Malaria diagnosis and deep learning

Aniss Acherar

aniss.acherar@gmail.com

Sorbonne Université

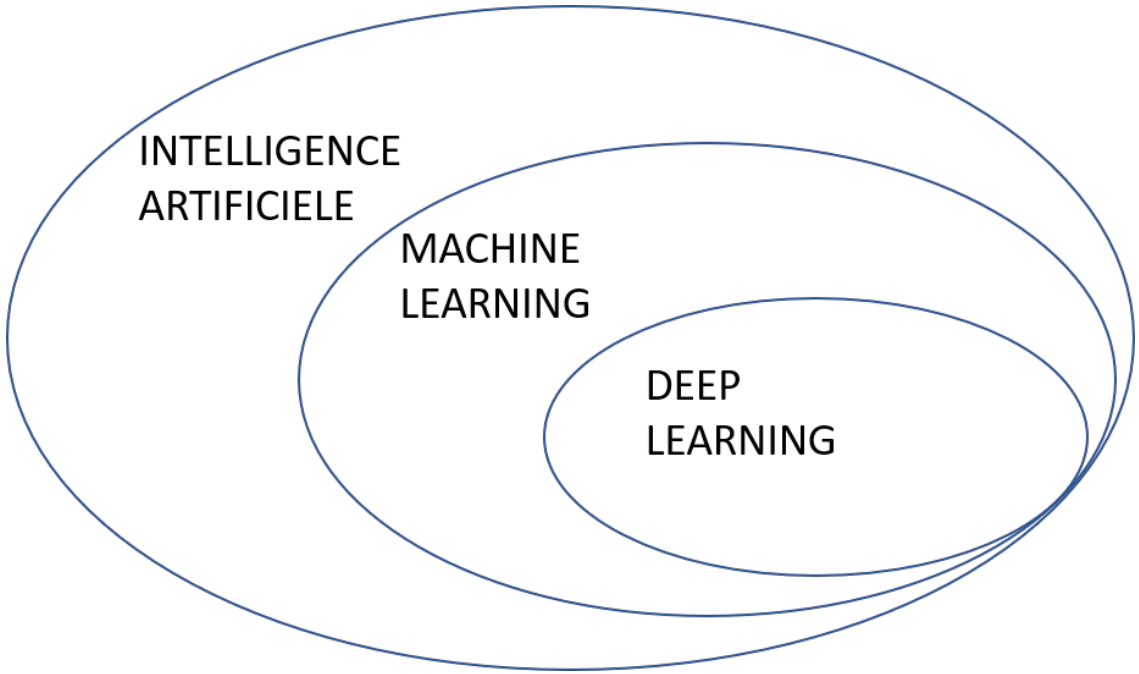
Tuesday, July 5 2022

1. Artificial intelligence, Machine Learning & Deep Learning
   - Some reminders : Learning, regression, classification

2. Image classification
   - What's an image ?
   - Convolutional neural networks (CNNs, ConvNets)
   - Malaria diagnosis using CNNs
   - Results
   - Limits & future improvements

Plan

Artificial intelligence, Machine Learning & Deep Learning



Computer vision

Thanks to AI, we have a technological area applied to unstructured data (face recognition, motion detection, self-driving cars navigation etc...)

Plan

Supervised or unsupervised learning

Supervised learning :

Training an algorithm on input data that has been annotated (labeled) :

- The inputs **x (features)** have a meaning **y (output/target)** ;
- For each input **x** : we have a value to predict **y** ;

Unsupervised learning :

Training an algorithm on input data that has not been labeled :

- The inputs **x (features)** doesn't have a meaning **y (output/target)** ;
- The goal is to **cluster** unlabeled data : find groups (clustering) or association ;

Reminders

Regression or classification ?

- Predict the sugar level of a patient is a **regression/classification** ;
- Predict that an MRI scan have a tumor is a **regression/classification** ;

Plan

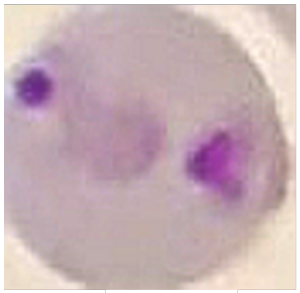
Plan

A grayscale image is a matrix of values

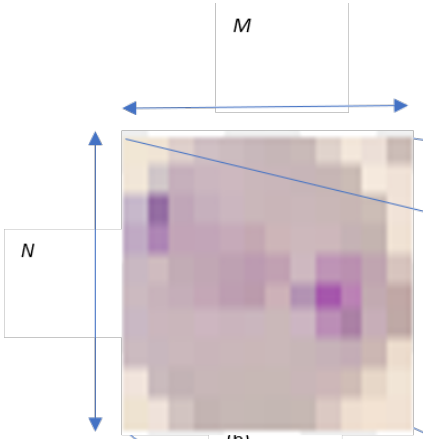
28 × 28
= 784 pixels

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	87	157	156	187	215	81	0	0	0	0	0
0	0	0	0	0	0	0	0	5	25	54	25	96	140	215	215	251	254	254	254	254	254	163	0	0	0	0	0
0	0	0	0	0	0	0	0	157	254	254	254	0	254	254	254	254	254	254	254	254	241	100	0	0	0	0	0
0	0	0	0	0	0	0	0	156	117	184	214	214	156	117	49	19	19	141	254	241	158	23	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	173	254	246	70	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	247	254	219	58	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	43	203	254	250	121	15	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	43	145	254	254	229	111	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	4	137	245	254	254	238	40	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	130	254	254	254	254	254	235	106	23	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	55	196	196	114	194	223	254	254	216	23	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	75	245	253	55	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	150	254	134	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	200	254	134	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	246	253	59	0	0	0	0	0	0	0	0	0
0	0	0	17	11	18	0	0	0	0	0	0	0	0	67	254	254	170	0	0	0	0	0	0	0	0	0	0
0	0	0	128	254	159	0	0	0	0	0	0	20	140	248	254	254	9	0	0	0	0	0	0	0	0	0	0
0	0	0	154	254	229	80	79	79	161	176	218	254	254	254	104	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	29	203	254	254	254	215	254	254	254	250	95	11	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	5	80	155	155	156	111	58	58	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A color image is also a matrix of values



(a)



(b)

