ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

HỌC MÁY - CO3117

Bài tập lớn 3

Phân loại hình ảnh động vật biển

Sea animals image Classification

Giảng viên hướng dẫn: TS. Lê Thành Sách
Lớp: TN01
Nhóm: DNA05

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 11 - 2025

BÁO CÁO KẾT QUẢ LÀM VIỆC NHÓM

STT	MSSV	Họ và Tên	Nhiệm vụ phụ trách	Mức độ hoàn thành
1	2310167	Tăng Hồng Ái	Thực hiện phân tích EDA: thống kê và trực quan hóa kích thước, độ sáng, độ tương phản, phân bố nhãn. Mô tả đặc trưng dữ liệu và viết phần báo cáo EDA.	100%
2	2310510	Phạm Khánh Duy	Tiền xử lý và tăng cường dữ liệu ảnh. Xây dựng pipeline trích xuất đặc trưng bằng **ResNet50, EfficientNetB0, ViT**. Lưu và quản lý dữ liệu đặc trưng và viết phần báo cáo tương ứng.	100%
3	2312506	Nguyễn Trần Yến Nhi	Huấn luyện và đánh giá các mô hình **Logistic Regression, SVM, Random Forest**. Tổng hợp kết quả, trực quan hóa và viết phần báo cáo phân tích mô hình.	100%

Bảng 1: Bảng phân công công việc và mức độ hoàn thành

Mục lục

1 EDA

1.1 Tổng quan dữ liệu

Bộ dữ liệu sử dụng trong bài này là `Sea Animals Image Dataset` gồm 13.711 hình ảnh ở định dạng .jpg và .png. Dữ liệu bao gồm nhiều hình ảnh của các loài sinh vật biển khác nhau như: cá, sứa, rùa, cá mập, cá heo, sao biển,... phục vụ cho bài toán nhận dạng và phân loại sinh vật biển. Bộ dữ liệu được chia thành các thư mục, mỗi thư mục tương ứng với một lớp (*class*) đại diện cho một loài sinh vật biển cụ thể.

Mục tiêu của phần EDA là phân tích đặc trưng tổng quan của dữ liệu hình ảnh bao gồm kích thước, tỷ lệ, độ sáng, độ tương phản, màu sắc và phân bố lớp — nhằm hiểu rõ cấu trúc dữ liệu trước khi huấn luyện mô hình phân loại.

1.2 Tổng quan dữ liệu ảnh

Tất cả các ảnh trong thư mục dữ liệu được duyệt qua và lưu lại các thông tin:

- **Class**: tên lớp sinh vật biển.
- **Image_Name**: tên tệp ảnh.
- **Path**: đường dẫn tuyệt đối.
- **Width, Height**: kích thước ảnh (pixel).
- **Channels**: số kênh màu.

	Class	Image_Name	Path	Width	Height	Channels
0	Clams	11127770525_eb487e975c_b.jpg	/kaggle/input/sea-animals-image-dataste/Clams/...	200	300	3
1	Clams	7409907104_e68910d92d_o.jpg	/kaggle/input/sea-animals-image-dataste/Clams/...	300	225	3
2	Clams	24659170965_ed13a20e49_o.jpg	/kaggle/input/sea-animals-image-dataste/Clams/...	300	200	3
3	Clams	134917699_add44eaf73_o.jpg	/kaggle/input/sea-animals-image-dataste/Clams/...	225	300	3
4	Clams	11587746256_6d931eb556_o.jpg	/kaggle/input/sea-animals-image-dataste/Clams/...	300	225	3

Total images loaded: 13711

Hình 1.1: Thông tin ảnh

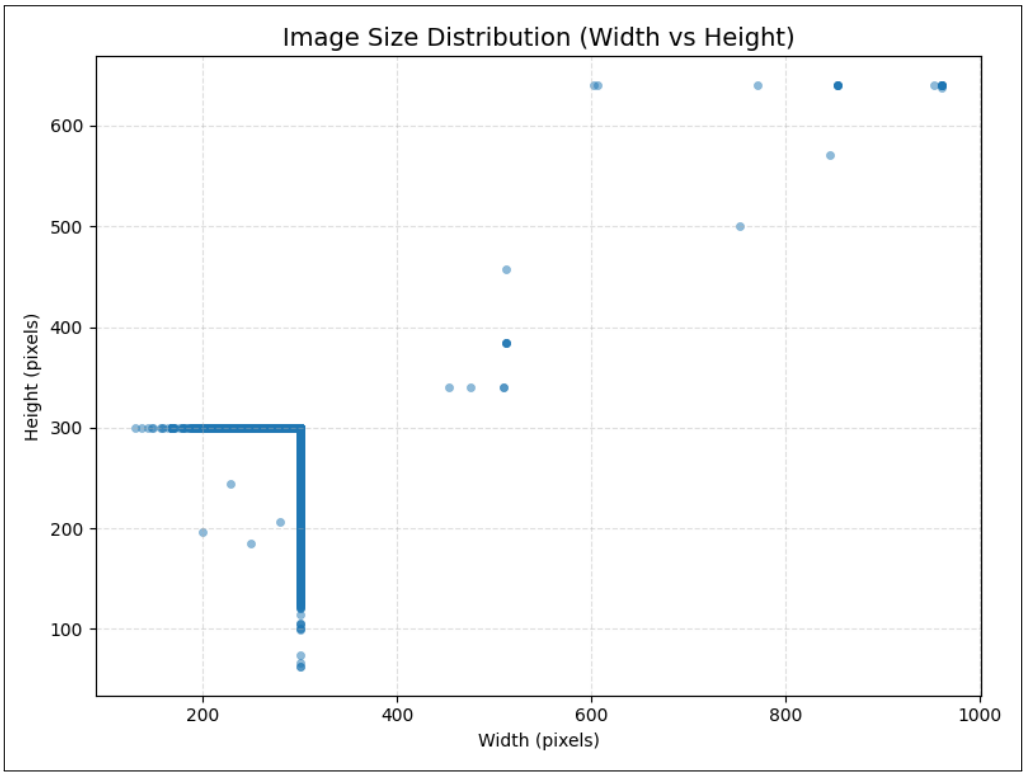
1.3 Phân tích kích thước ảnh

Dữ liệu gồm 13.711 ảnh với các thống kê chi tiết như sau:

Thông số	count	mean	std	min	25%	50%	75%	max
Width	13711.0	293.07	55.04	131.0	300.0	300.0	300.0	5120.0
Height	13711.0	222.41	49.81	63.0	200.0	218.0	225.0	3840.0
Channels	13711.0	3.00	0.05	1.0	3.0	3.0	3.0	4.0

Bảng 2: Thống kê chi tiết kích thước và kênh ảnh

Kết quả cho thấy kích thước trung bình của ảnh là khoảng 293 × 222 pixel, với độ lệch chuẩn lần lượt là 55 và 50. Điều này cho thấy phần lớn ảnh có kích thước tương đối đồng đều, tập trung quanh mức 300 pixel chiều rộng và khoảng 220 pixel chiều cao.



