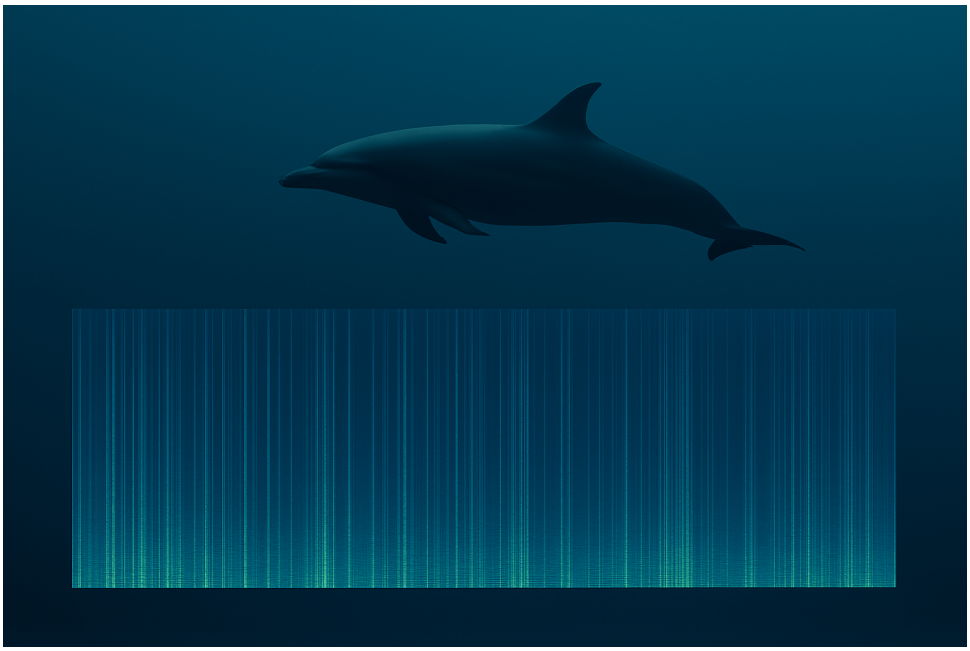
UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Détection de clics d'Odontocètes



TARVERDIAN Mariam, VO Nguyen Thao Nhi

Master 2 MoSEF 2024 - 2025

Sous la supervision de Mr Roman Yurchak et Ahmed Coulibaly

1 Introduction

Ce projet exploite les données de la plus vaste base d'enregistrements acoustiques sous-marins des Antille, recueillies dans le cadre du projet européen CARI'MAM (2017–2021). Avec plus de **25 To** d'enregistrements provenant de **17 stations**, la détection automatique des vocalises de baleines à bosse a montré de bons résultats. En revanche, l'identification des clics d'odontocètes (dauphins, cachalots...) reste difficile en raison du **bruit ambiant complexe**. L'objectif est de déterminer si un extrait audio contient des **biosonars** (clics de delphinidés) ou des bruits transitoires d'origine environnementale.

2 Données

La base se divise en deux ensembles : un ensemble d'entraînement issu de 8 sites, et un ensemble de test provenant de 2 autres sites, sans chevauchement. Chaque fichier audio, au format **WAV**, dure **200 ms** et est centré sur un clic potentiel. Les fichiers (**16 bits**, **256 kHz**) peuvent contenir des clics de delphinidés ou des bruits transitoires. L'ensemble d'entraînement comprend **23 168 fichiers annotés** (1 : clic présent, 0 : absence), tandis que le test contient **950 fichiers non annotés**, au même format. 1

3 Métriques d'évaluation

Pour évaluer et comparer objectivement les performances des différents modèles, nous utilisons le score AUC-ROC (Area Under the Receiver Operating Characteristic Curve). Cette métrique est particulièrement adaptée aux problèmes de classification binaire et mesure la capacité du modèle à distinguer les deux classes. L'AUC-ROC correspond à l'aire sous la courbe ROC, qui trace le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR) pour différents seuils de décision :

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) \tag{1}$$

avec $\text{TPR} = \frac{\text{VP}}{\text{VP+FN}}$ et $\text{FPR} = \frac{\text{FP}}{\text{FP+VN}}$, où VP, FN, FP et VN représentent respectivement les vrais positifs, faux négatifs, faux positifs et vrais négatifs.

4 Modélisation

Dans ce projet, trois approches de modélisation ont été explorées pour la détection des clics d'odontocètes : (1) l'utilisation d'extractions de caractéristiques manuelles, (2) l'analyse du signal audio brut complétée par quelques caractéristiques extraites, et (3) l'exploitation de spectrogrammes. Chaque méthode s'appuie sur des architectures de réseaux de neurones adaptées aux spécificités de la tâche.

