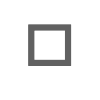
**Transcript**

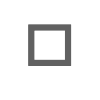
February 26, 2024, 11:03AM

 **Fabio Pierazzi** 0:10  
OK, let let's do the let me rephrase.  
Let's make it easier.  
We can record as you started now.  
But if someone feels uncomfortable, just tell me.  
I'm not.  
Just stop recording and there's a later, OK, going to be shared anywhere between us, so OK.  
So let's start with maybe June here. Yes.  
You shoot me typing?  
I'm just taking some notes for the meetings.  
Don't start to undress and there's some setting not.

 **Jun Shi** left the meeting

 **Fabio Pierazzi** 0:55  
Do you mind that if I go with someone else or we think it's?  
Yeah.  
Maybe ask you.  
OK, so do you want to start?

 **Yichen Zhou** 1:04  
OK.

 **Jun Shi** joined the meeting

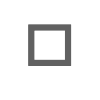
 **Fabio Pierazzi** 1:36  
Are you able to share the slides?

 **Yichen Zhou** 1:39  
Uh, yes, I'm I'm opening the file.

 **Fabio Pierazzi** 1:42  
OK.  
OK.  
Each one.  
Are you there?  
I think I have the feeling goes up.  
So you had no, no, I said that, of course I said that things will be very fine and but that's the ended up not being super fine.  
I.  
Maybe I can ask Zaid to go think this time.  
And then you, Chen would probably figure out the network issues and come back should be able to by the way, override the sharing.

 **Zaid Bin Hazzaa** 2:50  
Ah OK so.  
They told me just to share my slides.

 **Fabio Pierazzi** 2:54  
You.

 **Yichen Zhou** left the meeting

 **Zaid Bin Hazzaa** 3:14  
OK.  
Can you see the slides?

 **Fabio Pierazzi** 3:16  
Yes.

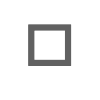
 **Zaid Bin Hazzaa** 3:18  
OK, first of all, I'm sorry it's not to present.  
It as a there's no, but I just wake up and feel like we'll cold.  
So I just prefer to stay in the flat.

 **Fabio Pierazzi** 3:30  
Oh, then yeah. No.  
OK, that's very wise then.

 **Zaid Bin Hazzaa** 3:34  
Yeah.  
Yeah.  
So sorry if all that OK can I start?

 **Fabio Pierazzi** 3:39  
Yeah.

 **Zaid Bin Hazzaa** 3:41  
Umm, OK and hello.  
In this presentation, I will just go briefly about what I did last two or three weeks.  
So I just started to do a annotated book view graphy and summarize some papers that related to my project.  
So far, the papers that I is most related to dedicated performance for model diction or some novel approach and UM some uh people use some I mean static analysis methods, some of them used anemic analysis and also some people talk about challenges, challenges to apply machine learning core.

 **Yichen Zhou** joined the meeting

 **Zaid Bin Hazzaa** 4:28  
Uh.  
Computer security and also some of the uh.  
A proposed some solution for that.  
So far I just.  
Read and some rice.  
Seven papers as following this table and I'm going to read more and try to analyze this analyze more peoples that's related to my project and the second part I want to mention about it about machine learning, class IO and that last two weeks I just load the entire that sit and do a simple classifier and also try to eliminate some feature spaces, some feature space and to classify on on this data set.  
Uh, I have some results, but I don't document it well, so I think that next meeting I will have well documented.  
Uh experiment that I can share.  
Share share with you.

 **Fabio Pierazzi** 5:37  
Mm-hmm.

 **Zaid Bin Hazzaa** 5:37  
Uh, and that's all.  
Thank you so much.  
This is the progress for the whole 2 weeks.

 **Fabio Pierazzi** 5:45  
OK.  
Thanks.  
And so I want to just mention one thing and for the.  
Just for the research review that you started, you mentioned that so far you have seven papers.  
Just I think I mentioned this to you already, but uh, of course, as a rule of thumb, but.  
In the final dissertation, you would be expected to have, like at least around 15 or 20.  
They did.  
They don't need to be all explored in depth, right?  
I mean that thing that the key point of the exercise is that you will become able to do a little bit of research yourself on related papers and maybe briefly understand their context, even from the abstract and by just skimming that right.  
So it's not expected that of course you will be 20 papers in very detail, otherwise it will take you.  
Months to probably be the understand 23% from top to bottom, especially if you're not experiencing these field, I mean but but it will take a lot of.  
I mean, if we take all the time and that's not the point.  
But the point is that to show that you can.  
Identified papers yourself and position them with your work and OK and and that's good that you've started, like playing around with the code already.  
I think it's a good day.  
More so to have something working by the next meeting, even if just very preliminary results, the earlier, I mean it's always a balance deep earlier you do experiments.  
Do you already have some results but at the same time?  
Uh, sometimes you need to read a little bit of papers and understand a bit.  
Delete ship I mean, or at least the problems called to make sure you do design experiments correctly.  
Or maybe have ideas that are valid so it's always a little bit of balance.  
I think it's probably anyway important for all of you to maybe try to do a very simple classification task and make sure it works, because that will prompt questions in your mind, right?  
So, but what is this matrix?  
What is this and that?  
And when you read the papers, you will try to look for that information after I think otherwise, the risk is that if you just start reading the papers, there is so much research happening that you get lost.  
So if you do the simple task with the code and then read it, understand better what you have done with a few lines of Python code, then you can probably feedback correctly.  
Do you have questions or?

