

# Artifact: “If security is required”: Engineering and Security Practices for Machine Learning-based IoT Devices

Nikhil Krishna Gopalakrishna  
Purdue University  
gopalakn@purdue.edu

Dharun Anandayuvraj  
Purdue University  
dananday@purdue.edu

Annan Detti  
Purdue University  
adetti@purdue.edu

Forrest Lee Bland  
Purdue University  
fbland@purdue.edu

Sazzadur Rahaman  
University of Arizona  
sazz@cs.arizona.edu

James C. Davis  
Purdue University  
davisjam@purdue.edu

## A APPENDIX

This appendix provides our data collection instruments, survey data, and interview questions.

### A.1 Survey instrument

Survey questions are classified based on the research questions and listed in Table 1.

### A.2 Interview protocol

The interview questions are listed in Table 2.

### A.3 Survey data

The survey data is illustrated in Figure 1.

### A.4 Interview transcripts

The interviews are transcribed in appendix A.4.1, appendix A.4.2, appendix A.4.3, and appendix A.4.4.

**Table 1: Survey Questions.**

Number	Question	Type
<b>Profile</b>		
Q1	What is (are) your current position(s)?	Multiple Choice
Q2	How many years of work experience do you have in software development?	Multiple Choice
Q3	What is the target sector of the product(s) developed?	Multiple Choice
Q4	What is the number of employees working in software engineering roles in your company?	Multiple Choice
Q5	What is the size of your typical software development team?	Multiple Choice
Q6	What is your university degree(s) in?	Multiple Choice
Q7	How many years of machine learning experience do you have in software development?	Multiple Choice
Q8	Where/how did you learn about Hardware Limitations on Edge devices used in IOT?	Multiple Choice
Q9	Where/how did you learn about Machine Learning Programming practices and developments?	Multiple Choice
<b>Theme 1 : Machine Learning for IoT devices</b>		
<b>RQ1: What are the Machine Learning Techniques, languages, tools, and best practices adopted for implementation on resource constrained edge devices?</b>		
Q10	What software development process do you follow?	Multiple Choice
Q11	Which primary languages do you use for development?	Multiple Choice
Q12	Which primary development environment/tools do you use?	Multiple Choice
Q13	What design patterns are useful when developing for edge devices? (Device Gateways, Device Wakeup Triggers, Shadow Objects, etc.)	Freeform Text
Q14	Which specific Machine Learning Algorithms do you use with consideration for the edge device it will be deployed on?	Freeform Text
Q15	For the Machine Learning Models, do you: develop models from scratch, adopt academic papers?	Multiple Choice
Q16	Which Machine learning framework do you use? (Select multiple)	Multiple Choice
Q17	How would you describe your company's maturity in terms of Machine Learning usage?	Multiple Choice
Q18	How do you validate your Machine Learning Code before deploying it to the final product?	Multiple Choice
Q19	What tools and techniques do you use to validate your machine learning models?	Multiple Choice
Q20	In what proportion of projects do you update your machine learning models in the final product using software updates?	Likert Scale
Q21	Do you collect data from the edge devices in field to improve your Machine learning model?	Likert Scale
Q22	How much data processing do you do on the edge device vs the cloud?	Multiple Choice
<b>RQ2: What are the challenges and consequences developers face due to resource limitations in developing machine learning software for edge devices?</b>		
Q23	At what point in your development cycle do you integrate your machine learning code with the rest of the embedded software?	Multiple Choice
Q24	What pruning techniques do you use for your machine learning models? (Select Multiple)	Multiple Choice
Q25	What challenges do you face when attempting to prune machine learning models?	Multiple Choice
Q26	Select the answer that is most true in your experience: choose hardware based on software or software based on hardware	Multiple Choice
<b>Theme 2: Secure IoT engineering</b>		
<b>RQ3: How do engineers incorporate security into the IoT Engineering Process</b>		
Q27	What regulations do you have to comply with during your software development process?	Multiple Choice
Q28	Describe the process your team follows for security analysis? (e.g. how do security considerations affect your Design, Review,Validation, and Maintenance stages?)	Freeform Text
Q29	Which of these tools and methodologies does your team use for security checking?	Multiple Choice
Q30	Which of these tools and methodologies do you find the most useful, and why?	Freeform Text
Q31	What aspects of security analysis do you find the mostchallenging, and why?	Freeform Text
Q32	Estimate how many CVEs has your current team dealt with during your tenure?	Multiple Choice

**Table 2: Interview Questions.**

<b>Theme 1: Machine Learning for IoT devices</b>		
<b>RQ1: What are the Machine Learning Techniques, languages, tools, and best practices adopted for implementation on resource constrained edge devices?</b>		
Q1	What development process do you follow when developing ML models for IoT devices?	
Q2	Do you use any academic papers to understand the current state of ML modeling? if so how do you identify which academic sources to use? do you implement code directly from these papers (code re-use)?	
Q3	What methods do you use to validate ML models in constrained environments?	
<b>RQ2: What are the challenges and Consequences developers face due to resource limitations in developing machine learning software for edge devices?</b>		
Q4	What challenges do you face with resource constraints when developing ML models for IoT devices?	
Q5	Do you use specific ML algorithms depending on the edge device it will be deployed on?	
<b>Theme 2: Secure IoT engineering</b>		
<b>RQ3: How do engineers incorporate security into the IoT engineering process?</b>		
Q6	Describe the process your team follows for security analysis?	
<b>RQ4: How do engineers reason about trust in their IoT systems?</b>		
Q7	What threat modeling process do you use?	
<b>RQ5: What other factors affect secure IoT engineering?</b>		
Q8	What aspect of security analysis do you find the most challenging and why?	

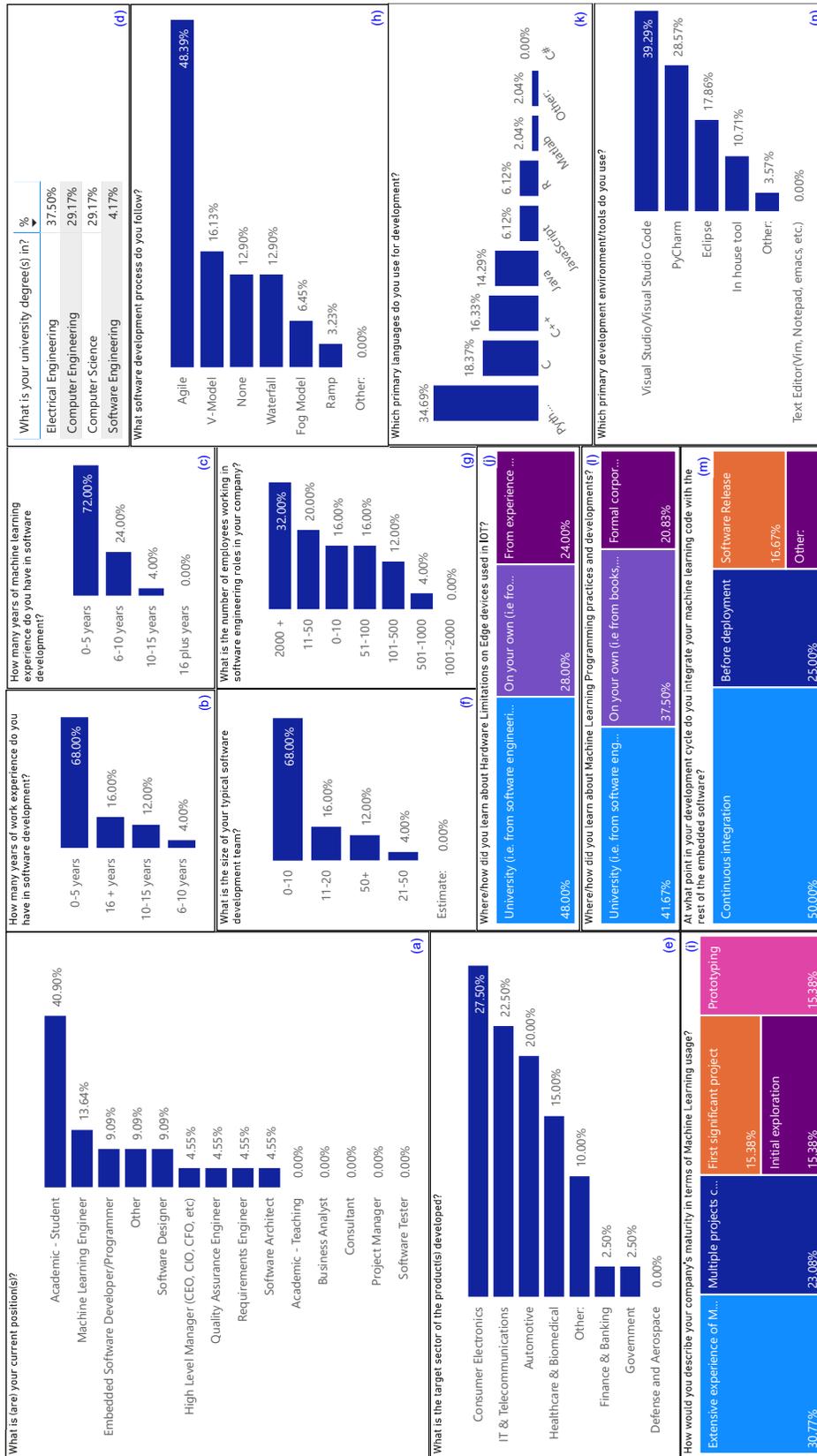


Figure 1: Survey data, part 1.



