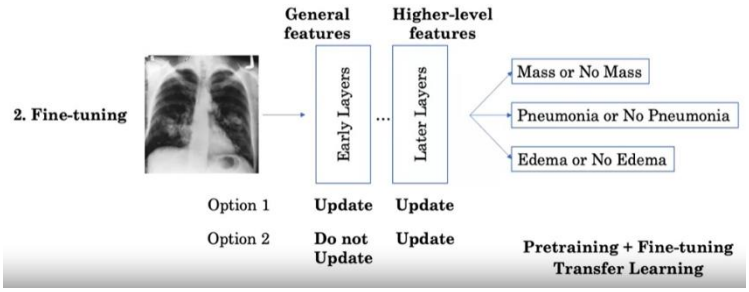
The challenge is that all of these architectures are data hungry and benefit from the millions of examples found in image classification datasets. In medical problems , how can we still apply these techniques when we don't have millions of examples? One solution, is to pre-train the network. Here the idea is to first have the network, look at natural images, and learn to identify objects such as penguins, or cats, or dogs, then use this network as a starting point for learning in medical imaging task by copying over the learned features. The network can then further be trained to look at chest X-rays and identify the presence and absence of diseases. The idea of this process, is that when we're learning our first task of identifying cats or dogs, the network will learn general features that will help it's learning on the medical task. An example of this, might be that the features that are useful to identify the edges on a penguin, are also useful for identifying edges on a lung, which are then helpful to identify certain diseases. Then when we transfer these features to our new network, the network can learn the new task of chest X-ray interpretation with a better starting point. This first step, is called pre-training, and the second step ,is called fine-tuning. It is generally understood that the early layers of the network, capture low-level image features that are broadly generalizable, while the later layers capture details that are more high-level or more specific to a task. So for instance, the early layer might learn about the edges of an object, and this might be useful for chest X-ray interpretation later. But the later layers, might learn how to identify the head of a penguin and may not be useful for chest X-ray interpretation. So when we fine-tune the network on chest X-rays, instead of fine-tuning all features we've transferred, we can freeze the features learned by the shallow layers and just fine-tune the deeper layers. In practice, two of the most common design choices are one, to fine-tune all of the layers, and two, only fine-tune the later or the last layer and not fine-tune the earlier layers. This approach of pre-training and fine-tuning, is also called transfer learning and is an effective way to tackle the small dataset size challenge.



Medical image datasets typically have 10 thousand to 100 thousand examples.

