

Biodiversity Image Quality Metadata Augments Convolutional Neural Network Classification of Fish Species

Jeremy Leipzig¹[0000-0001-7224-9620], Yasin Bakis²[0000-0001-6144-9440],
Xiaojun Wang²[0000-0002-2995-9050], Mohannad Elhamod³[0000-0002-2383-947X],
Kelly Diamond^{4,5}[0000-0001-8639-6795], Wasila Dahdul⁶[0000-0003-3162-7490],
Anuj Karpatne³[0000-0003-1647-3534], Murat Maga^{4,5}[0000-0002-7921-9018],
Paula Mabee⁷[0000-0002-8455-3213], Henry L. Bart Jr.²[0000-0002-5662-9444],
Jane Greenberg¹[0000-0001-7819-5360]

¹ Metadata Research Center, College of Computing and Informatics, Drexel University,
Philadelphia, PA 19104 USA jn147@drexel.edu

² Tulane University Biodiversity Research Institute, 3705 Main St., Building A3, Belle Chasse,
LA 70037 USA

³ Department of Computer Science, Virginia Tech, Rice Hall Information Technology
Engineering Building, 22904, 85 Engineer's Way Charlottesville, VA 22903

⁴ Division of Craniofacial Medicine, Department of Pediatrics, University of Washington,
Seattle, Washington

⁵ Center for Development Biology and Regenerative Medicine, Seattle Children's Hospital,
Seattle, Washington

⁶ UC Irvine Libraries, Department of Digital Services

⁷ National Ecological Observatory Network Program, Battelle, Boulder, Colorado

Abstract. Biodiversity image repositories are crucial sources of training data for machine learning approaches to biological research. Metadata, specifically metadata about object quality, is putatively an important prerequisite to selecting sample subsets for these experiments. This study demonstrates the importance of *image quality metadata* to a species classification experiment involving a corpus of 1935 fish specimen images which were annotated with 22 metadata quality properties. A small subset of high quality images produced an F1 accuracy of 0.41 compared to 0.35 for a taxonomically matched subset of low quality images when used by a convolutional neural network approach to species identification. Using the full corpus of images revealed that image quality differed between correctly classified and misclassified images. We found the visibility of all anatomical features was the most important quality feature for classification accuracy. We suggest biodiversity image repositories consider adopting a minimal set of image quality metadata to support future machine learning projects.

Keywords: image classification, convolutional neural networks, image metadata, quality metadata

1 Introduction

1.1 Quality Metadata for Species Image Repositories

The extensive growth in open science repositories, and, in particular, the underlying application of rich metadata has potential value for data mining, machine learning and deep learning (ML/DL). Metadata has historically been used for machine learning and automatic document classification [1] and there is growing attention to the role of metadata in reproducible research pipelines [2,3]. Less common, but of paramount importance is metadata that denotes the quality of the object being represented. Metadata addressing quality control characteristics of data can support the data cleaning steps common to virtually all ML/DL analyses. In fact, computer vision is one area of particular interest where quality-specific metadata can play an important role in the selection of training, validation and test image sets. For example, Ellen et. al found the use of context metadata, consisting of hydrographic, geotemporal, and geometric data, collected or extracted from plankton images improved the accuracy of a convolutional neural network (CNN) classifier [4]. Tang found a 7% gain in mean average precision after including GPS coordinates in a general image classification task [5]. These studies shed light on an important area of metadata research that has broad implications for leveraging collections of digital images across nearly every scientific discipline.

One area of particular interest is specimen images, particularly given their value as a data source for species identification and morphological study^{6,7}. The research presented in this paper, addresses this topic in the context of a NSF supported Harnessing the Data Revolution (HDR) project, Biology-Guided Neural Networks for Discovering Phenotypic Traits (BGNN). A team of information and computer scientists, biologists, and image experts are collaborating to develop a novel set of artificial neural networks (ANNs) for classifying fish species and extracting data on fish external morphological features from images of fish specimens. Unlike genomic data, specimen trait data is largely unstructured and not machine readable. The paucity of trait data for many groups of organisms has led to efforts to apply neural network-based

classification and morphological analysis to the extensive store of existing species photographic images to automatically extract trait data in an unsupervised manner. The product of these efforts, the focus of BGNN, should improve our ability to derive phenotype data from digital analogs of specimens. Metadata is recognized as an important aspect of this research, particularly in the selection of images for ML/DL.

The research presented in this paper demonstrates the value of image quality metadata for BGNN, and presents the results of baseline research on automatic specimen classification, depending on metadata quality metrics. The sections that follow provide contextual background; describe the research sample of digital images and the scheme of 22 image metadata attributes, and present the results of our baseline study. The conclusion highlights the importance of this work, specifically metadata that indicates image quality attributes important for selecting high quality images for, in our case, biology-guided training of neural networks to extract phenotypic trait data from diverse assortments of fish specimen images.