 **Zaid Bin Hazzaa** 8:49  
Umm, no thank you.

 **Fabio Pierazzi** 8:53  
OK, cool.  
Uh.  
Maybe can I go back to June if it doesn't work for you now?

 **Zaid Bin Hazzaa** 9:00  
Yeah, just so just.

 **Fabio Pierazzi** 9:07  
Yeah, it should be fine when I think when I mean, I don't know exactly the permissions that you have, but in principle you can override each other's presentation so that if someone else does sharing a between gesture go up even if someone else will sharing.  
So I I'm sorry I've been.  
It's not even slightly.  
It's not.  
It's not like.

 **Jun Shi** 9:42  
Microphones.  
No, wait.

 **Fabio Pierazzi** 9:43  
No, wait, wait, wait, wait, wait.  
You did. You just?  
No, just because my home for the OK.  
OK, I don't have teams open here.  
Ohh, OK.  
No, no, it's it's using a camera and microphone, so I can just so they can hear you internal.  
Yes, yes.  
So hello everyone.  
So I'm Jane, so I don't know.  
I'm gonna do my presentation for my projects variance 1 permissions, features.  
You could go through the full screen for some reason, but if you want to can also just go and I mean, I mean ask you have animations, you can also go, OK.  
It's PowerPoint.  
You can also upload PowerPoint in the share button.  
Of this one?  
Yeah.  
And so you're going.  
So there's there's some background for this for this times out topic and here I also need some challenges that we may face in the future work and some challenges that we need to figure out is or maybe we need to be rates are improved.  
So so for example, like some performance and efficiency and the.  
When we do some paper research, we found that there's some papers that mention some trips or errors, and then the important thing is the it's it's uh, it's a method.  
So here is the project problem definition.  
So the first one is the transitive fication permise can be misleading by the features of the data sides, and so there's some there's a problem because, UM, thousands drift.  
And so here's some the problems related to umm, simples and the then it's the.  
It has.  
It has vacation.  
Best question and so moves to the second one and in ohh yeah appreciate based on the method may find some performance and evaluation of the it's experiments so we.  
Umm.  
Some papers, most of the papers mentioned that we need.  
We need we use the simple classifier like SVM.  
Ohm bass and some random forest.  
And then, umm, when I so in the final step for the.  
Search experiments.  
We need to figure out why the classifier do the uh do this classification for the data sets or for the malware detection.  
Uh.  
Here's the three items, the 1st and is exploring and evaluates the effectiveness of different types of machine learning August in Android mobile detection.  
And second one is verify which feature size and features more suitable for mobile classification.  
Has been creation, so for me is I need to focus on missions features and third one is is called.  
How explanation method can be better?  
Support the design and the improvement of the classification algorithms.  
So here's the separate objective.  
So the for the objectives I just.  
That's it.  
I just separates A3 amps into the small.  
As more staff, it's more objectives as weak.  
I can achieve one step one by one step.  
So.  
So I can to join so there is the I I used some expanding protocols that I will be I will use for this project.  
So for some package, at least over here, but I haven't.  
Uh, I haven't released.  
Sorry, I rest the specifications so because I I right now I couldn't control.  
Ohh no which which is supporting yours something so that I will report out the speech version that I use in the future and for some tools I using the data fail so it's also it is is this also used the drifter environment but I think it's more convenience for me.  
But then I moved to return view on some of.  
The first part is about the application.  
Some application of the.  
I'm.  
Technology in more detection so so for dripping is very frameworks.  
Research research paper so.  
And then we used the, but you use the translator TRANSLATINGS.  
But for me, for just for for me, I I'm.  
I'm I'm a little confused for for the technique and various in this paper so that I uh, I'm clicking reading trying to understand this so.  
I list some.  
I read some confusion.  
So overall this two page 2 papers to just the boundaries of the model detection and analyzing the expandability and umm then the user trust off detection system on the one hand and importance of instant expandability and some technical evaluation of the others.  
So, so future words, we need, we hope to get straight A's balance between this through directions.  
So hoping to get the good understanding.  
Uh, that's true and good.  
And his friend beauty for the model or for the experiments.  
And then it's the some challenges on that show changes in ML, more detection.  
So in this paper is being shown some.  
Some arrows or some trips that we faced during the.  
Experiment during the coding parts, some trips or the arrows. Maybe.  
Uh.  
Give you some on.  
Or and trusted results.  
So this paper has been leased the whole workflow and maybe in some workflow that's uh in in the sorry the the workflow have been separated so we can see over here in this it's a feature and then the this figure is just shown that some evaluations parts does help with cause for a taxpayer for mobile detection.  
We couldn't just.  
Only will catch the accuracy.  
Uh, so we still need to check other.  
Other parts arc I've one score so long.  
Sorry.  
Ohh and then I also mentioned that here is the some in some step is for maybe training parts.  
We can use some some technique like uh, because validation technique, technology and.  
Yeah.  
So umm, so here is for the importance of the features.  
So everybody have different features.  
So here I just find some papers for my features permissions and therefore this paper is the only for the only researching and permissions.  
So I got some I I I can read the results of these papers and it just gave me some idea or some.  
Something that I can maybe after my work come out, I can compare with this uh results and that chat my my work is good or not.  
So the final one is the installation method learning station so.  
Uh, for this suit for us.  
I think the most point most important point I wanna mention is that I found some paper use a difference on top S sorry.  
So if we use shop framework and framework and so, but we still need, we can use difference on the same functionality for the shop like for for example if we use a random forest classifier, there's some approaches for just names.  
Uh to be sharp and so.  
And if we use other ammo classifier, there's also some deep shop.  
So, UM, yeah.  
So E every time you see France are interpreters for office.  
Same XAI framework to deal with a different classifier algorithms.  
I just wanna maybe mention this.  
Yeah.  
So there's some work that have been done.  
So I have been done with the this that's working world and some basic experimental frameworks, so I'm hoping during this one month I can so I can finish the I can just at least make the my model working.  
Just let's say run running training and get some.  
We talked maybe depend the technique and standing English through some paper, and I think that's all for me.  
Thank you.  
OK.  
Thanks. So.  
First of all, I have some feedback.  
I mean it's I have some feedback on this light.  
So and it's not only for the presentation of life here, it's in general also for the final presentation that you will do.  
Minor thing and the slide numbers.  
OK, sounds stupid, but I've taken some notes about things that all you had on the slides, but I don't know which.  
Ohh yeah yeah.  
I mean I I have to explain all this light with the paper or sorry, it's on the flag.  
I can say it's like once they choose, there will eventually something and I guess make this lights way more lightweight and concise.  
There is just too much text.  
I mean the sense that.  
You have to keep in mind that let's say I can do, but it's not specifically talking about me.  
It's something that you learned, right?  
But keep in mind like I'm in 200 projects you come here, you want to make it clear to me like snaps, what are the key points like for example this slide is relatively clear having because it's like kind of concise in what's contains.  
So actually have like done and challenges and.  