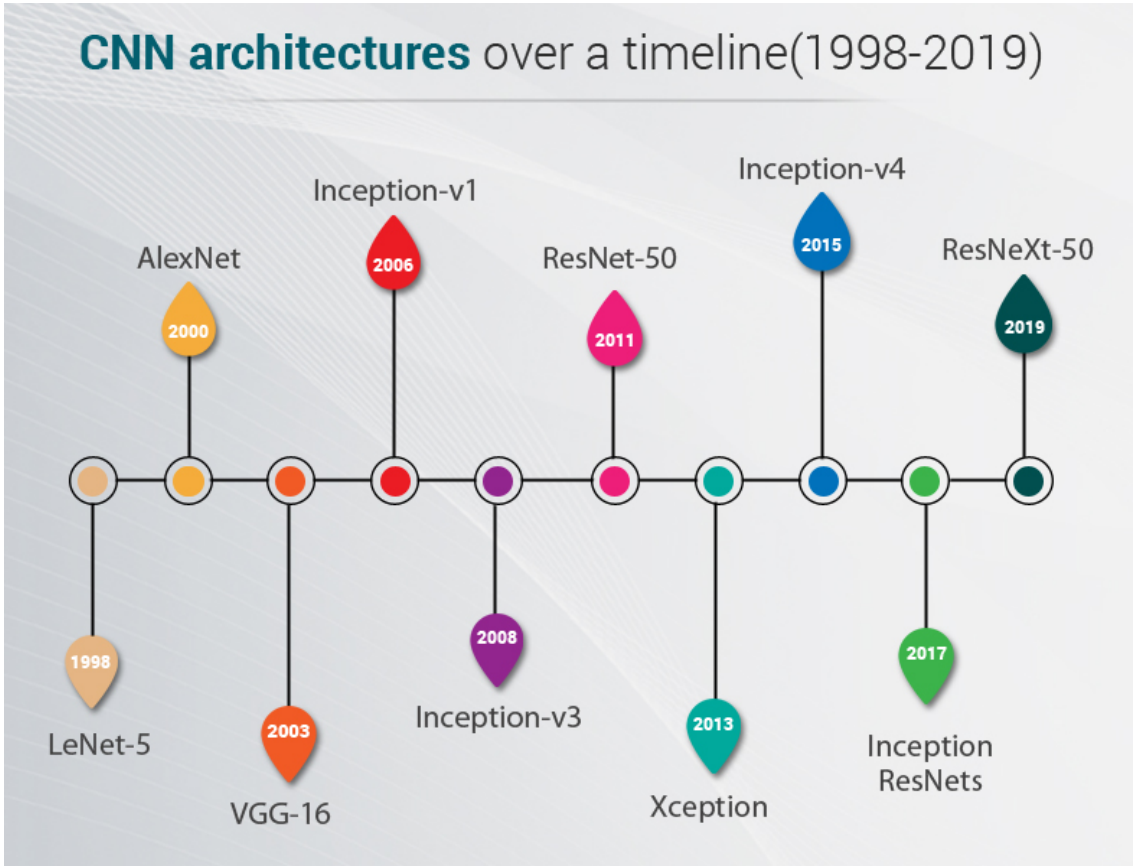
RGB intensity values

227.333	227.333	227.667	228.667	229.000	228.000	228.000	223.667	220.667	220.333	219.333	214.333	
228.667	228.667	228.667	229.667	228.667	227.667	226.333	222.333	218.000	215.667	213.333	207.333	
228.667	228.667	228.333	228.333	229.000	226.000	224.000	220.333	212.667	208.333	205.333	203.000	
228.667	228.667	227.333	227.333	227.000	224.000	220.333	212.333	208.000	203.000	200.000	198.667	
228.667	228.667	226.333	226.667	226.000	220.000	215.333	208.333	204.667	202.667	199.333	195.333	
229.667	228.667	227.000	226.000	220.667	214.333	206.000	201.000	196.667	195.667	195.000	193.000	
230.000	229.333	226.000	222.667	217.667	207.333	198.000	190.667	183.333	181.333	185.333	188.333	
232.000	231.333	222.333	220.000	210.000	200.333	189.000	182.667	181.000	180.000	180.667	180.333	
229.333	226.000	212.667	204.333	193.667	193.000	192.000	185.333	180.333	178.667	181.333	181.667	
219.667	210.333	203.000	199.000	191.667	190.667	190.000	187.333	183.000	181.667	182.333	182.667	
210.000	201.000	190.667	174.333	159.000	157.000	159.333	181.667	182.667	184.000	183.667	181.667	
204.333	191.000	172.333	148.667	133.333	132.000	137.000	159.667	172.000	181.000	180.333	179.333	
198.000	170.333	147.333	133.333	127.333	123.667	126.000	138.333	159.667	175.667	176.333	177.667	
204.667	171.333	145.667	133.000	128.667	123.333	128.000	137.333	159.333	175.333	178.333	178.667	
201.000	170.333	149.333	142.000	138.000	130.000	134.333	148.000	166.667	178.667	180.333	180.667	
193.000	180.333	165.000	148.667	140.333	144.000	147.333	165.333	174.667	180.667	180.667	181.333	
193.000	187.333	180.000	171.667	168.667	171.000	172.000	175.667	182.333	183.667	180.667	179.667	
192.667	192.333	193.000	193.667	194.000	196.333	195.667	190.333	190.667	187.000	180.333	179.000	
192.000	191.667	194.667	197.667	200.000	202.000	201.333	197.333	195.333	188.667	180.667	176.667	
191.667	191.000	193.667	193.667	195.667	198.333	198.333	196.667	193.667	188.667	181.667	176.333	
191.667	193.000	194.000	193.667	194.333	195.667	196.667	193.667	191.667	188.667	183.667	177.333	
192.000	193.000	194.333	193.333	192.667	193.667	194.667	190.667	189.333	188.667	185.000	178.333	
193.000	193.000	193.333	192.333	190.667	190.667	190.667	189.333	186.667	186.333	184.667	178.667	
194.000	194.333	192.667	191.333	190.333	189.333	187.667	188.333	186.667	183.667	180.667	177.667	
195.000	193.667	191.667	190.333	189.333	189.667	187.667	187.333	186.667	183.667	181.667	180.000	
194.333	193.333	191.333	189.333	188.333	189.000	188.000	186.333	186.333	185.333	184.333	183.000	
194.667	194.333	192.333	190.333	188.333	189.000	189.000	187.000	187.000	187.000	187.000	186.000	
195.667	195.333	193.667	192.000	190.000	188.333	188.333	188.333	188.333	187.333	187.333	187.333	
196.333	195.000	195.000	193.333	191.000	189.333	189.333	189.333	188.667	187.667	187.667	188.333	
195.333	194.000	194.000	194.667	192.333	191.000	190.333	189.667	189.667	188.667	188.667	189.333	
196.000	193.000	193.000	193.000	192.333	192.000	190.667	190.667	190.333	189.667	189.667	191.333	
195.000	193.000	193.000	193.000	192.333	191.333	191.667	191.667	190.000	189.000	189.667	190.667	

(c)

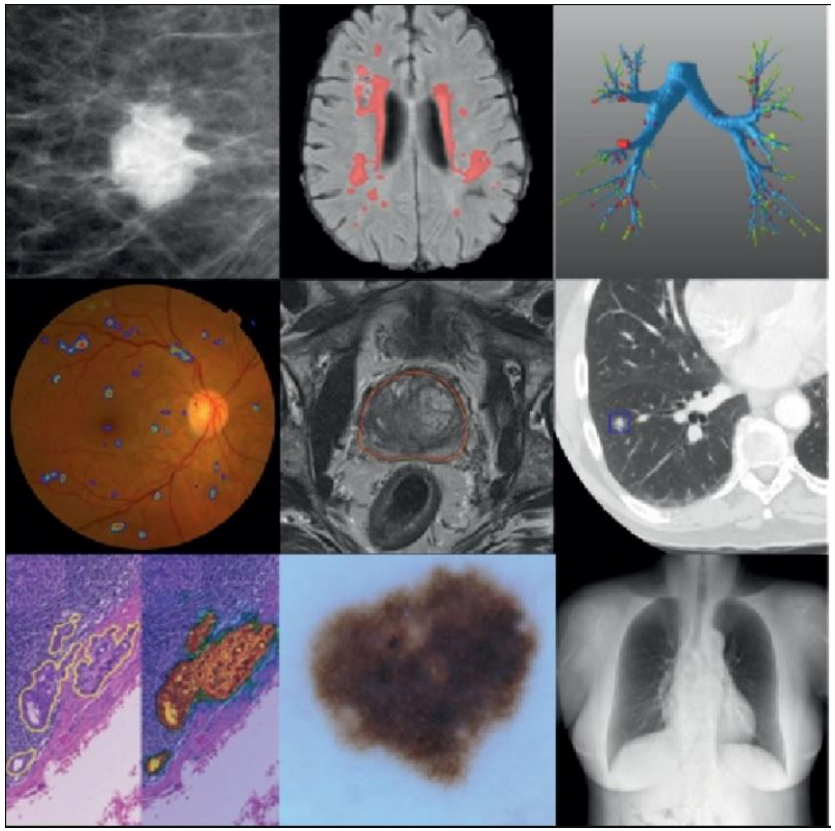
Plan

Evolution of the architectures



AISmartz.com : CNN Architectures Over a Timeline (1998-2019)

