WEBVTT

NOTE duration:"00:28:35.3300000"

NOTE language:en-us

NOTE Confidence: 0.668213725090027

00:00:00.130 --> 00:00:05.970 Elegance.

NOTE Confidence: 0.886512994766235

00:00:06.610 --> 00:00:38.470 Driving image Ng dates biomarkers for non small cell lung cancer patients using deep learning thanks thanks. Everyone for the opportunity to speak today. I'm excited to talk a little bit about the work that my lab has been doing for the past year that I've been on faculty here. Some of the work today. Drives from the research as resident and as well as the work that we've done for the last year, specifically in the area of deep learning and machine learning here are my relevant disclosures.

NOTE Confidence: 0.926800429821014

00:00:40.830 --> 00:01:11.280 So our group is a very, very unique. Interdisciplinary lab, we have clinicians data scientist computer scientists mathematicians and engineers that spend pretty much the entire campus, including a large group of undergraduates. Graduate students as well as medical residents are interested specifically in the application of machine learning to clinical oncology and our major research areas are in the areas of machine learning algorithm development, improving the efficiency of machine learning techniques and novel.

NOTE Confidence: 0.898173391819

00:01:11.280 --> 00:01:43.610 Clinical applications of existing machine learning techniques and algorithms and area that we've been particularly interested in which is sort of the most emerging area with the machine learning is deep learning so before I kind of go into the work that we've done with respect to non small cell lung cancer. I'll kind of give a little bit of a framework for discussion because I think these words are often times turn around and then very frequently. They're kind of used as equivalent, but they're actually somewhat different and so deep learning is really just a subset of machine learning. It is a method in which we can kind of.

NOTE Confidence: 0.914132356643677

00:01:43.610 --> 00:02:14.680 It's a method of machine learning and I think that's open Top spoken about in the context of artificial intelligence and the reason for that is because it's able to learn features with raw data and so it's allowed to utilize information end to end format, which in many ways, is very, very useful and can be utilized when we're developing AI technologies deep learning is kind of the underpin the engine for a lot of things that we're using today, so virtual assistance self driving cars natural language processing anything that pretty much is being used by Facebook or Amazon or anything like that is.

NOTE Confidence: 0.906234443187714

00:02:14.680 --> 00:02:42.890 Primarily is primarily used in the context. I mean, if it's primarily using deep learning in some way, shape or form, and so our interest is particularly trying to apply deep learning in a clinical context and one of the big areas in which deep learning is particularly useful in the areas of image analysis before we go into sort of the project. I'll talk a little bit about deep learning algorithm is and some of this might be a little bit basic, but I think it's always useful just because I'm not sure exactly how many people here applied math background.

NOTE Confidence: 0.893617212772369

00:02:43.460 --> 00:03:14.270 So I think that common way in which we apply deep learning technology is in the area of classification. And so when we look right here. This is just a simple classifier in which we have inputs weights. Anna prediction so this can be a regression or any sort of technique. And when we're thinking about inputs that could be pixel intensities. Clinical variables lab values words in a sentence numeric features so any sort of input data that you want to put in the weights are really parameters in which you optimize the function that is modeling your data so they could be coefficients in your regression.

NOTE Confidence: 0.887762188911438

00:03:14.270 --> 00:03:45.800 Or they could be what will see here is the weights in the neural network after the weights are optimized they go through a function a function that can be in that I think that really is a special sauce and what determines what machine learning techniques for using in deep learning. There's a specific type of function. That's utilized and I think that's probably what makes it unique compared to the logistic regression afterwards. You look at a prediction you get a class output. It gives you a probability of your classes so for example, if we were developing a classification algorithm that was take input data. Let's say clinical variables and determine whether or not, that patient had cancer or not we would develop a class probability.

NOTE Confidence: 0.895642280578613

00:03:45.800 --> 00:03:52.860 And would have 90% cancer yes and 10% cancer know and from there, you can actually make a prediction depending on what you created your cut off.

NOTE Confidence: 0.891600787639618

00:03:53.800 --> 00:04:25.090 Now, when we're thinking about deep learning specifically compared to regression techniques. In many ways. It takes the most important part of a classification and sort of replicates that in an edit format and So what we're doing is picking the most important part of that singular classification and we're kind of creating a nested format with multiple classifications with multiple parameters. They're able to model a very heterogeneous

function a very, very difficult function in a nonlinear space. And so this is a neural network. It's the fundamental deep learning out a fundamental unit of a deep learning algorithm.