4.1 Approche par extraction de caractéristiques avec MLP

L'approche par extraction de caractéristiques présente plusieurs avantages : elle est plus légère en calcul, nécessite moins de données d'entraînement, offre une meilleure interprétabilité grâce à des caractéristiques souvent liées à des propriétés physiques ou perceptuelles, et fonctionne bien avec des jeux de données restreints. En revanche, elle dépend fortement de la pertinence des caractéristiques extraites, peut induire une perte d'information, requiert une expertise du domaine, et se montre moins efficace pour modéliser des patterns complexes ou non linéaires.

Pour cette raison, nous avons opté pour un **perceptron multicouche (MLP)**, une architecture de réseau de neurones classique mais efficace pour des données tabulaires issues d'extractions manuelles. Le MLP permet de modéliser des relations non linéaires entre les caractéristiques extraites et les classes cibles, tout en restant relativement simple à entraîner.

Tous les fichiers audio au format `.wav` sont chargés avec leur fréquence d'échantillonnage native pour préserver l'intégrité spectrale. Un filtre passe-bande entre **5000 Hz** et **100000 Hz** est appliqué afin d'éliminer les basses fréquences liées au bruit ambiant tout en conservant la plage des clics d'odontocètes. Pour chaque enregistrement, nous extrayons des caractéristiques temporelles et fréquentielles à l'aide de `librosa`, `scipy.signal` et `scipy.stats` : statistiques d'amplitude (moyenne, écart-type, min, max), puissance RMS, centroïde spectral, largeur de bande, flatness spectrale, fréquence de pic (FFT), ICI (intervalle inter-clic), SNR, kurtosis et skewness. 2

Dans notre modélisation par **MLP**, les caractéristiques sont d'abord normalisées via un **StandardScaler** pour compenser les différences d'échelle, accélérer la convergence et améliorer la généralisation. L'optimisation des hyperparamètres repose sur une recherche par grille **(GridSearchCV)** avec validation croisée stratifiée **(StratifiedKFold)** garantissant un équilibre des classes. Les paramètres testés incluent : fonctions d'activation, architecture des couches, régularisation (`alpha`) et stratégie d'apprentissage.

La configuration optimale retenue comporte une architecture **(128, 64)**, une activation **'tanh'**, une régularisation **L2** avec **alpha = 0.01**, et un taux d'apprentissage **constant**, garantissant un bon compromis entre performance et robustesse. Le modèle obtient un score **AUC-ROC de 0,9852** sur le jeu de validation, mais chute à **0,8651** sur les données de test, révélant un possible surapprentissage.

4.2 Apporche par CNN 1D hybride pour l'analyse des formes d'onde

Notre approche hybride utilise directement les formes d'onde brutes comme entrée d'un réseau neuronal convolutif 1D, tout en l'enrichissant par des caractéristiques spécifiques aux odontocètes. Cette méthode préserve l'information temporelle complète du signal tout en exploitant des attributs acoustiques pertinents. Le CNN 1D permet de traiter directement le signal temporel sans perte d'information, contrairement aux approches basées sur le spectrogramme, tout en apprenant automatiquement les filtres optimaux pour la tâche de détection. Cette synergie a permis d'atteindre un score AUC de **0,94**, une amélioration significative par rapport à l'approche CNN pure (0,87).

Le prétraitement audio comprend : (i) conversion en mono, (ii) filtrage passe-bande entre 5-120 kHz correspondant aux vocalisations des odontocètes, (iii) normalisation entre -1 et 1, et (iv) redimensionnement à longueur fixe. Parallèlement, des caractéristiques spécifiques sont extraites via transformation de Hilbert pour l'enveloppe du signal, quantifiant le nombre de clics, leur amplitude et la régularité des intervalles. Cette extraction ciblée s'est avérée déterminante pour l'amélioration des performances.

L'architecture CNN comprend trois couches convolutives conçues pour extraire progressivement des caractéristiques de plus en plus complexes :

1. **Première couche convolutive** : 32 filtres, noyau de taille 15, détectant les motifs simples et variations rapides du signal.
2. **Deuxième couche convolutive** : 48 filtres, noyau de taille 9, capturant des motifs plus complexes tout en préservant la résolution.

3. **Troisième couche convolutive** : 64 filtres, noyau de taille 5, se concentrant sur les détails fins essentiels à la détection.

Pour contrer le surapprentissage, nous implémentons une normalisation par lots après chaque couche convolutive, un dropout de 0,3 dans les couches denses, et un scheduler ajustant le taux d'apprentissage selon les performances de validation. L'entraînement utilise la binary cross-entropy et l'optimiseur Adam, avec early stopping pour une convergence optimale.

