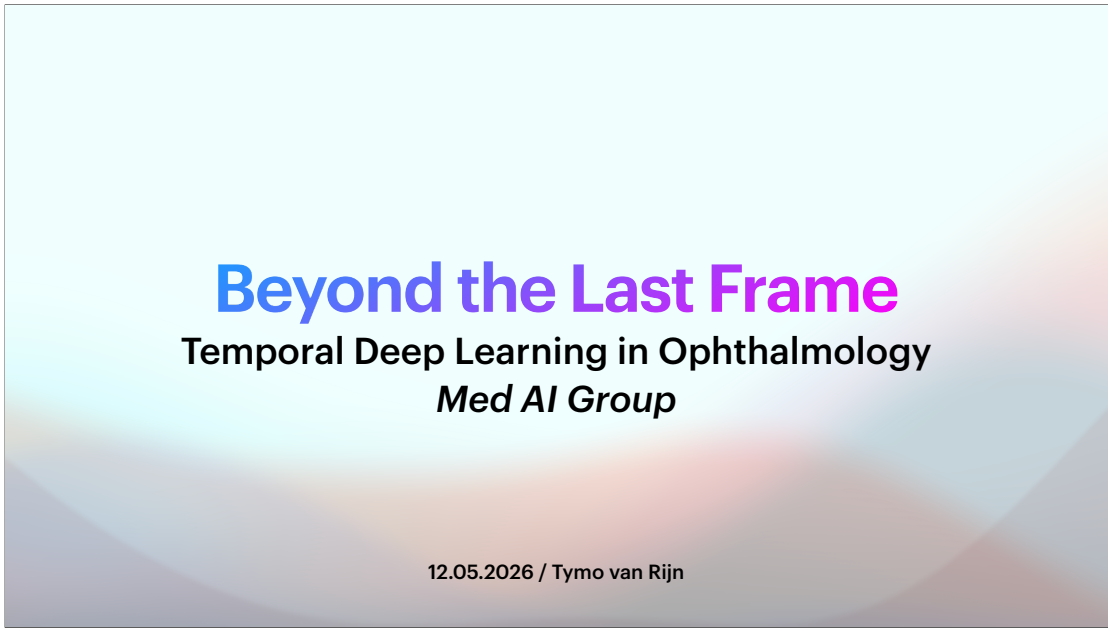
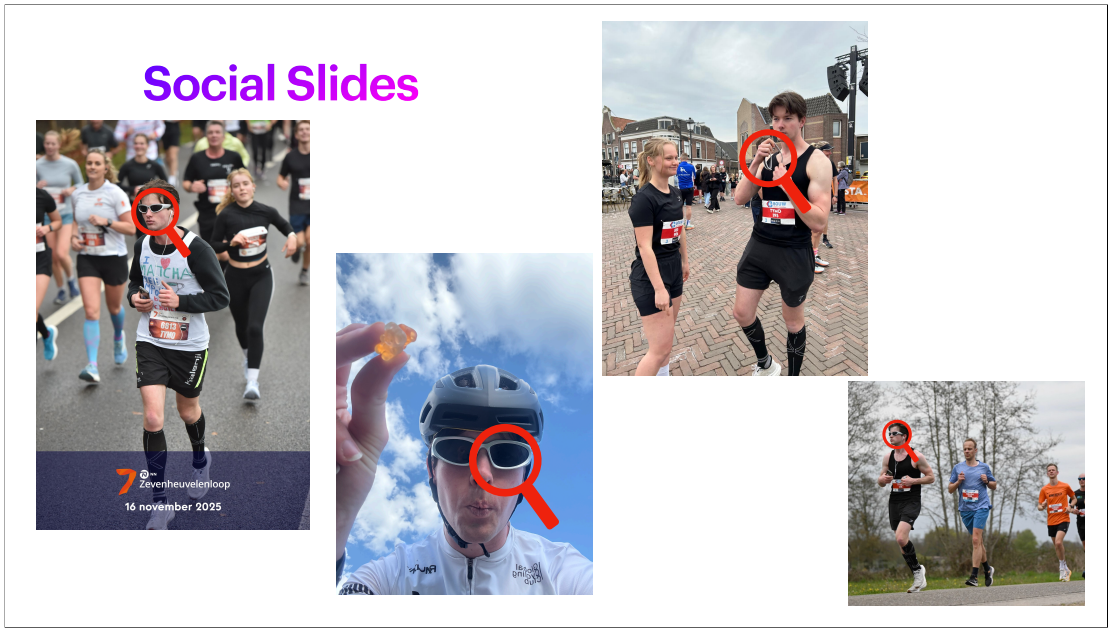


Hello everyone, welcome to my presentation about the study called: Beyond the Last Frame

Social Slides

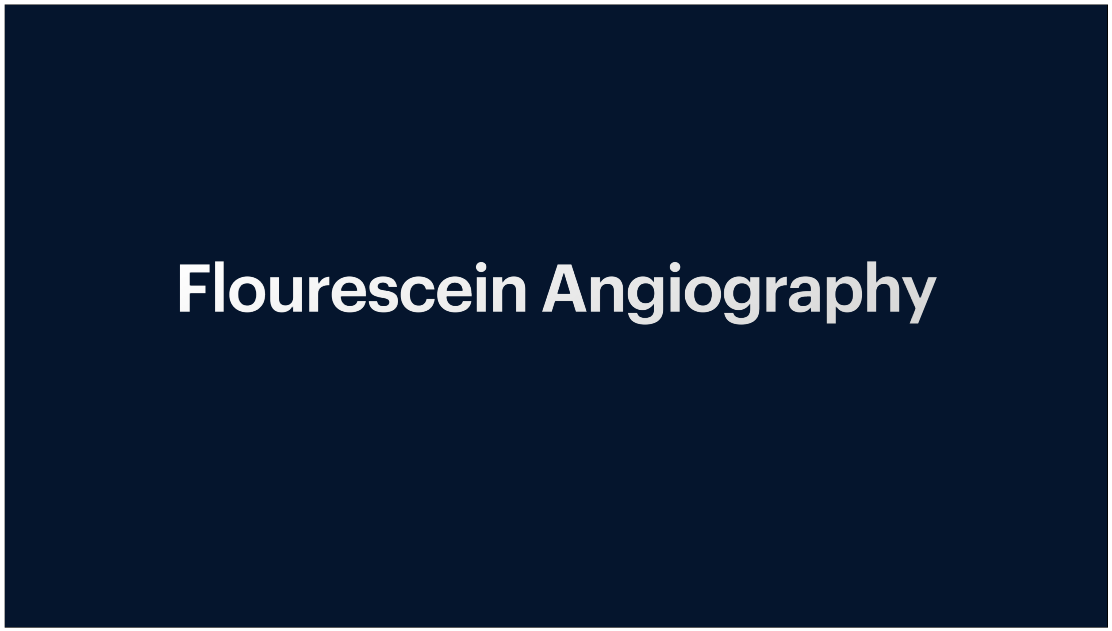
HOGESCHOOL ROTTERDAM



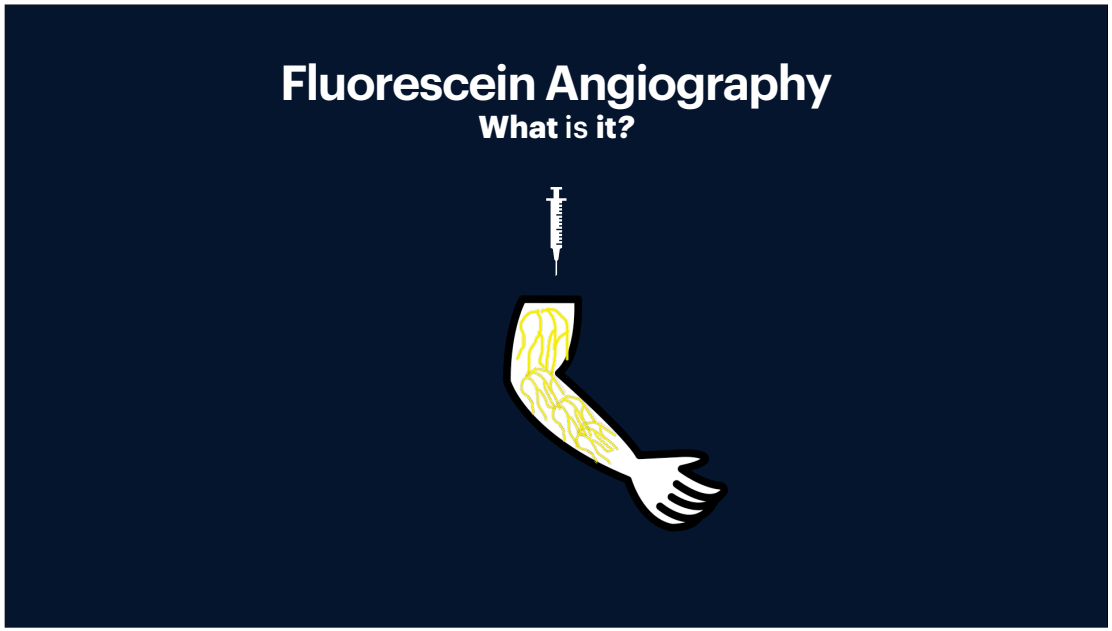
And now, I have a very interesting question for you all, since pattern recognition is a big deal in this field of work, what kind of pattern are you able to see going on here?

The glasses! Now you might ask why use the glasses? I honestly don’t know, they were 2 euros and they don’t work but they look really cool. They don’t even offer protection for my eyes, but luckily my eyes are still healthy.