Figure 2: Survey data, part 2.

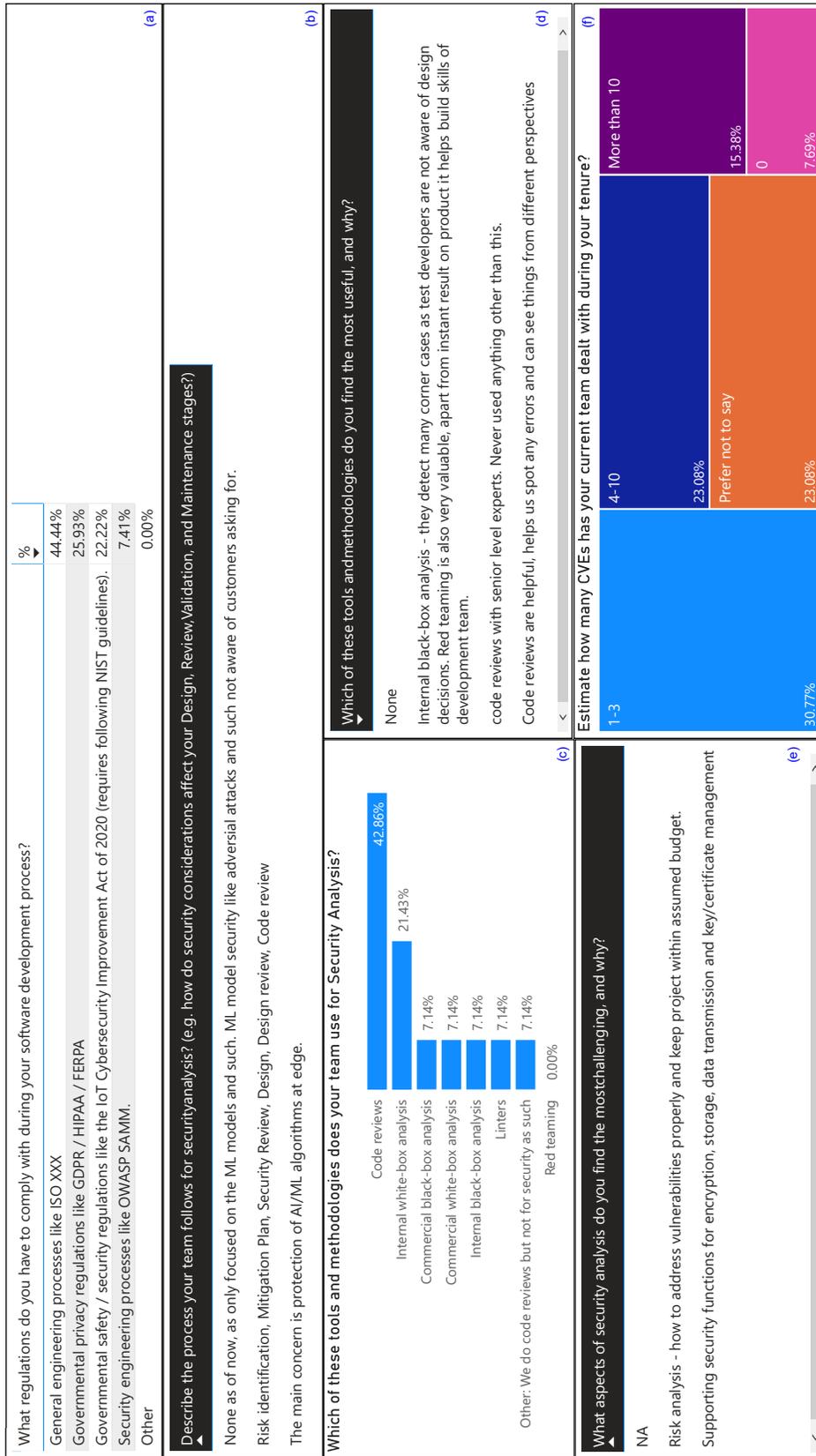


Figure 3: Survey data, part 3.

#### A.4.1 P1 interview transcript.

INTERVIEWER 1: Oh, go ahead. Yes.

P1: I'm P1. I've been working on system architecture team and from past, I would say 18 plus years, I have been ... I started working as a software engineer, mainly on the automotive chips, then slowly moved on to writing architectural models and then into an architect role. And I've been working on some parts of machine learning over the past a few years, I would say, not more than three or four years. I've been mainly focusing on trying to learn the [inaudible 00:00:57] and also try to see how best it can scale to a GPU.

INTERVIEWER 1: Okay. And you are also working at [inaudible 00:01:10], so that's good. Okay. So what we'll do today is we have ... doing a research paper on implementing machine learning, all the terms on IOT and edge devices. So that's the focus and what we are doing here is more from a software engineering perspective. See what's the landscape and study what are the best practices that have been used and how development happens in this field. Not more into the actual algorithms in itself, but more from the development. It's that focus. So what we'll do today is we have about eight or nine questions. Not too much. It should take us about 30 minutes, depending on how we go. And INTERVIEWER 2, we'll go ahead and get started with the questions and I will chime in and sit there quiet. So, INTERVIEWER 2, go ahead.

INTERVIEWER 2: Sure. Okay. So the first question is, describe the process that your team follows for security analysis. So how does security considerations affect your design review and your validation or maintenance stages?

P1: So what do you exactly mean by security in ... you mean security while developing the product or ...

INTERVIEWER 2: Yeah, Just throw at any stage of the product. So either through development or designing it or testing.

INTERVIEWER 1: And also when the product gets deployed, right. How do you ensure that you have a secure solution and you're not getting into trouble attacking them, any kinds of memberships access. Including, but not limited. [crosstalk 00:03:10]

P1: So definitely one of the things that we need to take care about is that when we release ... there are two kinds of releases that we do. One, we do an open source release, which is, there is very straightforward. We don't have to worry too much about it. What I mean by that is we have to ... any shortcomings, we get notified very quickly by the open source community and then we can, we can fix it. Of course, it's not a good thing to release something in open source, which is not adequately tested or verified. So as part of that, what we mainly do is ensure that, based on the previous occurrences of the way a security breach happens, we try to minimize as part of the design, when we go through that process.

P1: And in invalidation, we try to mimic the scenarios where we can try to see if we can reach the security. So these are the two phases where we try to do. And of course, when it's a closed source, then we have to be more careful that some of you don't accidentally release any code ... there are certain things where people can actually snoop out transactions, especially with the

denial of service attacks, kind of a thing. So we try to minimize that when we do that.

INTERVIEWER 2: Okay. That's interesting. That's kind of a good path into the next question, which is what threat modeling process do you use? So how do you determine what kinds of threats that you will face?

P1: So it mainly depends on where the application is getting deployed. Right? So I have not had too much experience with an IOT kind of a framework where we try to deploy things on an IOT kind of a framework, but mainly not on a cloud kind of framework. So we get some of the basic security we get from the provider itself. And apart from that, the only thing we try to do is, can I snoop out any data where, say, using targeted attack? So it could be something like, I try to push certain things into the system and whether it can detect that it is a malicious attempt or not.

INTERVIEWER 2: Okay. And then to kind of go off of the same topic. So what aspects of security analysis do you find to be the most challenging and why?

P1: Let me think about it a little bit.

INTERVIEWER 2: That's fine.

P1: So you're talking about anything that when we deploy any attacks on that, or ...

INTERVIEWER 2: Yeah, or even just in the overall process, so everything you were just talking about. So in any of the stages of the product, so whether it's design or review validation development, or even like you were just talking about testing it. Which area in there do you really find to be the most challenging in regards to security?

P1: I think that the testing part is the most challenging aspect of it. From a design perspective, there are certain things that we do, but did it really work or not is something very, very difficult. And there are always creative ways in which people can get through some of the aspects. Especially when handling data, it becomes more critical.

INTERVIEWER 1: I have one add-on question. So in terms of security, we have, if you broadly see, there are two kinds of attacks. One, we call us physical attacks. And the second one is the network-based attacks. You said you were working on the more cloud deployment based thing. So you will be focusing on the fact that when the data traffic that you receive from their end point, their cloud, do you consider physical attacks? Like somebody actually tapping off the physical device as a part of your security threat model. Or if it's physically compromised, it's anyways compromised, and you start more focusing at that point. You figured out that something has happened, and we start looking at ways to preempt further damage back into our neck.

P1: As I said before, most of them are machine learning applications. I'm mainly focused on the training part of it. So because of that, I'm very focused on cloud. On a physical ... say, if it's an inference kind of a thing which is running on your IOT, then I would say a lot of emphasis has to be given on the hardware aspect of it. The hardware design should definitely take care of doing a bunch of things. One very common area where you can really snoop things out is MMU. So if your MMU gets compromised, then you pretty much have physical address, and then you can do whatever you want with it. So I would say hardware design becomes more

critical when you're talking about a physical device kind of a thing.