NOTE Confidence: 0.889710307121277

00:04:25.090 --> 00:04:52.990 It's in you know, I guess it stylistically one can think of the layers as represent different levels of abstraction, so for example, if your input was an image classification problem. The input would be the pixels. The first layer might be representing lines and edges that you see the second layer might represent shapes that are drawn from those lines. The 3rd layer might represent facial features and then the 4th layer would be a facial recognition algorithm so similarly each layer kind of represent a different level of higher levels of abstraction.

NOTE Confidence: 0.873288631439209

00:04:53.660 --> 00:05:24.980 And so neural networks, the one maybe disadvantage or maybe one thing that's important to note is they require their supervised learning algorithms. So they require a significant amount of training and there's a training phase, which requires label data so it's not an unsupervised algorithm. The supervised algorithm and the weights are initially considered to be random and it's very time intensive and requires high performance. GPU servers, which we've been very fortunate to have here at yeah, we have 2 of them, too, and video DJ S ones, which are basically Top of the line. So what happens is that as the input layer takes the function the weights are randomized and it promulgates.

NOTE Confidence: 0.87793505191803

00:05:24.980 --> 00:05:55.080 Through the hidden layers and then you get to an output layer and that output, you generate a predicted output right using random weights and that predicted output is just a vector of what your data represents based on your neural network architectures. There's a true label that we have right because this is label data that we're training and so we have a true representation of that vector and then we use. An objective function to calculate the loss, so how much are true label is from what are predicted label is and that tells us how off our network is in the most basic form and then what we do is we adjust our parameters.

NOTE Confidence: 0.896019220352173

00:05:55.080 --> 00:06:25.710 Backwards and so we go back and minimize that loss and So what we can see here is that we go back and we adjust our weights and so initially our losses very high in our accuracy is very low when we start training. Our model is basically using random parameters random coefficients. If you will, but then as we kind of go to our second sample. We continue this process and overtime. We see that our loss function decrease in our accuracy increases and So what we're actually trying to do is optimize. It's an optimization problem of that objective function and So what we were doing in some azatian problem.

NOTE Confidence: 0.880976378917694

00:06:25.710 --> 00:06:42.260 And fundamentally overtime with more training data and iteratively overtime your loss decreases in your accuracy increases and so your training is essentially complete when you minimize your loss to hopefully 0 or close to it, and then you keep the parameter weights from that moment.

NOTE Confidence: 0.885079741477966

00:06:45.520 --> 00:07:16.740 Perfect great and so we're thinking about neural networks. I think the one that I've explained to you is just a feedforward Neural Network, which is a simple way in which we kind of explain them to the most basic form of a neural network, but there's actually very complex architectures. They also fundamentally rely on the same sorts of mathematics, but they they're more so tailored for the data that they're using so feedforward networks are very much used for very low dimensional data. That's already feature engineered convolutional networks are very, very common.

NOTE Confidence: 0.864894807338715

00:07:16.740 --> 00:07:34.240 Image analysis is primarily what we use in the context of our lab bra. Colonel networks are very useful for evaluating sequences because they have? What they call memory. So it can understand sequences, and then residual networks are something more complicated. But they're very useful and introducing random error so the networks don't memorize your data set.

NOTE Confidence: 0.904813647270203

00:07:34.950 --> 00:08:07.340 And so these networks actually can pump very, very complex with multiple multiple neurons, but I think there's some benefits to this sort of algorithm compared to a standard regression for one, you can predict multiple outputs simultaneously. These outputs do not necessarily have to be related because we're looking at a nonlinear. We're trying to model a nonlinear function or I guess it could be linear, but it doesn't have to be linear like a regression. Like many forms of regression. We can look at different disparate outcomes that are related or unrelated and also we can train them all simultaneously so feasibly you can develop an algorithm which I'll show you in a little bit.

NOTE Confidence: 0.877842605113983

00:08:07.340 --> 00:08:09.530 It predicts multiple outcomes at the same time.

NOTE Confidence: 0.898087322711945

00:08:10.190 --> 00:08:42.180 The other thing that I think is important to appreciate is also that these parameters, so those lines represent each parameters and so, if you think about regression when I was when I was doing undergraduate work in applied mathematics. I remember they told me that, like having 10 variables in a regression technique was like a lot. I remember that was crazy.