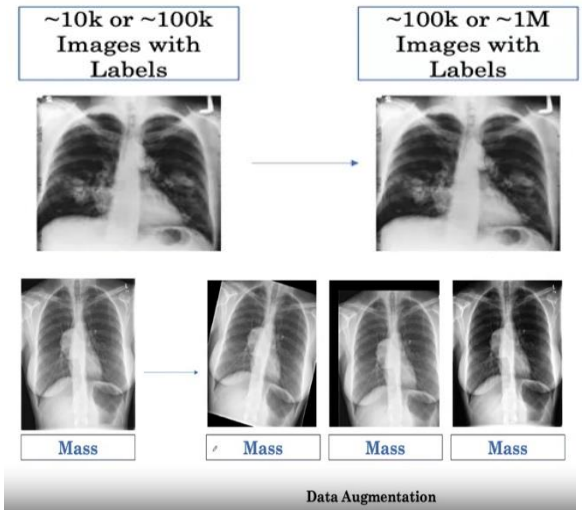


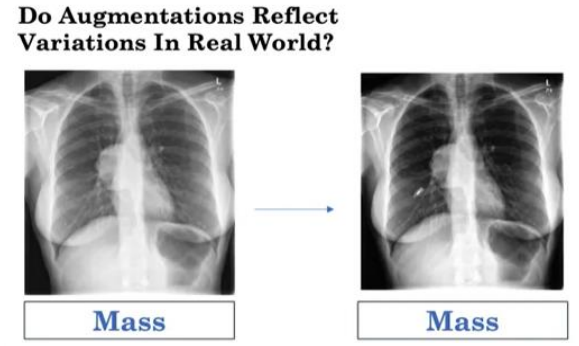


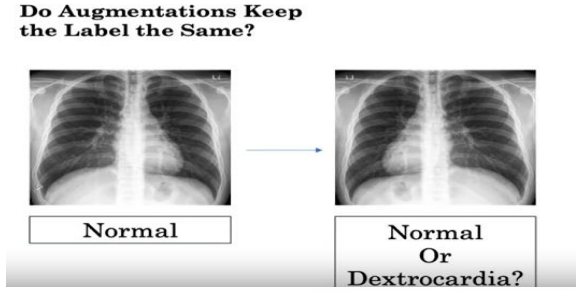


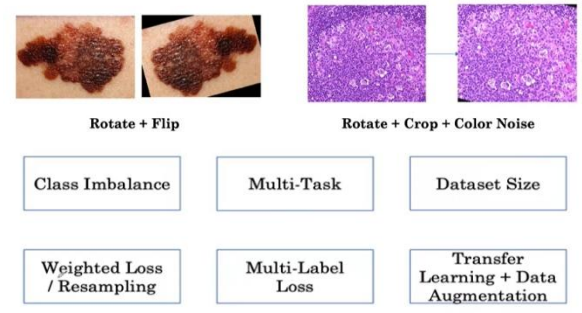
Generating More Samples

Let's talk about a second solution to the data set size challenge. The idea is to trick the network into thinking that we have more training examples that are disposal than we actually do. Right before we pass an X-ray image into the network, we can apply a transformation to it. We have several options here. We can rotate it and pass it in, or we can translate it sideways and pass it in, or we can zoom in, or we can change the brightness or contrast, or apply a combination of these transformation. This method is called data augmentation. In practice, there are two questions that drive the choice of transformations we pick. The first is whether we believe the transformation reflects variations that will help the model generalize the test set and therefore the real world scenarios. For instance, we might believe that we're likely to see variations in contrast in natural X-rays, so we might have a transformation that changes the contrast of an image. A second design choice is to verify that our transformation keeps the label the same. For instance, if we're laterally inverting a patient's X-ray, this means flipping the left over to the right and the right over to the left, then their heart would appear on the left-hand side of the image. This is the right of the body. However, the label of normal would no longer hold because this is actually a rare heart condition called dextrocardia in which your heart points to the right side of your chest instead of to the left side. **So this is not a transformation that preserves the label.** The key here is we want the network to learn to recognize images with these transformations that still have the same label, not a different one. Beyond X-rays, there are other useful data augmentation procedures for other tasks. Rotation and flipping are useful for algorithms trained to detect skin cancer, for instance. In histopathology, one large source of real world variation is different shades of pink and purple seen in these microscopic images. Color noise is often added and all this does is have slightly different shades of these pink and purple colors to help the network generalize. In addition, rotation and cropping are also useful data augmentation procedures for histopathology images. As a recap, we've looked at the weighted loss and the resampling methods to tackle the class imbalance problem. We've looked at the multi-label loss to allow the network to identify multiple diseases in a chest X-ray. We've also covered transfer learning and data augmentation procedures as ways to tackle the challenge of having a small training data set.



