Liv4D

19th of March 2019

Annotating Batch#1

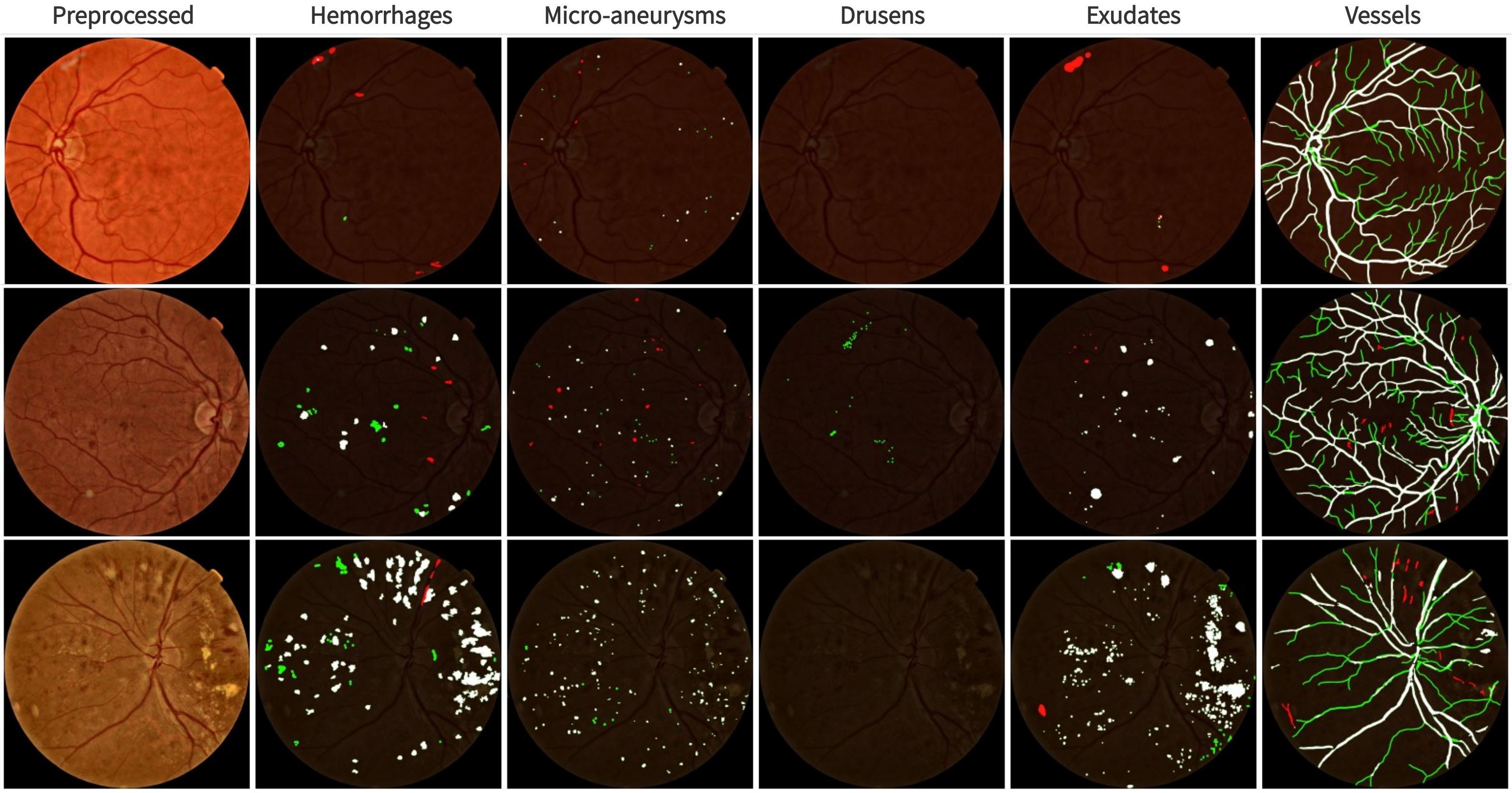
Quality Report

- Qualitative Analysis (All lesions)
- Quantitative Analysis (Red lesions)
- Machine Learning Analysis (Vessels)
- Conclusions & Guidelines

Qualitative Analysis (ALL LESIONS)

Lets start by analyzing your corrections visually. Figure 1 presents, for three images, these corrections (addition or deletion) on the two main red lesion types, the two main bright lesion types and on vessels.