1.2 Image Metadata Content Description and Quality

Over the last few decades, national and international support has targeted the digitization of analog specimen collections. Examples of key programs include the U.S. National Science Foundation's (NSF) Advancing Digitization of Biodiversity Collections (ADBC) program and the European Union's (EU) Distributed System of Scientific Collections (DiSSCo) project. Digital images produced by these initiatives have further supported the creation of large-scale specimen images repositories (e.g., The Phthiraptera Database <http://phthiraptera.info/sid>; Burke Museum Paleontology Database, <https://www.burkemuseum.org/>; Morphosource, <https://www.morphosource.org/>), where the images and their associated metadata are freely available for research and education purposes. The search, discovery, and use of images from these and other collections is highly dependent on the metadata associated with the image objects. A number of different metadata schemes are used across these projects, due, in part to the diversity of scope of the projects and the make-up of the project teams.

The MODAL framework [8, 9] provides a mechanism for understanding the range and diversity of metadata standards used to describe digital specimens, as their domain foci (general to specific), including the extent of support for image quality. Metadata applicable to analog images, which is our immediate focus, falls broadly into two classes, 1.) descriptive metadata that is humanly generated by a curator; and 2.) technical metadata that is automatically generated by technology used to capture the image. The Dublin Core Metadata Standard and the Darwin Core Standard are among the most popular schemes used for describing digital specimens. These metadata standards are more descriptive covering *specimen name* and *topical* and *geo-spatial aspects*, with limited coverage of technical aspects. On the technical end, Dublin Core’s *dc:type* or *dc:description* properties or Darwin Core’s *dwc:dynamic* property may be used to record information impacting object quality, but these properties are still limited coverage of quality measures. The recently developed Audubon Core metadata standard for multimedia objects includes a couple of metadata properties classed under the Service Access Point Vocabulary that support some aspects of quality assessment. For example, *Image-Height* and *Image-Width* can give an indication of quality, but without knowledge of the ideal height and width it is difficult to make a clear assessment of image quality.

A richer set of metadata properties giving insight into image quality is found in the more technically oriented metadata standards identified in **Table 1**. The example metadata properties need to be measured against parameters that define image quality to be of value. The Digital Imaging and Communications in Medicine (DICOM) standard is an exception; this extensive scheme with over 200 metadata properties includes *imageQuality* as a metadata property, and supports scoring on a scale of 1 to 100.

Table 1. Example Technical and Biomedical Metadata Standards

Metadata Standard	Primary Focus	Metadata quality property
Preservation Metadata: Implementation Strategies	Long term preservation	Fixity

Exchangeable image file format (EXIF)	Image formats	X/Y dimensions, compression, color space
DICOM ¹⁰	Medical imaging	imageQuality (1-100)

Semantically-oriented ontologies and even controlled vocabularies, can also be used to indicate value. **Table 2** identifies two ontologies, and example semantics, that indicate image quality, and **Figure 1** illustrates the class-hierarchy where the entity *Thing*, representing an anatomical feature and aspect of color (e.g., hue and saturation) can encode object quality.

Table 2. Selected Ontologies

Ontology	Primary Focus	Semantics/metadata values
Biomedical Image Ontology (BIM) ¹¹	Biomedical images	Image filters, ImagePreProcessing, ImagePostProcessing
OntoNeuro-Base ¹²	Neuro imaging	structure of interest, orientation, segmentation result

Class Hierarchy

```

Thing
+ continuant
+ specifically dependent continuant
+ quality
+ physical object quality
+ physical quality
+ radiation quality
+ electromagnetic \(EM\) radiation quality
+ optical quality
+ chromatic property
- color hue
- color saturation
- low saturation
- high saturation

```

Fig. 1. Phenotype And Trait Ontology: Example of Quality relating to color saturation and other factors.

Overall, the schemes identified here range in their focus on content description, supporting discovery with metadata properties such as a specimen's scientific name, geographic location, provenance, and collector's name, to technical aspects that aid in access and can help to determine aspects of quality, particularly when the parameters of what determines quality are known. Our assessment finds there does not yet exist a targeted metadata standard that captures the types of object - in our case, specimen image quality - necessary for our work in BGNN.

The need for adequate metadata to support our efforts to aggregate a sufficient quantity and variety of teleost fish images for experiments in species classification, trait segmentation and ultimately automated phenotyping in a supervised machine learning context, led us to examine a number of large curated image repositories. We initially explored using images from the iDigBio Portal (<https://www.idigbio.org/portal>) the national repository of NSF's Advancing Digitization of Biodiversity Collections program, a 10-year effort to digitize data on specimens in U.S. biodiversity collections. The iDigBio Portal reportedly hosts 160,000 media files on fishes. However, searching the images, visualizing them online and obtaining copies presented challenges. A random check of roughly 5,000 images subsampled from 100,000 media files representing a number of families and genera of teleost fishes detected misidentified fish species and non-fish specimens such as plants and crustaceans. Moreover, few of the images had metadata for filtering based on quality and the metadata were incomplete in most of these instances.

We ultimately settled on using images of fish specimens from the Great Lakes Invasives Network (GLIN), one of the NSF ADBC Thematic Collections Networks. The images used in this study were obtained from the Illinois Natural History Survey Fish Collection, one of six fish collections that participated in the GLIN project.