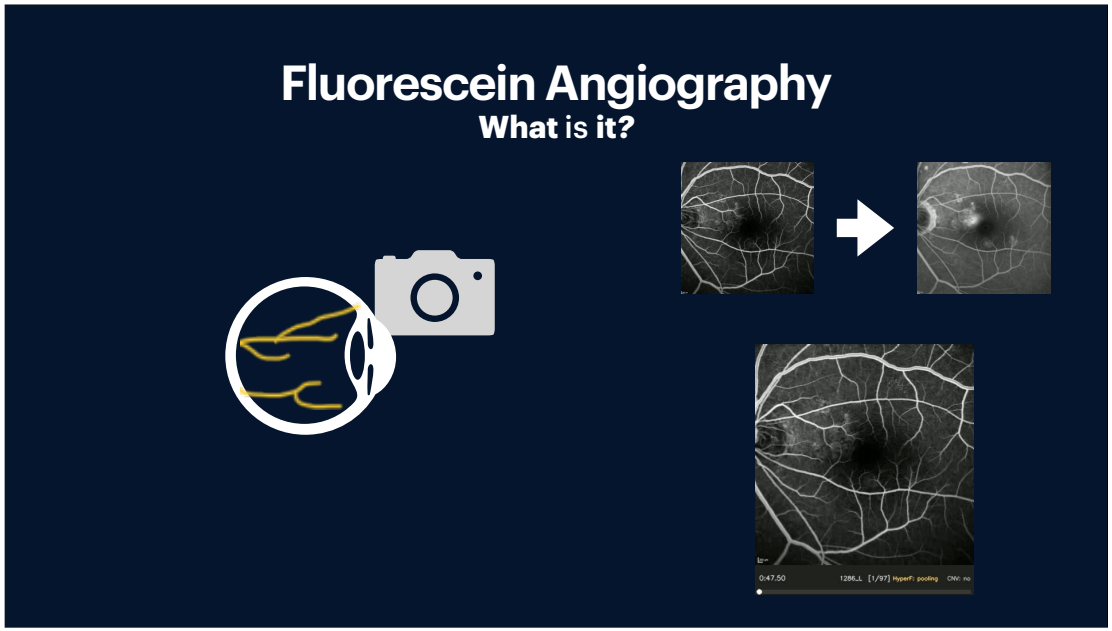
But not everyone has the privilege of healthy eyes, and one way to find out of your eyes are actually healthy is by undergoing…



When things go wrong with our vision, doctors need to look deep inside the eye to investigate eye health. And to do that, they use a fascinating procedure called a Fluorescein Angiography exam.

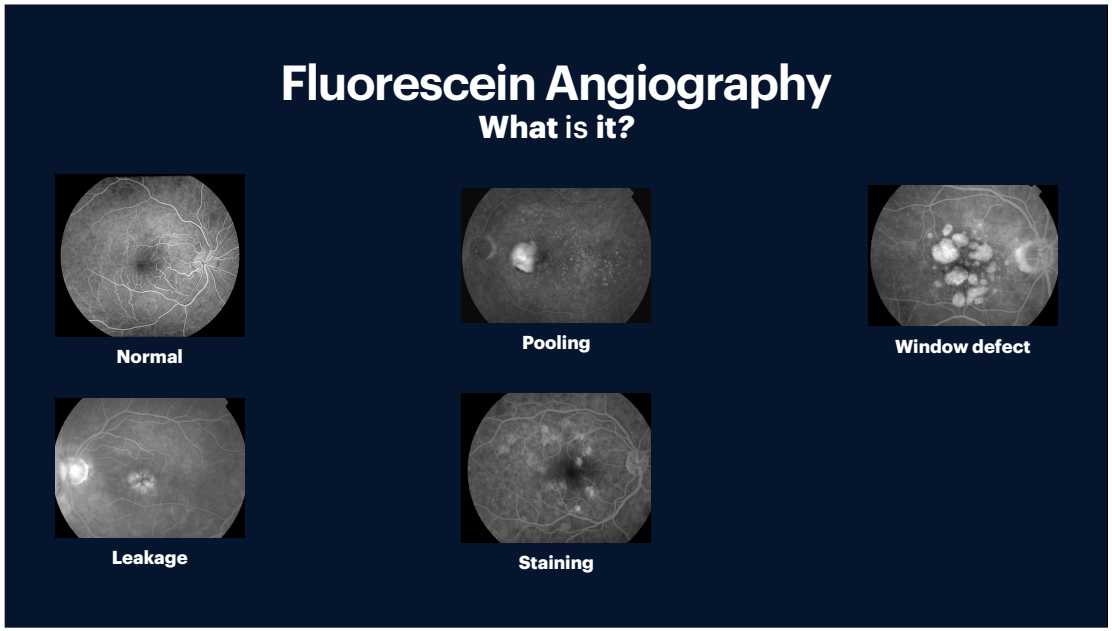


So what actually is FA? I will try to explain as simple as possible.
The eye doctor AKA ophthalmologist starts the examination by injecting a special dye into the bloodstream of the patient.
The dye travels from the arm to the eye in seconds. It’s a dynamic process. It doesn't just 'appear'; it flows, it leaks, and sometimes, it lingers.

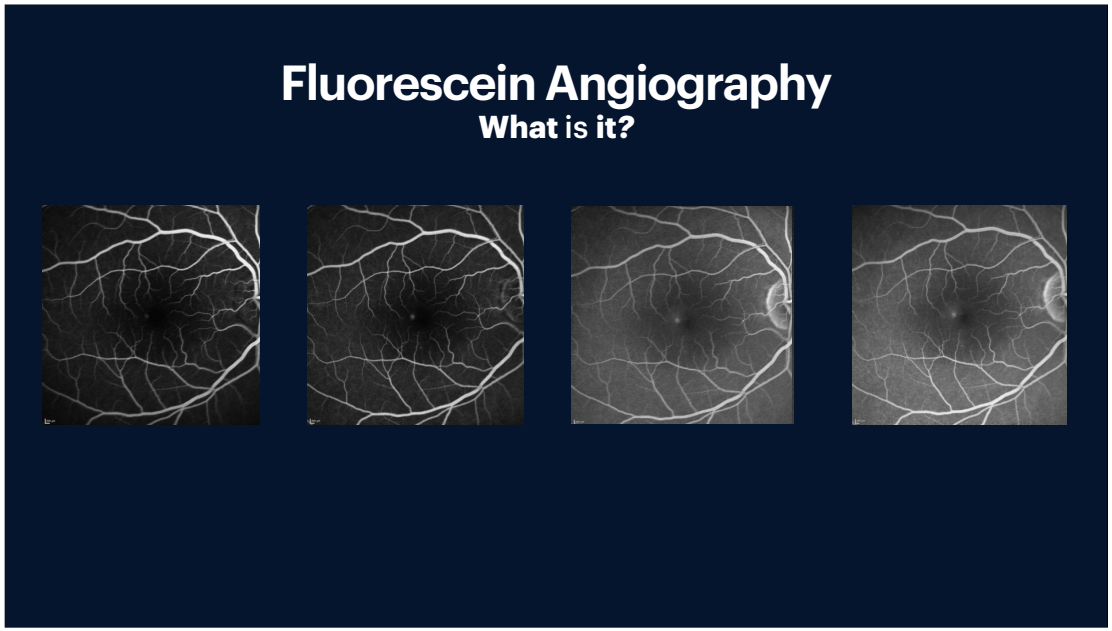


As the dye moves through the eye, the ophthalmologist takes a sequence of photos. We're looking for "Hyperfluorescence" — areas that are too bright — which can indicate various things from fluid pooling to structural staining.

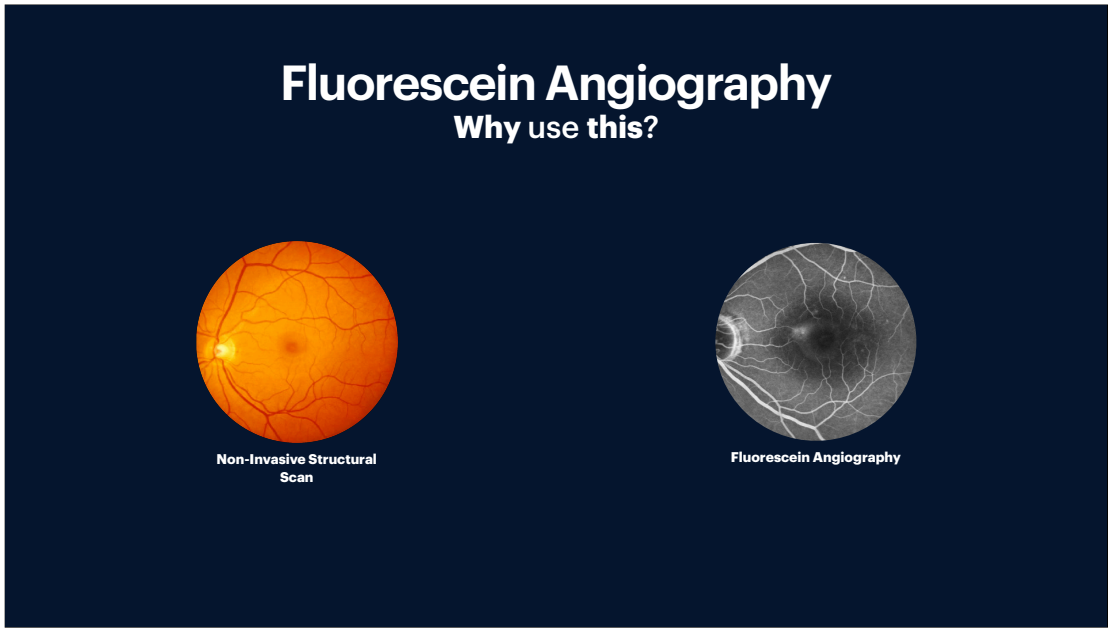
To give an idea of what it looks like, this is a first picture of an examination and this is the last one. I also made a timelapse to show how the dye evolves over time.



Ophthalmologists look for these patterns. To a human expert, the time it takes for the features to appear is a massive clue.