Figure 1: Biomarker Corrections by the Clinical Team

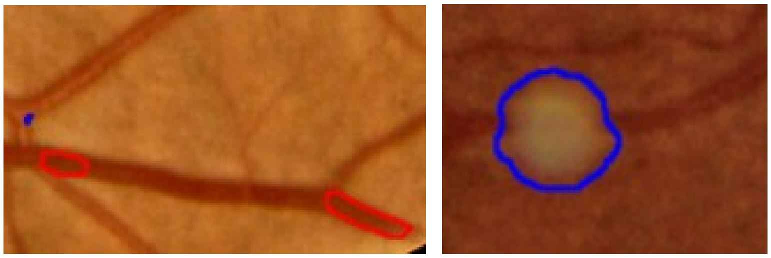


Preprocessed Hemorrhages Micro-aneurysms Drusens Exudates Vessels

White: biomarkers left as is. Red: Removed biomarkers. Green: Added biomarkers.

First we must note the significant amount of corrections you‘ve made to the database.
Visually, it would seems that your corrections consisted more of additions than deletions. This impression is confirmed by the numbers: the ratio of added pixels over all corrected pixels is **55%** for bright lesions and **62%** for red lesions. Thus, according to your corrections, the precision of our algorithms for lesions detection is **82%** (you erased 18% of the pixels initially detected), and their recall is **70%** (70% of pixels labeled as lesions on the final annotations were actually detected by our algorithms).

In fact, the real performances (particularly the precision) may be lower, as some lesions still present in the database are erroneous and should be removed in our opinion (see Figure 2). Having such errors in the training data may not be critical as they will probably not be generalized by the neural network. Nonetheless we should try to avoid them as much as possible.



◀ **Figure 2: Examples of f**alse positives of our algorithms not removed by your corrections.

Left: vessel segments detected as hemorrhages.
Right: reflection artifact detected as exudate.

Quantitative Analysis (Red Lesions)

In order to evaluate the effect of correcting of the red lesion labellings, DR gradings computed from our labeling were compared with those provided with the MESSIDOR database. To compute our own DR grades, we used an adaptation of the grading system used by MESSIDOR :

- **0 (healthy):** Micro-aneurysms **≤ 1**, Hemorrhages **≤ 1**, **no** Neovascularization;
- **1:** Micro-aneurysms **≤ 5**, Hemorrhages **≤ 1**, **no** Neovascularization;
- **2:** Micro-aneurysms **≤ 15**, Hemorrhages **≤ 5**, **no** Neovascularization;
- **3:** Micro-aneurysms **> 15 or** Hemorrhages **> 5 or** Neovascularization **> 0**;

Lesions with a radius lower than 10 pixels were removed to prevent over-detecting the pathology (see Appendices). The screening and grading performances are presented in Table 1.

Table 1: Agreement Matrices for Screening and Grading before and after Corrections of the Red Lesions.

Before Correction

All Grades

PREDICTION

TRUTH	0	1	2	3
0	9	0	0	0
1	0	1	10	5
2	0	0	5	3
3	0	0	2	6

Screening (Healthy vs Pathological)

PREDICTION

TRUTH	H	P
H	100.00%	0.00%
P	0.00%	100.00%

Grading (1 vs 2 vs 3)

PREDICTION

TRUTH	1	2	3
1	6.25%	62.50%	31.25%
2	0.00%	62.50%	37.50%
3	0.00%	25.00%	75.00%

After Correction

All Grades

PREDICTION

TRUTH	0	1	2	3
0	7	1	1	0
1	1	3	9	3
2	0	1	1	6
3	0	0	1	7

Screening (Healthy vs Pathological)

PREDICTION

TRUTH	H	P
H	77.78%	22.22%
P	3.13%	96.88%

Grading (1 vs 2 vs 3)

PREDICTION

TRUTH	1	2	3
1	20.00%	60.00%	20.00%
2	12.50%	12.50%	75.00%
3	0.00%	12.50%	87.50%

First and foremost, please note that the screening performance *Before Correction* is **rigged**: to reduce your work we purposely deleted all lesions from images classified as Healthy in the MESSIDOR database. In reality the specificity of our model is **not** 100%.