Medical applications [1]

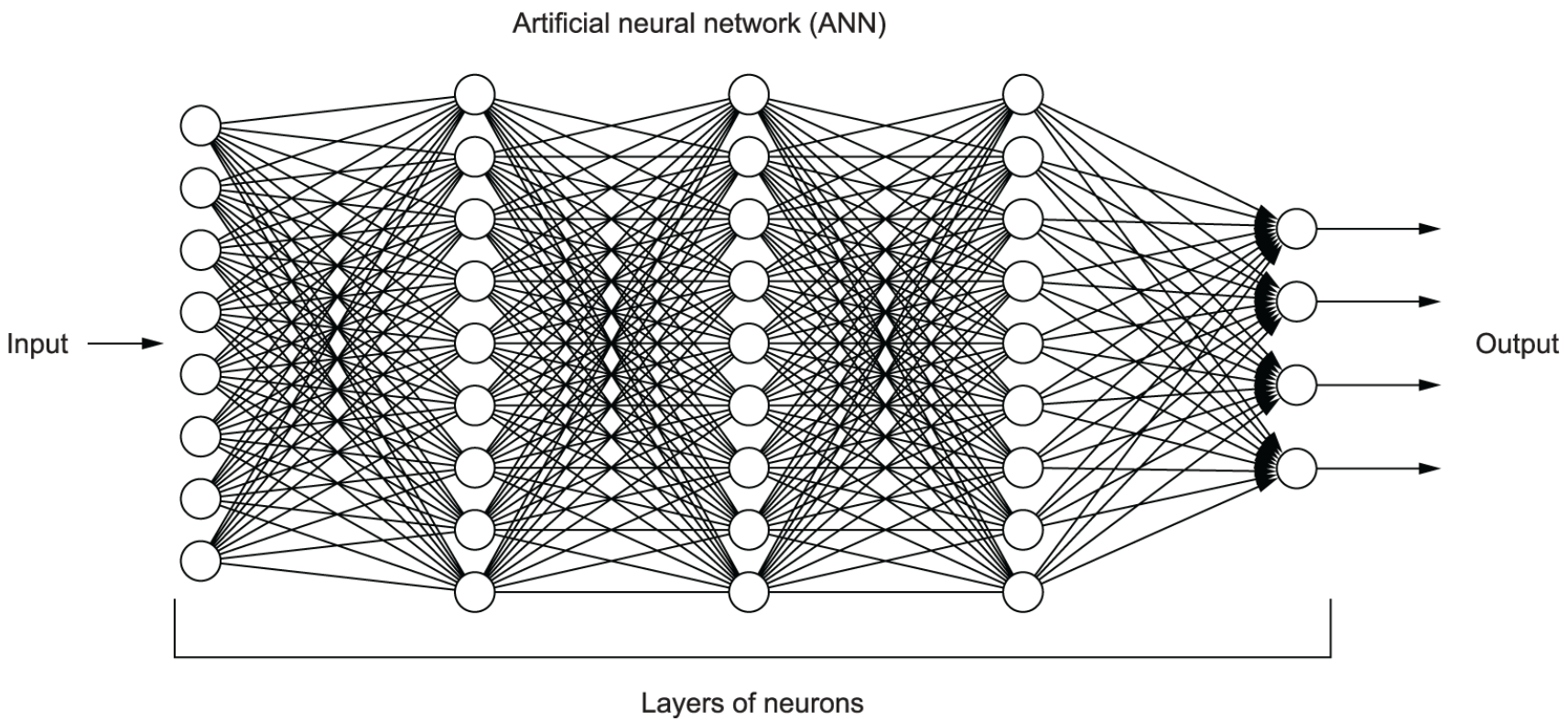


- Mammographic mass classification
- Segmentation of lesions
- Leak detection
- Diabetic retinopathy
- Prostate segmentation
- nodule classification
- Breast cancer detection
- Skin lesion
- Bone suppression