INTERVIEWER 1: One related question to you, security comes closely and play with privacy, right? So which privacy laws do you try to comply? How do you address? Because each country seems to have made their own choices with respect to GDPR, and then the U.S. Side, we have other protocols. And how do you try to design around privacy? You can choose to answer [inaudible 00:11:13]

P1: So I don't have any specific answer for that. And I seriously don't know what is the right way to go about it. But definitely, when we collect the data, that is where a lot of privacy issues map. Once you start working with the metadata, then you don't really need a lot of private information. So some of the projects that I've worked, I really don't need any private information. I just need a lot of metadata. And then these are like a static data, right? I mean, these are not something like you're trying to capture a live location or something like that. It's not related to that. So because of that, it becomes much easier. But if your project is dependent ...

INTERVIEWER 1: Also, your focus is more on training, right?

P1: Yeah. My focus is mainly on training. And in that, I am focused on certain things, which are very, very concerned with privacy. For example, anytime I'm working with the medical data, that becomes a very, very tricky situation, unless you have set up proper working environment and you ensure that the data is not leaving your trusted network. That is utmost necessary. In this case, it becomes not just personal data, but also certain trends and things like that. So you have to be extra careful when you try to work with that data. But for most of these things, there are pretty well established on this one.

INTERVIEWER 1: Okay. So ...

INTERVIEWER 2: I have a question ... [crosstalk 00:13:46]

P1: Just give me a minute. I'm getting a call. Just give me a minute.

INTERVIEWER 2: No problem.

P1: (silence)

P1: Sorry. It was the pharmacy.

INTERVIEWER 2: It's alright.

P1: Yeah, go ahead. Ask me the question.

INTERVIEWER 2: So since you deal with training, just as kind of a security question, do you have to be concerned with having data sets that could have, or could be, malicious to the model that you're training?

P1: Yes, definitely. We have to ensure that the data that we are getting is from a trusted source also. Otherwise, it becomes a nightmare to train certain things. There have been in the past where sometimes the craning model predict some things which are not politically correct, and things like that. So we have seen that in the news in the past, right. So we have to be extra careful when we provide the training data, not just for security or even the nature of the data, but also the diversity and other things that we bring in. So it highly depends on the source of that data.

INTERVIEWER 2: Okay. Thanks. And to go onto the next question. So what development processes do you currently follow to ensure the success in your IOT edge device?

INTERVIEWER 1: So let me give you a highlight. So by means of development process, we are looking to see how many people

are agile towards waterfall model [inaudible 00:16:04] CI/CD, what kind of software development processes do you follow?

P1: Right now, most of the things are agile based development.

INTERVIEWER 1: Okay. So, INTERVIEWER 2, you want to go to the next one?

INTERVIEWER 2: Sure. So how do you identify which academic papers to use for machine learning model modeling? And do you implement code directly from academic papers into your programs, or do you reuse code from like an open source platform?

P1: So for most part, the fortunate thing for me has been that [LARGE TECH COMPANY's] research does a lot of research around machine learning. And we try to use the frameworks that is developed by them. And generally, I don't use any open source.

INTERVIEWER 2: Okay.

P1: I'll go to them or something like that.

INTERVIEWER 2: So what methods do you use to validate machine learning models and constrained environment? So if you have something on the edge and you push an update to it or something, how do you verify that that was a good update?

P1: So I cannot answer that question mainly because I don't have ...

INTERVIEWER 1: I mean, that's okay. I think, because you're not much on the inference, but more on the training. So that kind of makes sense. So, one other thing that I will still touch upon since I know you also played around with IOT systems. So what do you think is the main challenge that you face with the resource constraint device, which you have done in the past, right? So if you are to develop an algorithm that fits to a smaller resource constraint systems, if memory, bandwidth, battery, life, computational power ... what is the number one problem that you see?

INTERVIEWER 1: So I think you dropped INTERVIEWER 2.

INTERVIEWER 2: Yeah. I'm still here.

INTERVIEWER 1: Okay. I think let's wait for him to return.

P1: Are you able to hear me now?

INTERVIEWER 1: Yeah, I know you [inaudible 00:18:59].

P1: Sorry.

INTERVIEWER 1: So I was just going to repeat my question. So if you are doing on a resource constraint implementation, which do you think is the problem for you? What aspects of resource constraints, memory bandwidth? Is it computational power? Is it the participation? What is the major aspect if you are to fit the machine learning model?

P1: I would say that memory and power, these are the two main things to consider. The computation power is directly proportional to the power consumption. So we know that if you put more assignments and things like that, then you could get the compute power, but then what is your power budget that you want to get to? And if you have more compute than would your memory scale too. We have seen in the past where GBU has done, and then it's the launch latencies that cause a cause of headlock.

INTERVIEWER 1: So INTERVIEWER 2, do you want to ... I will probably go to the last question. From a research community ... I mean, for a community of researchers at Purdue or any other institutions, where do you think the research community should focus their efforts on with so much of work happening around

machine learning, IOT, edge devices. Where do you think should the next set up focus should be, in your thoughts.

P1: Very, very tough question. I would say it depends on ... There are newer and newer areas of application of machine learning. I think one of the things that [LARGE TECH COMPANY] did very recently is, in our GTC talk, there was a presentation on trying to identify man in the middle attack where the batch has already got in. And then how do you identify, in the network, which node got compromised and things like that. I think that is a very, very new field where some of the AI research is going on. And that will be, I would say, a good area. Of course, medical research has always been a good area, but with newer frontiers.

INTERVIEWER 1: Will you be able to share the GTCs and DTC? Is that open?

P1: Yeah, I can. What I can do is, I'll try to find out the block on that and I'll share it with you

INTERVIEWER 2: I watched that recently and it was really good,

INTERVIEWER 1: So, INTERVIEWER 2, do you have any other questions that you have ...

INTERVIEWER 2: I have one more question on here. I don't know, since you don't deal with IOT devices as much, but ... So this question, just asking, which specific ML algorithms do you use with consideration for the edge device will be deployed on?

P1: I don't know the exact answer for that because I don't work in that framework.

INTERVIEWER 1: Okay. That is, I think, pretty much what we wanted to cover. Thanks a lot for your time, P1, and I'll just stop the recording and then we'll stop for a couple of minutes before.

#### A.4.2 P2 interview transcript.

INTERVIEWER 1: Second for it to connect. Okay. Now we are on and rolling. Go ahead. Yeah.

P2: So I'm software engineer working on embedded devices and also machine learning algorithms, particularly speech recognition. What else? I'm in the business for more than 20 years. In the domain, I was [inaudible 00:00:36] with... Is also audio/video, instance messaging, audio/video conferences, tools, let's say, and speech to text, text to speech, video protection for a client and embedded devices. Okay [crosstalk 00:01:19]

INTERVIEWER 1: So, that's good introduction [inaudible 00:01:20]. So thanks for that. So, well for today's talk and into our short discussion, we have eight specific questions. You can go ahead and take... I think I scheduled 30 minutes, but depending on how we manage time, so how much time it takes. So I'll get started.

P2: So we are currently doing our research around fitting some software development practices, especially for IoT and Edge devices, where the resource constraints are pretty scarce. That's our focus area. So we are looking at studying the current trends and what is being done in the industry in this space, how people develop software. So most of our questions will actually be around this process. It's more from learning what has been done.

INTERVIEWER 1: So let me get started with the first question. So, the first few questions are on security. So can you describe the process your team would follow for security analysis? So how do you consider security during the design review validation at various stages of development?

P2: So, first of all, you must decide if security is required. And if required, what level. So if you push security to the level that is hard to maintain, and it's adding significant value to the [BOM 00:03:11] cost, then it is a question if it will be accepted by the market. Okay.

P2: So first, I believe, argument when we talk about the security is how much it will be visible by the user and how much it will add to the BOM cost. And if it's really necessary, what we are protecting, what are the assets that should be protected? Sometimes, simply there's a decision, we do not have assets that should be protected, or if the device is breached and the assets that can be accessed by the attacker is simply not valuable to add additional cost to the BOM cost of the device. Because it is always development costs. It is sometimes additional [inaudible 00:04:29] costs.

P2: So first of all, you need to know what you are protecting, what assets are valuable, if valuable. And simply, what is the range of the security you would like to apply? And knowing that, the whole process start. You are adding, if required, hardware too, for example, and count of cipher data, and you can apply the whole secure boot process. You can add multiple things, and simply you need to find proper balance between the cost and the level of security for your use case.

INTERVIEWER 1: Okay. So what threat modeling process do you use? [crosstalk 00:05:41].

P2: Excuse me?

INTERVIEWER 1: What threat modeling processes do you use in [crosstalk 00:05:45] a threat attack, if there is... How do you go about deciding... Let's say, you said cost was one of the things,

but the security you decide to implement, how do you determine what types of security do you think will be required as it [crosstalk 00:06:06]

P2: Let's maybe discuss about, let's say, Google and Android phone. So you can imagine that the valuable assets, that every Google Android phone has is, for example, network, single network, that can be used for speech recognition. So this network can be really extracted because it is visible to everyone as a link, because you are downloading the model. And then you are using the single network on a phone. And you can load the network and use it on a phone, but also you can download the network and create, for example, speech recognition service that is directly using that network. You are not building your own, but simply hack and use network from Google.