While the metadata associated with the GLIN collections is not extensive and does not indicate image quality, the overall quality of the images was useful to serve the needs of the project, provided we gathered basic metadata to confirm the quality of individual images. Conversely image metadata may be rich and follow a standard, but that standard

may not include aspects of image quality relevant to machine learning. We found image quality varied substantially, and perhaps more importantly, unevenly with respect to taxa within image repositories (Figure 3), which may belie both individual variation in photography and batch effects associated with submitters. This challenge, ultimately, led BGNN research to identify a set metadata properties to track image specimen quality, and the analysis demonstrating the value of these properties.

1.3 Quality Metadata for Machine Learning

The challenges we faced obtaining images of fish specimens from image repositories for use in the BGNN Project have implications across neural networks using images. The process we undertook to determine which images would be useful to us and manually annotated facets of image quality affect species classification accuracy. This manual approach contrasts with the automated image quality assessment (IQA) which is the focus of most machine learning image quality work.

IQA is an established and active area of research within computer vision. IQA research is concerned with the automated assessment of image quality for perception by both humans and machines, and the creation of robust image classification, segmentation, or detection models that are resilient to low-quality images. Reference-based IQA techniques (full-reference or FR) attempt to quantify the effects of distortions such as those imposed by compression or other forms of blur, noise, or loss of contrast. No-reference IQA algorithms have no access to high-quality reference images during inference [13]. Both FR and NR approaches rely on ratings from several human observers to establish subjective measures for training. Databases such as the Tampere Image Database [14] serve as fixed gold standard references for these attempts at quantifying and modeling image quality.

Unlike IQA, domain-specific annotated quality assessment is conducted by trained annotators for the purposes of grading key relevant features. For example, in fishes the ability to see and count features like scales and ray-fins is crucial to species identification. This intense work, captured in a set of metadata properties, informed our research goals.

Purpose & Goals. The overall aim of our research was to examine the importance of *image quality metadata* for species classification. Our goals were the following:

1. Determine if the annotated image quality affected classification accuracy
2. Determine which specific quality annotations were most important for classification accuracy
3. Make recommendations for future image quality metadata in large image repositories

2 Methods

To address the above goals, we conducted an empirical analysis that involved the following computational steps. Except where noted, analyses were performed using R and Python. Raw data is located at <https://bgnn.org/iqm> and reproducible source code is available at <https://github.com/orgs/hdr-bgnn/iqm>.

2.1 Sample.

The dataset used for this study comprises 23,807 digital images of fish specimens from the Illinois Natural History Survey (INHS) Fish Collection that were produced for the Great Lakes Invasives Network Project [15]. After checking the images for file duplications, errors with image or file formats, institution code, catalog numbers and suffixes to file names, the images were transferred to a file server from which they could be shared with other researchers in the BGNN project. Specimen collection information (occurrence records) for the images were gathered from FishNet2 [16] and the scientific names were updated using Eschmeyer's Catalog of Fishes [17].

Based on our experience with other fish specimen image datasets available online, we defined a set of metadata properties to record image quality for the digitized fish specimens (**Table 3**). The set of properties

forms a metadata scheme for capturing image quality, and is based on the expertise of members at Tulane University's Biodiversity Research Institute, with feedback from members of the Metadata Research Center, Drexel University. Team members included informaticians, fish experts, and data entry technicians, who defined and refined the metadata scheme over a period of several months, as they processed specimen images. The scheme includes 22 metadata properties, requiring the content-value of a categorical concept, free text, a Boolean operator, or a score. A web-based form, and an underlying SQL-based database help to expedite capturing the metadata content (Figure 2). The main purpose of capturing this metadata is to derive a set of quality features (metadata properties) for filtering and retrieving digitized specimens and determining how the quality features impact image processing and machine learning.

The image quality score used in this study is an integer value between 1 to 10 that represents the evaluator's gestalt opinion of the quality of the image and its usefulness for further analysis. The evaluator assigns the score to an image after answering 21 quality-related questions about the image. However, the overall image quality score is independent of the other quality assessment questions and doesn't integrate data from responses to any of these questions.

A score of 8-10 would be for images that are good to excellent, a score of 5-7 would be for images that have some issues but may be usable and a score of 1-4 would be for images that have major problems and are unusable.

2.2 Descriptive statistical analysis of quality

A basic exploratory data analysis was performed on quality metrics. Quality averages by taxonomic groups (genus and species) were examined in order to understand potential biases.

2.3 Implementation of a CNN-based classification pipeline

A convolutional neural network image classification pipeline was developed using PyTorch [18] with Torchvision [19] extensions. Genera (genus groups) and species (genus + specific epithet combinations)

were trained and inferred simultaneously using a novel multi-classifier model, called a Hierarchy-Guided Neural Network (in submission). Several hyperparameters, including learning rate, regularization lambda, early stopping patience were tuned prior to this quality analysis.

