Avancement des IA Audio et Vision

Stéphane Chavin, Pierre Mahé, Hervé Glotin

PSIBIOM :

Avancement des IA Audio et Vision

Stéphane Chavin, Pierre Mahe°, Hervé Glotin
Université de Toulon, Aix Marseille Univ, CNRS, LIS, Toulon, France
Centre Int. d'IA en Acoustique Naturelle, Toulon, France,
https://cian.univ-tln.fr
(° former member)

RR DYNI LIS 20231208

Sommaire

Ce travail a reçu le soutien financier de France 2030, PIA3 PSI-BIOM, opéré par l'ADEME, subvention 2182D0406-A. Il s'inscrit également dans le cadre du Programme d'Investissements d'Avenir de l'Agence Nationale de la Recherche intégré à France 2030, projet Terra Forma, ANR-21-ESRE-0014. Il est également partiellement subventionné par l'ANR SYLVANIA ANR-21-CE04-0019.

1. Rappel des tâches

Le LIS a pour tâche la construction des modèles off line de référence pour visuel et audio, et de l'audio online sur le PIC de la carte QHB. On présente ici les courbes, scores des modèles appris sur les données les plus récentes au début décembre 2023.

2. Vision

2.1 Base de données

Nous avons utilisé la nouvelle version du dataset d'OCAPI sur nextcloud (https://psi-biom.terroiko.fr/index.php/apps/files/?dir=/OCAPI_images_database/V2_23_10_27&fileid=83180), qui est plus complet et comprend un dataset de test indépendant.

En perspective les images de la caméra développée par SiConsult dans un élevage de biche (https://psi-biom.terroiko.fr/index.php/apps/files/?dir=/OCAPI_images_database/2023_11_24_Psibiom_Vision_Prototype_1_PICAREL_LE_HAUT&fileid=86307), reçues début décembre, pourront être testées par la suite.

2.2 Vision offline YOLOv5

Pour l'IA VISION offline, le choix s'est tourné vers un modèle YOLOv5 [1] 'You Only Look Once' implémenté sur les GPU UTLN. Ce type de réseau a pour objectif de détecter et classifier plusieurs classes sur des images et est très utilisé dans les problématiques de reconnaissance sur des vidéos notamment.

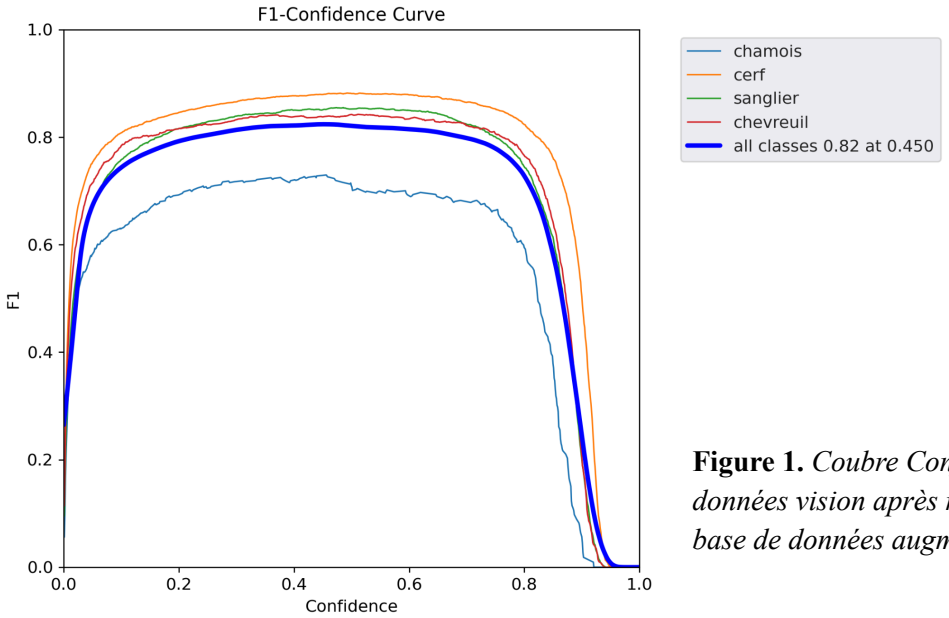

Figure 1. *Coubre Confiance-F1 sur les données vision après réentraînement sur la base de données augmenté*