Hình 1.2: Phân bố kích thước ảnh

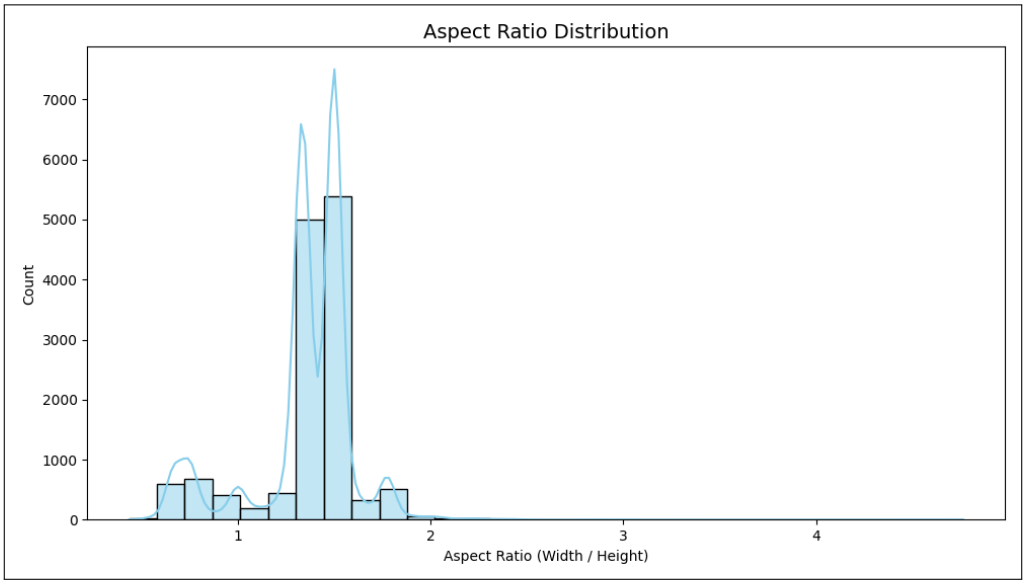
Biểu đồ phân bố kích thước ảnh minh họa rõ rằng đa số các điểm dữ liệu tập trung thành cụm dày ở vùng (Width ≈ 300, Height ≈ 200–230). Ngoài ra, có một số ảnh ngoại lệ (outliers) có độ phân giải rất cao, lên đến 5120×3840 pixel, tuy nhiên số lượng này rất nhỏ.

Đa số các ảnh có số kênh là 3, tương ứng với ảnh RGB thông thường. Điều này cho phép quá trình tiền xử lý và huấn luyện mô hình xử lý ảnh (CNN, ViT, v.v.) diễn ra thống nhất mà không cần chuyển đổi định dạng.

1.4 Tỷ lệ khung hình (Aspect Ratio)

Tỷ lệ khung hình được tính theo công thức:

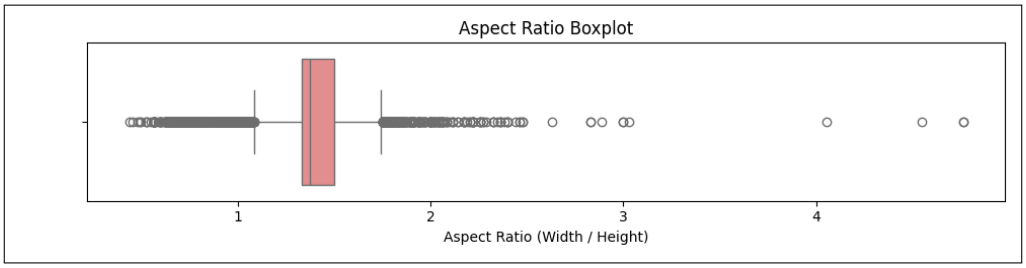
$$\text{Aspect Ratio} = \frac{\text{Width}}{\text{Height}}$$



Hình 1.3: Biểu đồ histogram tỉ lệ khung hình

Kết quả cho thấy:

- Phần lớn ảnh có tỷ lệ khung hình nằm trong khoảng 1.3–1.6, tương ứng với các dạng phổ biến như 4:3 hoặc 3:2.
- Hai đỉnh nổi bật trên biểu đồ histogram cho thấy dữ liệu chứa hai nhóm kích thước chính — một nhóm ảnh gần vuông (tỷ lệ ≈ 1.3) và một nhóm ảnh nằm ngang nhẹ (≈ 1.5).



Hình 1.4: Biểu đồ boxplot tỉ lệ khung hình

Kết quả cho thấy:

- Biểu đồ hộp cho thấy hầu hết dữ liệu tập trung chặt quanh trung vị, chứng tỏ tỷ lệ khung hình giữa các ảnh khá ổn định.
- Tuy nhiên, vẫn tồn tại một số ngoại lệ (outliers) có tỷ lệ cao vượt trội (>3) — đây có thể là các ảnh banner, ảnh toàn cảnh hoặc hình bị kéo dãn.

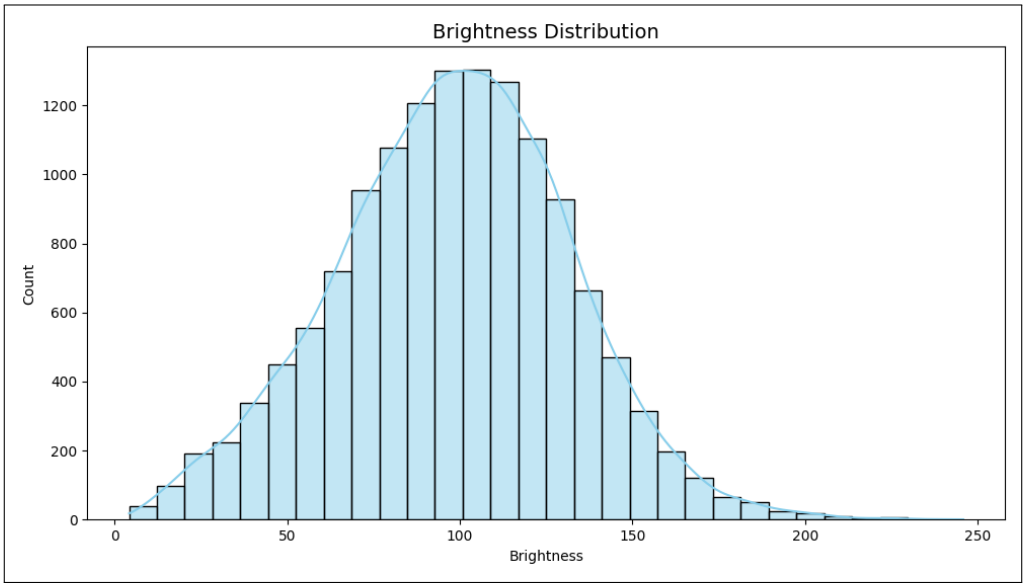
Sự phân bố tương đối ổn định của tỷ lệ khung hình giúp việc chuẩn hóa dữ liệu (resizing và cropping) trong quá trình huấn luyện mô hình học sâu trở nên thuận lợi hơn, vì không cần xử lý quá nhiều trường hợp biến dạng tỉ lệ.