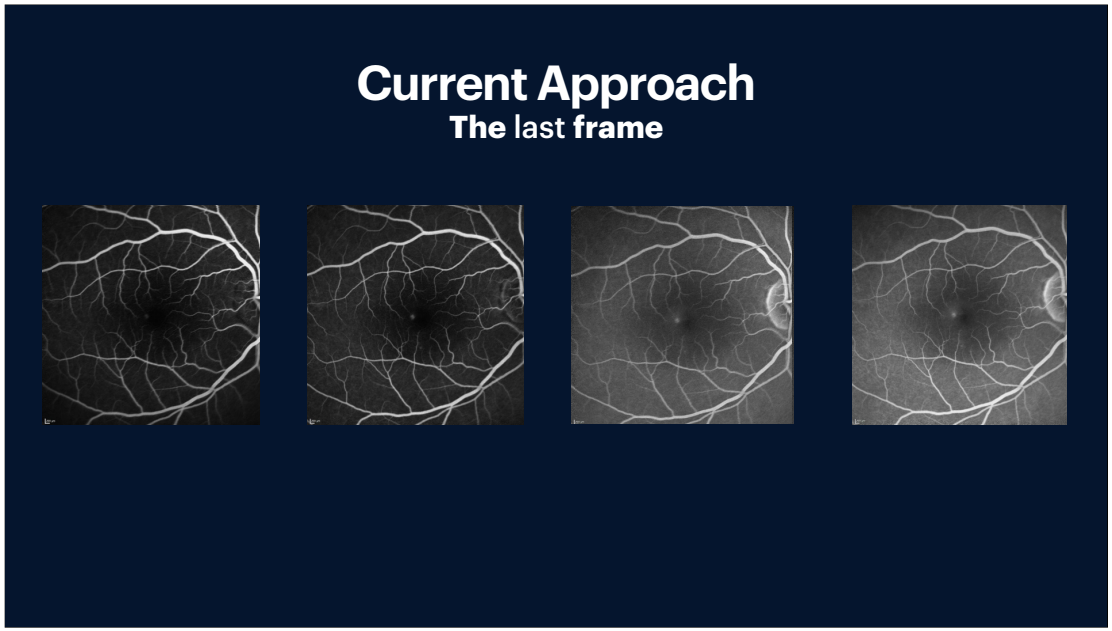
Look at this sequence. Between the first and last frame, the landscape changes entirely.



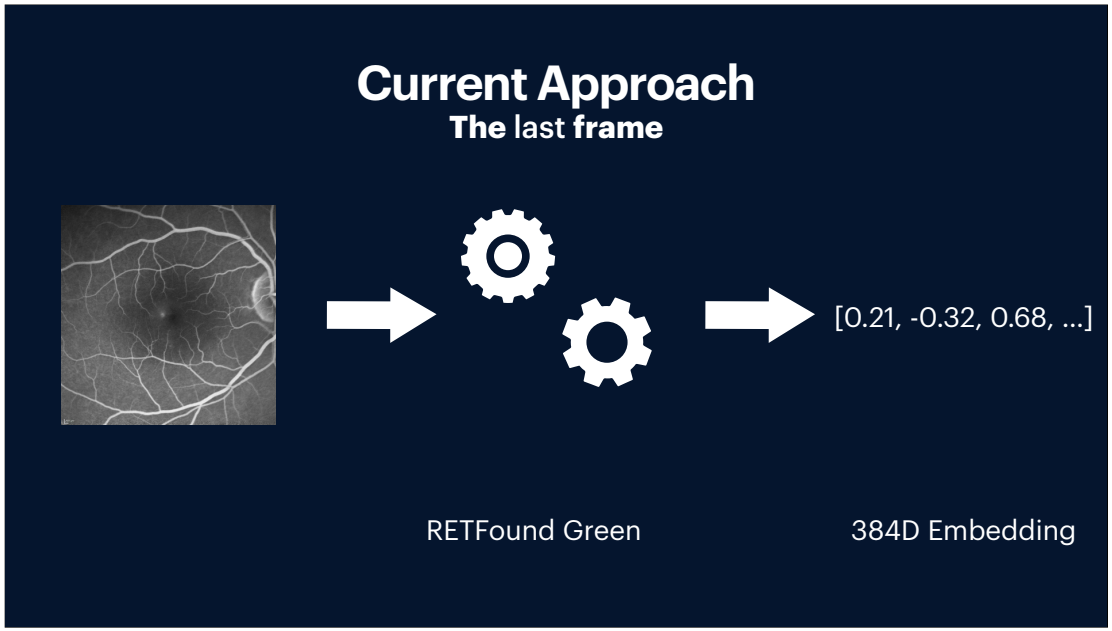
Now you might ask: why go through the trouble of an injection Why not use a standard non-invasive scan?

The reason is that non-invasive examinations often show us the **structure** of the eye, but FA shows us the **function**. The moment we inject that dye, we turn a static clinical observation into a live data stream of the patient's circulatory health.

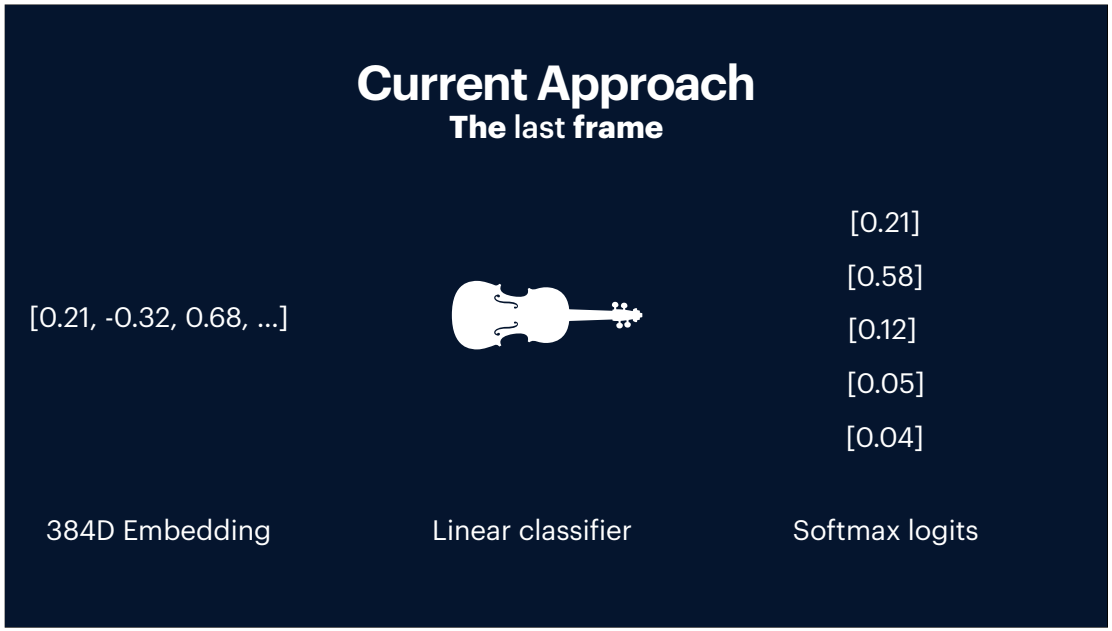
So let’s look at how we currently handle automated grading task.



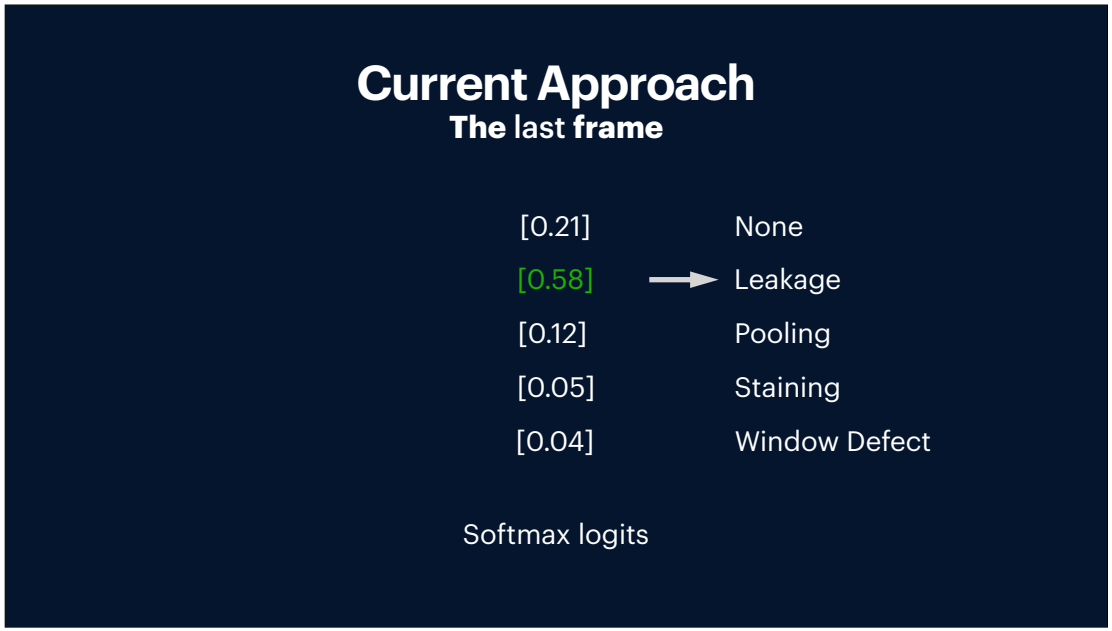
We take the very last frame —End of the examination— and we ask the AI to guess the whole story based only on that single moment



We use powerful models like RETFound Green. It turns that final image into a 384D embedding, that now represents the visual features of the eye.



We then feed those numbers into a linear classifier. Its elegant, but it’s making a decision based on a frozen moment in time.



Then the models gives us his best guess.