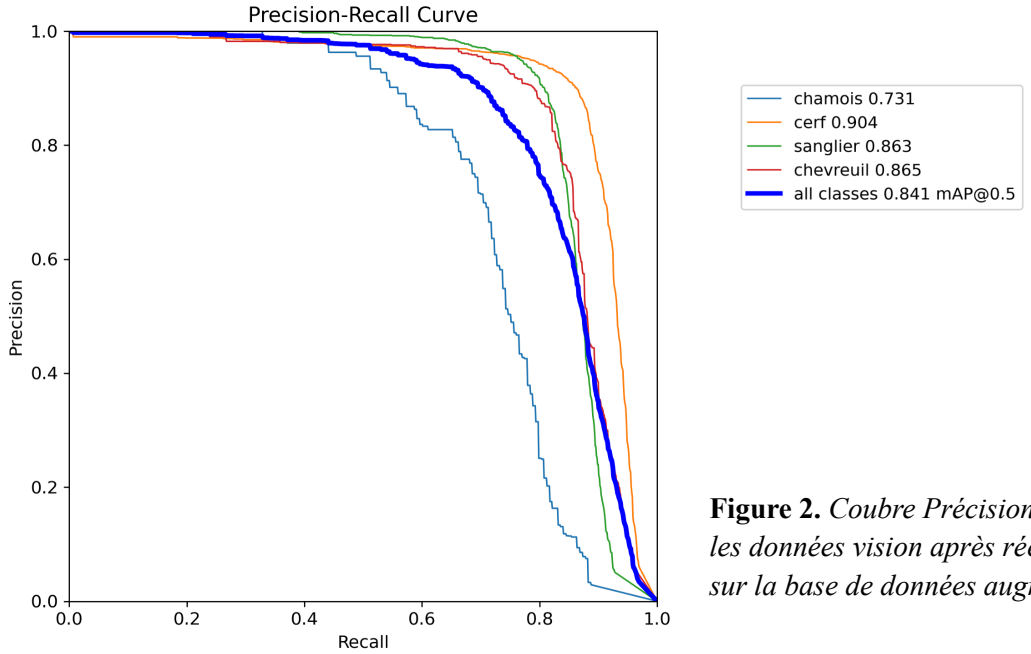

Figure 2. *Coubre Précision-Recall sur les données vision après réentraînement sur la base de données augmenté*

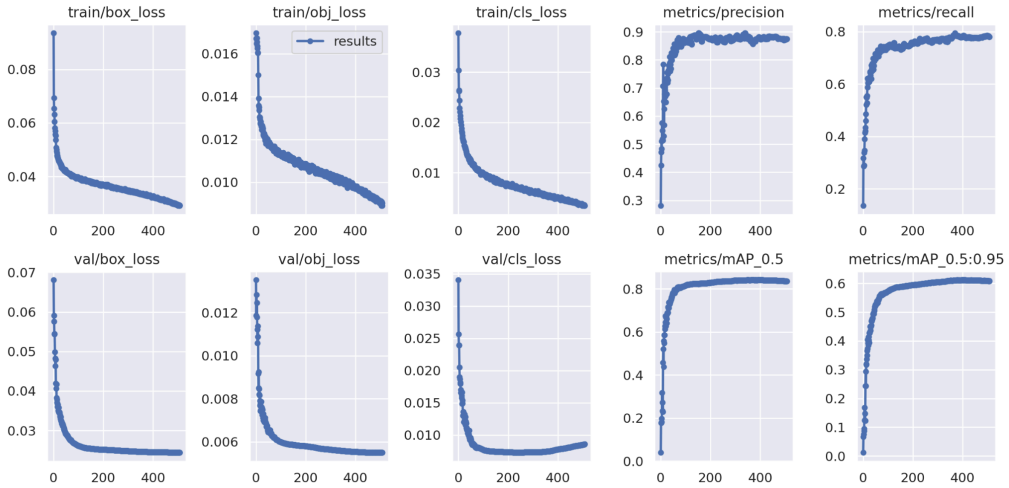

Figure 3. *Résultats de l'entraînement du réseau YOLOv5s sur la base d'annotation à jour en déc. 2023*

Comme le montrent les résultats ci-dessus (Fig. 1, 2 et 3), l'entraînement du modèle vision a permis d'obtenir de très bons résultats sur la détection des 3 classes d'intérêt (Sanglier, Cerf et Chevreuil). En effet avec des scores respectivement de mAP@0.5 à 0.863, 0.904 et 0.865, le modèle semble très performant. De plus, la courbe de précision-recall montre que mis à part les chamois, non considérés dans la liste des animaux d'intérêts à ce niveau là du développement, avec une précision de 0.8, on obtient un recall de 0.8 également. Ceci signifie donc qu'il y a une bonne précision sur les détections et que le modèle manque peu de détections.

a.

b.

