Towards the automatic classification of unlabeled insect images: a clustering approach based on visual embeddings using Dytiscinae beetles as a model

Patricia L. M. Torres[1,2], Matías R. Urcola[1,2], Luciano M. Barbalace[1], Efraín J. Calla Yana[3], María E. Buemi[4], Diego A. Rodríguez[5] y Cristian A. Martínez[5*]

[1]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento Biodiversidad y Biología Experimental, Laboratorio de Entomología, Buenos Aires, Argentina
[2]CONICET-UBA, Instituto de Biodiversidad y Biología Experimental y Aplicada (IBBEA), Buenos Aires, Argentina
[3]Universidad Nacional del Altiplano, Av. Floral 1153, Puno, Perú
[4]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Computación, Grupo de Procesamiento de Imágenes y Visión por Computadora
[5]Universidad Nacional de Salta, Facultad de Ciencias Exactas, Departamento de Informática, Salta, Argentina
*cmartinez@di.unsa.edu.ar

Abstract. Taxonomic identification of Dytiscidae beetles is challenged by intraspecific variability and morphological overlap among species. In this study, unsupervised learning approaches are evaluated to analyze visual similarity patterns from images. Two models are compared: (i) self-supervised deep embeddings reduced via UMAP and (ii) dimensionality reduction using PCA on raw pixel data. The results show that deep embeddings recover coherent structures in the feature space. In contrast, the PCA-based model produces a highly fragmented distribution with weak species–cluster correspondence. These findings highlight that representation quality is a critical factor for successful clustering in biological contexts.

Keywords: Unsupervised learning, Taxonomy, Dytiscidae, Clustering, DINOv2

Hacia la clasificación automática de imágenes de insectos sin etiquetado: un enfoque de agrupamiento basado en representaciones visuales usando escarabajos de la subfamilia Dytiscinae como modelo

Resumen. La identificación taxonómica de escarabajos de la familia Dytiscidae se dificulta por la variabilidad intraespecífica y el solapamiento morfológico entre especies. En este trabajo, se evalúan enfoques de aprendizaje no supervisado para analizar patrones de similitud visual a partir de imágenes. Para ello, se comparan dos modelos: (i) representaciones vectoriales (embeddings) de aprendizaje profundo autosupervisado reducidas mediante UMAP y (ii) reducción de dimensión mediante PCA aplicado directamente a píxeles. Los resultados muestran que las representaciones de aprendizaje profundo recuperan estructuras coherentes en el espacio de características. Por el contrario, el modelo basado en PCA genera una distribución altamente fragmentada con una escasa correspondencia entre especies y agrupamientos (clusters). Estos hallazgos remarcan que la calidad de la representación es un factor crítico para alcanzar un clustering de calidad en contextos biológicos.

Palabras clave: Aprendizaje no supervisado, Taxonomía, Dytiscidae, Clustering, DINOv2

1. Introduction

Coleopteran taxonomy relies on the identification of morphological characters, a process that requires high levels of expertise and presents limitations in scalability and reproducibility (Valan et al., 2019). Identification at the genus and species level is generally restricted to specialist taxonomists, who assess the validity of species on the basis of fine morphological characters. Although these traits are largely discrete and stable for each entity, they may show overlap between taxa as well as intraspecific variability. In most cases, identifying an insect remains a manual task: it requires a specimen and the use of optical and/or stereoscopic microscopy, and it proceeds through dichotomous keys that guide the taxonomist along a series of binary morphological decisions. The specialist must examine diagnostic characters such as wing venation, chaetotaxy, coloration patterns, the structure of the mouthparts, or the morphology of the genitalia, among others. This traditional approach is time-consuming per specimen, demands years of specialized training, and may vary according to the criterion of each specialist, which reduces the scalability of identification when large volumes of material must be processed or when trained professionals are scarce (Wägele et al., 2011). These challenges are particularly pronounced in groups such as Dytiscidae, where multiple species exhibit substantial phenotypic similarity (Holmgren et al., 2016; Yee, 2023).