1.5 Phân tích đặc trưng hình ảnh

1.5.1 Độ sáng (Brightness)

Độ sáng của ảnh được đo bằng giá trị trung bình cường độ điểm ảnh (pixel intensity) trên toàn bộ ảnh, với thang đo từ 0 (đen hoàn toàn) đến 255 (trắng hoàn toàn). Độ sáng được tính từ trung bình pixel của ảnh grayscale:

$$\text{Brightness} = \text{mean}(I_{\text{gray}})$$

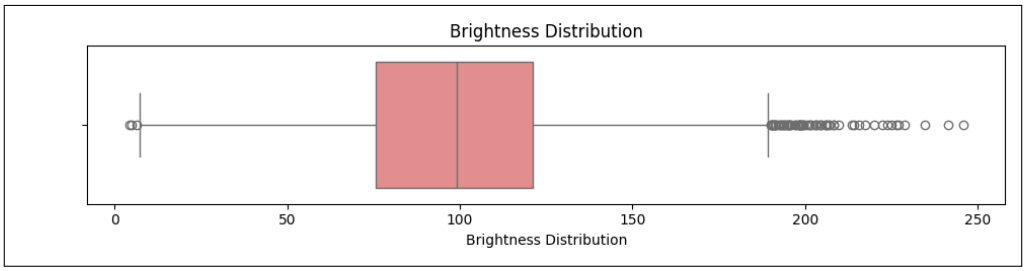


Hình 1.5: Biểu đồ histogram độ sáng

Kết quả cho thấy:

- Phân bố độ sáng có dạng chuông lệch phải (right-skewed), với phần lớn ảnh tập trung quanh giá trị 90–120, cho thấy bộ dữ liệu nhìn chung có độ sáng trung bình và cân bằng.

- Một số lượng nhỏ ảnh có độ sáng thấp (<50), biểu hiện qua phần đuôi trái của biểu đồ, là những ảnh tối hoặc có nền đen.
- Ngược lại, các ảnh có độ sáng >200 rất hiếm, chủ yếu là ảnh chứa vùng sáng mạnh hoặc độ phản chiếu cao.



Hình 1.6: Biểu đồ boxplot độ sáng

Kết quả cho thấy:

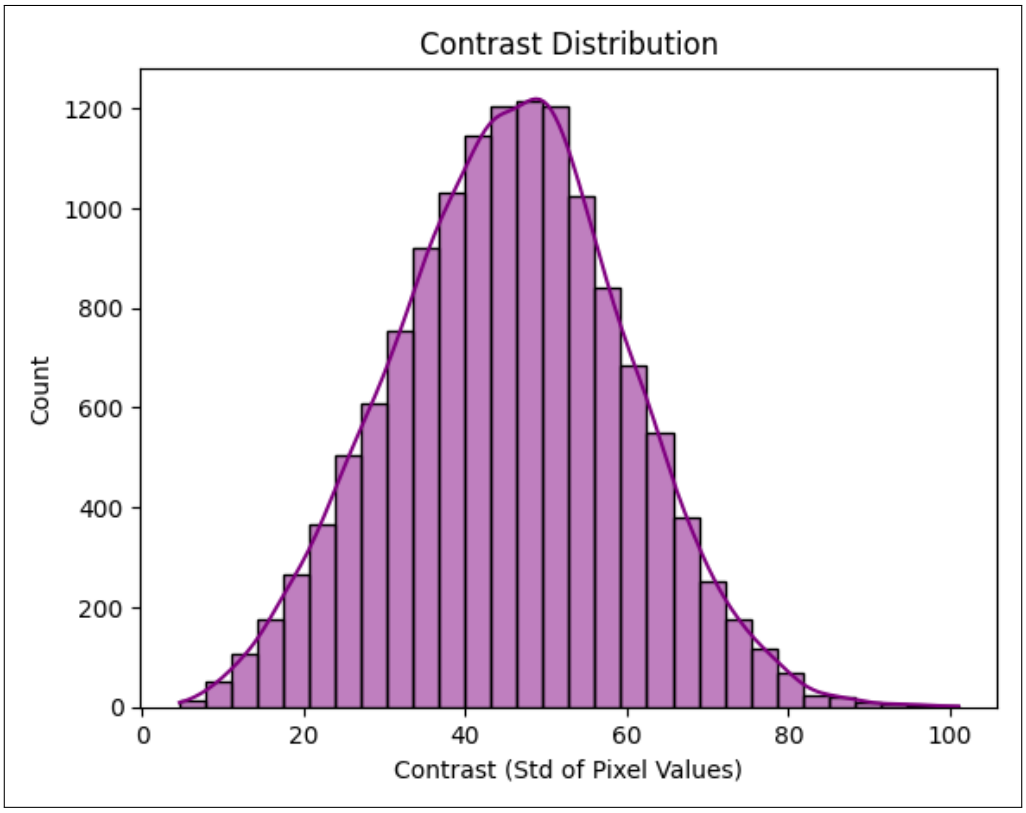
- Biểu đồ hộp (boxplot) cho thấy trung vị nằm gần giá trị trung bình, chứng tỏ độ sáng được phân bố khá ổn định.
- Một vài ngoại lệ xuất hiện ở hai đầu biểu đồ, nhưng không ảnh hưởng nhiều đến toàn bộ phân bố.
- Ngược lại, các ảnh có độ sáng >200 rất hiếm, chủ yếu là ảnh chứa vùng sáng mạnh hoặc độ phản chiếu cao.

Các kết quả cho thấy tập dữ liệu có cân bằng ánh sáng tốt, không cần tăng cường mạnh các kỹ thuật như "brightness augmentation" trong giai đoạn tiền xử lý.

1.5.2 Độ tương phản (Contrast)

Độ tương phản được tính dựa trên độ lệch chuẩn (standard deviation) của giá trị điểm ảnh trong mỗi ảnh. Giá trị độ tương phản cao cho thấy ảnh có sự khác biệt mạnh giữa vùng sáng và tối, trong khi giá trị thấp thể hiện ảnh có màu sắc đồng đều, ít chi tiết nổi bật. Độ tương phản được đo bằng độ lệch chuẩn của pixel ảnh grayscale:

