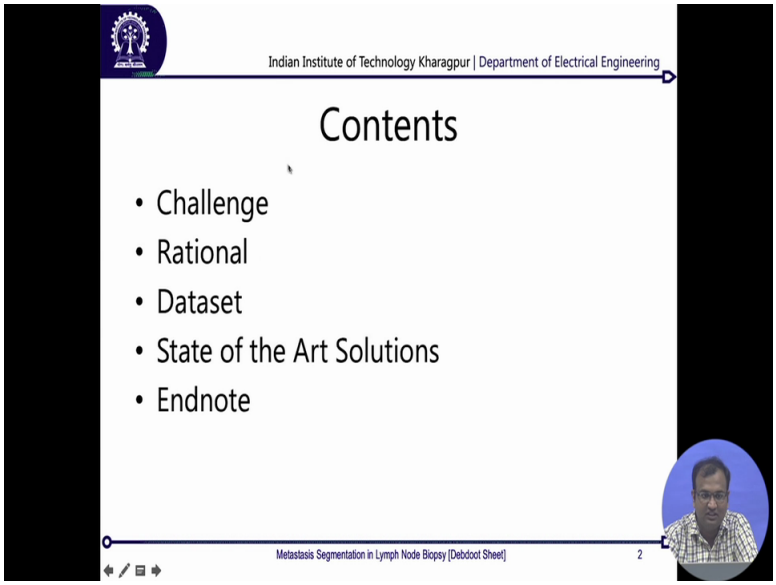
Professor Debdoot Sheet
Department of Electrical Engineering
Indian Institute of Technology Kharagpur
Module 4
Lecture No 20
Metastatic Region Segmentation in Lymph Node Biopsy

Welcome and so this is our last lecture for the season and today I would be discussing about the final one which is left behind and till now we have been discussing quite lot about radiological modality. So we started down initially we did start with an optical modality for fundus imaging on analyzing your blood vessels in the eye and then we went over to another interesting vascular structure which is present in radiology and that was about vessels within lung CTs. And then we did discussed about MR Lesion segmentation and following that we did have ultrasound tissue characterization. Now if you look at this whole pathway we started with some optical modality which had a bit of microscopy and which was more of a like in vivo application thing. So the human is alive and on them and all of these imaging were going on over there eventually.

So ultrasound was where we did discuss about computational techniques being brought down in order to favor something which is called as (()) (1:22) histology. And now what I would be discussing is a big data deluge which we are facing today. And this is about Biopsy and analyzing of biopsy. So one specific problem which I take down is from one of the challenges called as Camelyon and this one is about Metastatic Region Segmentation in Lymph Node Biopsy. So I will be coming down eventually one by one as to what is the disease pathology as have been discussing with the other application areas and cumulatively beyond that I would also be making you aware about what the big data deluge in this particular problem is and then what we can do in order to solve it.

(Refer Slide Time: 2:09)



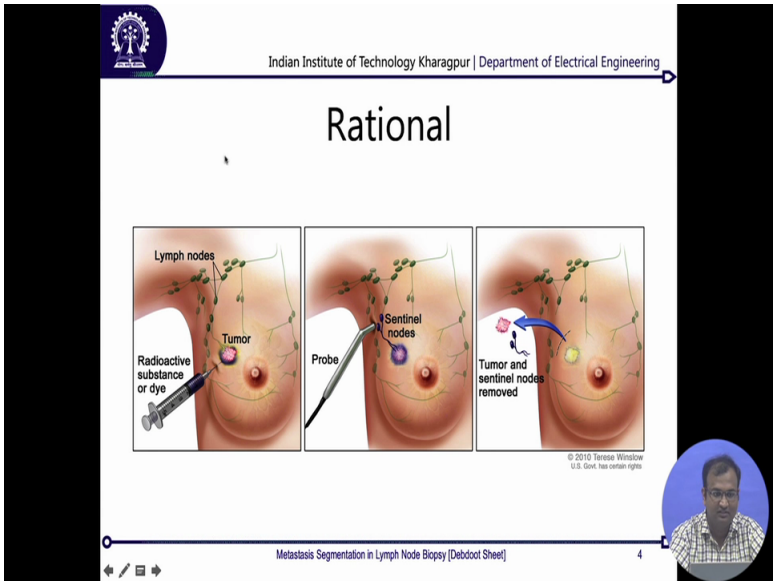
So overall the organization is arranged something like we have the challenges first and then I would be once the challenge is introduced to you then I would be speaking about the Rational and introduction to the problem which we are trying to deal in hand until we are quite sensitive about the pathology which we are dealing with and the actual biological problem to be solved on the clinical side, it would really be hard to understand as to what Rational was there when undertaking this sort of a problem statement solution.