Unsupervised learning (Zhang & Lu, 2021) enables the analysis of latent data structures without labeled annotations, offering a promising alternative for exploring morphological patterns. However, its effectiveness strongly depends on the quality of image representations. Because it does not depend on costly prior labels, unsupervised learning can act as a support tool for the taxonomist, automatically organizing specimens into groups on the basis of visual similarity. The expert can then use these groupings as an exploratory step to rapidly identify closely related specimens, detect outliers, or generate new hypotheses about species boundaries in groups where morphological divergence is minimal. Most studies reported in the literature focus on the development of machine learning and deep learning models for insect identification from images (Teixeira et al., 2023; Gao et al., 2024). However, these approaches rely on datasets that must be annotated in advance by domain experts. Recently, self-supervised learning models have been developed for automatic insect identification (Kar et al., 2023; Charoenpanyakul et al., 2024). Unlike traditional supervised approaches, self-supervised learning acquires visual representations directly from unlabeled images, substantially reducing the need for expert annotation. Despite these advances, there is still a lack of methods capable of automatically grouping morphologically similar insects to assist taxonomists in the identification of species and the initial organization of biological collections. Unsupervised approaches have been applied to related problems, but the way that morphological information is obtained differs fundamentally from the one explored here. In Coleoptera, clustering has been used on manually measured morphometric variables (Bin Azis & Sharif, 2019), while in other taxa unsupervised learning has been combined with geometric morphometrics based on expert-defined landmarks (Bellin et al., 2022) or with genetic data for species delimitation (Derkarabetian et al., 2019). To the best of our knowledge, however, the use of self-supervised deep learning embeddings extracted directly from dorsal habitus images of beetles (requiring neither manual measurement nor landmark digitization) remains unexplored for unsupervised morphological clustering in this group.

This study evaluates the ability of two unsupervised approaches to recover the taxonomic structure from beetle images. Specifically, we compare classical dimensionality reduction and clustering methods with a self-supervised representation of dorsal habitus images of ten Dytiscinae species, assessing which patterns emerge

from each approach, the degree of agreement with the prior taxonomic classification, and the type of variation—biological or technical—that may underlie the observed groupings.

2. Materials and Methods

2.1 Dataset

The dataset consists of dorsal-view images of male beetles from three genera belonging to the family Dytiscidae, subfamily Dytiscinae (*Hydaticus*, *Notaticus*, and *Thermonectus*) including multiple species per genus with varying sample sizes. All specimens were taxonomically identified by P.L.M. Torres and M. Michat (University of Buenos Aires), both experts in Dytiscidae systematics. Identifications were based on the examination of external morphological characters and male genitalia using a LEICA MZ6 stereomicroscope, following standard entomological techniques, including genitalia dissection, and consulting specialized literature such as revisionary works, dichotomous keys, and original descriptions. Images were captured using a LEICA MZ6 stereomicroscope (6.3× magnification) with ring LED illumination and a LEICA DMC2900 digital camera. All images were acquired at a resolution of 2000 × 2000 pixels. Only male specimens were used to avoid sexual dimorphism effects on elytra and pronotum texture. Fig. 1 presents representative examples, illustrating interspecific similarity.

2.2 Image Preprocessing

During preprocessing, the original 2000 × 2000-pixel images were resized to 256 × 256 pixels, followed by a center crop to 224 × 224 pixels. The cropped images were then converted into tensors by first scaling pixel intensity values from the [0, 255] range to [0, 1], and subsequently applying channel-wise normalization using the ImageNet mean and standard deviation for each RGB channel. This preprocessing pipeline produced standardized input images fully compatible with the feature extraction models employed in this study.

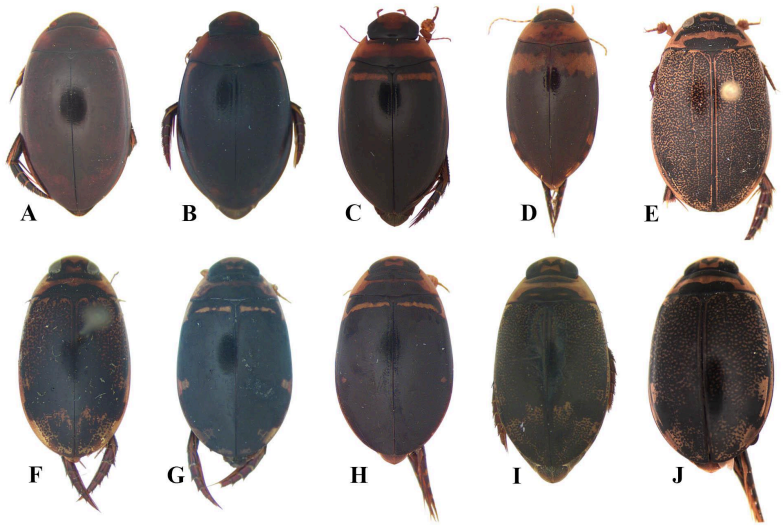

Fig. 1. Dorsal view of the 10 species of beetles of the Dytiscinae subfamily that comprises the analyzed dataset. A: *Hydaticus palliatus*, B: *Hydaticus tuyuensis*, C: *Hydaticus xanthomelas*, D: *Notaticus fasciatus*, E: *Thermonectus alfredi*, F: *Thermonectus circumscriptus*, G: *Thermonectus margineguttatus*, H: *Thermonectus nobilis*, I: *Thermonectus succinctus*, J: *Thermonectus tremouillesi*.