And now we're thinking about each of those lines represents a parameter in our model and so our deep learning. That's actually have upwards of thousands of thousands of parameters in which we were trying to optimize and so it's really doing some fast computations. This is actually made very easy using linear algebra and some stuff that you can learn to variable calculus.

NOTE Confidence: 0.858445703983307

00:08:42.180 --> 00:08:51.150 In the interest of time, not going talk about the mathematics of it. But if you are interested in. I love talking about. That part probably more than some other stuff and I'm happy to kind of discuss the mathematics of Optimization.

NOTE Confidence: 0.869221389293671

00:08:52.080 --> 00:09:22.170 So we have 3 major research efforts within our lab, 1st is imagine based biomarkers using image analysis using a deep learning platform to identify outcomes. The 2nd is that we analyse audio data for patients in a collaboration with Amazon in order to in order to model patient reported outcomes and the last is that we analyze clinical trial information to match clinical trial. Using text data. Now, I think the only reason I'm showing the slides to highlight that it doesn't necessarily mean that just because.

NOTE Confidence: 0.878111124038696

00:09:22.170 --> 00:09:43.000 Image analysis is the most popular way in which people using deep learning that that's the only way I think there's so many different methods in which we can leverage this sort of technique across different types of heterogeneous data streams. The focus of today is going to be the imaging component and specifically warned about 2 things 1st is our early stage non small cell lung cancer biomarker and Secondly. We're gonna talk about this new initiative that we have which is called Digital Twinning.

NOTE Confidence: 0.856535077095032

00:09:44.360 --> 00:10:16.490 So imagine based biomarkers so a lot of the work here, then I'm going to presenting is actually being is because of an undergraduate named Justin do and our former chief resident was now damn far Benjamin Khan and so a lot of the work that they've done is actually presented here today and a lot of the work. They did my post. Doc is on the image side and so our thought is that cancer biomarkers have a number of problems wonder inaccurate. They potentially could be very expensive require additional test testing. They sometimes are invasive and they're not necessary personalized and then maybe don't capture all the heterogeneity of a tumor and so we created a deep learning platform, which can.

NOTE Confidence: 0.874152064323425

00:10:16.510 --> 00:10:47.010 Essentially generate predicted a personalized outcomes using pretreatment diagnostic imaging and we've shown that this to be effective, and non small cell lung cancer and as well, and lymph nodes so there's

just schematic kind of make it a little bit more clear so tumor. There's preacherman imaging. The images are run through the platform, which is a lot more complex than this. And then we can develop a personalized outcome prediction and the reason why it's so vague is because we really can predict multiple different outcomes from that, from that network and So what we have looked at is overall survival. Local failure to failure treatment side effects an pathologic findings and all of those.

NOTE Confidence: 0.897981882095337

00:10:47.010 --> 00:10:50.860 And actually be looked at in the same model just by making multiple outputs.

NOTE Confidence: 0.878324806690216

00:10:52.000 --> 00:11:23.950 And so the first first step was when we looked at early stage not spa salon cancer so these are patients with small small tumors non small cell lung cancer tumors that are treated with radiation therapy alone. Static body radiation therapy. And so when we're thinking about it as a clinician. There's multiple variables that I do that. I think about whenever we're trying to predict outcomes for early stage not supposed to lung cancer first were thinking about the images? What does a tumor look like they were also think about the radiation therapy plan, which is basically whether or not we're giving enough radiation dose to the tumor and we also think about these clinical variables, like Oh? How old is the patient they smoke?

NOTE Confidence: 0.881111860275269

00:11:23.950 --> 00:11:53.960 Are they you know? Do they have uh? Multiple comorbidities and things of that nature. And so when we're trying to kind of sort of parse out all those different types of data streams. If you will. It's very difficult for us to kind of objectively? Do that and we developed a model? Which allows us to kind of merge all the data streams through different types of Neural Networks in order to kind of create outcome predictions were able to initially looked at just predicting overall survival. Local failure renal failure and distant failure and we compare it to individual so individually looking at one data stream versus another data stream.

NOTE Confidence: 0.867005884647369

00:11:53.960 --> 00:12:24.910 And then we also compared to what we consider to be a traditional Mount model, which is like a Cox regression model. So basically predictions that would be generated from a personal hazard model and so the important thing to appreciate from this figure is that the green represents the deep learning the combined deep learning model and we can see that it hasn't increased discriminate. Tori value and so this is an RSC curve essentially is a way for us to test how how well our model actually orders orders samples in based on their risk, but it in the shortest way.