Then I would be discussing about the dataset and also showing you some of the comparisons and at length trying to discuss at least one of the solutions. So I would be discussing about the winning team which had won this particular challenge last year at ISBI and what method they were using and then coming down to a endnote as to what more you can read and where you can read it out from.

(Refer Slide Time: 3:08)

So the challenge is what is called as Camelyon and if you look over here then there must banner over here says that it is a ISBI challenge on cancer metastasis detection in lymph nodes. So you can definitely go on this URL and this year it is Camelyon17.gand-challenge.org, so the new version is already there and up it is running. So the time when you are listening to this lectures, it is still accepting submissions hopefully around that time.

(Refer Slide Time: 3:35)



So what this whole challenge is about let us come down to basic about on the clinics. So there is a main problem, a major cause of cancers in women and one of the second highest leading cause of cancer related deaths and that is called as breast cancer. Now although a lot of women get affected by what is called as a breast tumor but then not all tumors are cancerous.

Now in case say a tumor ends up being cancerous then what happens is that this so a major appearance or a major attribute associated with a mass of tissue which is cancerous is that it starts proliferating or spreading out.

Now once it starts spreading out everywhere, then these cancerous cells they keep on spreading and affecting other organs as well and once they sort of start going down to other organs and colonize over there, so they release certain sort of chemicals which change the local ambience around different cells present over there and the other cells also start becoming a cancerous. So it is sort of a mob affect so if a mob is quite unruly and then it keeps on spreading everywhere then these few unruly people can actually seek down creation of much larger mobs and that would disturb the whole community in a standard way and over here even cancers lead to that.

So whenever you hear about a problem called as multiple organ failures due to cancer metastasis that basically means that these cancers were spreading down to multiple organs within the body and each organ kept on becoming unhealthy and cancerous and they kept on failing, which means they stopping their normal course of whatever actions they were having.

Now so here generally what is done for over here is that in breast, so if you look at the breast anatomy then around the breast there are lot of lymph nodes which are present and so in our body we have two sort of circulation systems. So one of the circulatory systems is the blood circulatory system on which you have oxygen and glucose and all of these transfers going on and other one is the lymphatic circulatory system which we have in our body.

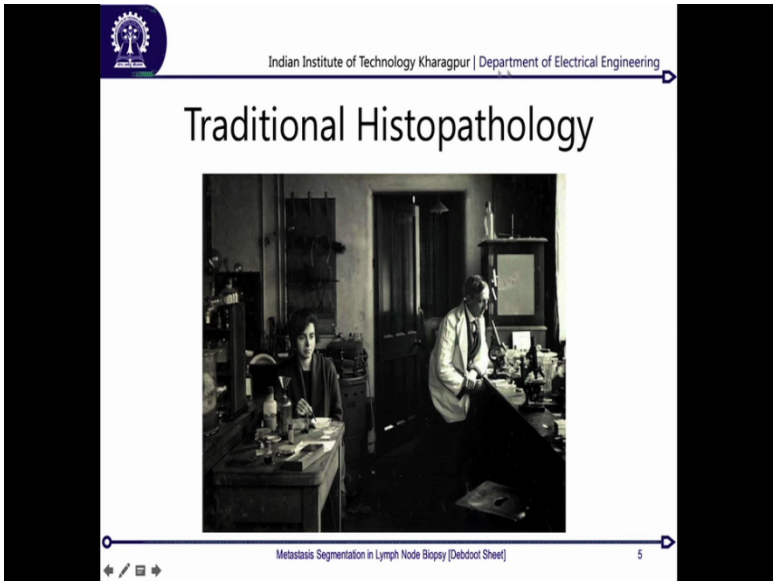
Now these lymphs so as in our blood circulatory system you basically have the heart which is keeping on pumping this constantly, but in lymphs you do not have that, so it is a much slow moving system and you need to secrete out lymph over there and that get secreted out from the lymph nodes. Now these lymph nodes are quite interestingly located around this groin area and if there is a tumor over here there is a high chance this being a soft tissue they actually spread.