2.3 Feature Extraction

Two approaches were evaluated:

- **Model 1 (DINOv2 + UMAP)**: Feature extraction was performed using DINOv2 ViT-B/14 (Oquab et al., 2023), a self-supervised pretrained Vision Transformer consisting of 12 Transformer blocks, 12 attention heads, a 14 × 14-pixel patch size, and an embedding dimension of 768 features. During inference, the normalized CLS token from the final layer (*x_norm_clstoken*) was used as the global representation of each image. To improve robustness against small geometric variations, embeddings were computed for three versions of each image: the original image, its horizontally flipped counterpart, and a center crop corresponding to 90% of the original image size, which was subsequently resized to 224 × 224 pixels. The final embedding for each sample was obtained by averaging these three representations, followed by L2 normalization. Finally, the resulting 768-dimensional embeddings were projected into a lower-dimensional latent space using Uniform Manifold Approximation and Projection (UMAP) (Allaoui et al., 2020) with 20 components, 20 nearest neighbors (*n_neighbors = 20*), cosine distance as the similarity metric (*metric = cosine*), and min_dist = 0.0. The resulting 20-dimensional latent representation was subsequently used as input for the clustering algorithms.
- **Model 2 (PCA on Raw Images)**: Each preprocessed image of size 224 × 224 × 3 was flattened into a one-dimensional feature vector, yielding a feature matrix of size N × 150,528, where N denotes the total number of images in the dataset. Before applying Principal Component Analysis (PCA) (Marukatat, 2023), the features were normalized using Min–Max scaling. The first 100 principal components were retained, explaining a cumulative variance of 0.9114. This dimensionality reduction provided a compact representation of the image information for subsequent clustering.

2.4 Unsupervised Clustering

The latent representations obtained from either **DINOv2 + UMAP** or **PCA** were used as input for four clustering algorithms to determine which method most accurately captured the intrinsic structure of the data.

K-Means

The K-Means algorithm (Ghosal et al., 2020; Singh & Srivastava, 2020) was evaluated. This method iteratively partitions the feature space by minimizing the within-cluster sum of squared distances between samples and their corresponding cluster centroids. To reduce sensitivity to random centroid initialization, 20 independent initializations (*n_init = 20*) were performed, retaining the solution with the lowest inertia. The algorithm was executed with random_state = 0 to ensure reproducibility.

Agglomerative Clustering

Agglomerative Clustering (Ghosal et al., 2020; Singh & Srivastava, 2020), a bottom-up hierarchical clustering method, was also evaluated. The algorithm starts by considering each sample as an individual cluster and iteratively merges the most similar cluster pairs until the desired number of clusters is reached. Inter-cluster similarity was computed using cosine distance (*metric = cosine*), while the average linkage criterion was employed, defining the distance between two clusters as the average pairwise distance between all samples belonging to each cluster.

Spectral Clustering
Spectral Clustering (Ghosal et al., 2020) was also considered. This algorithm constructs a similarity graph from the input samples and performs a spectral decomposition of the graph Laplacian to project the data into a lower-dimensional space where clusters become more easily separable. The similarity graph was built using the 20 nearest neighbors (*n_neighbors = 20*), and random_state = 0 was used to ensure reproducible clustering results.

BIRCH
The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm (Singh & Srivastava, 2020; Chaudhry et al., 2023) was also included. BIRCH is designed to efficiently summarize large datasets through a hierarchical structure known as a Clustering Feature Tree (CF-tree). In this study, a *threshold = 0.5* was used to control the maximum radius of the subclusters generated during tree construction, after which the final number of clusters was specified.

2.5 Experimental Configuration
Since the clustering algorithms considered require a predefined number of clusters, the following values of K were evaluated:

$$k \in \{3,5,7,9,10,11\}$$

These configurations included values below, equal to, and above the expected number of species in the dataset, allowing the sensitivity of each algorithm to the number of partitions to be systematically assessed.