Current Approach

The last frame

FFT Results (using only last frame)

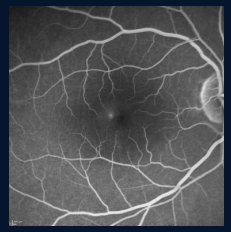
Metric	None	Leakage	Pooling	Staining	Window Defect	HyperF Type
AUC	0.95 ± 0.01	0.87 ± 0.02	0.76 ± 0.02	0.75 ± 0.03	0.78 ± 0.04	0.82 ± 0.02

And it’s quite good! Using just the last frame, we get a Hyperfluorescence type AUC of 0.82. But quite good isn't the same as ‘the best we can do’. We are leaving data on the table

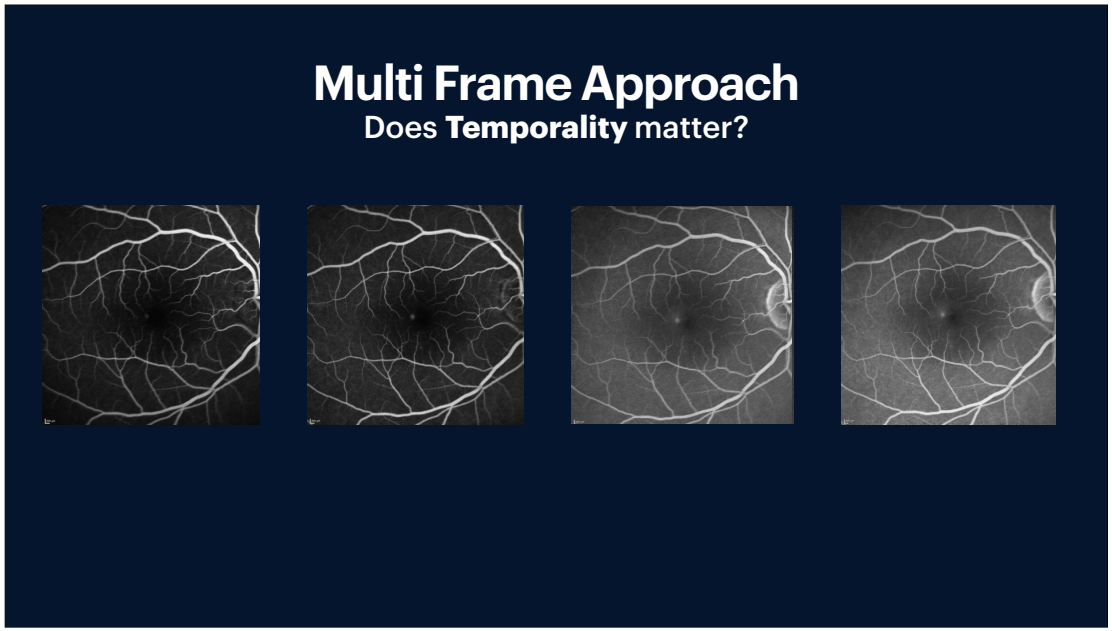
So we wanted to try a different path: The multi-frame approach

Multi Frame Approach

Does **Temporality** matter?

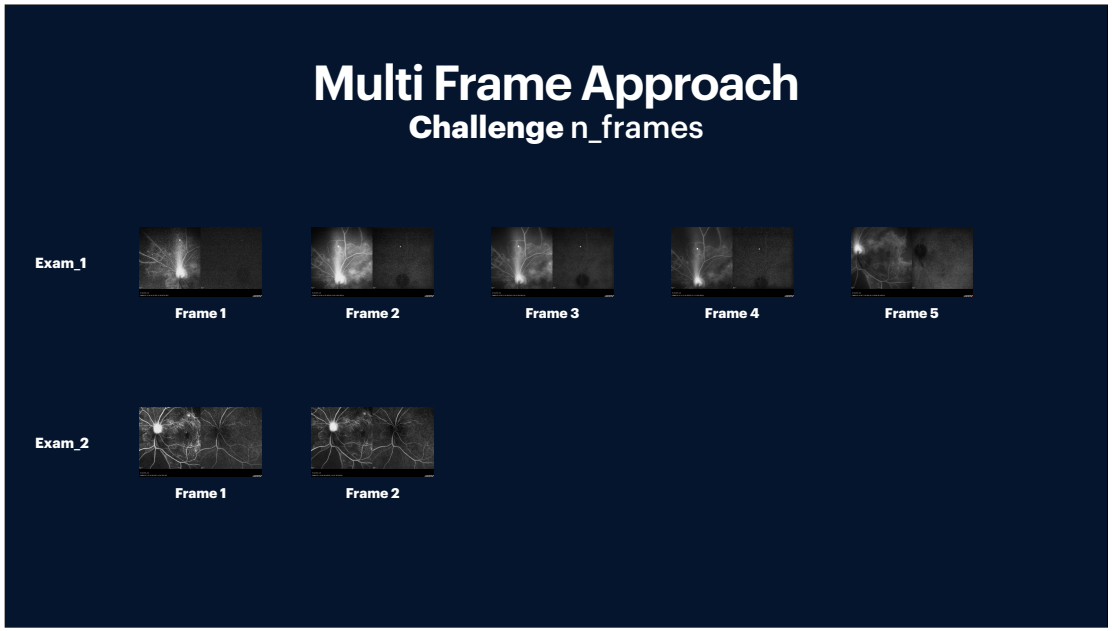


So if we instead of just looking at this last frame

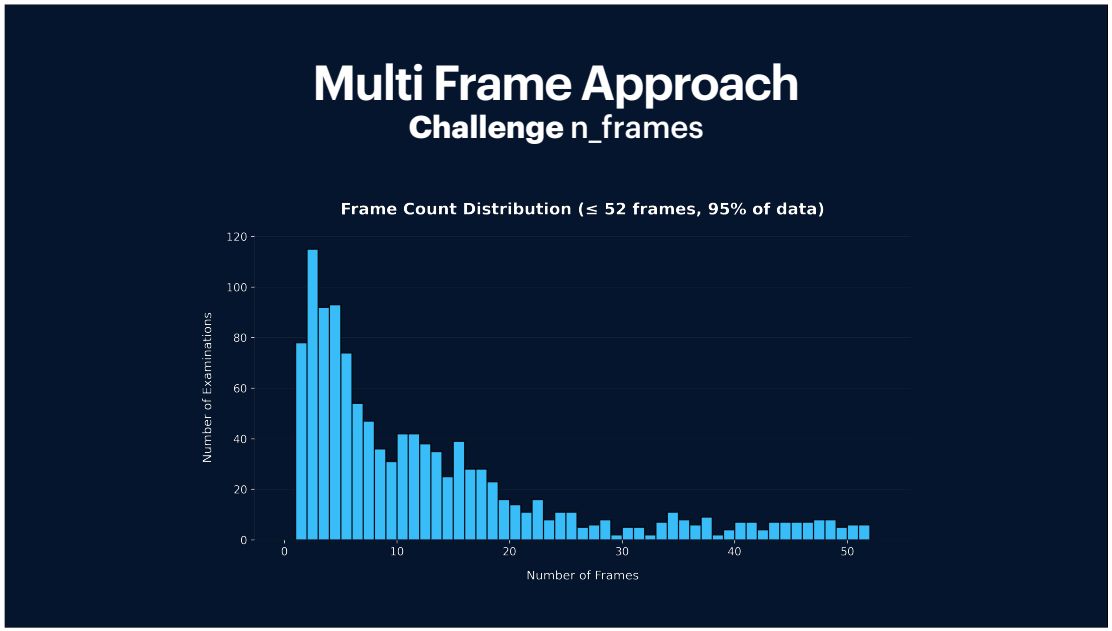


Take more data available from the examination there comes a question: Does the order and timing of these frames actually matter for the diagnosis?
To us it would seem like it, since the ophthalmologist also look at more than just the last frame to make their diagnosis, so why wouldn’t the AI benefit?

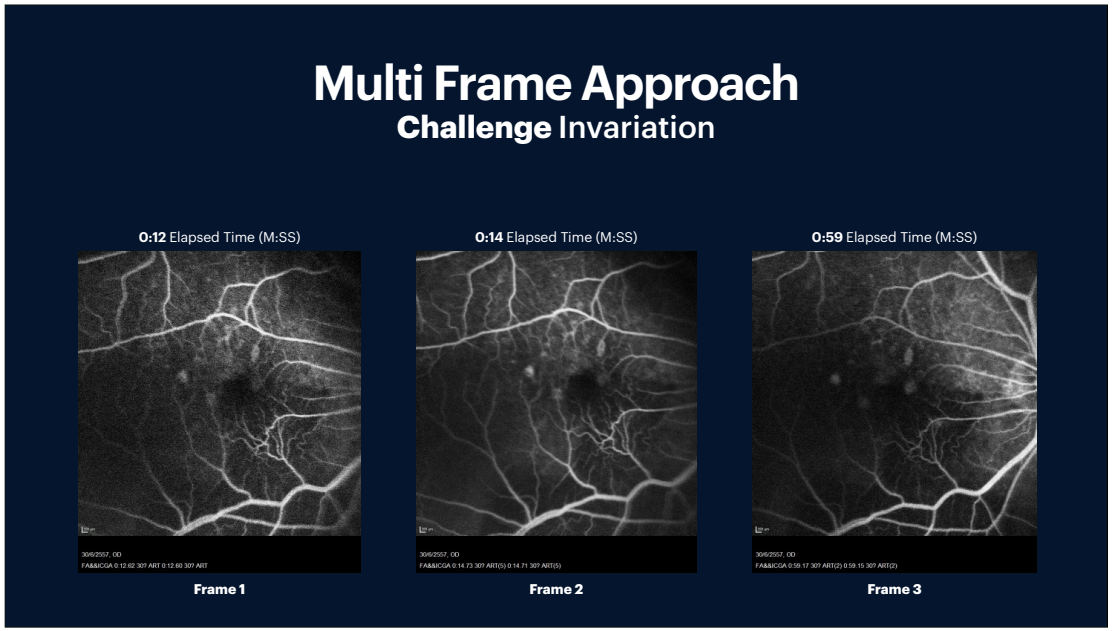
Research has shown, that ophthalmologists that had access to the entire examination were able to do a better job on the grading than by just looking at the last frame.



But there is a catch. In the real world, data is messy. One exam might have five frames; another might have fifty. AI likes consistency, and clinical reality is anything but that.

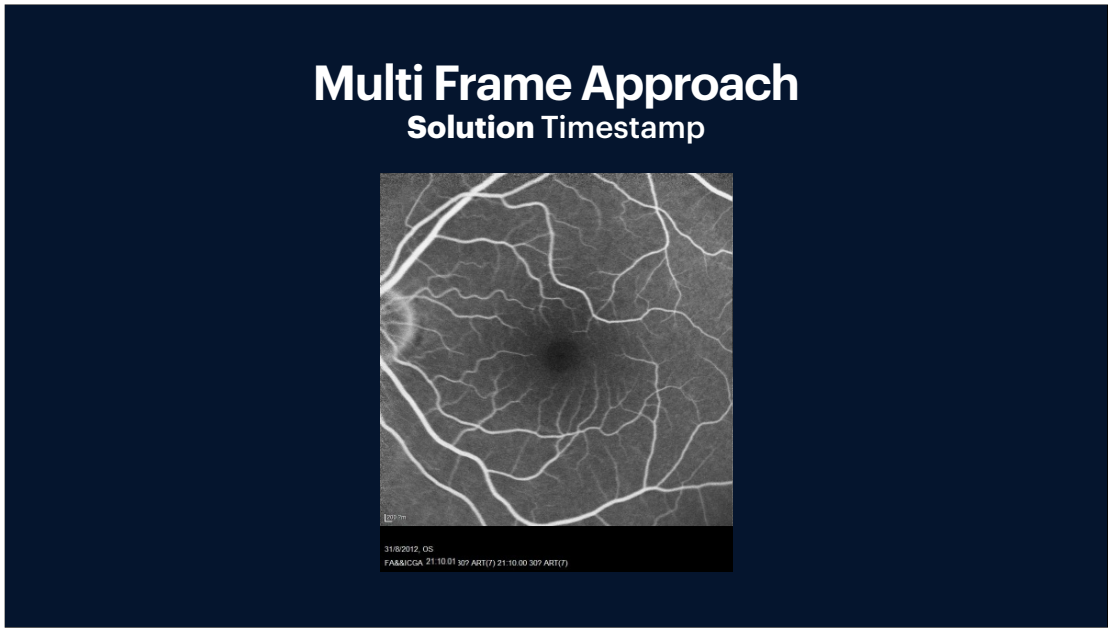


Look at this distribution. 95% of our examinations has 52 frames or fewer, but the variance is massive. We can’t just build a model that expects exactly 10 photos every time.