You in a sense that, OK, maybe this is even too concise, but you should follow like a more lightweight approach and when it would be appropriate, maybe use some figures and I'm not using random figures.  
I'm saying like if you have small diagrams of what you want to do right and also the literature review in this meetings and needs to be just.  
Very called size.  
Intuitive, right?  
So of course there will be the actual some preliminary reporting which would write down everything.  
You don't have to sell me to go in this meeting and do too much details of all the papers you might just say, hey, the scrapers wasn't something interesting.  
I'm going to use that or maybe newspaper like there are some papers for explanations and inside them, but you don't need to do it properly.  
Sure, if you because I mean usually in research you would have reading groups, right where people meet and would only together and even just discussing 2 papers may take a full hour.  
Yeah.  
Yeah.  
So just.  
And that I'm saying this as an advice also because.  
If there is just too much information into little time, so I knew most of these papers anyway, so I could get it right, but I'm not sure.  
For example, your colleagues.  
How much of it got from this?  
Unless they read it so sometimes too much information is too much.  
So just balance that, right?  
OK.  
And this shows feedback on these slides.  
Uh, which is anyway for everyone, right?  
Because I've seen, I mean it just maybe no one told you it, so it's fine.  
No worries, just just integrate it for next time and and don't worry if someone else has probably slides for today because I know you can't fix it in one minute.  
From the.  
No.  
OK, don't have the manual coming about the libraries, but it's fine.  
I mean, it's fine that you want, of course, to keep track of all the versions.  
There's also any way a command, once you do, I think it's called Deep Freeze.  
For example, but there are ways in which, especially if you create a Python virtual environment.  
So if you don't know what it is, just look it up.  
But it's that kind of.  
I think of supply chain, by the way.  
Does it now?  
Uh allows you to do very titrated by default.  
So it's very easy.  
I basically it's just a short of.  
It's a virtual environment where you have some versions of the libraries installed, and then in one stamp you can get all the versions that even at the end it's open, use those in configuration files, but like, don't go manually checking each version of the library because if you are in a virtual environment you can just take a snap before that.  
I've resisted them.  
Which version they are anyway?  
I want to mention also a few things about the theoretical part.  
So OK, in driving there's a minor comment that you also have seen minus as an author, but it's not an author.  
Just double checking.  
There's when you do, I think it could be sometimes Googles called larger guests.  
The authors, while just double check because she mess is not an author, it's an affiliation.  
So just double check when your side them, but again this is minor. Umm.  
It's good that you're reading this much, but I feel you're also getting lost and you're also looking at very complex approaches, for example, transcending transcend.  
I think so.  
As an advice I I would not.  
Read up now.  
I would not.  
I mean, it's fine if you want.  
It's related of course, right?  
But that is very that is very, very advanced and if your project was entirely on concept drift detection and mitigation that it would make sense to focus on transcendent, even if it's complex.  
But in the context of your project, I think at the moment transcend the can only confuse you.  
I mean it's confusing also for people like for top researchers in the community, it's a some extent they feel inside.  
It is confusing.  
Also ask that where the author sometimes I mean it's not that.  
Of course, we don't know what we wrote it, just that it's based on conformal prediction theory, which we didn't propose, which formally is fairly complex.  
I mean, in practice it's easy to understand because it uses very basic algebra like very very basic stuff, but it's a little bit of.  
There's a lot of relative comparisons of classes, and so on and distance metrics, so it then become quickly confusing and I think you don't need that complexity now.  
The same goes with the advanced approaches for explanations.  
So I think you should just start with Shae shop and as an advice to all of you I would also use sharp as a library at 1st and what I mean is that the data, the mathematical details behind shape are fairly complicated in the sense that it talks a lot about the simulating distributions of a black box approach at the end of the day what you need is to know that shop is just a perturbation based approach.  
You kind of, uh.  
First of, the feature should be able to see how the output changes and from that you can infer OK, these features are more likely typical.  
For good reclassifications, these features are more typical for malware classification, but I would just use the vanilla sharp.  
And it's fine if you want to read more advanced versions of course, but I like, Prime Minister said not to necessarily other application overcomplicate what you did.  
I thought that, for example, the papers that you found on explanations for malware and when you said that you would compare against what they obtained, I think that's like more maybe volleyball and related to what you're trying to do, right, for the future importance and so on.  
I think it's also more intelligible, more immediately understandable, right?  
Uh, but it's good that you.  
It's actually good that you started looking into your own papers, and I think that the papers that you found in general are all related.  
I'm just.  
Uh.  
Maybe suggesting you to narrow down, especially because you felt confused a bit, right?  
Give that to me that so keep it down to driving and a few other papers.  
Want explanations?  
First, let's start from there and if you want, I would can also temporarily scratch the sending transcend if you want up to you also release the code publicly if you want to use it.  
But uh UM.  
I'm just, uh, suggesting to keep it for later stage.  
Maybe because otherwise it will overwhelm you, especially if you don't know how to make a simple mean from the central semi transcend.  
You need to know DREBIN by part at the minimum, so it's kind of there's a dependency there, right?  
And UM, and they think it applies a little bit to some of the comments that you have made.  
I think you had a comma in a slide.  
You had a comment about the inappropriate baselines.  
And which is something you may take into consideration when you evaluate your approaches in the sense that but before you make a claim, you may want to ask yourself, is there a simpler version for doing this?  
Uh, but in general, so don't get too much confused about it.  
Does and donts?  
Because, I mean, I think it's good that to keep them in mind when you design your experience, but I don't think you should follow them this, I mean, OK, this is I'm going to change something a bit weird maybe.  
When you design your experiments, double check your experiments with those, but don't read all of those and try to integrate all of those in your experiments.  
If it makes sense, because to do I mean to really follow all the time pitfalls.  
It takes a lot of experience and understanding of the problem right, and it will overwhelm you.  
Where as if you design a little experiment and then you ask about the is this experiment fine with respect to those pitfalls?  
It could be some cases, yes.  
Some cases no, and you just note it down.  
OK, I think it's just important that yourself aware of why you're not doing something.  
You're or there may be something bias and stuff like that, but I wouldn't start from the paper and write resign from the paper because it would be overwhelming.  
They would design something simple and then double check with the paper.  
It said I different approach right where it should make things a bit clearer.  
OK, sorry if I lost a bit of spent a bit of time on this, but I think there were some.  
There were some things I wanted to comment that maybe that should be helpful also for the others here.  
So it's as I want to spend a bit more time on it for that reason as well, so.  
Darling, move ahead to yunchuan.