1. Litjens et al. "A Survey on Deep Learning in Medical Image Analysis" 2017

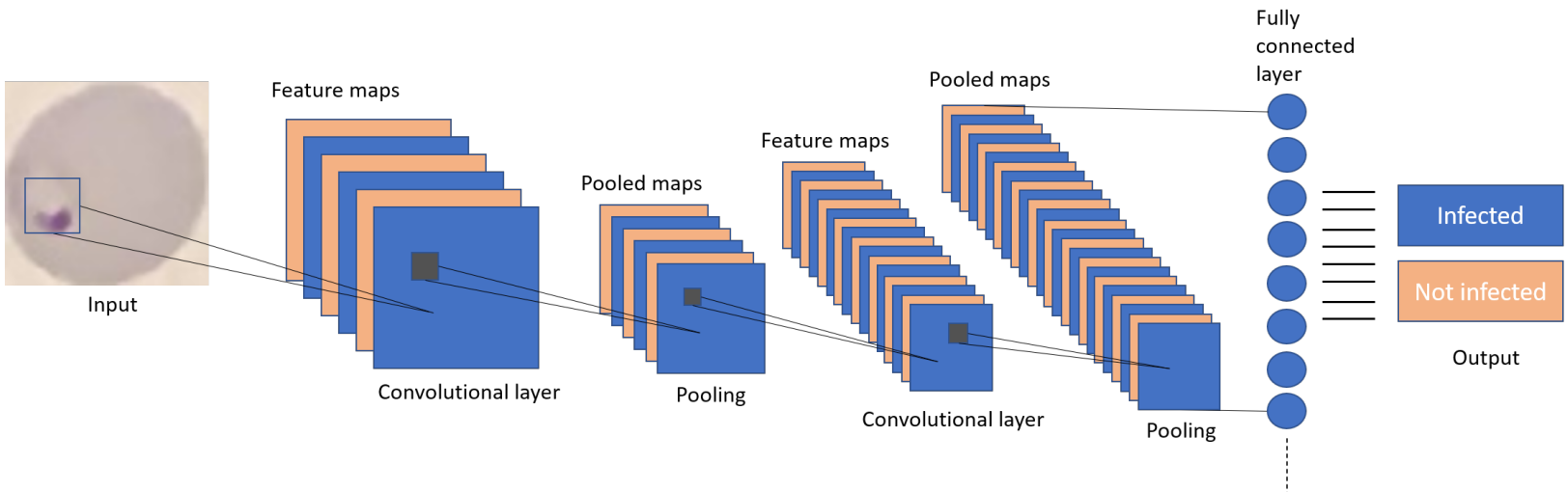
Simple ANN architecture

Different layers

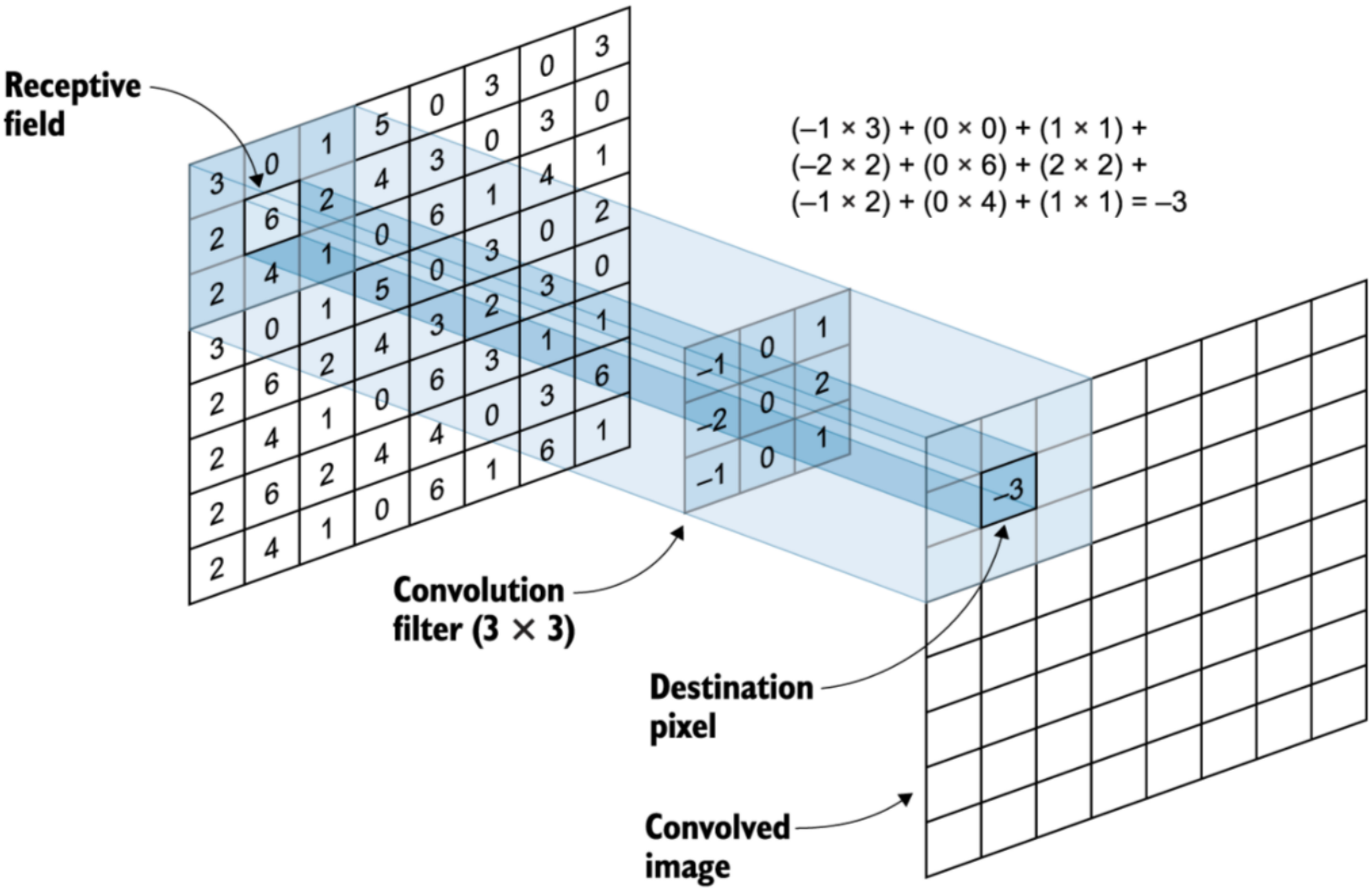


Simple CNN architecture

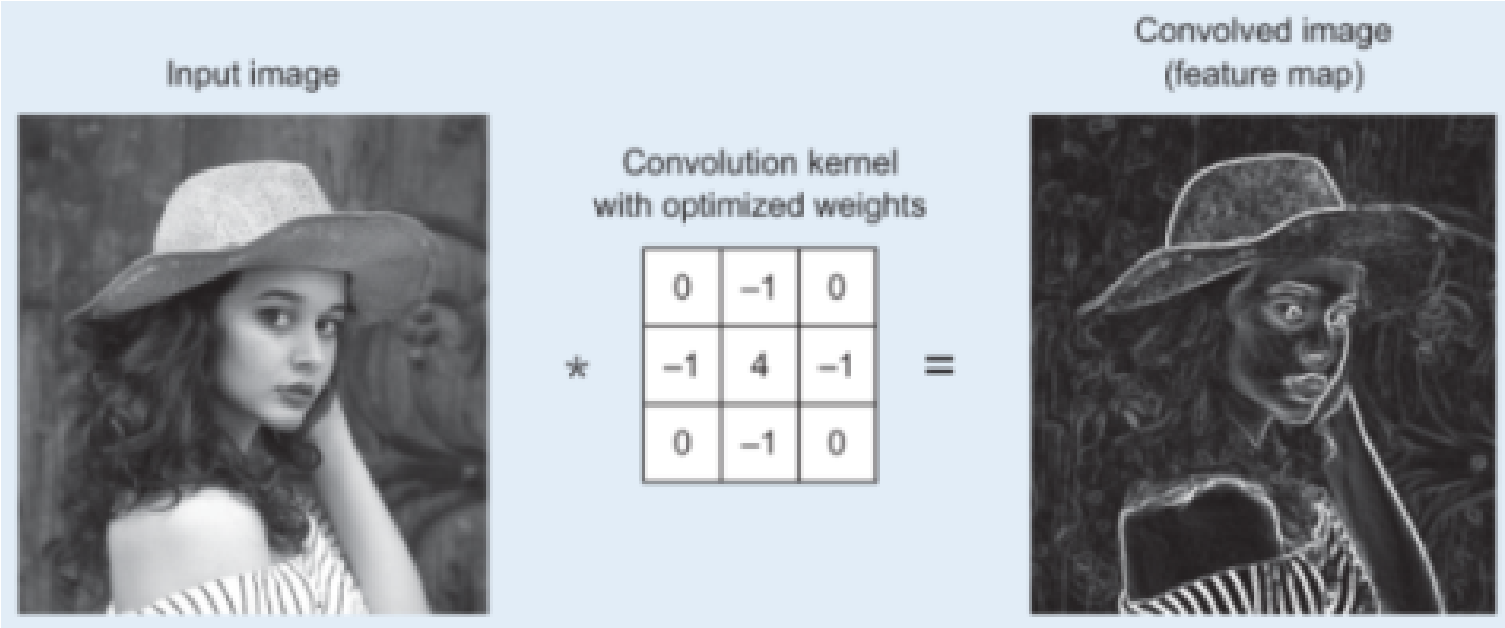
Different type of layers



Convolution

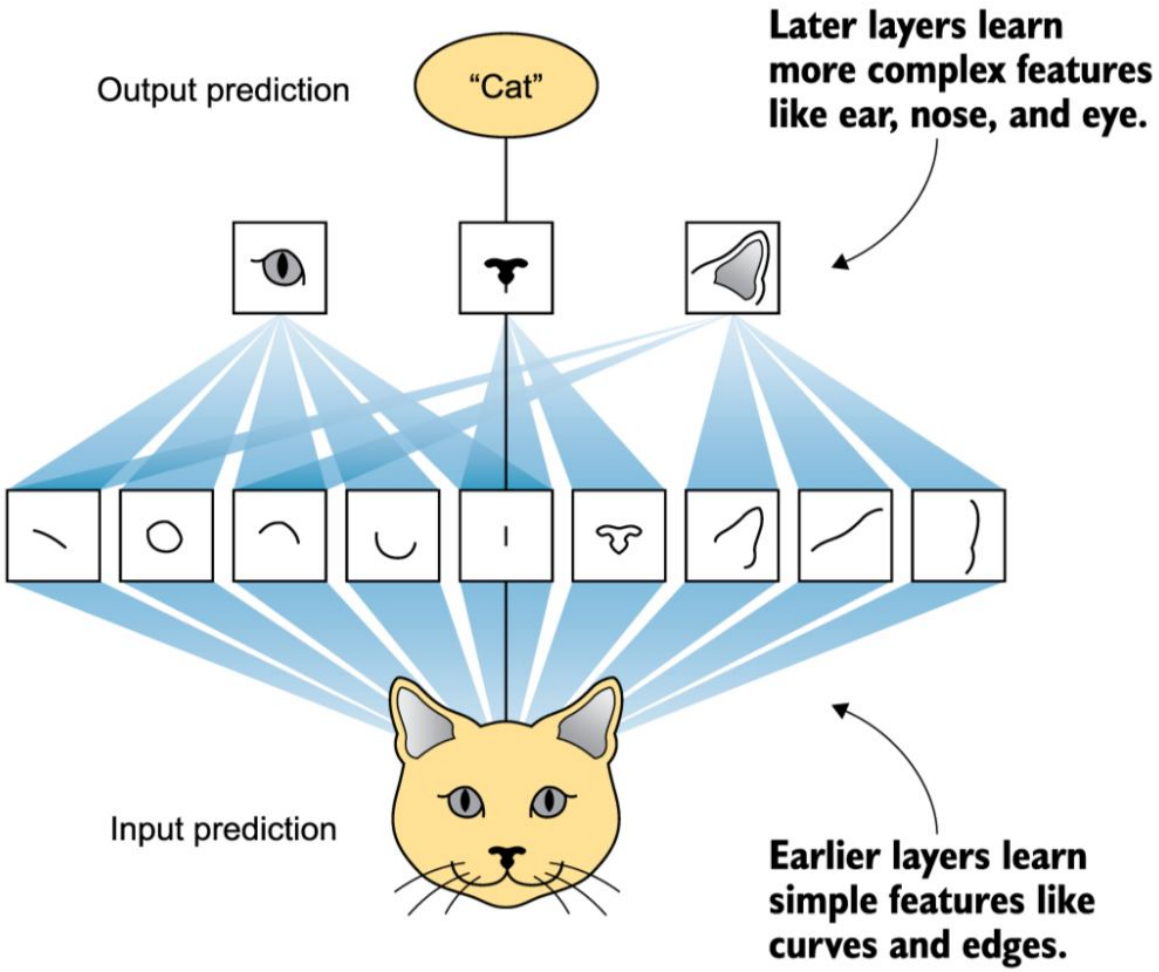


An edge detection kernel on an image (feature extraction)



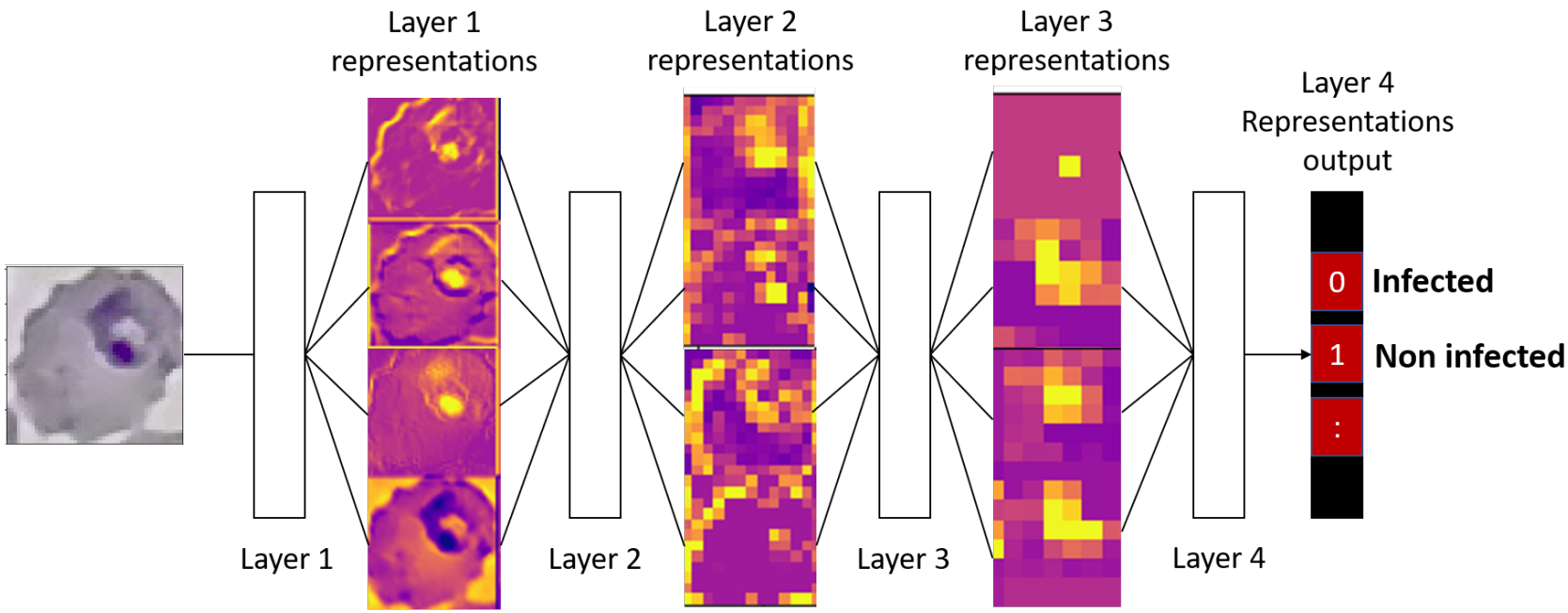
Feature extraction principle

During the training process, the model detect common features



Feature extraction

During the training process, the model identifies features (patterns)



Plan

Malaria

Some key facts

- Malaria is a life-threatening disease caused by parasites called "Plasmodium" that are transmitted to people through the bites of infected female Anopheles mosquitoes