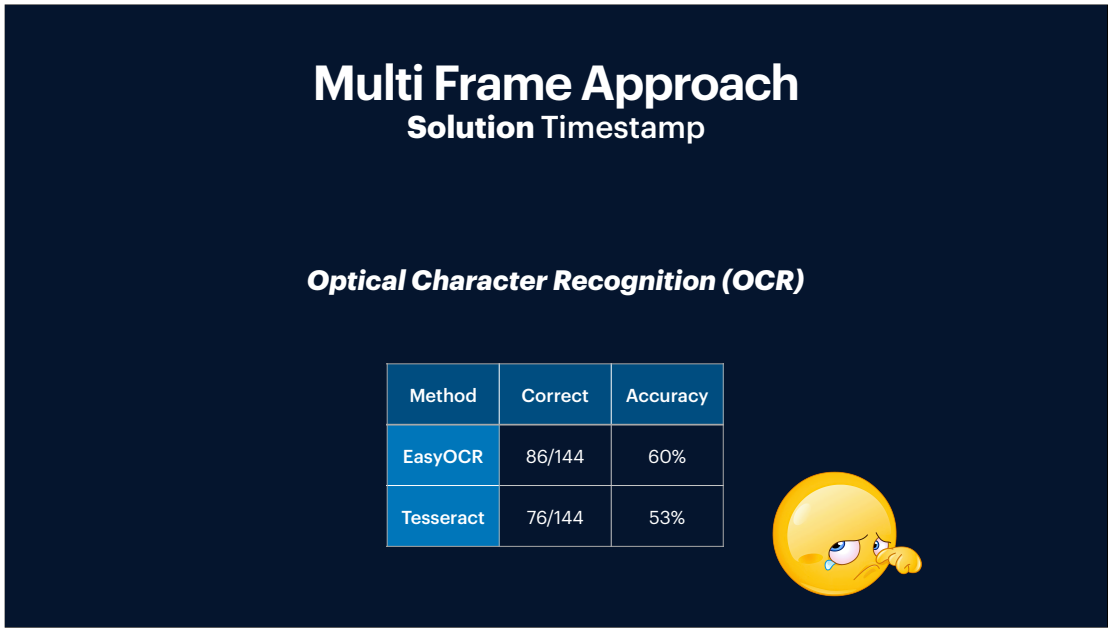


Another challenge, is that the elapsed time in between frames is also variating A LOT, so sometimes you could be looking at 2 frames that are in order next to each other, but have 3 minutes of space in between them. This does not make it a nice and easy uniform dataset.

Another BIG Temporal problem we were facing was that there were no timestamps in the metadata, so we couldn’t really put time itself to practice into our approach. And even if we were to have time, we still would not have phases discriminated from each other, which we could actually work with. So these were some of our main challenges.

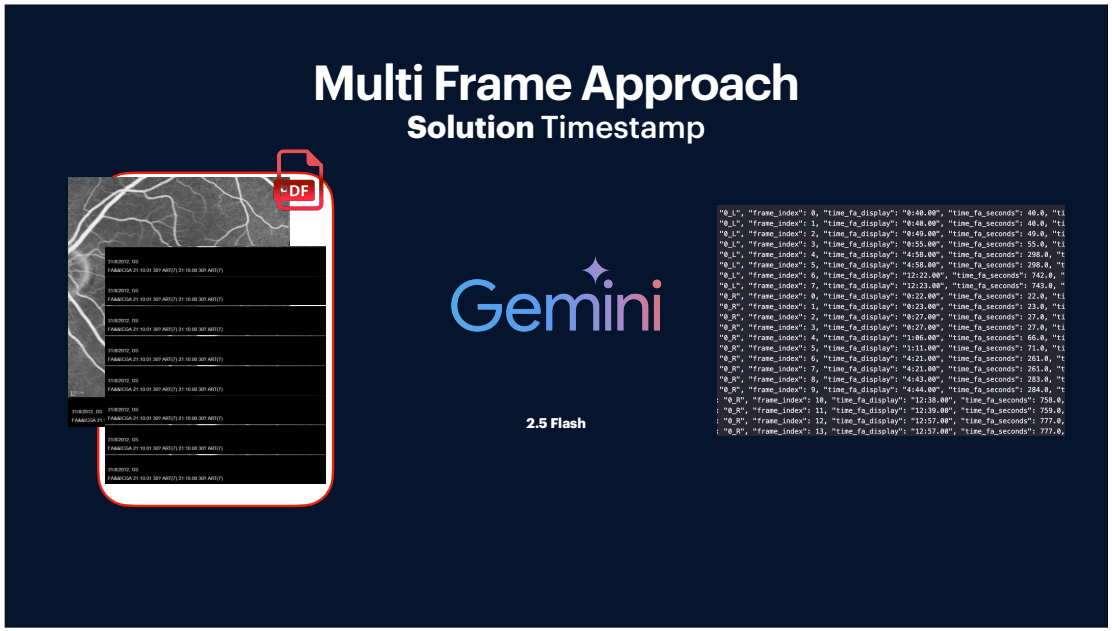


So, to make this presentations more positive I will now show some of our solutions to these challenges. So for our missing timestamps in the metadata we came up with the following solution. In the actual plain frame, there exists a timestamp! Which is good news for us, but how will we extract timestamps from 34000 frames you may ask.



The plan was to use OCR, which made it possible to insert an image, and then get the text extracted from it, this sounded very promising to me. So first of all I made a test set where I manually annotated 144 frames with correct timestamps. Then I let 2 different OCR's battle against each other to see how many of the frames they actually got the correct timestamps for.

Standard tools like Tesseract and EasyOCR are, frankly, a bit disappointing here. They get it right only about 53% to 60% of the time. In medicine, a coin flip isn't good enough.

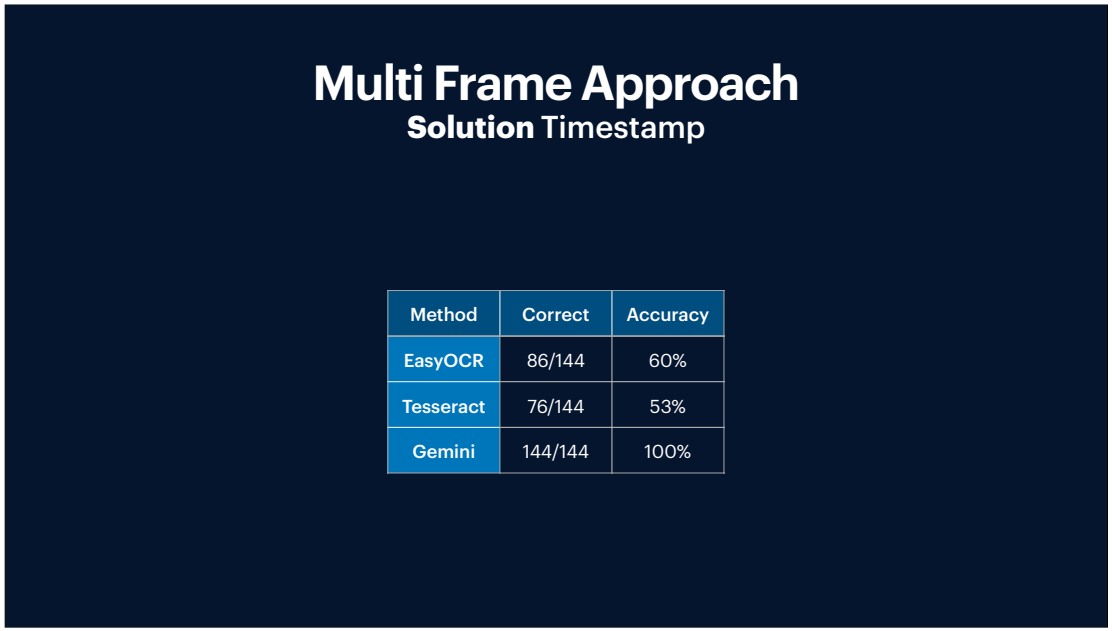


So, I got kind of sad that the simple OCR did not do its job well, and felt like this was not going to work out. So I started thinking in a very different way.

For each frame, of every examination, I cropped out just the bottom panel where elapsed time was.

