Subtitles and transcriptions
Subtitles and transcriptions are available for selected materials for purpose of helping users understand the contents of the educational sessions.
Uncertain words have been indicated with ?? before and after the part. Parts that could not be understood at all have been indicated as [Audio Not Clear].

Every effort has been made to faithfully reproduce the audio of the sessions as recorded. However, no responsibility is accepted for mistakes or omissions. ESO does not endorse any opinions expressed in the presentations.

Breast cancer CAD tools

Dr Polónia: Good morning, everybody. My name is António Polónia. I'm a pathologist. I work at Ipatimup which is in Portugal, Porto, Portugal, and I'm going to talk about "Breast cancer CAD tools". So, first of all, I would like to thank the invitation. For me, it's a pleasure and an honor to be here. These are my disclosures. So, I will, basically, talk about the work that we have been doing, believe it or not, since 2014. So, me and Catarina Eloy, we are both pathologists. We grouped with Paulo Aguiar which is what you would call a computer scientist. And we later teamed up with the team of professor Aurélio Campilho along with, at the time, two of his PhD students, which was Guilherme Aresta and Teresa Finisterra. And we organized the Bach Challenge. So, the Bach Challenge was a grand challenge on breast cancer histology images that were held at ICIAR 2018. And it had the purpose to develop software that was able to classify breast tissue into four different classes. And the classes were, normal tissue, benign, carcinoma in situ, and invasive carcinoma. So, at the time, this was new because usually the tasks that were been doing at the time were basically detecting invasive cancer from everything else. So, we increased the level of complex nature of this task into four different tasks. Now, we know that not all lesions will be included into these four classes, but the majority of the lesions will be. So, this work was published in 2019. It had, as I said, the purpose to classify breast tissue into four different classes, and it had two parts to test two different tests. We had Test A and Test B. Test A consisted in the classification of the four classes that I talked about, the classification of microscopic photographs. Test B, which was a little bit more difficult, it had the task of classifying whole slide images. Now, as you will see later, so, this is Table 2, the summary of the methods submitted for part A, and we see that 10 teams were able to achieve a performance above 0.8, which means 80% accuracy and this 80% accuracy, it's probably, as you will see later, the average accuracy of human experts. But we had 10 teams achieving or even surpassing this threshold of 80%. In part B, we have a lesser accuracy because it's a more difficult task. But funny enough, the team that won part B was also one of the teams that had higher classification on part A. So, we choose this team, team 248, which was composed just by one person, Scotty Kwok, which is here. And if you are interested to see the nature of his algorithm and, particularly, the way that he trained the algorithm, which I think was very smart on his part, you can see here, the paper he submitted on ICIAR 2019, it's on the web. So, we took a larger group of pathologists because we wanted to see, we wanted to compare the performance of a larger group of pathologists against the software, but also, the interaction of the human observers and the software. So, we put together a small team of four pathologists. And at the time four pathology residents, which was Sofia Campelos, Ierece Aymore, Ana Ribeiro and Rita Canas-Marques, they were the pathologists. And at the time, the residents were Daniel Pinto, Ricardo Veiga and Magdalena Biskup. So, the task was this. They had to classify, just based on HE, the microscopic photographs, which were 100 photographs, and also the Test B, which consisted in the classification of 152 regions of interest in just 10 HE whole slide images. So, in Phase 1, they had to make that classification based on morphology alone into these four classes, normal, benign, carcinoma in situ, and

