Validation of ML Libraries

With Professor Hasse , Prof. Dr. Christian Johner

Transcript

00:00:05 Speaker 1

Medical Device Insights.

00:00:08 Speaker 1

A podcast by the Johner Institute for medical device manufacturers, authorities and notified bodies.

00:00:18 Speaker 1

Our podcast here is certainly known for always drilling thin boards and we are planning a few for the future

00:00:27 Speaker 1

introductory topics for beginners, for beginners, but this time it won't be the case.

00:00:33 Speaker 1

This time we're drilling extremely thick boards and all those who think machine learning and regulations, I'm not particularly interested in that, they can skip this podcast episode for once.

00:00:47 Speaker 1

Everyone else can look forward to real thick content and I have Oliver Haase with me today,

00:00:55 Speaker 1

with whom I have been working for a while in the context of machine learning and also the validation of machine learning libraries.

00:01:02 Speaker 1

Oliver, do you want to introduce yourself very briefly so that our listeners know who I can have as a guest here?

00:01:07 Speaker 2

Yes, very much, dear Christian.

00:01:08 Speaker 2

Thank you very much for allowing me to be here today.

00:01:10 Speaker 2

I have a background in computer science.

00:01:12 Speaker 2

I am a computer scientist, have a professorship in software engineering and have been involved in software verification and software validation for many years.

00:01:22 Speaker 2

And for some time now, it has also been applied to machine learning, i.e. machine learning, especially in the context of regulated markets such as medical devices.

00:01:33 Speaker 2

And I am driven by the guiding question, especially the central question, of how to check and prove the correctness of a machine learning application.

00:01:44 Speaker 2

And this evidence goes far beyond model evaluation as we have seen it.

00:01:51 Speaker 2

from machine learning.

00:01:53 Speaker 1

I guess that you are now mainly alluding to the libraries that also have to be validated.

00:01:59 Speaker 1

What do you need them for, what, what kind of libraries are they that most people use and that probably also need validation?

00:02:07 Speaker 1

At least I just heard you when you said it's not just about the model itself.

00:02:12 Speaker 2

Yes, exactly, Christian, of course I'm also alluding to these libraries, which are also called frameworks, i.e. machine learning libraries or machine learning frameworks.

00:02:20 Speaker 2

And these are such things as PyTorch, TensorFlow, Carras for neural networks or X.G.

00:02:29 Speaker 2

Boost for gradient boosting.

00:02:31 Speaker 2

And these libraries are collections of machine learning algorithms that can be used to create models, machine learning models, if you have training data, if you have good training data.

00:02:44 Speaker 2

And these machine learning models, in turn, can then make predictions in the end by

00:02:50 Speaker 2

extract patterns from the training data, recognize and extract patterns.

00:02:54 Speaker 2

And these predictions are then good, for example, to rely on C.T.

00:02:59 Speaker 2

images.

00:03:02 Speaker 2

And this training of the models, that's a, that's a very complex task, an algorithmically very complex task.

00:03:10 Speaker 2

And that's why these machine learning libraries are also the real reason why it is comparatively easy for everyone to use

00:03:18 Speaker 2

models, machine learning models, if he has good training data available, because these algorithms are already there.

00:03:25 Speaker 2

Theoretically, you can do the same thing without libraries, so you don't necessarily need them, but that would mean years of development effort for people if you wanted to implement it all again by hand.

00:03:39 Speaker 1

Similar to the other libraries, the question naturally arises here, yes, how is this to be viewed from a regulatory point of view and

00:03:48 Speaker 1

Of course, we all know the sub-requirement, i.e. the requirement for the Software of Unon Provenance.

00:03:55 Speaker 1

62 304 already gives us a lot of what we have to do, what we have to specify in terms of requirements, how we have to validate or verify them.

00:04:06 Speaker 1

If I understand you correctly, however, we have a more complex scenario here.

00:04:11 Speaker 1

So this is not just about the 62 304 and the sub

00:04:16 Speaker 1

Validation.

00:04:17 Speaker 1

Can you then give our listeners an introduction to how this is to be viewed from a regulatory point of view and what you have to do in concrete terms to create the regulatory requirements?

00:04:27 Speaker 2

Yes, very much.

00:04:28 Speaker 2

Yes, that's exactly right.

00:04:30 Speaker 2

So, this is to be assessed differently than with traditional software.

