



SORBONNE UNIVERSITÉ

M2 INFORMATIQUE - DAC  
REDS - RAPPORT 1

---

## Classification Hippopotames / Rhinocéros

---

**Etudiants :**

Grégoire GOUJON

Amayas SADI

Ghiles OUHENIA

Amine YOUSSEF KHODJA

Koceila KEMICHE

**Encadrant :**

Olivier SCHWANDER

## Table des matières

<b>1 Description des objectifs du dataset</b>	<b>2</b>
1.1 Provenance des photos . . . . .	2
1.2 Nettoyage du dataset . . . . .	2
1.3 Présentation des données . . . . .	2
<b>2 Analyse descriptive des données</b>	<b>3</b>
2.1 Paramètres de nos images . . . . .	3
2.2 Biais possibles . . . . .	4
<b>3 Aperçu des techniques utilisées</b>	<b>4</b>
3.1 Traitement sur les images . . . . .	4
<b>4 Conclusion</b>	<b>5</b>

# 1 Description des objectifs du dataset

L'objectif principal de ce dataset est de développer un algorithme de classification capable de distinguer efficacement entre les images d'hippopotames et de rhinocéros. La tâche de classification binaire permettra de contribuer à la conservation de ces espèces menacées en automatisant la surveillance à partir de photographies par exemple. Ce dataset peut être particulièrement utile pour les parcs animaliers, les organisations de conservation de la faune, les chercheurs et les amateurs de photographie de la faune. Il permettra d'automatiser l'identification des animaux à partir d'images, ce qui pourrait accélérer la détection de menaces ou la collecte de données sur ces espèces.

## 1.1 Provenance des photos

Les données pour ce projet ont été acquises en utilisant une API spécialisée dans la collecte d'images. L'API a été configurée pour récupérer des images d'hippopotames et de rhinocéros à partir de diverses sources en ligne contenant principalement des contributions de photographes amateurs. Cette approche nous a permis de créer un ensemble de données diversifié et représentatif des deux espèces cibles.

L'utilisation de l'API a simplifié le processus de collecte de données en automatisant le téléchargement des images. Cependant, il était nécessaire de mettre en place des filtres pour garantir la qualité et l'éthique des images collectées. Toutes les images étaient liées à une source fiable et respectueuse des animaux.

## 1.2 Nettoyage du dataset

Une fois les images téléchargées, elles ont été soumises à une vérification manuelle pour s'assurer de la conformité aux critères de qualité, notamment l'exclusion d'images floues, bruyantes ou inappropriées. Le nettoyage des données est une étape vitale dans notre projet de classification entre hippopotames et rhinocéros. Il vise à préparer nos données brutes pour garantir la fiabilité de nos résultats et à nous permettre de transformer le bruit des données brutes en des signaux utiles, assurant ainsi la qualité de notre dataset et la précision de notre modèle de classification.

## 1.3 Présentation des données

Après un nettoyage à la main nous ne gardons que des images de "bonnes qualités". Un humain pourrait très facilement classer les différentes images.



FIGURE 1 – Hippopotame



FIGURE 2 – Rhinocéros



FIGURE 3 – Hippopotame

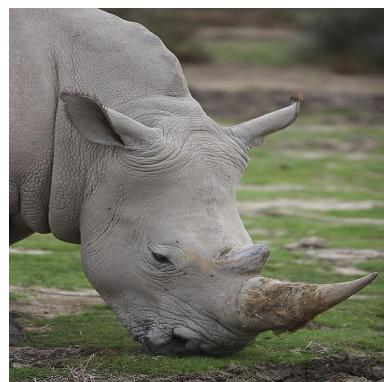


FIGURE 4 – Rhinocéros

## 2 Analyse descriptive des données

Une analyse exploratoire des données a été réalisée pour comprendre la distribution des classes, les caractéristiques des images, et pour identifier les éventuelles anomalies ou problèmes de qualité des données.