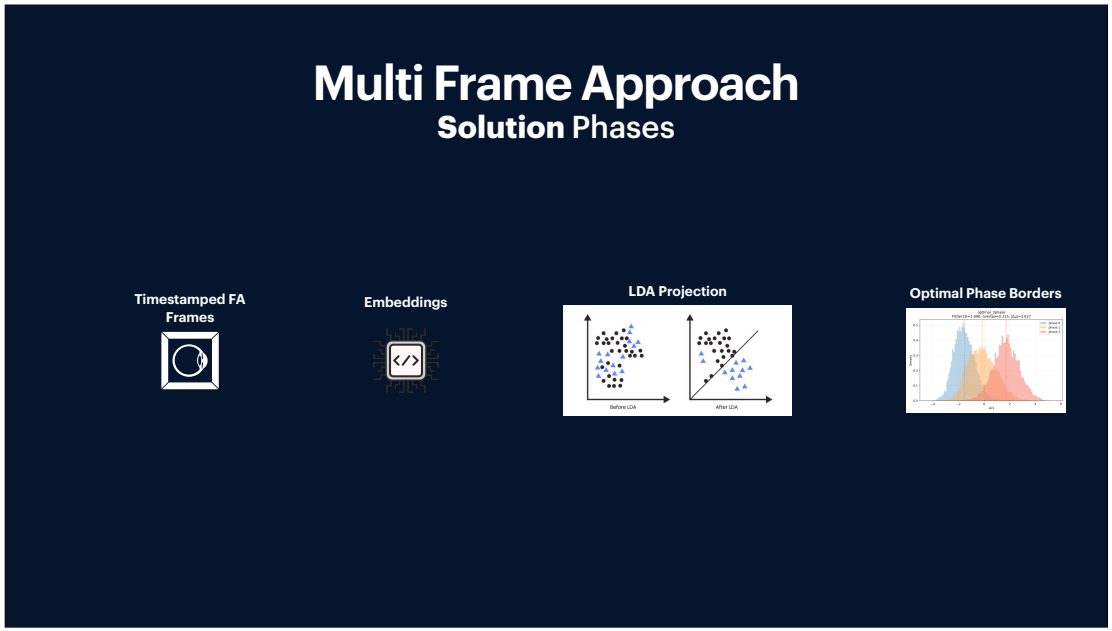
Then for all these bottom panels, I added them to 1 big PDF

Then, I got some help of a friend; Gemini. I uploaded the pdf and asked very kindly if he could please give me all the correct elapsed time timestamps. And guess what, he did it, I knew I could count on him/her

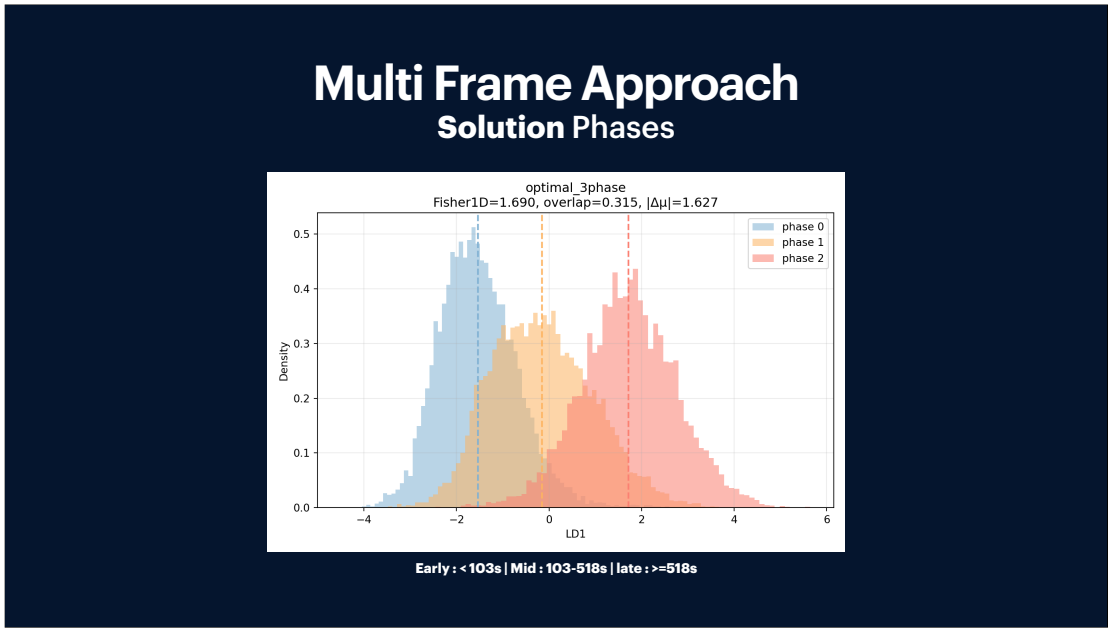


But were the timestamps actually correct this time? Lets compare.

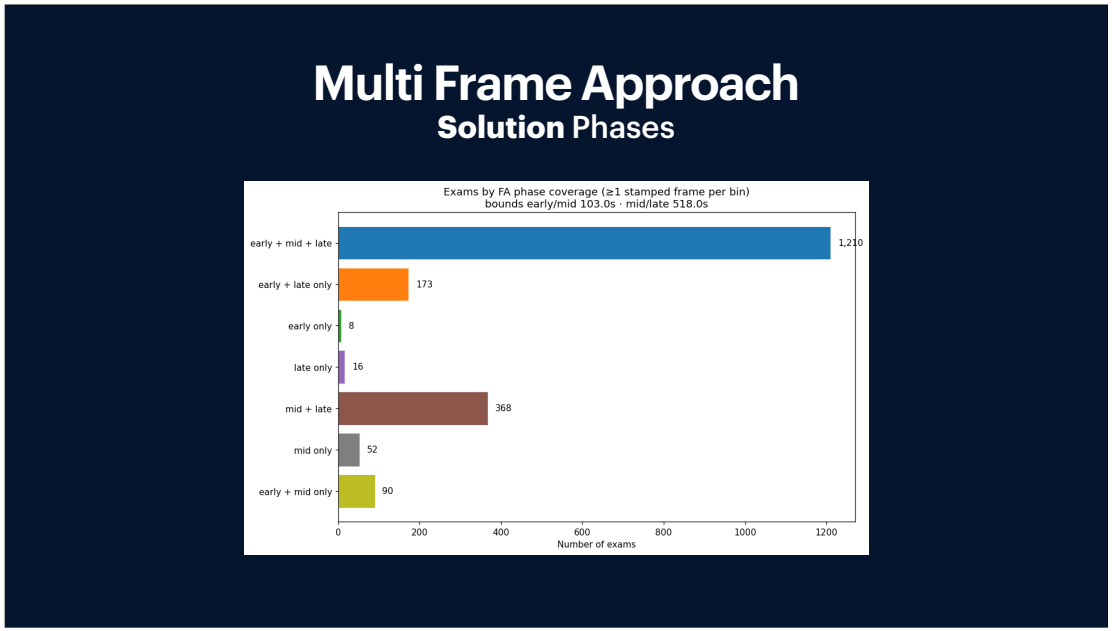
The result? 100% accuracy on our test set. We finally fixed the clocks and enriched the existing .csv



Now that we had the ‘when’, we could find the ‘what’. We used Linear Discriminant Analysis (LDA) to project our embeddings and find the optimal phase borders — the natural transitions in the eye’s response to the dye.



Mathematically, we defined three distinct windows: Early (before 103 seconds), Mid (up to 518 seconds) and Late (anything after). This turns a chaotic sequence into a structured story



Most of our exams —1,210 of them— cover all three phases. This means we have a full beginning, middle and end for the majority of the patients.
We use this cohort to run some experiments which will show us the difference in performance by using just 1 frame from each different phase