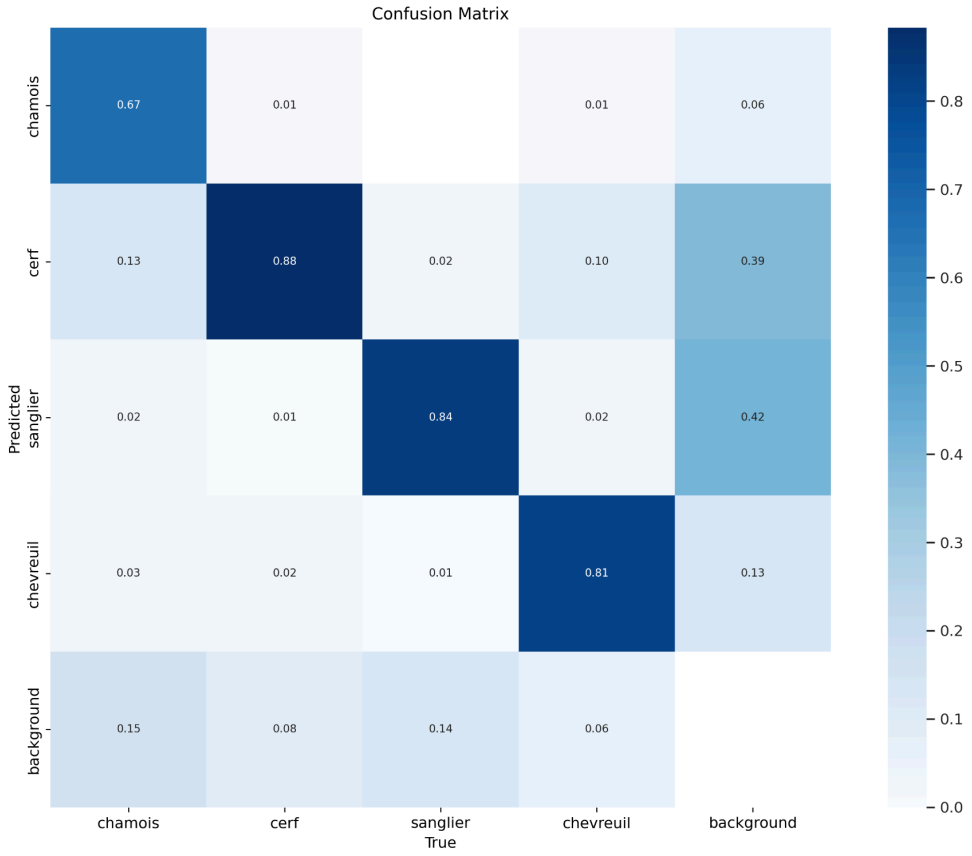

Figure 4. *Matrice de confusion du premier modèle (a) et du dernier modèle (b)*

La figure ci-dessus (Fig. 4) met en évidence la progression faite par le modèle grâce à l'ajout des nouvelles annotations de cerf. Avec un très mauvais score lors des premiers entraînements (16% de bonnes prédictions), c'est désormais la meilleure espèce détectée avec 88% de bonnes prédictions. Les sangliers et chevreuil eux sont constant dans la détection avec + de 80% de bonne prédiction.

2.3 Vision online

Du fait de la contrainte de la carte Brainchip pour le module VISION, le modèle Yolov5, ne peut être utilisé en IA embarquée. En effet, le modèle doit convenir à un système TensorFlow tandis que YoloV5 est en PyTorch, ce qui le rend non compatible. Afin d'avoir une compatibilité parfaite, il a été décidé d'utiliser le modèle YoloV2 produit par Brainchip directement.
Ce modèle n'a pu être entraîné avec la nouvelle base de données, ce sont donc les scores du précédent modèle qui sont données ci-dessous.