### 2.1 Paramètres de nos images

- **Les images :** Le dataset contient 10000 images d'hippopotames et 10000 images de rhinocéros. Elles présentent une variété de résolutions, allant de 672x480 à 5472x3648 pixels. Un redimensionnement est nécessaire, d'autant que certaines sont bien trop lourdes pour pouvoir travailler efficacement avec.
- **Répartition :** La répartition des classes est équilibrée, avec 50% d'hippopotames et 50% de rhinocéros. Cela permet de prévenir des biais potentiels et de garantir que notre modèle de machine learning sera capable d'apprendre efficacement à discriminer entre les classes.
- **Diversité :** Les images montrent des animaux dans divers contextes, y compris des plans rapprochés, des plans d'ensemble, des images de jour et de nuit. Une diversité insuffisante dans les images peut entraîner une spécialisation excessive du modèle sur des cas spécifiques, ce qui le rend moins capable de faire face à de nouvelles situations ou de traiter des variations naturelles.

## 2.2 Biais possibles

Il est nécessaire d'être critique sur notre dataset même si nous avons essayé au mieux d'éviter les biais lors de la sélection et du nettoyage.

- **Biais de sélection des sources** : Si les sources de données d'où sont collectés les images présentent des préjugés ou des préférences, cela peut se refléter dans le dataset. Exemple : si une source a principalement des images d'hippopotames dans un contexte de parc animalier, tandis qu'une autre source a principalement des images de rhinocéros dans la nature sauvage, cela peut créer un biais en faveur de certaines conditions.
- **Biais de qualité des images** : Si la qualité des images varie considérablement entre les deux classes, le modèle peut être influencé par la qualité plutôt que par les caractéristiques réelles des animaux.
- **Biais de sélection d'âge** : Si les images sont principalement d'animaux d'un certain âge, cela peut entraîner un biais envers cette tranche d'âge particulière, affectant la capacité du modèle à reconnaître des animaux de différents âges.
- **Biais d'annotation** : Si la classification des données d'entraînement contient des erreurs provenant de l'extraction, cela peut affecter la qualité des données et entraîner des résultats incorrects. Ce biais sera normalement évité puisque nous avons nettoyé le dataset nous-mêmes à la main.

Les biais les plus forts que pourrait avoir notre algorithme seraient principalement dûs au contexte dans lequel sont prises les photos. Un élément est il redondant dans la majorité des photos d'une classe et pas dans l'autre. Actuellement nous en identifions un qui pourrait poser problème : l'eau. En effet la majorité des photos d'hippopotames sont prises dans ou à côté d'un point d'eau alors que c'est rarement le cas pour les rhinocéros. Il sera nécessaire de prendre en compte ce biais possible lors du traitement des images.

## 3 Aperçu des techniques utilisées

Les images ont été collectées à partir de sources diverses, puis annotées manuellement pour garantir leur précision. Les techniques de prétraitement des images, telles que le redimensionnement et la normalisation, ont été appliquées pour garantir la cohérence des données.

### 3.1 Traitement sur les images

- **Redimensionnement** : Les images collectées étant de tailles diverses, il est nécessaire de normaliser leur taille pour qu'elles correspondent à l'entrée de notre modèle. Nous les redimensionnons toutes à une taille moindre (256x256), cela les rendra plus légères et permettra de mieux les manipuler.
- **Passage au noir et blanc** : Les images en couleur nécessitent plus de données pour l'entraînement de modèles, car il y a plus de variations possibles. Les images en couleurs ont trois canaux (rouge, vert, bleu) pour chaque pixel, tandis que les images en noir et blanc n'ont qu'un seul canal. De plus nous pensons que cela pourrait permettre d'éviter le biais de l'eau énoncé plus haut car sa couleur en noir

et blanc se rapprocherait de celle des paysages des autres photos. Mais l'objectif principal est de réduire la complexité des données.



(a) Image de base  
(3264x2648)



(b) Redimensionnement  
(256x256)



(c) Passage au noir et  
blanc

Nous pourrions éventuellement effectuer du recadrage (cropping), cela consiste à découper une région d'intérêt dans une image. Cela peut être utile pour éliminer des parties inutiles de l'image ou pour extraire une région spécifique d'intérêt. En effet sur la majorité des photos d'hippopotames ou de rhinocéros, l'animal est au centre de l'image puisqu'il est l'objet photographié. Nous pourrions également augmenter les données (variations des images existantes) si notre dataset n'est pas suffisant.

## 4 Conclusion

L'ensemble de données pour la classification d'hippopotames et de rhinocéros a été soigneusement construit à partir de sources fiables et annoté avec précision. L'analyse exploratoire des données révèle une distribution équilibrée des classes, ce qui est favorable pour la création d'un modèle de classification binaire. Les prochaines étapes consisteront à concevoir et à former un modèle d'apprentissage automatique pour cette tâche.