- In 2020, there were an estimated 241 million cases of malaria worldwide.
- The estimated number of malaria deaths stood at 627 000 in 2020.
- Children under 5 accounted for about 80% of all malaria deaths in the Region

The fight against malaria

2 strategies

- Direct strategy : Accurate diagnostic and good treatment
- Indirect strategy : Reduce or stop transmission by vectors

The momentum ?

- Data sharing
- Computational biology
- Better access to internet
- Better performing smartphones/cameras

Clinical needs

CNR du Paludisme

Clinical malaria

What information is expected by the physician

1/ Malaria infection?

2/ What species?

3/ What stage of development?

4/ What parasitaemia?

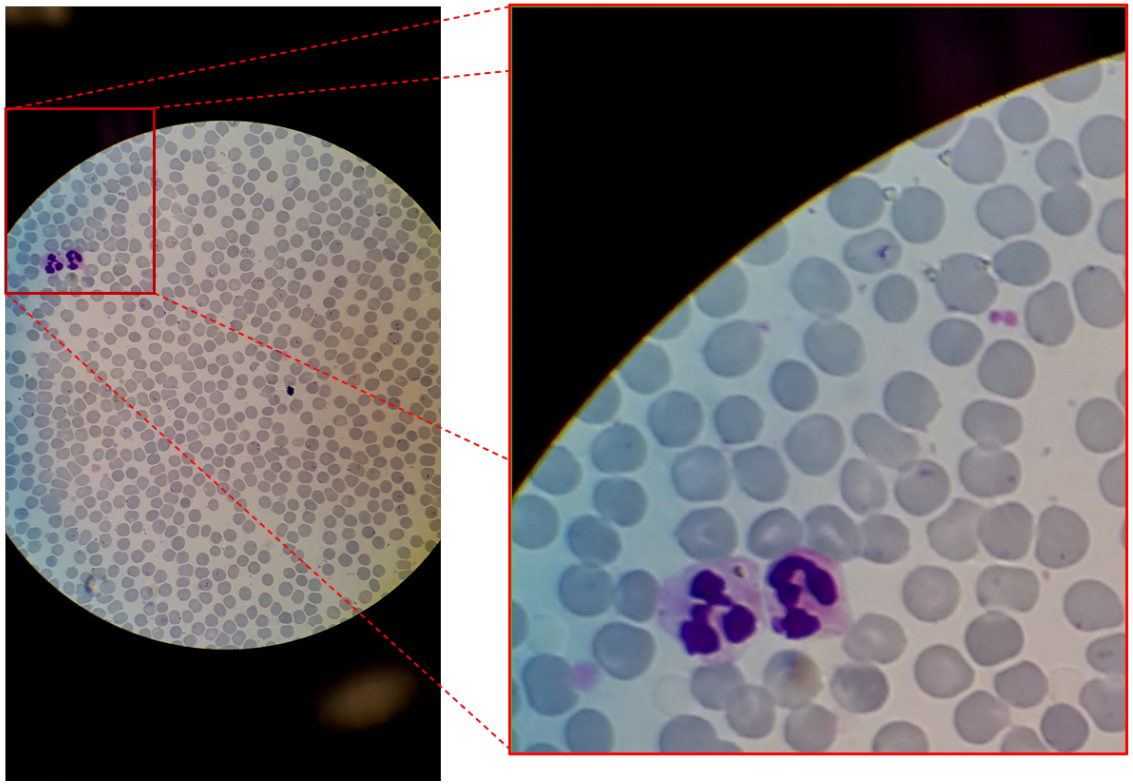
5/ How long is the result?