00:04:34 Speaker 2

In this respect,

00:04:35 Speaker 2

than that these machine learning libraries play 2 different roles.

00:04:40 Speaker 2

Not only do they play the roles, the role of a SOOP, when they are then used in the finished medical device, but they also play the role of a software tool during the training process.

00:04:53 Speaker 2

So, to be more specific, if a model, as long as a model is trained,

00:05:00 Speaker 2

With the help of a machine learning library, this library plays the role of a software tool in the training process and is therefore regulated according to ISO 13485, i.e. the harmonized standard for quality management.

00:05:15 Speaker 2

And the fully trained model that is used in the medical device plays the role of a SUB, as you mentioned, a software fun-on-provenance.

00:05:27 Speaker 2

and must then be validated according to 62304.

00:05:33 Speaker 2

And the good news overall is that if you validate the machine learning library according to these regulations, you can then also use it in the finished product and also for the development process, even though it is third-party software.

00:05:50 Speaker 2

And that's very good news, of course, because as you said, this is a

00:05:55 Speaker 2

a great source of reuse.

00:05:59 Speaker 1

So we actually have a mental division that we have to make with these libraries.

00:06:04 Speaker 1

I think we will talk about it again later that we have to look at one part as a computerized system regulated by Chapter 416 that has to be validated, i.e. that or as a process tool, and on the other hand another part that we then see from 62304 point of view as a sub

00:06:25 Speaker 1

and that's certainly a special feature.

00:06:29 Speaker 1

Do you see any other differences to classic software than in the sense of third-party libraries, what do I know, for example a statistics package that now goes beyond that?

00:06:41 Speaker 2

Yes, absolutely.

00:06:42 Speaker 2

So we'll actually go into this distinction in more detail later, what it means for validation.

00:06:49 Speaker 2

The further difference or a substantial further difference

00:06:53 Speaker 2

is simply the much bigger role that such a library plays in the overall product.

00:07:01 Speaker 2

With traditional software, the manufacturer develops a large part, usually a large part, of the software by hand and then takes care of certain partial functionalities through third-party libraries.

00:07:15 Speaker 2

And that's completely different with machine learning.

00:07:17 Speaker 2

There is the almost or not only almost, there is the complete

00:07:22 Speaker 2

The result is actually soup.

00:07:24 Speaker 2

In the machine learning development process, the manufacturer does not actually develop software itself, but uses its training data.

00:07:36 Speaker 2

He configures the machine learning library and in the end the code that is created is completely third-

party code.

00:07:46 Speaker 2

That is, the model

00:07:48 Speaker 2

is actually sub in its entirety.

00:07:50 Speaker 2

This is a completely different central role that this library plays compared to traditional software development, and so of course this subvalidation also has a completely different status.

00:08:05 Speaker 1

I think that's a very good keyword, because now we're already at validation and what I think most people realize is that it is required by regulation for the model to be validated afterwards is not something you have

00:08:18 Speaker 1

then the library is also indirectly validated.

00:08:24 Speaker 2

No, unfortunately no, not from 2 out of 2 different perspectives, once not in terms of regulation and once not in terms of content.

00:08:33 Speaker 2

So from a regulatory point of view, it is completely clearly specified by 62 304, there are clear requirements for SOOPs,

00:08:43 Speaker 2

under which they may be used.

00:08:45 Speaker 2

And this includes, above all, specifying the expected functionality and then validating it.

00:08:52 Speaker 2

And this is something that auditors will adhere to during an audit.

00:08:59 Speaker 2

And that is also quite justified and that is absolutely justified.

00:09:03 Speaker 2

And that brings us to the substantive reason why this is not enough.

00:09:07 Speaker 2

And that's the argument, it's the same as traditional software development.

00:09:12 Speaker 2

Not only is the complete software tested, but testing is carried out at various levels of software development.

00:09:21 Speaker 2

The individual units are tested, then the integration test is carried out up to the complete system, larger and larger components are tested and this is done because testing.

00:09:35 Speaker 2

can only ever show the presence and never the absence of errors.

00:09:40 Speaker 2

That is, it will never be a complete proof of the absence of errors, and that's why you do it on many levels, simply to increase the probability that you will find errors, because you will only ever be able to test parts of the actual inputs.

00:09:59 Speaker 2

And this is especially true for machine learning models, and we know that at the latest

00:10:06 Speaker 2

since the vulnerability of machine learning models to so-called adversarial attacks has been known, in which it is relatively easy to arrive at incorrect predictions due to small differences in the input data.