Les scores obtenus avec ce modèle sont moins satisfaisants que pour la version offline. En effet en s'intéressant uniquement au 6 classes, c'est-à-dire en supprimant les moutons, on obtient une mAP de 63 % avec un maximum pour les chèvres (79.8%), suivi par les chevaux (77.2%), les chamois (71.9%) et enfin les Cervidae (67.4%). La mAP des sanglier est de 37.7 % et celle des bovins de 45.9 %.

3. Audio

3.1 Base de données

La base de données utilisée pour l'entraînement des réseaux de neurones au vu de détecter des signaux acoustiques est composée par des annotations manuelles de Maxime Cauchois, Elodie Massol et David Funosas. Ces annotations ont été retravaillées pour correspondre au réseau utilisé.

3.2 Audio Offline YOLOv5

Pour les mesures d'activités taxonomiques, un algorithme de détection d'objet de type YOLO 'You Only Look Once' a été développé. Il a été décidé de travailler sur YOLOv5 [1] étant parmi ceux obtenant les meilleurs résultats. Cette méthode de traitement permet dans un premier temps d'éviter les fausses détections, du fait que le réseau n'est pas forcé à effectuer des prédictions sur chaque spectrogrammes, deuxièmement elle permet d'atténuer les problèmes liés aux chorus, c'est-à-dire la superpositions des chants de plusieurs espèces. Cette particularité a en effet été observée dans l'étude et s'explique par le fait que YOLO apprend des motifs en temps/fréquence, soit la représentation de la vocalise, et peut ainsi faire la distinction entre deux espèces qui émettent des vocalises sur une même fréquence et au même moment.

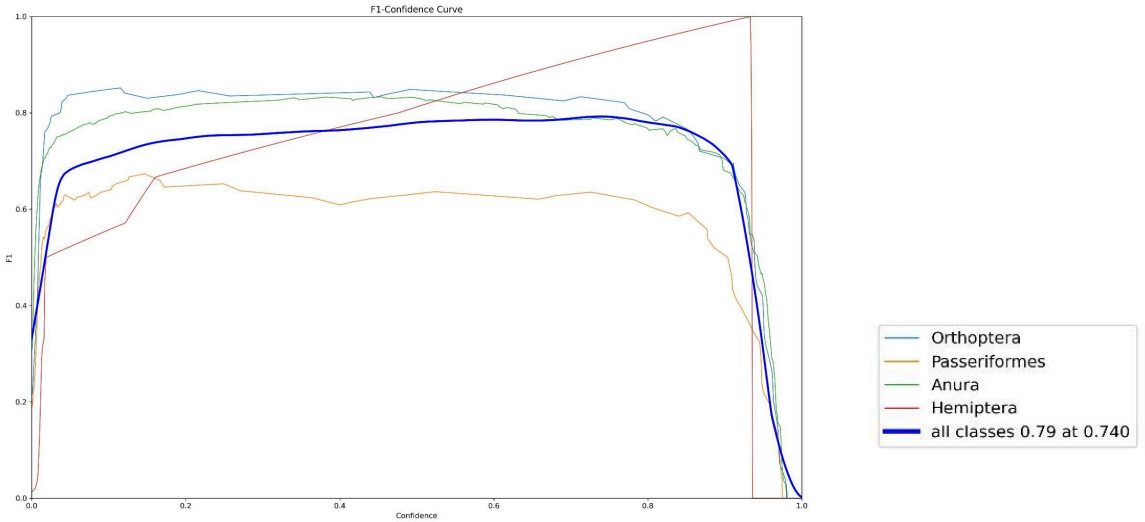

Figure 5. *Coubre Confiance-F1 sur les données acoustique après amélioration des annotation*