INTERVIEWER 1: Ah, okay. Okay. I think I got your point. So [crosstalk 00:07:30]

P2: And the asset is just the network and you can imagine that other companies that are providing the algorithms would like to protect the IP, protect fan networks, because the whole IP is just network.

INTERVIEWER 1: A little bit related to security, this is something that we were studying as part of the research. Do you have to comply with privacy, also? Privacy laws that are... How do you determine how to comply? And do you also look at privacy in different geographies and stuff?

P2: So if we are talking about the privacy, it is the hard problem. And when just to... Not be hard because it will be really visible to the market, but simply value of assets lost and protected by the law. We are trying to not store any private data that could be breached, extracted, and used by hacker in any way. So we are simply not trying to tackle such cases.

P2: And from my previous work, it was not... It was always an issue because it is a really hard problem. And it is really easy to... At one point, lose your name, lose your brand. Simply, once you are hacked, your brand will suffer. And another side of the stick, you have the cost. Simply applying additional security levels is costly. And it is not on a secure boot. It is secure storage with this sometimes secure [inaudible 00:10:45].

INTERVIEWER 1: Yeah, the whole end to end, almost.

P2: Yes, you are as secure as the weakest point, and you need to not only apply all of these security measures, but also you need to [inaudible 00:11:03] on every stage, that it is working.

INTERVIEWER 1: Okay. Okay. That's good to know. Slightly we'll change the gears from security to software development. So what development process do you currently follow when you develop software for the speech recognition [crosstalk 00:11:30] when I say development process, we are checking with folks, whether they stick [inaudible 00:11:37] agile or any such thing, and what do you follow at different stages, any light that you could shed on that.

P2: And so we are trying with building algorithms agile approach. So this will be a simplest answer.

INTERVIEWER 1: Okay. Okay. Now I will shift a little bit towards machine learning. So far machine learning, how do you identify which machine learning model to use? Do you do a lot of academic paper studies? Do you implement code from open source

or academic sources that's there? Are you completely do the research within your company? What those are...

P2: So, first of all, we are doing all kinds of papers, research, and maybe I will talk about myself.

INTERVIEWER 1: Yeah.

P2: I'm trying not to look exactly what each typology is bringing. I'm rather thinking about higher level, why this typology bringing the value. It's not... Something, for example, additional connection or hyper parameter. It's not something that is bringing my attention. I'm trying to understand where and where the additional value on previous work is coming from. And based on that knowledge, we are trying to improve our own typologies we are currently using, to also improve our results. So we are not directly using open source academic typologies. We are having our own, but we are trying to learn where we can improve our algorithms. So, it is important to [inaudible 00:14:42] all these papers. Just maybe run open source algorithms to get a reference where we are, compared to academia. But we are trying to go beyond that.

INTERVIEWER 1: Okay. Okay, good. So another question related to machine learning, how do you fit your machine learning model to the IoT or Edge devices, in this case, where the resources are pretty constrained? You don't have a large memory, could be [inaudible 00:15:28]. What process do you use if you develop one and doing a development to make sure it fits in with the...

P2: Yeah. So first of all, we need to know exactly what capabilities we have. So how much memory we have, what is the latency to the model, what is the compute resource? What are the capabilities of the compute resource? How much of the utilization we can take for our algorithms? So when we have a base information about what we can deliver.

P2: So we know exactly, for example, that we test on that particular hardware, fully connected layers, several types of convolutions. So we know what to expect. And simply, we are trying to fit our algorithm we have, for example, for big device. We are shrinking it, using the best primitives. And by best, I believe, I mean here size, efficiency of compute, and final accuracy.

P2: So knowing that, for example, and you can take a sparse, fully connected layer, but it may appear that on phone, you are getting, let's say, 70% reduction in memory footprint, the computation cost is much higher than [inaudible 00:17:32] representation because of the [inaudible 00:17:37] you operating on. And you also get, for example, small accuracy drop. And knowing that you are simply trying to find the primitives, you will build your network on point. And then knowing what primitives are available, what budget is available, the big model is shrink to that requirements.

INTERVIEWER 1: So another little bit of a related question on the validation of those models. You don't actually... Do you have the final hardware in place before you start development? Because sometimes it's not the case. So how do you go about [crosstalk 00:18:32]

P2: But you know what are the properties. You don't have hardware, but you have properties of the hardware. You know exactly where the hardware team is targeting. It's not something that you [crosstalk 00:18:47]

INTERVIEWER 1: Emulation and other vehicles to do this? Or do you just do simulation?

P2: So, it depends. I prefer Excel sheet because bringing emulator to a state that you can perform simulation is taking time. And also building machine learning algorithm is taking time. So it's better crude estimate what can be run using Excel sheet and get some [inaudible 00:19:36] and then simply prepare machine learning algorithms that simply relies on this crude estimate [inaudible 00:19:48] than waiting for simulation or emulation environment. Everything depends simply on what is available and what is the shape of this environment.

INTERVIEWER 1: Okay. That's good. So do you have a specific class of ML algorithms or ML models that you consider in your research you could share something like that?

P2: So you're thinking about use cases or particular typologies?

INTERVIEWER 1: Particular typologies.

P2: So I cannot disclose.

INTERVIEWER 1: Okay. That's fine. So, any other thoughts? Forrest, do you have any other specific questions?

INTERVIEWER 2: Nothing for me.

INTERVIEWER 1: Okay. So there any other thoughts [foreign language 00:20:43] before we conclude? I think we touched about more or less all the topics [crosstalk 00:20:49] wanted to talk about. Anything else that you would think the research community should look into, should [inaudible 00:20:57].

P2: So, first of all, when you think about embedded or Edge devices, you need to understand one major thing, cost. And whenever you think about cost, it is building hardware, enabling hardware by building [inaudible 00:21:28] software. Then you have building machine learning algorithm, and then you have deployment of the machine learning algorithm to the software or hardware. And believe me that the hardest part... sorry.

P2: Okay. So I'm going back to the cost. So I may give you an example of one SOC provider who had a working hardware, so it was even close to the selling the new SOC to the market. And they gave up on distributing the hardware because it appeared that bringing app software, in a way that it is usable by the customers, was too costly to be accepted by the SOC company.

P2: So the biggest cost currently I see is in building user-friendly environment for deployment of new algorithms. And whenever you are building new SOC, new hardware for embedded, you need to understand how many of each chips, without any change, you will be shipping to the market, so you will get a return of your investment.

P2: And if you calculate, for example, if you are selling a chip, for example, one dollar, and you would like to ship three millions of such chips, you will have three million of your revenue. And if you look how much you must pay for a single developer for one year of his work, it is, let's say, \$200,000. Okay. In California, it is, for the developer it is [crosstalk 00:24:38].

INTERVIEWER 1: I know. I know, yeah.

P2: So, you want to sell three million of chips. And 10% of your revenue will be just taken by single developer to bring up your SOC software. And do you think single developer is enough?

INTERVIEWER 1: Yep. Certainly not. Yeah.

P2: So this is the biggest, I would say, challenge of building a new embedded SOC. You need to have great scale to get a return on your investment.

INTERVIEWER 1: Yep. That's a good point. I think that's a really good insight.

P2: And whenever you are planning, and this is my first point about the security, I think security is always a cost. And you must well understand what is good enough. Simply, [inaudible 00:25:54] means that you will not sell because we will be not providing the correct software support, or you will be simply too expensive.

INTERVIEWER 1: Good. So I don't think we have further questions. This was good insight.

INTERVIEWER 1: So INTERVIEWER 2, anything else from you, before I stopped recording and have a few final thoughts.

INTERVIEWER 2: No, thanks for taking questions. I appreciate it.

INTERVIEWER 1: Thank you. So I'm going to stop the recording.

#### A.4.3 P3 interview transcript.

INTERVIEWER 1: Now you could go ahead. Thanks P3. So why don't you introduce yourself, then we can get it started.

P3: Yeah. I'm P3. I've been in the industry, I think this year I will complete about 30 years. We used to work together, INTERVIEWER 1 and myself in our old days in LARGE INDIAN COMPANY. I happened to be the Chief Architect of the [inaudible 00:00:27] group at LARGE INDIAN COMPANY. After that I went to LARGE STORAGE FIRM. I was the worldwide head of architecture for the removable products group of LARGE STORAGE FIRM, where we did all kinds of things related to wireless flash devices and things like that. Subsequently, well, the startup bug, also known as old age, hit and some of my old buddies, ANONYMIZED who happened to be INTERVIEWER 1's boss once upon a time, so we all got together and started this company, SMALL COMPANY NAME. We are in the eighth year of existence. We do a fair bit of consulting on one side. But I work on the products side of SMALL COMPANY NAME where we have a few products.

P3: One of them, a flagship component, there's some things we call [inaudible 00:01:26], which is heavily AI and computer vision based, which we use for largely anomaly detection. Originally we used to focus on the manufacturing industry, but of late we have pivoted to the infrastructure segment. So we've done all kinds of AI deployments on cloud, on standalone in-premise servers and tiny devices, including \$1 microcontrollers. So we've done all kinds of stuff of CV and AI. That's it. Short intro for myself and good to meet you guys, INTERVIEWERS 1 and 2.