00:10:24 Speaker 2

And these adversarial attacks alone show that we can only ever depict a small section of reality when testing.

00:10:33 Speaker 2

And that's why it's important to

00:10:35 Speaker 2

that we do this on a wide variety of levels.

00:10:39 Speaker 1

But now a manufacturer might say, yes, they are somehow already used thousands or millions of times, why do I have to do a validation again now, can't I argue somehow risk-based and say that the probability that I will find something is so arbitrarily low?

00:10:57 Speaker 2

Yes, unfortunately no, here too

00:10:59 Speaker 2

the fact that these libraries are used frequently is certainly an important building block in the overall view, but it is not enough on its own.

00:11:08 Speaker 2

And since you mentioned risk-based, Christian, that's where the difference between 13 485 tool validation and 62 304, namely subvalidation, strikes.

00:11:22 Speaker 2

In tool validation, you can certainly do that, you can argue risk-based, i.e. risk-based.

00:11:28 Speaker 2

With sub-validation, you are simply not allowed to do that.

00:11:31 Speaker 2

It is not part of the concept that you weigh up the risks.

00:11:36 Speaker 1

Except if you argue about the software security class, of course, but otherwise there is indeed nothing about risk-based in the context of Sub.

00:11:44 Speaker 2

Yes, exactly, exactly.

00:11:45 Speaker 2

When it comes to the actual validation, then it's simple, then it's closed, then the sub has to be specified and validated.

00:11:51 Speaker 2

And by the way, this has not only a regulatory background, but also an actual substantive background.

00:11:56 Speaker 2

So, these regulations are justified here as well.

00:11:59 Speaker 2

Some of you may still remember the so-called Heartbeat Bug from 2012.

00:12:05 Speaker 2

that was a mistake in an Open SSL implementation and Open SSL is a library used by web servers that was very, very widespread, is still widely used today and has at least one level of awareness and distribution like the machine learning libraries mentioned.

00:12:25 Speaker 2

And this error, which had the result or it had the consequence that over half a 1000000 web servers became vulnerable and it was only discovered after a long time,

00:12:35 Speaker 2

This was due to the mistake of a PhD student who programmed a new feature and exactly one member of the Open SSL community.

00:12:44 Speaker 2

That was a developer who accepted exactly this feature and included it in the Open SSL code.

00:12:50 Speaker 2

.

00:12:50 Speaker 2

By the way, some web servers are still susceptible to this bug today, which is also a good argument for the importance of post-market surveyance.

00:13:02 Speaker 2

This means that even in such extremely widespread libraries, it happens that there are errors that are only detected after a long time and that then have devastating consequences.

00:13:16 Speaker 2

And that brings me to the next sub-aspect, namely, if you look at TensorFlow, for example, there are currently about 4000 so-called open issues that have been created within the last 5 years.

00:13:31 Speaker 2

Well, the library has been around for about 5 years and in that time about 4000 open issues have been recorded.

00:13:39 Speaker 2

These are all problems that have not been solved to this day.

00:13:44 Speaker 2

And most of these problems are not a serious threat.

00:13:50 Speaker 2

Many of them have something to do with the fact that TensorFlow shows strange behavior in particularly rare cases under certain configuration conditions.

00:14:01 Speaker 2

But this shows that despite this diverse use, even a library with TensorFlow is still far from being completely error-free.

00:14:13 Speaker 1

So we have neither the excuse that the libraries would be indirectly validated with the model, nor do we have the excuse that they can say, yes, they are in use millions of times and everything is fine, we obviously have to roll up our sleeves and we have to validate.

00:14:31 Speaker 1

Yes, now we come, I think, to the interesting and essential question and how do you do that now?

00:14:36 Speaker 1

So, how do you validate a machine learning library?

00:14:41 Speaker 2

Exactly, that is actually the central question and I would like to come back to this distinction again or I have to come back to the distinction between tool validation according to ISO 13485 and subvalidation according to IEC 62304.

00:14:58 Speaker 2

I'll start with the subvalidation.

00:15:00 Speaker 2

For subvalidation, it must be shown that the inference functionality of the library works correctly.

00:15:09 Speaker 2

That means you have to show

00:15:11 Speaker 2

that the predictions made by the model correspond to its weights.

00:15:16 Speaker 2

This has nothing to do with the prediction quality of the model and this is also one of the reasons why I said earlier that this correctness check goes beyond the model evaluation.