NOTE Confidence: 0.869323194026947

00:12:24.910 --> 00:12:41.680 Discriminate Tori measure and what we also see that it outperforms the Cox regression model and it also performs each of those individual data streams alone and So what we get when we generate from this is that we should be 1 combining all of the available data that we have to us and Secondly this deep, learning algorithm is potentially deep learning potentially useful way in which we can do that.

NOTE Confidence: 0.890256762504578

00:12:42.510 --> 00:13:03.120 And so just for those who are interested in Kaplan Meier curve based on our predictions of overall survival. But we can see is that if we look at the ones who are predicted to have predicted to have death versus the ones who are not predicted death and we track them out over years and years overtime in a more traditional Kappa Mayer format. We see that there's a statistically significant difference and overall survival as well as disease specific survival.

NOTE Confidence: 0.893322348594666

00:13:04.490 --> 00:13:35.540 Now, one big sort of a drawback of deep learning models is this idea that there's so many parameters and they're not really kind of linked to coefficients like you know variables like age and and brakes and things like that, so how do we know what? The network looking at? How do we know exactly what this model is doing. There's a lot of ways in which we can kind of Fact Check like regression based models by looking at sort of at the predictor makes sense. But I think there's a big concern about the ability for us to interpret deep learning models and so we've worked with some.

NOTE Confidence: 0.88606333732605

00:13:35.540 --> 00:14:06.670 With some collaborators at the time who are Georgia Tech score now acquired by Facebook as a lab was acquired by Facebook but they are looking at they did. They developed a technique called gradient waited class activation Maps with essentially is doing is identifying areas of the pixel image at different layers of our network and So what we can think about is that if it's highlighting anatomically relevant areas as it goes down the network. That means that it's looking at the places that have position would look so for example, if this model was looking at if it was highlighting this patients are more their spine.

NOTE Confidence: 0.898407638072968

00:14:06.670 --> 00:14:15.940 And we know that we're just getting lucky that we were predicting correctly. But we can see when we're predicting local failure that we actually are looking as the network is learning an isolating the tumor.

NOTE Confidence: 0.901514530181885

00:14:17.190 --> 00:14:48.590 Now the interesting part about this is actually so this panel is a little bit complicated. But I'll explain so this is the original image. This is our local failure prediction when we're looking at regional failure. This

is the final prediction. And so when we're looking at regional figure out we see that the network is actually looking at the media. Steinem, which to us makes sense that there's some relevance to the model is actually looking at what we would look at it as as clinicians. This is the overall survival plot. I've heard different theories on it. I think some people think it's because it's looking at the heart and things of that nature that it's potentially.

NOTE Confidence: 0.860941886901855

00:14:48.590 --> 00:15:01.720 That's that's what's highlighting and then those organs are important. I think that actually what it probably means is that overall survival. Not multifactorial that really the pictures only a piece of the puzzle, but but either way.

NOTE Confidence: 0.869481325149536

00:15:02.590 --> 00:15:33.100 And so now we've collaborate we build a collaboration with with four other cancer centers. OHSU University, Colorado University. Braska and Jefferson to externally validate this algorithm because one thing that we always worry about is whether not this is actually generalizable across different patient populations and so some of the work that Justin's been doing so, we have currently the Nebraska cohort in the Jefferson court and what we see is in the Nebraska cohort that are combined model. It does not perform as well as our model here trained at Yale, but it still performs fairly well and.

NOTE Confidence: 0.850459694862366

00:15:33.100 --> 00:15:58.330 And it still out performs a Cockrel Hazard model and I think the big thing about the regression techniques. That's always been thought about is that they generalize quite well, meaning that they can work very well on unseen data what we found is actually at the Cox works will have trained at Yale actually works significantly worse. Then then then if you compare the differences between the combined deep learning model and the external validation set and that's actually BeenVerified both in the Braska Cohort and the Jefferson Court.

NOTE Confidence: 0.889212906360626

00:15:59.100 --> 00:16:27.860 We've also started looking at side effects side effects that are little more rare makes me think this could be potentially useful in predicting something else. That's a little bit more useful in the clinic at that moment. And so when we're looking at Grade 3 pneumonitis, which is a common side effect that we worry bout for radiation therapy or a great 3. Asafa Gitis or we find is again that the deep learning model, which combines all the data streams is actually particularly better than any of the other models, including a generalized linear model or Ardo symmetric parameters, which is in green of what we actually been taught to use as our cutoff clinically.