Now once they end up getting spread up and affect down these lymphatic nodes over here then all of them are going to spread down via this lymphatic node network and that is going to create a major problem for us. So in general what is done is that when say there is a tumor and a needle is inserted in order to pull out some aspiration or take a small Biopsy via a cold needle biopsy method, then along with that the standard procedure is that after the treatment a

pathologist also draws out a sample from the lymph nodes itself. And from this lymph node they draw out samples in order to understand whether there is some sort of a metastasis or cancer spreading out in the lymph node over there.

Now for that so what needs to be done is as a stage over here so there is a mass of tissue from the lymph node taken out and this lymph node is sort of a very small structure it is about say centimeter or 2 centimeters in its diameter and that is what has to be taken out. And now you have to analyze very carefully whether all the cells present over there and it is pretty densely packed which has the whether all the cells are perfect or there are some traces which are cancerous in nature.

(Refer Slide Time: 7:29)



Now the traditional practice is what is called as histopathology or like in standard term “histo” is basically tissue is a name for tissues and pathology understands the disease pathway, so it is disease pathway analysis and understanding within tissues itself. So the old technique over here was when say there is a technician who takes out these, so there would be a Doctor who pulls out the mass of tissue and there is a technician who prepares them into slides.

So, as we had done learnt on the microscopic site that you will have to do some sort of a processing in order to do that we had also seen down the histology slides for oral Biopsies, so over here you will have similar kind of slides being prepared which are for lymph node Biopsies. And then there is a pathologist who is a trained clinician, he is a medical doctor

who is going to look down through different regions of the tissue mass over there and then find out how it is.

So this is a traditional practice which is till date followed in major parts of India and major hospitals in chains. But the problem so the major issue over here you see is that it is a manual procedure, so you will have to take a slide put it down and then observe with the microscope and the slides they are not everlasting, so you cannot preserve them forever. So the movement you see it and then within a few days they might get corrupted and you have to discard the whole thing.

(Refer Slide Time: 8:50)

So in order to get rid of that and with the modern practices coming down where we would like to preserve medical records, so we have one of the preserved medical records which we have is radiology reports because your MRCT all of them are digital and you can always take a CD drive or a hard disk on which you have your dicom files which are saved and you can preserve them for long.

So you have some disease and you are treated and then you can keep those records for referral may be even 20 years later on. So 20 years later on if there are some traces of that disease and on setting or may be a side effect then you can again find it out over there. Now in the modern practice what happens is that for these histologies as well we have a similar kind of an option, which is through something called as a whole slide imager or a whole slide digitizer, ok.

So what it does is you take that whole that complete slide and you mount it onto a cartridge over here, so there is basically a robotic microscope inside this box and what that does is that it takes image of one portion and then slides the slide by a few millimeters or whatever resolution you need and then takes it. So you have basically a tile of such images created, so imagine there are multiple tiles in which you have the whole slides scanned down and then all of them are stitched together to get down a panoramic projection.

So the same way as you do with your normal photography which is you take photographs multiple photographs by rotating your camera and then you stitched it down using any of your panorama stitching softwares or freewares which is available and then you have a 360 degree view. So over here we have a similar view of the whole slide coming down and that is what is called as a whole slide image.

Now, it is pretty good because what you would see down and this kind of a whole slide viewing workstation is this is a virtual slide on which this is where the mass of tissue is and since you could scan down the whole slide, so you have a thumbnail and then you can pick up whichever slide you want to see and then you can magnify to whichever region you want to see and since it is scanned over different magnification different optical magnification, so you can emulate the same kind of behavior and the good thing is that you do not need to anymore sit down at a microscope physical microscope.