2.4 Classification accuracy using high vs low quality subsets

Using the composite median image_quality score of 8, we divided the data set into low-quality and high-quality subsets. Some species are inherently more visually similar to others, so in a classification scenario, an unbalanced distribution of taxa would confound our aim of measuring the isolated effect of image quality. To address this we sampled images based on equal distributions of individuals by species (*Esox americanus*, *Gambusia affinis*, *Lepomis gibbosus*, *Lepomis humilis*, *Lepomis macrochirus*, *Lepomis megalotis*, *Lepomis cyanellus*, *Notropis atherinoides*, *Notropis blennioides*, *Notropis buccata*, *Notropis hudsonius*, *Notropis stramineus*, *Notropis volucellus*, *Noturus gyrinus*, and *Phenacobius mirabilis*) totaling 106 individuals both high and low quality subsets.

2.5 Quality features distinguishing correctly and incorrectly identified species

Using a dataset of 1703 quality annotated images with 20 or more individuals per species (in order to achieve enough training and test data), the holdout test set of 341 images (17 species over 8 Genera) was then divided into correctly classified and misclassified images. Quality features between these two subsets were compared, and pairs of correct/incorrect within species were examined closely.

Image Metadata Entry



Size:

Image Name:

Image Size: width: px

Image Size: height: px

1. Is a fish specimen visible in the image? ☒ Yes ☐ No

2. How many fish specimens are in the image?

3. Does the image include a ruler? ☒ Yes ☐ No

4. Does the image include a specimen label? ☒ Yes ☐ No

5. Does the number in the image name (INHS_FISH_9999.jpg) match the number on the label? ☒ Yes ☐ No

6. Is the catalog number on the label numerical or some other format (e.g., a barcode)? ☒ Yes ☐ No

7. Is the specimen label detailed (multiple lines of information) or does it have just a specimen number? ☒ Detailed ☐ Basic

8. Does the image have a color scale bar? ☐ Yes ☒ No

9. Is there a specimen label for each fish specimen in the label? ☒ Yes ☐ No

10. Does the image contain any objects other than fish, ruler, specimen label, color bar? ☐ Yes, like "plant", "rock", ☒ No

11. Do any of the objects in the image overlap or hide any parts of the fish? ☐ Yes ☒ No

12. Use a clock face (click "show/hide clock"), which direction is the fish's head pointing? 1-12:

13. Which surfaces of the specimen are visible (left side visible when head points left)? ☒ Left side only ☐ Right side only ☐ Left and some of top ☐ Left and some of bottom ☐ Right and some of top ☐ Right and some of bottom ☐ Top ☐ Bottom ☐ the view is more complicated

14. Is the body of the fish specimen straight or curved? ☒ straight ☐ curved

15. Are any parts of the fish cropped from view? ☐ Yes ☒ No

16. Are all parts (including fins) of the fish visible? ☒ Yes ☐ No

17. Are any of the fins lying down or folded oddly? ☐ Yes ☒ No

18. Rate the brightness of the fish specimen. ☐ dark ☐ bright ☒ normal

19. Is the background of the image uniform? ☒ Yes ☐ No

20. Is the fish specimen in focus? ☒ Yes ☐ No

21. Does the specimen have any color issues? ☐ Contrast ☐ Saturation ☐ Shifted colors ☒ No issues ☐ Others :

22. Rate overall image quality from 1-10:

Fig. 2. Web form for quality metadata entry

At the time of this publication, metadata annotations indicating quality have been created for a total of 1935 images. Table 3, column 1, lists the metadata properties that form the bases of quality assessment; column 2 and 3, follow with each metadata property's data type, and provides a brief description.

Table 3. Quality Metadata Fields

Metadata: Quality property	Data type	Description
1. if_fish	boolean	Whether or not fish are in the image.
2. fish_number	integer	Number of fish in the image.
3. if_ruler	boolean	Whether or not a ruler in the image.
4. if_label	boolean	Whether or not a label in the image.

5. if_label_name_correct	boolean	Whether or not the image name matches the number on the label.
6. if_label_catalog_number_correct	boolean	Whether or not the catalog number is correct.
7. label_detailed	categorical	Whether or not the label is detailed and has full information.
8. if_colorbar	boolean	Whether or not a colorbar is in the image.
9. if_each_fish_label	boolean	Whether or not each fish specimen has a label.
10. non_specimen_objects	freetext	Identify objects in image that are not fish specimens.
11. if_overlapping	boolean	Whether or not objects in the image overlap or hide any parts of the fish.
12. specimen_angled	categorical (1-12 based on clock face)	Number on a clock face the fish's head is pointing to.
13. specimen_view	categorical	The surface(s) of the specimen

		visible (left side visible when head points left)
14. if_curved	categorical	Whether or not the body of the fish specimen is curved or straight.
15. if_missing_parts	boolean	Whether or not any parts of the fish are cropped from view.
16. if_parts_visible	boolean	Whether or not all parts of the fish (including fins) are visible.
17. if_fins_folded	boolean	Whether or not any of the fins are folded.
18. brightness	categorical	The brightness of the fish specimen.(dark, bright, normal)
19. if_background_uniform	boolean	Whether or not the background of the image is uniform.
20. if_focus	boolean	Whether or not the fish specimen is in focus.
21. color_issues	categorical	Identify color-related issues with the specimen (e.g.

