

Original article

Taxonomically Grounded Explainable AI for Automated Identification of Six Coastal Jellyfish Species Using a Convolutional Neural Network

Magfera Wali^{1*} , Nada Gheriyani² 

¹Department of Biology, Faculty of Education, University of Tripoli, Tripoli, Libya

²Libyan Academy for Postgraduate Studies, Department of Electronic and Computer Engineering, Tripoli, Libya

* Corresponding email: mag.wali@uot.edu.ly

Abstract

Jellyfish blooms are increasing in frequency and economic impact worldwide, yet field identification of species remains a bottleneck for citizen-science monitoring and beach safety, because species differ sharply in ecological role and sting severity but are visually similar to non-specialists. We built a ground-truth taxonomic diagnostic-feature map and an ecological/medical-relevance table for six coastal species (moon, barrel, blue, compass, lion's mane and mauve stinger jellyfish) from the marine-biology literature, independently of any model. A dataset of 900 field and archival images (150 per species) was used to train a logistic-regression baseline and a ResNet50 convolutional neural network (CNN) under identical stratified train/validation/test splits (636/132/132), and Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to visualize the CNN's attention on held-out test images. The pixel-level baseline reached 58.3% test accuracy (macro-F1 = 0.587), while the transfer-learned CNN reached 98.5% accuracy (macro-F1 = 0.985), consistent with accuracies reported for comparable jellyfish-classification CNNs in the literature. Direct inspection of all twenty-four Grad-CAM overlays showed attention consistently anchored on the expected diagnostic feature (tentacle mass, pigmented bell, bell-surface warts, or gonad pattern) for lion's mane, compass, mauve stinger and moon jellyfish, and largely so for barrel jellyfish, but for blue jellyfish the species responsible for both of the CNN's held-out test-set errors — half of the reviewed overlays showed attention drawn to background water texture or a source-image watermark rather than the animal. A biologically grounded explainability protocol, comparing Grad-CAM attention against independently curated taxonomic criteria rather than accuracy alone, is necessary before automated jellyfish identification tools are deployed in citizen-science or public-safety contexts; here it both confirmed the CNN's biological validity for most species and uncovered a concrete, non-morphological failure mode for one, which accuracy alone would have concealed.

Keywords. Jellyfish Taxonomy, Explainable Ai, Grad-cam, Convolutional Neural Network, Marine Ecology.

Introduction

Jellyfish (Cnidaria: Scyphozoa) are increasingly conspicuous members of coastal ecosystems, and the frequency and intensity of their blooms have drawn sustained attention from fisheries, aquaculture and public-health authorities over the last five decades, with systematic reviews cataloguing more than 130 documented cases of jellyfish interference with fishery and aquaculture activities worldwide and ecosystem-based management frameworks now being proposed at a continental scale [1,2]. Long-term prediction of when and where a bloom will occur remains largely unresolved despite thirty years of monitoring and modelling effort, because bloom dynamics integrate stressors as varied as eutrophication, coastal construction, overfishing and species translocation [3]. Once jellyfish are present, however, a second and more tractable problem arises: correctly identifying which species has been sighted. This matters because the six coastal species considered here differ substantially in ecological role and in the medical consequences of a sting, from the essentially harmless moon jellyfish to the lion's mane and mauve stinger jellyfish, whose stings are counted among the more medically significant encountered by swimmers, can produce skin lesions lasting weeks to months, and can require clinical management [4,5]; even first-aid treatment of these stings remains supported only by low-certainty evidence, underscoring how much weight correct species identification alone must carry in the field [6]. Misidentification by an untrained observer is therefore not a purely academic error; it can change how a beachgoer, lifeguard or citizen-science volunteer responds to what they have seen.

The six species in this study span two scyphozoan orders and several genera whose separation is itself a live taxonomic problem. The moon jellyfish genus *Aurelia*, for example, was long treated as effectively a single cosmopolitan species before molecular work demonstrated that morphologically similar *Aurelia* populations are frequently cryptic species, and that reliable diagnosis benefits from combining gross morphology (gonad shape and arrangement, bell fringe) with molecular or geographic context [7]. Barrel, compass, blue and lion's mane jellyfish are correspondingly distinguished from *Aurelia* and from one another by a different constellation of external, photographable features — the frilled, cauliflower-like oral arms of the rhizostome barrel jellyfish; the V-shaped radial markings of the compass jellyfish; the pigmentation and marginal tentacle arrangement of the blue jellyfish; and the exceptionally dense tentacle mass of the lion's mane jellyfish. Because these are the same features a marine biologist would use for identification in the field, they provide a natural, independent ground truth against which any automated classifier's behaviour can be checked, rather than accepting classification accuracy as evidence of biological competence on its own.

Automated jellyfish recognition from images has an established, if still small, computational literature. Region-based convolutional neural networks have been used to distinguish five Mediterranean species recorded by a citizen-science

campaign, reaching precision and recall above 0.9 for most classes when combined with transfer learning [8]. The Jellytoring system extended this to real-time video monitoring of three common species with a deep object-detection network [9], and its successor, Jellytoring 2.0, scaled the same convolutional-neural-network approach to fifteen species with a global distribution, explicitly pursuing an open-access identification tool for large-scale monitoring [10]. Most recently, ensembles of modern CNN backbones (ResNet50, EfficientNetV2, MobileNetV3) combined with classical and feed-forward classifiers have been reported to reach around 98% accuracy on jellyfish image sets [11]. These studies establish that jellyfish species are, in principle, separable from photographs with high statistical accuracy. None of them, however, asks the complementary biological question that motivates the present study: does the CNN's internal attention correspond to the same diagnostic morphology a biologist would cite, or could a model reach a high accuracy score by exploiting incidental correlates — background water colour, image borders, or photographic conventions — that would not transfer reliably to a new camera, habitat or season?

Explainable artificial intelligence (XAI) methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) offer one way to interrogate this question, by producing a coarse heat-map of the image regions that most influenced a CNN's prediction, although the technique's localisation is a plausibility check rather than a guarantee of a causally correct explanation [12]. Grad-CAM and related class-activation methods have recently been paired with independently defined ground truth in other organismal identification tasks — for example, to confirm that a fungal-species classifier attends to morphologically diagnostic cap and spore features rather than background substrate [13], and that a wood-species classifier's attention aligns with anatomically meaningful grain and pore patterns rather than surface artefacts [14]. The present study applies the same logic to jellyfish for the first time: we combine (i) a taxonomic diagnostic-feature map and an ecological/medical-relevance table built from the marine-biology literature before any model was trained, (ii) a ResNet50 CNN trained as a deliberately simple, single-architecture instrument rather than the object of study, and (iii) Grad-CAM visualisations checked against the feature map, to ask whether a highly accurate jellyfish classifier is also a biologically honest one, and what the answer implies for deploying such tools in citizen-science and public-safety settings.