2.6 Clustering Evaluation
Each clustering configuration was evaluated using a combination of internal and external validation metrics.
The internal metrics (Chaudhry et al., 2023) assessed the structural quality of the clustering based exclusively on the distribution of samples in the feature space:

- Silhouette Coefficient, which simultaneously measures intra-cluster cohesion and inter-cluster separation; higher values indicate better-defined clusters.
- Calinski–Harabasz Index, which quantifies the ratio between between-cluster dispersion and within-cluster dispersion; higher values indicate better cluster separation.
- Davies–Bouldin Index, which measures the average similarity between clusters; lower values indicate more compact and better-separated clusters.

To enable an objective comparison of the clustering algorithms within an experimental framework and to provide a robust criterion for selecting the best clustering solution, two external validation metrics were also computed using the ground-truth taxonomic labels available in the dataset. These labels were not used during the clustering process and were employed exclusively for the evaluation and comparison of the resulting clustering solutions. The external metrics (Arinik et al., 2021) considered were:

- Adjusted Rand Index (ARI), which measures the agreement between the obtained partition and the reference taxonomic classification while correcting for chance agreement.
- Normalized Mutual Information (NMI), an information-theoretic metric that quantifies the amount of shared information between the predicted clusters and the ground-truth labels.

2.7 Selection of the Best Clustering Solution

To avoid selecting the optimal clustering solution based on a single evaluation metric, a composite performance index was defined.

First, the five evaluation metrics were normalized to the [0, 1] interval. For the Silhouette Coefficient, Calinski–Harabasz Index, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI), normalization was performed by dividing each value by the maximum value observed across all evaluated configurations. For the Davies–Bouldin Index, whose interpretation is inversely related to clustering quality, normalization was computed as the ratio between the minimum observed value and the value obtained for each configuration.

A global performance score was then calculated as the arithmetic mean of the five normalized metrics (Eq. 1):

$$Score = \frac{Sil_{norm} + CH_{norm} + DB_{norm} + ARI_{norm} + NMI_{norm}}{5} \quad (1)$$

This approach assigns equal weight to all evaluation metrics, simultaneously integrating measures of cluster cohesion and separation (internal validation metrics) and agreement with the reference taxonomic classification (external validation metrics). Finally, all clustering configurations were ranked according to this composite score, and the configuration achieving the highest score was selected as the best clustering solution.

3. Results

3.1 Model 1

Table 1 summarizes the numerical results obtained with Model 1 on the study dataset. Specifically, it reports the internal and external validation metrics for different values of K across the clustering solutions generated by the evaluated clustering algorithms using the DINOv2 + UMAP feature representation. The best performance is achieved with K-means (K=10).

Table 1. Numerical results obtained with Model 1 on the study dataset. The row highlighted in gray indicates the best-performing configuration according to the highest Score value.

Algorithm	K	Silh.	C-H	D-B	ARI	NMI	Score
K-Means	3	0.53	437.93	0.66	0.14	0.32	0.642
	5	0.51	565.22	0.70	0.30	0.52	0.794
	7	0.55	756.33	0.64	0.37	0.56	0.917
	9	0.55	808.09	0.61	0.42	0.56	0.970
	10	**0.54**	**824.13**	**0.64**	**0.43**	**0.62**	**0.973**
	11	0.55	847.93	0.64	0.41	0.62	0.969
Ag. Clustering	3	0.53	437.93	0.66	0.14	0.32	0.642
	5	0.51	554.70	0.69	0.29	0.51	0.784
	7	0.52	525.21	0.64	0.32	0.56	0.826

	9	0.53	794.17	0.66	0.44	0.62	0.961
	10	0.51	749.63	0.66	0.44	0.63	0.944
	11	0.48	696.07	0.69	0.43	0.62	0.908
Spectral Clustering	3	0.44	254.00	1.12	0.17	0.38	0.528
	5	0.48	440.85	0.75	0.31	0.55	0.752
	7	0.56	733.95	0.67	0.38	0.57	0.914
	9	0.55	735.19	0.68	0.40	0.60	0.928
	10	0.55	740.43	0.70	0.42	0.62	0.934
	11	0.51	736.31	0.73	0.39	0.60	0.891
BIRCH	3	0.53	437.93	0.66	0.14	0.32	0.642
	5	0.51	563.43	0.70	0.31	0.53	0.800
	7	0.53	646.21	0.65	0.36	0.57	0.875
	9	0.53	705.36	0.66	0.37	0.57	0.890
	10	0.54	773.37	0.66	0.42	0.61	0.941
	11	0.54	785.19	0.67	0.40	0.61	0.937