		Contrast, Saturation etc.)
22. image_quality	Integer (1-10)	The overall quality of the image.

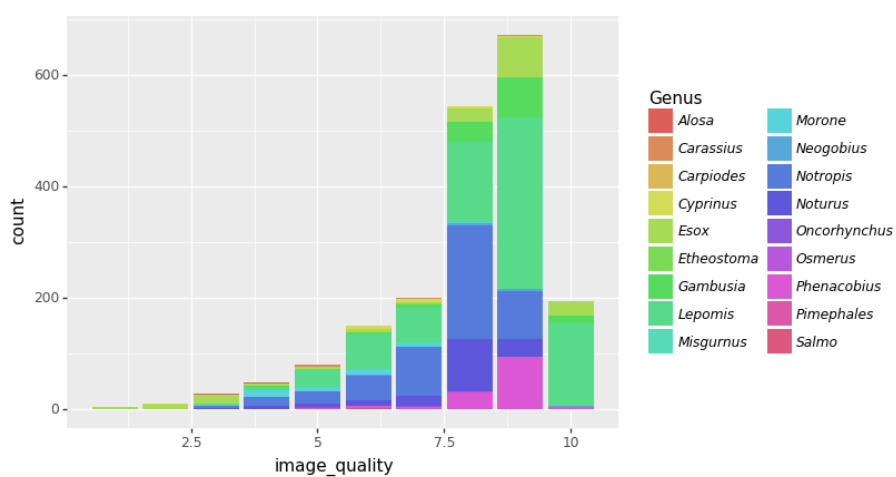


Fig. 3. A histogram of manually annotated image quality scores across the 18 genera

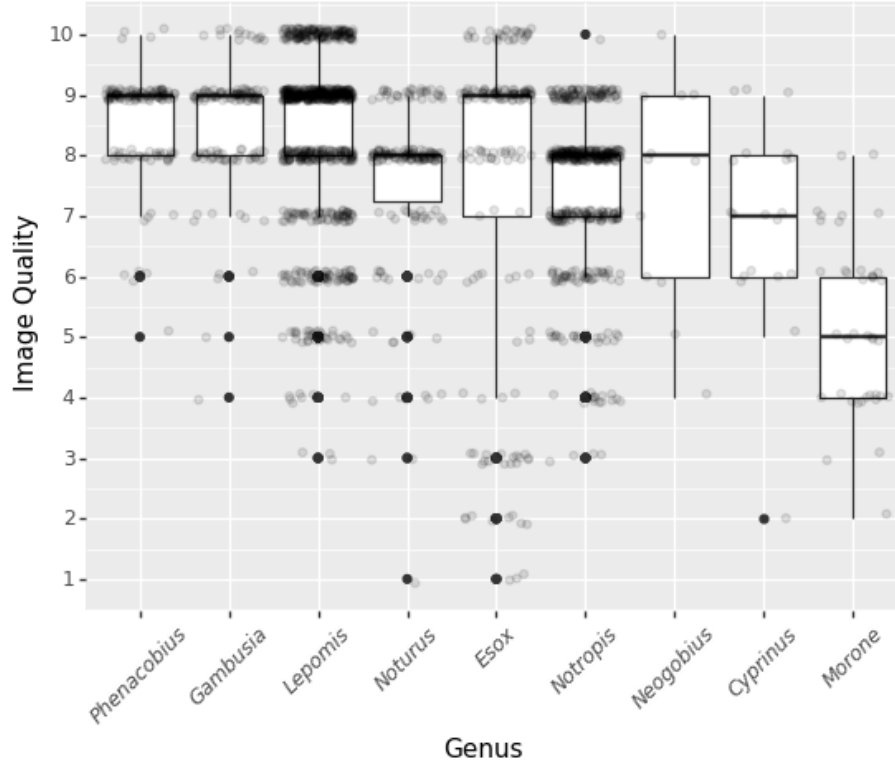


Fig. 4. Quality scores by genera with more than 10 individuals clearly shows variability within and between genera

In order to prepare images for classification, Advanced Normalization Tools in R (ANTsR)[20] was used to subtract background shading, rulers, and labels using a supervised segmentation with a training set developed in 3D Slicer [21]. This "fish mask" was used to isolate fish images for downstream steps.

Early attempts at classification using a subset of 53 species belonging to 13 genera, with each species having 50 images, produced a mean F_1 score ($2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$) of 0.757 (± 0.017 SD) for species classification and 0.910 for genus classification.

Our analysis focused on comparing low and high-quality images that were roughly balanced by genus and species composition, in order to

control for the effect of inherent differences in identification difficulty that vary among taxa. We noted that image quality varied non-randomly among species (Figure 3), perhaps due to batch effects as well as anatomic differences between fish taxa that affect photographic fidelity.



Fig. 5. Examples of very low (1) and very high (10) quality images of *Esox americanus*

3 Results

3.1 Low/High Subset Comparison

A t-test of F1 scores generated by several runs on the small balanced high and low quality subsets showed a small but significant difference in accuracy (0.41 vs 0.35, p -val=0.031) (Figure 6).