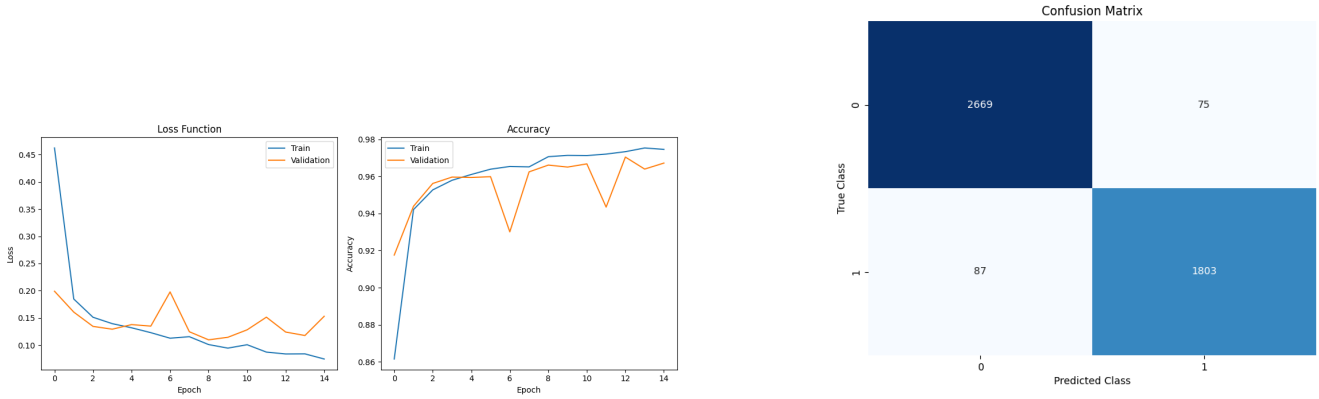
L'analyse des performances du modèle révèle plusieurs observations importantes :

— Le modèle réduit constamment ses erreurs sur les données d'entraînement (bleu) et de validation (orange), avec quelques fluctuations acceptables sur la validation.

— La précision du modèle augmente rapidement puis se stabilise à des valeurs élevées (94%), démontrant une excellente performance de classification des clics d'odontocètes.

— La matrice de confusion indique 2669 vrais positifs et 1803 vrais négatifs, contre seulement 75 faux négatifs et 87 faux positifs.

— Les performances sur notre ensemble de test interne sont légèrement supérieures au score AUC de 0,94 obtenu sur la plateforme d'évaluation, suggérant un léger surapprentissage.



(a) Courbes d'apprentissage du modèle

(b) Matrice de confusion

FIGURE 1 – Évaluation des performances du modèle CNN 1D hybride

Ces résultats indiquent que cette architecture hybride, combinant l'apprentissage automatique des représentations avec l'expertise du domaine, offre une approche efficace pour la détection des odontocètes.

4.3 Approche par CNN 2D sur spectrogrammes Mel

La troisième approche transforme les signaux audio en spectrogrammes Mel, qui représentent l'évolution des fréquences dans le temps, et utilise un réseau neuronal convolutif 2D pour analyser ces images spectrales. Cette méthode exploite la structure temps-fréquence des signaux audio, similaire à la manière dont l'oreille humaine perçoit les sons. Elle permet au modèle de capturer les relations complexes entre les différentes fréquences et leur évolution temporelle. Toutefois, cette approche présente certains inconvénients : la conversion en spectrogrammes entraîne une perte d'information de phase, ce qui peut affecter la précision dans certains cas. De plus, cette méthode nécessite une étape de transformation du signal avec des choix de paramètres spécifiques et est généralement plus coûteuse en termes de calculs.

Le prétraitement des données audio constitue une étape cruciale. Les enregistrements sont échantillonnés à une fréquence de 256 kHz, puis découpés en extraits de 200 ms. Ces extraits sont ensuite convertis en spectrogrammes mel avec une taille de FFT de 1024, une longueur de saut de 128 et 128 bandes mel. Nous avons choisi une fréquence minimale de **1000 Hz** pour cibler les caractéristiques des signaux de biosonar, avec la fréquence maximale fixée à 128 kHz. Une normalisation est appliquée aux spectrogrammes pour faciliter l'apprentissage du réseau. 3 4

L'architecture CNN se compose de deux parties principales :

1. **Couche d'extraction de caractéristiques** comprenant quatre blocs de convolution avec une profondeur croissante (16→32→64→128). Chaque bloc intègre une couche de convolution 2D suivie d'une normalisation par lots, d'une activation ReLU, d'un max pooling et d'un dropout pour réduire le surapprentissage. Un pooling adaptatif global finalise cette partie en réduisant la dimensionnalité.
2. **Couche de classification** constituée de deux couches linéaires (128→64→2) avec un dropout intermédiaire (taux de 0.5), produisant une sortie à deux classes correspondant à la présence ou l'absence de biosonar.