INTERVIEWER 2: Thank you. If INTERVIEWER 1 doesn't have anything, we can jump into the interview questions.

P3: Please.

INTERVIEWER 2: So we're going to start with the security analysis. So describe the process your team follows for a security analysis, for example, how do security considerations affect your design review, validation and maintenance stages of your software development. By security analysis, I'm sure you know, but we mean the tools and methods your team uses to investigate your software for malware, to mitigate the risk of breach of security, maybe things like bugs and vulnerabilities in your software.

P3: So for these, we've got two categories. One is the deeply embedded systems, which is probably what you guys are interested in, and perhaps not the fully cloud based solutions.

INTERVIEWER 2: Mm-hmm (affirmative).

P3: So in the fully cloud based solutions we are largely dependent on, in some sense, the goodness of the cloud. It's almost impossible to see what Azure and AWS, et cetera, are doing under the hood. So there's a large level of dependence on their security procedures. Having said that, of course we ensure that all the TLS, SSL certificates, et cetera, are valid. Now for the embedded case it's a little different. In fact, when I say embedded, I meant the truly in-premise version. So here there are two bits to it. One is the deployment part and second is the development part. Now in the development part of it, we literally tend to use the standard tool chains that are available, now especially if you look at the ML or the deep learning world, there's a heavy dependence on opensource components, whether it is TFPF, like micro, et cetera.

P3: So here we're really not doing a whole analysis into the weaknesses of some of these tools. Simply because there's not much I can do about it with TensorFlow, as a whole there is not much you can really do about it. But where we do security analysis is when we put these things together. For example, I've got a microcontroller, I've got a collection from that microcontroller to my ... So maybe I should reverse the flow. So, let's say I start with the IDE of a particular tool, above that I have my TF or TF Lite micro or my Glow tool chain. Now, once this is ... Let's say a model or an image processing component has been generated, we have our own methods of encryption. It could be as simple as [inaudible 00:05:20] vital code, or it could be a key based encryption that we put in place to ensure that the entire pipeline that we built, even though there's heavy dependence on opensource tools is whole free, if we use a loose terminology. So independently, ITE, if it's got a hole there's nothing we can do about it.

P3: If TF has got a hole, Caffe has a hole, Torch has a hole, there is not much we can do about it. But with the components that come out of it, we ensure that we have wrappers to ensure that there is some levels of encryption unhackability before it gets into a binary that goes on to the eventual edge IoT device. When I say IoT device, it could be a Nvidia device or it could be a tiny \$1 microcontroller. So this is the method that we follow. But if the question is about, can you prove that the IDEs or the tool chains don't have bugs? There's nothing we can do about it and we are not [crosstalk 00:06:27].

INTERVIEWER 2: Yeah. That makes sense. Thank you. INTERVIEWER 1, do you have any follow-up questions on question one?

INTERVIEWER 1: No. Go ahead INTERVIEWER 2.

INTERVIEWER 2: That's good. So, we're going to talk about threat modeling in the second question. So, as you know threat modeling is the practice of identifying potential threats, and security mitigation is to protect something of value. It can be something confident, like confidential data or intellectual property. So what threat modeling process do you guys have in place?

P3: So mainly we are concerned about in-memory re-engineering, that is our biggest fear. Again, I'll go back to the embedded system deployment, which I believe is what you guys are interested in because of the IoT terminology that you used. So let's assume that we are deploying one of our AI ... For convenience I'll just call it model, typically it's not just model, also stuff in front. I mean, there's a pre-processor, there'll be crackers there'll be ... Then the model. Then there'll be a post-processor. All of this, let's say, we are wrapping together into one block. Now this block, or let's call it as a library or a binary eventually. Now, all this development happens in-house. So which means there is no question of anybody hacking into it at that point. Our biggest fear is reverse engineering or re-engineering of the algorithms that we've got once it has got deployed into a actual system.

P3: So here I'll go a little more specific. During execution, and we're talk about execution on an embedded system, the code is effectively decrypted into memory and the models are executing literally in-memory. So our biggest fear is in this memory ... There's not much we can do about this because eventually it has to execute from memory. Whether the layers and the weights, less to do with weights, but that too. Largely to do with the layer B

composition, whether somebody can do a layered decomposition after reading it from memory, this is our single biggest fear post deployment.

P3: Now, to mitigate this, as I said, we have some obfuscation techniques where the model, even though it is developed as one monolithic model, which of course during the development phase you can open Pencil Board, or Netron, or whatever it is, and you can figure out what is in there. But finally we'll obfuscate it and the reload in-memory, we kind of do it in a slightly disparate kind of fashion. Meaning if I've got a 50 layer model, these 50 layers would be sitting in blocks in various parts of memory, just to make life a little harder for somebody who wants to reverse engineer this. Then finally these blocks would be put together, like real time in-memory composed-

INTERVIEWER 2: I see.

P3: [crosstalk 00:10:00] the model executes. So this is our biggest fear. This is a few levels and measures that we take. During development we're not worried, because the walled garden in which development happens.

INTERVIEWER 2: I see. That's good. I have a follow-up question. In your team, is there a specific software team that deals only with software side of things? Sorry, what I meant is security ... A security focused team that deals with security threats and stuff, or is it everybody's job, including the developers, the security analysis?

P3: I would like to believe that all of our guys are super security aware but unfortunately [crosstalk 00:10:52]. So we would have, say, a key couple of architects who would be aware of all these holes and we do in-house reviews. Like I said, not really for the development part, because like I said, we're not too worried about it. But when this whole pipeline is put together and pre-library creation, pre-bandwidth creation, we have the intense reviews to prove that this in-memory reverse engineering is not possible. So when I say not possible, there is nothing that is impossible, given enough time somebody can do this. We just put in the effort to make it as hard as possible for somebody who spends the effort to try to do this in-memory reverse engineering.

P3: So, yeah, [crosstalk 00:11:41] your question, yeah, it's not as if every member in the team is super aware of this. People are relatively aware and they are aware of the pitfalls and the need for this. But I will say everybody is an expert, a few people would be experts [inaudible 00:11:57] review.

INTERVIEWER 2: So, the next question is still security related. We've spoken about security analysis. So what do you find the most challenging about security analysis? Maybe you have touched on this before too, but so what is the most challenging aspect of security analysis, and why?

P3: So to us it's exactly what I described. This prevention of the in-memory reverse engineering of a model is our biggest worry. But again, if I were going to go back, largely triggered by your question, are there intrinsic holes due to third party tools? As all of you guys know ML is all about third party tools from Facebook and Google. I mean, the industry wouldn't have been what it is today if these guys had not released all these nightly builds literally every single day. That very point of the tier [inaudible 00:13:14] nightly build, we have no way of validating what is

lurking in there. So, I would say that is one worry, but I would say 90% of the worry is that post-deployment hack.

INTERVIEWER 2: Got it. I'm going to move on to the next question. We're going to talk about your development processes. So, what development processes do you currently follow to ensure the software you're developing for your IoT device is successful? In other words, what does your development process look like, from starting with requirements up to the end delivery of the software for your customer or for use?

P3: So, we are ISO 9001, 2015 company. So essentially we are rigorously following the ISO format for a simple reason. I mean, one, we will keep getting audited. One, out of necessity and other as part of surveillance. So we absolutely need to be compliant to this. Now, having said that, we tend to follow the agile flow. So till maybe two years ago we used to be mostly the waterfall, the old fashioned way. But now I would say 95% of all programs that we run internally would be agile based. Again, the usual, you'll have requirements, you have user stories, you have backlogs, you've got ... And we got intense reviews, that happens.

P3: So, I would say in one word, the sheer necessity of being compliant to ISO 9001, and shows that we need to be very, very organized. So, yeah, I think that's it. But if you want any specifics, I can go into it, but maybe I'm not clear what specifics you ... If you could give me a hint as to what specifics, I can maybe go into it.

INTERVIEWER 2: I guess my main question was asking you about agile versus waterfall and stuff. But if INTERVIEWER 2 has any follow up question, he can add.

INTERVIEWER 1: No, I think we're good. I think we've got that idea of [crosstalk 00:15:46] that you do follow it. That's good enough.

INTERVIEWER 2: Cool. Good.

P3: The significant difference is the change from waterfall to agile [inaudible 00:15:56] about a couple of years ago. It's become extremely important for us as customer base increases and [crosstalk 00:16:03] are getting added, the old way is hard to manage [inaudible 00:16:08].

INTERVIEWER 2: So we're going to move on to Machine Learning modeling. So do you use any academic papers as sources for understanding current state of Machine Learning modeling? If so, how do you identify which academic papers to use for modeling and then do you implement code directly from those papers, like code reuse?