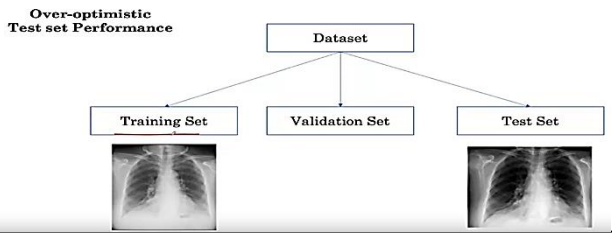


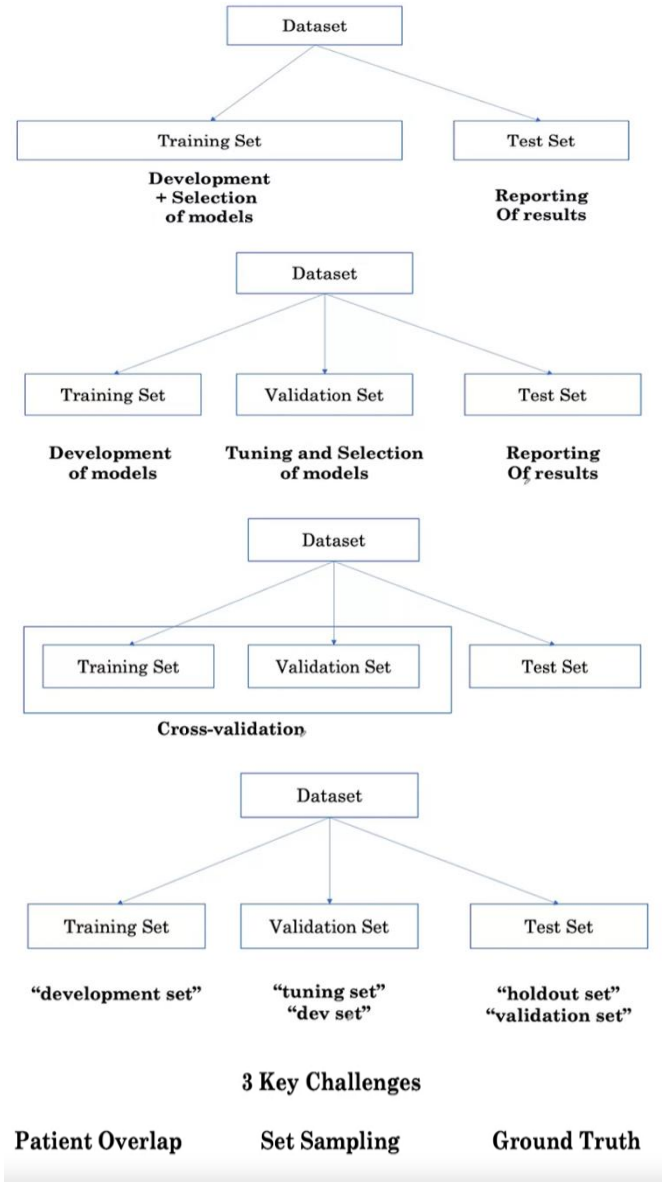


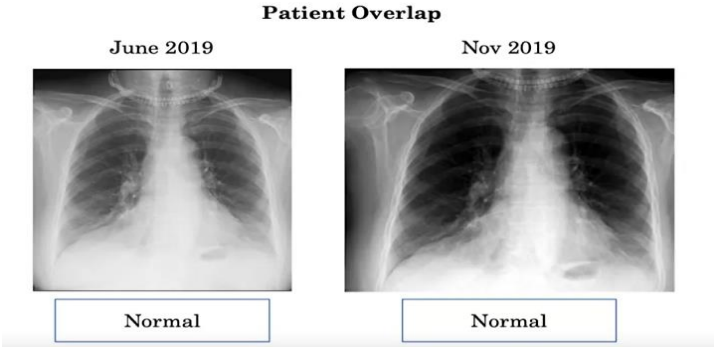


Model Testing

Now that you've seen how you would go about training a model for medical diagnosis, let's talk about how you would go about testing such a model. You will learn about the proper use of training, validation, and test sets. And about the need for strong ground truth in order to evaluate your models. When we apply machine learning to a dataset, we usually split it into a training and a test set. Our training set is used for the development and selection of models and our test set for the final reporting of our results. In reality, the training dataset is further split into a training set and a validation set, where the training set is used to learn a model and the validation set is used for hyper parameter tuning and giving an estimate of the model performance on the test set. Sometimes the split into a training and validation set is done multiple times in a method called cross validation to reduce variability in our estimate of the model performance. These sets also go by different names sometimes like validation can be called tuning or depth set, the training set can be called the development set, and the test set can go by holdout or even more confusing the validation set. We will stick to the terms training, validation, and test set for our purposes. We'll cover three challenges with building these sets in the context of medicine. The first challenge relates to how we make these test sets independent, the second relates to how we sample them ,and the third relates to how we set the ground truth. Let's cover the problem of patient overlap first. Let's say a patient comes in twice for an x-ray, once in June and once in November. Both times, they're wearing a necklace when they have their x-ray taken. One of their x-rays is sampled as part of the training set and the other as part of test. We train our deep learning model and find that it correctly predicts normal for the x-ray in the test set. The problem is that it's possible that the model actually memorized to output normal when it saw the patient with a necklace on. This is not hypothetical ,deep learning models can unintentionally memorize training data, and the model could memorize rare or unique training data aspects of the patient, such as the necklace, which could help it get the right answer when testing on the same patient. This would lead to an overly optimistic test set performance, where we would think that our model is better than it actually is.

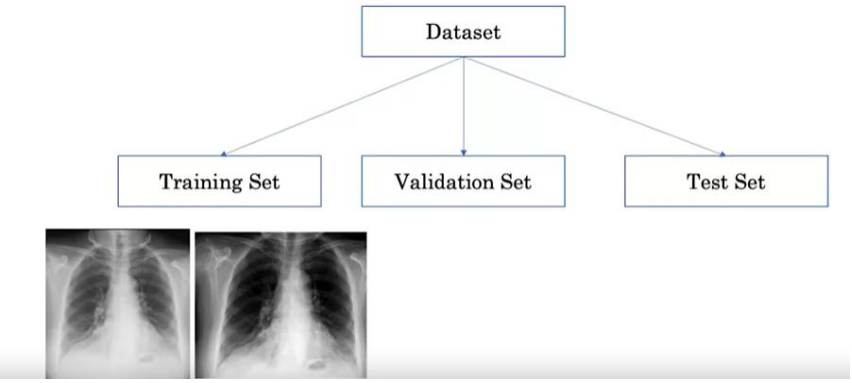






Splitting data by patient

To tackle this problem in our data set, we can make sure that a patient's X-rays only occur in one of the sets. Now, if the model memorizes the necklace on the patient, it does not help it achieve a higher performance on the test set because it doesn't see the same patient. Let's see this in action over all patients. When we split a data set in the traditional way, images are randomly assigned to one of the sets. Notice that this way we get X-rays that belong to the same patient in different sets. For instance, X-ray one belongs to patient 20 and is part of training. X-ray two also belongs to patient 20 and is part of validation. And X-ray zero also belongs to patient 20,which is part of test. This is the problem of patient overlap. Instead, when we split a dataset by patient, all of the X-rays that belong to the same patient are in the same set. For instance, X-ray ,X-ray two, and X-ray zero, which all belong to patient 20 ,are all part of training. Here seven, eight, and nine,which are all part of patient 32, are all in the validation set and we don't see 20 and 32 here in the test. This way, we can make sure there's no patient overlap between the sets. This covers our solution to the first challenge.