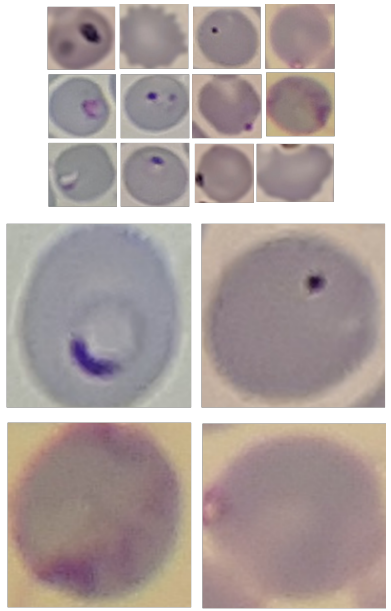
Less than 2 hours

2. From Dr Marc Thellier (AP-HP)

Problem statement

Identify infected RBCs





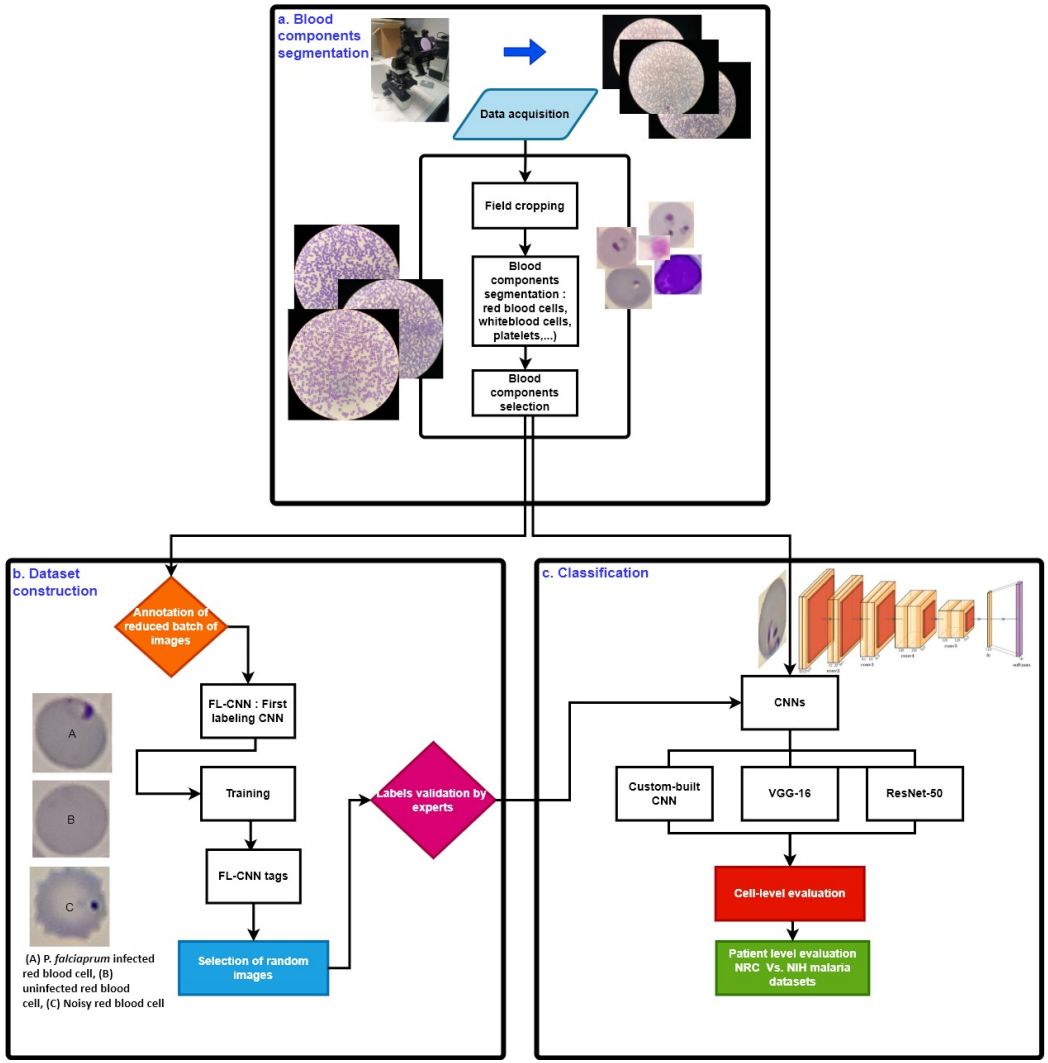
State of the art

A field of interest !

	Source	Accuracy	Sensitivity	Specificity	Dataset
	Ross et al. (2006)	73.0	85.0	NR	Private
	Das et al. (2013)	93.2	94.0	87.9	Private
	Adi et al. (2016)	87.1	NR	NR	Private
CNN 6 convs	*Liang et al. (2017)*	97.3	96.9	97.8	NIH
	Dong et al. (2017)	98.1	97.3	98.7	Private
	Peñas, Rivera & Naval Jr (2017)	92.4	95.2	84.7	Private
	Gopakumar et al. (2017)	97.7	NR	NR	Private
ResNet-50	*Rajaraman et al. (2018)*	98.6	98.1	99.2	NIH
VGG-16	*Rahman et al. (2019)*	97.7	97.4	97.9	NIH
Ensemble	*Rajaraman, Jaeger & Antani (2019)*	99.5	NR	NR	NIH

System design

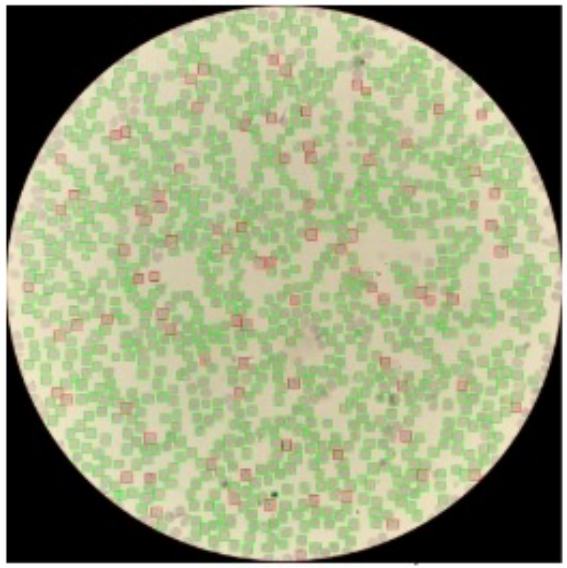
Different modules



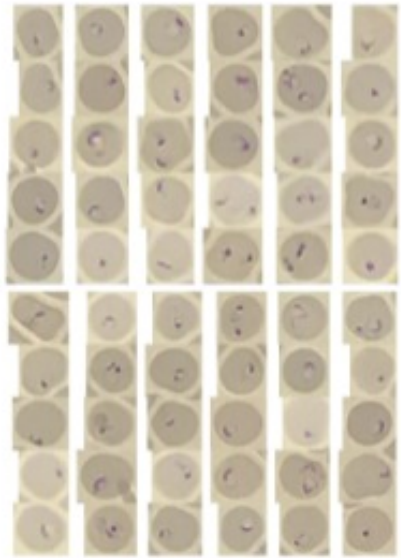
How it works ?

