Data augmentation & transformations

- One of GaussianBlur and MedianBlur
- One of Sharpen
- Flip
- Rotation between (10, 20)deg

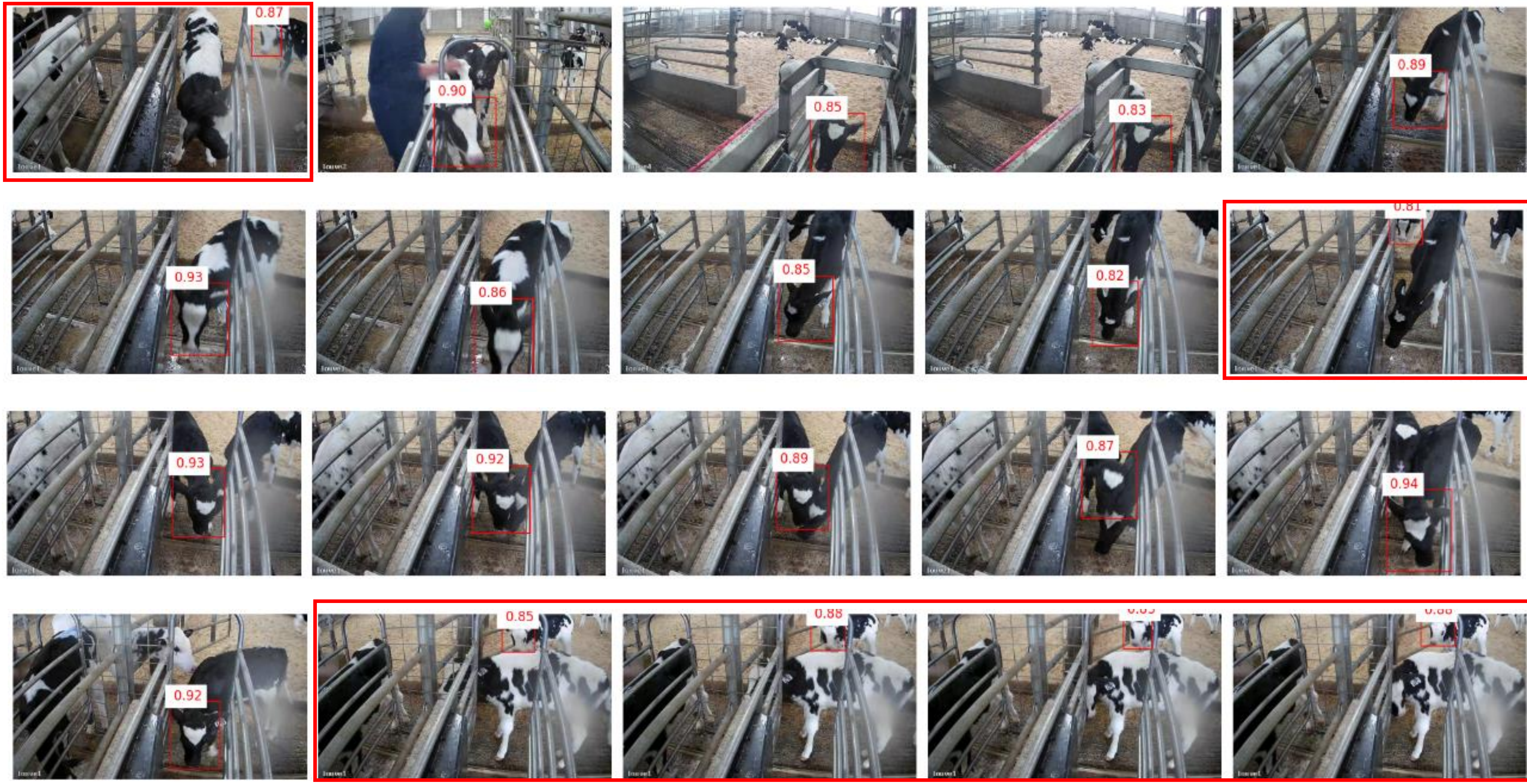
Performances

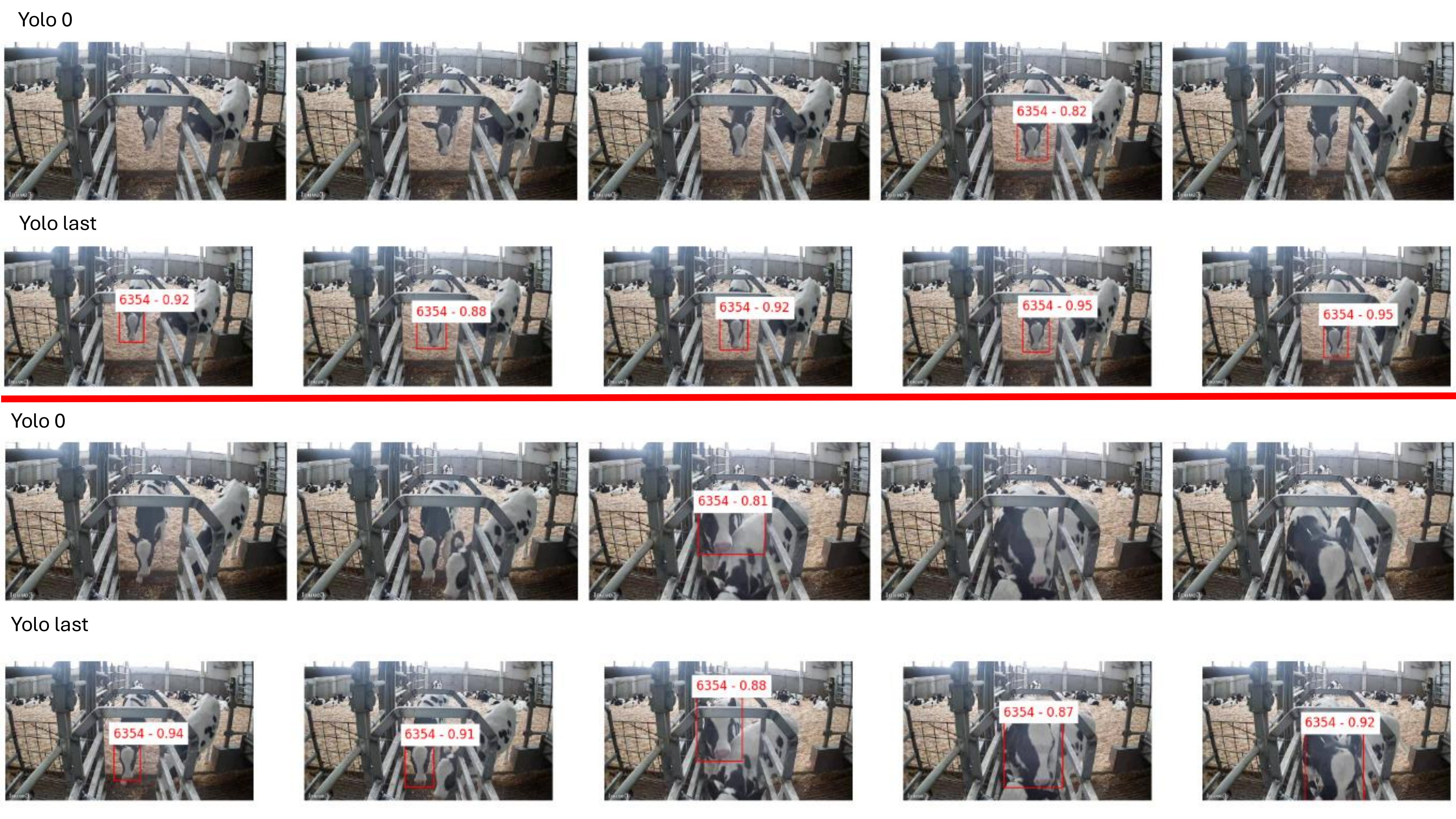
- Generate 5 transformed version of each image in original training set
- From 178 to 890 images

P	R	mAP50	MAP50 - 90
1.0	0.769	0.891	0.502

```
# total videos extracted, then total videos with a least one detection by y_face, and number of images extracted by y_face 3
records[0].shape, records[0][records[0]['nfaces'] > 0].shape, records[1].shape
```

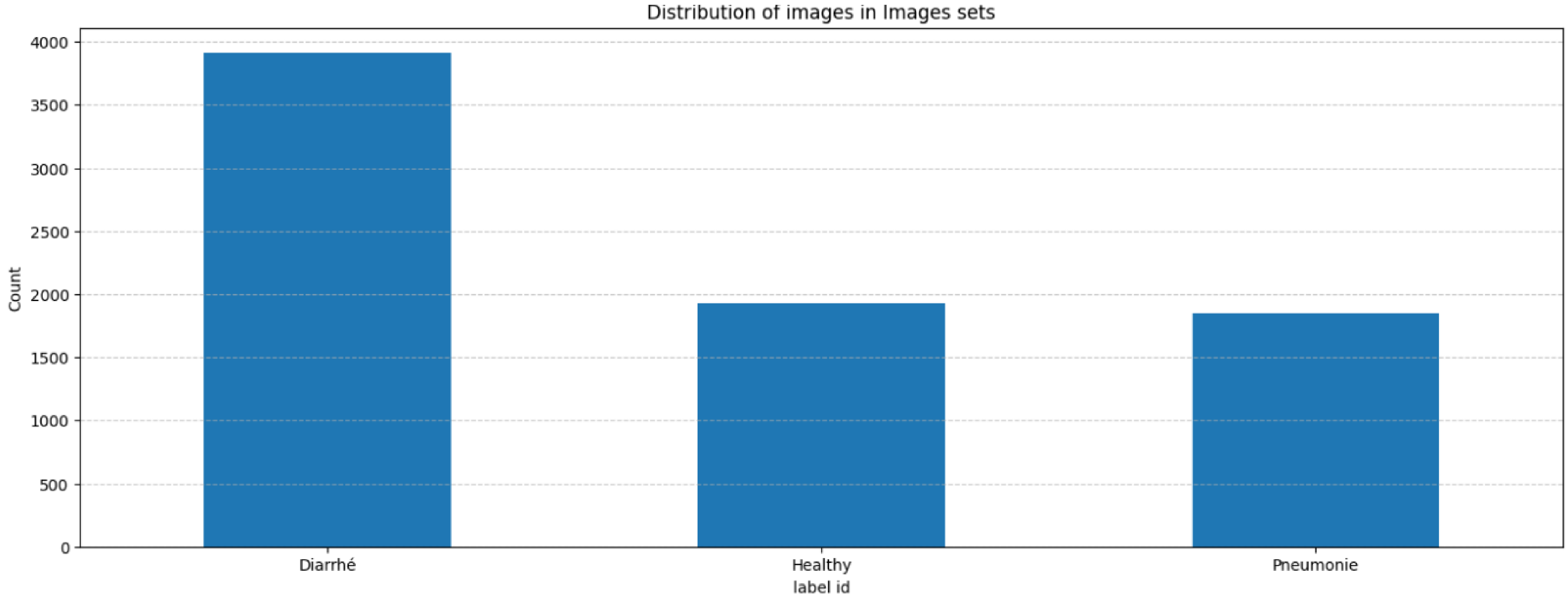
```
((96, 9), (49, 9), (346, 12))
```


0.87
0.90
0.85
0.83
0.89
0.93
0.86
0.85
0.82
0.93
0.92
0.89
0.87
0.94
0.92
0.85
0.88


Yolo 0
6354 - 0.82
Yolo last
6354 - 0.92
6354 - 0.88
6354 - 0.92
6354 - 0.95
6354 - 0.95
Yolo 0
6354 - 0.81
Yolo last
6354 - 0.94
6354 - 0.91
6354 - 0.88
6354 - 0.87
6354 - 0.92

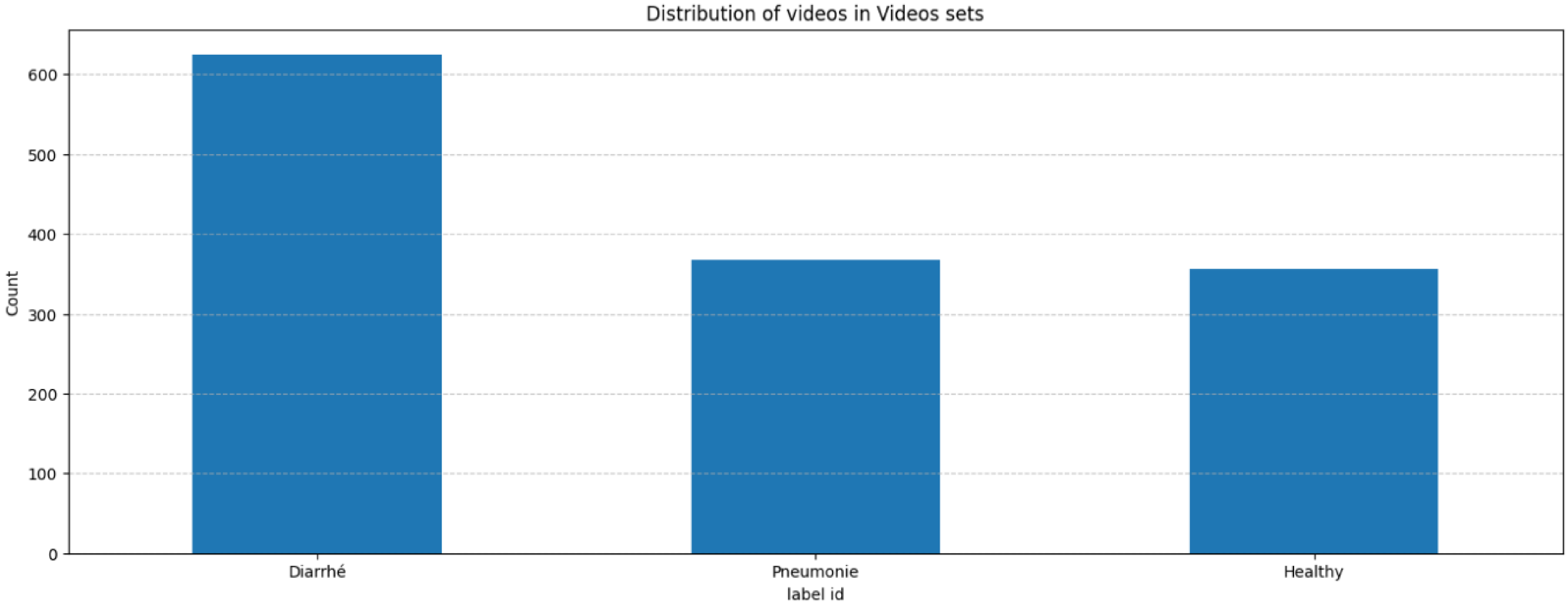
Données extraites:

- 1349 Vidéos et 7687 Images
- Contenant 76 veaux uniques



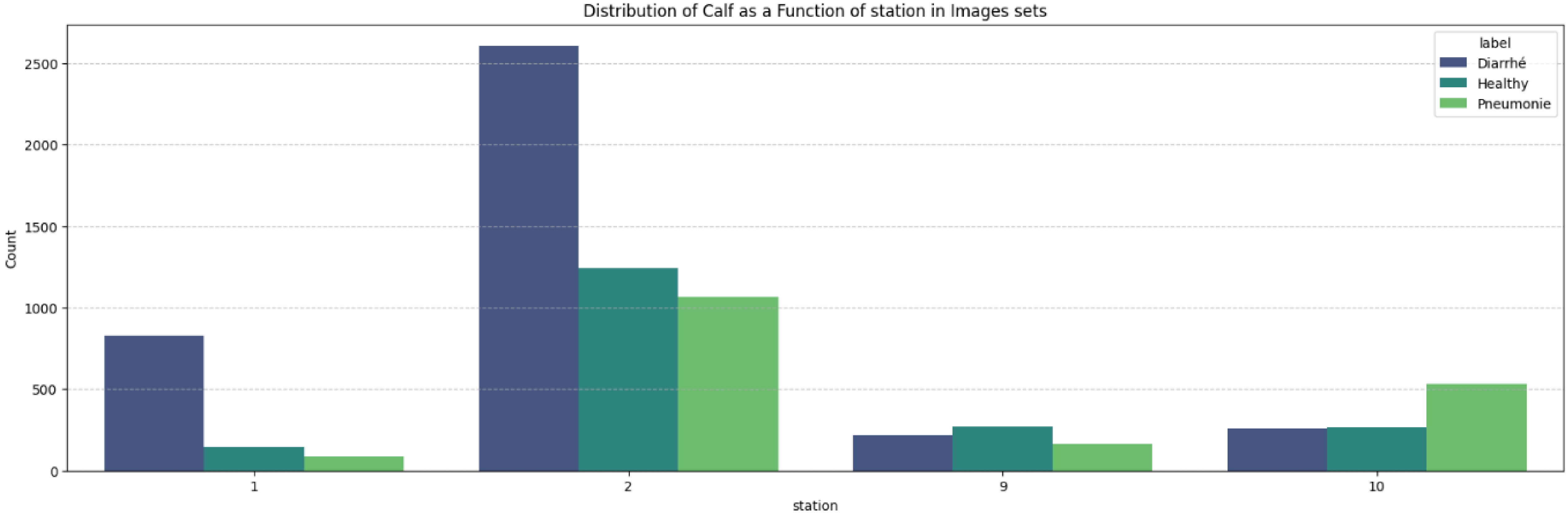
Données extraites:

- 1349 Vidéos et 7687 Images
- Contenant 76 veaux uniques



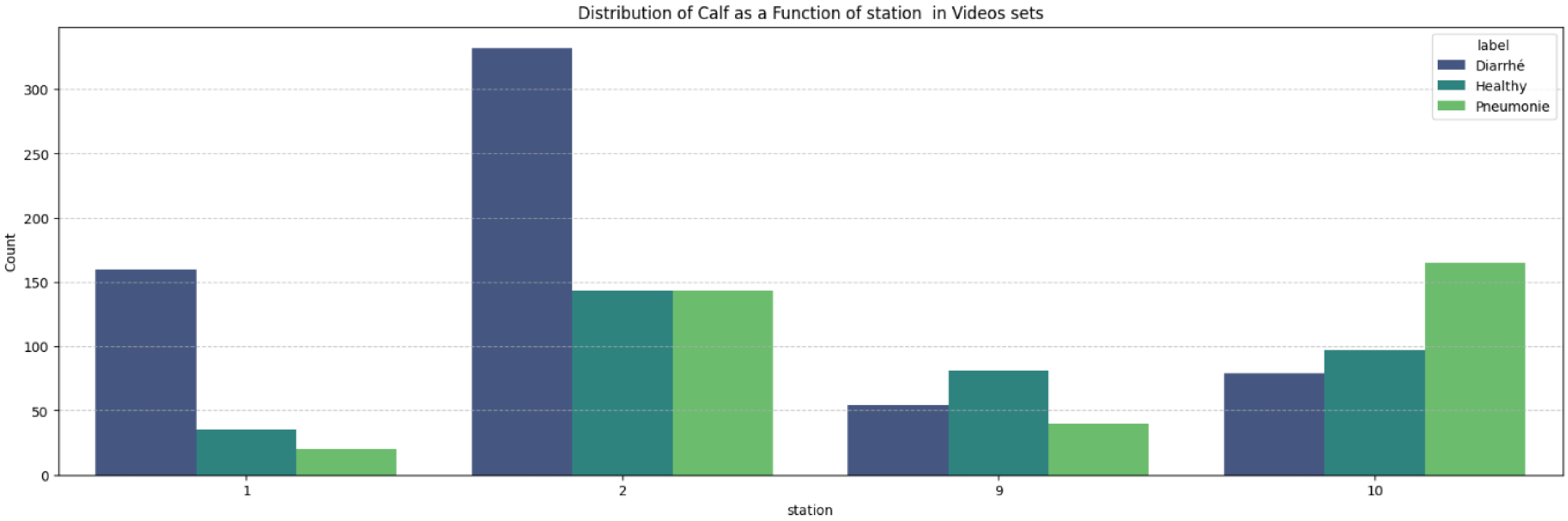
Données extraites:

- 1349 Vidéos et 7687 Images
- Contenant 76 veaux uniques



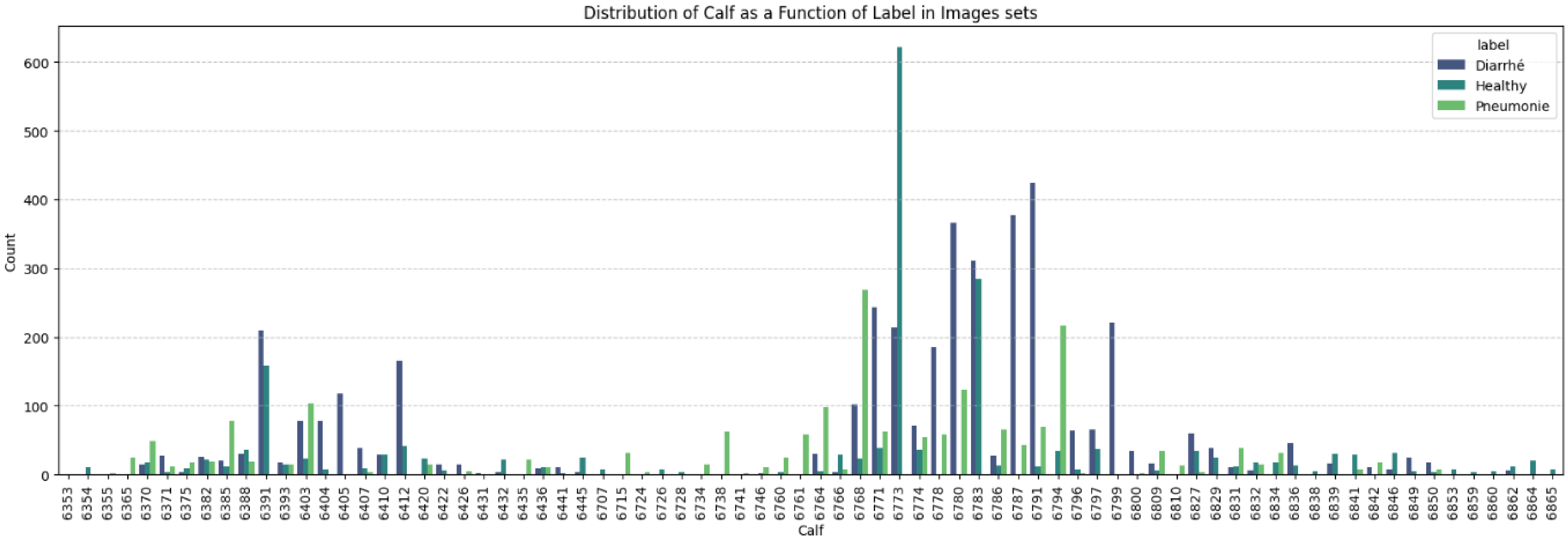
Données extraites:

- 1349 Vidéos et 7687 Images
- Contenant 76 veaux uniques



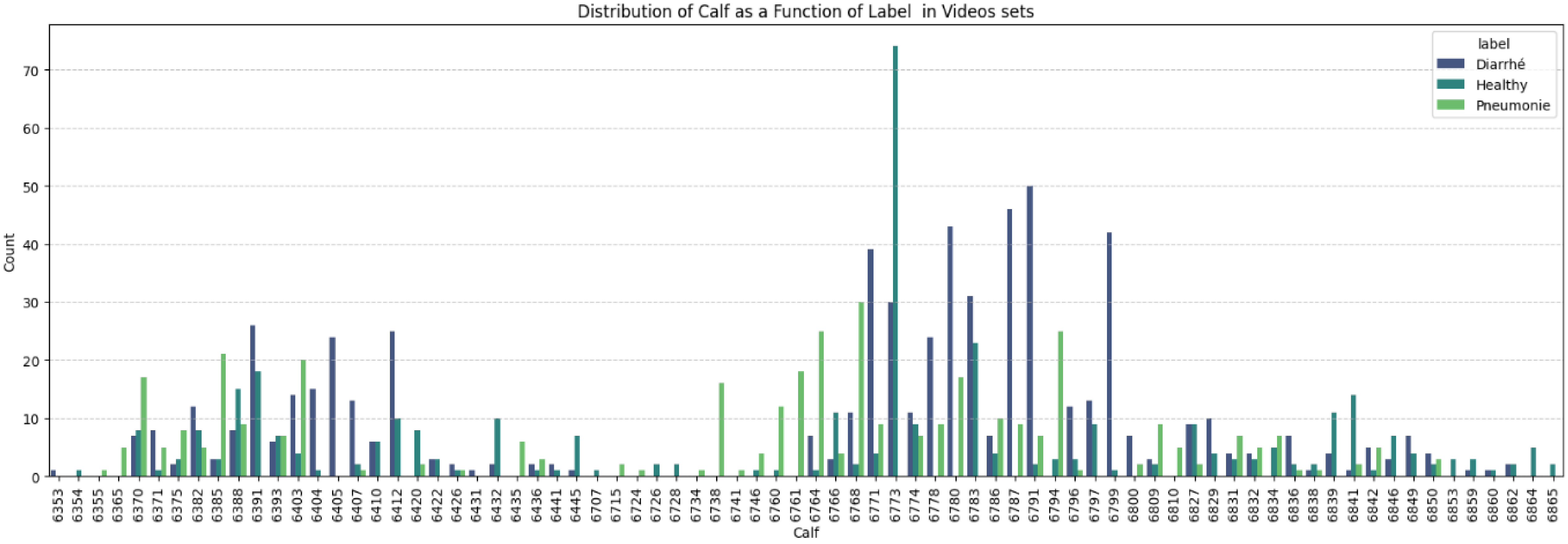
Données extraites:

- 1349 Vidéos et 7687 Images
- Contenant 76 veaux uniques

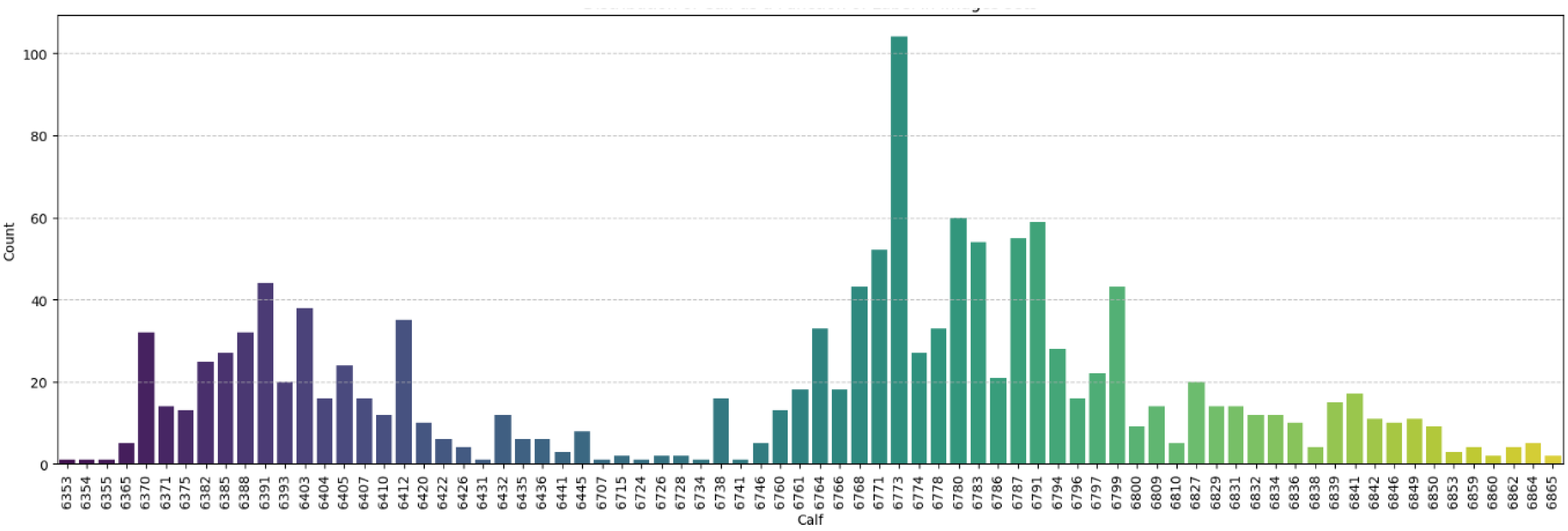


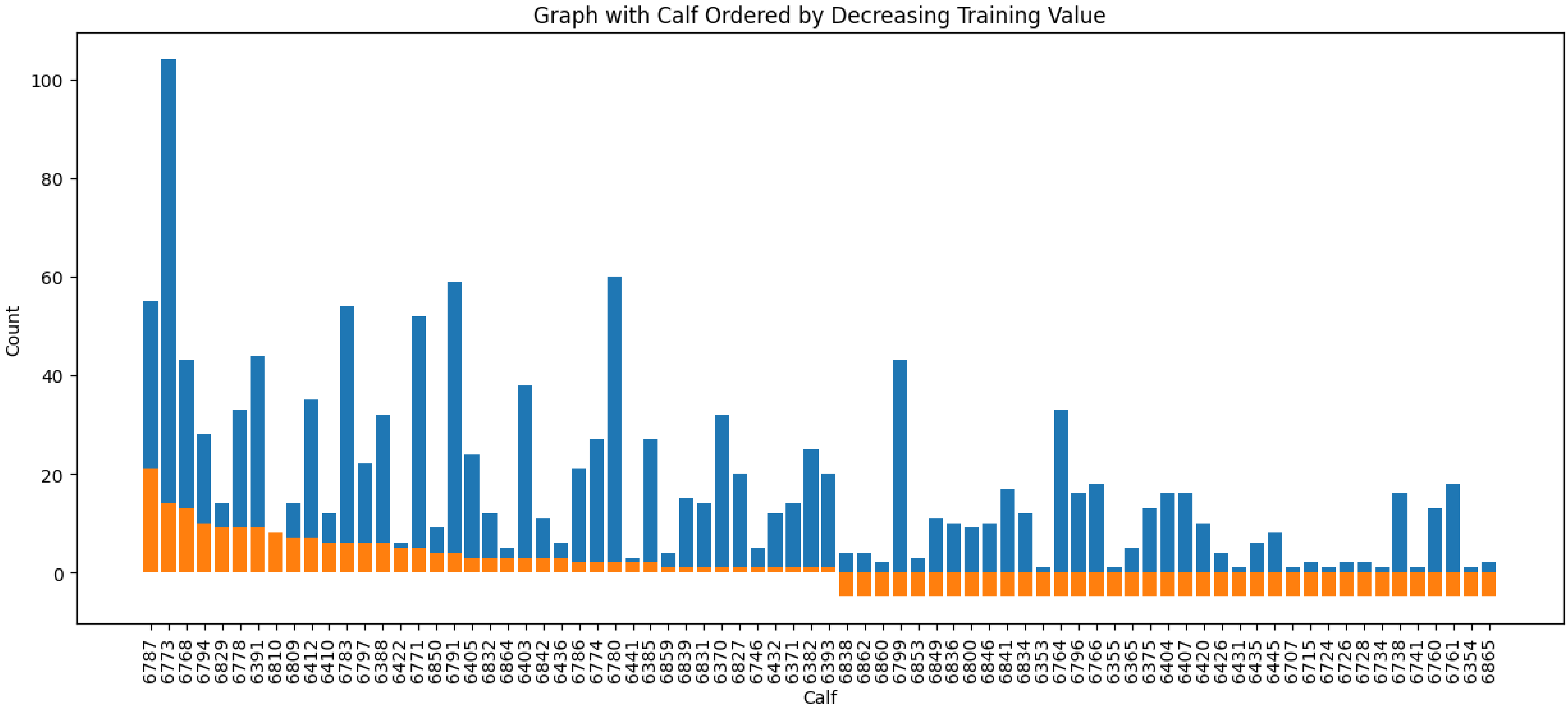
Données extraites:

- 1349 Vidéos et 7687 Images
- Contenant 76 veaux uniques

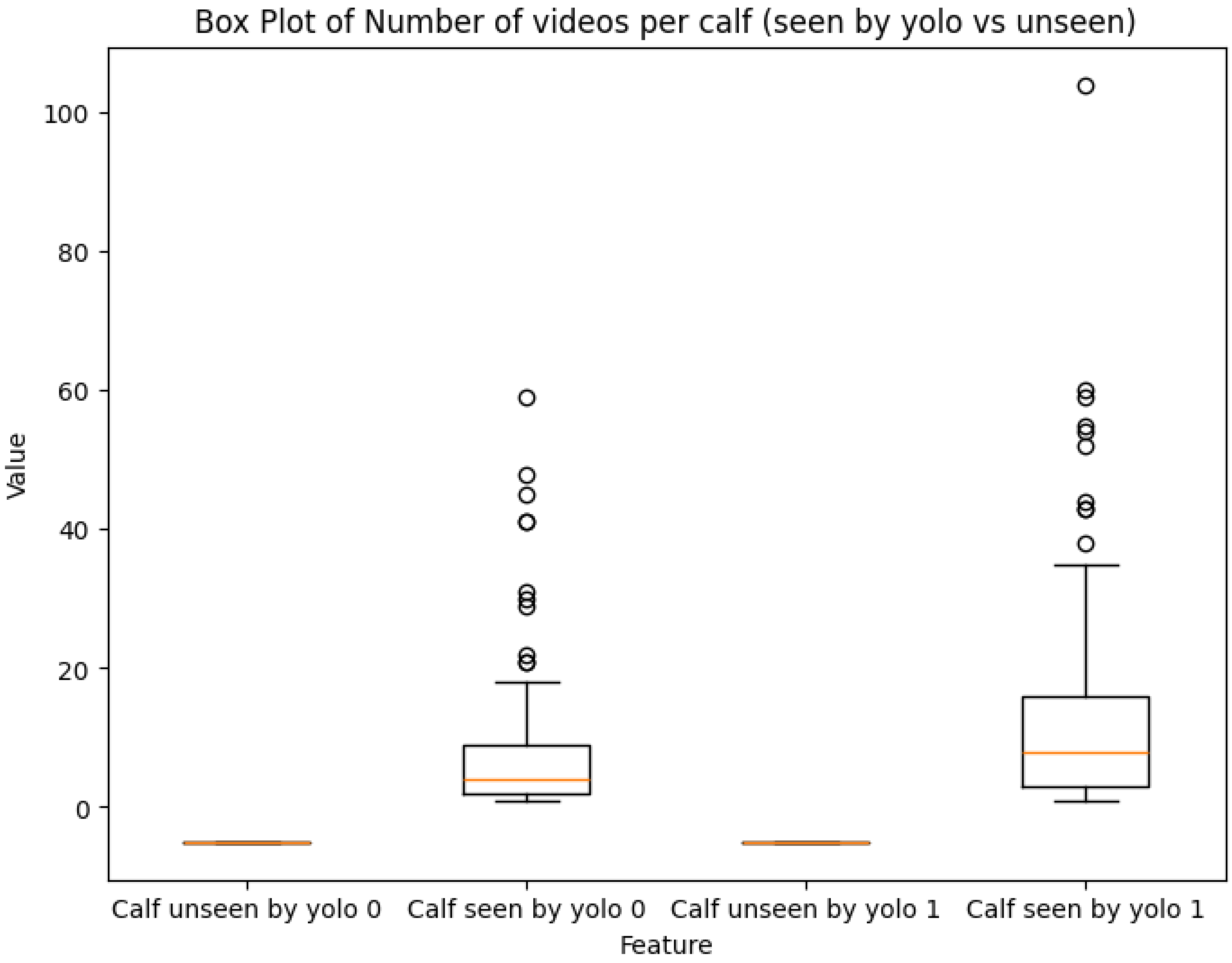


Number of videos containing each calf





Graph with Calf Ordered by Decreasing Training Value
Count
100
80
60
40
20
0
Calf
6787
6773
6768
6794
6829
6778
6391
6810
6809
6412
6410
6783
6797
6388
6422
6771
6850
6791
6405
6832
6864
6403
6842
6436
6786
6774
6780
6441
6385
6859
6839
6831
6370
6827
6746
6432
6371
6382
6393
6838
6862
6860
6799
6853
6849
6836
6800
6846
6841
6834
6353
6764
6796
6766
6355
6365
6375
6404
6407
6420
6426
6431
6435
6445
6707
6715
6724
6726
6728
6734
6738
6741
6760
6761
6354
6865




```
w = stats.mannwhitneyu(join_df[join_df["count"] < 0]["count"].to_list(), join_df[join_df["count"] >= 0]["count"].to_list(), alternative='two-sided')
w.pvalue, w.pvalue < 0.05
```

```
(3.2651912110895236e-20, True)
```

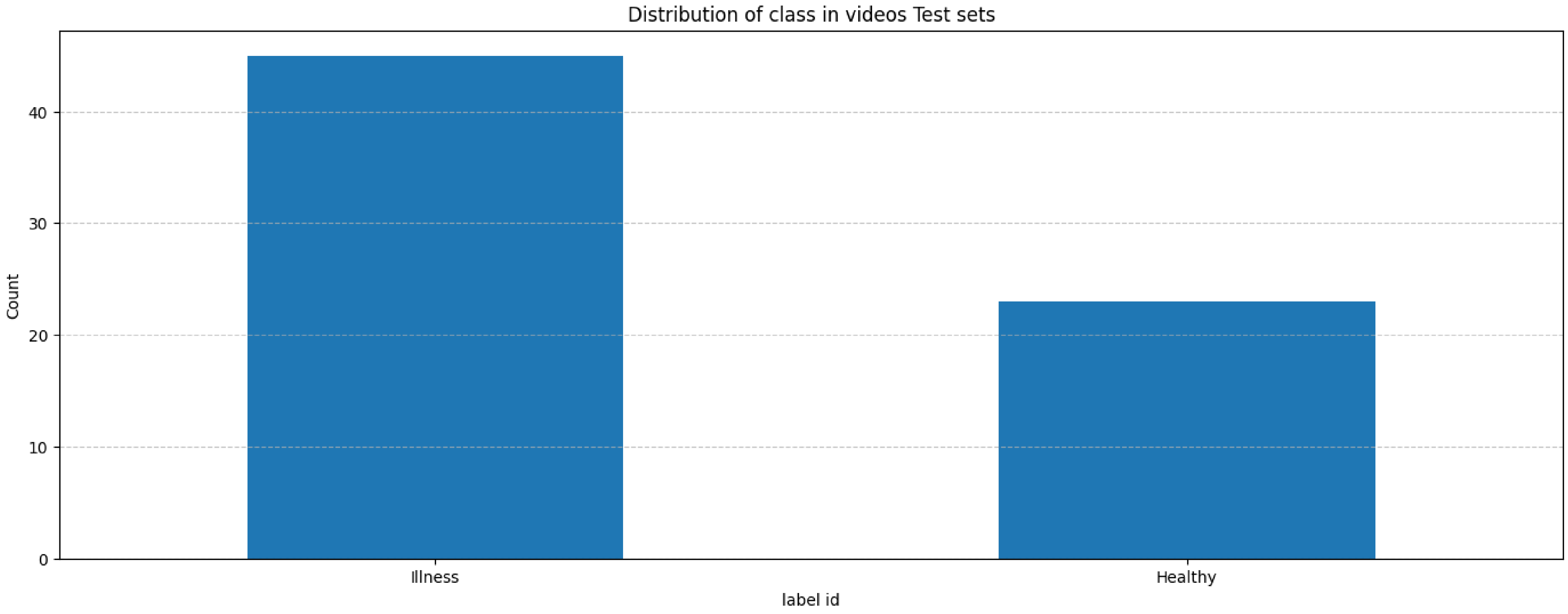
Videos Training set details

Videos	Leave One Calf out	All
Training set	207	80% of 1215
Validation set	1008 (only use 10-20%)	20% of 1215
Test set	68	
Calf number in Train + Val set	44	
Calf number in Test set	24	

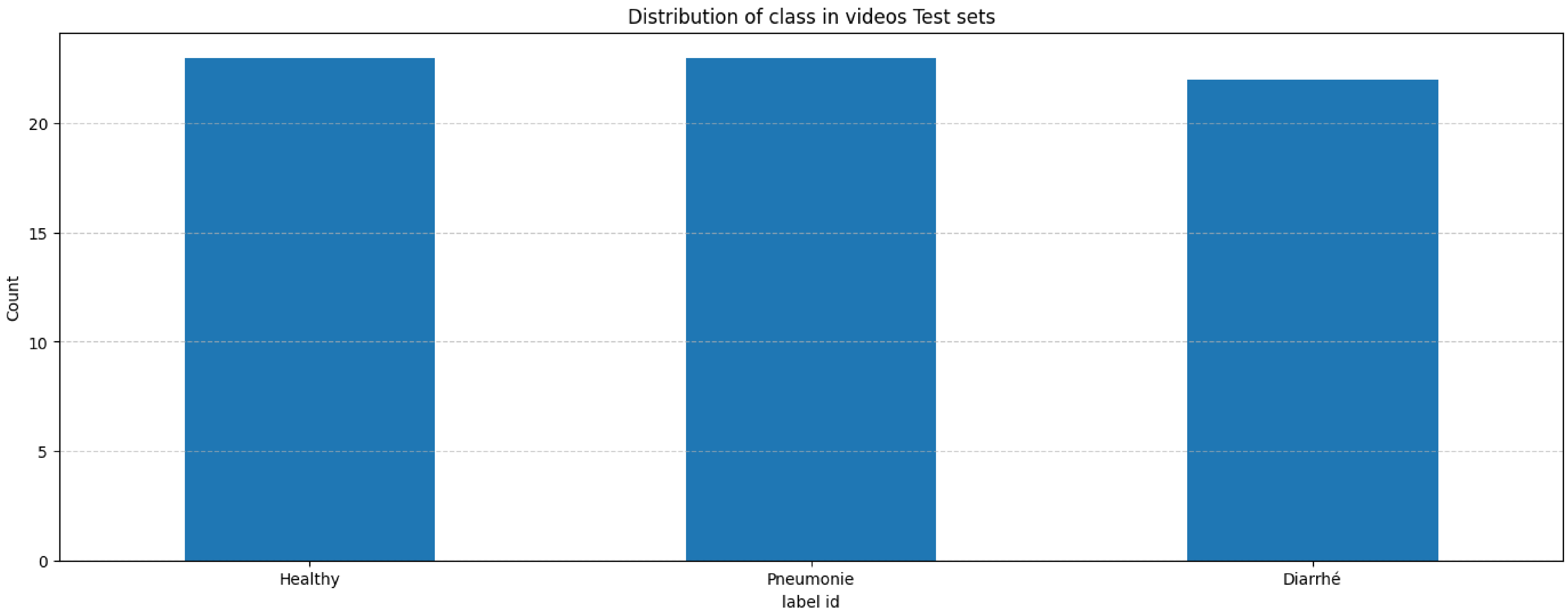
Images Training set details

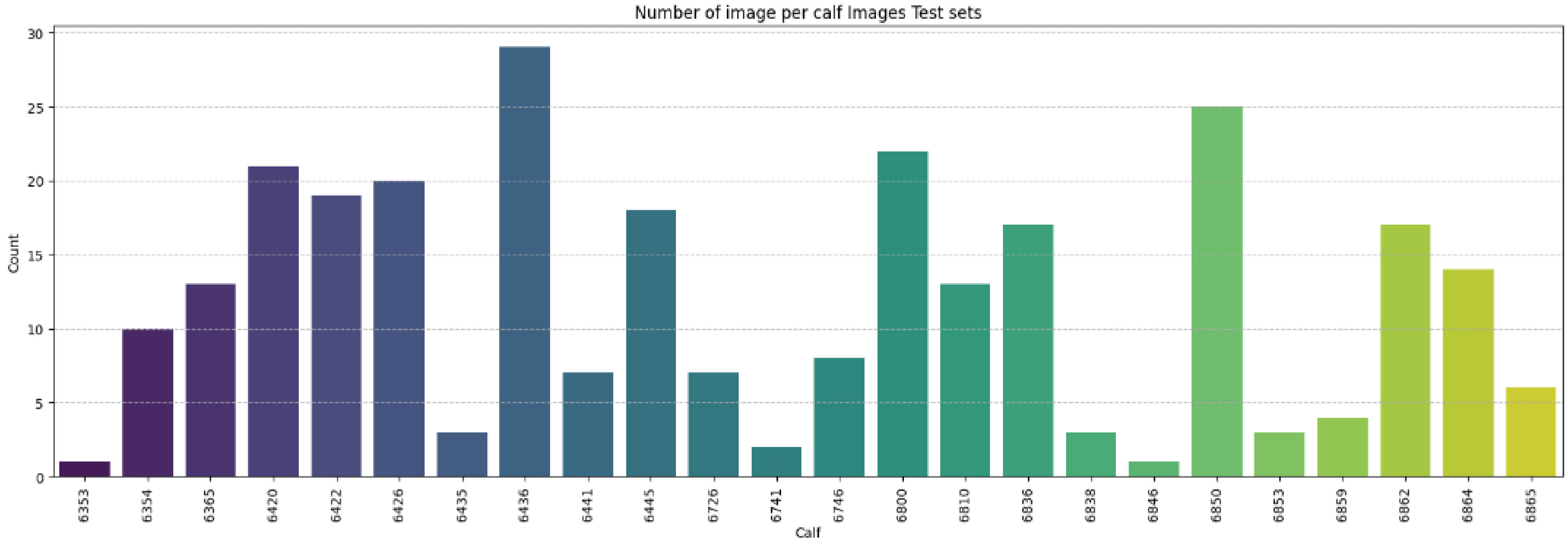
Images	Leave One Calf out	All
Training set	213	80% of 7149
Validation set	6936 (only use 10-20%)	20% of 7149
Test set	283 from videos test set	
Calf number in Train + Val set	44	
Calf number in Test set	24	