$$\text{Contrast} = \text{std}(I_{\text{gray}})$$



Hình 1.7: Biểu đồ boxplot độ sáng

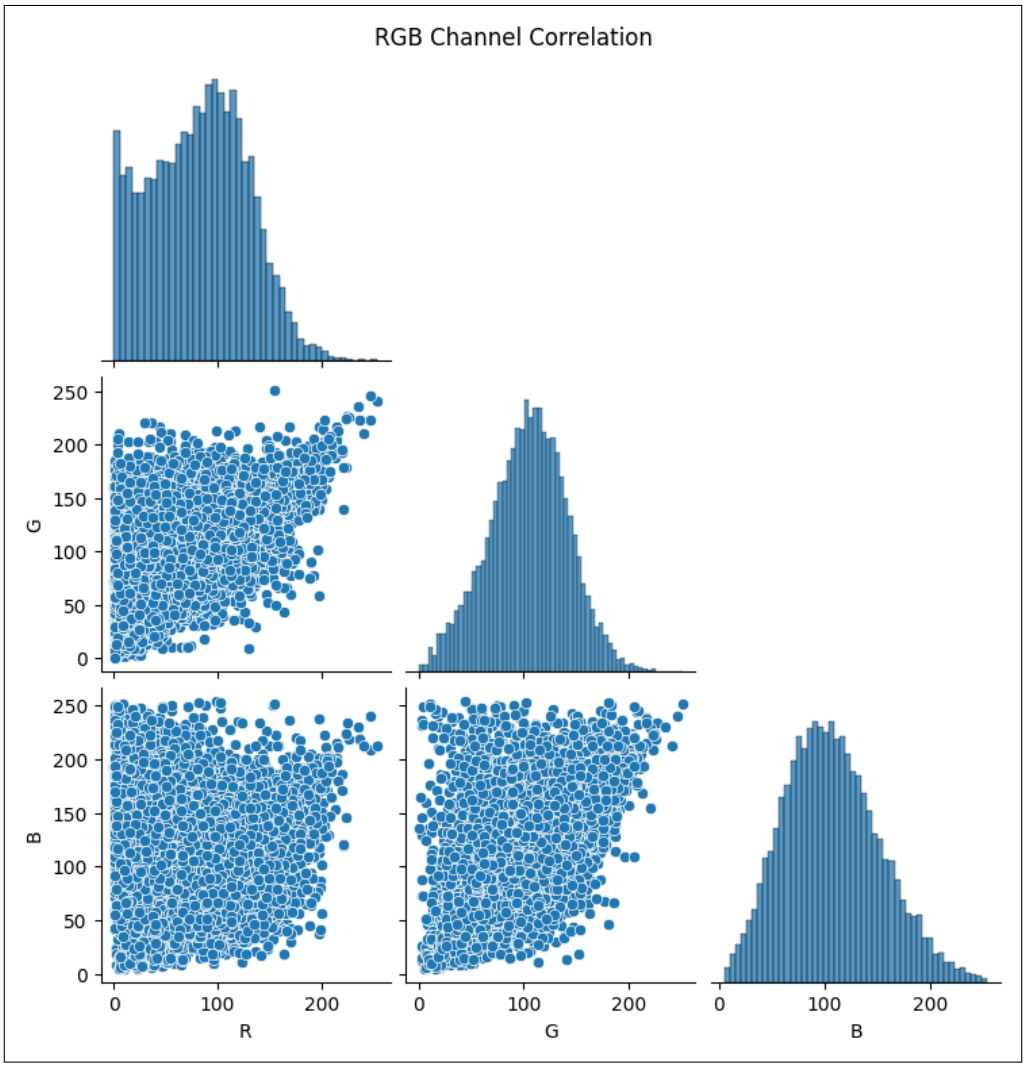
Phân bố có dạng chuông cân đối (gần Gaussian), với trung tâm nằm quanh độ lệch chuẩn 45–50, cho thấy phần lớn ảnh có độ tương phản vừa phải.

Rất ít ảnh có độ tương phản thấp (<20) hoặc quá cao (>80), chứng tỏ tập dữ liệu không chứa nhiều ảnh mờ hoặc bị cháy sáng.

Nhìn chung, mức độ tương phản ổn định giúp các mô hình thị giác máy tính (đặc biệt là CNN và ViT) dễ dàng trích xuất đặc trưng cạnh, vùng sáng–tối mà không cần các bước hiệu chỉnh phức tạp như histogram equalization.

1.5.3 Phân tích kênh màu RGB

Biểu đồ trên thể hiện mối tương quan giữa các kênh màu RGB thông qua cặp đồ thị phân tán (scatter plot) và phân bố tần suất (histogram). Một số nhận xét chính:



Hình 1.8: Biểu đồ tương quan giữa kênh màu RGB

Phân bố cường độ màu:

- Kênh **Red (R)** có xu hướng phân bố khá rộng, tập trung chủ yếu trong khoảng giá trị từ 50 đến 150.
- Kênh **Green (G)** có phân bố lệch phải nhẹ, cho thấy màu xanh lá thường chiếm ưu thế trong hình ảnh sinh vật biển, có thể do ảnh hưởng của ánh sáng môi trường nước biển hoặc đặc điểm sinh học của sinh vật.
- Kênh **Blue (B)** có giá trị trung bình cao hơn hai kênh còn lại, phản ánh rõ đặc trưng môi trường biển với sắc xanh đặc trưng.

Tương quan giữa các kênh:

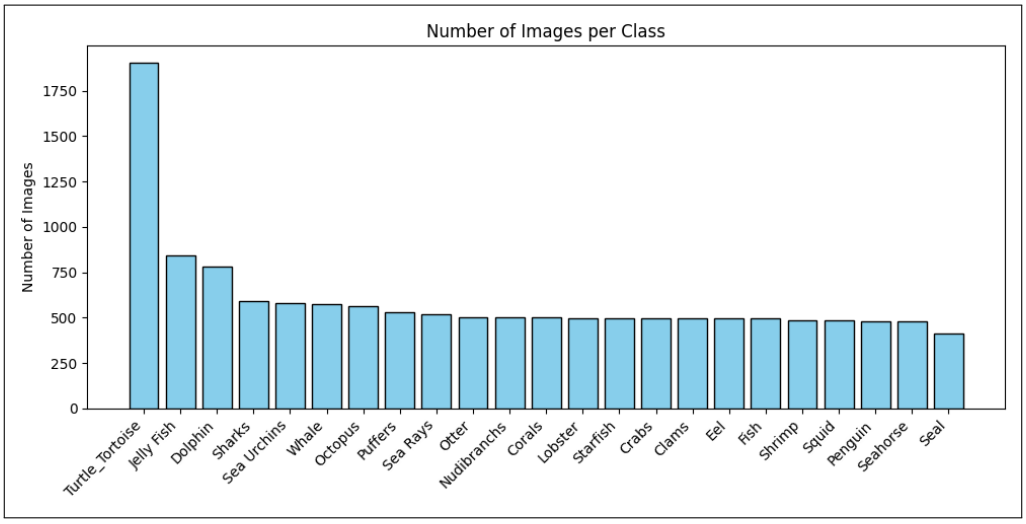
- Các điểm dữ liệu cho thấy có **mối tương quan dương nhẹ** giữa các cặp kênh R–G và G–B, tức là khi cường độ của một kênh tăng thì hai kênh còn lại cũng có xu hướng tăng theo.
- Tuy nhiên, mức độ phân tán vẫn khá lớn, chứng tỏ tập dữ liệu có **độ đa dạng màu sắc cao**, phù hợp với đặc điểm phong phú của các loài sinh vật biển (san hô, cá, rùa, động vật thân mềm, v.v.).

Ý nghĩa đối với mô hình phân loại:

- Việc phân tích kênh màu giúp xác định các đặc trưng trực quan quan trọng của từng lớp sinh vật. Những loài có màu sắc đặc trưng (ví dụ: cá đỏ, san hô tím, rùa xanh, v.v.) có thể được nhận dạng hiệu quả thông qua thông tin RGB.
- Kết quả này gợi ý rằng các mô hình học sâu như *Convolutional Neural Network (CNN)* có thể tận dụng sự khác biệt màu sắc giữa các loài để cải thiện độ chính xác của quá trình phân loại.