Pour améliorer la robustesse du modèle, deux techniques d'augmentation sont appliquées : ajout de bruit gaussien (niveau 0.005) et décalages temporels (jusqu'à 10% de la longueur). L'entraînement utilise l'entropie croisée et l'optimiseur Adam (taux initial de 0.001) avec un scheduler ReduceLROnPlateau. Un arrêt anticipé (patience de 3 époques) limite le surapprentissage, avec un maximum de 20 époques et des batchs de taille 16. Sur le classement public, cette approche a obtenu un score AUC-ROC de 0,65, inférieur aux autres méthodes testées.

5 Conclusion

Notre étude comparative des trois approches pour la détection des clics d'odontocètes a montré que le MLP (AUC 0,87), le CNN 1D hybride (AUC 0,94) et le CNN 2D sur spectrogrammes (AUC 0,65) offrent des performances variables. Le CNN 1D hybride s'est révélé le plus efficace grâce à la combinaison du signal brut et des caractéristiques spécifiques aux clics.

Plusieurs pistes d'amélioration ont été identifiées pour chaque modèle :

— **MLP :** Pour cette approche, l'extraction de caractéristiques supplémentaires ciblant spécifiquement les propriétés temporelles des clics pourrait enrichir la représentation des signaux. Une analyse d'importance des features permettrait également de sélectionner un sous-ensemble optimal de caractéristiques, réduisant ainsi le risque de surapprentissage observé.

— **CNN 1D hybride :** L'implémentation d'une validation croisée stratifiée sur 3 partitions améliorerait la robustesse du modèle face à la variabilité des données. L'augmentation des données via des techniques comme time_stretch() et pitch_shift() permettrait de diversifier les exemples d'entraînement et de mieux généraliser aux variations naturelles des clics.

— **CNN 2D :** Malgré ses performances inférieures, ce modèle pourrait être amélioré par une optimisation des paramètres de génération des spectrogrammes pour mieux capturer les caractéristiques distinctives des clics. L'application de techniques de débruitage spécifiquement adaptées aux environnements marins pourrait également réduire l'impact du bruit sur la qualité des spectrogrammes.

À l'avenir, un modèle d'ensemble combinant les prédictions des différentes approches pourrait exploiter leurs forces complémentaires et améliorer la robustesse globale du système de détection des clics d'odontocètes dans les environnements acoustiques complexes des Antilles.