invasive carcinoma. Now in Phase 2, they had to do the same thing, but they knew the output of the system. They have no idea if the system was making the correct diagnosis, or classifications, but they knew the output of the system and they could use it as they felt. There are some rules of engagement that I will speak later. In Phase 3, they had to do the same task as in Phase 2, but not only they knew the output of the algorithm, they also knew the accuracy of the algorithm and the accuracy of the observers, which meant they had no idea what was the correct response, the correct classification, but they knew how good they were and how well the system was. So, the rules of engagement were just these two rules. If the classification of the observer was equal to the classification of the algorithm, no change was allowed, but if the classification did not match the classification of the algorithm, changes were allowed. So, this was to simulate the hypothetical situation in which if an expert thinks it's one class and the system, the output of the system is the same, why would they change it? But if it's different, they can think about it. So, how good was the system? So, here you have P for pathologists, R for residents, and AI for artificial intelligence. And the system was really good because it was the second-best result. You see that the average of pathologist on Phase 1 was 80%. The 80% threshold that I talked about earlier, and the system was really good. However, if you, instead of evaluating the accuracy, which is the number of correct responses made by the algorithm, you make the concordance with the ground truth, you actually have two pathologists better than the system, but even in this situation, the system performs really well. Now, in Phase 2, they knew the output of the system. And if the classification matched the initial classification, they could not change it, but if it was different, they could reconsider it. And funny enough, and although in this phase, they had no idea if the system was good or not, they already kind of knew that the system was performing well and they changed their initial classification. They changed the initial qualification to the suggestion made by the output of the algorithm. And you kind of see this pattern of pathologist on the left, residents on the right, and AI on the middle. Now in Phase 3, after knowing that the algorithm was really good and how good they were and compared themselves to the algorithm, people got to trust more the algorithm and change more their initial classification, and they agreed more with the algorithm. So, you kind of see these almost all pathologists, almost all the residents on the left, having higher accuracy than the algorithm... But what I find remarkable, because initially I thought that the observers that were below the algorithm would probably increase their accuracy up to the level of the algorithm. That was my initial expectation, but in the end, that's not what happens. Almost all observers were below the algorithm and now, almost all observers are above the algorithm, so they increase their accuracy and they increase their accuracy above the accuracy of the algorithm. I find this remarkable. Now in Test B, which was more difficult, the algorithm didn't perform quite well, and everybody was better than the algorithm, but what we measure was the effect of a bad algorithm on the accuracy of the human experts. And what we see is that some have slightly increased, some, usually they have a slight decrease. On average, they decrease slightly. It was not statistically significant, which means that if you are assisted by a CAD tool that is worse than you, you won't decrease your accuracy. And this is also very important. We also try to classify the pictures into easy cases and difficult cases. So, we did this by classifying the images into images that are correctly classified by less than 50% of the experts, which is a difficult image and images that are correctly classified by more than 75% of the observers are what we call easy pictures. And we actually see that the accuracy of the algorithm is higher in the easy pictures, rather than on the difficult pictures. In Algorithm B in Test B, you don't see that, you see more or less of a plateau in which there's almost the same accuracy in difficult or easy pictures. We also notice that the accuracy of the observers was higher in larger ROIs because in whole slide images, ROIs have different sizes. So, we could measure the accuracy with respect with different sizes of ROIs. So, if the ROI is bigger, is larger, the accuracy is higher, and the system also has the same trend. So, it has higher accuracy in larger ROIs than in smaller ROIs. So, small lesions are always more difficult. Now, we have a couple of examples. Here, we have three images of inflammation. Now, the first three are fat necrosis, which we know, as pathologists, we know that this is a pitfall for invasive cancer. And the fourth picture is just peritumoral inflammation. Now, you can see that the output of the system was invasive carcinoma in all of these pictures, which is incorrect. And, apparently, it's one of the pitfalls for the system. It's a pitfall for us, and, apparently, it's a pitfall for the system. Now, in a different situation, all of