The lower screening accuracy *After Correction* is due to two images labeled Healthy by MESSIDOR clinicians in which you found lesions (images #12 and #13) and one image annotated as grade 1 in MESSIDOR from which you removed most of the micro-aneurysms (image #85). In those 3 cases, I wouldn't say your corrections are erroneous but rather that those images were arguably misclassified by MESSIDOR clinicians. At least in that debate, we have the local (i.e. biomarker level) annotations to justify our diagnosis: that's the whole point of local annotations!

In our opinion, the grading disagreements are more concerning (overall accuracy decreased from 37.5% to 35.5% after correction). There are two causes for those disagreements. **1.** Corrections only marginally improved grade 1 accuracy: 60% of them are still classified as grade 2, mainly because some hemorrhages falsely detected by our algorithm were not erased. **2.** Grade 2 images were classified as grade 3 (images #2, #20, #88) after correction because you annotated these images with more precision than MESSIDOR clinicians (and than our algorithm) did.

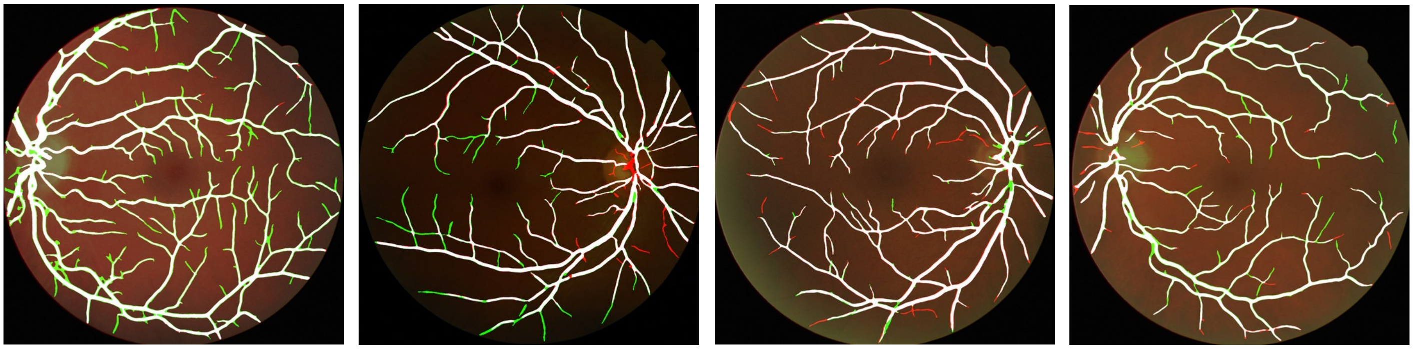
To sum up, these comparisons confirm that local labeling is well suited for screening. They also highlight two biases of our current process which may cause "over-labeling":

- Unlike when labeling images from scratch, the large amount of lesions detected by the AI algorithm can saturate the clinician's attention, who is then less likely to validate every one of them individually.
- Mixing images of different grades during the annotation process may influence the user to unconsciously try to reproduce the high lesion density of pathological images on all images. Then, when labeling healthy images, he/she may be more prone to classify uncertain ensembles of pixels as lesions.

Machine Learning Analysis (VESSELS)

Our neural network for vessels segmentation has been quickly retrained with the 50 newly annotated images (thereby doubling the number of training images). In this section, we compare those two models (the initial pre-labeling model and the retrained one). The following figures show the improvement in segmentation owing to the retraining.

Figure 3: Differences in automatic vessels segmentation when retraining the model with the new labels from batch #1.



White: Segmentation common to both model. Green: Added by retraining. Red: Removed by retraining.