 **Yichen Zhou** 32:53  
OK.  
So now my window is shared right?

 **Fabio Pierazzi** 33:09  
What?

 **Yichen Zhou** 33:11  
Is my screen share.

 **Fabio Pierazzi** 33:11  
There's no.  
Now yes.

 **Yichen Zhou** 33:20  
OK.

 **Fabio Pierazzi** 33:24  
OK, we see full screen. Yeah.

 **Yichen Zhou** 33:27  
So the project aims is, is it affective to only use UI feature set to detect malware?  
The programming language is that I'm going to use is Python And the software package include the loading function on feature loading function on GitHub and sklearn.  
And to Android allows downloading third party apps, so there's a risk of malware infection.  
Driving is a lightweight method for detecting Android malware achieved by machine learning.  
It is lightweight, it costs low time, it has low time and resource consumption.  
It is accurate.  
It has low positive rate and it is reliable.  
It has detection rate of 94% performed better than none of the 10 antivirus scanners.  
Uh jabeen performance?  
Static analysis on software.  
One of the feature sets of Japanese to detect malware to detecting malware is based on URL.  
While 60% of the cyber attacks are based on malicious URLs, where we regularly establish network connections to retrieve command to exfiltrate data collected from the device.  
Therefore, all IP addresses, host names and URLs.  
Found in the disassembled code I included in the last set of features.  
That's it.  
Thank you.

