Herbert Y. Kressel, MD  Hi. This is Herb Kressel and welcome to the Radiology podcast. Today I’m joined by Dr. Luciano Prevedello who is Assistant Professor of Radiology and Division Chief of Imaging Informatics at The Ohio State University Wexner Medical Center. Dr. Prevedello, welcome.

Luciano M. Prevedello, MD, MPH  Thank you. Thank you very much for having me here today.

H.Y.K.  Sure. Today we’ll be discussing the paper that you and your colleagues wrote on “Automated Critical Test Findings Identification and Online Notification System Using Artificial Intelligence in Imaging.” It’s a very promising topic indeed and I must share with you that yesterday as I normally do I woke up to National Public Radio, but I woke up very abruptly because what I heard on the radio was some informatics expert pontificating that artificial intelligence meant the end of radiology as we know it. Happily the piece also included a discussion by members of the data science group at Mass General Hospital who put it into better perspective noting that artificial intelligence was a tool that would enhance radiologists function but not a primary threat at this time. So I’m sure everybody will be interested in your paper and the discussion of it. So first what actually is deep learning and if you could define that that would be helpful.

L.M.P.  Sure so deep learning is actually a one method using machine learning and machine learning is within a larger and broader concept of artificial intelligence. If we consider the old method the methods that have been developed over many years with are neuro networks all deep learning is a new methodology that uses neuro networks but with many hidden layers and with that the algorithm to find very complex (inaudible) within the info provided.

H.Y.K.  Got it. So in your study let’s talk a little bit about the methods. So you’re looking at identifying critical brain findings on non-enhanced CT so how did you actually begin to determine the number of cases that would be included and to develop the function?

L.M.P.  So that is a very good question because many people in the space of deep learning they say that it’s required to have thousands or even millions of images to be able to run these algorithms. What we were able to find with this research is that for you to prepare the data so to do some processing that allows the computer to recognize those patterns at a much faster pace. So it was – it’s very difficult to determine ahead of time what is the number of cases that the computer will need, but we tested the learning rate so the computer can graph in which the computer is learning so that learning process takes several iterations you know that it’s going to take many data sets for you to do that. But in our case we were able to see that given the pre-processing that we did that 100 cases was gonna be enough for it to determine.

L.M.P.  So for the (inaudible) the 100 cases for the initial training set, is that correct?

L.M.P.  100 cases for hemorrhage mass for hemorrhage right.

H.Y.K.  Okay and then you had separate numbers of cases that’s how you got to the 246.

L.M.P.  Correct.

H.Y.K.  Now for each of these exams it looks like you generated a total of 2500 images for the total group and so what is the effect of you know do you do the whole exam is that what it is or do you do variations on the key sections to generate these additional images that are used for training?

L.M.P.  So for the purpose of this study we concentrated the analysis on the mid portion of the head wherein most of the pathology is. So after removing the areas...
with air in non-important information over the skull base which doesn’t have a lot of variability and is concentrated in that area, that number of images relates to the single slice in every exam which is about 4.8 mm or 5 mm each.

H.Y.K. Okay so you wanted to get the whole – the key volume of interest. Now my understanding is that sort of one of the key drivers of the effectiveness of this approach is actually how the initial tagging or annotation was done and so who actually tagged or annotated the images that were used to train the system in your study?

L.M.P. So I worked alongside with the radiology resident and we tagged them all going through the consecutive, all the consecutive cases in a specific time frame we evaluated report, technology report, and we had a methodology to tag them described in the paper which is very – the description is detailed in the paper, but it was myself and a radiology resident tagging them all.

H.Y.K. So you know a lot of times when we do diagnostic performance studies one of the things we study is inter-observer variability, do you have any notion if you used a different radiologist or more experienced radiologists do you think the tagging would have been the same or might it have been a bit different?

L.M.P. For this paper we didn’t evaluate the – we didn’t do any kappa analysis or agreement inter-reader variability. It is an important question to ask ourselves if it’s actually part of a larger group that needs to determine some standards of how to tag those cases because at the end of the day as you mentioned tagging is a very important activity for these machine learning processes because the computer will learn based on what we’re providing to it.

H.Y.K. So in the paper you mention that you looked at two different window settings and I know that’s related in part to the different types of clinical applications you were anticipating, but can you tell us specifically why you chose to look at two different types of window settings?

L.M.P. So basically this is what you would do in a clinical setting. When I’m reviewing a stroke case I cannot detect that stroke in the CT of the head with the same setting that I used to detect hemorrhage. So when we were fitting those cases to the algorithm it became clear that it was not going to be able to recognize the pathology the stroke disease using the standard setting for hemorrhage because it was missing those cases. So we replicated what we do in the clinical setting and then we were able to obtain a much better performance.

H.Y.K. So if this progresses and you actually have a viable clinical application for this program, do you think you’ll need to generate separate settings or will the machine be trained to basically replicate the two settings from a single data set?

L.M.P. I think there are two approaches to this. The approach that we did was we were using, we were converting the DICOM file to a JPEG which in that process you decrease from 60 bit to 8 bit and because of that, because of that limitation that we had, we had to create the settings. If you go to the – and just go back a little bit – in that setting then you have two different windows and yes you would need two different algorithms to identify all those pathologies, but if the algorithm is capable of going to the densities themselves the Hounsfield units, may not need to go that route.

H.Y.K. I see. Good. So now that we sort of understand what did you do, how did it work? What did you find?