The contingency matrix (Table 2) reveals a heterogeneous but taxonomically informative structure, with pure, single-species-dominated, and mixed clusters. The Dominance column reports the proportion of each species recovered by its dominant cluster (P(cluster | species)), computed directly from the absolute counts. Several species are recovered as compact units: *Thermonectus alfredi* (95% in C2), *T. tremouillesi* (88% in C6) and *T. circumscriptus* (79% in C5). In contrast, *Hydaticus palliatus* and *T. succinctus* show the lowest dominance (45% and 48%). At the cluster level, C3 is composed entirely of *H. palliatus* and C8 is dominated by *Notaticus fasciatus*, whereas C0 is shared almost equally by *T. margineguttatus* and *T. nobilis*.

Table 2. Species-by-cluster contingency matrix for the best clustering solution obtained with Model 1 (K-Means, K = 10). Each cell reports the number of images assigned to a given cluster. For each species, the dominant cluster (i.e., the cluster containing the largest number of images of that species) is highlighted in blue. The Dominance (%) column indicates the percentage of images of each species assigned to its dominant cluster.

Species/cluster	**C0**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**#imgs**	**Dom. (%)**
Hydaticus palliatus	0	0	0	**18**	6	0	0	9	0	7	40	45
Hydaticus tuyuensis	0	12	0	0	0	0	0	1	0	**16**	29	55
Hydaticus xanthomelas	0	**22**	0	0	5	5	0	8	0	0	40	55

Notaticus fasciatus	0	0	0	0	16	0	0	1	**23**	0	40	57
Thermonectus alfredi	0	0	**38**	0	0	1	1	0	0	0	40	95
Thermonectus circumscriptus	0	0	6	0	1	**31**	1	0	0	0	39	79
Thermonectus margineguttatus	**30**	6	0	0	2	0	0	2	0	0	40	75
Thermonectus nobilis	**29**	2	0	0	9	0	0	0	0	0	40	72
Thermonectus succinctus	0	0	8	0	**19**	7	2	0	4	0	40	48
Thermonectus tremouillesi	0	0	4	0	0	0	**29**	0	0	0	33	88
Total number of images	59	42	56	18	58	44	33	21	27	23	381	

At the genus level (Table 3), the three genera differ markedly. *Hydaticus* spreads across several clusters (mainly C1, C9, C3, C7). *Notaticus* concentrates in only two clusters (C4 and C8). *Thermonectus* is broadly distributed, in C0, C2, C5 and C6. Some clusters comprise a single genus (C0, C2, C6 exclusively *Thermonectus*; C9, exclusively *Hydaticus*; C8, predominantly *Notaticus*), while others combine several.

Table 3. Image distribution by cluster at genus level in the best clustering of the Model 1 (K-Means, K=10).

Genus / Cluster	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
Hydaticus	0	34	0	18	11	5	0	18	0	23
Notaticus	0	0	0	0	16	0	0	1	23	0
Thermonectus	59	8	56	0	31	39	33	2	4	0

3.2 Model 2

Table 4 presents the numerical results obtained with Model 2 on the study dataset. In this case, clustering performance was evaluated using the clustering algorithms applied to the preprocessed images represented by the first 100 principal components extracted through Principal Component Analysis (PCA). The best performance is achieved with Spectral Clustering (K=9).

Table 4. Numerical results obtained with Model 2 on the study dataset.

Algorithm	K	Silh.	C-H	D-B	ARI	NMI	Score
K-Means	3	0.12	54.88	2.45	0.04	0.09	0.744
	5	0.12	41.71	2.28	0.06	0.11	0.759
	7	0.13	36.22	2.06	0.08	0.15	0.845
	9	0.12	32.05	2.06	0.07	0.15	0.822
	10	0.12	30.41	2.15	0.07	0.16	0.793
	11	0.11	28.09	2.16	0.07	0.17	0.790