NOTE Confidence: 0.889416694641113

00:16:29.850 --> 00:17:00.520 Now it isn't only applied to lung tumors. We've actually seen this is a very, very effective way in which we can evaluate lymph nodes as well. And so this is some work. That was done by Benjamin Khan, who was our former chief resident who is now at Dana Farber, which was essentially identifying lymph nodes that one have cancer and Secondly have cancer that spread outside of the lymph node. This is particularly important in the context of head and neck cancer patients because it is a way in which we determine whether not the patient needs chemotherapy or not, and So what he was able to determine first on internally validated is that is that a deep learning Model Outperforms.

NOTE Confidence: 0.881428122520447

00:17:00.520 --> 00:17:31.310 A logistic regression as well as a more random forest model, which is another common machine learning technique and Secondly when we do an external validation from patients at Mount Sinai. Mount Sinai Medical School as well as in the cancer genome. Atlas the TCI a cancer imaging archive. An we also compared it to radiologists neuro radiologist trained here. One is that our models continues to have good performance, but also is pretty much at the same level of the radiologist.

NOTE Confidence: 0.903409421443939

00:17:31.310 --> 00:17:57.320 The Blue is actually a very interesting, so the blue is actually indicating when we gave the radiologist. The deep learning algorithm. Whether or not, their performance improved and what we found is that one of the radiologist. His performance improved very well and then one was slightly less. They didn't change a lot of their predictions. But I think this is an area of ongoing interest where whether not we can kind of combine human intelligence with the intelligence of machines to sort of improve production overall.

NOTE Confidence: 0.858929693698883

00:17:58.290 --> 00:18:28.300 So what are the future directions with respect to are not small so long cancer 1st and foremost. I think that we want to advance this in the locally advanced setting and with our collaborators. We have developed a data sharing relationship which has been very fruitful. This is just what we call TCT stochastic neighbor, embedding Plaza nonlinear form of dimensionality reduction away for you to look at the clustering algorithm sorts of imaging features and what we can see is that there's a imaging features between partial responders incomplete responses with with in Stage 3 disease.

NOTE Confidence: 0.8878054022789

00:18:28.300 --> 00:18:58.310 And even within patients who have driver mutations. There's a complete response or a partial response up complete versus partial response. There's some sort of imaging features and so this tells us that the signal within the images should be able to help us identify which patients are responders and which ones are not Secondly were interested in looking at our grad camps for recurrence prediction. And so this is a figure that highlights.

This is a patient who had treatment. But the grad camp showed this area of highlight highlighted area that we didn't think was actually that important. And then a year later. We saw that that was actually where the tumor.

NOTE Confidence: 0.892424762248993

00:18:58.310 --> 00:19:10.080 Occured and so, if we were able to theoretically look at this and appreciate that this was actually an area of risk and we could have potentially done something different potentially treated this little spot here, which actually I guess turned out to be cancer.

NOTE Confidence: 0.84659481048584

00:19:10.790 --> 00:19:35.660 And the last thing is, we want to integrate pet see T into into the model just because we've been noticing is that pet actually has a very, very good indicator deep learning analysis of pet images sort of gives us a good signal. So we can see the pet imaging of this pet. This patient who survived for very long time. The painting is very crisp and this was a lot more erratic and that sort of Radisys am I mean that sort of erratic pet signal on the deep learning models suggest the potentially have a lower life expectancy.

NOTE Confidence: 0.879769861698151

00:19:36.290 --> 00:20:09.020 OK, so in the last 5 minutes. I'm going to talk about this. This ongoing project, which is pretty interesting partially because two of the strongest undergrads I've ever met in my tire life are working on this. Gabriel sushi and who are who are working on this idea of digital Twins and so basically the idea is that can we leverage deep learning to identify patients are previously looking similar to us how digital Twins actually? How do their outcomes differ from ours and so similarly we employed are diploid platform with the techniques that are oftentimes used for facial recognition software.