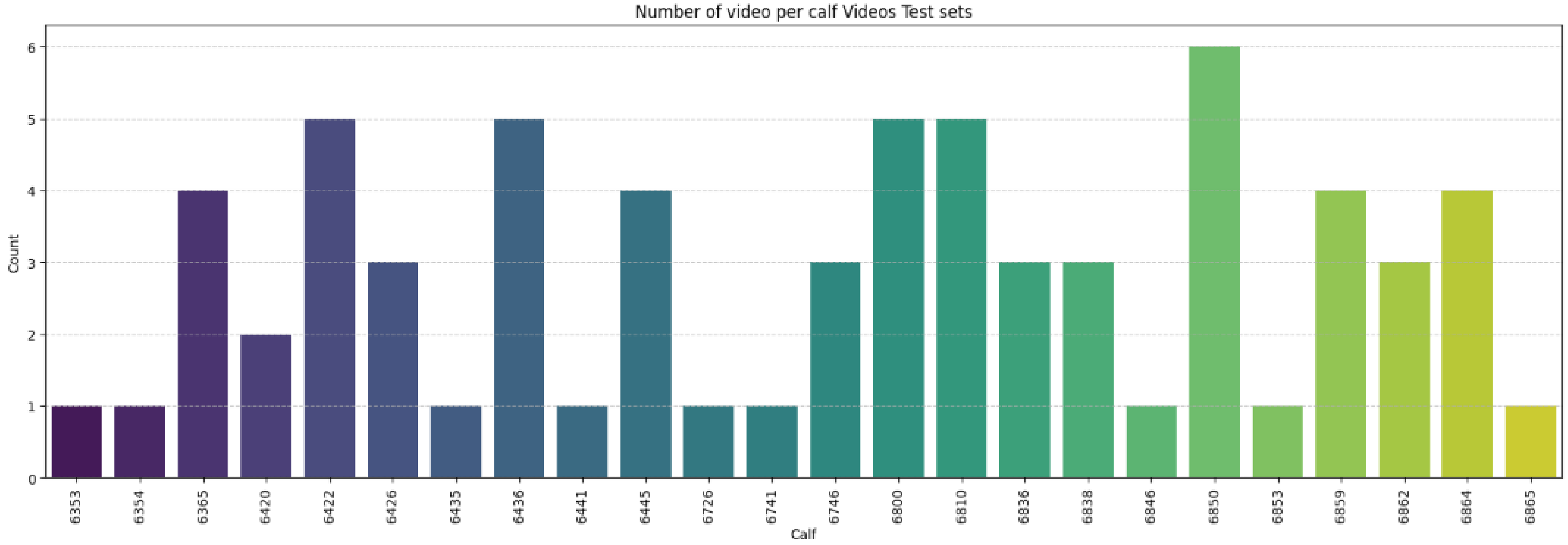
Training details

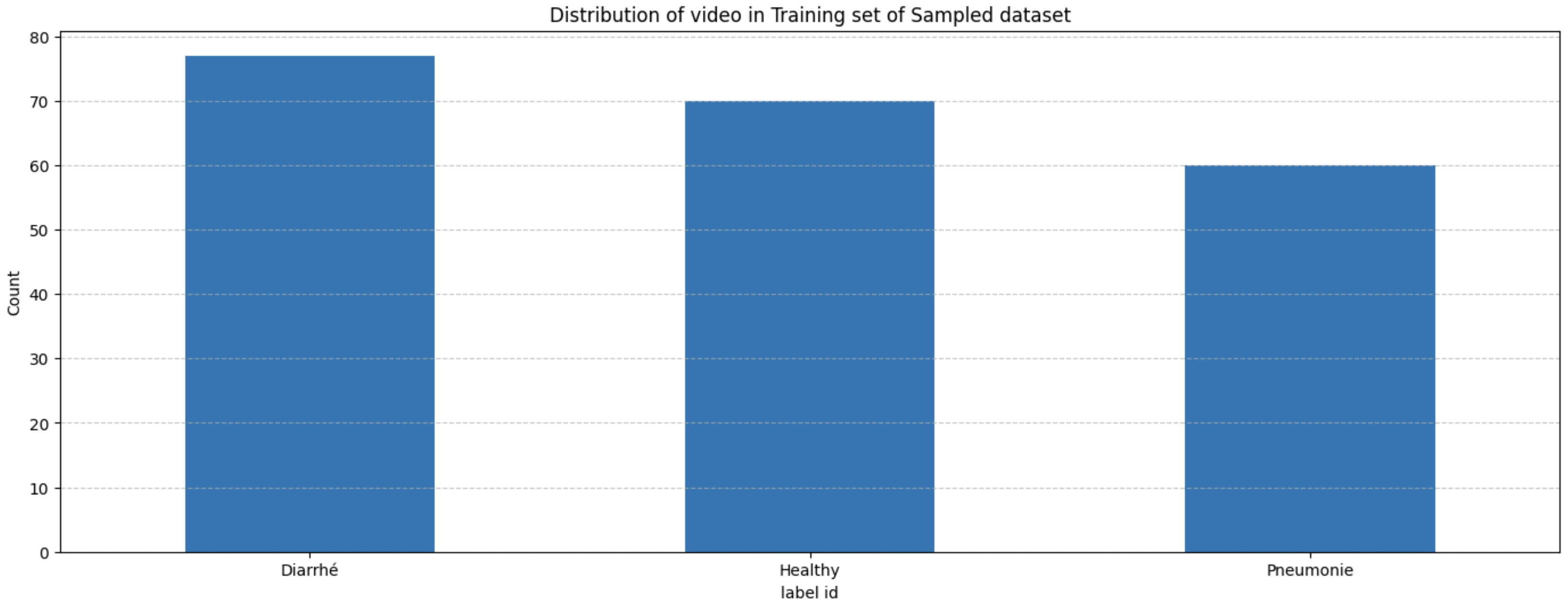


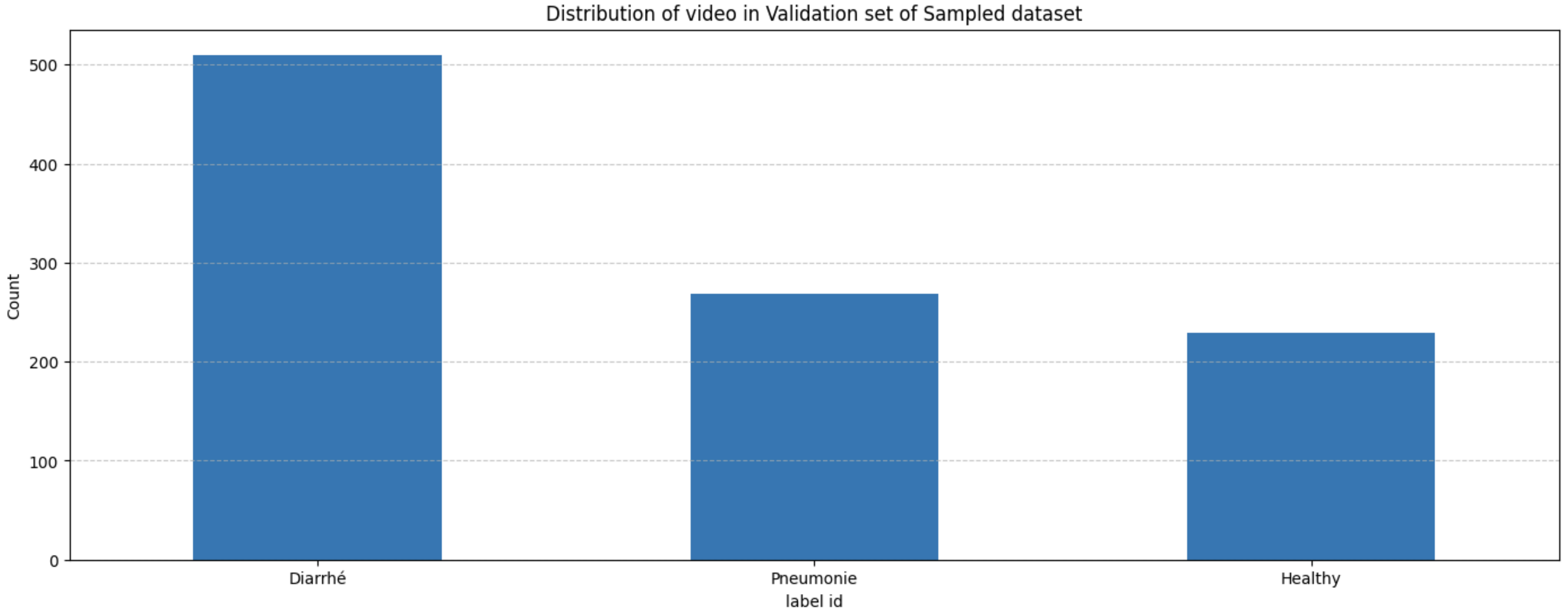
Training details




Number of image per calf Images Test sets
Count
0
5
10
15
20
25
30
Calf
6353
6354
6365
6420
6422
6426
6435
6436
6441
6445
6726
6741
6746
6800
6810
6836
6838
6846
6850
6853
6859
6862
6864
6865


Number of video per calf Videos Test sets
Count
0
1
2
3
4
5
6
Calf
6353
6354
6365
6420
6422
6426
6435
6436
6441
6445
6726
6741
6746
6800
6810
6836
6838
6846
6850
6853
6859
6862
6864
6865


Distribution of video in Training set of Sampled dataset
Count
0
10
20
30
40
50
60
70
80
Diarrhé
Healthy
Pneumonie
label id


Distribution of video in Validation set of Sampled dataset
Count
0
100
200
300
400
500
Diarrhé
Pneumonie
Healthy
label id

VideoMAE

- Solution aux redondances entres les frames des vidéos de notre dataset
- Bonnes performances sur de petits datasets

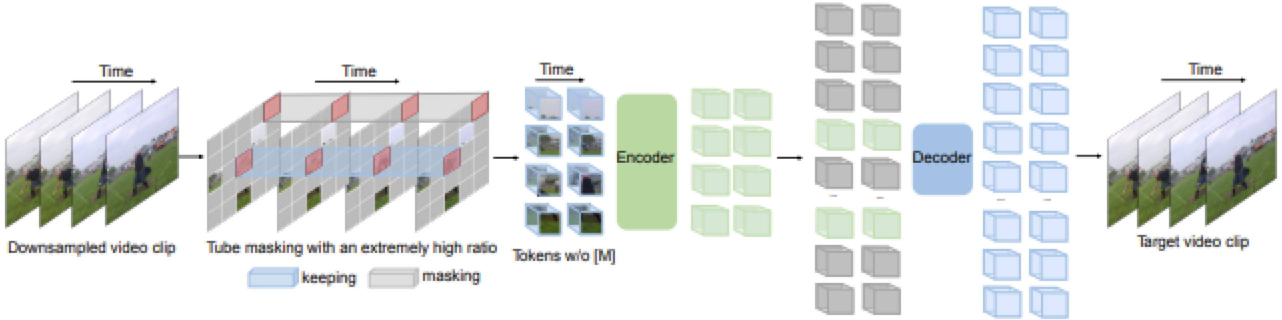


Figure 1: **VideoMAE** performs the task of masking random cubes and reconstructing the missing ones with an asymmetric encoder-decoder architecture. Due to high redundancy and temporal correlation in videos, we present the customized design of tube masking with an extremely high ratio (90% to 95%). This simple design enables us to create a more challenging and meaningful self-supervised task to make the learned representations capture more useful spatiotemporal structures.

Figure 2: Slowness is a general prior in (a) video data [88]. This leads to two important characteristics in time: temporal redundancy and temporal correlation. Temporal redundancy makes it possible to recover pixels under an extremely high masking ratio. Temporal correlation leads to easily reconstruct the missing pixels by finding those corresponding patches in adjacent frames under plain (b) frame masking or (c) random masking. To avoid this simple task and encourage learning representative representation, we propose a (d) tube masking, where the masking map is the same for all frames.

Modèle VideoMAE

Training details

- Used a pretrained model on SSV2
- 10s of videos
- 16 frames per videos seperate by 15 frames each
- Balance each batch
- Use a weighted loss
- Train over 10 epochs

-	**Two Class:** Illness(Negatif) and Healthy (Positif)		**Three class:** Healthy, Pneumonie and Diarrhea	
-	Sample	Whole	Sample	Whole
Accuracy en %	45.59	48.53	33.82	29.41
F1-score en %	55.42	47.76	33.82	29.41
Balanced Accuracy en %	58.89	53.67	33.53	29.18

Modèle VideoMAE

Whole dataset			
In %	F1	Précision	Rappel
Healthy	31.11	31.82	30.43
Pneumonie	33.33	27.03	43.48
Diarrhea	19.35	33.33	13.64

Sample dataset			
In %	F1	Précision	Rappel
Healthy	51.35	37.25	82.61
Pneumonie	7.69	33.33	4.35
Diarrhea	16.67	21.43	13.64

-	**Two Class:** Illness (Negatif) and Healthy (Positif)		**Three class:** Healthy, Pneumonie and Diarrhea	
-	Sample	Whole	Sample	Whole
Accuracy en %	45.59	**48.53**	33.82	29.41
F1-score en %	**55.42**	47.76	33.82	29.41
Balanced Accuracy en %	**58.89**	53.67	33.53	29.18

Figure 7. Visualization of space-time attention from the output token to the input space on Something-Something-V2. Our model learns to focus on the relevant parts in the video in order to perform spatiotemporal reasoning.

Timesformer

- Faster to train than 3D CNN
- higher test efficiency (at a small drop in accuracy)

Modèle Timesformer

Training details

- Used a pretrained model on Kinetics dataset
- 10s of videos
- 16 frames per videos seperate by 15 frames each
- Balance each batch
- Use a weighted loss
- Train over 10 epochs

-	**Two Class:** Illness(Negatif) and Healthy (Positif)		**Three class:** Healthy, Pneumonie and Diarrhea	
-	Sample	Whole	Sample	Whole
Accuracy en %	44.12	51.47	39.71	32.35
F1-score en %	36.67	40.00	39.71	32.35
Balanced Accuracy en %	45.02	50.58	39.39	32.08

Modèle Timesformer

Whole dataset			
In %	F1	Précision	Rappel
Healthy	34.62	31.03	39.13
Pneumonie	36.36	31.25	43.48
Diarrhea	20.69	42.86	13.64

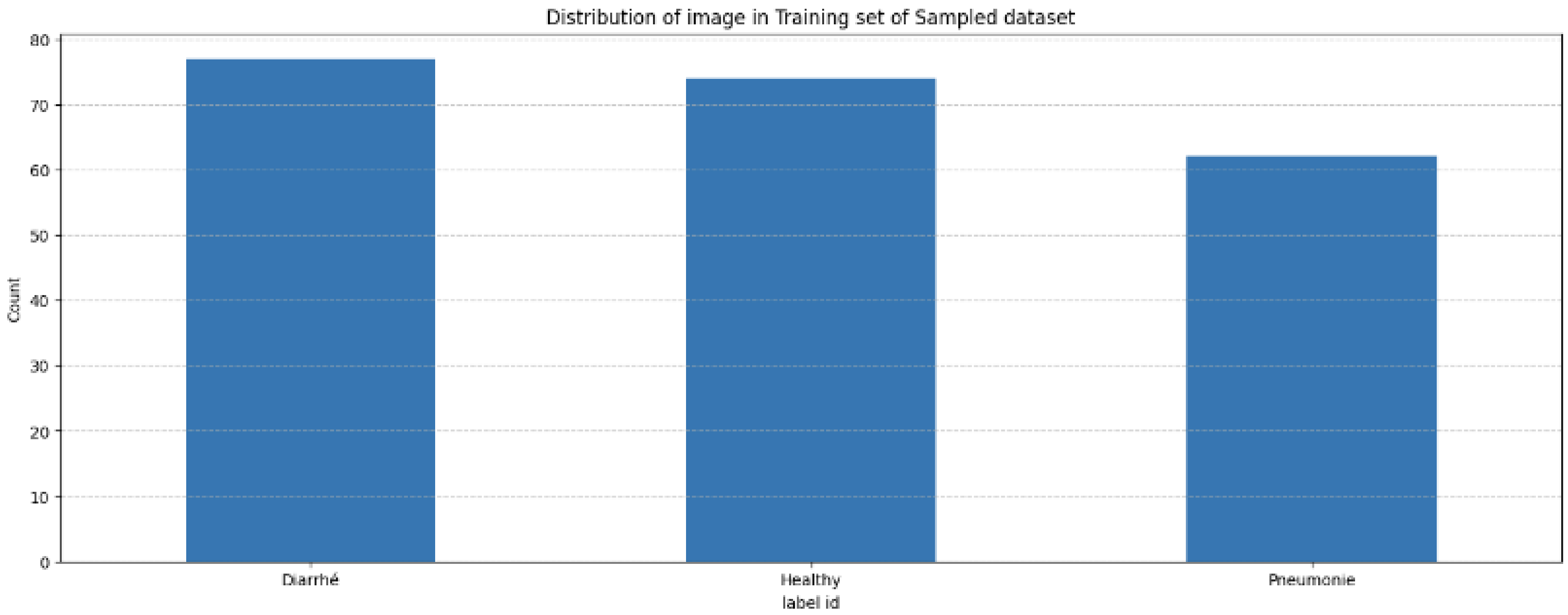
Sample dataset			
In %	F1	Précision	Rappel
Healthy	56.72	43.18	82.61
Pneumonie	24.24	40.00	17.39
Diarrhea	22.22	28.57	18.18

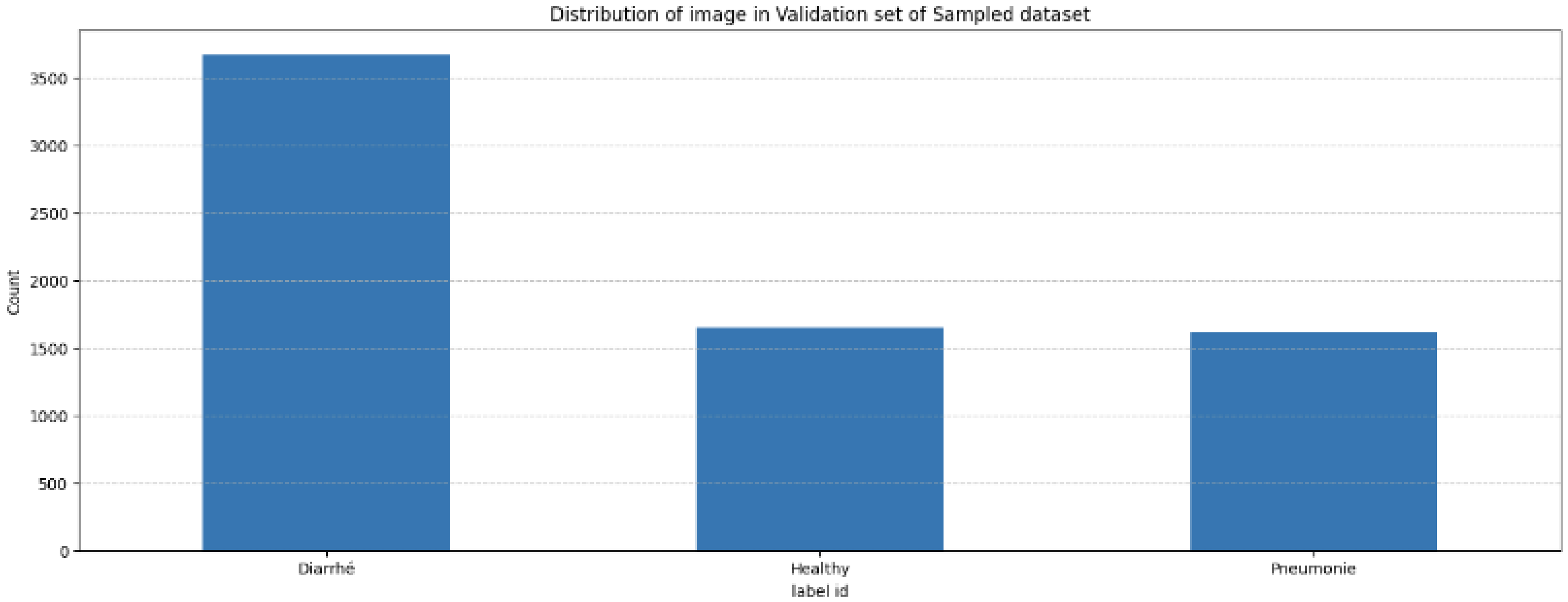
-	**Two Class:** Illness(Negatif) and Healthy (Positif)		**Three class:** Healthy, Pneumonie and Diarrhea	
-	Sample	Whole	Sample	Whole
Accuracy en %	44.12	**51.47**	39.71	32.35
F1-score en %	36.67	**40.00**	39.71	32.35
Balanced Accuracy en %	45.02	**50.58**	39.39	32.08

Modèles d'Image

Training details

- Balance each batch
- Use a weighted loss
- Train over 10 epochs with early stop
- Use the best model base on lower loss on training