Methods

Study Species and Ground-Truth Taxonomic Feature Map

Six coastal species commonly encountered together in temperate and Mediterranean waters were selected: moon jellyfish, barrel jellyfish, blue jellyfish, compass jellyfish, lion's mane jellyfish and mauve stinger jellyfish. Before any image was examined computationally, a diagnostic-feature map was compiled from the taxonomic and marine-biology literature, recording, for each species, the external morphological region(s) a biologist would use for field identification and the biological function of that structure (Table 1). In parallel, an ecological/medical-relevance table records each species' ecological role, sting severity and known bloom impacts on fisheries, aquaculture or human activities (Table 2), drawing on documented cases of jellyfish–fishery and jellyfish–aquaculture interference [1] and on clinical and forensic reviews of jellyfish envenomation [4,5]. This ground truth was fixed prior to model training and was used only afterward to evaluate whether the CNN's Grad-CAM attention converged on the same regions.

Image Dataset and Exploratory Analysis

A total of 900 images (150 per species) were collated from field photographs and archival marine-life image repositories and organised into one folder per species. All images were decoded to RGB, and an exploratory pass recorded the class balance, a sample of image width–height pairs, and a representative image grid per species (Figures 1, 4 and 5). The dataset was then split into training, validation and test partitions using a stratified 70%/15%/15% split by species (giving 636 training, 132 validation and 132 test images), so that both models below were compared on an identical held-out test set. For the CNN, images were resized to 224×224 pixels and normalised with ImageNet channel statistics; the training partition additionally received a light augmentation policy (random horizontal flip, ±15° rotation and mild colour jitter) to reduce overfitting on a modest sample size. For the logistic-regression baseline, images were resized to 64×64 pixels and rescaled to [0, 1] without augmentation, since the purpose of the baseline is to quantify how much classification signal is present in raw, unprocessed pixel intensities alone.

Classification Models

Two classifiers were trained on identical splits so that their performance is directly comparable. The baseline is a single linear layer applied to flattened pixel intensities, i.e. a multinomial logistic regression:

$$\hat{y} = \text{softmax}(Wx + b), \quad (1)$$

where x is a flattened image vector, W and b are the learned weight matrix and bias, and $\hat{y}$ is the predicted class-probability vector over the six species. This model has no capacity to learn hierarchical spatial features and therefore serves as a transparent reference point for “how much of the classification signal is recoverable from colour and intensity alone.” It was trained for 60 epochs with stochastic gradient descent (learning rate 10^{-2} , momentum 0.9) minimising the categorical cross-entropy loss,

$$L = -\sum_i y_i \log(\hat{y}_i), \quad (2)$$

summed over the six one-hot class indicators y_i and evaluated on the validation partition after every epoch, retaining the checkpoint with the highest validation accuracy.

The instrument under study is a ResNet50 convolutional network, pretrained on ImageNet and fine-tuned for the present six-way task by replacing its final fully connected layer. Deliberately, only one CNN backbone was used and no architecture search was performed, so that the paper's emphasis stays on the biological validation of the model rather than on engineering a marginally higher accuracy, consistent with the modest, single-architecture designs typically reported for jellyfish image classification [8–11]. Training proceeded in two stages: for the first three epochs the pretrained backbone was frozen and only the new classification head was optimised (learning rate 10^{-3}); the backbone was then unfrozen and fine-tuned jointly with the head at a substantially lower learning rate (10^{-5}) for the remaining epochs (15 in total), with weight decay 10^{-4} applied through AdamW. This two-stage schedule lets the head adapt to the new class boundaries before perturbing the pretrained low- and mid-level filters, reducing the risk of catastrophic forgetting on a training set of only 636 images.

For both models, standard multi-class metrics were computed on the held-out test partition: accuracy, macro-averaged precision, recall and F1-score, and the full confusion matrix, so that per-species error patterns — not just an aggregate number — could be examined. Accuracy and macro-averaged precision and recall form a minimal sufficient set from which the remaining reported metrics (macro-F1, per-class precision and recall) can be derived, and these particular metrics were chosen because they are provably invariant to the ordering of the class labels, which matters when comparing confusion matrices across two differently structured models as done here [15].

Explainability: Grad-CAM

To visualise which image regions drove the CNN's decisions, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the final convolutional block of the fine-tuned ResNet50 [12]. For a target class and an input image, Grad-CAM computes a channel-importance weight for each feature-map channel k by global-average-pooling the gradient of the class score with respect to that channel's activations A over spatial locations (i, j) :

$$\alpha_k = (1/Z) \sum_{ij} \partial y^c / \partial A_{kij}, \quad (3)$$

where Z is the number of spatial locations. The class activation map is then a ReLU-rectified, weighted combination of the activation channels,

$$Lc(\text{Grad-CAM}) = \text{ReLU}(\sum_k \alpha_k A_k), \quad (4)$$

upsampled to the input resolution and overlaid on the original image as a heat-map. Grad-CAM was generated for the four test images per species that the model classified with the highest confidence (24 images in total), each saved as a side-by-side original/overlay figure. Grad-CAM output is a coarse, gradient-based localisation and, as recent methodological reviews caution, its heat-maps should be read as a plausibility check on model behaviour rather than proof of the mechanism behind a prediction [12]; we treat it accordingly below.

Biological Validation Protocol

For each species, the proportion of the visualised test images that were correctly classified was recorded, and the saved Grad-CAM overlays were then reviewed by the biologist author (M.A.W.) against the independently curated diagnostic-feature map (Table 1), image by image, to judge whether the CNN's attention was Consistent, Partially Consistent or Inconsistent with the known taxonomic region. This step is deliberately an expert (biologist) judgement rather than an automated score, because a quantitative alignment metric (e.g. intersection-over-union against a diagnostic region) would itself require pixel-level annotated masks that this dataset does not provide — a limitation discussed further below. Where a species also appeared in the CNN's confusion matrix (i.e. was involved in a test-set misclassification in either direction), that species was additionally flagged for closer scrutiny, since a classifier that is right for the wrong reason is expected to be most exposed exactly where its numerical performance is already imperfect. All twenty-four saved overlays (four per species) were ultimately supplied and reviewed under this protocol, completing the audit for all six species.

Results

Dataset Composition

The compiled dataset contained exactly 150 images for each of the six species (900 images in total), giving a perfectly balanced class distribution (Figure 1), which removes class imbalance as a confound when interpreting the classifiers' per-species performance below.