00:15:28 Speaker 2

The point here is to show that the model makes the predictions it would have to make according to its weights.

00:15:38 Speaker 2

And these can be good predictions and they can be bad predictions.

00:15:41 Speaker 2

That depends on whether this model is well trained or not well trained.

00:15:45 Speaker 2

But that comes, that is not the consideration in this context.

00:15:50 Speaker 2

And for this verification you need 3 things: First you need a specification of the expected behavior.

00:15:57 Speaker 2

What would actually have to be predicted?

00:16:00 Speaker 2

You have to specify that.

00:16:02 Speaker 2

Secondly, you need a test oracle, which means that you have to

00:16:07 Speaker 2

be able to show what this expected behavior should actually look like.

00:16:12 Speaker 2

So, you need a comparative value to what you observe and thirdly, you need suitable test data.

00:16:20 Speaker 2

These are these 3 components that are needed for subvalidation.

00:16:25 Speaker 2

It's a bit different with tool validation, as we briefly touched on earlier.

00:16:31 Speaker 2

Tool validation can and should also be risk-based.

00:16:36 Speaker 2

This means that it is not necessary to validate the full training functionality of the library.

00:16:45 Speaker 2

That wouldn't be plausible at all, it wouldn't be practical at all, because training makes up the majority of these libraries.

00:16:53 Speaker 2

For tool validation, one can now actually use suitable techniques of model evaluation to show:

00:17:03 Speaker 2

that the result of the training corresponds to the training data.

00:17:07 Speaker 1

So, we have 2 validations that we actually have to do here now, namely we have to make sure that the library really actually trains and that afterwards, that when the model is trained, then also the Predict function, maybe if you concentrate this on that, then also the Predicted according to how the model has just been trained.

00:17:27 Speaker 1

And what makes it so difficult is that one part of the

00:17:31 Speaker 1

is ultimately regulated by the 13485 and the other part by the 62304.

00:17:37 Speaker 1

I would say that this will make some auditors sweat and will certainly lead to exciting discussions during the audit, so that one is always clear which regulation applies here now, because this demarcation curve runs through the middle of this library.

00:17:55 Speaker 1

I think that's really a special feature that we haven't had in any other way.

00:18:01 Speaker 1

But it sounds a bit like some work that manufacturers have to do here and I think that at least part of this validation is now relatively independent of the specific medical device where the model is used.

00:18:17 Speaker 1

This raises the question, does everyone have to do it again, can you somehow share these efforts, can you maybe even buy it ready-made, what would be your recommendation?

00:18:28 Speaker 2

Yes, exactly.

00:18:28 Speaker 2

Yes, thank you, thanks for the question, Christiana.

00:18:30 Speaker 2

This is actually the case that this validation means work in these 2 different ways.

00:18:39 Speaker 2

The good thing is that not everyone has to do it completely from scratch or reinvent the wheel, at least not at all.

00:18:45 Speaker 2

Then maybe some bad news.

00:18:48 Speaker 2

It's not as if you would validate a machine learning library, such as TensorFlow, for example.

00:18:56 Speaker 2

and then it's done for use as a sub or as a tool.

00:19:00 Speaker 2

That would be, that would not be practicable at all, these libraries are far too big, far too extensive, much too complex.

00:19:08 Speaker 2

In the end, that would mean doing the test activities that the Tensorflow community, for example, is doing, and making it better, and that's not realistic.

00:19:21 Speaker 2

But the good news is that you don't have to,

00:19:25 Speaker 2

that for a specific product you don't have to validate the library generically, but for a specific product you only have to show that the library works correctly for the specific model.

00:19:40 Speaker 2

This means that you have to show that this specific model has been trained correctly and you have to show that this concrete model predicts correctly.

00:19:49 Speaker 1

Because I can intervene very briefly, so that means, for example, that there is no need now if you decide on a very special type of model, no, on architecture.

00:20:00 Speaker 1

I'll say now maybe somehow a convolutional neural network with a certain architecture, you don't have to worry about the X.

00:20:07 Speaker 1

G.

00:20:07 Speaker 1

Boost variant, you don't have to worry about the normal, possibly not all variants of neural networks, but about the architecture you have chosen here.

00:20:19 Speaker 2

Exactly and even further.

00:20:20 Speaker 2

So you don't even have to take care of all the models of this architecture, but only this one concrete model.