 **Fabio Pierazzi** 35:00  
Oh, OK.  
Uh, thanks you.  
China did.  
You can I ask if you uh.  
Because I remember you also sent me some uh emails about the, well, maybe it was about the project, but I replied to you also with uh the GitHub called in from some GitHub called information that is on the GitHub that I share with you all.  
Did you?  
Did you check out the repository and try to analyze?  
I don't know.  
Like, did you try to run code from the repository and load the features and stuff like that? Yeah.

 **Yichen Zhou** 35:46  
I have.  
I have a red about it, but I haven't started to run the code.

 **Fabio Pierazzi** 35:52  
OK, my advice is as I was telling what should be others, I think my advice is to.  
Try to do it as soon as possible, at least as a first run.

 **Yichen Zhou** 36:12  
OK.

 **Fabio Pierazzi** 36:12  
It will then help, I mean and then of course, this job tried to do it first to run.  
You should have the ingredients to just stack security it and make it work right.  
Of course you may.  
There may be some library setup issues that you may have, like regular programming issues, but once you run it and maybe you understand them a little bit, what the code is doing, it will then help you for the research or review and there's a saddle.  
So to I think was a Zaid, but I mean, uh, for the visual review, it's good for now.  
You're focused on driving, should understand the driving in depth.  
You are expected.  
After you also understand driving to read more papers right, and find more papers yourself by following also the guidelines in my how to review which arch or article.  
Do you have questions or?

 **Yichen Zhou** 37:21  
Uh, I have a question.  
So how is running the code related to the literature review?

 **Fabio Pierazzi** 37:28  
I think it is in the sense that when you run the code you will probably.  
So OK, unless you of course run it blindly, but if you read the code.  
You will see that maybe there is a feature about the function that maybe there's a random split function and I think.  
If you understand the steps that you need to run the code to do a malware classification, you will realize what are the parts that you need to be looking for in the papers.  
How they do things, for example?  
In the code, if there is like a linear SVM loaded and there is a random St conductor, then you would like go back to the virtual store and see whether other papers are doing a random split or the fact that a linear SVM is used with a certain parameter.  
You may then read up what that parameter means.  
You may look at other papers, whether they're doing things related to that and another thing is that if you start looking at the features in the data set, for example, after the data set is loaded and you play a little bit with the data, then when you go back to the driving paper you will make maybe more sense to try to map the features back to that force you are focused on the URLs, right?  
But there could be other features that are so in the full data set.  
There is not only one of the features, so it would be good to have an understanding of the others as well.

 **Yichen Zhou** 39:08  
Uh, sorry.  
What's the last sentence that you said?

 **Fabio Pierazzi** 39:12  
That in the data set that there are things beyond the URL, so there are also you know, activities, permissions and so on.

 **Yichen Zhou** 39:17  
Ah.

 **Fabio Pierazzi** 39:20  
So I think an idea of those would anyway help you do a comparison.  
I think that the reason that running the code will be helpful for research to review.  
Is because by running the code you will start asking and and thinking about what you need to do for the project you will start thinking about.  
Well, maybe I should read papers about what?  
Other explanations that have made, or maybe it should read papers about what other classifiers they used.  
It's maybe for things like this.  
I think the idea is that if you run the code first.  
Probably it will restrict your literature review a bit in the good way, because otherwise if you start reading papers about malware detection with machine learning, you have thousands, right?  
But it's up to you in a sense.  
If you want to do at church review beforehand, I I just think that especially if they're doing the driving paper, it's going to be very important to try to.  
Make the code run because the code is run based on the driving paper.  
The features are based on the driving paper.  
It's one of those situations in which doing the actual exercise will help you understand better.  
How they approach works?  
Sort of just reading it.  
Or should because sometimes when you read things you have an impression of understanding how they work.  
But then when you have to do it yourself or modify anything in that approach, you are lost until you try it.

 **Yichen Zhou** 41:04  
OK.

 **Fabio Pierazzi** 41:04  
Is it clear?

 **Yichen Zhou** 41:06  
Uh, I think I.

 **Fabio Pierazzi** 41:07  
I mean that did I ask her?

 **Yichen Zhou** 41:09  
Yes.

 **Fabio Pierazzi** 41:14  
OK, cool.  
So do you have other questions that for now?

 **Yichen Zhou** 41:19  
No, no.

 **Fabio Pierazzi** 41:20  
Alright, cool.  
And I guess all we we can move to changjoon.

 **Changjoon Park** 41:29  
OK.  
Umm.  
Then can I share my PPT slides OK?

 **Fabio Pierazzi** 41:34  
Yes.

 **Changjoon Park** 42:07  
E can you see my slides?  
Can you see my the the the shared the shared windows?

 **Fabio Pierazzi** 42:17  
I'm seeing everything white.

 **Changjoon Park** 42:22  
Of that.  
OK, this is it.  
Let me share my screen just.

 **Fabio Pierazzi** 42:31  
If it's, if it's in PowerPoint, the.  
You click the share button.  
If you upload the slides, it should share and now I see this could OK now I see the slides, yes.