Ag. Clustering	3	0.10	40.97	2.18	0.04	0.09	0.651
	5	0.10	34.25	2.27	0.05	0.12	0.679
	7	0.11	30.72	2.23	0.06	0.14	0.730
	9	0.09	24.61	1.96	0.06	0.14	0.712
	10	0.09	23.00	1.91	0.06	0.17	0.756
	11	0.10	22.38	1.95	0.06	0.18	0.754
Spectral Clustering	3	0.11	48.08	2.20	0.05	0.12	0.761
	5	0.11	38.09	2.26	0.06	0.13	0.760
	7	0.12	35.07	2.05	0.07	0.14	0.807
	9	**0.12**	**31.35**	**2.09**	**0.09**	**0.18**	**0.860**
	10	0.11	29.85	2.17	0.09	0.19	0.852
	11	0.11	27.26	2.06	0.08	0.20	0.849
BIRCH	3	0.09	41.62	2.84	0.04	0.08	0.610
	5	0.10	34.62	2.39	0.05	0.10	0.651
	7	0.10	30.40	2.40	0.06	0.13	0.687
	9	0.09	26.31	2.32	0.07	0.15	0.729
	10	0.10	25.99	2.21	0.07	0.16	0.753
	11	0.10	25.47	2.25	0.08	0.17	0.780

For the best configuration of Model 2 (Spectral Clustering, K = 9), the contingency matrix (Table 5) reveals a highly dispersed distribution, with no species reaching clear dominance within any cluster. The highest concentration observed is that of *Notaticus fasciatus*, which concentrates 70% of its images in cluster 1, and *Thermonectus alfredi*, which reaches almost 50% in cluster 6 (Table 5). With the exception of *Hydaticus tuyuensis* (45% in C0), all remaining species show maximum proportions below 30% in any single cluster.

Table 5. Species-by-cluster contingency matrix for the best clustering solution obtained with Model 2 (Spectral Clustering, K = 9). Each cell reports the number of images assigned to a given cluster. For each species, the dominant cluster (i.e., the cluster containing the largest number of images of that species) is highlighted in blue. The Dominance (%) column indicates the percentage of images of each species assigned to its dominant cluster.

Species/cluster	**C0**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**#imgs**	**Dom. (%)**
Hydaticus palliatus	**11**	2	10	9	0	6	0	1	1	40	28
Hydaticus tuyuensis	**13**	0	7	4	1	1	2	1	0	29	45
Hydaticus xanthomelas	9	4	6	3	0	**11**	1	3	3	40	28
Notaticus fasciatus	0	**28**	2	4	1	3	0	2	0	40	70
Thermonectus alfredi	7	1	4	1	3	1	**20**	3	0	40	50
Thermonectus circumscriptus	4	9	2	4	4	**10**	2	2	2	39	26
Thermonectus margineguttatus	3	**10**	4	2	2	9	2	8	0	40	25
Thermonectus nobilis	1	6	4	4	5	1	7	4	**8**	40	20
Thermonectus succinctus	0	9	7	4	2	2	5	1	**10**	40	25
Thermonectus tremouillesi	7	5	**8**	2	3	1	2	5	0	33	24
Total number of images	55	74	54	37	21	45	41	30	24	381	

At the genus level, every cluster mixes at least two genera (Table 6). All clusters contain multiple genera, with no clear separation. Even where a genus dominates, it never does so alone: *Notaticus* reaches its highest count in C1 (n = 28), yet that same cluster absorbs 40 *Thermonectus* images and 6 of *Hydaticus*. *Thermonectus*, on the other hand, appears in all nine clusters.

Table 6. Image distribution by cluster at genus level in the best clustering of the Model 2 (Spectral Clustering, K=9).

Genus / Cluster	**C0**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**
Hydaticus	33	6	23	16	1	18	3	5	4
Notaticus	0	28	2	4	1	3	0	2	0
Thermonectus	22	40	29	17	19	24	38	23	20

3.3 Model Comparison

Tables 1 and 6 highlight contrasting patterns between the two models:
Model 1:

- Higher intraspecific concentration
- Presence of nearly monospecific clusters
- Partial recovery of taxonomic structure

Model 2:

- High intraspecific dispersion
- Strong interspecific overlap
- Low structural coherence

4. Discussion

The results indicate that the primary difference between both approaches lies in the data representation rather than the clustering algorithm.