The qualitative analysis of biomarkers corrections revealed that, for vessels, most of these corrections consisted in adding small vessels (see Figure 1). Logically, retraining the model on those images allowed it to better detect small vessels. This was tested on another public database (DRIVE): the algorithm increased its performance on small vessels (diameter < 13 pixels) from **42%** to **46%**. Globally, the model gained in sensitivity on this database (see Table 2).

Initial pre-labeling model:

		PREDICTION	
		Background	Vessels
TRUTH	Background	98.60%	1.40%
	Vessels	27.73%	72.27%

Retrained model:

		PREDICTION	
		Background	Vessels
TRUTH	Background	98.26%	1.74%
	Vessels	25.65%	74.35%

▲ **Table 2:** Vessels segmentation confusion matrices.

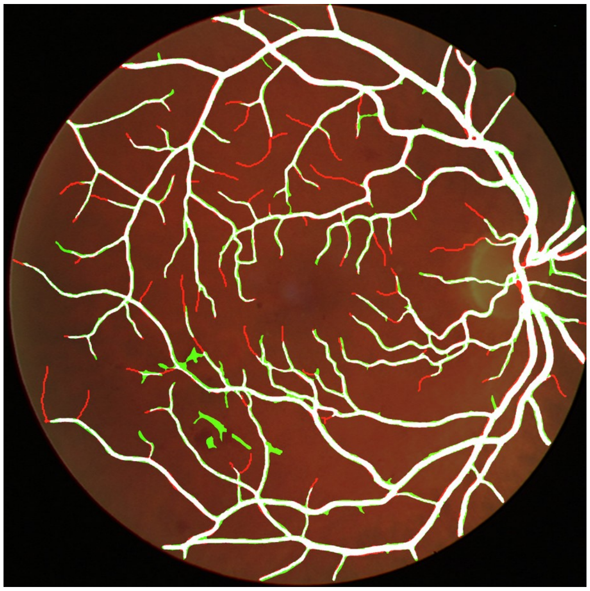


Figure 4: Comparison between the segmentation from the retrained model and the DRIVE ground-truth. ▶

Despite this improvement, the model still misses a lot of small vessels (causing a false negative rate of 25%). It also commits minor false positive errors (e.g. hemorrhages or choroidal vessels detected as retinal vessels) but these should disappear as our database grows. On the other hand, the retrained model sometimes over-segments small vessels (causing the green shadow around the ground-truth small vessels in Figure 4). I would blame the use of a brush size too wide for those tiny vessels... Though, I'm sure the recognition of pencil pressure (when using tablet and stylus) recently added to the annotation interface will be of great help in that respect.

Finally we've noticed some misaligned vessels on the corrected labels (see Appendices). Those small errors can make the annotations unusable as a validation reference, but shouldn't cause any problems as training labels. Indeed, as long as the large majority of the vessels are well aligned, the network should average out those errors and not learn them.

Conclusions & Guidelines

In brief...

You've performed an impressive amount of work on the 50 first images: in total you spent **16 hours** correcting them (20 minutes per images) over 5 weeks. Overall, our algorithms seem to provide good enough lesion detection, as more than two thirds of lesions in the final annotations were detected automatically. Regarding the specificity, the algorithms also show satisfactory performance since you added more lesions than you removed. Although, we found some cases where erroneous lesions slipped you attention and were not removed.

Those kind of errors are not critical in small numbers, but they are likely to cause poor sensitivity performance on screening or grading tasks if they are learned by the neural network (because then it will "over-detect" lesions). Unfortunately, they are not simply caused by moments of inattention. Our labeling process is cognitively biased (see the quantitative analysis above) and is likely to report more lesions than expected if the biases are not actively taken care of. Hence the updated annotation guidelines below.

Still, the corrections you've made on our data are really valuable to us and are already improving the performance of our algorithms! Indeed, the vessels segmentation gained 4% of accuracy on small vessels.

Guidelines for labeling future batches

Red Lesions

- Most grading systems give significant weight to the presence or absence of neo-vessels, but none were annotated even though Dr. Duval noted their presence in one image when annotating vessels. Please look out for **neo-vessels** and roughly circle them with the fill brush.
- Be careful to remove **erroneous hemorrhages**.
- Even though the number of micro-aneurysms detected by the AI algorith can be overwhelming, try as much as possible to remove the erroneous ones.