ViT

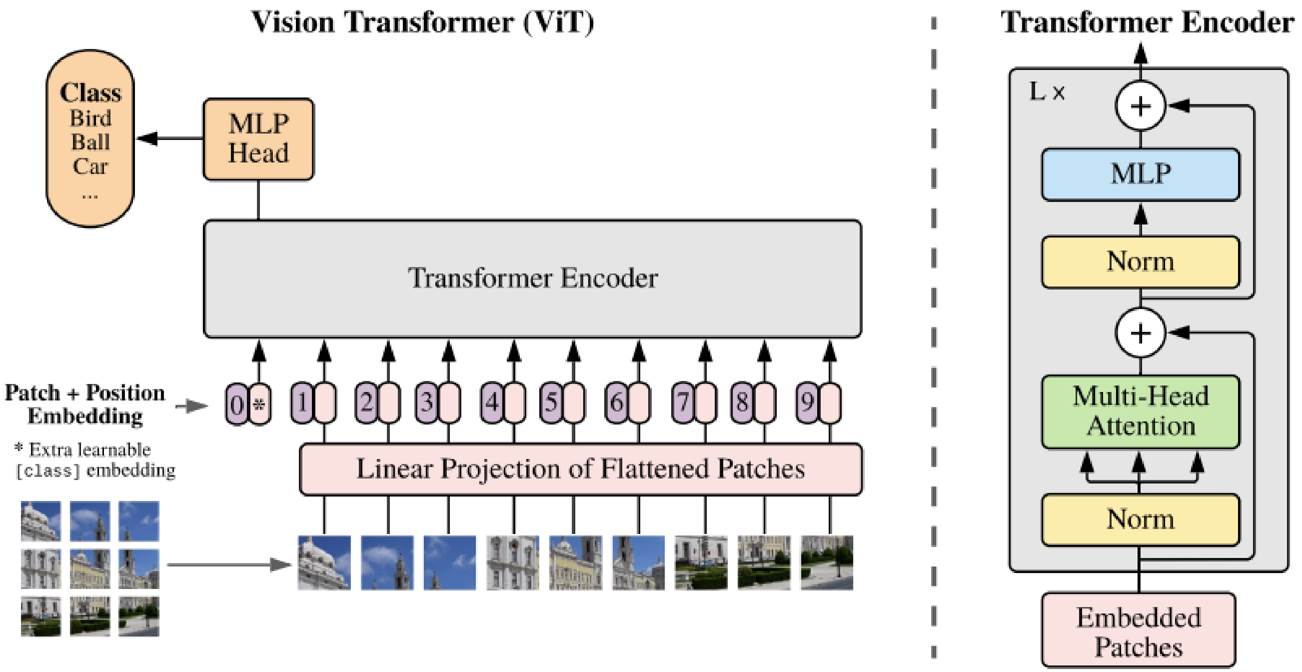


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

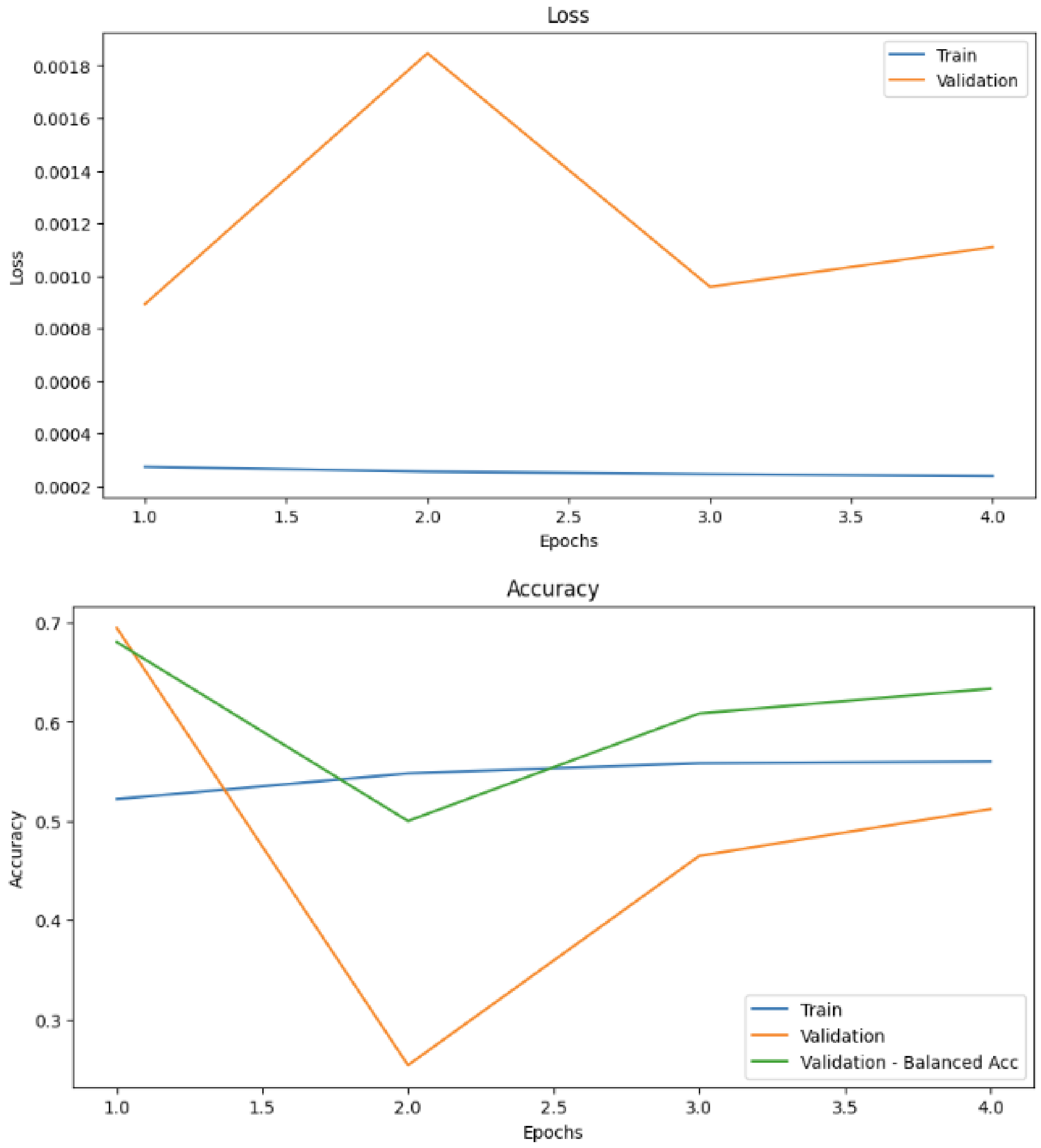
Modèle VIT

Whole dataset			
In %	Precision	Recall	F1
Healthy	49.59	55.05	52.17
Pneumonie	30.91	45.33	36.76
Diarrhea	53.85	28.28	37.09

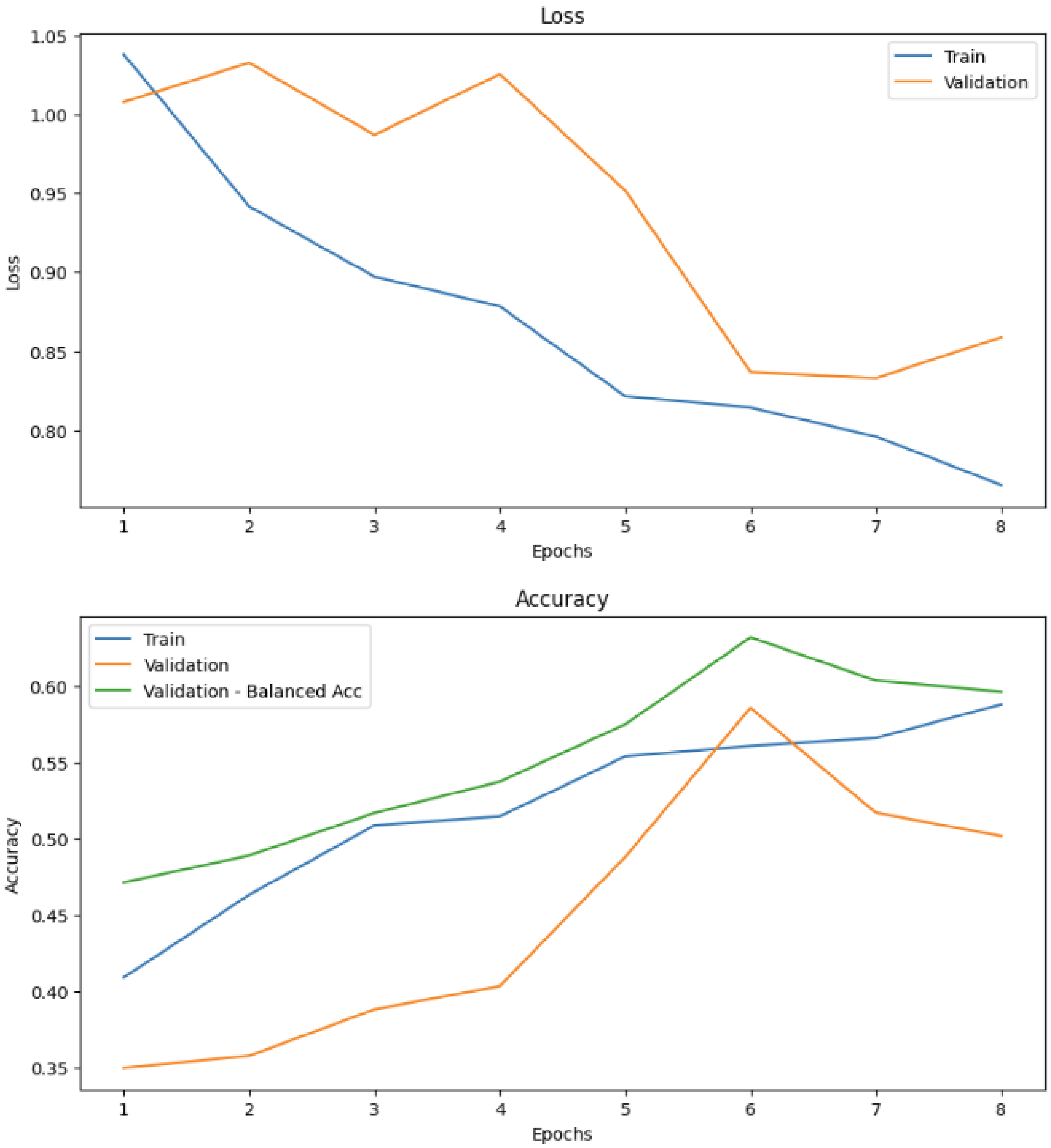
Sample dataset			
In %	Precision	Recall	F1
Healthy	49.06	71.56	58.21
Pneumonie	26.61	44.00	33.17
Diarrhea	0	0	0

-	**Two Class:** Illness(Negatif) and Healthy (Positif)		**Three class:** Healthy, Pneumonie and Diarrhea	
-	Sample	Whole	Sample	Whole
Accuracy en %	46.29	**63.60**	39.22	43.11
F1-score en %	54.76	**58.63**	31.21	42.81
Balanced Accuracy en %	53.41	**64.23**	38.52	42.89

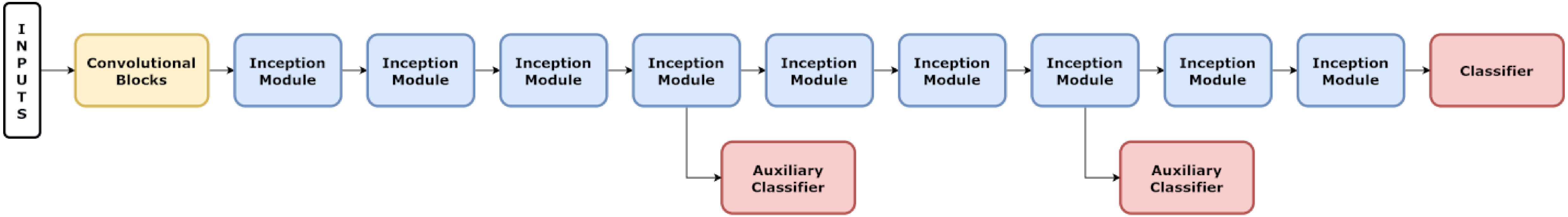
Modèle VIT on Training: Best



Modèle VIT on Training: Whole dataset and multi-class problem



InceptionV3



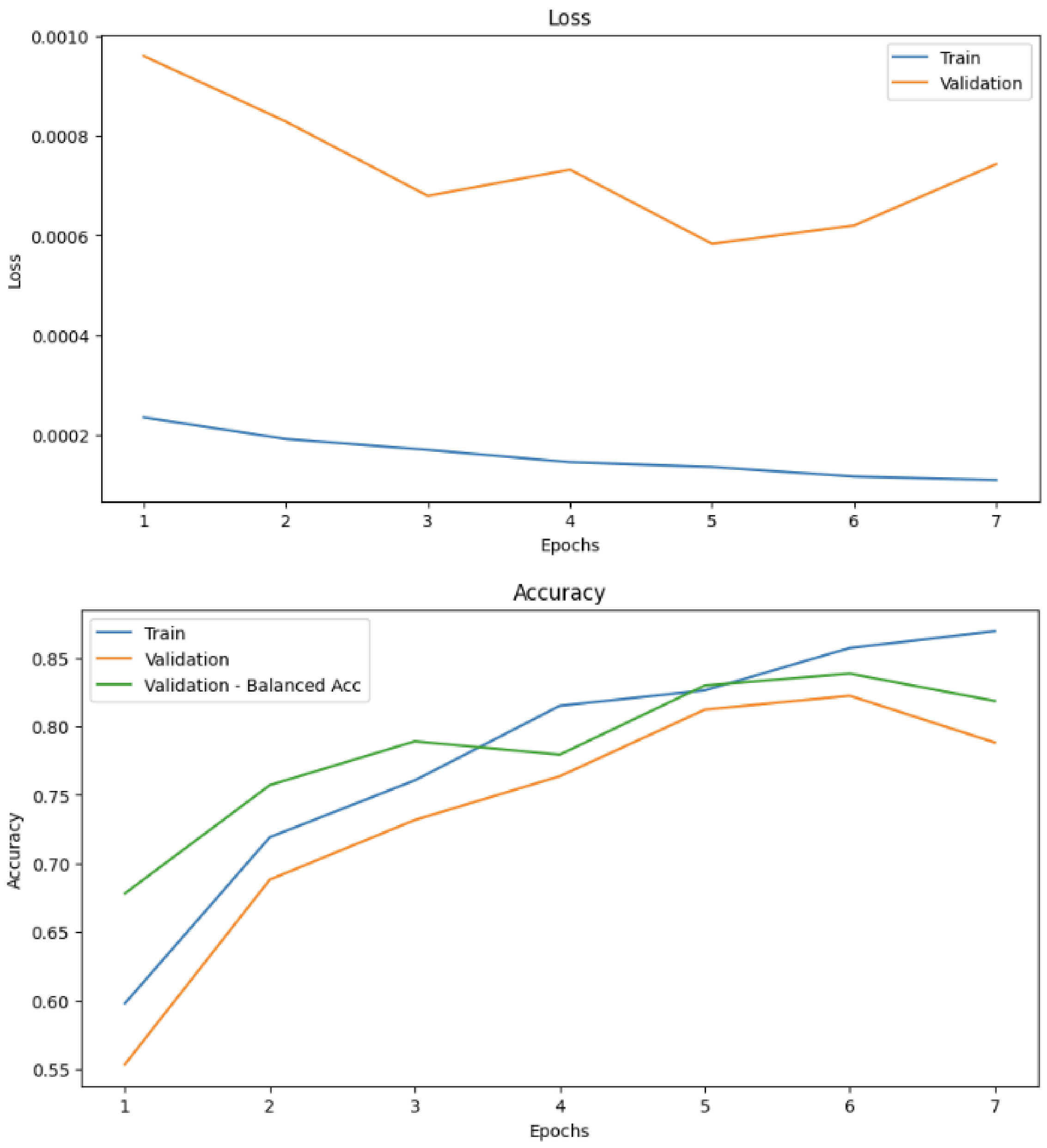
Modèle InceptionV3

Whole dataset			
In %	Precision	Recall	F1
Healthy	44.72	66.06	53.33
Pneumonie	18.33	14.67	16.30
Diarrhea	53.23	33.33	40.99

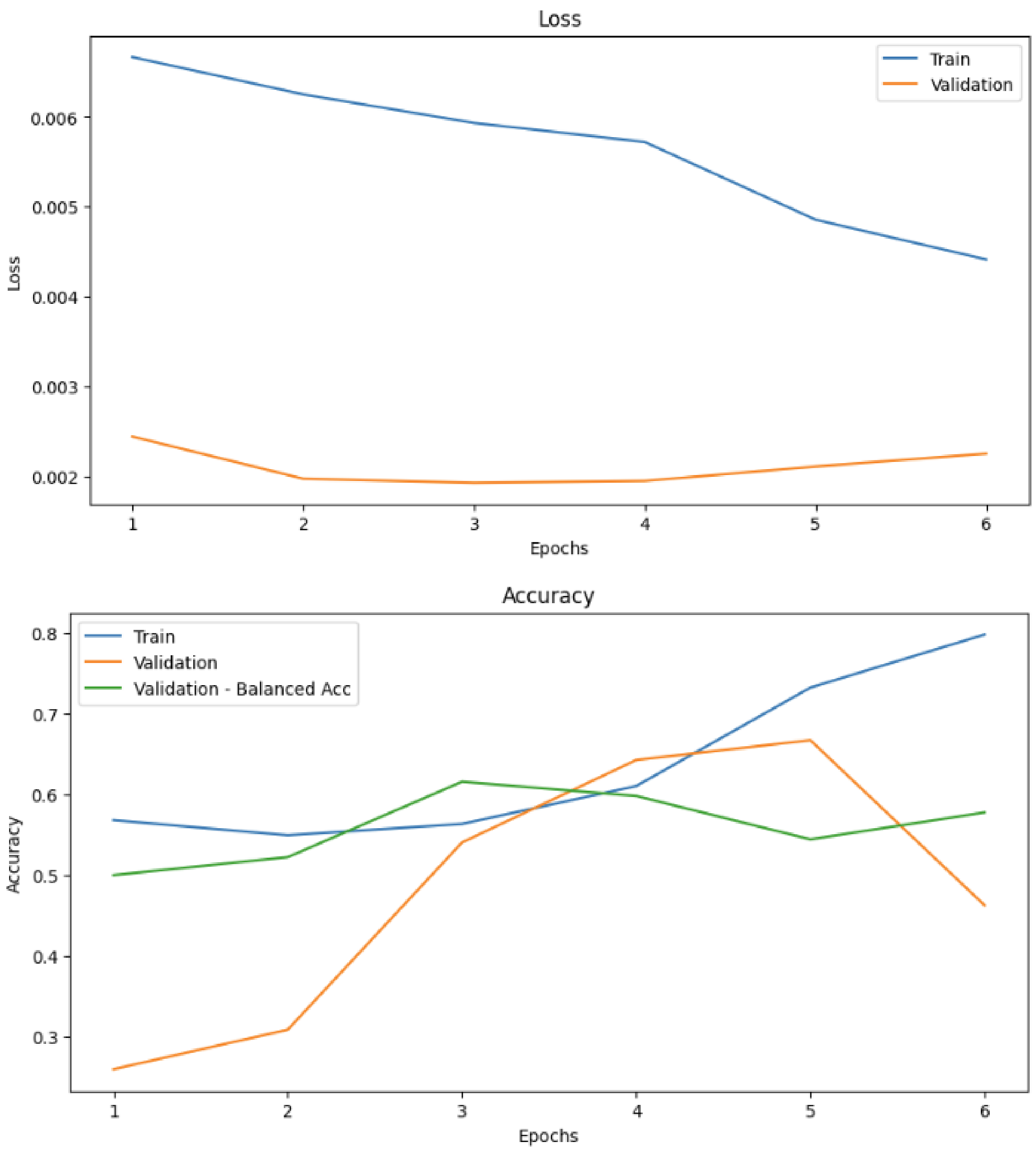
Sample dataset			
In %	F1	Précision	Rappel
Healthy	44.27	53.21	48.33
Pneumonie	22.94	33.33	27.17
Diarrhea	19.72	32.56	14.14