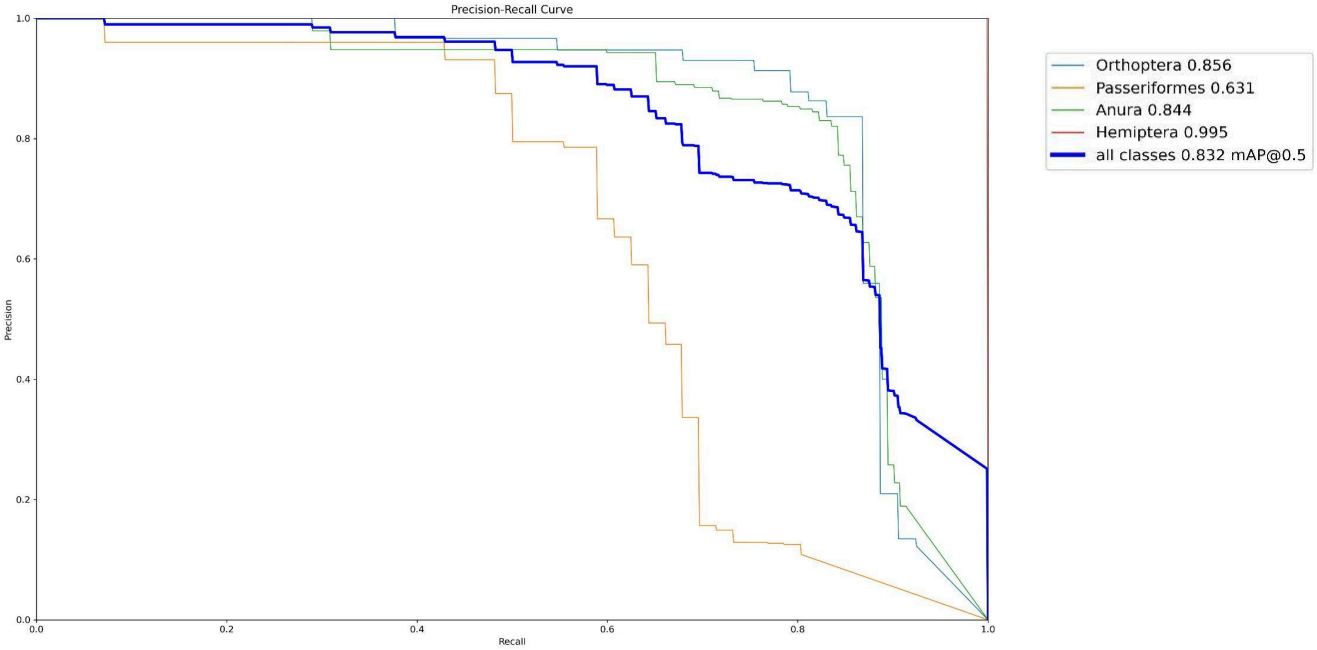

Figure 6. *Coubre Précision-Recall sur les données acoustiques après amélioration des annotations*

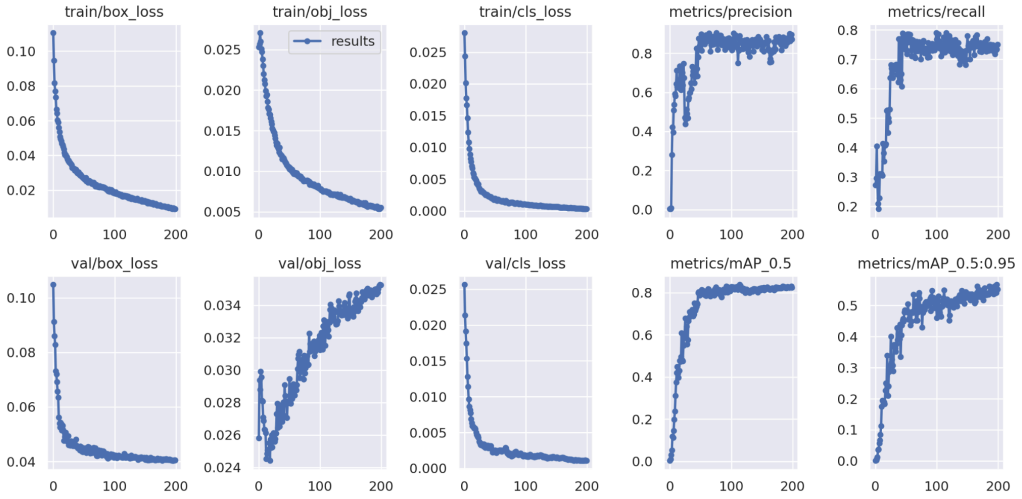

Figure 7. *Résultats de l'entraînement du réseau YOLOv5s sur la base d'annotation à jour*

Avec le jeu d'entraînement utilisé pour réaliser de la détection des 3 classes : Orthoptères, Anoures et Passeriformes, le réseau semble avoir bien appris à faire la différence et ainsi à détecter sur de nouveaux enregistrements.