these pictures represent carcinoma in situ. They represent the CIS. And the output of the system was correct. It was the CIS for all of these pictures, but you can see that initially, only half of the people were making the correct diagnosis. And on the last phase, after they saw the output of the system, more people were able to correctly classifying the lesion as carcinoma in situ with the system of the algorithm. This is just to prove that this is in fact the CIS, the lack of cytokeratin 5/6 and strong ER expression. This is another example. This is the same lesion, but from two different cases. This is a very difficult lesion. Some of you might have already made the diagnosis, but this is a very difficult lesion. What was the output of the system? It's normal, which is incorrect. This is a difficult picture. Only one pathologist was able to correctly classifying this image. Here, more people did it on Test B because of whole slide, because of context, because of larger regions of interest. But the output of the system was normal which was incorrect. And to prove to you that this is in fact not normal, this is an E-cadherin which you can see, there is a lack of E-cadherin. So, this is in situ lobular neoplasia. Now, final example, also the same lesion, but two different cases. And this is very interesting because this lesion, this photograph in Test A was incorrectly classified by all of the observers. All observers said, in this picture, invasive carcinoma, which is not. It's a sclerosing lesion, but it's really difficult. And in here, it was a little bit more easy, but still difficult. So, what was the output of the algorithm? Benign, which is correct. And, apparently, the pitfall that we have on sclerosing lesions, the system does not have. It does not recognize in sclerosing lesions the features of invasive carcinoma and we can take advantage of that. So, this is just to prove that it is, in fact, sclerosing lesion with the presence of CK5/6, P63 all over the place and heterogeneous ER expression. So, in our work the most clinically relevant incorrect classification was between benign and carcinoma in situ. Usually, the differential diagnosis between hyperplasia and the CIS. And we saw about 7% of this error rate, but with the assistance of the algorithm, we could decrease this error rate almost a half, about half of that which is amazing. In Test B, it was more or less the same because the algorithm was not really good, so, we couldn't benefit from it. So, what we saw was that from Phase 1 to Phase 3, the concordance between experts and the system increased. And initially, there was this more or less linear relationship between concordance between experts and the system and accuracy. If you agree more with the system, you will have in the end, more accuracy. And in Phase 3, we had this outlier which was an observer that just didn't trust the algorithm. And it was also the observer with less accuracy on the last phase. So, we then asked, what is the probability of being correct if we don't agree with the system? And this is very interesting because in Phase 1, if the observer and the system did not agree, the likelihood of being correct was on the side of the AI system. The system was more likely to be correct rather than the human expert. Now, in Phase 3, in which the human expert was able to know the output, consider the output, and change their initial classification, we see that the likelihood of being correct if the system and the human expert don't agree is higher for the human expert rather than the AI system. Now, I know that these differences are not statistically significant, but the increased likelihood of being correct from Phase 1 to Phase 3 was statistically significant. Now, you probably are asking, "So what is the likelihood of being correct if they agree, if both human experts agree with the system?" And the likelihood is really high. It's not a hundred percent, but it's close to that. So, if the likelihood... if the system agrees with a human observer, they are probably correct. So, the question is, "Will machines replace humans?" That's what everybody is really afraid of. And we know that machines are really good at pattern recognition. Some people will say, better than humans. And because pathologists do a lot of pattern recognition, they could be, in theory, be replaced. Unfortunately, that's not what pathologists just do. They do much more than that. And even if they just did that, which they don't, you have to teach every little task, pattern recognition that pathologists actually do and that takes a long time. But what do machines really need to replace us? Well, machines need the “holy grail”. They need the “holy grail” to be able to replace us. And the holy grail is the ground truth. However, the ground truth is somehow of a utopia. Sometimes the ground truth does not exist. For instance, in borderline lesions, in small lesions, in difficult lesions, rare lesions, even experts don't agree with each other. So, as in Matrix, “there is no Spoon”, maybe, there is no ground truth. Now, finally, I just want to say that it's time to choose, but it's not to choose the “red pill” or the “blue pill”. That choice has already been made. You have to choose when are you going to take the pill that has already been chosen and to choose when do you

enter this, the digital revolution. So, some people will enter sooner. Some people will enter later. You just have to decide when. When do you want to enter and when do you want to make the choice. But the choice has already been made. So, thank you very much for your attention. And I will be glad to answer to any question that you have. Thank you very much.