A Annexes

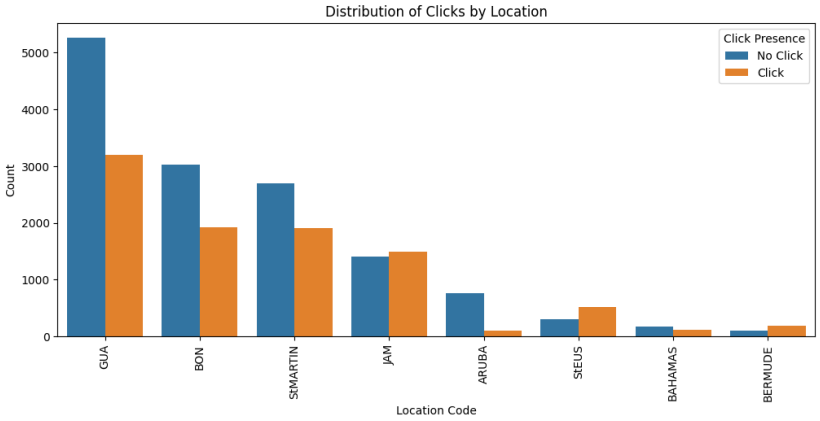


FIGURE 1 – Repartition des clicks

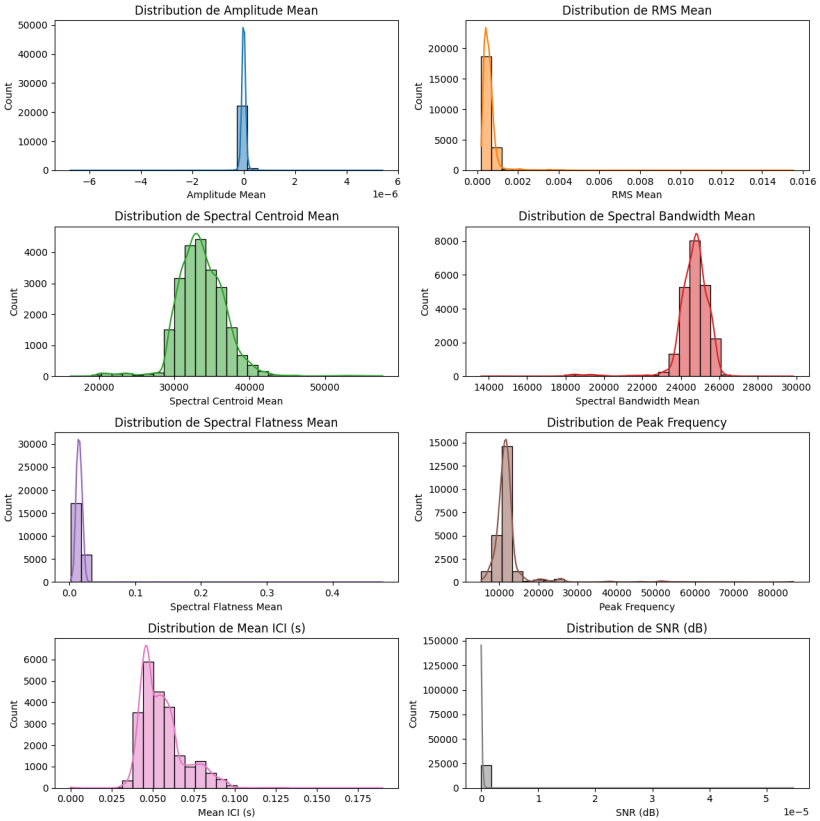


FIGURE 2 – Distribution de quelques caractéristiques extraites

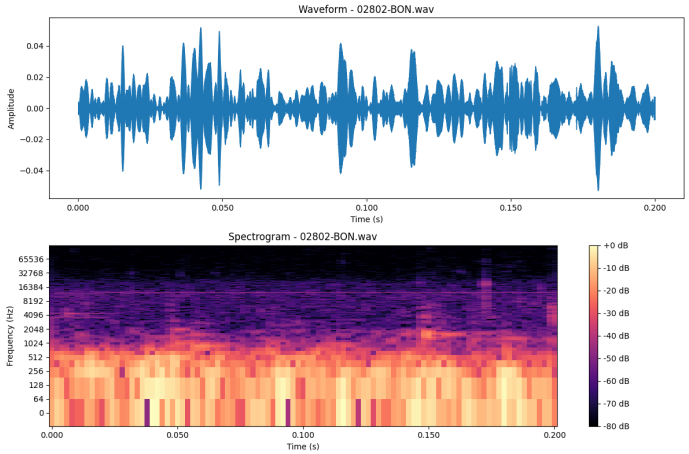


FIGURE 3 – Exemple d'analyse temporelle et spectrale (Forme d'onde et Spectrogramme)

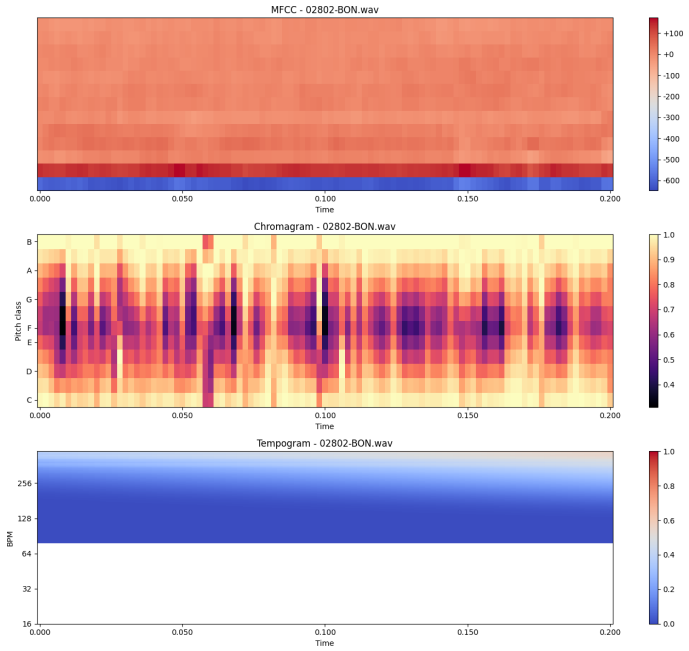


FIGURE 4 – Exemple de représentations fréquentielles MFCC, Chromagramme et Tempogramme)