Bright Lesions

- Be careful to remove reflection **artifacts** detected as exudates.
- Even though the amount of exudates detected by the AI algorithm is a bit overwhelming, try as much as possible to remove the erroneous ones.

Vessels

- Continue to annotate small vessels with as much care.
- When possible, without increasing the annotation time too much: use a minimum brush size of 5 px for small vessels. (You can now use the pen pressure to annotate medium vessels without manually changing the brush size, when using a tablet with stylus.)

Matters open to discussion

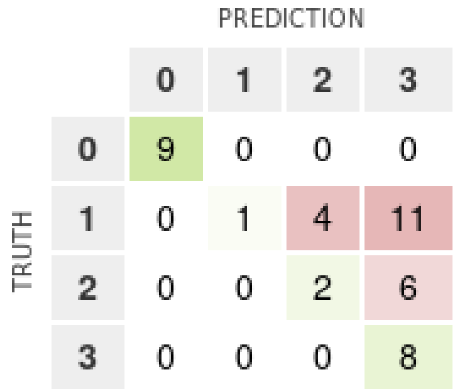
- Should we define a minimal lesion size (for example by limiting the zoom level)?
- Should the diagnosis from MESSIDOR be made available during the annotation?
- When and how should we test inter-operator variability?
- Ultimately, what grading system will we use?

Appendices

Qualitative Analysis

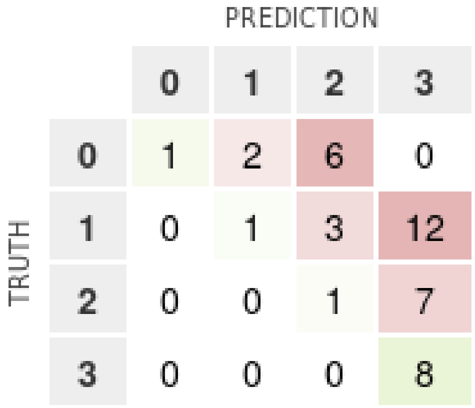
Table 1: Agreement Matrix without processing and rule tolerances

Before Correction



		PREDICTION			
		0	1	2	3
TRUTH	0	9	0	0	0
	1	0	1	4	11
	2	0	0	2	6
	3	0	0	0	8

After Correction



		PREDICTION			
		0	1	2	3
TRUTH	0	1	2	6	0
	1	0	1	3	12
	2	0	0	1	7
	3	0	0	0	8

Machine Learning Analysis

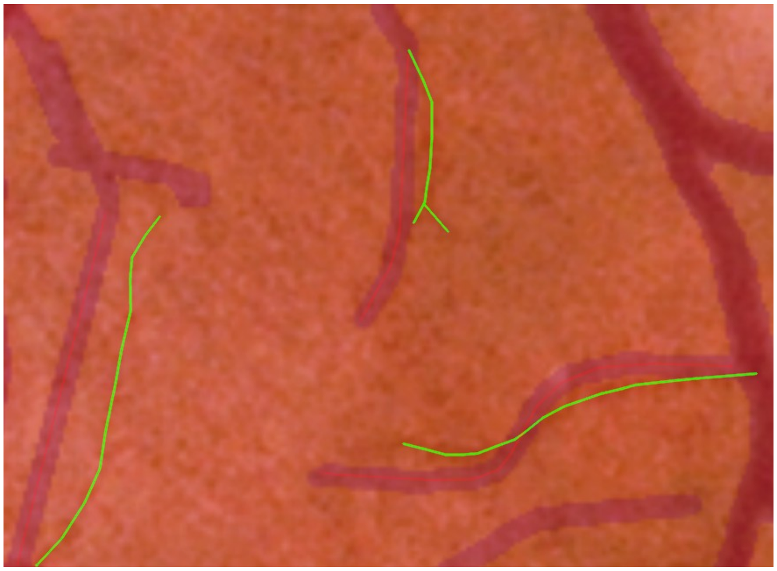


Figure 1: Misaligned Vessels Annotations