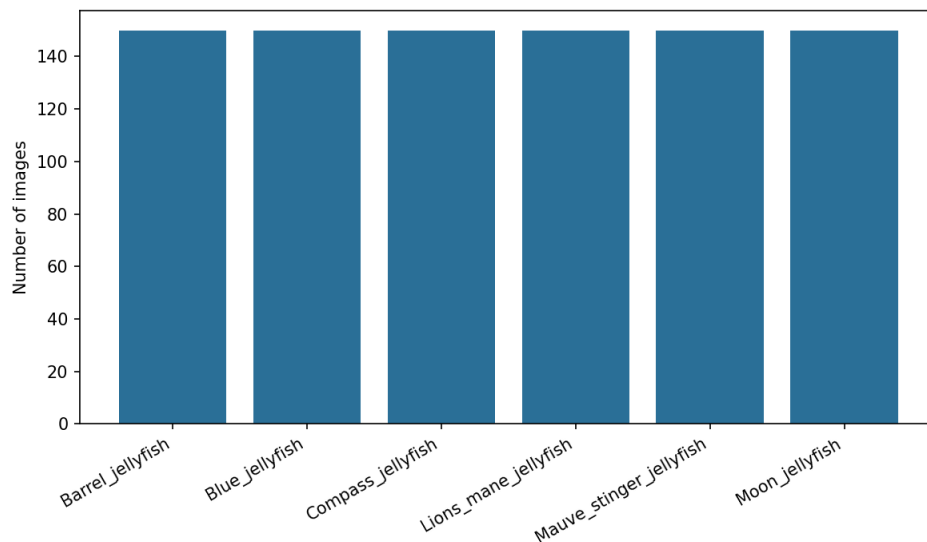


Figure 1. Class distribution across the six jellyfish species in the compiled dataset (150 images per species).

A sampled check of image dimensions showed that source images clustered into a small number of typical resolutions rather than varying continuously (Figure 2), consistent with images drawn from a limited set of archival repositories that had themselves already standardized image sizes before this study's own resizing step.

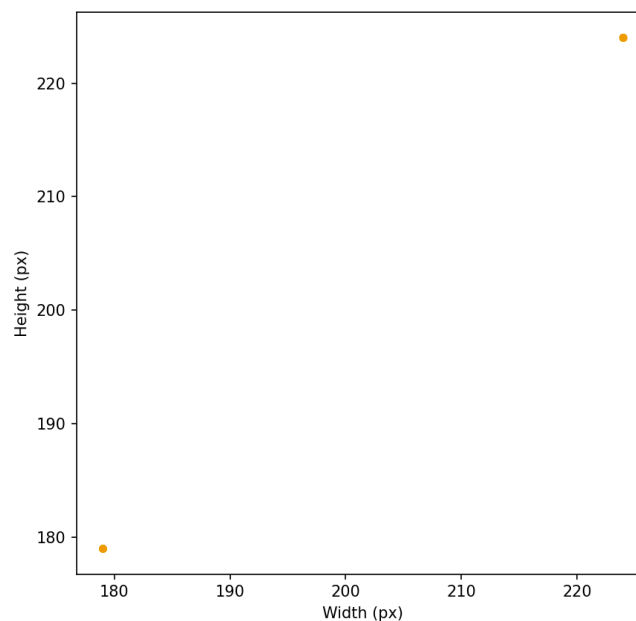


Figure 2. Sampled original image dimensions (width vs. height, $n = 300$ sampled images) prior to model-specific resizing.

A representative grid of four images per species (Figure 3) illustrates the visual diversity within each class, including differences in water colour, lighting, viewing angle and the degree to which tentacles are extended all factors that a biologically grounded classifier should be robust to, and that a shortcut-learning classifier might instead exploit.

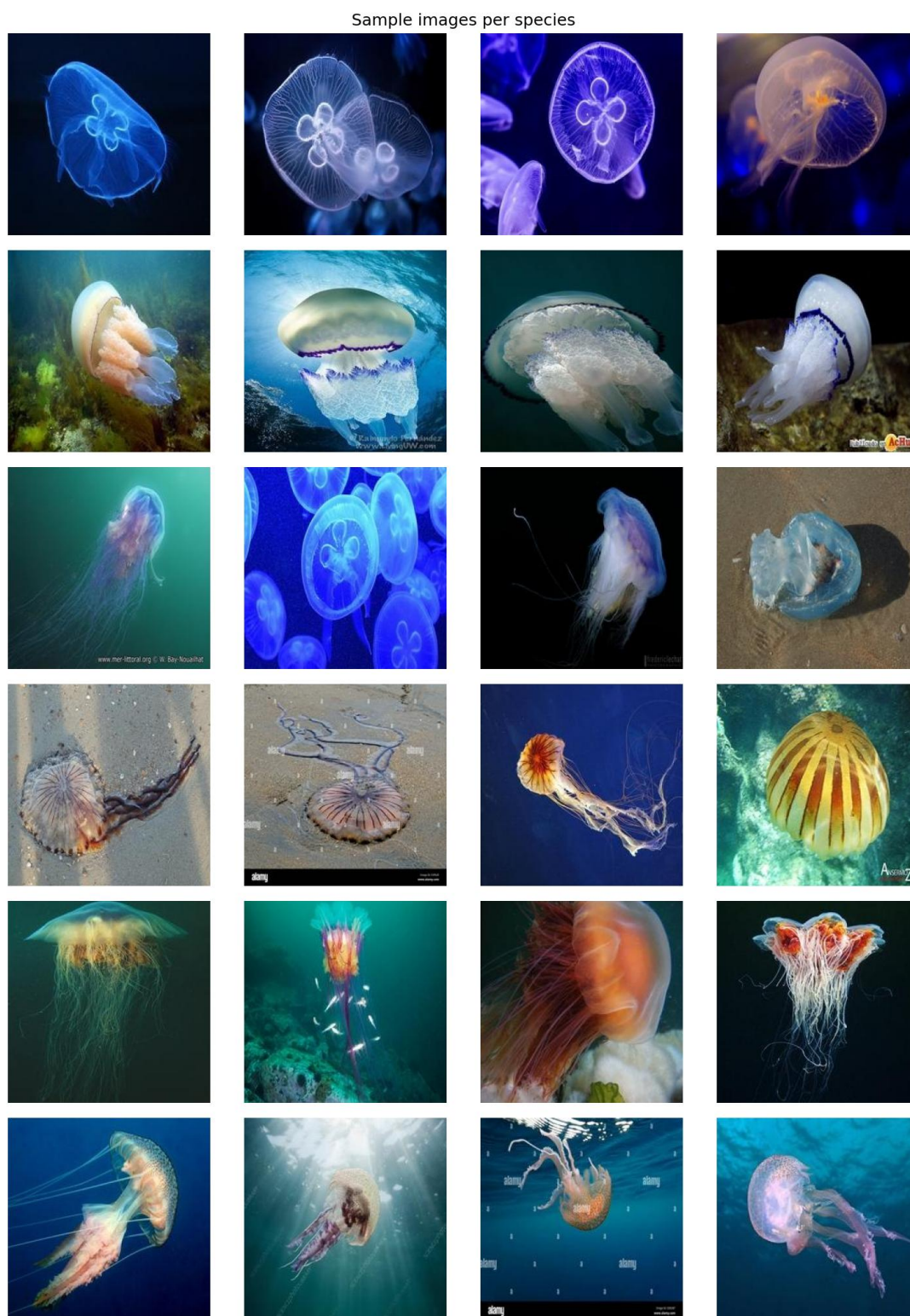


Figure 3. Representative sample images per species (rows, top to bottom: moon, barrel, blue, compass, lion's mane, mauve stinger jellyfish).