L.M.P. So after evaluating all those cases the machine learned from the previous tagging that we did. For the hemorrhage, mass effect, and hydrocephalus the performance of the algorithm was 0.91 in the area under the curve, the ROC curve. Stroke it was 0.81 for the detection of those cases.

H.Y.K. I see. So that’s I think pretty promising. So you then in the study we didn’t talk about this, but after you have developed this tool you then tested it with a separate independent validation set of 226 exams and again what was the rationale for doing that and do you think this is a large enough number of exams for the independent validation set?

L.M.P. So we – I think that the main reason for that (inaudible) to be separate from the training and the validation set was because we needed to be sure that this was generalizable. So we evaluated the performance of algorithms in a totally different time frame to see okay we developed the algorithm in the beginning of 2015 all the cases in the beginning of 2015 let’s see six months of that time frame whether it’s gonna perform the same way. In terms of the number of cases, I believe it was, it’s one of the limitations of the study. The more cases you have the more you’re gonna be able to test some of the limitations of your own algorithm and I think future studies will have to test this methodology on a larger scale.

H.Y.K. So you know I’m certainly just starting to learn and understand these types of studies, but what I’ve learned is that with deep learning research there are, particularly with this kind of clinical application, there are generally two major types of concerns; spectrum bias which really is asking was the case distribution truly representative of the total population that would potentially have the abnormality, and then dependent of that was the did the technique lend itself to over fitting the data and they are somewhat related but separate. So how would you say these two types of concerns were addressed in your study?

L.M.P. To answer that question, number one I think we did something that is it’s important to emphasize that
you don’t see in many of the research out there. So what
we did was we evaluated consecutive cases of available
(inaudible). Doing that we’re not doing any selection bias
because what you have there is what you’re gonna get
in the clinical setting. That relates also to the fact that
we’re dealing with (inaudible) optimal data sometimes
and you have to move around and those were fed into
the algorithm. My concern of some of the research that
does not do that is that (inaudible) gonna apply those al-
gorithms in the clinical scenario it will not know what to
do with those motion-degraded studies. So that’s on the
selection, data selection side. On the over fitting there
are ways in machine learning to deal with over fitting
and basically if you’re evaluating your validation set and
comparing to the training set and at the same time or
towards the end of analysis you’re evaluating your test set
so you have (inaudible) and if the performance of your
algorithm is similar in all those sets, then the likelihood
of over fitting is it’s minimal. That’s what you’re trying to
do by doing that type of analysis.

H.Y.K. So in looking at your results, the diagnostic per-
formance was quite promising, but do you think this is
actually good enough for a robust clinical application?

L.M.P. Well and then I think here it’s important to dis-
tinguish what type of clinical application. If you’re asking
me whether this is ready to go for clinical practice and
start reading studies, I would say absolutely not. This is
not ready for that. But if the question is can this assist
us as a radiology community in going through all the high
priority cases, one example is that there are reports out
there that says that about 40% of all our inpatient studies
are considered STAT. So when you have that number of
STAT and the list keeps growing and growing, how you’re
gonna sub-prioritize those high priority studies. And for
that type of situation I think this is very promising.

H.Y.K. Great. So let’s look into the future of this be-
cause I think it would be very helpful in the ER setting
for someone to have a machine tool that can really high-
light the stroke patients who need to be identified or the
(inaudible) or whatever and immediately so what else do
you think is needed to develop the tool that you’re work-
ning on to be a mainline you know tool for evaluation?
What are the next steps that you foresee for your group
for this work?

L.M.P. And we are already working on this. We are try-
ing to make the application imbedded into, at least in the
beginning, a research setting, but there’s a lot of – when
you develop an algorithm in the lab it’s very different to
make it to production style. So in production is not nec-

essarily like a tool that you can install everywhere, but
make it so it can ingest the DICOM files, all the images
and convert them automatically and then have a final al-
gerithm that spits out the outcome. All that is something
that we’re working on to make it more workable in that,
more translational research environment and to answer
your second portion of the question is I think what is
needed to make that transition is gonna be a validation.
This is a very important feat of any research through
okay it worked well in this setting but let’s see in the day-
to-day application will this hold true as well and that’s
what we’re trying to do as a second stab – very important
piece of the puzzle.

H.Y.K. So would your advice to me the next time I wake
up with a shocking statement that AI is the end of radiol-
ogy I should just shut off my alarm clock immediately and
not listen any further?

L.M.P. No I think actually it’s important for us to dis-
cuss this.

H.Y.K. Sure.

L.M.P. We shouldn’t ignore artificial intelligence we
should embrace it. But what is important to understand
here is that there are a lot of other things that we do as
radiologists that are not gonna be able to be replicated
by machines and we’re not a one trick pony and there’s
several things that we need to do in order to leverage
these technologies to allow us to do a better job, but I
don’t think you need to be worried any time in the future,
in our future at least.

H.Y.K. I’m turning 70 I just have to worry about the
next six months.

L.M.P. So I think it’s something that we need to pay
attention. It will change the way that we practice in the
future, that’s why it’s important that we embrace, under-
stand it, have an educational component to our residents
about this, but not be afraid of it because it won’t take
our jobs.

H.Y.K. Great. Well Dr. Prevedello I really want to thank
you for taking the time to join us and thank you so much
for this fascinating paper. I certainly learned a lot from
the conversation and I hope our readers and listeners
will as well. Thank you so much for participating.

L.M.P. It was my pleasure. Thank you for inviting me.

H.Y.K. You’re welcome.