Split By Image

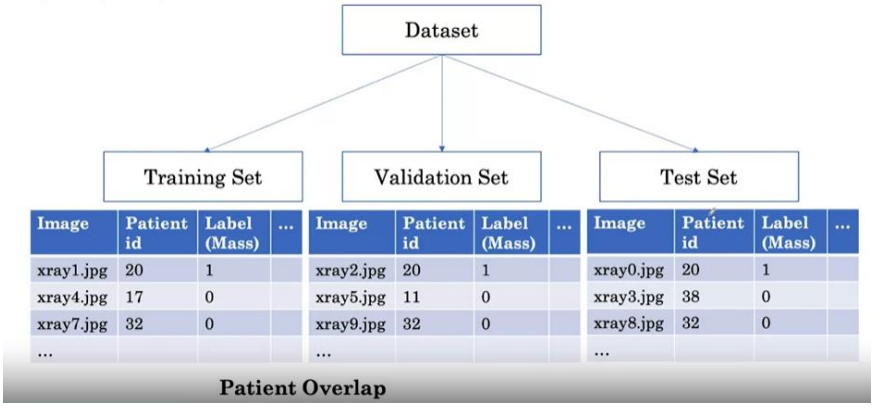


Image	Patient id	Label (Mass)	...
xray1.jpg	20	1	
xray4.jpg	17	0	
xray7.jpg	32	0	
...			

Image	Patient id	Label (Mass)	...
xray2.jpg	20	1	
xray5.jpg	11	0	
xray9.jpg	32	0	
...			

Image	Patient id	Label (Mass)	...
xray0.jpg	20	1	
xray3.jpg	38	0	
xray8.jpg	32	0	
...			