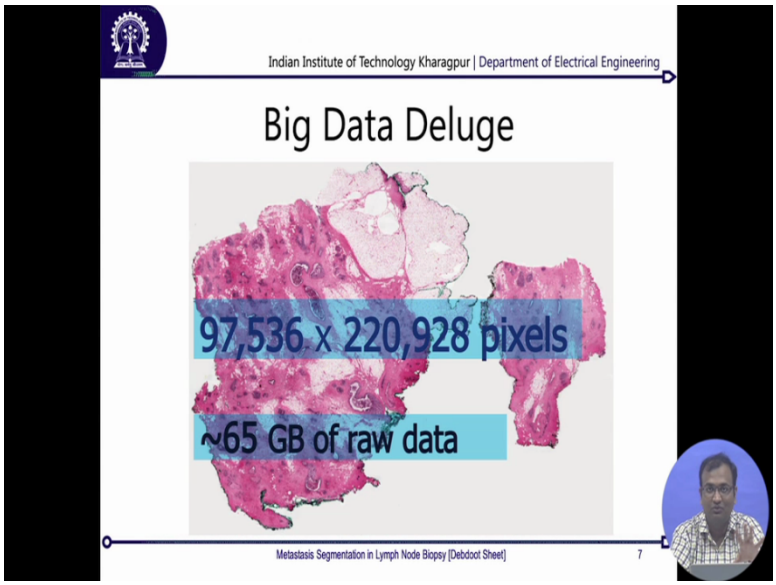
So the pathologies can actually say get down digitized images coming over the internet to him at some other location and the images might be digitized at some other place. So this would bring down, this is what is enabling telepathology in today's world and is not restricting any more that your samples have to be sent on a physically to a different city.

But then a collection center in your small city, rural places and everywhere can take down these can actually collect your sample and then make them into slides and then digitize everything and then a super-specialist pathologist may be located at some other nodal center at a major metropolitan city and still can do the reporting without you having to travel or your samples having to travel down to that major city which has other risks of sample degradation as well.

So now that we have this kind of a major solution available, it should really make our lives much easier, but the only challenge is it does not work out so easy and it is not so empirically straight forward, because what we have today what we are facing is what is called as the big

data deluge and this whole big data deluge is around a condition, now while this is supposed to solve out but then the problem is that there is a cumulative affect associated with that and that is what restricts us from making use of this one and that whole problem is what is called as the big data deluge.

(Refer Slide Time: 12:19)



Now here if you typically consider one of this examples then one virtual slide will have something around 97500 cross 220000 pixels and that together comes down to about 65 gigabytes of raw data in itself. Now that is not a small number because imagine this one, one single slide is equal to one pen drive one of your 64 GB pen drives of data itself, so you would have to carry down one pen drive for every single slide.

Now while the problem was about (()) (12:52) all of these large number of slides over time you basically have the same sort of a cumulative problem, just digitizing does not physically remove out your storage barrier because you are still requiring that huge amount of space and the other major problem is that all the images which you have predominantly till now analyze say for CT or MR, there were just 640 480, or 800 600, 512 cross 512, they the numbers were still in the range of less than a thousand.

And now we are speaking about numbers which are in the order of a million on one single dimension, so this is about 1 million and so this is roughly about 1 million and this so 0.1 billion and this whole thing is also quite close to the that. So now imagine that that these huge numbers are what are going to cause down problems in the coming days and so large number

of images, so large number of scales over there is not something which we can very easily tackle up.

(Refer Slide Time: 14:01)

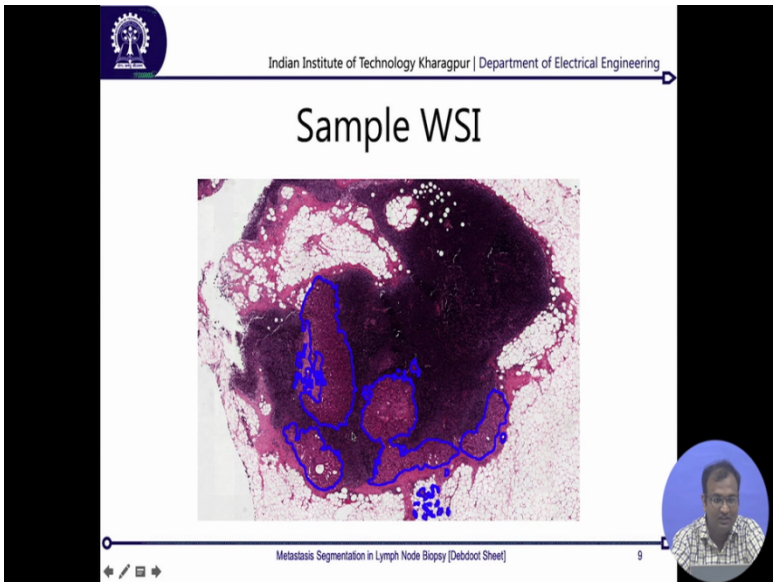
And with that problem in hand is what this whole challenge was designed over there. So what we started down is quite interesting because how though we had the images provided is on a standard way of pyramidal decomposition format. So typically, images are acquired at multiple magnifications as we had seen down in the lab scale demonstration that you can have different objectives with different magnification.