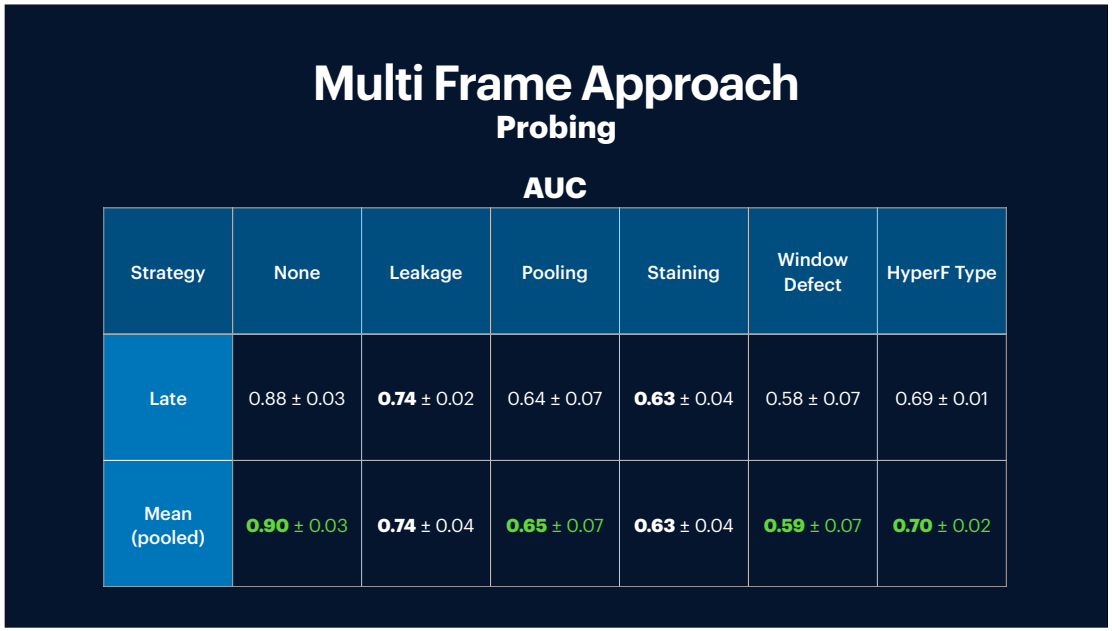
Multi Frame Approach

Probing

AUC

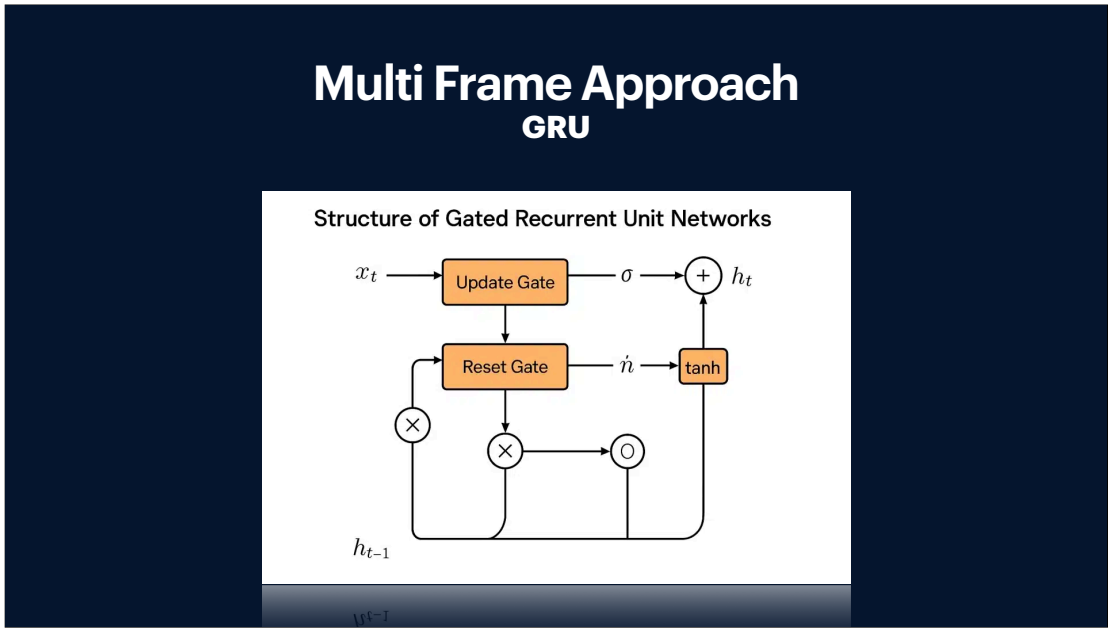
Strategy	None	Leakage	Pooling	Staining	Window Defect	HyperF Type
Early	0.84 ± 0.02	0.64 ± 0.03	0.60 ± 0.03	0.51 ± 0.03	0.54 ± 0.04	0.63 ± 0.03
Mid	**0.88** ± 0.03	0.71 ± 0.02	0.60 ± 0.07	0.57 ± 0.04	**0.60** ± 0.07	0.67 ± 0.02
Late	**0.88** ± 0.03	**0.74** ± 0.02	**0.64** ± 0.07	**0.63** ± 0.04	0.58 ± 0.07	**0.69** ± 0.01

So when we ‘probe’ these phases, meaning that we make embeddings out of a frozen backbone and then passing those embeddings through a linear classifier, we see that different pathologies reveal themselves at different times. Even though using the late phase frame gives us the best classification results, I think this shows that there is also information residing in the other phases that could attribute to making a better classification



So now, what happens when we actually try to use more than just 1 frame to make our classification? Our first most simple approach was making an average of all the embeddings inside of 1 examination and then making a classification from that mean pooled embedding.

It performed better, but not by a lot, so this was not really showing anything yet. This also seemed as a kind of dumb way to use the temporality, since taking the average lets a lot of important data possibly disappear.

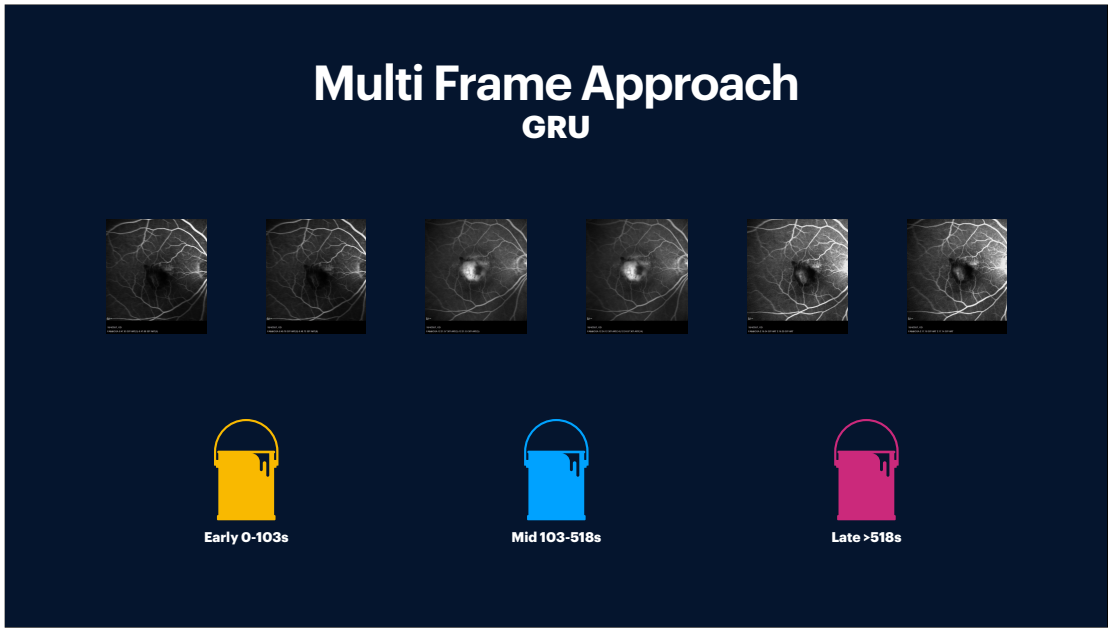


Now this is the heart of our current solution: The GRU. Unlike a static classifier, a GRU has a memory. It processes one frame at a time, updating its internal hidden state as it goes.

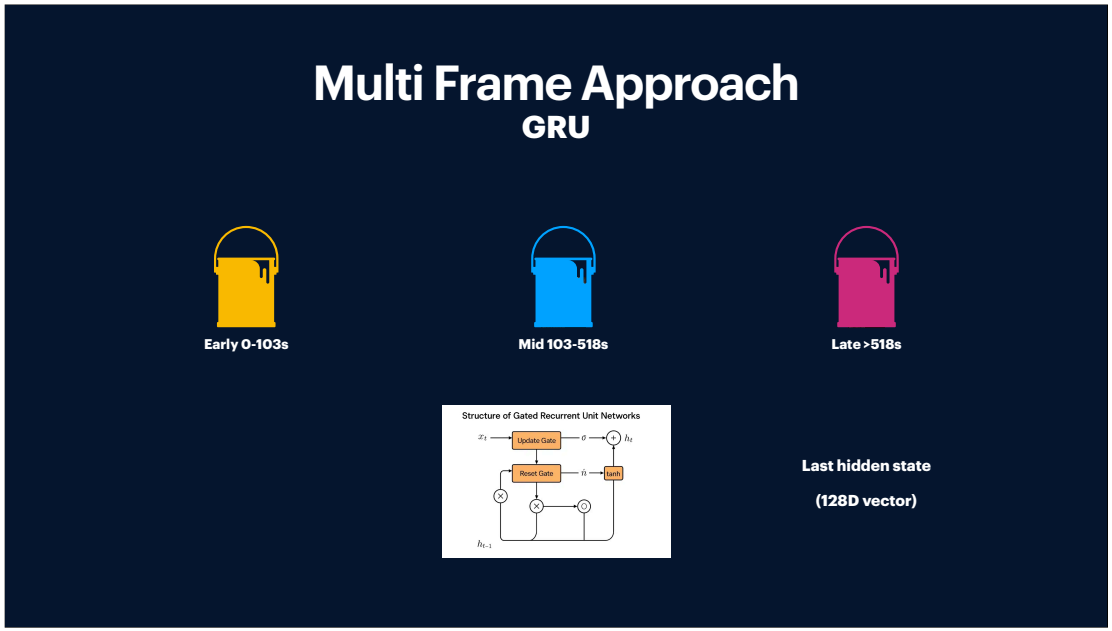
Instead of the algorithm starting fresh with every single frame, it used the GRU to decide what is worth carrying over.

There is a Update gate, which decides if there is something in the new frame that should be kept in memory
And there is the Reset gate, which decides whether something should be cleared from memory.

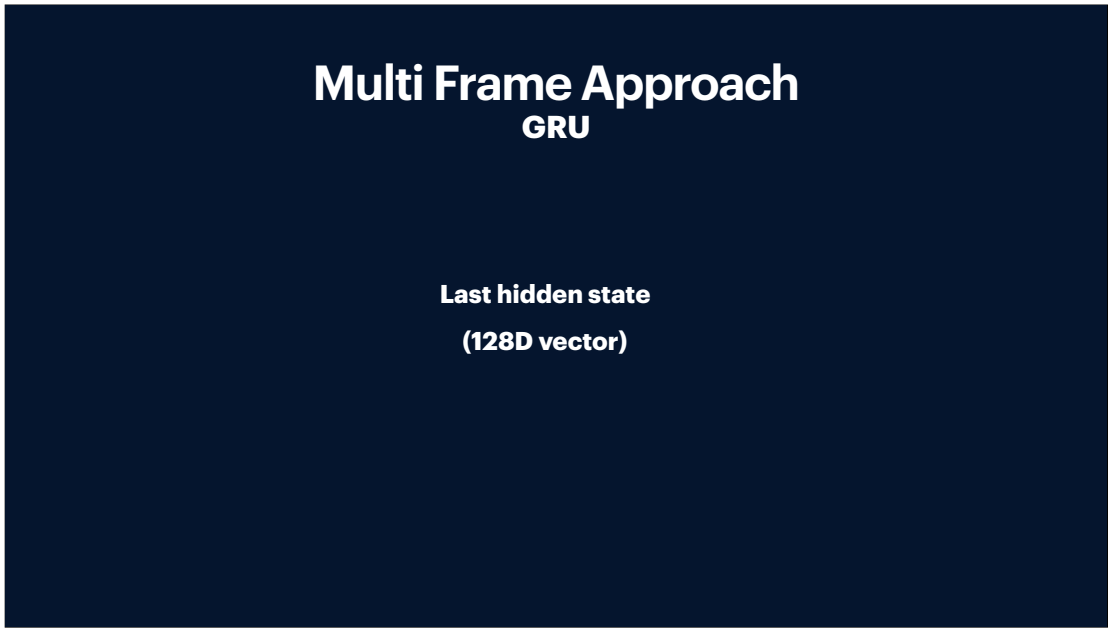
A standard neural network treats every frame as an isolated island. A GRU turns those frames into a narrative.