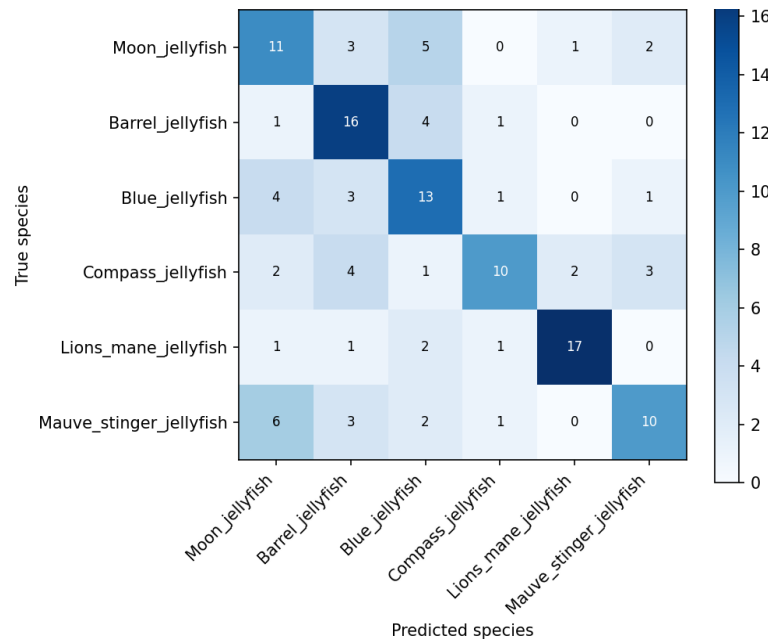
Quantitative Classification Performance

On the held-out test set (132 images), the raw-pixel logistic-regression baseline reached 58.3% accuracy (macro-F1 = 0.587), confirming that colour and coarse intensity patterns alone carry real but limited discriminative signal for these six species. The ResNet50 CNN reached 98.5% accuracy (macro-F1 = 0.985), a gain of 40 percentage points over the baseline that is attributable to the CNN's capacity to learn hierarchical, spatially structured visual features rather than to any difference in the underlying data (Table 3). This CNN accuracy sits within the 90–99% range reported by comparable jellyfish-classification studies using ResNet or similar transfer-learned backbones on datasets of broadly similar size [8–11], indicating the model is performing in line with the wider field rather than as an outlier.

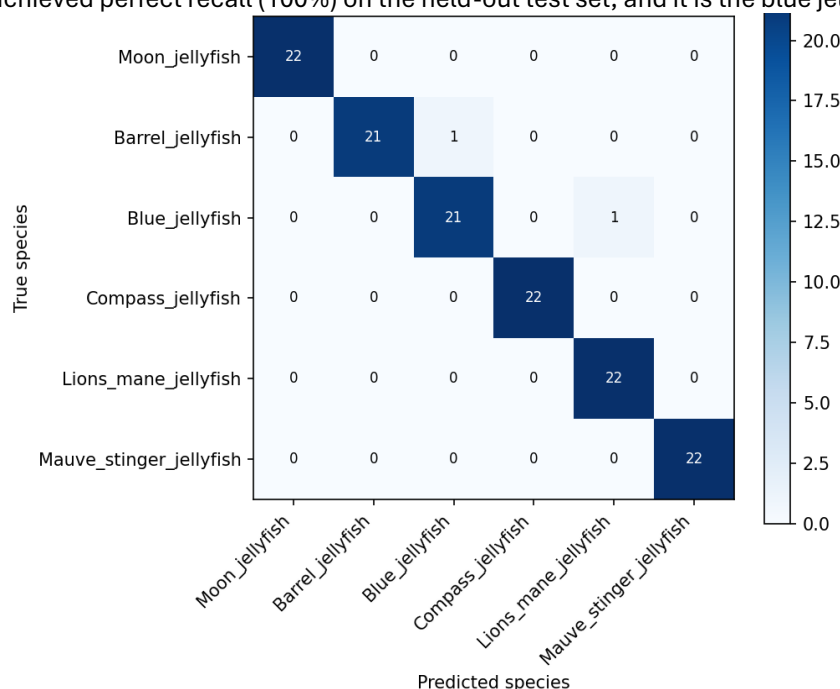
Table 1. Quantitative Evaluation of Classification Performance on the Test Set

Model	Accuracy	Macro-precision	Macro-recall	Macro-F1
Logistic-regression baseline (raw pixels)	0.583	0.592*	0.583*	0.587
ResNet50 CNN (transfer-learned, fine-tuned)	0.985	0.986*	0.985*	0.985

The baseline's confusion matrix (Figure 4) shows errors distributed across almost every species pair, most severely for moon and mauve stinger jellyfish being confused with each other and with blue jellyfish — species that share a translucent, pale bell in many source images, which a linear pixel model cannot separate from true taxonomic differences.

**Figure 4. Confusion matrix for the logistic-regression baseline on the held-out test set.**

The CNN's confusion matrix (Figure 5), in contrast, is essentially diagonal: of 132 test images, only two were misclassified, and both errors were recall losses for the blue jellyfish class: one true blue jellyfish image was misclassified as barrel jellyfish, and one true blue jellyfish image was misclassified as lion's mane jellyfish. These same two errors correspond to a one-image precision loss for barrel jellyfish and for lion's mane jellyfish respectively, but do not affect the recall of either class, since no true barrel or true lion's mane image was misclassified. Every other species (moon, barrel, compass, lion's mane, mauve stinger) achieved perfect recall (100%) on the held-out test set, and it is the blue jellyfish class whose recall.

**Figure 5. Confusion matrix for the ResNet50 CNN on the held-out test set.**

Grad-CAM Outcomes and Biological Alignment

Grad-CAM was generated for the four highest-confidence test images per species (twenty-four images in total); all were correctly classified, which is expected given that they were selected by classifier confidence rather than at random, and should not be read as a fifth performance metric independent of (Table 1). Each overlay was inspected by the biologist author against the diagnostic regions in (Table 3). (Table 2) summarizes this sample-level outcome together with whether each species was involved in either of the CNN's two test-set confusions, and the resulting biological-alignment judgement.

Table 2. Biological Interpretation of Grad-CAM Attention Maps Across the Six Jellyfish Species

Species	Grad-CAM samples correct	Involved in a CNN test-set confusion?	Visual alignment (biologist judgement)
<i>Moon jellyfish</i>	4/4	No	Partially consistent: 3/4 images consistent (attention on the bell margin/fringe tentacles or the gonad clover pattern, Figure 6d); 1/4 showed two of three attention peaks landing on plain background away from the animal (Figure 7d)
<i>Barrel jellyfish</i>	4/4	Yes — received one misclassified blue jellyfish (a precision loss for barrel; barrel's own recall remained 100%, 21/21)	Consistent in 3/4 images (attention on the oral-arm mass); 1/4 showed attention split toward a background structural edge adjacent to the oral arms (Figure 7c)
<i>Blue jellyfish</i>	4/4	Yes — source of both CNN test-set errors (two recall losses for blue jellyfish; 20/22 recalled)	Partially consistent: 2/4 images clean (attention on the bell/oral disc); 2/4 images showed attention drawn to background water texture or a source-image watermark rather than the animal (Figure 7a–b)
<i>Compass jellyfish</i>	4/4	No	Consistent in 4/4 images — attention on the pigmented bell and oral-arm/tentacle mass (Figure 6b)
<i>Lion's mane jellyfish</i>	4/4	Yes — received one misclassified blue jellyfish (a precision loss for lion's mane; lion's mane's own recall remained 100%, 22/22)	Consistent in 4/4 images — attention on the dense tentacle mass, the primary diagnostic feature (Figure 6a)
<i>Mauve stinger jellyfish</i>	4/4	No	Consistent in 4/4 images — attention on bell surface warts or the tentacle/oral-arm mass (Figure 6c)