P3: Yeah, so it's all of the above. So it's everything. I mean, in fact, in the ML world, you don't read a paper every single day, you're in trouble. It's literally pretty much every member of the team. So we have a IEEE login for the company as well. So we do everything from IEEE papers, to archive papers, to the CV papers. There are a number of forums that we track. In addition to the regular publications and results that come out of Google, Facebook, Amazon, Microsoft. Essentially these four outfits are where the maximum amount of really pushing the needle action kind of happens. Of course, there are lots of startups like [inaudible 00:17:37] et cetera, where [inaudible 00:17:39]. But I would say it's a combination of IEEE papers and other publications that we track literally on a daily basis. Yeah, pretty

much every employee in the ML group, at least in my solutions group, is extremely familiar with the need to track papers.

P3: Now, to your second question, do we use a third party completely open source code? Absolutely, yes. But provided they are truly permissive licenses, like the Apache plus licenses. We are again, very, very conscious about that, so we evaluate to see if the [inaudible 00:18:24] code or the associated code that the papers reference, if they are permissive we would certainly use it. For example, majority of the backbone networks that allow very, very prevalent, is literally open source, like the MobileNets and the [inaudible 00:18:46] and the VGTs of this world. Having said that, we do a lot of modifications to this. I mean, we do heavy customization to these code bases that are out there. In addition to completely from scratch models.

P3: Like for example, for the embedded devices, really embedded devices, we've got literally three layer models to end layer lite models which are developed completely from scratch. But bigger models, there is a relatively large levels of reuse of the meta-architecture with heavy customization that we do. But the key here is these will all be ... Whichever backbone is used, that would be permissive license based.

INTERVIEWER 2: I'm going to move on to the next question. It's related to validation. So what methods do you use to validate your Machine Learning models when developing for constraint environments, such as embedded systems?

P3: So this is an interesting question. So when we say validation, and that to an embedded device, it is a combination of, of course, your usual matrices of accuracy, which would be, again the MAPs and IOUs and the top three, top five. Whichever matrices based on the kind of model that we're using, based on if there are segmentation, if you're using a detector regressor [inaudible 00:20:14]. It would have its own standard matrices. But, the matrices are well-defined, but eventually it is about the dataset on which you are claiming an F1 score, for example. So, here is the hardest topic. See, if it's a standard use case, like people detection, face detection, et cetera, your matrices of accuracy would be based on equally standard databases, like the ImageNet's and the [inaudible 00:20:47].

P3: But, in our world, it's a little tricky now, because of the industries that we work with. For example, we work heavily with the railroad industry in the US. So here you will not use a public dataset. The dataset will always be very, very closely held. So, we need to have partnerships with these companies, the US railroad outfits, and on those datasets we have our own matrices of accuracy. So accuracy, of course is the most well-known part, but on an embedded system, it is not just that, it is footprint. Footprint is our biggest problem in most of the embedded systems. In fact, and especially for video. Audio is not that hard, in terms of footprint. Trying to do, say, 98% F1 score on a difficult dataset with 100KB of RAM, and maybe 2MB of flash, that is a challenge. So the validation would include not just accuracy, but [inaudible 00:22:09] based on the megahertz or the MCPS available and the footprint available.

P3: So what we do is when we define a particular program, we will define these matrices right off the bat. Accuracy, of course, like I said is a given. FPS latency and a footprint, both in terms of flash and RAM and the megahertz part of it, we would define

it. At every stage of our agile flow we will compare as part of the so-called CI/CD flow, continuous integration, continuous development flow. We would check against these matrices and figure out how far, how close we are. So this is the validation flow, it's a very CI/CD based, completely CI/CD based validation flow where we are constantly, as part of your weekly sprints trying to meet accuracy, latency, throughput, and flash MCPS and RAM criteria. So these five or six criteria would be validation criteria for us.

INTERVIEWER 2: So, moving on to some of the challenges you face when developing for resource constraint IoT systems, what are the biggest challenges you face? I know you've already touched some of them, but things like memory, bandwidth, battery life, other computational resources. So what would you consider the biggest challenges?

P3: I would say all of the above, but battery less so, at least in the applications that we are working with. So most of them we are going after, in some sense [inaudible 00:24:04] applications where battery has not been a problem. But, yeah, it's still a problem. It's really the cost part. I mean, let's say somebody wants to deploy on a microcontroller costing \$5 versus doing it on a \$50 application processor, that puts a big constraint on MCPS and memory. So that is the biggest concern from a, how do you say, a market or what you choose perspective. But from a technical perspective, one of the biggest problems that we face is the inability of standard tools to be able to squash a model into something that fits with a push of a button. I'll give you some examples. Let's say I use a regular TF flow.

P3: I pick the latest and the greatest 2.4 lightly built TensorFlow. Then I develop my floating point [inaudible 00:25:04] model, which is, of course, your starting point of your development. Now, let's say I meet all my accuracy targets, but finally I got to squash it into this \$5 device, which has got 100KB of memory. Now, in an ideal world, I would have liked the tool chain to be able to do this with a push of a button. By which I mean, automatic quantization, quantizational training. Then in some ways in a platform agnostic fashion, if this could be fit into the device of choice. So I've started with TF, but I've chosen a TI device or an XP device. In a platform, independent fashion, if I could do the squashing easily, that would have been very, very, very valuable.

P3: However, what we have found is it's not very easy. I mean, it's actually mature as certain layers not being convertible automatically, to as bad as we having to write every layer in C before it goes into the device. So this is, from a technical perspective one of our biggest challenges. From an overall perspective, it is like I said, fitment based on cost, which will translate back into MCPS and footprint. If somebody could solve this problem, platform agnostic automatic quantization flow, where I can get a four-bit model or an eight-bit model, or even a two-bit model, push of a button, independent of the platform, I'll be thrilled. But I think we're years away from that.

INTERVIEWER 2: So what are some techniques you use to fit a large model on a small, like maybe a \$5 microcontroller? What type of techniques do you use, like pruning and stuff?

P3: I think the techniques are kind of well known. It is, of course, everything from ... In fact, the way we look at it is, first thing is you

try to get the smallest floating point model itself. In other words, layer reduction, filter size reduction, filter count reduction, as the most no-brainer things to do. When I say no-brainer, obvious but hard to do. Then the relatively easier bits, which is pruning and quantization. So all of the above. The model architecture too. The simpler thing of reduction of nodes and filters and so on and so forth. So this is a challenge. Again, tools are improving, for example, TensorFlow Lite can create quantized versions, but it has limitations on layer support for example.

P3: For example, batch norm is supported but layer norm is not supported by TensorFlow today for date quantization automatically. So there are lots of limitations based on [inaudible 00:28:13]. So, to, again, reset a bit and answer your question more precisely, the architecture choice, filter choice, layer choice, followed by the quantization and the pruning techniques [crosstalk 00:28:30].

INTERVIEWER 2: This is going to be our last question, at least for me. INTERVIEWER 1 will have follow-up questions, but as far as Machine Learning algorithms, which specific machine learning algorithms do you use for your development process, considering that you're developing for an IoT device?

P3: Maybe I haven't got the questions. So when you say algorithm, did you mean ... I mean, there are millions of algorithms, so I could go ... It will be first by saying I choose a higher level algorithm. I mean, I shouldn't even say an algorithm, an architecture. Let's say a CNN architecture versus an LSTM versus a GRU versus [crosstalk 00:29:25], right? So, that just keeps changing. I mean, there's no one size fits all thing. Our work, if you look, it is more of an application basis application way. I'll give you some examples. So let's say if it is a video based implementation on a IoT device, I would say it is largely a CNN based implementation today, largely. I mean, I would say maybe 90, 95% would still be a CNN base. However, if it's an audio application that you are sticking into an IoT device, it is mostly LSTM plus [inaudible 00:30:06] time series data. So it's LSTM plus CNN combo is what we do.

P3: So typically through one supervised ... Through GAN category, we don't see real deployments. I mean, we might use a GAN to do some kind of an augmentation/training, but a long story short, I would say, for video plus audio it's CNN plus LSTM combo, that finds [crosstalk 00:30:38] of deployments. Under that, of course there are multiple algorithms, right. I have a segmentation network which is CNN based. I can have a regressor which is CNN base. I can have a [inaudible 00:30:53] detector which is CNN based. So top level CNN, LSTM, beneath it, SegNet, classifiers and detectors and time series models. These four would be the sub-architectures today.

INTERVIEWER 2: Cool.

INTERVIEWER 1: I have just one final question. I think, P3, it was good that you covered most of the topics. I just have one final thought question. Do you choose hardware before you decide ... Let's say a customer walks in with a requirement to do something and you know you're going to use Machine Learning on the IoT. So apart from the cost consideration, do you go looking for a hardware first or do you first look at the software module and then go look for the hardware? I mean, which comes first for you?

P3: It's a good question, actually. Okay, maybe I'll give you ... If cost is not a criteria for us, an Nvidia box is our defacto choice. But cost is always a consideration. So it's like this, if cost is absolutely not a consideration, then we will ... Cost and power and space, et cetera, then the easiest thing is to deploy this on an Nvidia server class device. The next is to deploy on a Nvidia, you know the Jetson family class. These would be in the \$1,000 plus once you do a deployment. I mean, the modules are only \$100 or \$300, but once it gets into a industrial grade device it becomes \$500, \$2,000. There are a number of deployments in this category as well. However, there might be use cases, especially in audio where, as I said earlier, people want this to go into a coffee machine or a ceiling fan or things like that.