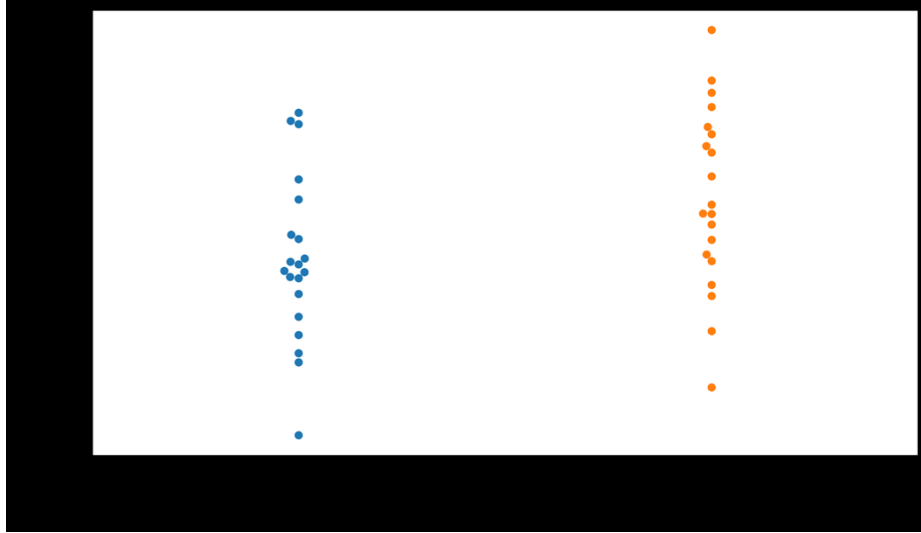


Fig. 6. F1 test score across 19 trials on genus classification using low quality (mean 0.35) and high quality (mean 0.41). Using high quality images produced better F_1 scores (0.41 vs 0.35, $pval=0.031$).

3.2 Quality by Classification Outcome

Here we compared correctly classified vs misclassified images using a test set of 341 images (278 correctly classified and 63 misclassified). A confusion matrix (Figure 6) shows that most misclassifications occur between species within the same genus, although these misclassifications are not symmetric (e.g. 53% of *Notropis blennius* images were misclassified as *Notropis atherinoides*, while no *Notropis atherinoides* were misclassified as *Notropis blennius*).

Comparing the means of quality scores between correctly classified images reveals five quality features correlated with classification accuracy: *if_curved* (0.0179 vs 0.0634), *if_parts_visible* (0.8669 vs 0.8413), *if_overlapping* (0.1043 vs 0.1270), and *image_quality* (8.230 vs 8.079), and two negatively correlated: *if_background_uniform* (0.6151 vs 0.6984), and *if_fins_folded* (0.046512 vs. 0.025641). While *image_quality* is the strongest variable, a logistic regression which includes all features except *image_quality* (to avoid collinearity), reveals *if_parts_visible* ($p\text{-val} = 0.0001$) as the sole significant covariate.