-	**Two Class:** Illness(Negatif) and Healthy (Positif)		**Three class:** Healthy, Pneumonie and Diarrhea	
-	Sample	Whole	Sample	Whole
Accuracy en %	41.70	**53.36**	34.28	40.99
F1-score en %	**56.69**	51.47	32.72	39.20
Balanced Accuracy en %	52.41	**55.39**	33.56	38.02

Modèle InceptionV3 on Training: Best



Modèle InceptionV3 on Training: Best 2



Efficientnet-b3

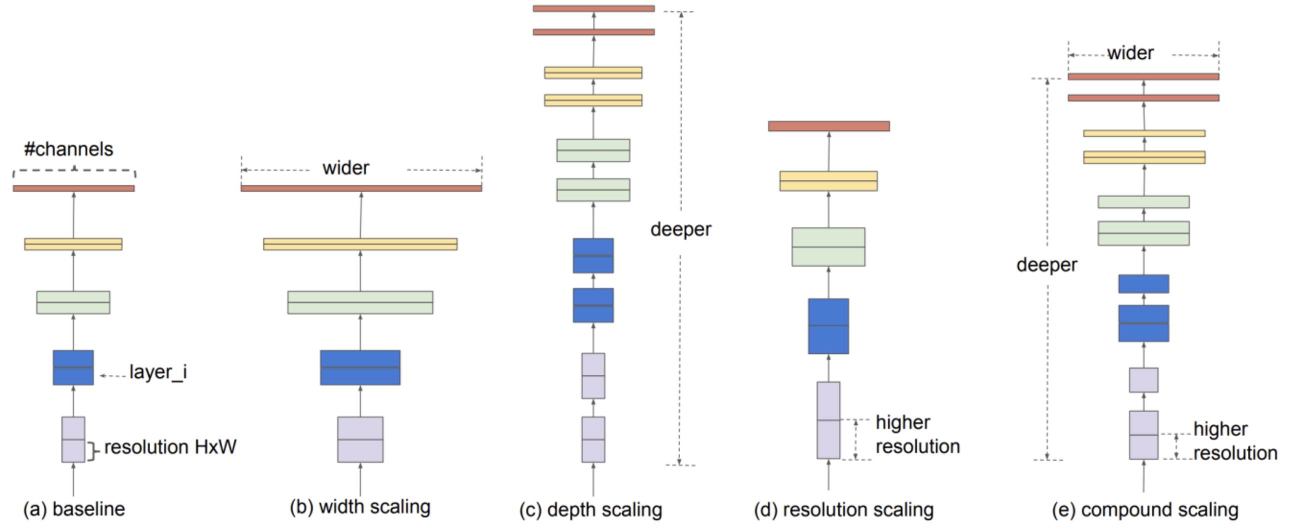
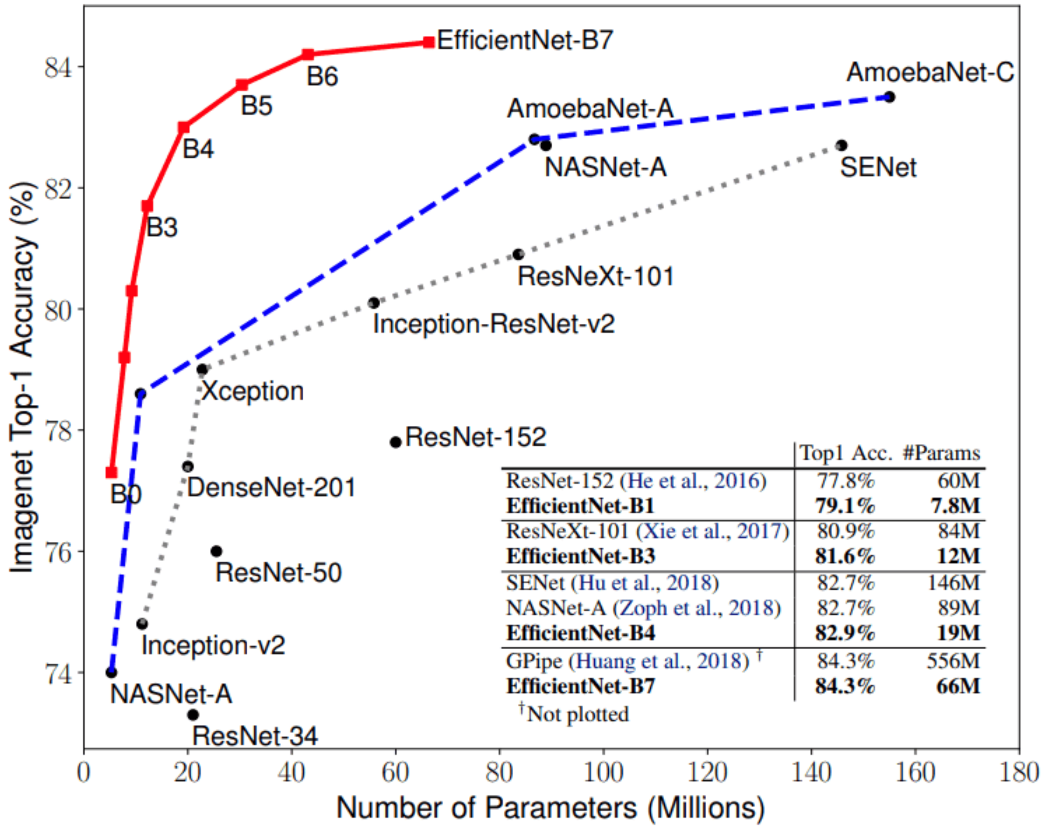


Figure 2. **Model Scaling.** (a) is a baseline network example; (b)-(d) are conventional scaling that only increases one dimension of network width, depth, or resolution. (e) is our proposed compound scaling method that uniformly scales all three dimensions with a fixed ratio.



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted

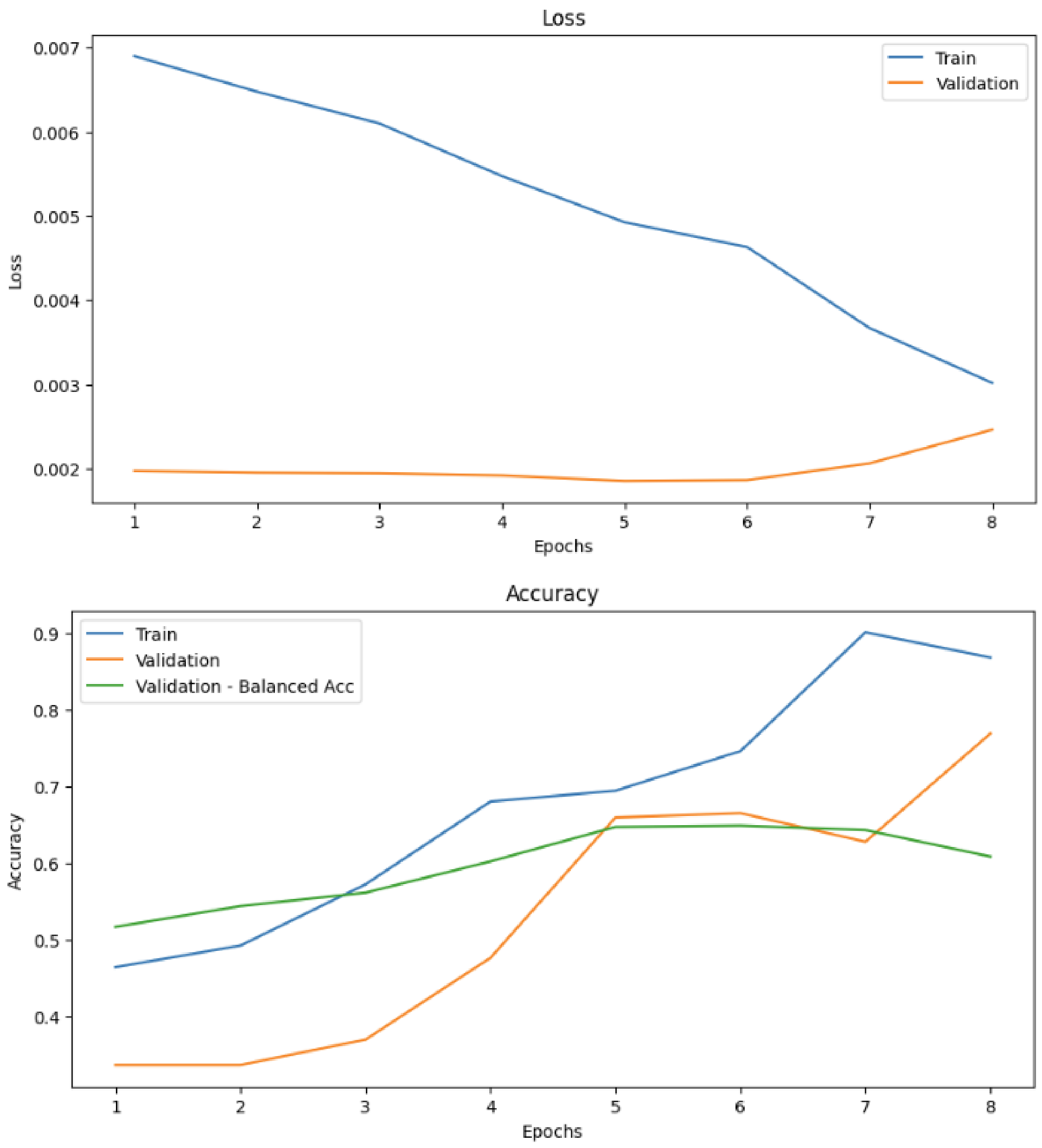
Modèle Efficientnet-b3

Whole dataset			
In %	Precision	Recall	F1
Healthy	44.70	54.13	48.96
Pneumonie	29.41	26.67	27.97
Diarrhea	43.37	36.36	39.56

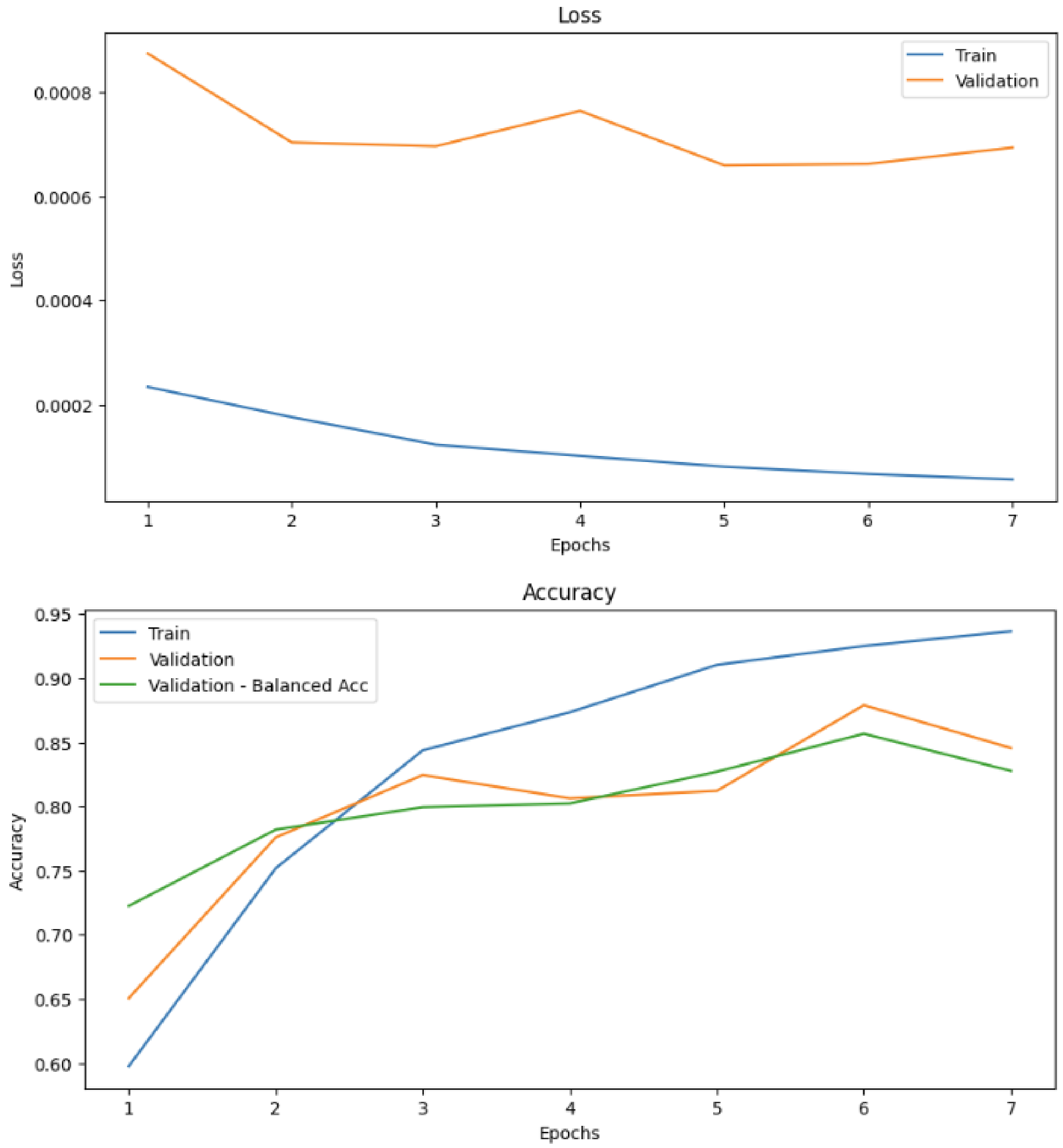
Sample dataset			
In %	Precision	Recall	F1
Healthy	45.52	60.55	51.97
Pneumonie	32.43	16.00	21.43
Diarrhea	41.58	42.42	42.00

-	**Two Class:** Illness(Negatif) and Healthy (Positif)		**Three class:** Healthy, Pneumonie and Diarrhea	
-	Sample	Whole	Sample	Whole
Accuracy en %	**53.71**	49.82	42.40	40.64
F1-score en %	**58.93**	56.17	40.39	40.11
Balanced Accuracy en %	**59.79**	56.11	39.66	39.05