1.6 Phân tích theo lớp (Class-level Analysis)

1.6.1 Số lượng ảnh mỗi lớp

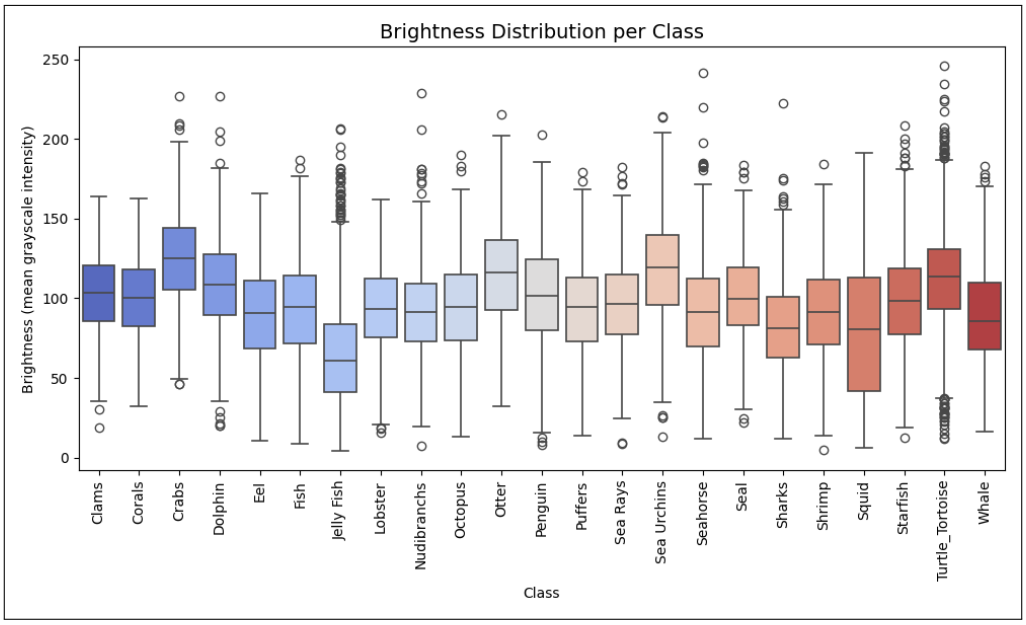


Hình 1.9: Số lượng ảnh trong từng lớp sinh vật biển.

Phân bố số lượng ảnh giữa các lớp trong tập dữ liệu không đồng đều. Cụ thể, một số lớp như *Turtle_ Tortoise* có hơn 1800 ảnh, trong khi các lớp khác chỉ khoảng 400–500 ảnh. Tỷ lệ giữa lớp lớn nhất và nhỏ nhất đạt khoảng **4.6 lần**.

Điều này cho thấy tập dữ liệu **không cân bằng**, vì vậy cần áp dụng các kỹ thuật như *data augmentation* hoặc *class weighting* trong giai đoạn huấn luyện mô hình để tránh hiện tượng lệch dự đoán (bias) về các lớp có nhiều mẫu hơn.

1.6.2 Độ sáng (Brightness) theo lớp

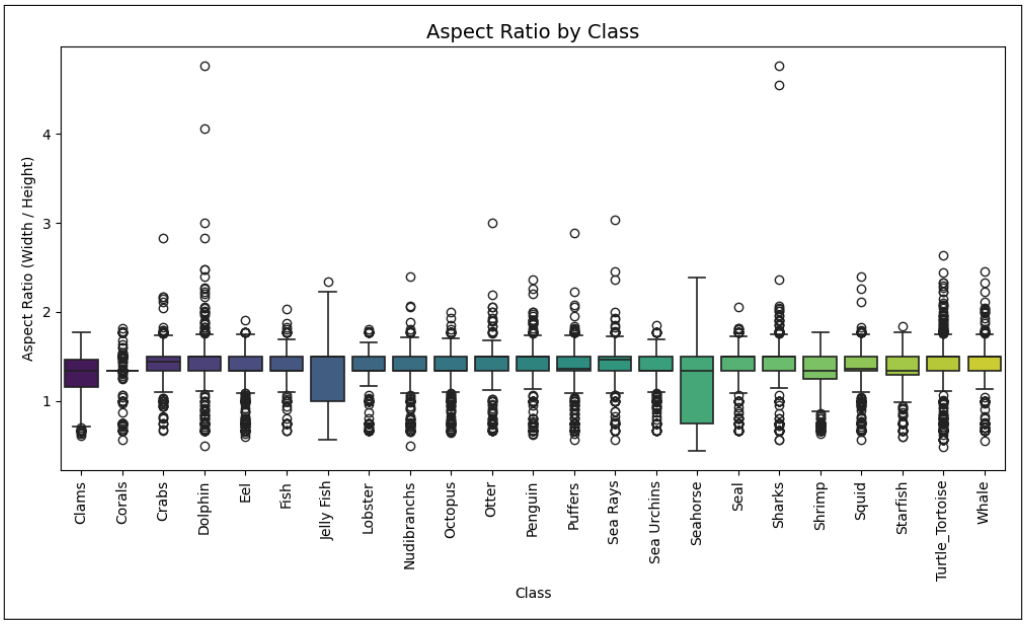


Hình 1.10: Phân bố độ sáng (Brightness) của từng lớp.

Boxplot thể hiện độ sáng trung bình cho thấy sự khác biệt rõ rệt giữa các lớp. Một số lớp như *Fish*, *Turtle_ Tortoise* có độ sáng ổn định, trong khi các lớp như *Jelly Fish*, *Shrimp* có độ biến thiên lớn. Nguyên nhân có thể đến từ điều kiện ánh sáng môi trường khác nhau khi chụp — ví dụ, ảnh chụp gần mặt nước sáng hơn so với ảnh chụp dưới độ sâu lớn.

Điều này cho thấy độ sáng có thể ảnh hưởng đến việc trích xuất đặc trưng và cần được chuẩn hóa (*brightness normalization*) trước khi huấn luyện.

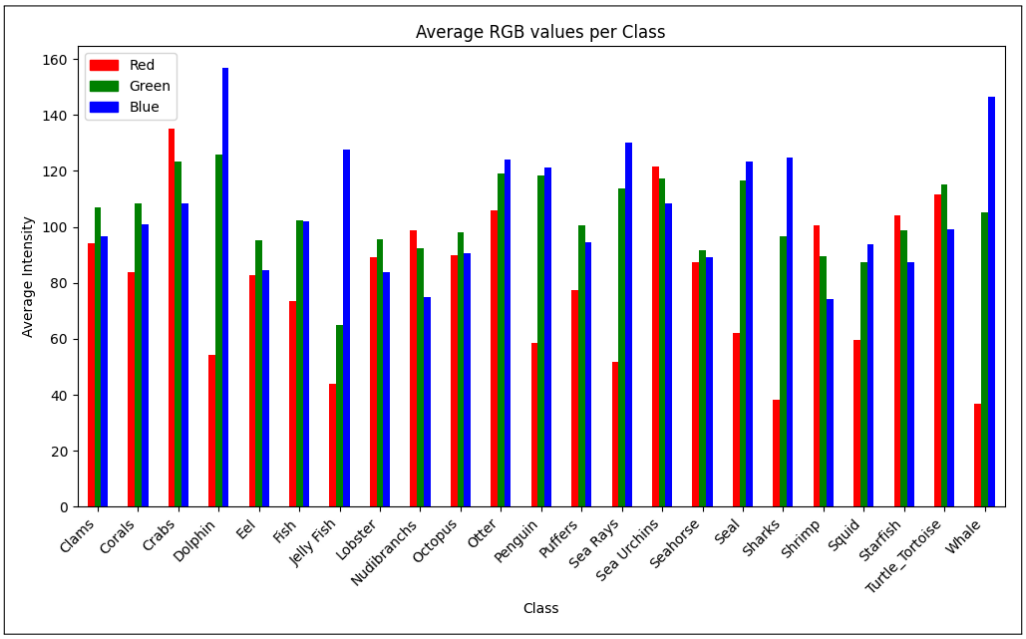
1.6.3 Tỷ lệ khung hình (Aspect Ratio) theo lớp



Hình 1.11: Tỷ lệ khung hình (Aspect Ratio) của từng lớp.