Patient Overlap

Split By Patient

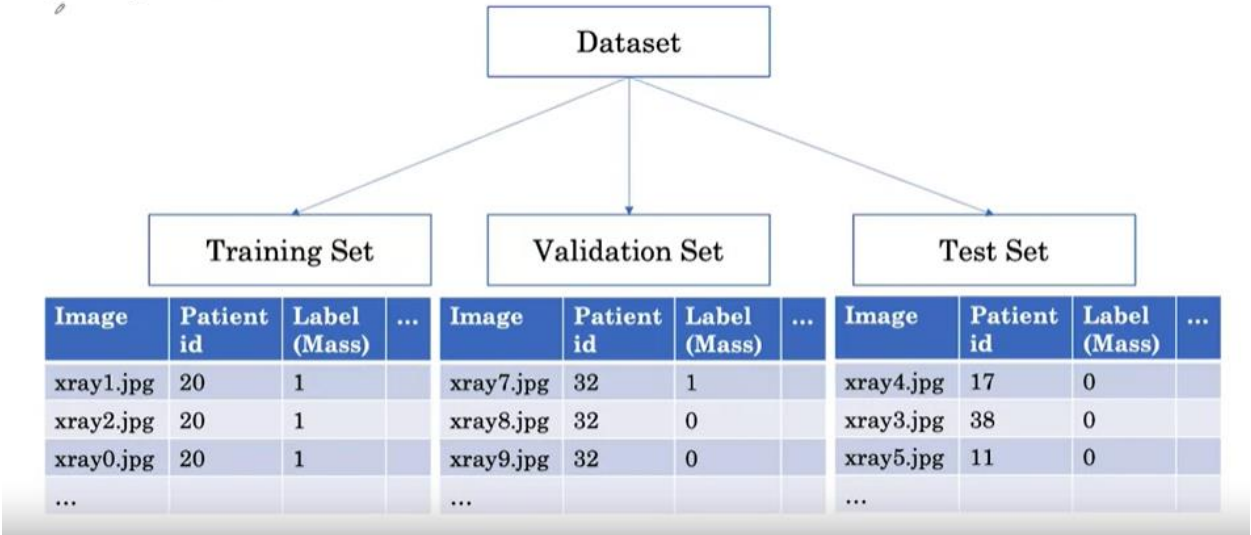


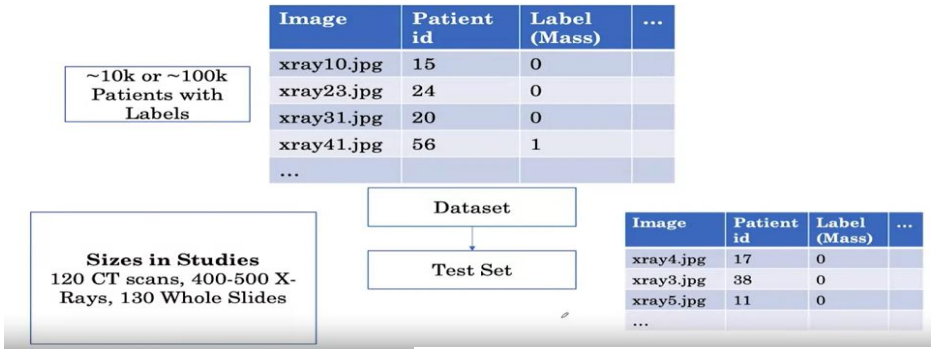
Image	Patient id	Label (Mass)	...
xray1.jpg	20	1	
xray2.jpg	20	1	
xray0.jpg	20	1	
...			

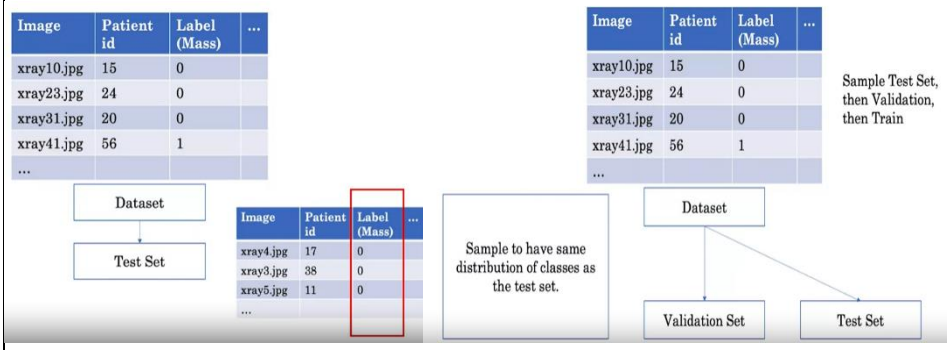
Image	Patient id	Label (Mass)	...
xray7.jpg	32	1	
xray8.jpg	32	0	
xray9.jpg	32	0	
...			

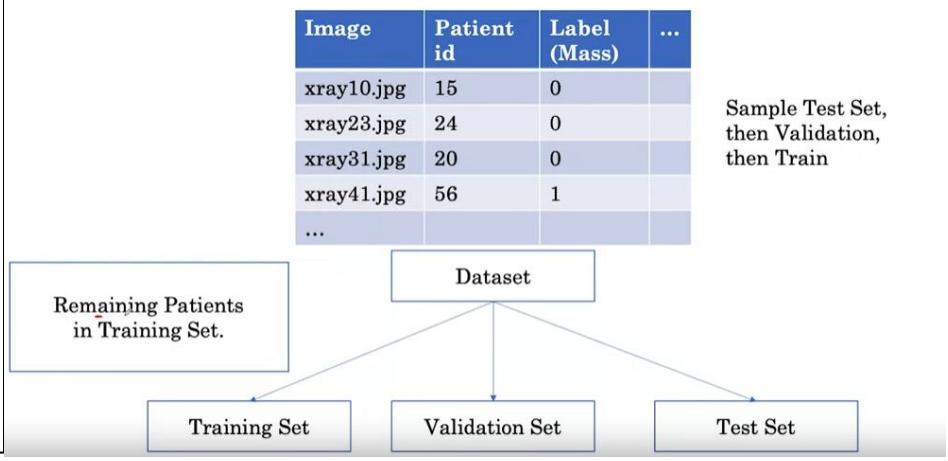
Image	Patient id	Label (Mass)	...
xray4.jpg	17	0	
xray3.jpg	38	0	
xray5.jpg	11	0	
...			