Model 1 effectively captures relevant morphological patterns, as evidenced by clusters partially aligned with species, coherent groupings at the genus level (Table 3), and dominant concentrations in some clusters (Table 2). Nevertheless, separation is not complete. Some clusters are nearly monospecific while others combine several species in comparable proportions, showing that the model organizes visual similarity well but does not convert that structure into a perfect species-cluster partition. This limitation can be attributed to intraspecific variability, morphological similarity across species, and the continuous nature of phenotypic space. At the species level, two patterns can be noted. The first one involves relatively pure groupings (*Hydaticus palliatus* in C3, *Notaticus fasciatus* in C8, *Thermonectus tremouillesi* in C6). The second involves morphological overlap between species, particularly in C0, C1, C4 and C7. The nearly 1:1 coexistence of *T. margineguttatus* and *T. nobilis* in C0 — which together account for the whole cluster (75% and 72% of their respective images) — is consistent with their dorsal similarity (Fig. 1 G, H): comparable pronotal markings and a yellowish band at the base of the elytra. Cluster C9, in turn, groups two *Hydaticus* species without any other genus, indicating that at least a subset of *Hydaticus* specimens resemble each other more than any *Thermonectus*. At the genus level (Tables 2, 3), the feature space partially captures taxonomic structure, though heterogeneously: clusters largely dominated by a single genus (C9 for *Hydaticus*, C8 for *Notaticus*) coexist with a dispersed distribution of *Hydaticus* across several clusters, reflecting limited cohesion. At the species level, only 45% of *H. palliatus* images fall in its dominant cluster. Notably, some *H. palliatus* images with evident differences in coloration or illumination are still assigned to the same cluster, suggesting that the model groups by global shape and texture rather than local color variation. Conversely, *T. alfredi* (95%, C2) and *T. tremouillesi* (88%, C6) appear to possess distinctive characters that the model interprets as discrete entities, so that K-Means recovers boundaries closely matching the systematics of these beetles.

In contrast, Model 2 exhibits structural limitations, including fragmentation of species across clusters (Table 5), systematic coexistence of genera (Table 6), and the absence of taxonomically consistent groupings. This suggests that PCA applied to raw pixels fails to capture meaningful invariances, producing a feature space dominated by noise and low-level variations. The contingency matrix (Table 5) reinforces this interpretation, showing uniform distributions without clear dominance patterns. The low dominant-cluster proportions for most species (Table 5) point to a limited capacity to group specimens by species. The exception is *Notaticus fasciatus* (70% in C1), which behaves similarly to Model 1, suggesting more distinctive visual features. The fragmentation of *H. palliatus* and several *Thermonectus* indicates that the PCA representation does not robustly recover intraspecific similarity. At the genus level (Table 6), every cluster contains at least two genera, so no clear generic separation is achieved. Relative frequencies vary between clusters but not enough to make consistent groupings: C1 combines 28 *Notaticus* images with 40 *Thermonectus* and 6 *Hydaticus*, while *Thermonectus* occurs in every cluster. As in Model 1, the cluster where *Notaticus* concentrates shows null or very low *Hydaticus* representation,

indicating a consistent separation between these two genera in at least part of the feature space.

5. Conclusions and Future Work

As discussed in the Introduction, this study is, to our knowledge, the first to apply self-supervised deep learning embeddings to unsupervised morphological clustering in beetles.

The findings demonstrate that:

- Unsupervised learning can reveal morphological patterns without labeled data, recovering coherent structure at least partially
- Representation quality is the key factor determining clustering performance
- Deep embeddings reduced via UMAP significantly outperform PCA on raw pixels
- Species-level classification remains challenging, with fragmentation of several species across multiple clusters
- The structure present in the embeddings is not always robustly recovered, since performance depends strongly on the clustering scheme applied
- Despite these limitations, Model 1 shows relevant potential for taxonomic practice, helping to explore intra- and interspecific variation, generate taxonomic hypotheses, and assist identification

Future work will focus on:

- Evaluating more advanced self-supervised models
- Exploring deep clustering approaches
- Expanding and improving the dataset, particularly in terms of sample size per species

Acknowledgments

We thank four anonymous reviewers whose comments significantly improved the manuscript. We also thank Dr. M. Michat (University of Buenos Aires) for the loan of several specimens and for his contribution to the taxonomic identification of the material used in this study. This work was partially funded by the Fundación Ciencias Exactas y Naturales (FUNDACEN) under project +4i “Use of machine learning for taxonomic identification of microscopic samples of economic and environmental interest,” as well as by the “Program to Support Mandatory Experimental Undergraduate Theses (+Tesis)” of the Facultad de Ciencias Exactas y Naturales (University of Buenos Aires)-FUNDACEN, and by the Research Council of the Universidad Nacional de Salta through project 2735 “Technological Innovation Applied to Real-World Problems.”