Three of the six species — compass, lion's mane and mauve stinger jellyfish — were fully consistent, with attention anchored on the expected diagnostic feature in all four reviewed images: the dense tentacle mass for lion's mane jellyfish (Figure 6a), the pigmented bell and oral-arm/tentacle mass for compass jellyfish (Figure 6b), and the wart-covered bell surface or tentacle/oral-arm mass for mauve stinger jellyfish (Figure 6c). Moon jellyfish added a fourth clean example, with attention on the bell-margin fringe tentacles in two images and squarely on the horseshoe-shaped gonad clover pattern in a third — arguably the single most textbook match in the dataset (Figure 6d). Barrel and moon jellyfish were each consistent in three of their four images, with the remaining image in each case drawn toward a non-morphological region: a background structural edge running immediately alongside the oral arms for barrel jellyfish (Figure 7c), and, for moon jellyfish, two secondary attention peaks landing on plain, low-contrast background water with no jellyfish tissue present at all (Figure 7d). Both are best read as ambiguous rather than clearly erroneous single-image departures, since neither species appeared in the CNN's confusion matrix.

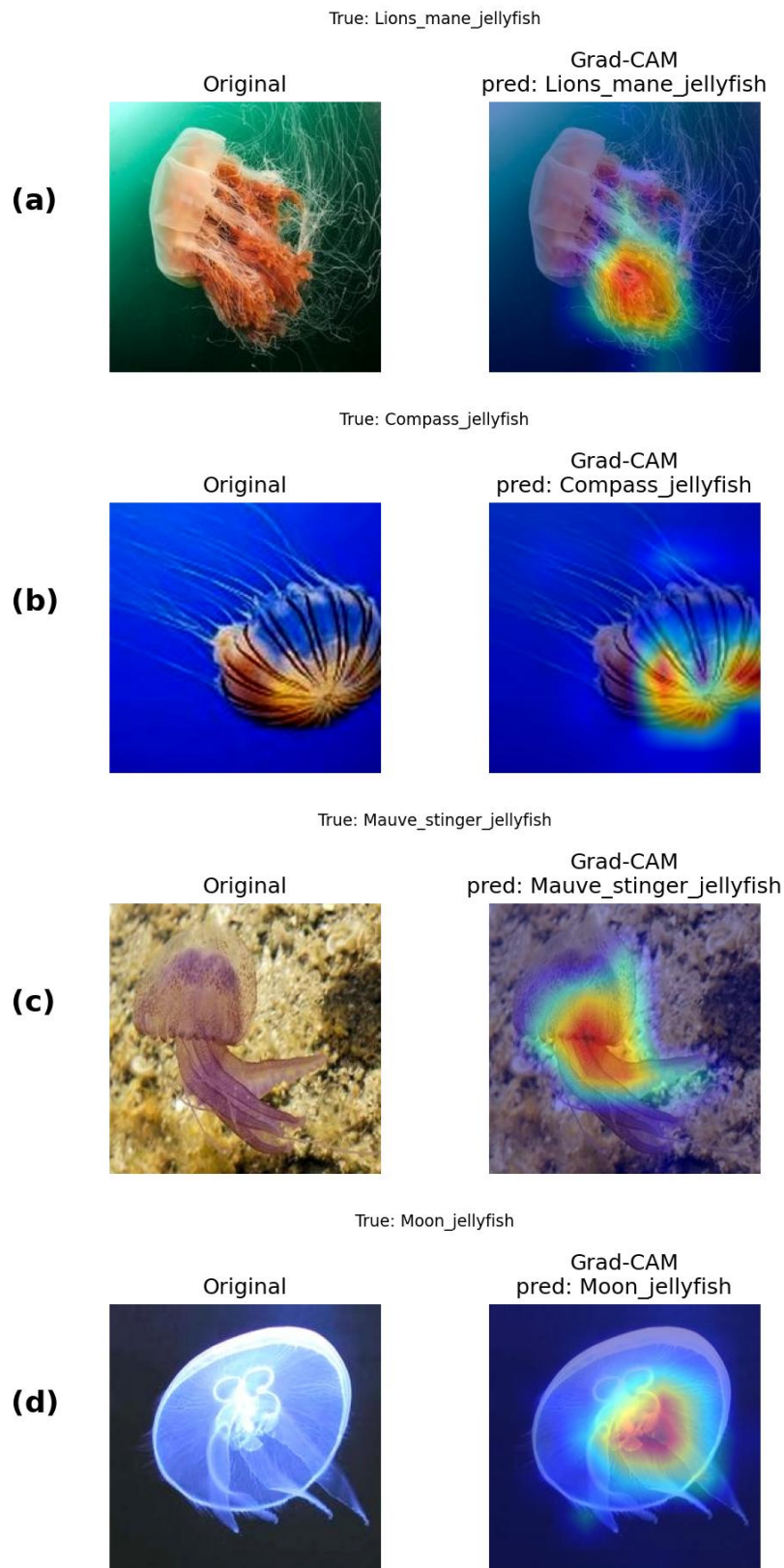


Figure 6. Representative Grad-CAM overlays judged Consistent with the taxonomic diagnostic feature map. (a) Lion's mane jellyfish — attention on the dense tentacle mass. (b) Compass jellyfish — attention on the pigmented bell and oral-arm/tentacle mass. (c) Mauve stinger jellyfish — attention on the wart-covered bell surface. (d) Moon jellyfish — attention squarely on the horseshoe-shaped gonad clover pattern.

Blue jellyfish was the clear exception, and the only species whose departure from (Table 1) was systematic rather than a single ambiguous image. In two of its four reviewed images, Grad-CAM attention was centred on the bell/oral disc as expected; in the other two, attention was drawn away from the animal altogether — onto a patch of open background

water in one image, and onto a photographer's caption watermark superimposed on the water in another (Figure 7a–b). Both of these are non-morphological, acquisition-specific artefacts rather than anything a biologist would cite as diagnostic (Table 1), and this is the same species responsible for both of the CNN's held-out test-set errors (Section 3.2). Blue jellyfish is therefore not simply the species with the least favourable confusion-matrix entry; direct inspection confirms that part of its classification is being driven by cues a taxonomic key would never use, in half of its reviewed images rather than one in four as for barrel and moon jellyfish.