NOTE Confidence: 0.871518433094025

00:20:09.020 --> 00:20:40.340 To find patients that are similar and so we have 2 things one is digital Twins, which is the best match among our cohort and 2nd is a digital family, which is kind of an average of the 10 the 10 Best 10 similar to you and so this is just a schematic. Similarly, we take the freemen images. We identify the digital twin in it, using an algorithm. I'll show you in a little bit and then we kind of see how their outcomes. Look and so this is the idea of a digital Twins. We take 3. Three of our deep learning models and we have 3. We have the parameters are shared between the 3 but we actually have to minimize.

NOTE Confidence: 0.900915026664734

00:20:40.360 --> 00:21:10.550 The distance between the similar images, so the vectors that are created from the similar ones. They need to be close together and the ones are very different have to be very far apart so there's an anchor image, which is basically an image that's true. This is what you want to look at that are cancer patients. There's a positive image, which is that image. And

it's in a different different orientation. And So what we often times have been doing is looking at them on a different scan that they've had or an alternative slice so it's kind of a different perspective of that image. And then there's a negative image, which is basically an image that is the wrong one, it's a different image so.

NOTE Confidence: 0.850642085075378

00:21:10.550 --> 00:21:32.900 I change the slide, which is a little too politically charged. Yes, these undergrads you gotta be careful. And so and so the negative images? Are is a wrong in different image and so we used for the wrong images or normal lung and then non cancer findings and then not benign nodules and so we did in a trained format to make sure that it was learning the easy stuff, 1st and then the hard stuff later on.

NOTE Confidence: 0.899291217327118

00:21:33.730 --> 00:21:58.820 And then I think it's important to understand what the objective function isn't this fun, beating this, where the true innovation is what we're trying to do is so this is your so this is a loss function. This is the function that you're trying to optimize which is taking your actual image your positive image in your negative image and you want to maximize you want to minimize the distance is that is that exist between your actual image and the positive image and you want to maximize the distances between your negative image and your image.

NOTE Confidence: 0.879184067249298

00:21:59.730 --> 00:22:30.110 And so I think it's important to look at sort of whether not they actually look like Twins like Are they looking the same and what we found is actually they do not so often times their gender is different than what their digital twin is there races, oftentimes different. The things that did seem to be kind of sort of correlated were Histology and driver mutations as well as age is something that we're going to continue to look at but interesting nonetheless. So I wish I could show you guys. The photos of the people just to kind of show you how different some of these digital Twins look compared to some of their I guess.

NOTE Confidence: 0.864463865756989

00:22:30.110 --> 00:22:37.330 Siblings but but it's pretty stark and I think it kind of helps us understand that may potentially there something else underlying that.

NOTE Confidence: 0.864671766757965

00:22:38.090 --> 00:23:08.190 And So what we did is we actually took the digital twin and the digital family, we looked at their overall survival estimate based off of the average of those ones, and then we subtract that to the actual estimate so the actual overall survival. And So what you can appreciate from this and we compared to what would be predicted from matching algorithm stage or

Cox Hazard algorithm and So what we can see is on the right is when we would underestimate potentially how much we would underestimate how much survivor would be and then the left is it that we would.

NOTE Confidence: 0.886435627937317

00:23:08.190 --> 00:23:41.320 Overestimate so we basically too optimistic and we found is the digital Twins, especially the digital family is actually associated with a very, very good actual estimate of your prognosis versus the stage, which is actually pretty erratic and I think that we have a wide variation in terms of how we're predicting patients, which is typically what we would kind of look at, I know in my clinical practice. That's pretty much what I would kind of look at is the mention of people and they tend to be actually sort of underestimating the survival of the actual patients. I think it's also important to note that the matching algorithms, which are based off of Euclidean distance and it's more based off of.

NOTE Confidence: 0.895329892635345

00:23:41.320 --> 00:23:43.690 Demographic features are also somewhat effective.

NOTE Confidence: 0.832258641719818

00:23:45.070 --> 00:24:15.830 Great so just to close deep learning methods appear to have utility and diagnostic image analysis. Deploring models represent a unique way for us to merge multiple streams of data and we continue to do external validation just to make sure that we can appreciate that these are prognostic. Even on datasets that were not used here at Yale and I just want to moment to think all the members of my lab that apartment therapy radiology and the Department of statistical data science and as well. The Center for outcomes research in valuation. Riding my postdoc and you guys can follow us on our GitHub. We have all the code available for you guys.

NOTE Confidence: 0.918475925922394

00:24:21.940 --> 00:24:24.510 Some questions there's a great talk.