Hầu hết các lớp có tỷ lệ khung hình trung bình xấp xỉ 1.2, cho thấy đa phần hình ảnh có dạng gần vuông. Tuy nhiên, một số lớp như *Dolphin* hoặc *Shark* có tỷ lệ khung hình lớn hơn, phản ánh đặc điểm hình thể dài và thuôn của các loài này. Điều này là đặc trưng quan trọng giúp mô hình học sâu nhận diện được hình dáng tổng thể của từng loài.

1.6.4 Màu sắc trung bình theo lớp



Hình 1.12: Giá trị trung bình của các kênh màu RGB theo từng lớp.

Biểu đồ thể hiện giá trị trung bình của ba kênh màu **R (Red)**, **G (Green)** và **B (Blue)** cho từng lớp sinh vật biển. Một số xu hướng đáng chú ý:

- Các lớp sinh vật sống gần bề mặt hoặc rạn san hô (như *Starfish*, *Turtle_ Tortoise*, *Corals*) có cường độ kênh **Red** cao hơn, thể hiện tông màu ấm (đỏ/nâu).
- Các lớp sinh vật sinh sống ở vùng nước sâu hoặc môi trường mờ (như *Shark*, *Jelly Fish*, *Seal*) có cường độ **Blue** trội hơn, phản ánh sắc xanh đặc trưng của đại dương.

Điều này chứng minh rằng **màu sắc là đặc trưng hữu ích** cho mô hình phân loại sinh vật biển, vì sự khác biệt về môi trường sống và cấu tạo cơ thể được phản ánh rõ qua các kênh màu RGB.

1.7 Kết luận

Qua quá trình phân tích thăm dò dữ liệu (EDA), có thể rút ra một số nhận xét chính như sau:

- **Cấu trúc dữ liệu:** Bộ dữ liệu gồm 13.711 ảnh thuộc nhiều lớp sinh vật biển khác nhau. Tuy nhiên, phân bố số lượng ảnh giữa các lớp không đồng đều, lớp Turtle_Tortoise có số lượng mẫu gấp nhiều lần so với lớp ít nhất.

- **Kích thước ảnh:** Hầu hết ảnh có kích thước trung bình khoảng 300×220 pixel, tuy nhiên vẫn tồn tại một số ảnh có kích thước rất lớn (trên 4000 pixel), cần được chuẩn hóa để đảm bảo tốc độ và tính ổn định khi huấn luyện.
- **Tỷ lệ khung hình:** Phần lớn ảnh có tỷ lệ khung hình 4:3 hoặc 3:2, thuận lợi cho việc chuẩn hóa kích thước mà không làm méo hình.
- **Đặc trưng màu sắc (RGB):** Kênh màu **Blue** chiếm ưu thế rõ rệt trong hầu hết các ảnh — phản ánh đặc trưng của môi trường biển. Ba kênh RGB có tương quan nhất định, cho thấy thông tin màu là đặc trưng hữu ích cho phân loại sinh vật biển.

Tổng thể, bộ dữ liệu có chất lượng hình ảnh tương đối tốt, tuy nhiên cần thực hiện một số bước xử lý trước khi đưa vào mô hình học sâu để đảm bảo tính đồng nhất và tối ưu hiệu quả huấn luyện.

2 Tiền xử lý dữ liệu

Dựa trên kết quả EDA, quá trình tiền xử lý dữ liệu được thiết kế và thực hiện theo các bước sau.

2.1 Chia tập dữ liệu (Dataset Splitting)

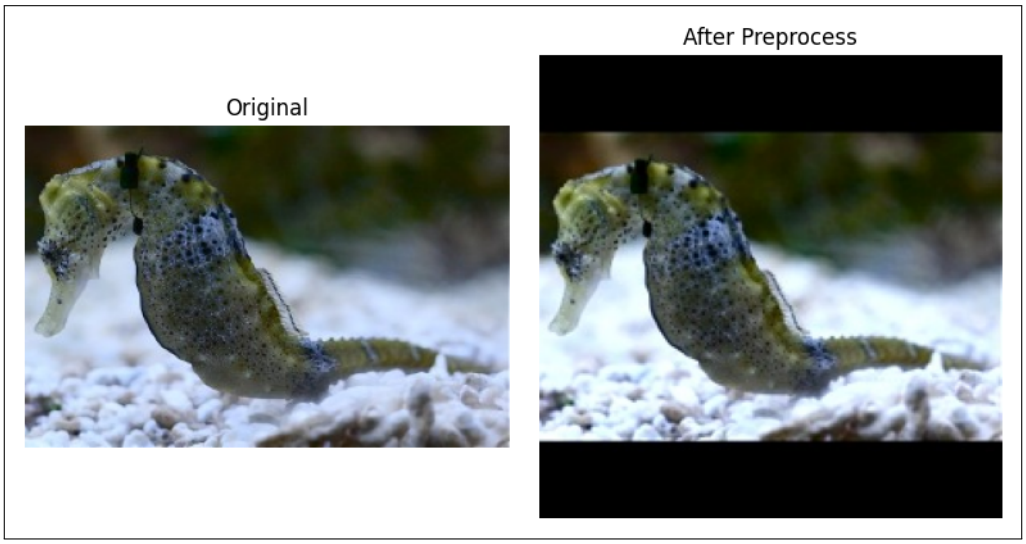
Bộ dữ liệu ban đầu được chia thành hai phần:

- **Tập huấn luyện (Train set):** 10,968 (80% tổng số ảnh)
- **Tập kiểm thử (Test set):** 2,743 (20% còn lại).

Quá trình chia dữ liệu được thực hiện bằng phương pháp *stratified split* — đảm bảo mỗi lớp sinh vật biển đều được phân bố đồng đều trong cả hai tập, giúp đánh giá mô hình khách quan hơn.

2.2 Chuẩn hóa kích thước và tỷ lệ khung hình

Để đảm bảo tính nhất quán cho đầu vào của mô hình, toàn bộ ảnh được chuẩn hóa về cùng kích thước cố định 224×224 pixel. Đối với các ảnh không vuông, kỹ thuật *padding* được áp dụng để thêm viền đen sao cho ảnh trở thành hình vuông trước khi *resize*. Phương pháp này giúp tránh làm méo hình hoặc biến dạng đối tượng, đồng thời phù hợp với đầu vào của các mô hình học sâu phổ biến như CNN và Vision Transformer.