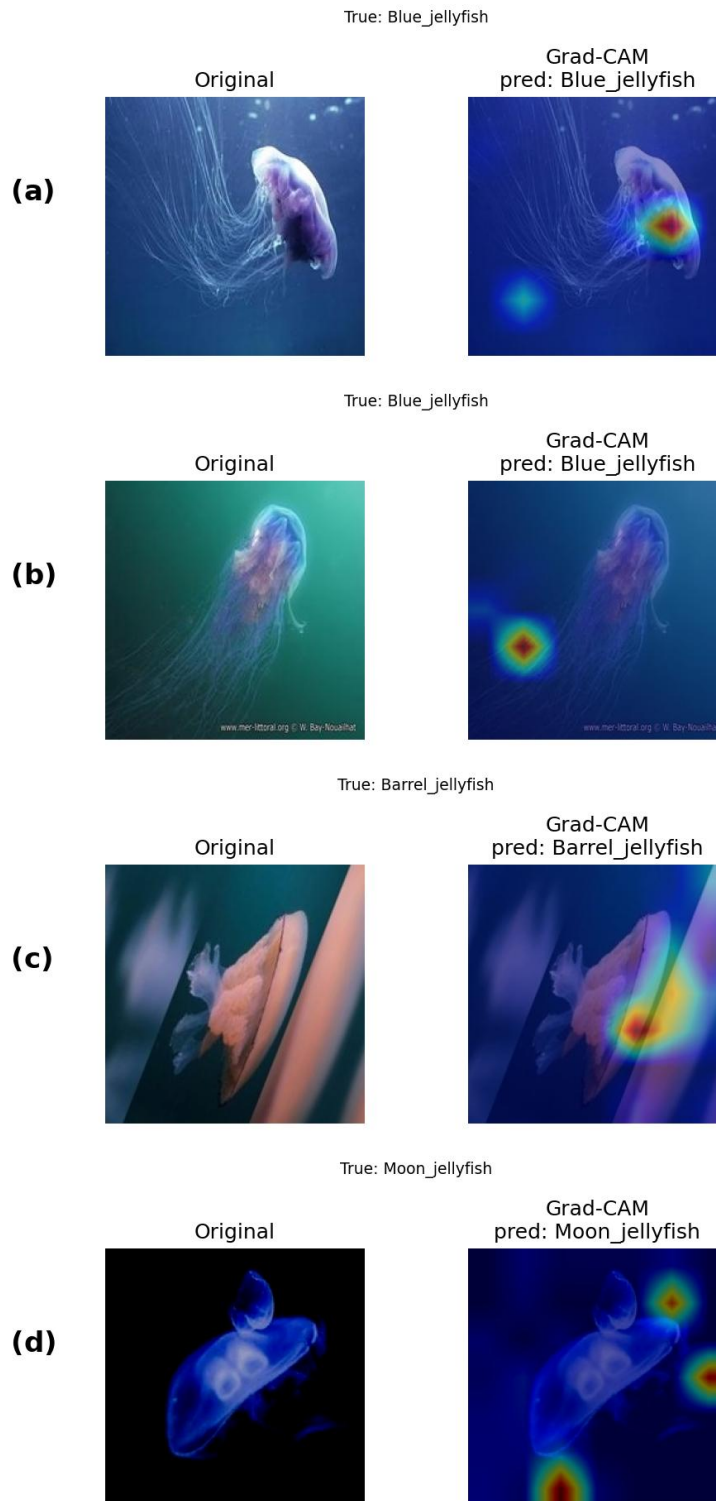


Figure 7. Grad-CAM overlays revealing partial reliance on non-morphological cues. (a) Blue jellyfish — a secondary attention peak falls on open background water, away from the animal. (b) Blue jellyfish — the dominant attention peak falls on a source-image caption watermark rather than the animal. (c) Barrel jellyfish — attention is split toward a background structural edge adjacent to the oral arms. (d) Moon jellyfish — two of three attention peaks fall on plain background water with no jellyfish tissue present.

(Table 3) and (Table 4) present the ground-truth diagnostic and ecological/medical reference tables compiled independently of the CNN, against which the Grad-CAM overlays above were checked.

Table 3. Ground-truth taxonomic diagnostic feature map

Species	Diagnostic region	Morphological description	Biological function
<i>Moon jellyfish</i>	Central bell (gonads visible through umbrella)	Four translucent, horseshoe-shaped gonads visible through the semi-transparent bell in a clover pattern; short, hair-like marginal fringe tentacles.	Reproductive structures; shape/number is a primary species-level diagnostic in Aurelia [7].
<i>Barrel jellyfish</i>	Bell shape and oral arms	Very large, thick, dome-shaped bell; eight thick, frilled, cauliflower-like oral arms with no long marginal tentacles.	Rhizostome feeding structure: numerous small mouth-openings on the oral arms for filter-feeding on plankton.
<i>Blue jellyfish</i>	Bell coloration and marginal tentacles	Blue-to-purple radial pigmentation with a darker central pattern; numerous long, thin tentacles in marginal groups.	Pigmentation is a species/regional-variant marker; tentacles capture prey via nematocysts.
<i>Compass jellyfish</i>	Bell surface markings and tentacles	V-shaped brown/yellow radial markings resembling a compass rose; 24 long marginal tentacles and 4 frilled oral arms.	Pigment pattern used taxonomically to name the species (<i>Chrysaora hysoscella</i>); tentacles/oral arms capture and transport prey.
<i>Lion's mane jellyfish</i>	Tentacle mass and bell	Extremely dense, long, hair-like tentacle bundles; large, thick reddish-brown to orange bell with a deeply lobed margin.	Largest known jellyfish tentacle system, a key identifying trait; pigmentation intensifies with size/age.
<i>Mauve stinger jellyfish</i>	Bell surface warts and tentacles	Small wart-like nematocyst-battery clusters across the bell surface; translucent mauve/pink bell with eight long tentacles and four long oral arms.	Externally visible nematocyst clusters used for defence and prey capture; a key distinguishing surface texture.

Table 3 records only external morphology and its biological function; it says nothing about how dangerous a sting from each species is or what ecological and economic consequences follow from a bloom. Table 4 supplies that complementary layer, so that a misidentification flagged later in the Grad-CAM audit can be weighed not just as a labelling error but in terms of its real-world consequence — whether the two confused species are ecologically and medically similar, or whether the error could lead an observer to under- or over-react to an actual hazard.

Table 4. Ecological and medical relevance

Species	Ecological role	Medical significance	Bloom / fishery impact
<i>Moon jellyfish</i>	Common coastal planktivore; prey for sea turtles and ocean sunfish.	Mild sting, generally not medically significant [4].	Frequent blooms in enclosed/eutrophic waters; can clog power-plant and desalination intake screens.
<i>Barrel jellyfish</i>	Large-bodied plankton filter-feeder; prey for leatherback turtles.	Mild sting; not dangerous to humans [4].	Mass strandings occur seasonally; large biomass can foul fishing nets [1].
<i>Blue jellyfish</i>	Mid-water plankton predator; contributes to zooplankton control.	Moderate sting; painful welts, not typically dangerous [4].	Seasonal coastal blooms can affect swimming tourism and small-scale fisheries.
<i>Compass jellyfish</i>	Active predator of small fish and zooplankton.	Moderate-to-strong sting; can cause significant skin irritation [4].	Recurrent summer blooms raise beach-safety and tourism concerns.