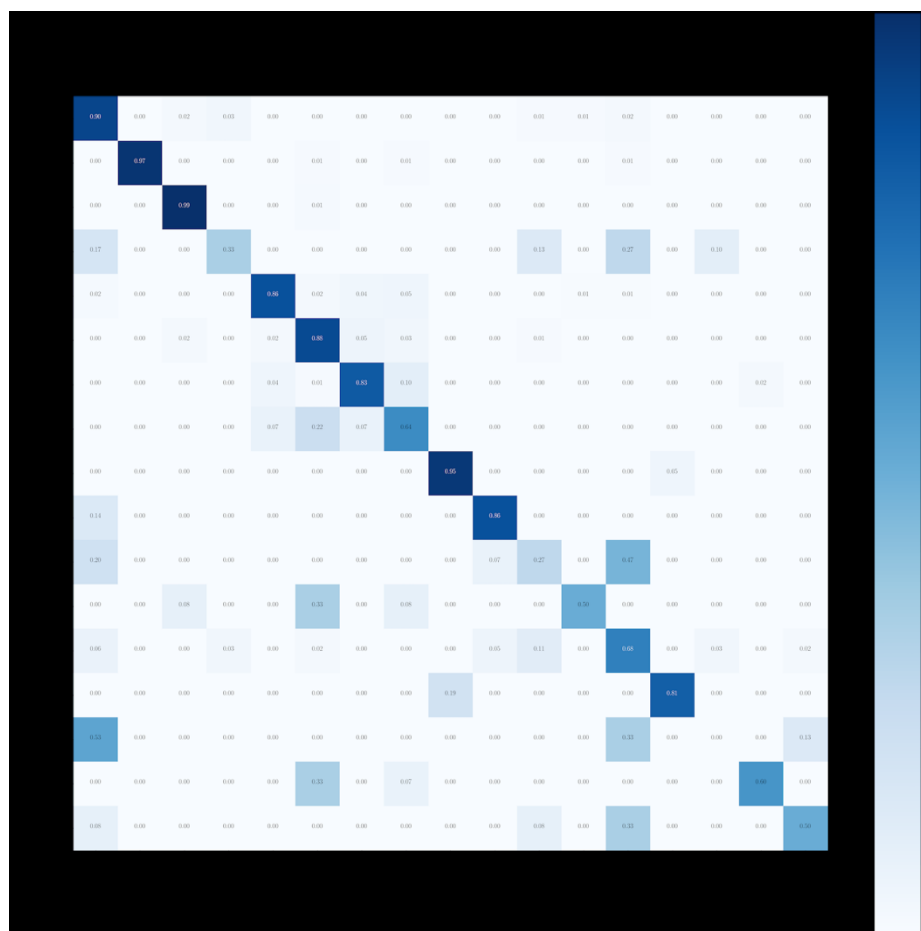


Fig. 7. A confusion matrix of 341 species classifications. The diagonal represents correct classifications. The y-axis represents true labels, and x-axis are predictions.

classified	correct	misclassified
probability of true label	0.837085	0.105787
if_ruler	1.000000	1.000000
if_colorbar	0.003597	0.000000
if_label	1.000000	1.000000
if_overlapping	0.104317	0.126984
if_curved	0.017986	0.063492
if_missing_parts	0.010791	0.015873
if_blur	0.000000	0.000000
if_color_issue	0.000000	0.000000
specimen_angled	8.830935	8.841270
if_background_uniform	0.615108	0.698413
if_parts_visible	0.866906	0.841270
image_quality	8.230216	8.079365
if_fins_folded	0.046512	0.025641

Table 3. Mean of quality features by correctly classified and misclassified images

4 Discussion

In this paper we show that image quality measures do affect classification accuracy. This was demonstrated using two approaches - a dichotomous split of the image corpus using the manually-annotated `image_quality` metric, and a comparison of correctly and incorrectly classified images from the entire quality-annotated data set. Our abilities to discern the importance of quality are hampered by three factors: 1) a relative paucity of low-quality images in our dataset, 2) the nature of classification - some fish are simply more similar to their brethren - but we have attempted to control for this where possible, and 3) some taxa are inherently more difficult to position or illuminate for photography. Our results lead to a number of recommendations for assessing quality of images from biodiversity image repositories to support machine learning analyses.

The `quality_score` assigned by curators, while based on a rubric, does lend itself to some inter-rater error. We surmised that a composite metric of the binary quality items (e.g. `if_curved`, `if_fins_folded`, etc..) could represent a more objective score, and explored this, but it ultimately did not prove substantially better than "image_quality". The quality scores generated by our curators included some that are strictly technical (blur, color issues), those that would apply to any biodiversity catalog ("`if_parts_visible`") and those that are specific to fish (e.g. "`if_fins_folded`"). We contend that all three types of quality (technical, biospecimen, taxon-specific) are important to include for biorepositories. The automated measurement of technical image quality, and possible higher-level judgments, can help accelerate the collection of this metadata. Metadata librarians may also find it useful to distinguish local and global image quality characteristics [22, 23] from semantic quality features. Local features might include fin-level textures that would indicate lepidotrichia, global features such as large segmented areas as well as basic image characteristics such as color and shape. These characteristics are logically distinct from semantic quality

judgments made by the curators in this project ("folded fins"/"label obstruction"), though automated semantic quality annotations are within the capabilities of neural networks.

Another metadata property that we have mentioned tangentially is provenance, particularly because of the batch effects introduced by disparate labs collecting and photographing specimens with different settings and equipment. Batch effects are a huge issue in all repositories and data coordinating centers, and due to the geographically localized nature of certain species and associated labs, this can sometimes present a "complete separation" problem, where controlling for the random effects of collection centers becomes impossible. This would suggest each center be encouraged to collect certain common species to serve as controls for normalization. Both provenance and quality are essential for large image repositories to address bias and confounds for downstream analyses.

Overall, observed that machine learning is hampered by its dependence on classification accuracy instead of more direct intermediate measures, for example, the number of features detected. Though this is an active area of research within the deep learning community, species detection will continue to present some confounds because of inherent heterogeneity discussed above. Certainly, one consideration specimen curators should be aware of is whether the intended use of the biorepository images is to classify real-world specimens. Certain types of low-quality images may serve to augment robustness in computer vision - a technique called "noise injection". These uses suggest annotating for quality, rather than simply culling, is a preferable strategy. Quality metadata to aid robustness and generalizability in machine learning, rather than a narrow focus on pristine specimens, is an open area for future work.

5 Conclusion

The main objective of this research was to determine if annotated image quality metadata impacted generic and species-level classification accuracy of a convolutional neural network. We conducted an empirical

analysis to examine which specific quality annotations were most important for classification accuracy. We worked with a set of 22 metadata properties developed at Tulane University to evaluate the quality of large caches of 2D fish specimen images. Our key finding was that images with high-quality metadata metrics had significantly higher classification accuracy in convolutional neural network analysis than images with low-quality metadata metrics.