Hình 2.1: Minh họa khung hình sau khi chuẩn hóa

2.3 Tăng cường chất lượng và tính đa dạng của dữ liệu

2.3.1 Tăng cường độ sáng và tương phản

Sau khi chuẩn hóa kích thước, ảnh được tăng cường nhẹ độ sáng và độ tương phản (tăng khoảng 10%). Mục đích của bước này là:

- Cải thiện khả năng nhận diện chi tiết của đối tượng.
- Giảm ảnh hưởng của sự khác biệt về điều kiện chiếu sáng giữa các ảnh.

Nhờ đó, dữ liệu đầu vào trở nên rõ nét và đồng nhất hơn, giúp mô hình học đặc trưng ổn định hơn.

2.3.2 Tăng cường dữ liệu (Data Augmentation)

Trong giai đoạn huấn luyện, quá trình tăng cường dữ liệu được thực hiện tự động bằng các phép biến đổi ngẫu nhiên để mô phỏng sự đa dạng tự nhiên của hình ảnh. Các phép biến đổi bao gồm:

- Lật ngang (*RandomHorizontalFlip*)
- Xoay ngẫu nhiên trong khoảng $\pm 15°$
- Cắt và phóng ngẫu nhiên (*RandomResizedCrop*)
- Thay đổi nhẹ độ sáng, tương phản và bão hòa (*ColorJitter*)

Những kỹ thuật này giúp mô hình học được các đặc trưng ổn định hơn, giảm hiện tượng *overfitting* và tăng khả năng khái quát hóa (*generalization*) trên dữ liệu thực tế.

2.3.3 Chuẩn hóa giá trị pixel (Normalization)

Sau khi ảnh được chuyển đổi thành tensor, quá trình chuẩn hóa pixel được áp dụng theo thống kê của bộ dữ liệu ImageNet — tiêu chuẩn phổ biến trong các mô hình học sâu đã được huấn luyện sẵn. Cụ thể:

$$\text{mean} = [0.485, 0.456, 0.406], \quad \text{std} = [0.229, 0.224, 0.225]$$

Chuẩn hóa này giúp mô hình hội tụ nhanh hơn và đạt hiệu quả cao hơn khi sử dụng trọng số khởi tạo từ các mô hình *pretrained*.

Hình 2.2: Minh họa 8 hình sau khi tiền xử lý

2.4 Kết luận

Quy trình tiền xử lý đã tạo ra một bộ dữ liệu ảnh đồng nhất, có kích thước chuẩn và được tăng cường hợp lý. Bộ dữ liệu sau xử lý có chất lượng cao hơn, giảm nhiễu giúp cho quá trình huấn luyện mô hình phân loại sinh vật biển bằng học sâu tốt hơn.

3 Trích xuất đặc trưng ảnh

3.1 Tổng quan phương pháp

Ta triển khai cả hai hướng trích xuất đặc trưng:

- **Phương pháp dựa trên CNN (Convolutional Neural Networks)**: Sử dụng các mô hình CNN đã được huấn luyện trước trên ImageNet như ResNet50 và EfficientNetB0.
- **Phương pháp dựa trên Transformer cho hình ảnh**: Sử dụng các mô hình Vision Transformer (ViT), với khả năng học biểu diễn toàn cục từ ảnh.

Việc lựa chọn hai hướng này nhằm so sánh khả năng biểu diễn đặc trưng cục bộ, dựa trên bộ lọc CNN với biểu diễn toàn cục, attention của Transformer, từ đó đánh giá hiệu quả các loại đặc trưng cho tác vụ phân loại sinh vật biển.

3.2 Trích xuất đặc trưng bằng CNN

Nhóm sử dụng hai mô hình CNN phổ biến: ResNet50 và EfficientNetB0.

Ưu điểm

- CNN có khả năng học các đặc trưng cục bộ như cạnh, kết cấu, hình dạng vật thể.
- Các mô hình pretrained trên ImageNet đã học được nhiều đặc trưng phổ quát, giúp giảm thời gian huấn luyện và cải thiện hiệu suất trên tập dữ liệu nhỏ.

Nhược điểm

- Tập trung nhiều vào các đặc trưng cục bộ, có thể bỏ sót các mối quan hệ toàn cục trong ảnh.
- Mô hình có thể nhạy cảm với các biến đổi không gian hoặc điều kiện ánh sáng khác so với dữ liệu huấn luyện gốc.

3.3 Trích xuất đặc trưng bằng Transformer

Nhóm sử dụng Vision Transformer (ViT) để trích xuất đặc trưng toàn cục từ ảnh.

Ưu điểm

- Transformer học biểu diễn toàn cục, chú ý đến mối quan hệ giữa các vùng khác nhau trong ảnh.
- Thích hợp để nắm bắt các đặc trưng phức tạp, có thể vượt trội khi vật thể có cấu trúc lớn hoặc nhiều lớp chi tiết.

Nhược điểm

- Cần nhiều dữ liệu và tài nguyên tính toán hơn so với CNN.
- Trên các tập dữ liệu nhỏ hoặc khi fine-tuning không cẩn thận, Transformer có thể không học được các đặc trưng hữu ích.

3.4 Lưu trữ và quản lý đặc trưng

Tất cả các feature vector đều được lưu theo định dạng HDF5 (.h5), mỗi mô hình tạo một file riêng. Việc này đảm bảo:

- Dễ dàng tái sử dụng các đặc trưng trong các bước huấn luyện mô hình phân loại phía sau.
- Tiết kiệm thời gian, tránh phải chạy lại quá trình trích xuất nhiều lần.
- Cho phép so sánh trực tiếp hiệu quả các loại feature từ CNN và Transformer.

4 Huấn luyện mô hình phân loại

Ta tiến hành huấn luyện các mô hình phân loại truyền thống dựa trên các đặc trưng (feature vectors) đã được trích xuất từ các mô hình học sâu ở bước trước.

4.1 Chuẩn bị mô hình phân loại

Ba mô hình được lựa chọn bao gồm:

- **Logistic Regression**: sử dụng solver `liblinear`, hệ số regularization $C = 0.1$, và cân bằng nhãn bằng `class_weight='balanced'`.
- **Linear SVM**: sử dụng kernel tuyến tính, bật chế độ dự đoán xác suất (`probability=True`), và cân bằng nhãn.
- **Random Forest**: gồm 100 cây, độ sâu tối đa 10, và sử dụng `class_weight='balanced'` để điều chỉnh dữ liệu mất cân bằng.

Các mô hình được cài đặt bằng thư viện `scikit-learn` (sklearn), hỗ trợ trực tiếp bài toán phân loại đa lớp. Dữ liệu đặc trưng được lấy từ các file `.h5` được sinh ra trong bước trích xuất đặc trưng (từ ResNet50, EfficientNetB0 và ViT).

4.2 Quy trình huấn luyện và đánh giá

- Với mỗi loại đặc trưng (`resnet50`, `efficientnetb0`, `vit-base-patch16-224`), mô hình được huấn luyện trên tập `train` và đánh giá trên tập `test`.
- Các chỉ số đánh giá bao gồm:
  - **Accuracy**: độ chính xác tổng thể.
  - **Precision**, **Recall**, **F1-score**: tính trung bình có trọng số (`weighted`) cho từng lớp.
- Ngoài ra, **ma trận nhầm lẫn (Confusion Matrix)** được sử dụng để trực quan hóa kết quả dự đoán đúng/sai giữa các lớp.

