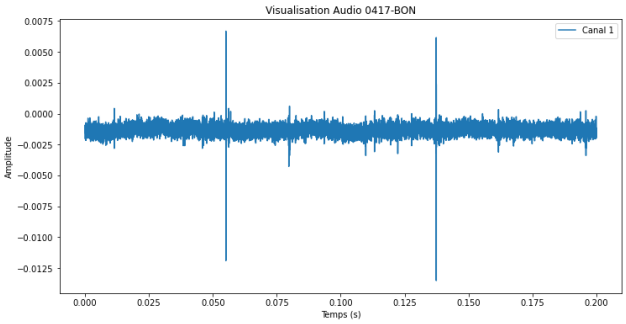
Data challenge ENS

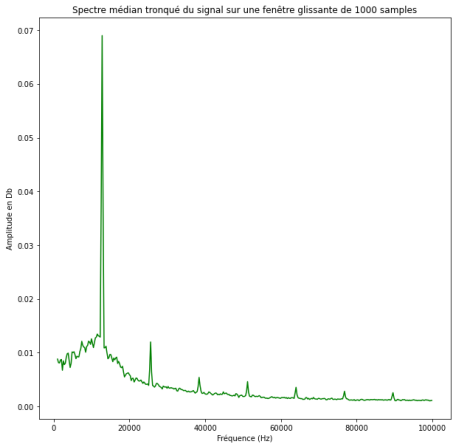
Compte-rendu de travail

I Biosonar - Détection de clics d'Odontocètes

*C*e dernier projet, un data challenge en collaboration avec l'université de Toulon, s'est révélé être l'aspect le plus captivant de ce semestre. Le défi consistait à détecter la présence de biosonars dans des enregistrements audio marins, sur un vaste ensemble de plus de 30 000 échantillons. Notre objectif était d'identifier la présence de dauphins ou de cachalots en utilisant des modèles de Machine Learning et de Deep Learning.

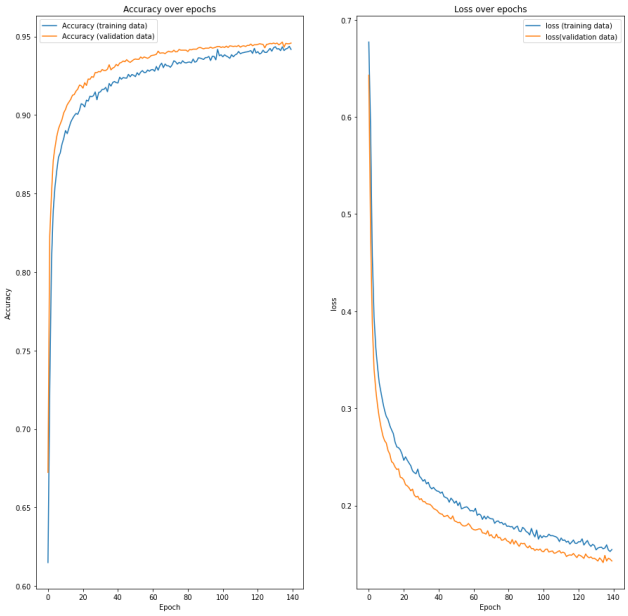
*L*a première étape consistait à maîtriser le jeu de données imposant. Nous avons utilisé la bibliothèque Librosa pour l'analyse préliminaire des données audio. Notre approche s'est divisée en quatre parties distinctes, chacune explorée dans un Jupyter Notebook. Nous avons commencé par examiner les fréquences communes de chaque signal et leur amplitude moyenne en utilisant une transformée de Fourier roulante et d'autres caractéristiques spectrales telles que le centre spectral et la platitude spectrale, extraites avec Librosa.