So here also on the slide digitizer which we had shown on the earlier slides you have similar kind of objectives and one objective with a different magnification is used at a time. So initially images are acquired with a 1X magnification with a rough scan, so that will give you more or less the overall architecture of the tissues present over there and then as you keep on increasing the objective magnification as we had seen on the week one lecture on microscophy, so you get down much more detail about say the whole cell and then eventually on the nuclear cytoplasmic border and then internuclear content.

So here this image was provided down in three different magnifications and that was put down into a tiff wrapper and then you can use the open slide framework APIs in order to read all of them. So please feel free to go through the website of the challenge and then you can download this data by just doing a nominal registrations which is for free, so it is just some information you have to put down and you can download the whole data. Now, if you carefully look over here I have it at 2 different magnifications what is shown over here. Now

at a much lower magnification you would be seeing down a whole jumble of cells and nuclei and just one level up of magnification, it becomes much clearer over here. And this is the kind of interesting way in which the whole data is available.

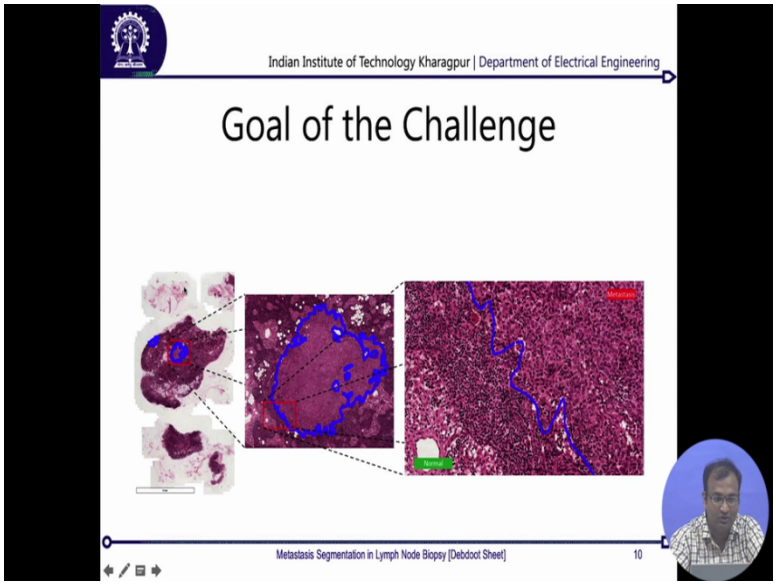
(Refer Slide Time: 15:43)



Now, on the sample side of it we have the whole slide image sample as well presented over here and this is where you can see that there is a lot of diversity over here. So if I am recording this image, so this was typically recorded at 1X of an optical magnification and then shown over here and the regions which are marked with this blue contours are the ones which exhibit distinct unsteady state metastasis within the lymph of Biopsy. Now if you look, majority part over here is just this white areas which do not even contain any mass of the tissue and then you have to look down in a perfect way in order to find out where the actual metastatic part within the tissues is present over there.

So this is the challenge which is faced down by today's community and although you might have a rough guess that may be the textures are different, may be the image intensity is different, but that is not always so because look over here the image intensity and nature of texture is quite similar to the intensity and nature of textures over here, but these regions are still not metastatic, whereas these regions are metastatic.