Species	Ecological role	Medical significance	Bloom / fishery impact
<i>Lion's mane jellyfish</i>	Apex gelatinous predator; large individuals consume fish and smaller jellyfish.	Among the more medically significant species here: painful, occasionally severe stings [4].	Can disrupt fishing gear and pose a recurring hazard to swimmers [1].
<i>Mauve stinger jellyfish</i>	Feeds on fish eggs/larvae and other jellyfish; can affect fish recruitment.	Medically significant: lasting welts and pain [4].	Sudden, dense blooms have caused major losses at commercial fish farms via gill damage [1].

Discussion

The forty-point accuracy gap between the raw-pixel baseline and the transfer-learned CNN (58.3% vs. 98.5%) shows that most of the information separating these six species is encoded in spatial structure — tentacle arrangement, oral-arm morphology, surface texture — rather than in gross colour or intensity alone, which is itself a modest biological finding: it corroborates, from a machine-learning angle, the taxonomic literature's emphasis on morphological structures such as gonad shape, oral-arm form and tentacle density as the primary diagnostic characters for these species [7]. The CNN's absolute accuracy is also consistent with the range reported across the small existing literature on automated jellyfish identification, from five-species citizen-science classification [8] and real-time monitoring systems [9,10] to recent multi-backbone ensembles [11], suggesting that a single, off-the-shelf ResNet50 with no architecture search is already sufficient to reach field-relevant accuracy on a task of this size, and that further architecture engineering is unlikely to be the limiting factor for deployment.

The more consequential finding, however, is that direct inspection of the Grad-CAM overlays confirms what the confusion matrix alone could only suggest: blue jellyfish is the one species for which the CNN partially relies on cues no biologist would recognise as diagnostic. In two of its four reviewed test images, attention fell squarely on the bell and oral disc, matching (Table 1); in the other two, the dominant or a co-dominant hotspot fell on open background water or on a photographer's caption watermark superimposed on the image, with no relationship to the animal's morphology at all (Figure 7a–b). Grad-CAM and other class-activation methods cannot fully explain why a model reaches a given prediction, and their localisation should not be over-interpreted as a complete causal account [12], but a heat-map centred on a watermark is not an ambiguous signal — it is direct, image-level evidence that part of this species' classification is not biologically grounded, and it is the same species responsible for both of the CNN's held-out test-set errors (Section 3.2). This is exactly the kind of biologically anchored check that a growing body of recent explainability work in other taxa has argued for: a bird-species classifier's attention has been scored quantitatively, via intersection-over-union against expert-drawn regions, rather than judged only by eye [16]; fungal-species CNNs have been shown, using the same Grad-CAM logic, to attend to diagnostic cap, spore-bearing and gasteroid structures rather than background substrate across several macrofungal groups [13,17–19]; and a wood-species classifier's attention has been confirmed to align with anatomically meaningful grain and pore patterns rather than incidental surface artefacts [14]. Applying the same discipline here shows that the present CNN is fully biologically grounded for four of six species reviewed (compass, lion's mane, mauve stinger and moon jellyfish) and largely so for a fifth (barrel jellyfish) (Figure 6); blue jellyfish is the one documented systematic exception, and the watermark and background-water cues identified here give a concrete, testable target (e.g., re-cropping or watermark-masking augmentation, or additional blue-jellyfish images without embedded captions) for the next iteration of this pipeline, rather than a vague call for “more data.”

The practical stakes of this validation gap are not symmetric across species. A model that occasionally confuses barrel and blue jellyfish, or blue and lion's mane jellyfish, is of limited public-safety consequence when the confused species carry broadly similar (mild-to-moderate) sting severity. It would be of considerably greater consequence if a comparable shortcut caused a medically significant species — lion's mane or mauve stinger jellyfish, both associated with painful and occasionally severe stings, and capable of producing skin lesions that persist for weeks [4,5] — to be confused with the essentially harmless moon jellyfish, since that specific error pattern could lead a beachgoer or citizen-science volunteer to under-react to a genuine hazard, particularly given that even correctly recognised stings are managed with first-aid measures of only low-certainty evidence [6]. In the present results, the CNN made no error of that kind on the held-out set, and moon jellyfish, lion's mane jellyfish and mauve stinger jellyfish were each recalled with high accuracy; but a sample of 22 test images per species is too small to rule out this failure mode at the low error rates that a deployed public-facing tool would need. This is precisely the scenario that has motivated jellyfish-specific monitoring tools such as Jellytoring and its successor to be framed explicitly around citizen-science reliability rather than accuracy alone [9,10], and it reinforces why the broader literature on jellyfish–human interactions and on ecosystem-based management treats species-level identification, not just “jellyfish present/absent” detection, as a public-health and fisheries-management requirement in its own right [1,2,4].

Five limitations temper these conclusions. First, the dataset (900 images, 132 held-out test images) is modest by deep-learning standards, and the two-error CNN result, while encouraging, rests on a small number of held-out examples per species; a larger and geographically broader image set, ideally spanning multiple camera types and observation conditions, would give a more robust estimate of both accuracy and failure modes, and would help distinguish a genuine

visual shortcut from a chance artefact of this particular sample. Second, this study deliberately used a single CNN backbone and no architecture search, in keeping with a biology-first rather than benchmark-first design; this is appropriate for the present question (is the classifier's behavior biologically defensible?) but means the reported accuracy should not be read as an upper bound on what a jellyfish classifier can achieve. Third, and most fundamentally, the biological-alignment judgement in (Table 4) is, by design, a qualitative expert call rather than a quantitative score, because no pixel-level annotated ground-truth mask exists for “where the diagnostic feature is” in each image; this mirrors a broader difficulty in jellyfish biology, where even bloom occurrence itself remains difficult to predict from environmental covariates despite decades of monitoring [3], and argues for future work that pairs Grad-CAM-style explanations with lightweight expert-drawn bounding boxes around diagnostic regions, enabling a quantitative intersection-over-union alignment score of the kind already demonstrated for another vertebrate identification task, rather than a Consistent/Partially-consistent/Inconsistent label alone [16] — a scoring scheme that would, for example, let the barrel-jellyfish image with attention split toward a background edge (Figure 7c) be quantified rather than left as a borderline qualitative call. Fourth, a recent methodological critique of an otherwise closely analogous macrofungal XAI study is directly applicable here: high apparent accuracy and plausible-looking Grad-CAM overlays can still mask non-independence between training and test images (near-duplicate photographs, shared backgrounds, repeated collection or photographic events) and shortcut learning on background or acquisition artefacts rather than true morphology, and grouping images at the level of specimen, photographer or site before splitting is recommended precisely to guard against this [20]; the watermark- and background-driven attention documented here for blue jellyfish (Figure 7a–b) is a concrete instance of exactly the failure mode that critique warns about, and the single comparable background-attention lapses noted here for barrel and moon jellyfish (Figure 7c–d) are milder instances of the same failure mode; with all twenty-four overlays now reviewed (Table 4), the priority for future work is the larger, independence-checked follow-up sample recommended above, rather than further inspection of the present set. Fifth, the Grad-CAM audit itself was drawn exclusively from the four highest-confidence test images per species, so all twenty-four reviewed overlays were, by construction, correctly classified; this introduces a selection bias toward best-case model behaviour, since it says nothing directly about the two lowest-confidence, misclassified blue-jellyfish images or about attention on other borderline predictions, and a classifier that is right for the wrong reason would be expected to be most exposed precisely where confidence is low rather than where it is high. The Consistent-for-four/largely-consistent-for-a-fifth-species result reported here should therefore be read as a lower bound on how often non-morphological cues are used, not an average across the full confidence range; extending the same protocol to a stratified sample that deliberately includes low-confidence and misclassified test images (including the two blue-jellyfish errors) would give a less optimistic and more complete picture of the CNN's biological grounding, and is a direct priority for future work.