 **Changjoon Park** 42:44  
OK.  
Umm this then.  
So let me get started.  
OK.  
I'm yeah, it's actually I'm.  
I saw it introduce you guys to the last time.  
I am responsible to a research on the API, but which is a very nice too.  
And uh, let me start with the perspective of your gains and to resume our this project.  
So actually I'm in this project, at least the the part that I am responsible.  
And I want to introduce the these point am.  
It's parted at the Shorty.  
This advantage of the.  
Which is the which is the one of the way which is the main way amend the methodologies that Robbins this adopted for their the machine learnings so.  
Umm, I spotted there's this fantasy that I that party did.  
Disadvantages said actually these these disadvantages.  
This is kind of the inheritance limitations so.  
My research goal is to overcome these static analysis.  
Umm.  
So in order to enhance the capabilities of starting malware analysis i.e.  
I am planning to integrate API usage pattern analysis, yes covering gaps but covering gaps between these aesthetic analysis and the dynamic analysis.  
So the main problem is, as I mentioned, it's just ago the study guide, static analysis and in terms of malware analysis, malware, malware detection is straightforward, easy and effective.  
However, it has a lot of disadvantages when it comes to detection rate or the accuracy.  
Compared to the dynamic dynamic or not dynamic analysis?  
Yeah.  
For example, the A many of dynamic and many of them malwares are using the methodology that they they are the they are men losing which is which is which functions malware which is malfunctions.  
Implemented our our functioned during the run times.  
But these static analysis, the drive-in trappings?  
Umm.  
Courtney, Ethan.  
Enabled to catch this malware at dynamic market dynamic malware analysis methodologies.  
So my aim is to umm, I understand the the procedures and procedures and these malware behaviors when it comes to the AK Course and the detect.  
I'm the API call the patterns.  
So in order to in order to approach this methodology, there are some key objectives and first of all, we need to call the cutter look catalog Android iPad calls frequently used or manipulated by malwares and we detect the patterns and all.  
I I'm I'm playing the detect we need to detect the patterns and umm yeah.  
But detect patterns and guys, we're gonna log that as we, of course, we've got already.  
You've got already the.  
The data sets and which has which includes API calls already, but I've I've found I found out there are at the the API calls included in the our data set which never whose name of the apps.  
But as we already know, UMI. Yeah.  
Is it umm, we can identify as I I found that, uh, I I found out that I those API calls can be identified as kind of a patterns, patterns of malwares patterns.  
So we're gonna detect and design A classifier to along the detect and sort of like for example, we can't put a kind of weight.  
Umm, according to the the patterns or the procedures of Marvis.  
So.  
In order to do that, we need the tools and equipments.  
Yeah.  
For example, I am planning to.  
I am planning to on the used the Android studios as a as our tools because we need to understand the how malware we need to understand how malware use the API calls.  
Uh to implement a model malfunctions.  
In order to do that, if we if we got, if we can obtains, if we can get the source code, can you can you can look into the source code.  
Umm, this understanding job?  
Job approval languages.  
Or if we can't, we can use.  
We can use disassemblers like cotters which which I'm which I'm gonna introduce the lighter at the latest slides called this or other disassemblers and by understanding assembler languages.  
So this is the the Android studio.  
And yet the cultures and yeah, it's Curtis.  
Here is one of the reverse engineering tools of course.  
Ohh it would be nice if you can get if we could get the source code model source code itself, but we can we can analyze the procedures how the malware, how the malware implement their logics to do the malformations.  
Just just look at their source codes, but if we can't get the source code itself, we we should the disassemble the source codes and look into the assembler codes.  
Of course.  
Umm these these stage my my sounds are a little bit complicated so I might I might not go in with these methodologies as what but we need to consider this methodology as well.  
So after after looking into uh, after finding a figuring out all the methodologies and we're gonna we we have to the figure out we have to the at this school we have this code it's a precision procedures and patterns in order to generate our the machine learning models.  
Uh.  
Rebuilding our machine learning models.  
Of course, we are gonna use the the the Python packages like tensor flow and Python which. Yeah.  
Yeah, this is the the disassembler called Cutter this there are some disassemblers, there are other disassemble disassemblers like the ID, ID, pros But these these these disassembler is is a free and these open source so every so everyone can use the freely no cost.  
So.  
So we've got, uh.  
And so a phone a for now.  
And so far we got.  
Yeah.  
So I'm sorry.  
I'm got uh, the patterns of the patterns or procedures of malwares and we caught AD detector set.  
So now, but there are some challenges we can we might encounters during the course during these procedures.  
So first of all, umm, actually, you know the most of most of malwares are using at office cations in order to hide their logics and they are malfunction logics.  
So you know, in order to look into these in order to detect these obfuscation, of course we use the.  
We are gonna detect it.  
We are.  
We are gonna detect the other all the API calls, API calls and all of them is load class which is mentioned in the driving paper.  
But as I as far as I know that is not included in our data set.  
That was my that that is one of had given data set.  
I will given data set, you know unloaded data days.  
So the which will be.  
What about one of my challenges, which will be we might be one of my challenges and other challenges might be, yeah, that we can't we we are hardly at the distinguishing my malicious form from the legitimate legitimate.  
I can use this.  
Yeah.  
And so on.  
So my future works.  
So to sum up, with my future work is going to be, uh, incorporate a dynamic analysis so as as possible as I can.  
Of course it's it's.  
It's really hard work to incorporate the static analysis and dynamic analysis, but I E we I personally I posted I planning to umm to incorporate these two, these two the the vantage is from these two methodologies by the the analysing patterns or patterns of the API use patterns as is put the weight in each patterns or its apex calls called based on our data sets that that's umm there's will be my future work.  
Yeah, that's all.  
Thank you.