00:20:29 Speaker 2

That means you do the tests, for example, it has comparisons with the oracles I mentioned earlier, you do them specifically for this one model.

00:20:40 Speaker 2

That sounds a bit like bad news again, because it looks very project or product-specific.

00:20:47 Speaker 2

And it is indeed the case that this has to be done anew for each model and for each project.

00:20:53 Speaker 2

But the way you do it is always the same.

00:20:57 Speaker 2

This means that you always have to make the specification of the desired behavior.

00:21:04 Speaker 2

You always need a test oracle and you always need the test data.

00:21:07 Speaker 2

And the way you create these 3 components always follows the same procedure.

00:21:15 Speaker 2

and for this we have developed a blueprint and building block that you can reuse and that you can adapt to your concrete, to your concrete models and thus to your concrete products.

00:21:27 Speaker 1

If I understand you correctly, you have now developed a process here or work instructions on how to

proceed here in concrete terms.

00:21:36 Speaker 2

Yes, but it also goes beyond the process, but we actually already have code building blocks.

00:21:45 Speaker 2

So we have building blocks that can be used to assemble test oracles, for example for neural networks.

00:21:55 Speaker 2

We have

00:21:57 Speaker 2

are generators that can be used to generate suitable test data.

00:22:01 Speaker 2

We have exemplary test validation plans, i.e. textual documents in which you describe what this validation looks like, i.e. which then summarizes and describes these components again, so to speak.

00:22:14 Speaker 2

Exactly, so we have a concept and beyond that, we already have concrete building blocks for the solution.

00:22:22 Speaker 1

What would you make now

00:22:25 Speaker 1

also recommend just as concretely what they should do now if they are already using libraries or are planning to use them.

00:22:33 Speaker 2

Yes, so for the subvalidation, as I said, these 3 components have to be created.

00:22:40 Speaker 2

You need a specification of the desired behavior, i.e. specifically that of the Predict function in the vast majority of cases.

00:22:49 Speaker 2

You need a testorrakel, you have to create it and you need suitable test data for it.

00:22:55 Speaker 2

And this test data then has to run through and control both the model and the oracle in the broadest possible way, so to speak, and trigger it.

00:23:07 Speaker 2

That's what you need for subvalidation, tool validation, because it's risk-based, it depends on

00:23:16 Speaker 2

depends more strongly on the specific product.

00:23:19 Speaker 2

In the end, however, this always means that model evaluations and model transparency measures are used to show that the model has been properly trained with the corresponding training data.

00:23:36 Speaker 1

Because that still sounds a bit like work, it would be O.

00:23:40 Speaker 1

K., if we publish your contact details there, that I

00:23:43 Speaker 1

the manufacturers can also contact you to maybe learn or maybe even reuse some things.

00:23:50 Speaker 2

Yes, of course, very much.

00:23:52 Speaker 2

You, you're right, of course, Christian, that's work, that's actually work that shouldn't be underestimated.

00:23:57 Speaker 2

But I would like to emphasize again that this is simply also an important part of the overall product.

00:24:04 Speaker 2

So these, these libraries are the essential core,

00:24:09 Speaker 2

of the model and must therefore also be carefully validated.

00:24:14 Speaker 2

And as I said, a large part of the work has already been done.

00:24:19 Speaker 2

So, we have these examples that are well reusable, that is.

00:24:24 Speaker 2

these libraries, the machine learning libraries, are a great source of reuse not only at the functional level, but also at the validation level, because this validation always looks very similar.

00:24:40 Speaker 2

And of course, we are happy to help you adapt and apply our solution to other products.

00:24:48 Speaker 1

Yes, then my recommendation would be that you might also go to the website.

00:24:52 Speaker 1

We already have a

00:24:53 Speaker 1

article, where a lot of things are described.

00:24:57 Speaker 1

We also link it again at the bottom of the description.

00:25:00 Speaker 1

And if you are looking for a shortcut and don't want to do all the work, then my recommendation would be that you simply contact my Professor Rabbit.

00:25:09 Speaker 1

Oliver, thank you from the bottom of my heart for the really very valuable insights.

00:25:14 Speaker 1

Thank you also for your assessment of my sympathetic Baden influence and in this sense see you next Züßdick.

00:25:22 Speaker 1

Thank you.

00:25:23 Speaker 2

Yes, thank you very much, dear Christian.

00:25:25 Speaker 2

It was a pleasure.