Workflow

Smartphone capture of *P. falciparum infected* thin blood smear with 500x magnification

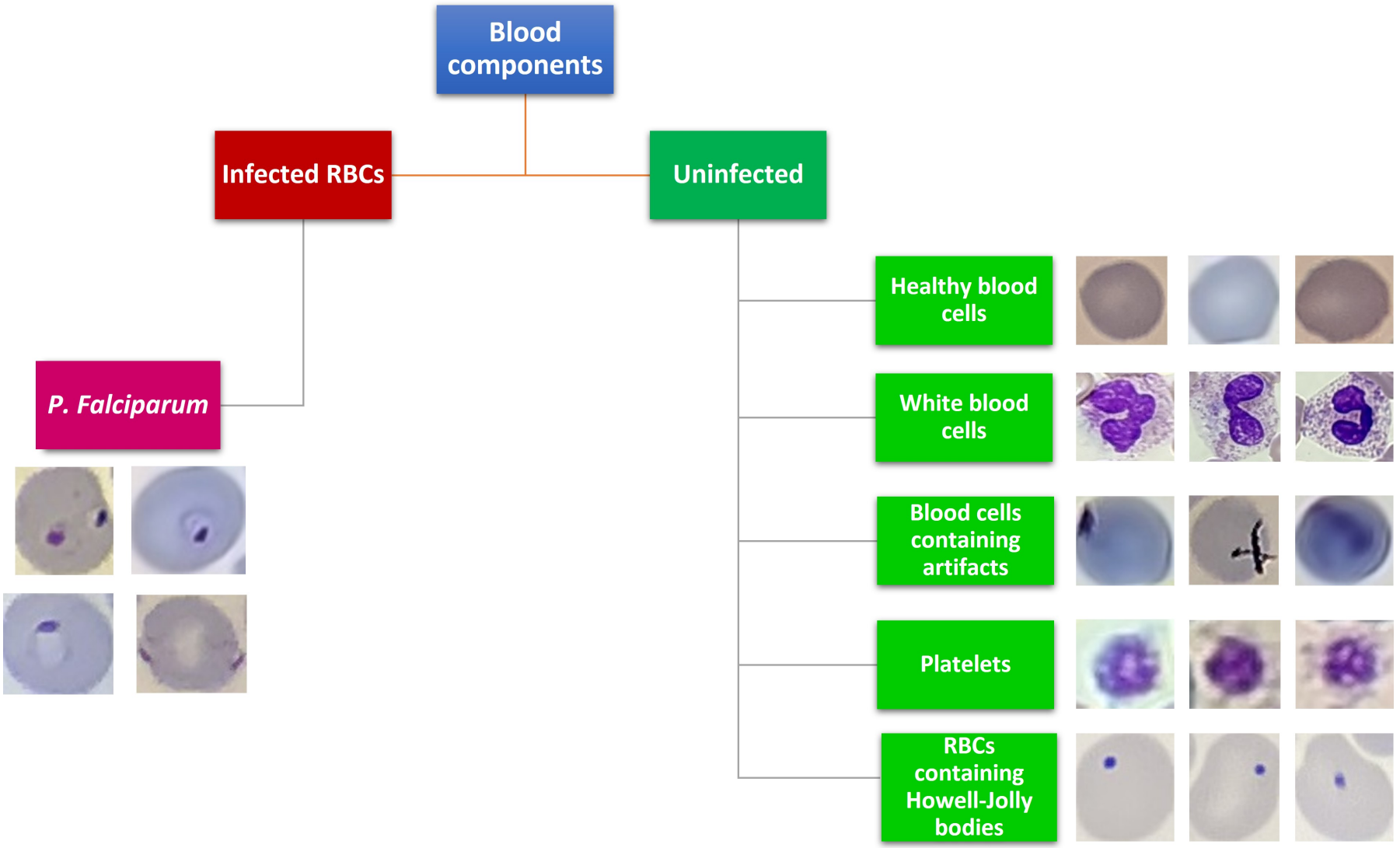


Boundig boxes of blood cells classification into infected red blood cell and uninfected blood component



Samples of *P. falciparum* infected red blood cells images

Biological diversity of the NRC dataset



Plan

Training

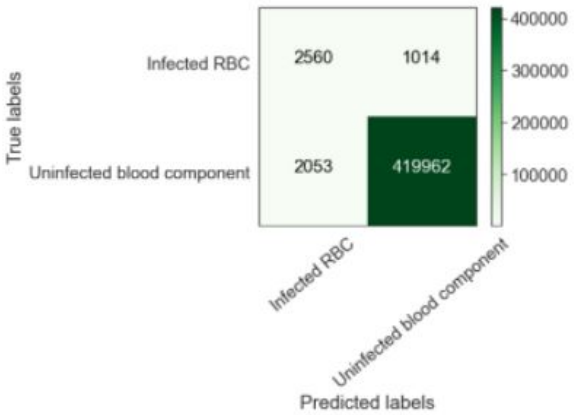
- 101 p.f infected and 101 uninfected patients
- 13,453 P. falciparum infected red blood cells and 54,705 uninfected blood components

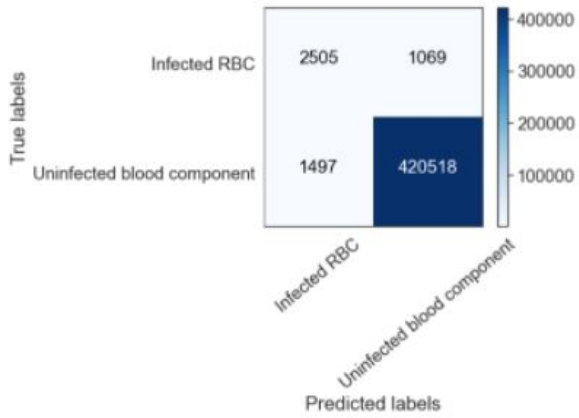
Method	Accuracy		Precision (ppv)		Sensitivity		Specificity		F1-Score	
	NRC	NIH	NRC	NIH	NRC	NIH	NRC	NIH	NRC	NIH
C-CNN	0.966 ± 0.0014	0.953 ± 0.0036	0.90 ± 0.0109	0.966 ± 0.0048	0.932 ± 0.0135	0.938 ± 0.0039	0.974 ± 0.0037	0.952 ± 0.0040	0.916 ± 0.0048	0.952 ± 0.0040
VGG-16	0.988 ± 0.0044	0.985 ± 0.0005	0.971 ± 0.0116	0.986 ± 0.0120	0.974± 0.0185	0.984 ± 0.1356	0.992 ± 0.0026	0.986± 0.0097	0.971 ± 0.014	0.986 ± 0.0120
ResNet-50	0.993 ± 0.0018	0.982 ± 0.0005	0.978 ± 0.0040	0.98 ± 0.0089	0.988 ± 0.0048	0.984± 0.0080	0.994 ± 0.0015	0.980 ± 0.0078	0.983 ± 0.0048	0.980 ± 0.0089

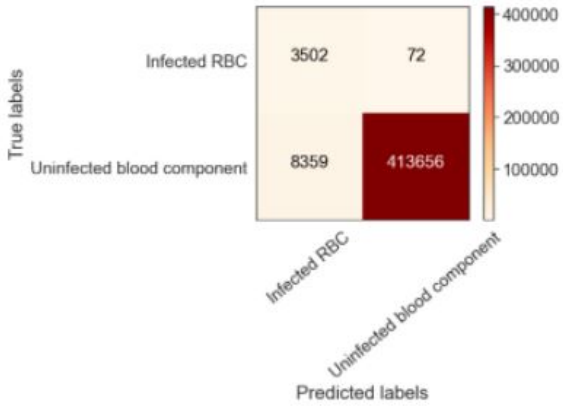
NIH malaria dataset :
https ://lhncbc.nlm.nih.gov/LHC-publications/pubs/MalariaDatasets.html