Malgré un score plus faible pour les oiseaux, en raison notamment de la diversité des chants d'oiseaux et la différence entre cris et chants, la mAP@0.5 dépasse les 0.8. La courbe précision-recall suggère tout de même une difficulté pour obtenir toutes les détections de la classe Passeriformes car malgré un précision faible, au-delà d'un recall de 0.8 le réseau ne fait plus vraiment de détections.

L'avantage de YOLOv5 dans ce cas est qu'il ne fera pas de fausses détections, mais ne fera simplement pas de détections. Pour ce qui est des Anoures et des Orthoptères, avec une précision de 0.8 on obtient un recall de 0.84, ce qui est très bon. Les sons de ces deux classes sont en effet plus stéréotypés que les chants d'oiseaux et donc plus simples à apprendre (Fig. 8).

Pour améliorer les scores de détection des oiseaux, il faudra donc entraîner sur une plus grande base de données qu'actuellement.

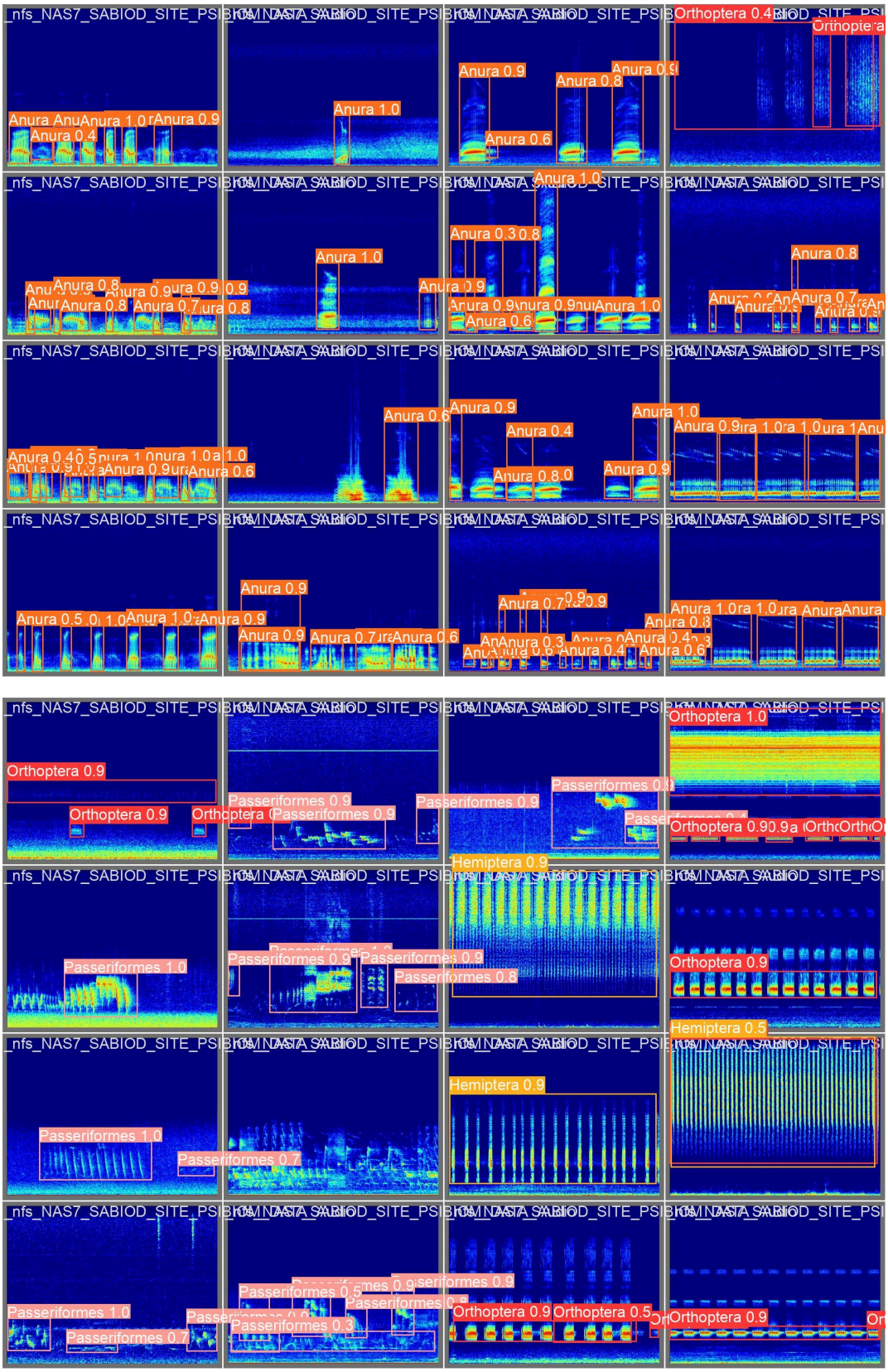

Figure 8. *Exemples de détections d'orthoptères, oiseaux et anours avec le dernier modèle*

a.

b.