We offer a number of recommendations for assessing the quality of images from biodiversity image repositories to support machine learning analyses. This investigation serves as a baseline study of useful metadata for assessing image quality. The methodology and our approach also serves to inform other research that may examine image quality and impact on classification for other fishes, other specimens, and even other disciplines where the image is a central object. Overall the research conducted serves the needs of the BGNN project, and biologically-focused, machine-learning projects generally, for determining whether images for biodiversity specimen image repositories are useful for higher-level analysis.

6 Acknowledgements

We thank Tulane University Technology Services for assistance setting up servers for hosting images and sharing them with other BGNN investigators. We thank Justin Mann, Aaron Kern, Jake Blancher for help with metadata form and gathering the metadata used in the project. The fish specimen images used in the project were obtained from Chris Taylor, Curator of Fishes and Crustaceans at the Illinois Natural History Survey (INHS). INHS is one of six fish collections participating in the Great Lakes Invasives Network (GLIN), five of which shared fish specimens images for use in the BGNN project.

7 Data & Code Availability

Raw data is available at bgnn.org/INHS. Reproducible code is available at <https://github.com/orgs/hdr-bgnn/iqm>

8 Funding

Research supported by NSF OAC Office of Advanced Cyberinfrastructure (OAC) #1940233 and #1940322.

9 Competing Interests

The authors have declared that no competing interests exist.

References

1. Hui Han, Giles CL, Manavoglu E, et al (2003) Automatic document metadata extraction using support vector machines. In: 2003 Joint Conference on Digital Libraries, 2003. Proceedings. pp 37–48
2. Schelter S, Boese J-H, Kirschnick J, et al (2017) Automatically tracking metadata and provenance of machine learning experiments. In: Machine Learning Systems workshop at NIPS
3. Leipzig J, Nüst D, Hoyt CT, et al (2020) The role of metadata in reproducible computational research. arXiv [cs.DL]
4. Ellen JS, Graff CA, Ohman MD (2019) Improving plankton image classification using context metadata. *Limnol Oceanogr Methods* 2:17
5. Tang K, Paluri M, Fei-Fei L, et al (2015) Improving image classification with location context. In: Proceedings of the IEEE international conference on computer vision. pp 1008–1016
6. Borges LM, Reis VC, Izbicki R (2020) Schrödinger’s phenotypes: herbarium specimens show two-dimensional images are both good and (not so) bad sources of morphological data. *Methods Ecol Evol*. <https://doi.org/10.1111/2041-210X.13450>
7. Hernández-Serna A, Jiménez-Segura LF (2014) Automatic identification of species with neural networks. *PeerJ* 2:e563
8. Willis C, Greenberg J, White H (2012) Analysis and synthesis of metadata goals for scientific data. *J Am Soc Inf Sci Technol* 63:1505–1520
9. Greenberg J (2005) Understanding Metadata and Metadata Schemes. *Cataloging & Classification Quarterly* 40:17–36
10. Bidgood WD Jr, Horii SC (1992) Introduction to the ACR-NEMA DICOM standard. *Radiographics* 12:345–355
11. Bukhari AC, Nagy ML, Krauthammer M, et al (2015) BIM: an open ontology for the annotation of biomedical images. In: ICBO
12. Temal L, Lando P, Gibaud B, et al (2006) OntoNeuroBase: a multi-layered application ontology in neuroimaging. In: Second Workshop: Formal Ontologies Meet Industry (FOMI 2006)
13. Bosse S, Maniry D, Muller K-R, et al (2018) Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. *IEEE Trans Image Process* 27:206–219

14. Ponomarenko N, Jin L, Ieremeiev O, et al (2015) Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* 30:57–77
15. Great Lakes Invasives Network Home. <http://greatlakesinvasives.org/portal/index.php>. Accessed 2 Aug 2020
16. FishNet2. <http://www.fishnet2.net/>. Accessed 2 Aug 2020
17. Eschmeyer’s Catalog of Fishes. <https://www.calacademy.org/scientists/projects/eschmeyers-catalog-of-fishes>. Accessed 2 Aug 2020
18. Paszke A, Gross S, Massa F, et al (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, et al (eds) *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., pp 8026–8037
19. Marcel S, Rodriguez Y (2010) Torchvision the machine-vision package of torch. In: *Proceedings of the 18th ACM international conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, pp 1485–1488
20. Avants BB, Kandel BM, Duda JT, et al (2019) ANTsR: ANTs in R: quantification tools for biomedical images. R package version 0542}
21. Kikinis R, Pieper SD, Vosburgh KG (2014) 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. In: Jolesz FA (ed) *Intraoperative Imaging and Image-Guided Therapy*. Springer New York, New York, NY, pp 277–289
22. Shyu CR, Brodley CE, Kak AC, et al (1998) Local versus global features for content-based image retrieval. In: *Proceedings. IEEE Workshop on Content-Based Access of Image and Video Libraries* (Cat. No.98EX173). pp 30–34
23. Vogel J, Schwaninger A, Wallraven C (2006) Categorization of natural scenes: local vs. global information. *Proc Conf Army Physicians Cent Mediterr Forces*
24. Wicczorek J, Bloom D, Guralnick R, et al (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* 7:e29715