Modèle Efficientnet-b3 on Training: Best



Modèle Efficientnet-b3 on Training: Best 2



INterpretable TRansformer

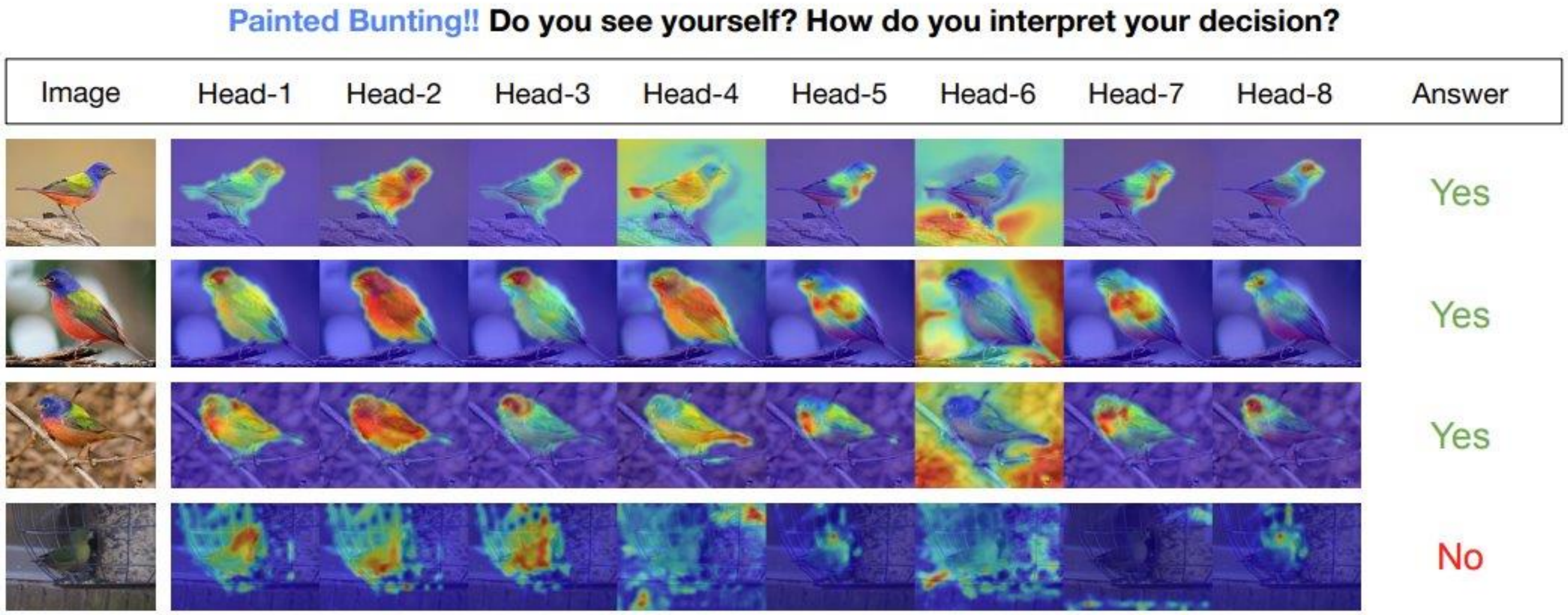


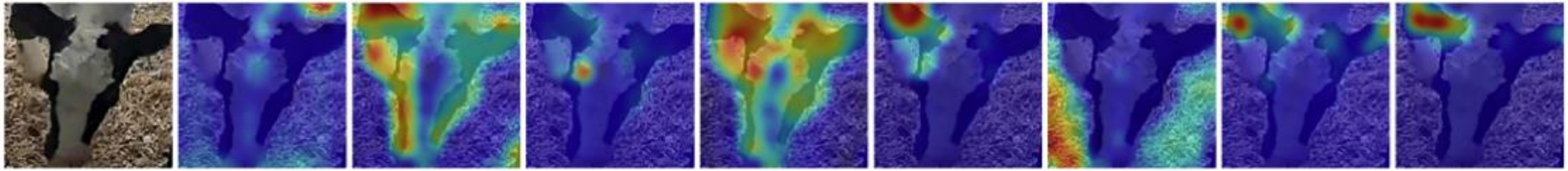
Figure 1: **Illustration of INTR**. We show four images (row-wise) of the same bird species Painted Bunting and the eight-head cross-attention maps (column-wise) triggered by the query of the ground-truth class. Each head is learned to attend to a different (across columns) but consistent (across rows) semantic cue in the image that is useful to recognize this bird species (e.g., attributes). The exception is the last row, which shows inconsistent attention. Indeed, this is a misclassified case, showcasing how INTR interprets (wrong) predictions.

Modèle INTR

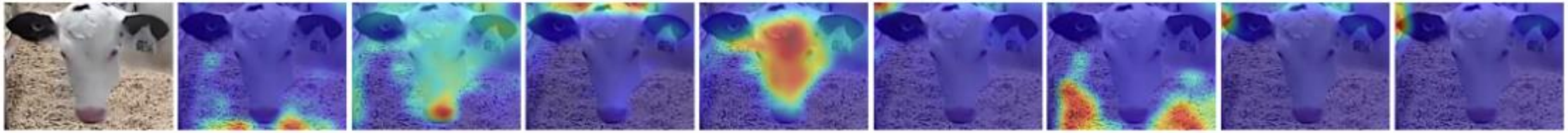
-	**Two Class:** Illness(Negatif) and Healthy (Positif)		**Three class:** Healthy, Pneumonie and Diarrhea	
-	Sample	Whole	Sample	Whole
Accuracy en %	44.16	**55.12**	40.98	42.40
F1-score en %	41.48	**60.92**	38.31	42.64
Balanced Accuracy en %	45.51	**61.79**	40.68	42.98

Modèle INTR

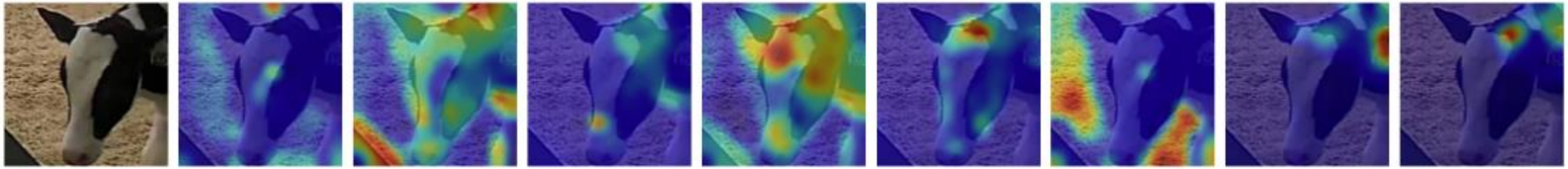
```
Species predicted by INTR is: Healthy
Species class  is: Healthy
```



```
Species predicted by INTR is: Healthy
Species class  is: Healthy
```

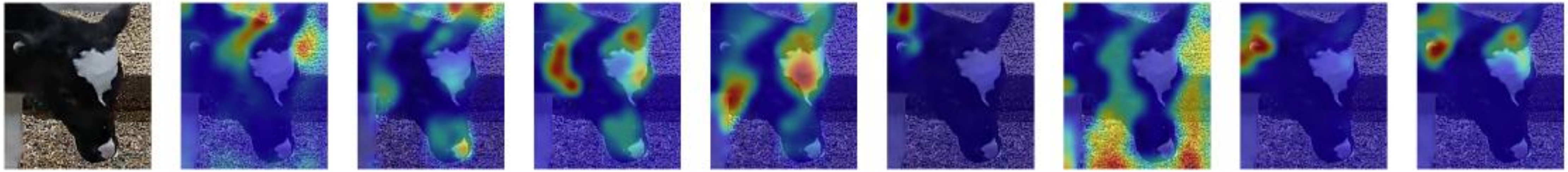


```
Species predicted by INTR is: Healthy
Species class  is: Illness
```

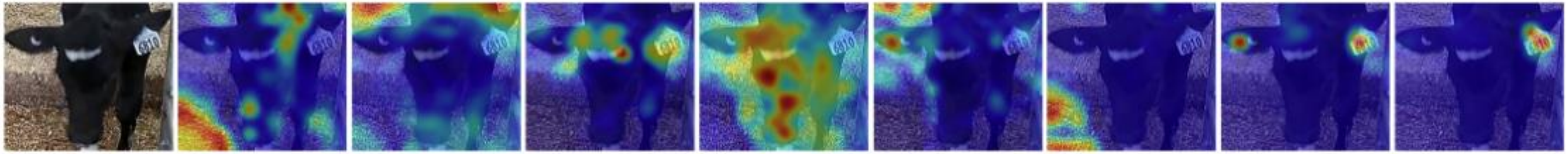


Modèle INTR

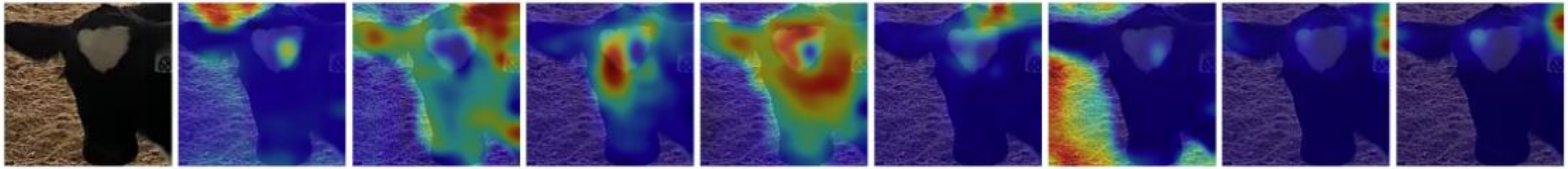
Species predicted by INTR is: Illness
Species class is: Illness



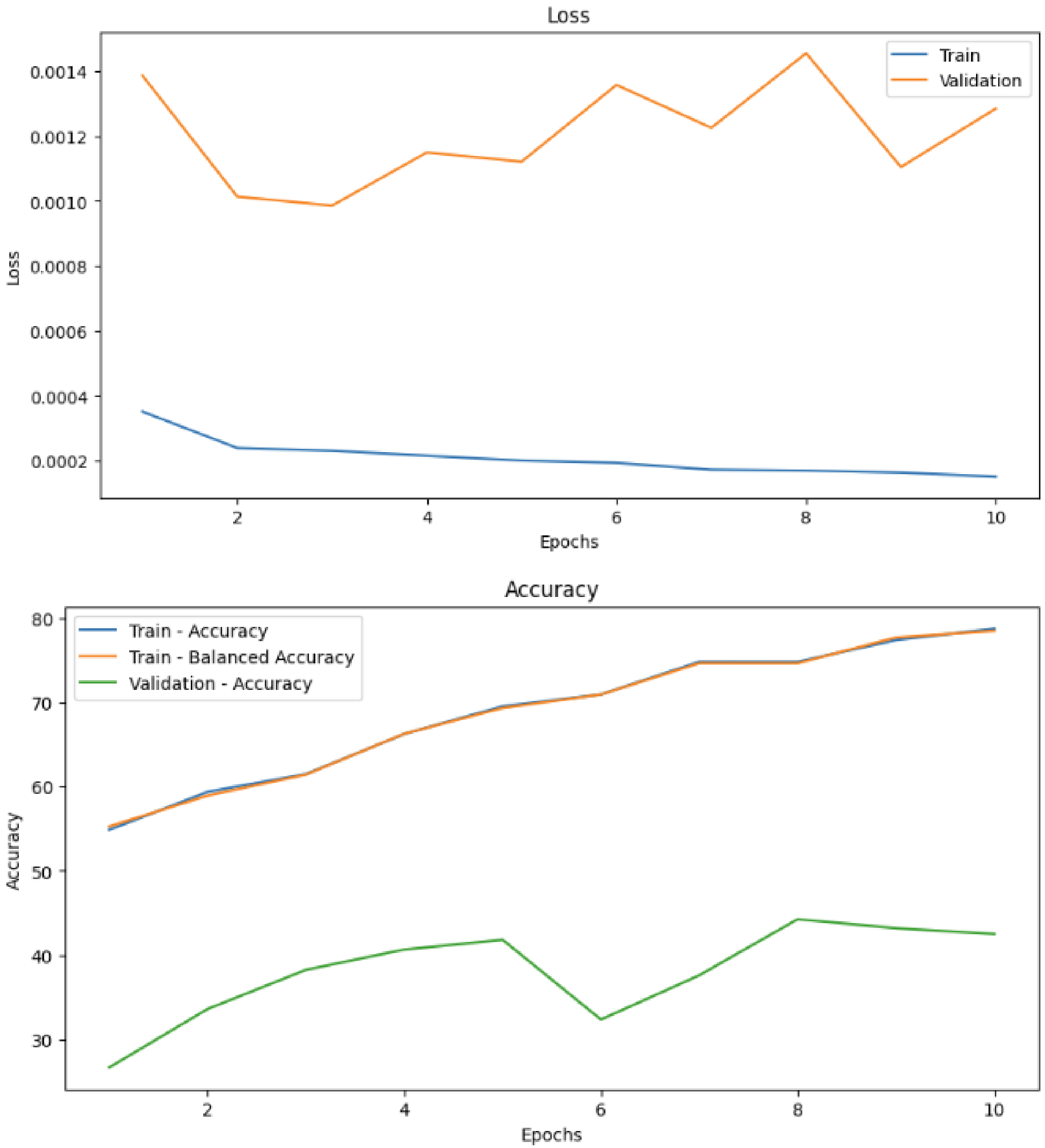
Species predicted by INTR is: Illness
Species class is: Illness



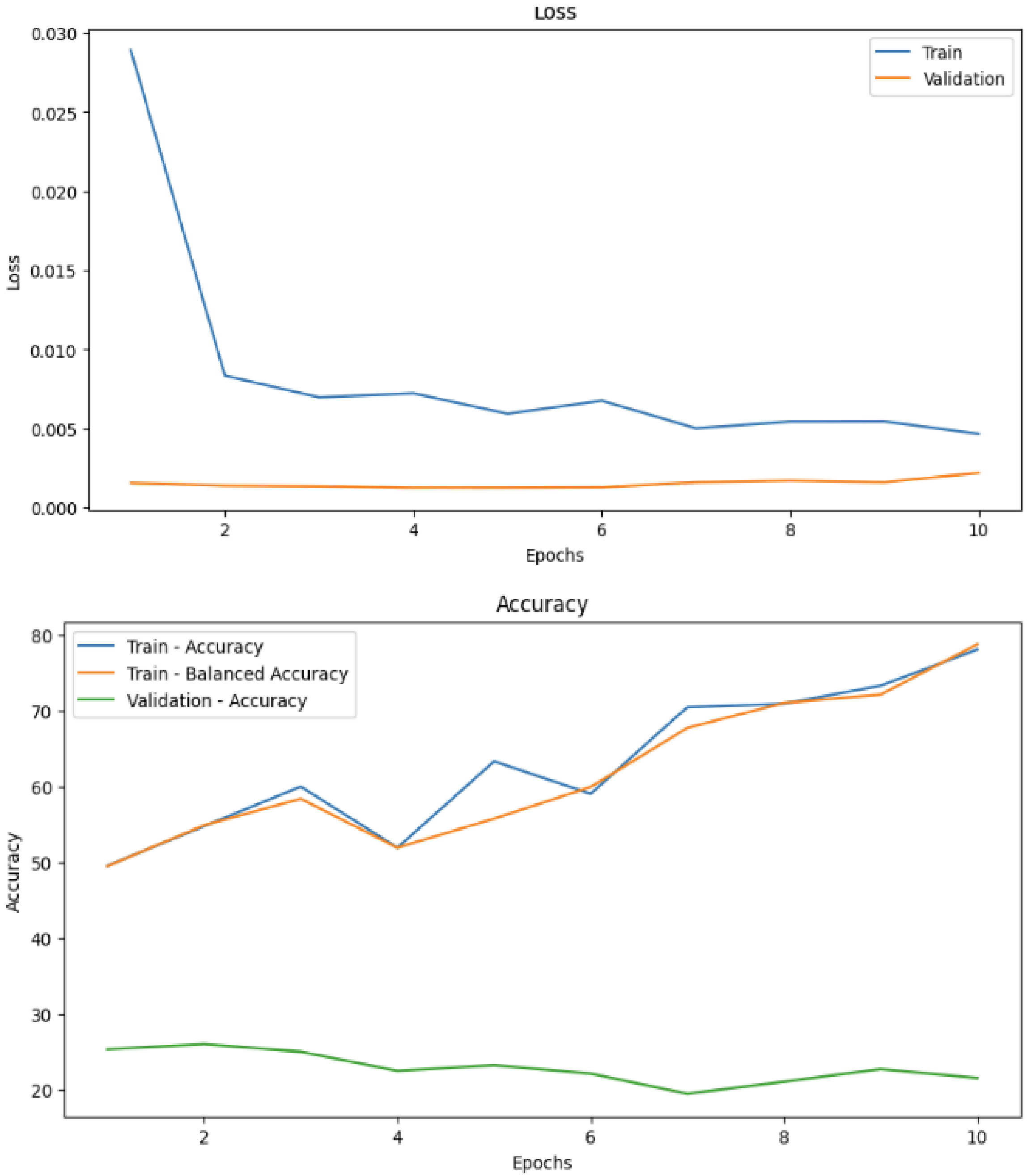
Species predicted by INTR is: Illness
Species class is: Healthy



Modèle INTR on Training: Best



Modèle INTR on Training: Best 2



Recap

	Modèle INTR – Whole set - Two Class	Modèle Efficientnet-b3 – Sampled set - Two Class	Modèle VIT – Whole set - Two Class	Modèle Timesformer – Whole set - Two Class – Video Model	**Previous performances (for multi-class and without LOCO)**
Accuracy en %	55.12	53.71	63.60	51.47	66.66
F1-score en %	60.92	58.93	58.63	40.00	66.46
Balanced Accuracy en %	61.79	59.79	64.23	50.58	-