References

Allaoui, M., Kherfi, M., Cheriet, A. (2020). Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In: El Moataz, A., Mammass, D., Mansouri, A., Nouboud, F. (eds). Image and Signal Processing. ICISP 2020. Lecture Notes in Computer Science, vol. 12119, pp. 317–325. Springer, Cham. https://doi.org/10.1007/978-3-030-51935-3_34

Arinik, N., Labatut, V., Figueiredo, R. (2021). Characterizing and Comparing External Measures for the Assessment of Cluster Analysis and Community Detection. IEEE Access, 9, 20255–20276. https://doi.org/10.1109/ACCESS.2021.3054621

Bellin, N., Calzolari, M., Magoga, G., Callegari, E., Bonilauri, P., Lelli, D., Dottori, M., Montagna, M., Rossi, V. (2022). Unsupervised machine learning and geometric morphometrics

as tools for the identification of inter and intraspecific variations in the *Anopheles maculipennis* complex. Acta Tropica, 233, 106585. https://doi.org/10.1016/j.actatropica.2022.106585

Bin Azis, T. M. F., Sharif, S. (2019). Clustering analysis of Coleopteran stored product pest based on morphometric structure. AIP Conference Proceedings, 2138, 050005. https://doi.org/10.1063/1.5121110

Charoenpanyakul, R., Kittichai, V., Eiamsamang, S., Sriwichai, P., Pinetsuksai, N., Naing, K., Tongloy, T., Boonsang, S., Chuwonginet, S. (2024). Enhancing mosquito classification through self-supervised learning. Scientific Reports, 14, 27123. https://doi.org/10.1038/s41598-024-78260-2

Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D., Thompson, E., Ashraf, I. (2023). A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective. Symmetry, 15, 1679. https://doi.org/10.3390/sym15091679

Derkarabetian, S., Castillo, S., Koo, P. K., Ovchinnikov, S., Hedin, M. (2019). A demonstration of unsupervised machine learning in species delimitation. Molecular Phylogenetics and Evolution, 139, 106562. https://doi.org/10.1016/j.ympev.2019.106562

Gao, Y., Xue, X., Qin, G., Li, K., Liu, J., Zhang, Y., Li, X. (2024). Application of machine learning in automatic image identification of insects - a review. Ecological Informatics, 80, 102539. https://doi.org/10.1016/j.ecoinf.2024.102539

Ghosal, A., Nandy, A., Das, A., Goswami, S., Panday, M. (2020). A Short Review on Different Clustering Techniques and Their Applications. In: Mandal, J., Bhattacharya, D. (eds). Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_9

Holmgren, S., Angus, R., Jia, F., Chen, Z. N., Bergsten, J. (2016). Resolving the taxonomic conundrum in Graphoderus of the east Palearctic with a key to all species (Coleoptera, Dytiscidae). ZooKeys, 574, 113–142. https://doi.org/10.3897/zookeys.574.7002

Kar, S., Nagasubramanian, K., Elango, D., Carroll, M. E., Abel, C. A., Nair, A., Mueller, D. S., O'Neal, M. E., Singh, A. K., Sarkar, S., Ganapathysubramanian, B., Singh, A. (2023). Self-supervised learning improves classification of agriculturally important insect pests in plants. The Plant Phenome Journal, 6, e20079. https://doi.org/10.1002/ppj2.20079

Marukatat, S. (2023). Tutorial on PCA and approximate PCA and approximate kernel PCA. Artificial Intelligence Review, 56, 5445–5477. https://doi.org/10.1007/s10462-022-10297-z

Oquab, M., Darcet, T., Moutakanni, T. et al. (2023). DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193. https://doi.org/10.48550/arXiv.2304.07193

Singh, S., Srivastava, S. (2020). Review of Clustering Techniques in Control System. Procedia Computer Science, 173, 272–280. https://doi.org/10.1016/j.procs.2020.06.032

Teixeira, A. C., Ribeiro, J., Morais, R., Sousa, J. J., Cunha, A. (2023). A Systematic Review on Automatic Insect Detection Using Deep Learning. Agriculture, 13(3), 713. https://doi.org/10.3390/agriculture13030713

Valan, M., Makónyi, K., Maki, A., Vondráček, D., Ronquist, F. (2019). Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. Systematic Biology, 68(6), 876–895. https://doi.org/10.1093/sysbio/syz014

Wägele, H., Klussmann-Kolb, A., Kuhlmann, M., Haszprunar, G., Lindberg, D., Koch, A., Wägele, J. W. (2011). The taxonomist – an endangered race. A practical proposal for its survival. *Frontiers in Zoology*, 8(1), 25. https://doi.org/10.1186/1742-9994-8-25

Yee, D. A. (2023). Ecology, systematics, and the natural history of predaceous diving beetles (Coleoptera: Dytiscidae). Springer, Cham. https://doi.org/10.1007/978-3-031-01245-7

Zhang, C., Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224