Sampling

Let's cover the second challenge of set sampling. Let's say we sampled a test set from the dataset. Sometimes, the size of the test set is a fraction of the full dataset like 10%. Other times, in human comparison studies, because the test set needs to be annotated by human readers, the bottleneck for the test set size is how many examples can readers be expected to read? Typically, test sets contain at least hundreds of examples in medical AI studies. The challenge with sampling a test set is that when we're randomly sampling a test set of hundreds of examples from the dataset, we might not sample any patients that actually have a disease. Here, we might not sample any examples where the label for mass is 1. Thus, we would have no way to actually test the performance of the model on these positive cases. This is especially a problem with medical data where we might already have a small dataset and not that many examples of each disease. One way that this is tackled when creating test sets is to sample a test set such that we have at least X% of examples of our minority class. Here, the minority class is simply the class for which we have few examples like here examples where mass is present. One common choice of X is 50%. So for sampling a dataset of 100 examples, we would have 50 examples of mass and 50 of not mass. This ensures that the study will have sufficient numbers to get a good estimate of the performance of the model both on non-disease and on disease examples. Once we sample the test set, typically, the validation set is sampled next before training. Because we want our validation set to reflect the distribution in the test set, typically, the same sampling strategy is used. We might decide to have once again 100 examples in the validation set of which 50 are mass and 50 are non-mass. Finally, the remaining patients can be included in the training set. Because the test and validation set have been artificially sampled to have a large fraction of mass examples, the training set will have a much smaller fraction of mass examples. You have seen that we can still train a model in the presence of imbalance data. So this covers our second challenge of set sampling.