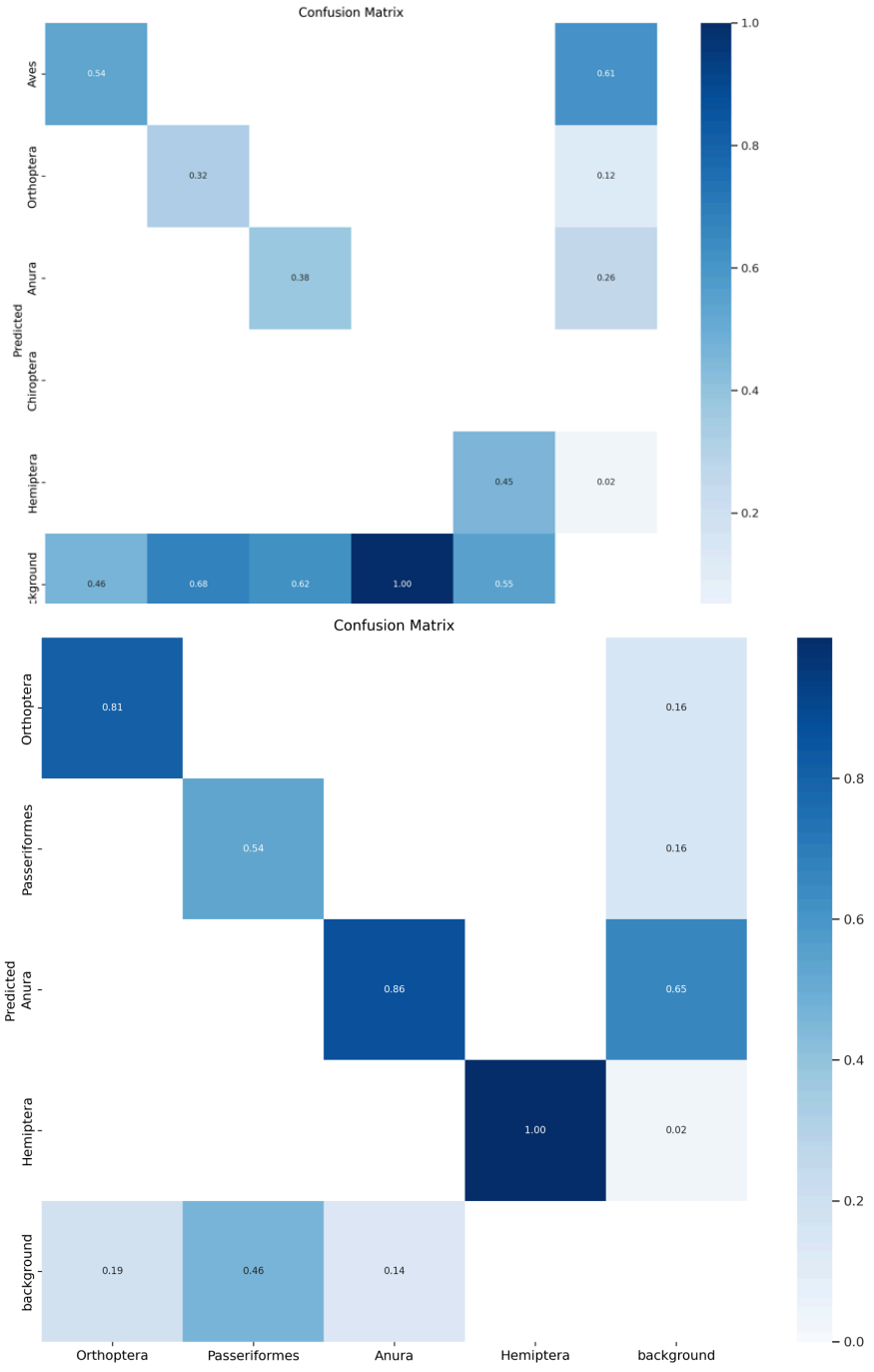

Figure 9. *Matrice de confusion du premier modèle (a) et du dernier modèle (b)*

Comme le montrent les deux matrices ci-dessus (Fig.9), une grosse amélioration à été faite dans le cas de la détection par acoustique. Alors qu'aucun scores ne dépassaient 54% de bonnes prédictions avec le premier modèle, ce sont désormais 81% des orthoptères qui sont bien classés, 86% des Anoures et 54% des annotations d'oiseaux qui sont classés. À noter que 46% des annotations d'oiseaux passent dans la case background, c'est-à-dire qu'ils ne sont pas considérés par le réseau comme des signaux sonores d'intérêt. Cela peut-être lié à un mauvais rapport signal à bruit dans les enregistrements,

rendant ainsi le signal peu visible dans le spectrogramme et donc non détectable avec une bonne précision.

3.3 Audio Online CNN

Une nouvelle architecture à seulement 6 couches (3 convolutions) pour 3 taxons (contre +/- 35 convolution avec YOLOv5) a été utilisée pour répondre à la problématique de l'embarquabilité du réseau. En effet, pour être utilisable sur carte et durable dans le temps au point de vue énergétique, le réseau se doit d'être léger et de ne pas consommer beaucoup d'énergie pour réaliser la détection. Dans le cas de ce nouveau réseau, la compression des poids et des calculs de convolution permet l'embarquement sur le PIC 32 de la carte QHB.