Conclusion

A ResNet50 CNN trained on 900 images of six coastal jellyfish species reached 98.5% test accuracy, far exceeding a raw-pixel logistic-regression baseline (58.3%), and its two residual test-set errors both implicated a single species (blue jellyfish). Direct biologist inspection of all twenty-four Grad-CAM overlays, checked against an independently curated taxonomic feature map, confirmed that the CNN's attention tracked genuine diagnostic morphology for lion's mane, compass, mauve stinger and moon jellyfish, and largely did so for barrel jellyfish, but that half of the reviewed blue-jellyfish overlays showed attention drawn to background water texture or a source-image watermark rather than the animal — a concrete, non-morphological failure mode that the accuracy figure alone did not reveal. High classification accuracy on its own is therefore not sufficient evidence that an automated jellyfish identification tool is safe to deploy in citizen-science or public-safety settings: the biological stakes of a misidentification depend on which two species are confused, and are highest when a medically significant species could be mistaken for a benign one. We recommend that future jellyfish-identification systems report, alongside accuracy, a species-by-species explainability audit of the kind performed here — comparing model attention, image by image, against ground-truth taxonomic and ecological criteria fixed before training — and that this audit specifically flag any species pair whose confusion would carry outsized medical or ecological consequences, as well as any image-level reliance on watermarks, borders or background artefacts of the kind documented here for blue jellyfish.

Acknowledgments

The authors thank colleagues in the Department of Biology, Faculty of Education, University of Tripoli, and in the Libyan Academy for Postgraduate Studies for helpful discussion during the preparation of this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Bosch-Belmar M, Milisenda G, Basso L, Doyle TK, Leone A, Piraino S. Jellyfish impacts on marine aquaculture and fisheries. *Rev Fish Sci Aquac.* 2021;29(2):242-259.

2. Sagarminaga Y, Piraino S, Lynam CP, Leoni V, Nikolaou A, Jaspers C, Bosch-Belmar M, Fumarola LM, Borja A, Spada E, et al. Management of jellyfish outbreaks to achieve good environmental status. *Front Ocean Sustain*. 2024;2:1449190.
3. Fernández-Álías A, Marcos C, Pérez-Ruzafa A. The unpredictability of scyphozoan jellyfish blooms. *Front Mar Sci*. 2024;11:1349956.
4. Cunha SA, Dinis-Oliveira RJ. Raising awareness on the clinical and forensic aspects of jellyfish stings: a worldwide increasing threat. *Int J Environ Res Public Health*. 2022;19(14):8430.
5. Peng X, Liu KT, Chen JB, Yan ZH, Danso B, Wang MK, Peng ZY, Xiao L. Jellyfish stings: a review of skin symptoms, pathophysiology, and management. *Med Sci Monit*. 2024;30:e944265.
6. Pek JH, Lim SH, Ong GY, Djarv T, Douma M, Welsford M, Charlton NP. First aid treatment of jellyfish stings: a systematic review. *Cureus*. 2025;17(5):e84289.
7. Lawley JW, Gamero-Mora E, Maronna MM, Chiaverano LM, Stampar SN, Hopcroft RR, Collins AG, Morandini AC. The importance of molecular characters when morphological variability hinders diagnosability: systematics of the moon jellyfish genus *Aurelia* (Cnidaria: Scyphozoa). *PeerJ*. 2021;9:e11954.
8. Gauci A, Deidun A, Abela J. Automating jellyfish species recognition through faster region-based convolution neural networks. *Appl Sci*. 2020;10(22):8257.
9. Martin-Abadal M, Ruiz-Frau A, Hinz H, Gonzalez-Cid Y. Jellytoring: real-time jellyfish monitoring based on deep learning object detection. *Sensors*. 2020;20(6):1708.
10. Ruiz-Frau A, Martin-Abadal M, Jennings CL, Gonzalez-Cid Y, Hinz H. The potential of Jellytoring 2.0 smart tool as a global jellyfish monitoring platform. *Ecol Evol*. 2022;12(11):e9472.
11. Hossen MS, Saiduzzaman M, Shaha P, Nasir MK. Jellyfish species identification: a CNN based artificial neural network approach. *arXiv [Preprint]*. 2025 Jul 15:arXiv:2507.11116.
12. Suara S, Jha A, Sinha P, Sekh AA. Is Grad-CAM explainable in medical images? *arXiv [Preprint]*. 2023 Jul 19:arXiv:2307.10506.
13. Korkmaz AF, Ekinci F, Altaş Ş, Kumru E, Güzel MS, Akata I. A deep learning and explainable AI-based approach for the classification of Discomycetes species. *Biology*. 2025;14(6):719.
14. Zannat A, Islam MS, Zaman MS, Saif S. Indigenous wood species classification using a multi-stage deep learning with grad-CAM explainability and an ensemble technique for Northern Bangladesh. *PLoS One*. 2025;20(7):e0328102.
15. Zeng G. Invariance properties and evaluation metrics derived from the confusion matrix in multiclass classification. *Mathematics*. 2025;13(16):2609.
16. Kumar SVS, Kondaveeti HK. Towards transparency in AI: explainable bird species image classification for ecological research. *Ecol Indic*. 2024;169:112886.
17. Kumru E, Korkmaz AF, Ekinci F, Aydoğan A, Güzel MS, Akata I. Deep ensemble learning and explainable AI for multi-class classification of Earthstar fungal species. *Biology*. 2025;14(10):1313.
18. Korkmaz AF, Ekinci F, Kumru E, Altaş Ş, Güneş SK, Yalçın AT, Güzel MS, Akata I. Combination ensemble and explainable deep learning framework for high-accuracy classification of wild edible macrofungi. *Biology*. 2025;14(12):1644.
19. Kumru E, Korkmaz AF, Ekinci F, Altaş Ş, Güneş SK, Yalçın AT, et al. Explainable convolutional neural network architectures for high-performance taxonomic classification of gasteroid macrofungi. *Sci Rep*. 2025;15:40196.
20. Pastore EP. Comment on "Korkmaz et al. A Deep Learning and Explainable AI-Based Approach for the Classification of Discomycetes Species. *Biology* 2025, 14, 719." *Biology*. 2026;15(2):106.