Việc huấn luyện và đánh giá được thực hiện thống nhất trên cùng một tập dữ liệu kiểm thử nhằm đảm bảo công bằng khi so sánh giữa các mô hình.

5 So sánh và đánh giá kết quả

5.1 Bảng kết quả tổng hợp

Bảng sau trình bày kết quả của ba mô hình Logistic Regression, Linear SVM và Random Forest khi huấn luyện trên ba bộ đặc trưng được trích xuất từ các kiến trúc khác nhau (ResNet50, EfficientNetB0, ViT-Base).

Feature	Model	Accuracy	Precision	Recall	F1-score
ResNet50	Logistic Regression	0.818	0.818	0.818	0.818
ResNet50	Linear SVM	0.793	0.795	0.793	0.792
ResNet50	Random Forest	0.724	0.731	0.724	0.716
EfficientNetB0	Logistic Regression	**0.845**	**0.846**	**0.845**	**0.844**
EfficientNetB0	Linear SVM	0.825	0.826	0.825	0.824
EfficientNetB0	Random Forest	0.763	0.766	0.763	0.760
ViT-Base	Logistic Regression	0.031	0.028	0.031	0.026
ViT-Base	Linear SVM	0.087	0.062	0.087	0.036
ViT-Base	Random Forest	0.138	0.032	0.138	0.040

Bảng 3: Kết quả huấn luyện và đánh giá các mô hình trên các tập đặc trưng khác nhau

5.2 Phân tích kết quả

5.2.1 Hiệu quả theo loại đặc trưng

Các đặc trưng trích xuất từ **EfficientNetB0** cho kết quả vượt trội so với ResNet50 và ViT. Cụ thể, Logistic Regression kết hợp với EfficientNetB0 đạt độ chính xác cao nhất **84.5%**, trong khi ResNet50 đạt khoảng 81.8%, còn ViT chỉ đạt hiệu suất rất thấp (khoảng 3–13%).

Điều này cho thấy đặc trưng của ViT khi trích xuất trực tiếp từ mô hình pretrained chưa được fine-tune có thể chưa phù hợp với dữ liệu ảnh sinh vật biển, do sự khác biệt về miền dữ liệu (domain gap) so với tập huấn luyện gốc (ImageNet).

5.2.2 So sánh theo thuật toán phân loại

Trong cả hai bộ đặc trưng từ CNN (ResNet và EfficientNet), **Logistic Regression** consistently outperform các mô hình khác, đạt F1-score cao hơn SVM và Random Forest từ 2–8%. Điều này có thể giải thích bởi Logistic Regression có khả năng tổng quát hóa tốt trên feature space đã được phân tách rõ ràng bởi các mô hình CNN pretrained.

5.2.3 Độ chênh lệch giữa Precision, Recall và F1-score

Các chỉ số Precision, Recall và F1-score gần tương đồng, chứng tỏ mô hình không bị lệch mạnh giữa việc dự đoán thiếu và dư mẫu. Đặc biệt, EfficientNetB0 + Logistic Regression đạt độ cân bằng rất tốt giữa các chỉ số, chứng minh đây là lựa chọn ổn định cho bài toán này.

5.3 Tổng hợp mô hình tốt nhất

Bảng sau tóm tắt các mô hình đạt kết quả cao nhất theo từng tiêu chí.

Metric	**Best Model**	**Feature**	**Best Score**
Accuracy	Logistic Regression	EfficientNetB0	0.845
Precision	Logistic Regression	EfficientNetB0	0.846
Recall	Logistic Regression	EfficientNetB0	0.845
F1-score	Logistic Regression	EfficientNetB0	0.844

Bảng 4: Tổng hợp mô hình và đặc trưng có kết quả cao nhất

5.4 Kết luận

Kết quả thực nghiệm cho thấy việc kết hợp giữa **đặc trưng CNN (EfficientNetB0)** và **mô hình Logistic Regression** mang lại hiệu quả cao nhất trong bài toán phân loại ảnh sinh vật biển. Mô hình này đạt độ chính xác **84.5%** và duy trì độ ổn định tốt giữa các chỉ số Precision, Recall, F1-score.

Trong khi đó, đặc trưng từ ViT chưa thể hiện tốt do thiếu bước fine-tuning phù hợp với miền dữ liệu cụ thể. Do đó, hướng phát triển tiếp theo có thể là **fine-tune mô hình Transformer** trên tập ảnh này hoặc kết hợp đặc trưng CNN và Transformer để tận dụng ưu điểm của cả hai hướng trích xuất đặc trưng.

6 Kết luận

Trong bài báo cáo này, nhóm đã xây dựng một quy trình (pipeline) hoàn chỉnh cho bài toán phân loại ảnh động vật biển, bắt đầu từ việc khảo sát dữ liệu (EDA), tiền xử lý, tăng cường dữ liệu, trích xuất đặc trưng bằng các mô hình học sâu hiện đại, cho đến huấn luyện và đánh giá các mô hình phân loại truyền thống.

Kết quả phân tích EDA cho thấy tập dữ liệu *Sea Animals Image Dataset* có sự đa dạng về kích thước, tỷ lệ khung hình và độ sáng, song vẫn đảm bảo chất lượng hình ảnh tốt cho quá trình huấn luyện. Việc chuẩn hóa và tăng cường dữ liệu giúp mô hình thích nghi tốt hơn với các biến đổi tự nhiên của hình ảnh trong thực tế.

Ở giai đoạn trích xuất đặc trưng, nhóm đã khai thác sức mạnh của các mô hình học sâu như **ResNet50**, **EfficientNetB0** và **Vision Transformer (ViT)**. Kết quả cho thấy các đặc trưng trích xuất từ mạng CNN (đặc biệt là EfficientNetB0) mang lại hiệu suất cao hơn rõ rệt so với mô hình Transformer trong bối cảnh dữ liệu có kích thước vừa phải và phân bố không đồng đều.

Khi kết hợp với các mô hình phân loại truyền thống, **Logistic Regression** đạt hiệu năng tốt nhất trên tập đặc trưng của EfficientNetB0 với độ chính xác (*Accuracy*) đạt khoảng **84.5%**, cùng các chỉ số Precision, Recall và F1-score đồng đều, chứng tỏ mô hình học được đặc trưng tổng quát và ổn định. Trong khi đó, Random Forest và SVM cho kết quả thấp hơn, chủ yếu do giới hạn trong khả năng phân tách biên quyết định phi tuyến của dữ liệu trích xuất.

7 Phụ lục

Dataset: Sea Animals Image Dataset
Github: DNA05's github page
Colab Notebook: DNA05-BTL3

Tài liệu tham khảo

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R and Python*. Springer, 2nd edition, 2021.

[2] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[3] Scikit-learn developers. Classification metrics — scikit-learn documentation. https://scikit-learn.org/stable/modules/model_evaluation.html, 2025. Accessed: 2025-10-25.

[4] Scikit-learn developers. Preprocessing data — scikit-learn documentation. https://scikit-learn.org/stable/modules/preprocessing.html, 2025. Accessed: 2025-10-25.