P3: These are all use cases that we're working with today. But here we cannot afford a \$1,000 of one cost of a Nvidia device. So therefore it has to be an MCU class most likely. Then the question is, do I define the software for a giant system and then try to optimize it down to this device? No, it doesn't make any sense. Off the bat, you need to know that I cannot have a 500 layer network. I can afford only maybe a 10 layer network. So it's not a question of optimizing a 500 to a 10. Off the bat we will start with a 10 layer, or a two layer, or a three layer network, knowing that I have to fit this into an MCU class device. Now, here again, we do some amount of experience based narrowing down. If I know it is an MCU class, obviously I will not be checking every single MCU on a ... There are 30 different vendors out there.

P3: We will have the usual suspects of the NXPs and the PIEs and the Silicon Labs and the Renesas of this world. We would directly pick one of these, rather than investigating all other 25 of them. Again, long story short, we don't blindly start with the software and then willy-nilly fit that into the hardware. We will make a choice of what category of hardware this might go into and then make the software architecture according to that.

INTERVIEWER 1: So, P3, thanks a lot for your time. I'm going to stop recording and then just ...

#### A.4.4 P4 interview transcript.

INTERVIEWER 1: Okay. So go ahead. P4, introduce yourself.

P4: Yeah. My name is P4 and I am an incoming M.S. student. ECE branch. I have worked in the industry for two years and from last September, I am working as a research intern. My research is mainly in the Federated Learning. Basically deep learning for the IOT devices and also in multi objective optimization, self supervised learning and all those topics. In the industry I have worked on embedded devices like deploying deep learning algorithms on embedded devices like NVIDIA Kx2 and all those platforms.

INTERVIEWER 1: Fantastic. That sounds really relevant to exactly what we are looking for, to talk to the kind of people. So go ahead INTERVIEWER 2 you want to get started with..

INTERVIEWER 2: Yeah, so it will be...I..

INTERVIEWER 1: To give you a bag of it. We have about eight questions.

P4: Okay.

INTERVIEWER 1: So we have seen it take 30 minutes. Some experience people spent about 45 minutes to an hour. But it totally depends on the answers.

P4: Okay.

INTERVIEWER 2: Well, thank you again P4 for being willing to answer our interview questions. So now I am going to go in and ask the interview questions. So the first question is going to be regarding security analysis. Can you describe the process your team follows for security analysis? For example, how does security considerations affect your design review validation and maintenance stages? And by security analysis, we mean the tools and methods your team uses to investigate your software for malware and to mitigate the risk of breach of security? Maybe things like bugs and vulnerabilities in your system.

P4: Okay. So in my experience, I was not a part of the team that worked on the security you are talking about. So when it comes to machine learning or even let's say deep learning, the main security or concern about is model security. What if someone tomorrow steals your model and they know the application. If they can get access to the device, then they can easily steal your model. That is a very most basic thing. So there are methods like encryption of weights, then there is jumbling of your weights of the network. All those kinds of things and platforms like Nvidia, they offer you, storing the model in memory space that cannot be over-written or copied by anyone. It can only be used. So such kind of things we had that is not like core security that you're talking about.

P4: So my work is only related to security for the deep learning models. So other than that, whatever I have read in the research, basically it is the various kinds of attacks that can happen on the model, adversarial attacks. No model is a foolproof, you can change the inference like the output generated by the model significantly just by manipulating an image, generating an image and sending it to the model. Those are the kinds of things that I am aware of. But since I have not worked in main security part, I don't know much about that.

INTERVIEWER 1: Okay. That is perfectly fine.

INTERVIEWER 2: So I guess the security aspect of your project is left for a different team that focuses mainly on security issues? Is that correct?

P4: Yeah. It was not like if you are talking in the context of IOT devices, it was not a pure IOT device you can say. It was connected to internet for sure, everything used to happen, but it also had a functionality where it can function without internet. So there was less risk when the device is not connected to internet [inaudible 00:04:51] other malware and all these problems.

INTERVIEWER 2: Okay. Yeah. So the second question is kind of related to this. If you have any idea how the security team or your own team, what type of threat modeling process do you use to understand the security threats you spoke about?

INTERVIEWER 1: If you want to skip the security, that's also fine. We can go towards ML if you want to. It's your choice P4.

P4: Yeah. So I may not be able to help you much with the security part, but with the research that I have done in federated learning, if you are talking in the context of IOT, you are more likely to use federated learning as a mechanism. So federated learning, basically that framework has like multiple devices. So there is a concept called Non-IID. I think you are aware of that. When you are dealing with a huge amount of data, but data distribution doesn't match from each client to client. Because your IOT device may be exposed to different kinds of data distributions at different locations. So in that case, first of all, more than security accuracy is a biggest concern. Second is a security part where in federated learning framework you exchange the parameters of your machine learning or deep learning model with the centralized server.

P4: So in that case, the threat model is like a curious server. So it is not like malicious server. So there are two concepts in the research where they use malicious server and there is a curious server. Curious server is the one which does not deviate from its given set of instructions. So here the role of the server is to just aggregate. Aggregate the different parameters coming from all the clients, which are IOT devices, and then send back the updated parameters to each device. So that is the role of the server. In that case the server will not do anything suspicious apart from aggregation. It can still infer some statistics about the data.

P4: Through the weights it can infer various statistics, but it will not do anything like attacks on the clients or anything. It will not try to manipulate the weights. Then there is a security threat model wherein there is a malicious server, which can manipulate your data, which can manipulate the parameter sent by your IOT device. So in that case, it will deviate your accuracy and all those things. It can also violate user's privacy. If you visualize the feature maps generated by these deep learning models, you can clearly get to know what is the underlying data that these IOT devices are using for training purpose.

P4: So here in the federated learning framework, the actual training happens on the IOT device. So there is only exchange of parameters. So even there is effort in direction of like you can use cryptography to encode your weights and all those things, but it is not efficient for IOT devices when they have so less compute power. There is some research going on in that direction, decentralized deep learning that I am currently working

on which uses blockchain. And coming to security, there is one more thing that I wanted to say. One second.

INTERVIEWER 2: Take your time.

P4: Yeah. The concept of differential privacy. That is one thing that is not explored as much as federated learning because of its less accuracy. So in differential privacy, what you do is before sending the model parameters from each IOT devices, they add noise to those parameters. So that the server will not be able to gauge exactly what are those parameters. Should it trust those parameters or not. But adding the noise will hamper your accuracy. So adding the noise like Gaussian noise or any noise is the basic thing we can do. But there is another like selectively adding the noise. So you can selectively add noise to only those parameters, which are very much impactful on your model output. Not adding to each and every parameters in your model. So that can boost your accuracy a little bit, but still it's a very under-explored area differential privacy. They have a theoretical guarantees, that's why it's much more preferred than federated learning which doesn't give you a theoretical guarantee's. So some work is in even in that direction also. That is all I know about the security aspect.

INTERVIEWER 2: Okay, that's very good. Actually. It was very detailed. Thank you for that. I am going to move to development processes you used. What sorts of development processes do you currently follow to develop for IOT devices? In other words, what does your development process look like from gaining their requirements for the product or for the software you are building to releasing the software for use?

P4: So generally we use agile methodology for building the software and regarding the testing part of the software, we do AB testing and all those things. Since I am a core machine learning developer, not tester on anything. I have very little idea about how they do, but AB testing is a standard one where you deploy to only few clients and see the response of it, see the experience of the client and then you deploy to the whole. It's a standard practice so that is all idea I have about that aspect of it.

INTERVIEWER 2: Okay. And then did you switch to Agile recently or has that been the main development method for..?

P4: Agile has been main.. Like since 2018 I started working. So Agile was the methodology. Basically most of the software engineering projects use agile. I worked also as a hardware engineer for IOT device. It was a medical IOT device. So in that we did not follow the agile methodology. But for software related, we used to follow agile methodology.

INTERVIEWER 2: Okay. That makes sense. I am going to move on to the next question. Do you use any academic papers to understand the current state of machine learning modeling? If so, how do you identify which academic papers to use for your models?

P4: Okay. So is your question specific to IOT or is it like in general?

INTERVIEWER 2: Yeah. That's for a machine learning development. Do you depend a lot on like academic papers to track what's going on right now? Maybe if there is a new feature introduced by researchers from different academic institutions.

P4: Yeah. We refer to papers which is published in top conferences. CVPR, NuerIPS and ICML ICLR these are the confidences that we refer to. Also we refer to the journal papers, but journal papers

generally tend to publish late. So whatever research you read in journal papers are not the latest it is like six to eight months old. Even the conferences paper published in the conferences are quite old because they publish like six months late after the author has submitted the paper. So you can even find a good papers on archive, but archive doesn't guarantee the quality of the paper.

P4: But if you could refer to any CBPR ICML papers, then I think those are very high quality papers. They even give you the code. And also in our first project that I worked on in AI was on bacteria enumeration. For that we had referred to object detection was the thing that we used. For that we had to refer to a lot of papers. That time in 2017 was the start of the project. That time Faster R-CNN was the highest performing one. So we go by which is performing as the best accuracy. Because the client requirement is like, they want at least 95% of human level accuracy. So human level accuracy was around lets say 99%. So we need to hit at least 91%- 93%. That was their requirement.