 **Fabio Pierazzi** 53:59  
OK.  
Thanks.  
Uh Chang, John umm so.  
I think so.  
You are buying to do way too much.

 **Changjoon Park** 54:17  
Yeah.

 **Fabio Pierazzi** 54:17  
I think, uh, I think even if you were just full time on this.  
For one year, no other exams, no other commitment.  
It will be.  
Very difficult to do all that you have outlined.  
And there are also some key misconceptions.  
Umm.  
For example, I think that so Android does not have binaries, so you cannot even use cut there because Android.  
As bytecode, which is a bit different and you can dip compile it to source code.  
So it's not like Windows binaries.

 **Changjoon Park** 55:11  
Ohh.  
I mean, I was.  
I'm sorry.  
I was meeting like, Umm the the job, the bytecode.  
Yeah, we can.  
We can add the the the code anyways.  
Actually this is not important, but I was planning to if if I need to look into those sampling code.  
I was meaning to the Java, Java this machine code.  
The object code Java code.

 **Fabio Pierazzi** 55:35  
But but in principle I mean.  
The Assembly should only be in the native libraries if there is any, which is a different type of problem in Android apps that you may be integrating some libraries compiled from C in assembly, but that's the corner case problem.  
I mean my.  
Understanding of what you presented is that you probably started reading the literature and you tried taking a little bit here and there and you want it to address everything.  
I think you're also mixing up, uh, dynamic analysis with symbolic execution with program analysis.

 **Changjoon Park** 56:23  
Yeah.

 **Fabio Pierazzi** 56:24  
Let me take a step back.  
So at the moment you also don't even have the.  
Umm short applications for like 2 main reasons.  
Uh.  
While one is a also security like you don't necessarily.  
I mean, you don't want to execute malware on your own laptops.  
I mean or show the research that you're doing also for extracting the features is in an Azure later environment.  
And but secondly is also a size problem, right?  
Would just be too large.  
It will be like I think that the the apps corresponding to the data they preprocessed dataset I shared with you will be the order of terabytes.  
Which will also be a bit complicated to share for money, right?  
My advice?  
Is that you can keep a note of all these ideas in mind when I would go back to the simple route of the project description.  
In the sense that the very root of your project description is that if you read the driving paper, you are given the driving preprocessed features.  
So you already have a preprocessing of the APIs appearing in the apps and you compare using API only versus using all the features.  
I would anyway advise to do that as a first step because it's much much easier than any of the other path you were exploring here.  
And if you can do that quickly, I think you can then move to something advanced afterwards and we can maybe discuss in the next group meetings as well about diving deeper into one of these techniques.  
My concern is just that even if let's say that you do Even so static analysis is like means any form of code analysis where you.  
One thing that you mentioned at some point, maybe think about symbolic execution.  
Sorry, I mean I I know I interrupted my sentence, but I was trying to formulate my thoughts, but one thing that you mentioned at some point made me think about symbolic execution because you were talking about relationship with APIs and how they may result in execution patterns.  
Just to clarify, dynamic analysis is when you actually run the apps.  
You have logs and you analyze the logs.  
Static analysis is whatever you do without executing the code and in symbolic execution you pretend that input some values.  
So you look at the code, if there is an if statement, you make an assumption.  
OK, let's say that this is the condition for this if statement is true upon this set of variables and and you can go on but and then you analyze the offering that way to see what execution patterns may end up.  
You may may happen without really executing the app.  
It's all kind of the prophetical analysis and uh, but even doing a project on symbolic execution alone, even doing symbolic and section for about 10 apps.  
So it's going to be an enormous amount of work because these techniques are fairly complex.  
Uh, unless you know them already.  
And I think I would not be feasible in the time of the project to get something meaningful as an output as well, because if you want to employ symbolic execution for malware detection.  
Or even thermic analysis will become a bit too much, I think, yeah.  
So I would not necessarily focus on the limitations of the API call set first.  
The very first I would just look at the code that I share with you.  
Try to wise that idapi features and compare the API feature results with the uh, other.  
With using all the features sense right?  
So you use either the whole feature set or the API is only and you look at the malware detection accuracy performance because that is also using a very simple machine learning model which is a linear SVM.  
And once you have done that, maybe you can then think of how you would expand it.  
I mean, if you want to do.  
Some code analysis of some case studies for explanation.  
We can explore that I can share some individual apps, but you need to take also precautions because I know it's unlikely that it executes anything, but there are some.  
There are some Android malware apps that app specifically designed to be triggered on Windows, even if it's not mint, there are some more apps that are specifically designed to be triggered if you analyze them on a laptop, right, for example.  
So I mean, we need to take care of some precautions if you want to analyze the code, but I think there needs to be a clear objective if you want to like if you have a vanilla version of doing the classification with API features versus the other features which are preprocessed in that they gave you and you do some explanation on top of that, then I think that for example could analysis could be one way of doing.  
It in more in depth explanation for example or but but but but you need some preliminary results anyway to decide that I feel because otherwise if you dive into one of the topics that you mentioned, I think it could be a rabbit hole where you there is there risk that you put in a lot of work but you don't come out with a report.