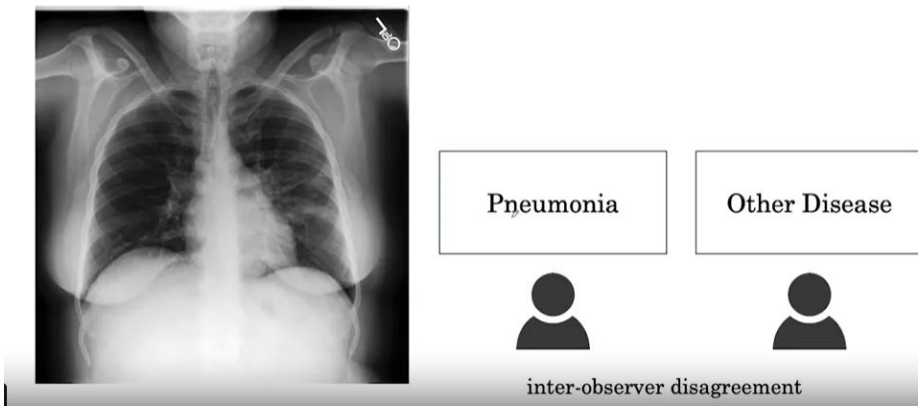


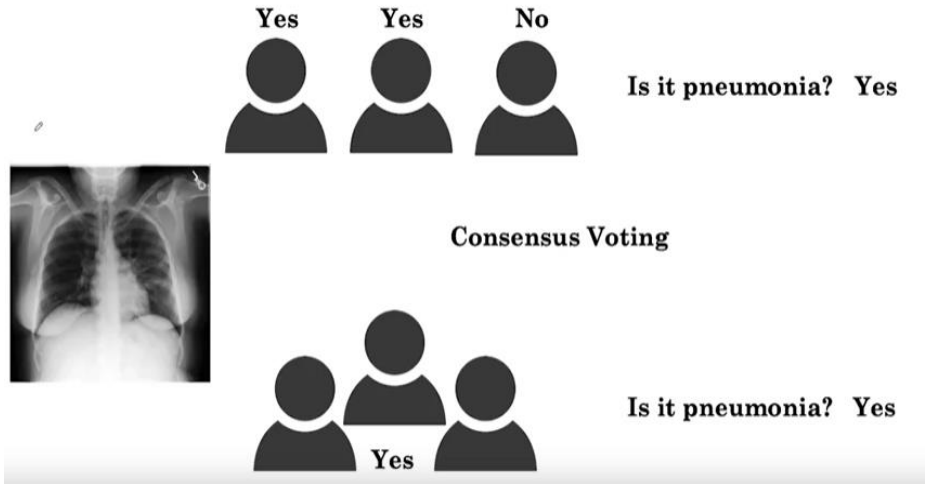




Ground Truth and Consensus Voting

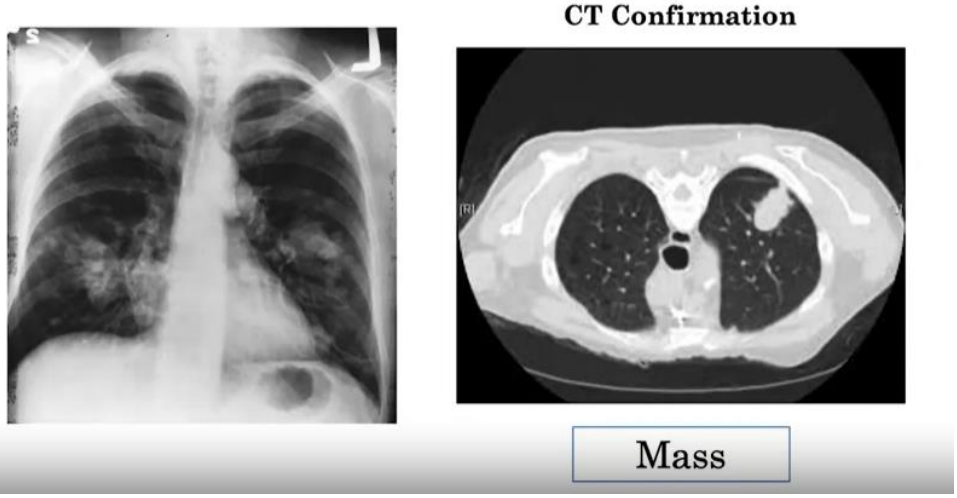
One major question in testing a model is how we determine the correct label for an example. The right label is more commonly called the ground truth in the context of machine learning or the reference standard in the context of medicine. On a chest X-ray, differentiating between some diseases might be complex. We might have one expert say this is pneumonia, another experts say it's another disease. This is called inter-observer disagreement, and is common in medical settings. The challenge here is how we can set the ground truth required for the evaluation of algorithms in the presence of inter-observer disagreement. So how do we determine the ground truth in the presence of inter-observer disagreement? We can use the consensus voting method. The idea behind consensus voting is to use a group of human experts to determine the ground truth. In one setting, we can have three radiologists look at a chest X-ray and each determine whether there is pneumonia present or not. If two out of the three say, yes ,then we would say the answer is yes. In general, the answer will be the majority vote of the three radiologists. Alternatively, we can have the three radiologists get into a room and discuss their interpretation until they reach a single decision, which can then be used as the ground truth.

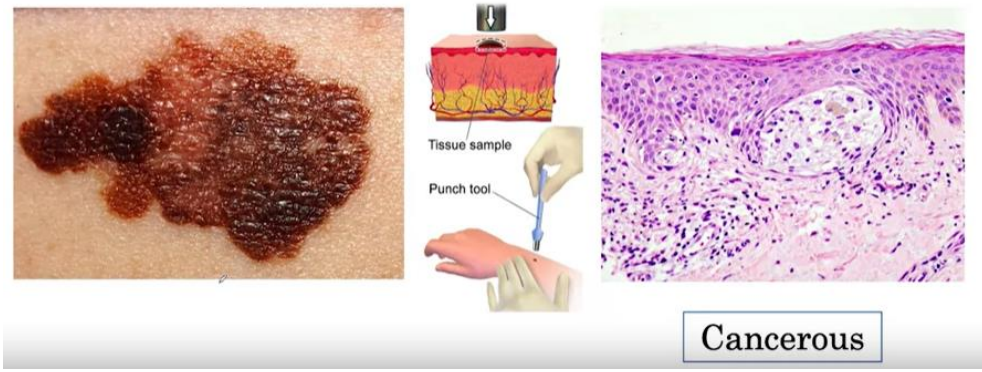




Additional Medical Testing

The second method is to use a more definitive test which provides additional information to set the ground truth. For example, to determine whether a patient has a mass using a chest x-ray, a more definitive test that can be performed is a CT scan. The CT scan shows the 3D structure of the potential abnormality, thus giving the radiologist more information. If a mass is confirmed on the CT, we can then assign that ground truth to the chest x-ray. In the dermatology study we saw earlier, the ground truth for the test set was determined by a skin lesion biopsy. This is a medical procedure in which a sample of the skin with the suspected cancer is removed and tested in a lab. The results of this test are then used to set the ground truth for the photo image. We've now looked at two methods to set the ground truth or reference standard, consensus voting and having a more definitive test. With the second method, the difficulty is that we might not have these additional tests available. Not everyone who gets a chest x-ray gets a CT scan, and not everyone who has a potentially suspicious skin lesion gets a biopsy. Thus, having a reliable ground truth with an existing data set often has to use the first method of getting a consensus ground truth on the existing data, which is the strategy that many medical AI studies use. Thus, we have covered our three challenges with algorithm testing and some solutions in the context of medicine. We've seen how we can split by patient in the creation of training validation and test sets, how we can sample a minimum percent of a minority class examples in the test set, and how we can use consensus voting or a more definitive test to set the ground truth. Congratulations on completing the first week. This week you've learned about examples of deep learning models for tasks across several fields of medicine and have learned about all the techniques needed to both train and test your own high-performance model for chest x-ray interpretation. You'll get to implement these ideas in the assignment and have your own chest x-ray interpretation model. In the next week ,you'll learn about ideas and methods to evaluate the models that you've built. See you then





3 Key Challenges

Patient Overlap	Set Sampling	Ground Truth
Split by Patient	**Minority class Sampling**	**Consensus voting / more definitive test**

Quiz- Week-1

Week 1 Quiz: Disease detection with computer vision
Practice Quiz • 30 min

✓ **Congratulations! You passed!**
TO PASS 80% or higher

Keep Learning

GRADE
100%

Week 1 Quiz: Disease detection with computer vision

TOTAL POINTS 10

1. Which of the following is not one of the key challenges for AI diagnostic algorithms that is discussed in the lecture? **1 / 1 point**

- ○ Multiple tasks
- ◉ Inflexible models
- ○ Dataset size
- ○ Class imbalance

✓ **Correct**
This was not discussed as one of the key challenges, but more complex models can be used to fit data. to avoid underfitting.