NOTE Confidence: 0.910207390785217

00:24:25.900 --> 00:24:56.680 Well, I'll start you, um, walked in a little bit into one of the general questions that come up with these approaches is the difficulty in knowing what's going on in that black box and I know, people are developing forensic methods basically for going back in and you talked about some of those what are the long term prospects have really being able to get a handle these algorithms are doing image analysis. I think that we're pretty close. I think so. The Great way to class activation Maps. I think are very, very effective.

NOTE Confidence: 0.897865951061249

00:24:56.680 --> 00:25:27.330 I think that it is uhm, but the most important thing is that I don't know if we can associate the next step. Where we can actually look at an image see what parts of the image are important and then

know that that is correlated with some sort of biological next step for translation. I think the more complicated pieces unpacking the black box for non image in data so for the audio study that we're doing, it's very difficult for us to sort of unpack why certain things are getting predicted one way versus another and I think that's going to be more difficult. The images are nice because you can kind of look at what parts of the pixel pixels were most important in generating.

NOTE Confidence: 0.846199691295624

00:25:27.330 --> 00:25:32.270 In output and so I think that it's more difficult for the non than an image Ng sort of data.

NOTE Confidence: 0.928927659988403

00:25:35.960 --> 00:26:06.250 OK, I've got another question so recently in the New York Times. There was an article about the potential development of 1/2 and half not situation in which only few centralized industrial groups. They had the collections of data sets and we had the computational power to really push the ball forward on machine learning and yeah. That's true. I think that's the reason why a lot of the groups don't necessarily want are not necessary.

NOTE Confidence: 0.883117735385895

00:26:06.250 --> 00:26:38.380 Want them to but they're not really necessarily invested in academic groups. If you have a deep learning model. That's very basic if you have the most data. It also model correctly and so I think that that's why it's important for us to sort of R groups and a lot of time investigating different unique methods in which we can kind of compressed data streams to make them smaller. I don't feel like computationally were particularly behind compared to other groups. I mean, especially Yale. Specifically, we have basically the two Top of the line computers and their relatively available for everybody. Just just in case anyone is interested. I don't feel necessarily that's going to, I think the bigger issue is that?

NOTE Confidence: 0.895123600959778

00:26:38.380 --> 00:27:09.980 Whoever gets controlled the data is going to be the ones who are who are going to be the leaders of the pack. and I think it's actually not going to necessarily industrial group. I think it's more of other other international groups that have Open Access. So we have a collaboration with the musical partner, Shenzhen China and they have a complete electronic medical record in order to get care, there, you have to provide your data for their data. Their data group and so like the ability for them to have an electronic medical record. They have like almost thousands and thousands of images every day that are uploaded. It's pretty impressive and so I think that's the more worried that we would have is not necessarily.

NOTE Confidence: 0.862124800682068

00:27:10.380 --> 00:27:14.050 Industrious more other governments that are maybe less regulated about data control.

NOTE Confidence: 0.947343349456787

00:27:16.840 --> 00:27:18.440 Any questions from the audience.

NOTE Confidence: 0.766785085201263

00:27:19.180 --> 00:27:21.270 And he started out.

NOTE Confidence: 0.90455287694931

00:27:22.880 --> 00:27:31.750 A good question so for the early stage non small cell lung cancer patients. We traded on 500 but we use data augmentation with bootstrapping to make it about a 5000 patients sample.

NOTE Confidence: 0.831633746623993

00:27:34.210 --> 00:27:36.170 Uh you mean for the.

NOTE Confidence: 0.848288655281067

00:27:38.300 --> 00:27:42.150 Yeah, so though the snow validations for each of those, 200 patients.

NOTE Confidence: 0.884891510009766

00:27:42.850 --> 00:27:48.990 Yeah, so those are completely test blind tested no parameters have changed their not trained at all on that it's just a test set.

NOTE Confidence: 0.879180073738098

00:27:53.030 --> 00:28:25.320 OK, well, one last question for Maine do you have any way of knowing from the structure of the data that you're working with? What scale of data set size. You'll need to do the analysis. You're trying to do with a good question. So the good so there's no way to calculate like power that you would need for it. I think that a lot of times people do learning curve analysis. That's what we do in our group and so you take batches of training data smaller batches of training data and you see whether or not your error function.

NOTE Confidence: 0.842378675937653

00:28:25.320 --> 00:28:35.330 Um optimizes on smaller sets and then you could have increased that overtime as long as the curve is asking to did then. Then you know that you're at the right number for see T scan.