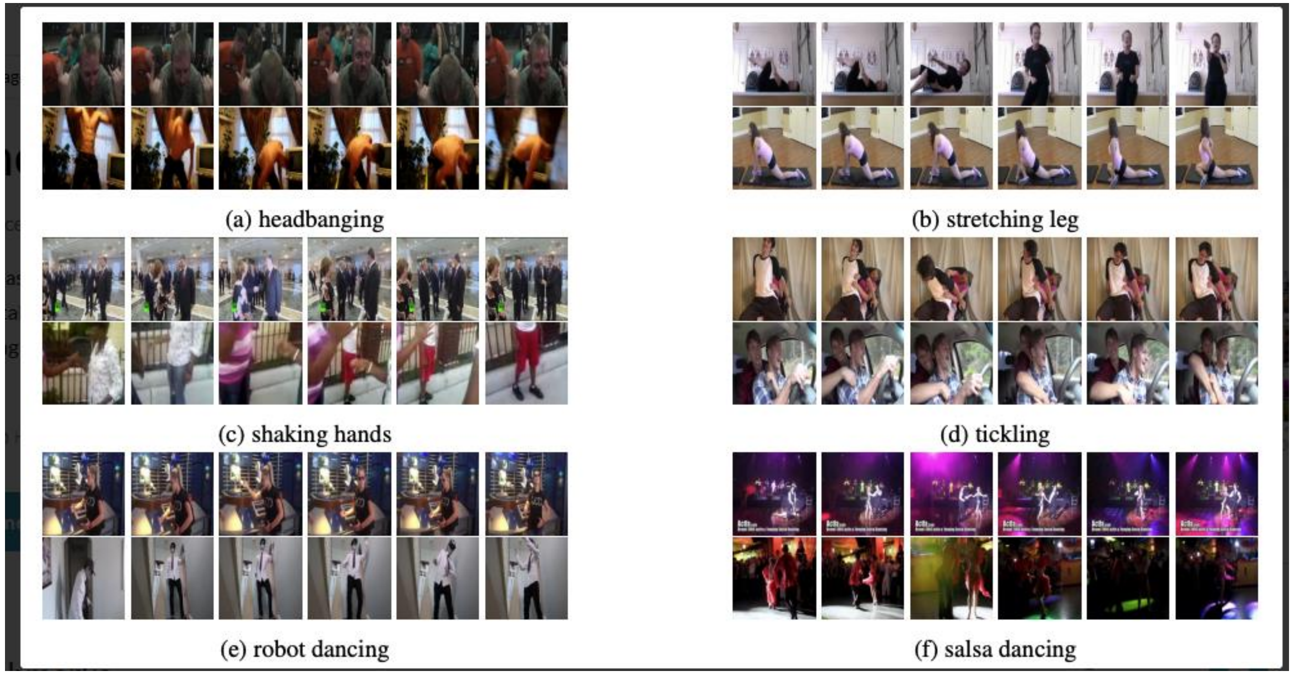
Autres

Leaving videos

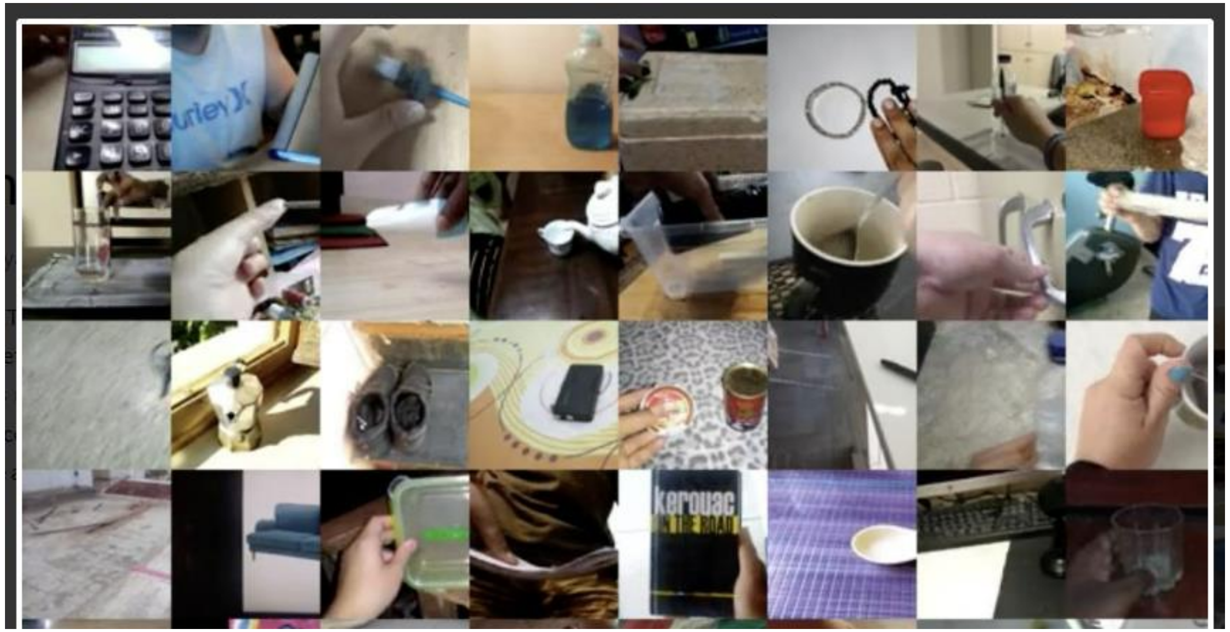
- Duration doesn't match the real duration of the calf at feeder

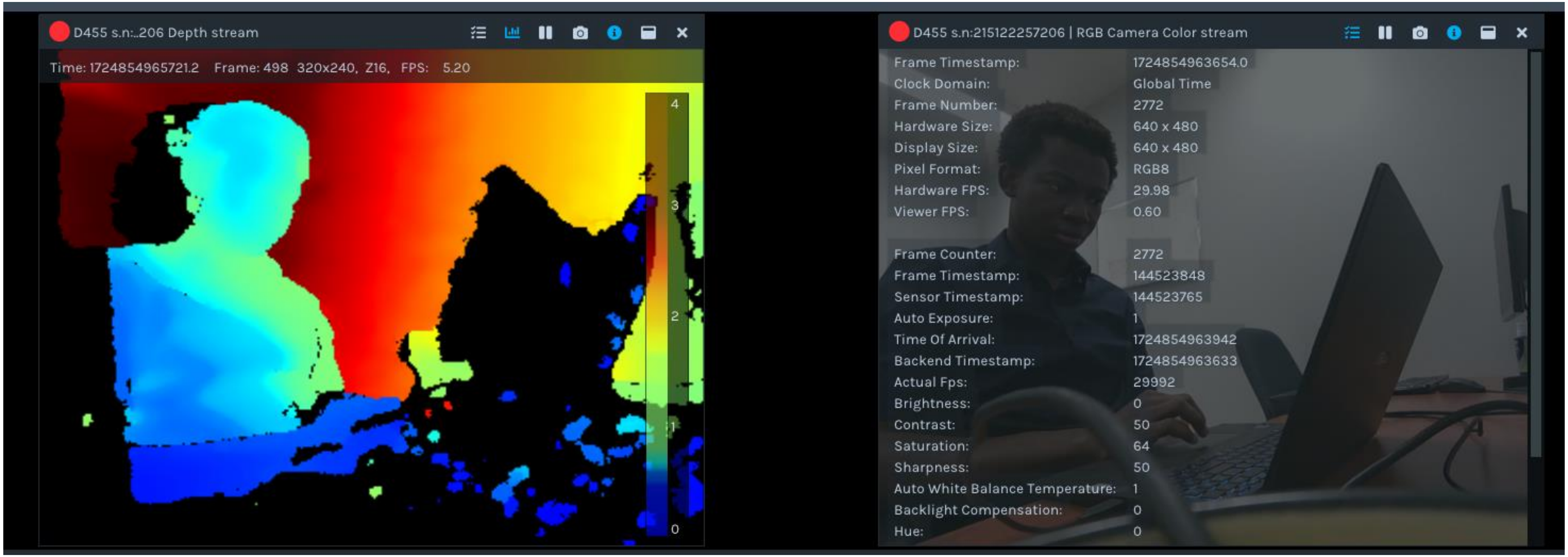
Modèle VideoMAE

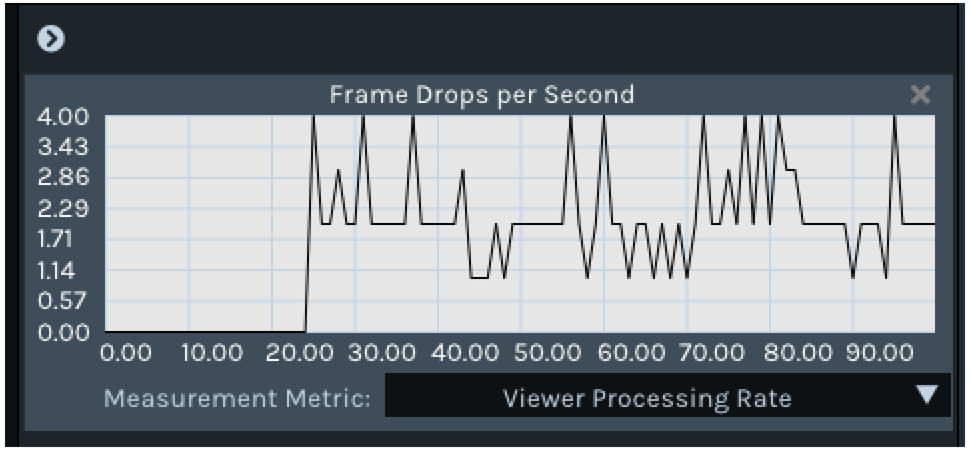
Kinetics dataset

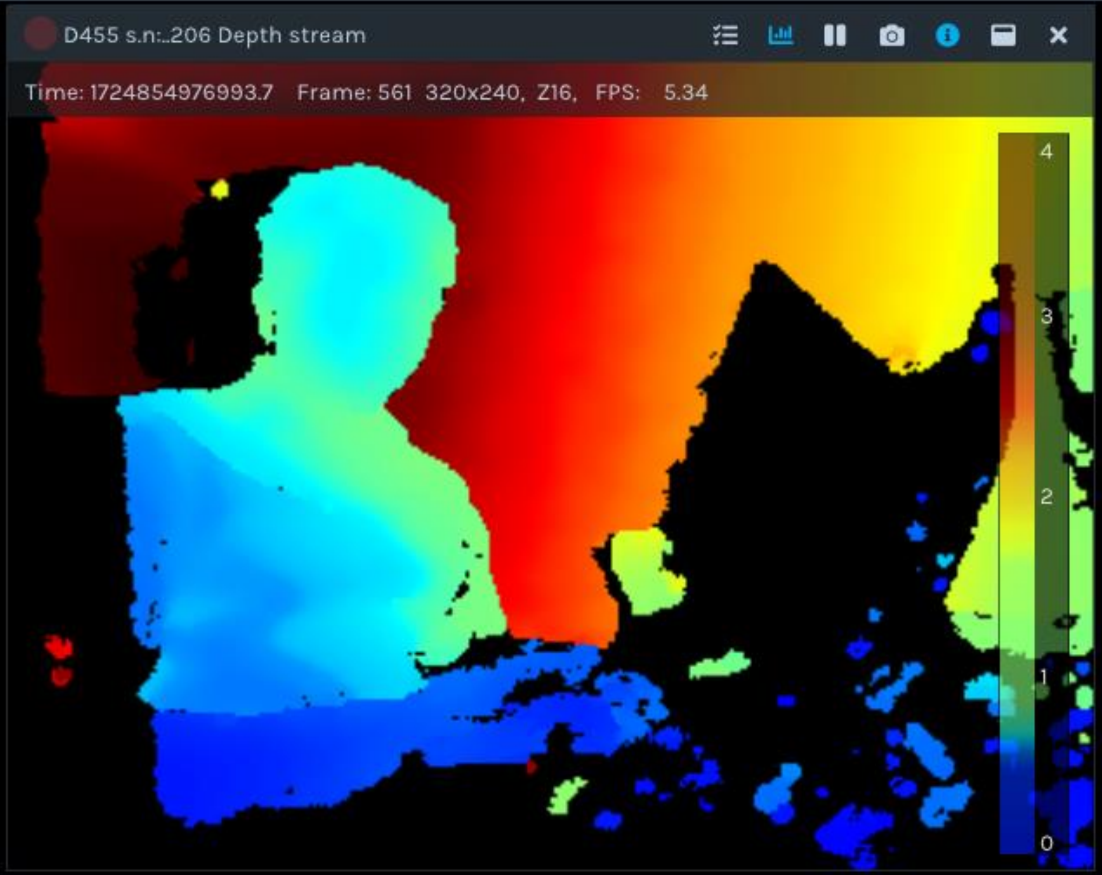


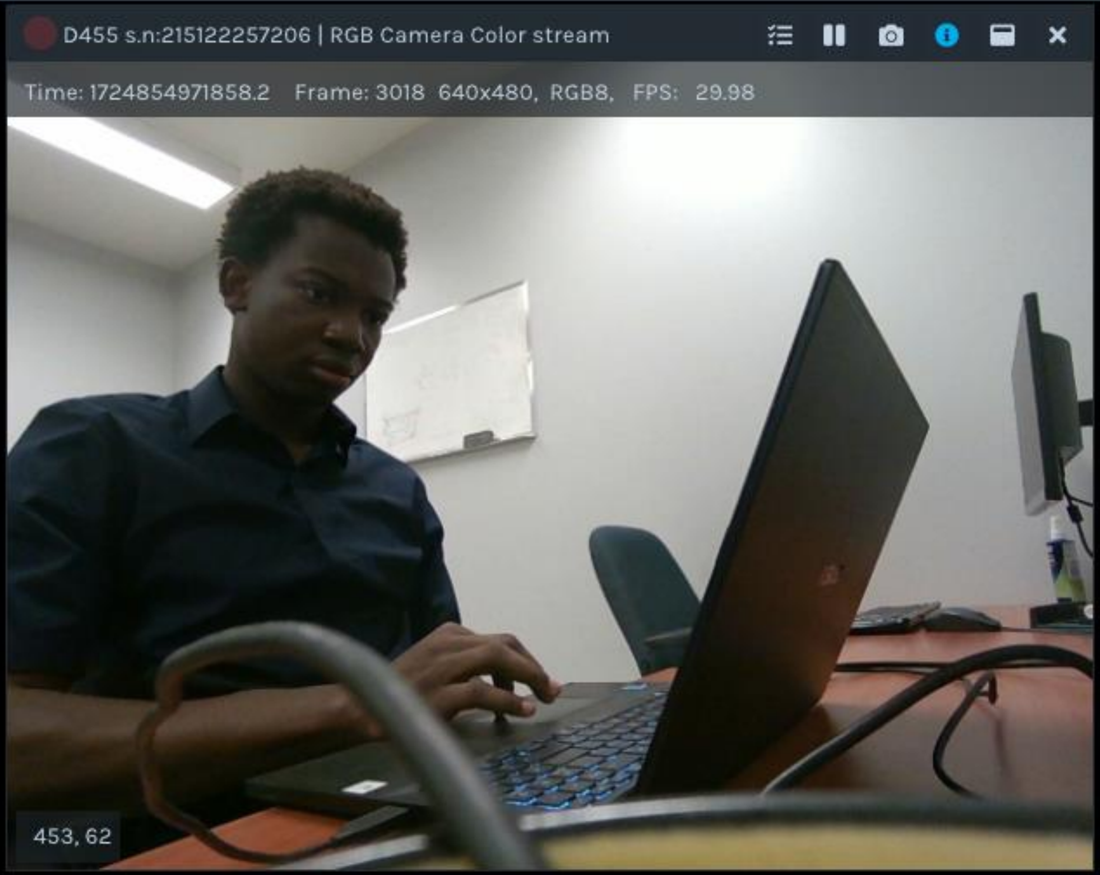
SSv2 dataset




D455 s.n:..206 Depth stream
Time: 1724854965721.2 Frame: 498 320x240, Z16, FPS: 5.20
4
3
2
1
0
D455 s.n:215122257206 | RGB Camera Color stream
Frame Timestamp: 1724854963654.0
Clock Domain: Global Time
Frame Number: 2772
Hardware Size: 640 x 480
Display Size: 640 x 480
Pixel Format: RGB8
Hardware FPS: 29.98
Viewer FPS: 0.60
Frame Counter: 2772
Frame Timestamp: 144523848
Sensor Timestamp: 144523765
Auto Exposure: 1
Time Of Arrival: 1724854963942
Backend Timestamp: 1724854963633
Actual Fps: 29992
Brightness: 0
Contrast: 50
Saturation: 64
Sharpness: 50
Auto White Balance Temperature: 1
Backlight Compensation: 0
Hue: 0


Frame Drops per Second
4.00
3.43
2.86
2.29
1.71
1.14
0.57
0.00
0.00 10.00 20.00 30.00 40.00 50.00 60.00 70.00 80.00 90.00
Measurement Metric:
Viewer Processing Rate


D455 s.n:..206 Depth stream
Time: 1724854976993.7 Frame: 561 320x240, Z16, FPS: 5.34
4
3
2
1
0


D455 s.n:215122257206 | RGB Camera Color stream
Time: 1724854971858.2 Frame: 3018 640x480, RGB8, FPS: 29.98
453, 62

Recap

	Modèle INTR – Whole set - Two Class	Modèle Efficientnet-b3 – Sampled set - Two Class	Modèle VIT – Whole set - Two Class	Modèle Timesformer – Whole set - Two Class – Video Model	**Previous perfor mances (for multi-class and without LOC O)**
Accuracy en %	39.39				
F1-score en %	50.0				
Balanced Accur acy en %	43.60				