2. You find that your training set has 70% negative examples and 30% positive. Which of the following techniques **will NOT help** for training this imbalanced dataset? **1 / 1 point**

- ○ Reweighting examples in training loss
- ○ Oversampling positive examples
- ◉ Oversampling negative examples
- ○ Undersampling negative examples

✓ **Correct**
Given that the model is being trained on more negative examples, sampling even more negative samples will bias the model even more towards making a negative prediction.

3. What is the total loss from the normal (non-mass) examples in this example dataset? 1 / 1 point

Please use the natural logarithm in your calculation. When you use numpy.log, this is using the natural logarithm. Also, to get the total loss, please add up the losses from each 'normal' example.

Example	P(positive)
P1 Normal	0.6
P3 Normal	0.3
P5 Mass	0.4

- ○ 2.19
- ○ -0.4
- ◉ 1.27
- ○ 0.00

✓ **Correct**

Since these are negative examples, the losses will be $-log(1 - P(positive))$.

For P1, $-log(1 - 0.6) = 0.91$.

For P3 $-log(1 - 0.3) = 0.36$.

The sum is $0.91 + 0.36 = 1.27$.

4. What is the typical size of medical image dataset? 1 / 1 point

- ○ ~1 million or more images
- ○ ~ 1 hundred to 1 thousand images
- ◉ ~10 thousand to 100 thousand images
- ○ ~1 to 1 hundred images

✓ **Correct**

Most often datasets will range from 10,000 to 100,000 labeled images. Fewer than 1000 is typically too few to train, validate and test a classifier, and very few datasets will have millions of images due to the cost of labeling.

5. Which of the following data augmentations would be best to apply? 1 / 1 point

None of the above

Correct

This rotation is most likely to help. This is a realistic transformation. Also, it does not risk changing the label.

6. Which of the following are valid methods for determining ground truth? Choose all that apply. 1 / 1 point

- [x] Consensus voting from a board of doctors

Correct

Consensus is considered less reliable than biopsy verification. However, the limited availability of biopsy data means that consensus voting may still be the best (or only viable) option.

- [x] Confirmation by CT scan

Correct

A CT scan can provide an objective ground truth. Keep in mind that there are likely fewer data examples where patients have both the chest x-ray and an additional diagnostic test for the same disease.

- [x] Biopsy

Correct

Biopsy is definitely a valid method. Keep in mind that there are likely fewer data examples where patients have both the chest x-ray and an additional diagnostic test for the same disease.

7. In what order should the training, validation, and test sets be sampled? 1 / 1 point

- ○ Training, Validation, Test
- ○ Validation, Training, Test
- ○ Validation, Test, Training
- ◉ Test, Validation, Training

✓ **Correct**
First the test dataset should be sampled, then the validation set, then the training set. This is so that you can make sure you can adequately sample the test set, and then sample the validation set to match the distribution of labels in the test set.

8. Why is it bad to have the same patients in both training and test sets? 1 / 1 point

- ○ None of the above
- ○ Leaves too few images for the training set
- ◉ Overly optimistic test performance
- ○ Leaves too few images for the test set

✓ **Correct**
Having images from the same patient is bad because it has been shown that the model may learn patient-specific features that are not generalizable to other patients.

9. Let's say you have a relatively small training set (~5 thousand images). Which training strategy makes the most sense? 1 / 1 point

- ☑ Retraining the last layer of a pre-trained model

✓ **Correct**
By using a pre-trained model, you can make use of its ability to recognize lower level features, and then fine tune the last few layers using your dataset.

- ☐ Retraining the first layer of a pre-trained model
- ☐ Retraining all layers of a pre-trained model
- ☐ Train a model with randomly initialized weights

10. Now let's say you have a very large dataset (~1 million images). Which training strategies will make the most sense? **1 / 1 point**

- [x] Training a model with randomly initialized weights.

  ✓ **Correct**
  Given a very large dataset, you have the option of training a new model instead of using a pre-trained model.

- [ ] Retraining the last layer of a pretrained model
- [ ] Retraining the first layer of a pretrained model
- [x] Retraining all layers of a pretrained model

  ✓ **Correct**
  Given the large dataset, you have the option of training all layers of a pre-trained model. Using a pre-trained model may be faster than training a model from randomly initialized weights.