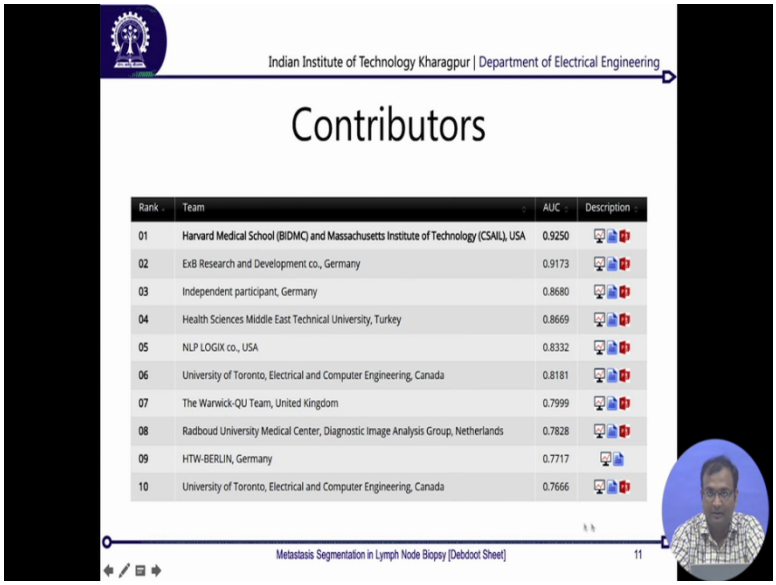
(Refer Slide Time: 16:49)



And that is what creating the main challenge for us and that is where the goal of this particular challenge comes. So the problem how it was defined is that given that you have a whole slide available at three different magnifications. So this is taken out at 1X, this is at 10X and this is at 40X. And at different magnification if you keep on magnifying you would be seeing a different kind of manifestation. So if you look carefully over here, this part is what the metastatic part is and this is what the normal tissue part is over there. So just by looking down at say textures or intensities it is really hard to find it out. But then as you go down to granularity of information you will be able to make out distinctly as to where the border is located of segregating them and this is what forms down the main challenge.

So assuming that say you want to apply some sort of a patch paste method, so now you will have numerous such numbers of patches because the smaller access over here is about 97000 and the larger access over here is about 220,000, so if you are even considering taken down 1000 cross 1000 image patches over here, you would be getting down 97 such image patches and on this side you will be getting 220 such image patches. So a 1000 cross 1000 pixel tile you will be having 97 cross 220 such tiles present and that is not a small number in any way and you will have to run down analytics on them. So this is where comes the big data deluge for you, so you can do your basic math over here as to what is the pixel size and operator size, what will be the total throughput and what will be the total computer power which you would be requiring in order to solve this problem.

(Refer Slide Time: 18:35)



So that is where this challenge was initially started down in 2016 and it was there was a lot of skepticism within the community as to how good they would be able to tackle this particular problem given the scale of compute which is associated with it. And the good news is that it is not so tough actually we have contributors a multiple of those contributors and this is just a list of the top 10 contributors from the day of the challenge result dissemination. So today there has been update on the contributors over here, so the winning team was actually from Harvard Medical School and MIT together joint team, which had a AUC of segmentation, so AUC is your area under ROC curve if you remember from the first weeks lecture on evaluation criteria. So this was all evaluated based on AUC scores and based on the best AUC score they were all ranked one by one.

So the best winning team is who had a AUC score about 0.925 and that is not a small number because you are nearing 1, so you do image that although this is a compute problem at scale at a much larger scale because the amount of data you will have to deal with and the total throughput of power which you will have to put down is quite something which is herculean. But then teams have been able to solve it and this has been solved within a human lifetime so not so complicated with say within a week time of compute is what would be required in order to solve it.

(Refer Slide Time: 19:50)

So without taking and if you see there are multiple teams over there and there have been teams who have been participating from across the world onto this one. Now the winning team on this one from MIT and Harvard together, so they had actually proposed a deep learning based solution and what they were using is they have a pre segmentation method taken down. So remember this point that I was telling you that the whole slide is about 97000 cross 220,000 pixels over there. But then majority of the area is just white blank area over there and you need to you can actually just say a small portion about 20, 25 percent of the area is what where the tissue is located, the actual Biopsy tissue. And you can actually choose out that particular located small part of the tissue.