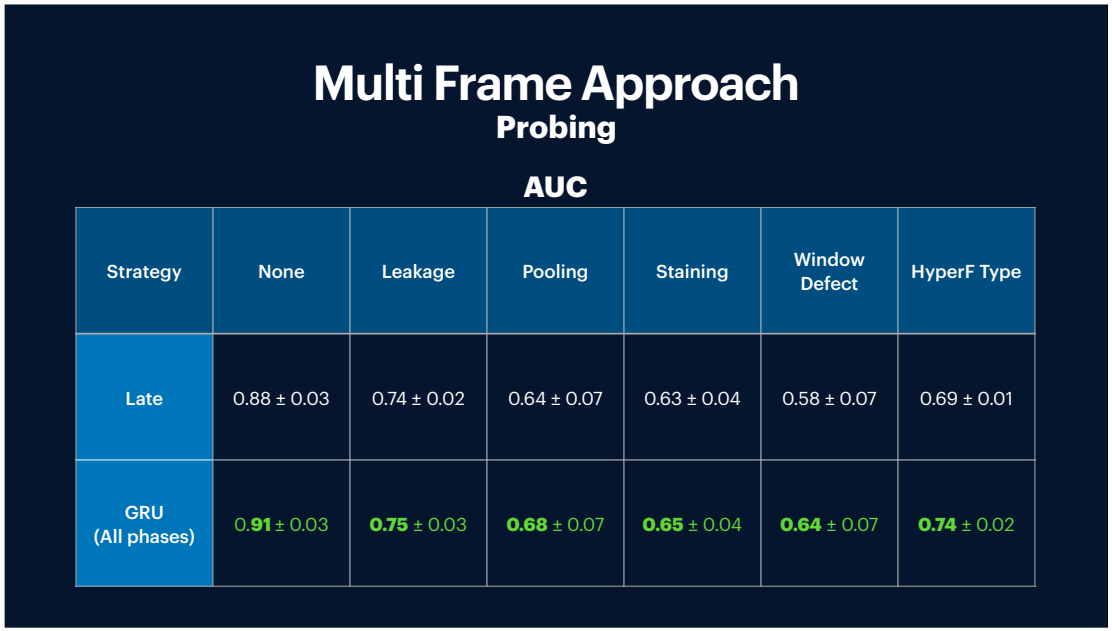
So to use this sequential architecture, we need sequences of frames. It was up to us to decide which frames we wanted to pick and feed to the GRU. Since we had defined our time phases, we could make buckets of the 3 separate phases and fill them with the associated frames from an examination.



After filling this buckets with the associated frames, we picked one frame from each bucket, and fed it into the GRU, then the GRU created the Last hidden state which is a 128D vector, when the next frame was being picked and fed to the GRU, the last hidden state changed accordingly, and so on, the same happened for the last bucket.



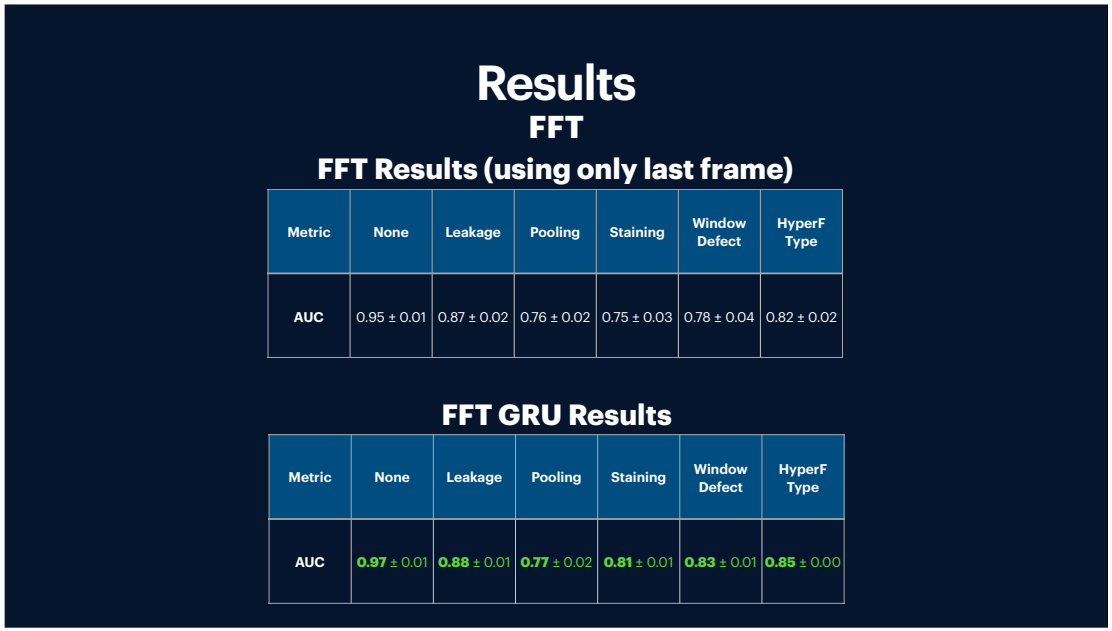
After all phases were shown to the GRU, we now had a last hidden state which kind of acted as a summary for our examination. And using this Last hidden state we made an actual classification using a linear classifier.



The impact is clear. By using all frames through a GRU, our diagnostic AUC jumps from 0.69 to 0.74. We are seeing more because we are looking at the whole story.

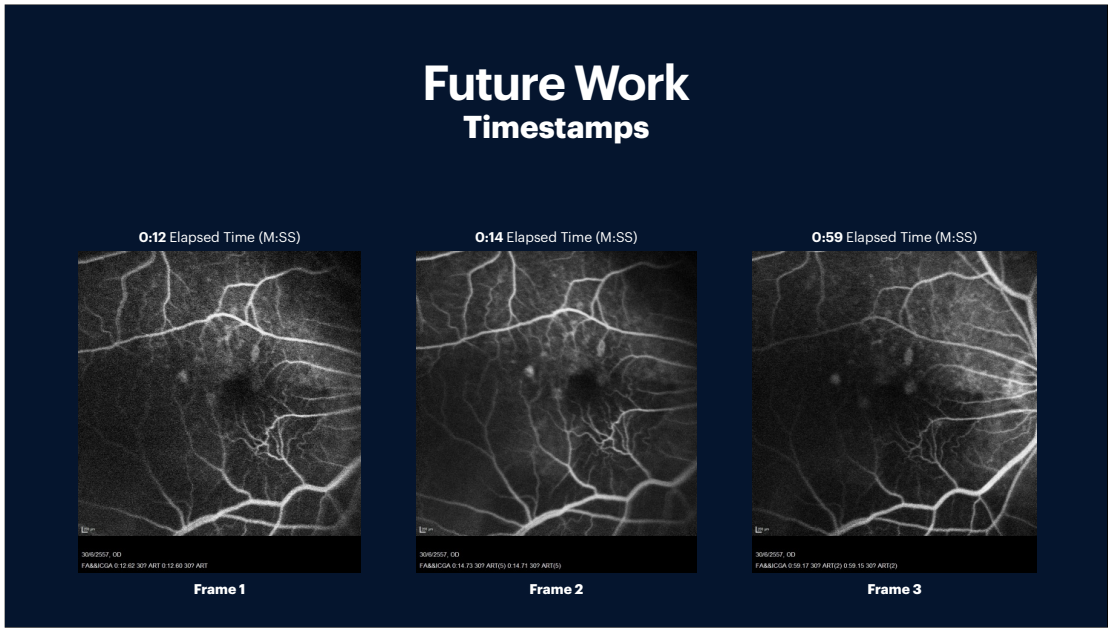
Results

Now let’s look at the final numbers (for now)

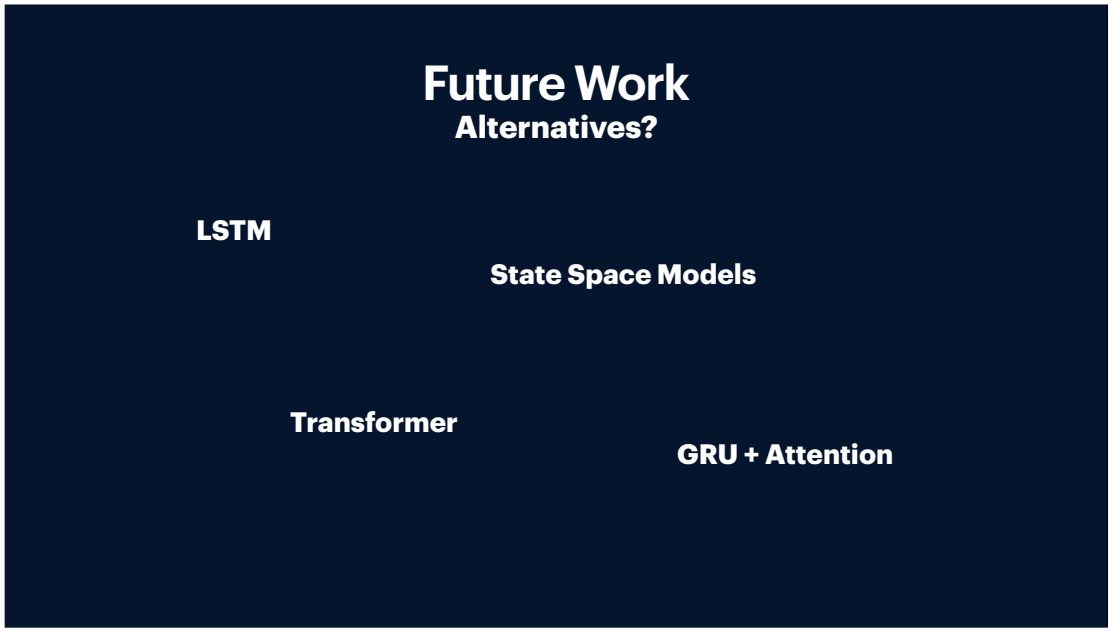


Our baseline ‘Last Frame’ approach gaves us a hyper fluorescence AUC of 0.82. Our new FFT GRU approach, which uses the entire sequence pushes that to 0.85. It isn’t just a marginal gain; it’s a more robust, clinically relevant wa of seeing

Future Work



The thing is, the GRU still has no idea of how much time has actually passed since the previous inserted frame, I think being able to use this information will increase the performance for the temporal task.



The thing is, the GRU still has no idea of how much time has actually passed since the previous inserted frame, I think being able to use this information will increase the performance for the temporal task.

Conclusion

Beyond the Last Frame

Temporal Deep Learning in Ophthalmology

Key Results

Baseline AUC : 0.82 (Single Frame)

GRU AUC : 0.85 (Sequential Data)

Contributions

Resolved Temporal Uncertainty (**OCR**)

Phase-Aware Feature Engineering (**LDA**)

Sequential Intelligence (**GRU**)

Future Work

Add Temporal Encoding

Experiment with different sequential architectures

Research what frames should be used

Tymo van Rijn / Idiap Research Institute / Med AI Team / Tvanrijn@idiap.ch