Le traitement des enregistrements se fait par fenêtres de 5 secondes avec prédiction de présence ou absence d'un ou plusieurs taxons = classification multilabel.
Avantages de ce nouveau réseau :

- Faible consommation d'énergie,
- Calculs rapides.

Pour le moment, la classification ne prend pas en compte les chiroptères dans la détection mais le fait que les enregistrements soient en expansion de temps fait qu'il n'est dans tous les cas pas possible d'utiliser le même réseau pour cette tâche.

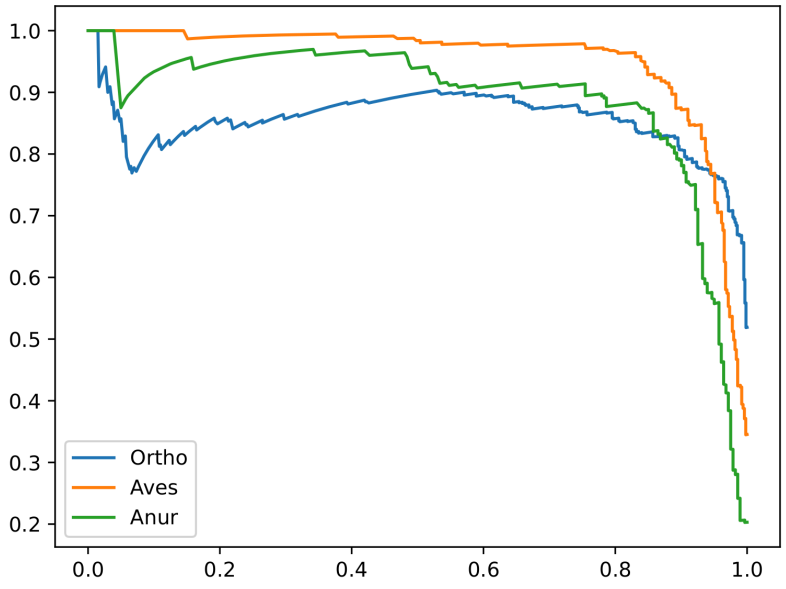

Figure 10. *Courbe Précision-Recall des 3 classes*

- Scores entraînement

classes : [Orthoptera, Aves, Anoures]
mAP 0.96; ROC AUC 0.93
Average Précision detailed : [0.97, 0.95, 0.96]

- Scores Test PSIBIOM

mAP 0.90; ROC AUC 0.94
Average Precision detailed : [0.85, 0.95, 0.89]

4. Discussion et conclusion

On mesure que les trois modèles appris sont dans des intervalles fonctionnels. Une progression est encore possible en vision offline avec une correction des labels de la base et un active learning. Le offline audio sera aussi augmenté avec plus de conditions bruitées ajoutées au training. Ce qui profitera aussi au online audio en fin janvier

.