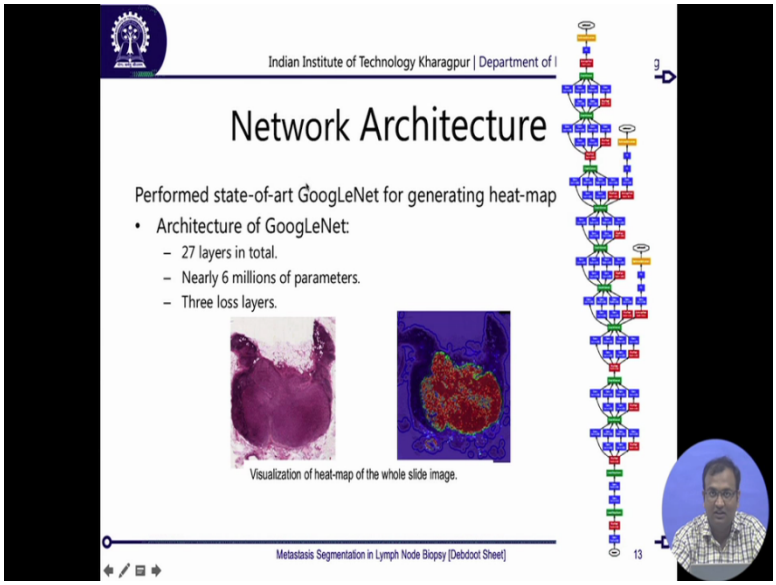
So what they had done was a very quick hack around that so used just a simple Otsu method by which you can iteratively you can find out, which is the best possible threshold and then do a binary thresholding and this is going to roughly give you the area where the tissue mass is present. Now once you have this rough area around where the tissue mass is present, then what they do is they randomly selected about multiple patches, so there were 1000 positive patches and 1000 negative patches of 256 cross 256.

Now, these positive patches are what are the regions where you have a metastasis present, and negative patches taken from a region where you do not have metastasis present, since you have the ground truth available with you as to where this metastasis is located that mark already available. So now you can sample down randomly over some regions from that metastatic region and other from the non-metastatic region.

On top of that they since the dataset also had normal slides over there for training, so they have taken down 1000 such negative patches from the normal slides like anywhere you sample from the normal slide within the tissue region you are going to get a negative patch because that is negative they do not have any kind of a metastatic region present in anywhere.

So using all of this, they trained down a one particular deep learning network in order to do it and what they report over here is that once you are doing this Otsu based thresholding coming down, you can actually remove 82 percent of the region on an average which is useless where you do not have any tissue mass present at all. And that is going to bring down your compute scale from that large onto a much smaller scale problem over here. So these are some very intelligent preprocessing techniques very simple but wise preprocessing techniques you can always put down in order to scale down the computational overload for your system.

(Refer Slide Time: 22:15)

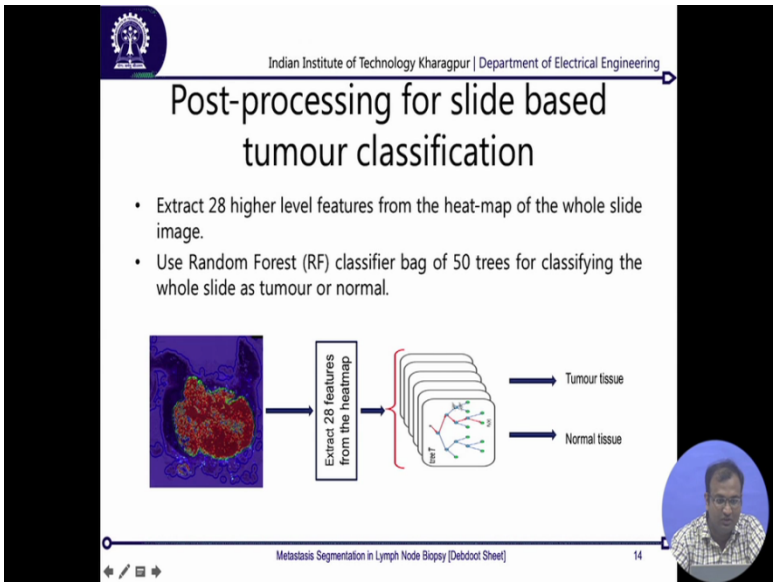


So the network which they use is what is the state of art GoogLeNet, so this is the modified Googles modified version of the standard LeNet architecture and what they have is it is a 27 layer network in total and including some of these layers which are called as inception layers. So you can read much more details on the actual GoogLeNet paper and on this one so since this is much beyond the coverage what I just wanted to point out is that since it is a very deep network with 27 layers, so one problem is that you will have vanishing gradient issues coming down when you are trying to backpropogate.