 **Changjoon Park** 1:02:07  
Yeah.

 **Fabio Pierazzi** 1:02:25  
So in a sense, my advice is for you to have a a possible dissertation together, right?  
And by the way, it's also kind of a double check, right?  
Maybe you're really interested in those topics, so in a sense, my challenge here is like, OK, then it's quite to solve what I gave you as an initial sketch quickly but quickly, but well.

 **Changjoon Park** 1:02:36  
OK.

 **Fabio Pierazzi** 1:02:49  
And if you feel comfortable with that, I think we can explore more of these other techniques.  
We wanted to explore because if you struggle with what I gave you, I think it will be even it will be just much, much harder with the techniques that you have described.  
It won't be maybe harder to return a single app, but it would be definitely harder to automate and scale it to a data set, both from an engineering and theoretical perspective.

 **Changjoon Park** 1:03:19  
OK. Yeah.

 **Fabio Pierazzi** 1:03:19  
I mean, I I don't.  
I don't want to throw you off, right?  
I think I'm here to just say it's greater to want to do so much, but I'm here also I guess to put in a kind of.  
Sort of.  
Period.  
I don't.  
I don't like calling it a reality track, but it's more like a supervised many students.  
I know I have worked with these things.  
I know that they're very challenging, so my advice is first to get the mini prototype that can become a dissertation and then we can have a discussion on whether you want to dive deeper into maybe starting to dynamic analysis.  
But I mean, I'm just.  
I'm talking from experience, right?  
And I don't want to discourage you.  
I'm it's really just advice for you to succeed, right?  
I would say let's be on the safe side.  
Do first, there seem like the compared to what you proposed like this simple task that I told you at the beginning of at the at the beginning of the project.  
Well, as the project description, if you want to call it and that's first showed that and from there I think you can then expand in different areas and can then discuss that, right.  
But I I would start from that because with that you would have a dissertation anyway.  
But if you start diving in these techniques, it will become very very complex.  
So I think unless you have a lot of expertise on these topics, but from your presentation at least I felt that you had it's bit of confusion about around some key causes because maybe it's the first time you are approaching this topic.  
So that's why we have my advice comes from.

 **Changjoon Park** 1:05:09  
But thank you.  
I yeah.  
Yeah, I think I was too ambitious.  
Yeah, but yeah, let me.  
Yeah.  
Then just finish it.  
The pre preliminary task that you gave me and it is the yeah, OK.

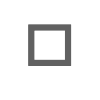
 **Fabio Pierazzi** 1:05:25  
OK, you know, but I mean and then I'll want to stop your ambition, right?  
I'm just here to resize it so that, uh, you can you can succeed so that you can succeed in the problem because you will have also a lot of other things going on in the middle, right.  
You will have the main exams and the courseworks and so on, right?  
So I'm also conscious of the limited time that you all have to work on the project in a sense, until some work well was after they make some.  
So I guess you will be full time on this, but.  
OK, so I I know where you're in a bit over time, but I'll was nice to see you all.

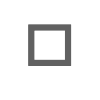
 **Changjoon Park** 1:05:56  
OK.

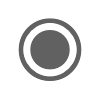
 **Fabio Pierazzi** 1:06:02  
Thanks for coming.  
Yeah.  
Further questions that we didn't cover in the meeting, feel free to follow up by email, otherwise I'll see you next time, which should be in the calendar.  
I just don't remember when or where but.  
It's probably March 25th.  
They were the 30 give meeting, yes.  
Again, in this meeting room.  
But if you want to join online at this stage it's fine.  
OK.  
Bye. Bye, everyone.  
See you later.

 **Changjoon Park** 1:06:37  
A copy.  
Thank you.

 **Zaid Bin Hazzaa** 1:06:38  
I think it's actually.

 **Yichen Zhou** left the meeting

 **Changjoon Park** left the meeting

 **Changjoon Park** stopped transcription