Patient-level inference : 200 patients (150 uninfected et 50 P.f infected) : VGG-16-NIH Vs VGG-16-NRC

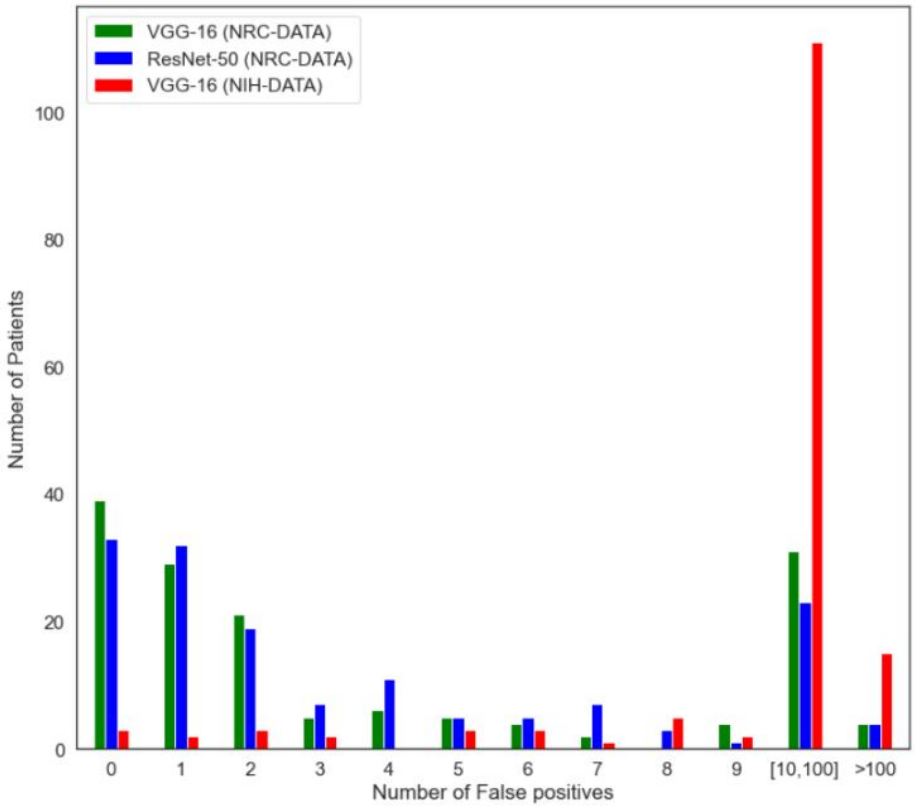
Confusion matrices







Patient-level validation 200 patients (150 uninfected et 50 P.f infected) : VGG-16-NIH Vs VGG-16-NRC



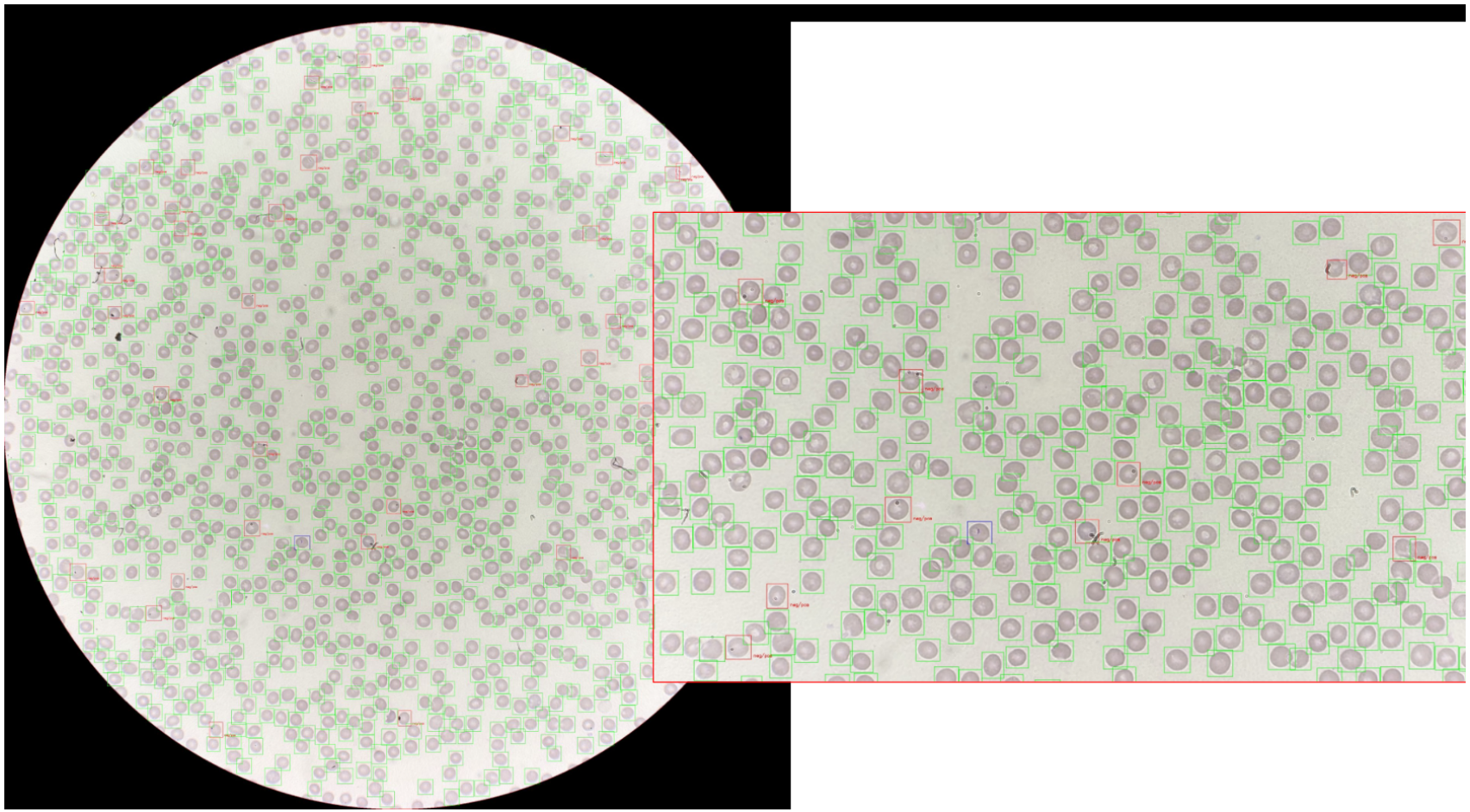
(a) Samples of false positives images of VGG-16-NRC

(b) Samples of false positives images of ResNet-50-NRC

(c) Samples of false positives images of VGG-16-NIH

Patient-level validation 200 patients (150 uninfected et 50 P.f infected) : VGG-16-NIH Vs VGG-16-NRC

Positif



Plan

What's next ?

Some limits

- The lack of quality
- Staining variability
- Artifacts diversity (false positives)

In progress

- Parasite density estimation
- Data augmentation
- Label free

Some tips

- Always check misclassified samples
- Change image representation (RGB to grayscale for example) to reduce computational complexity
- Use your best trained model to label new data
- Start training with a small dataset to explain your models
- Always think about real-life conditions (unbalanced data, generalizability...)
- Communicate with experts and data producers

"You can have data without information, but you cannot have information without data." (Daniel Keys Moran)

Questions ?