P4: So base accuracy was the first thing. Second thing was it has to run on a NVIDIA Kx2 platform. So we have to ensure that the inference time is less than 3.5 seconds. So all those things. So based on those parameters we used to explore which is the one. Maybe we can get like from 93%, we can go to 96% with a new model. After that project started, we started exploring more models seeing if we can improve the accuracy. But the timing was like seven seconds, which is quite double. So it is not a good trade-off for us. So based on those parameters, it is very specific to project, how much compute you have, because you can upgrade the underlying device and then you can even have the latest model working on it. That also works. So it depends on your project requirement generally, but initially this accuracy and timing, these are most critical things. So depending on your application you can prioritize timing or accuracy.

INTERVIEWER 2: Okay. So can we say that, I guess if I am understanding you correctly, can we say that customer requirements or the project requirements, it takes what sorts of academic papers you look for?

P4: Yeah.

INTERVIEWER 2: Okay. And then staying on academic papers, do you ever implement code directly from those academic papers into your program? I mean was it like minor modification to make it work?

P4: Yeah, generally that is the process we follow during POC- Proof Of Concept. So within like one month or 15 days, we have to demonstrate to the client that our methodology works or whatever we are taught that object detection works fine on this. So we have to demonstrate as fast as possible. So if the code is available, we just run it on our data and try to fine tune it. Do hyper parameter tuning and all those things and try to get the accuracy. So once the client accepts that, okay this is the way to go. Then we try to optimize everything. For Nvidia platform you can use Tensor RD to optimize your underlying computational graph. All those things we can do. Then you can build customized deep learning pipelines, customized augmentation strategies based on the underlying data. So almost like every decision you take after client approves the project, it is based on the data. Even, I can

talk specifically about object detection, but it is also true for any other kind of deep learning model you use.

P4: It majorly depends on.. Major chunk of the work goes into data annotation. We have seen multiple times, like I have developed 12 deep learning models. Majority of the time it is the data that solves the problem. Data as in not like adding more images or adding more data, it is the quality of the data. So you see a image that is confusing. You see a image where annotation is wrong. You see image where annotation is not done properly, and also data balance and all those. Those are normal for any deep learning algorithms. This kind of things we have seen that adding data has given us more promising results sometimes. Sometimes it hasn't. Sometimes it has backfired. So there is no concrete conclusion regarding that. Even in the literature you won't find any concrete evidence of that. So that is how we have dealt with some of these problems.

INTERVIEWER 2: Okay, perfect. I am going to move on to validations. So what methods do you use to validate your machine learning models, one developing for IOT related devices?

P4: Okay. So this may not be really for IOT devices because, the one we have worked did not actually use the non-IID data distribution. The project that I worked in the industry. It was generally like client used to have a certain portion of the data, but they used to give us the training dataset. They used to give us the validation data set, but not the test dataset. So we used to train and validate our model and validation dataset. And we used to report that accuracy. Then they used to run the inference at their end on the same device and validate if it works really well on the test dataset. So this is when you are developing each and every model. So at the end, there is reliability testing and all those things. So if a single image is passed multiple times through the device, whether can it produce consistent outputs?

P4: Theoretically it should. But because we had a optical system. So even slight changes like you turn off a LED in that optical system which used to capture this bacteria's, or there is a distortion due to the lens or anything. So the model used to give bad accuracy. So your model should be robust to such kind of things. So if your model should be robust to such kind of things, then you need to have such kind of data in your training dataset. So these are certain things that we have realized over working on that project. So it was you can say a closed environment. It is not like real world, not like autonomous car where it is exposed to very real world image. It was a closed environment and we could easily adjust lot of things, lighting, exposure, all those aspects. And the quality of the image, we could even select, we used to use 1200\*1200 size of image so that we could get a fine-grained object detection and also we used to have such kind of things.

INTERVIEWER 2: Okay. And then you mentioned you try your best to make your model as robust as possible, right. Do you mind talking about, some of the techniques you use to guarantee that your model is robust enough?

P4: Okay. So first simple thing we tried was to augmentation strategies. So normally you do flip and all these things, but you can do much more sophisticated augmentation techniques, contrast enhancement and all. Basically the understanding in the deep learning community is that if augmentation strategy is not a replication of real world scenario, then it is not useful. So that is

the understanding currently in the community. So you can't have like a high exposure image and send it to the model. Obviously the model is going to underperform, but you know that you designed the optical system so robust, so every time the device used to start, like whenever the user wanted to, it was basically used by microbiologists.

P4: So when they used to start, it used to run a set of algorithms. So I worked on mainly, calibration is one thing that I worked on, LED calibration. So that you cannot detect bacteria, there were different types of bacteria's, and you cant have a single lighting configuration for each kind of bacteria. So you need to have come up with a different kind of lighting. So some analysis was done before I joined the team, we had given, at this PWM we have to run red light or green or blue light and capture the image. Then we apply flag field normalization and all those things. So these were the kind of things that were followed initially. So whenever user turns on the device, it used to do certain checks saying that this calibration and distortion and everything, they had a threshold for all these algorithms. If we go over the threshold, then it used to inform the user that there is something wrong with the device.

P4: Another thing specifically that we added was the device used to suffer from dust generated by the, I can't mention the actual thing because it's a patented thing. But it used to have dust present on the area where we were imaging the bacteria's. So that dust was interfering with our model capability and it used to halt the device functionality. So then for that we added the dust detection kind of thing, where there were other methods that we had tried with, if the motor jams or motor draws more current, then we can assume that there is something wrong. So it could be, the first guess would be, it is the dust that is causing this. But through imaging techniques we had applied dust detection algorithm. Using image processing it was done, but it was not that efficient and robust. So we went for deep learning based algorithm. That's when I quit my job. So I did not work really much on that, but that's how we handled some of these things.

INTERVIEWER 2: Okay. Sounds good. Thank you for the detailed answer. I am going to ask you about the challenges. Some challenges you faced when developing. So what challenges do you face when you develop for resource constraint devices?

P4: Yeah. First thing is the compute. Second thing is the heating issues. Third thing is you can't really upgrade the models and you can't really upgrade the algorithm also. Once you deploy it in the real world, because the new algorithms, even though they may tend to claim that it is more accurate, but they will take a hit in the timing. So if you have a stringent requirement on the timing, then you can't really upgrade the algorithm over a period of time. What you can do is something like weight pruning and all those things. So we can do all such things, but RAM constraint, then the GPU, and then the timing are the three things we faced. And we had to do a lot of optimization to get a good result. Optimization part I can't really talk about.

P4: So yeah, if you talk about IOT devices in the research I have done, the majority of the challenges is in dealing with non IID data. So non-IID data will hamper your model. It is basically unbalanced data. Suppose you want to classify a image among 10 classes. One IOT device may have only images for like three

classes. Another IOT device may have all the image or images for each and every class. So in that case, how do you handle? That is first thing, second thing in IOT devices is the annotation part, how are you going to annotate the data? Like suppose there is a concept of Cross- Silo federated learning.

P4: So there is a good amount of literature, survey paper that was published back in 2018. They have all this new terminologies that were introduced basically for IOT devices. So Cross-Silo federated learning is when you have the number of clients within like hundred, and you are assured that you have enough compute. Then there is a generic federated learning wherein even your mobile device is a IOT device. So such kind of thing.

P4: So I worked on Cross-Silo federated learning. It was basically, to solve the problem of labeling. So user cannot, you can't ask the user to label like thousands of images for your deep learning model. They can only generate the data. So how are you going to annotate the data? So can we do that? Let's say 20% of the data is annotated. You can give some reward to the users. So then they will annotate 20% of the data. Can your model perform better? Which is 20% of the annotated data. So there you are self-supervised learning comes into the picture. So using supervised learning you can learn good amount of good weights and then you can annotate the rest 80% of the data, then train the model. So that was a concept that was used in my second research paper.

INTERVIEWER 2: Okay. Yeah. Good thing. You mentioned a supervised learning because my next question is, like what sorts of machine learning algorithms do you use what is the consideration for edge devices in your project?

P4: It will really depend on the problem at hand, right? The kind of machine learning algorithms we use. So first thing is we would not, unless it is an image or anything, we would not go with any deep learning. We go with general traditional SVMs and all those things, because they are computationally efficient. But if the problem really demands deep learning, like you want to classify, segment a medical image or anything, then we obviously have to go for this deep learning based algorithms. And there is also a lot of recent work published by Purdue itself in ICLR. So they have designed adversarial attacks on these models in federated learning setting. So you can look into that as well. That is also a good research paper. Apart from that, instead of using just a cloud, you can use fog nodes for efficient transfer of the data and all those. Those are like really communication related areas, not my area of research or anything. I am just aware that, there is some research regarding these kind of things.

INTERVIEWER 2: Okay. So it all depends on what you are doing and what you're trying to achieve with your.. So those are the questions I have. INTERVIEWER 1 probably has a follow-up question. So I will let him take over.

INTERVIEWER 1: No nothing much. It was a fantastic discussion. I was just listening to it. It's very nice to know that you have got such a vast experience at such early stage of your career. Thanks a lot for the interview with us and I will be stopping the recording and then we will talk for a couple of minutes before we drop.