Now for that one they have these inception layers over there which keep on injecting extra errors on to the system and the way these inception layers work is these are very weakly supervise classifier layers which use all the features on the particular layer and then they try to run a classifier and see whether it is working or not. So the performance of one of these inception layers which inject down extra error might be much weaker than the whole network as such but cumulatively they do a much better job in order to preserve a gradient to be transferred across the network while training such a big network. And this network approximately has 6 millions parameters to be tuned down over there.

And finally what you get out of this network is what they show is a heat-map, so this is the probability map of a region belonging to metastasis. So, typically if this is what you give over here and then train it and then on this particular test sample you would be seeing a heat-map coming down. So this is a probability that those particular regions or those particular patches they belong to a metastatic region. Now, it appears more or less continuous over here because you have very small patches 256 cross 256 taken down from a whole slide which is 97000 cross 220 samples. So they would approximately at a much smaller scale (()) (24:02) to almost as if there is one single pixel which we are looking down but it is a collection of pixels which we are looking.

(Refer Slide Time: 24:18)

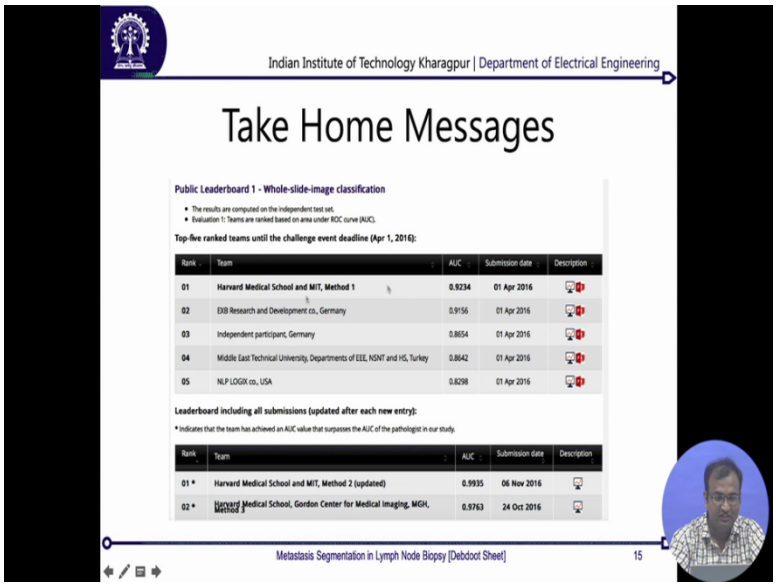


Now with this initial evaluation or the heat-map coming down as to where the location of metastatic regions can be located, what they do is they extend this one so they use this initialization coming down from the CNN GoogLeNet and then they start extracting multiple features from the patches present over here. And using those features they again train a

random forest with 50 trees and then these predicts out whether a particular slide over there has metastatic or is perfectly normal.

So there are two problems which were being addressed over here, one is whether the slide is metastatic or it is not, the other problem was what region over there is a metastatic. So you need to solve out both the problems over there. So together this is what they have solved out and this is where I would just draw a conclusion because there are multiple methods who have been over there you can look through all the presentations and PPTs are available clearly over there, so you can go through them as well.

(Refer Slide Time: 25:21)



And where you can go down is if you go on the website over there and then click on results you would be directed down to the Leaderboard and on this Leaderboard you would be seeing multiple of them and then you can click down on each of these links for a power point presentation and that would be downloading an appropriate PPT on which the details are present of the whole method in which they were implementing this.

So, on the conclusive note for this one as we come towards an end of the lecture as well as this season of the course on medical image analysis, I would encourage strongly everybody to actually participate on these kind of challenges, which are currently on going including the Camelyon 2017 challenge which is now up and do definitely make a note that these references provide a much detailed appreciation for how you can do including the software tools and how to set up your systems and since we have already done some hands on programming experience as well, I would strongly encourage all students to really go ahead

and participate in these grand challenges for just not for the sake of learning but also for providing a community driven challenges to major medical image analysis problems we are facing today. So with that I come to an end and thanks and all the best for your upcoming assignments and examinations, so thanks, bye.