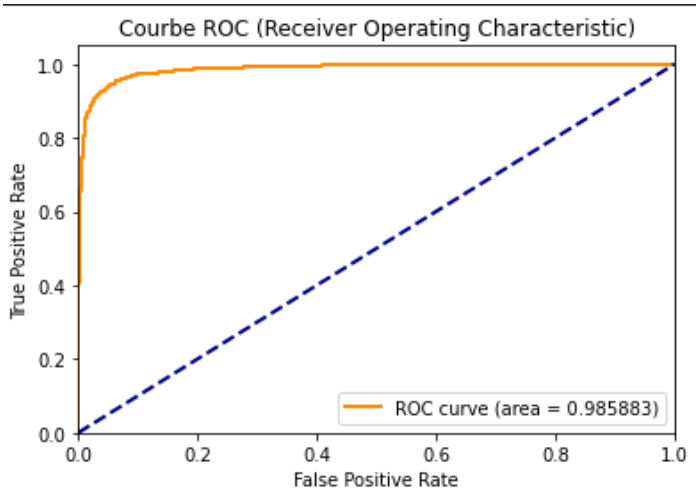




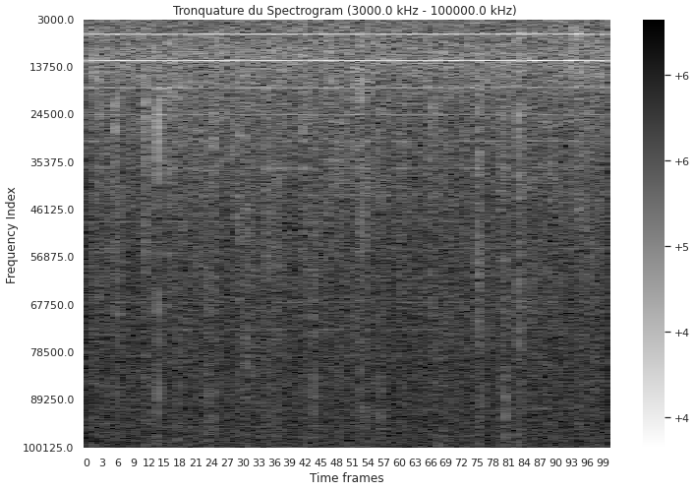
*F*ace à l'ampleur des données, nous ne pouvions pas charger toutes les caractéristiques d'un signal dans un tableau pour le traitement. Nous avons donc synthétisé l'information de chaque enregistrement audio en un nombre restreint de variables explicatives, stockées dans un DataFrame Pandas après s'être assurés de leur homogénéité.

*N*ous avons d'abord utilisé un modèle simple de Random Forest, qui a surpassé le benchmark mais sans atteindre un niveau de performance satisfaisant. Nous avons ensuite élaboré un modèle de Deep Learning utilisant des réseaux de neurones multicouches avec TensorFlow et Keras, parvenant à un score proche des 90 % sur la plateforme.





Nous avons également travaillé avec les spectrogrammes des enregistrements audio, une tâche complexe en raison de leur volume en mémoire. Après avoir ajusté les spectrogrammes pour se concentrer sur l'amplitude des fréquences pertinentes (3KHz - 150KHz) et les avoir transformées en valeurs logarithmiques, nous avons élaboré un modèle de réseau de neurones convolutif (CNN) pour traiter ces images de spectrogrammes. Malgré les longues durées d'entraînement (5h à 6h parfois), nous avons atteint seulement un score de 91 %.



Dans la troisième partie du challenge, nous avons expérimenté avec des modèles hybrides, combinant des données de Fourier et de spectrogrammes, mais ces approches se sont avérées moins performantes, avec un taux de réussite de seulement 76 % et un manque flagrant de capacité à généraliser.

En conclusion, ce data challenge a été extrêmement stimulant. Bien que de nombreux modèles n'aient pas atteint les performances escomptées, l'expérience acquise dans l'implémentation de réseaux de neurones a été très enrichissante. La progression du score de validation a révélé l'importance d'explorer différentes architectures et approches dans le traitement des données complexes.