Computational and Systems Biology

Identifying images in the biology literature that are problematic for people with a color-vision deficiency

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint version 1
March 14, 2024 (this version)

Sent for peer review
January 11, 2024

Posted to preprint server
November 30, 2023

Harlan P. Stevens, Carly V. Winegar, Arwen F. Oakley, Stephen R. Piccolo

Department of Biology, Brigham Young University, Provo, UT, USA

Abstract

To help maximize the impact of scientific journal articles, authors must ensure that article figures are accessible to people with color-vision deficiencies. Up to 8% of males and 0.5% of females experience a color-vision deficiency. For deuteranopia, the most common color-vision deficiency, we evaluated images published in biology-oriented research articles between 2012 and 2022. Out of 66,253 images, 56,816 (85.6%) included at least one color contrast that could be problematic for people with moderate-to-severe deuteranopia ("deuteranopes"). However, after informal evaluations, we concluded that spatial distances and within-image labels frequently mitigated potential problems. We systematically reviewed 4,964 images, comparing each against a simulated version that approximates how it appears to deuteranopes. We identified 636 (12.8%) images that would be difficult for deuteranopes to interpret. Although still prevalent, the frequency of this problem has decreased over time. Articles from cell-oriented biology subdisciplines were most likely to be problematic. We used machine-learning algorithms to automate the identification of problematic images. For a hold-out test set of 879 additional images, a convolutional neural network classified images with an area under the receiver operating characteristic curve of 0.89. To enable others to apply this model, we created a Web application where users can upload images, view deuteranopia-simulated versions, and obtain predictions about whether the images are problematic. Such efforts are critical to ensuring the biology literature is interpretable to diverse audiences.

eLife assessment

In this **important** study, the authors manually assessed randomly selected images published in eLife between 2012 and 2020 to determine whether they were accessible for readers with deuteranopia, the most common form of color vision deficiency. They then developed an automated tool designed to classify figures and images as either "friendly" or "unfriendly" for people with deuteranopia. While such a tool could be used by publishers, editors or researchers to monitor accessibility in the research literature, the evidence supporting the tools' utility was **incomplete**. The tool would benefit from training on an expanded dataset that includes different image and figure types from many journals, and using more rigorous approaches when training the tool and assessing performance. The authors also provide code that readers can download and run to test their own images. This may be of most use for testing the tool, as there are already several free, user-friendly recoloring programs that allow users to see how images would look to a person with different forms of color vision deficiency. Automated classifications are of most use for assessing many images, when the user does not have the time or resources to assess each image individually.

Introduction

Most humans have trichromatic vision: they perceive blue, green, and red colors using three types of retinal photoreceptor cells that are sensitive to short, medium, or long wavelengths of light, respectively. Color-vision deficiency (CVD) affects between 2% and 8% of males (depending on ancestry) and approximately 0.5% of females[1]. Congenital CVD is commonly caused by mutations in the genes (or nearby promoter regions) that code for red or green cone photopigments; these genes are proximal to each other on the X chromosome[2].

CVD is divided into categories, the most common being *deutan CVD*, affecting approximately 6% of males of European descent, and *protan CVD*, affecting 2% of males of European descent[1]. Both categories are commonly known as red-green colorblindness. Within each category, CVD is subclassified according to whether individuals are *dichromats*—able to see two primary colors—or *anomalous trichromats*—able to see three primary colors but differently from *normal trichromats*. Anomalous trichromats differ in the degree of severity with which they can distinguish color patterns. Individuals with deuteranopia (“deuteranopes”) or protanopia do not have corresponding green or red cones, respectively[3]. Individuals with deuteranomaly do not have properly functioning green cones, and those with protanomaly do not have properly functioning red cones. People with any of these conditions often see green and red as brown or beige colors. Thus, when images contain shades of green and red—or when either is paired with brown—parts of the image may be indistinguishable. Furthermore, it can be problematic when some pinks or oranges are paired with greens. These issues can lead to incorrect interpretations of figures in scientific journal articles for individuals with CVD.

Efforts have been made to ensure that scientific figures are accessible to people with CVD. For example, researchers have developed algorithms that attempt to recolor images so that people with CVD can more easily interpret them[4–6]. However, these tools are not in wide use, and more work is needed to verify their efficacy in practice. In the meantime, as researchers prepare scientific figures, they can take measures to improve accessibility for people with CVD. For example, they can avoid rainbow color maps that show colors in a gradient; they can use color schemes or color intensities that are CVD friendly[7]; additionally, they can provide labels that

complement information implied by color differences. However, for the millions of images that have already been published in scientific articles, little is known about the frequency with which these images are CVD friendly. The presence or absence of particular color pairings—and distances between them—can be quantified computationally to estimate this frequency. However, a subjective evaluation of individual images is necessary to identify whether color pairings and distances are likely to affect scientific interpretation.

In this paper, we focus on deuteranopia and its subtypes. To estimate the extent to which the biological literature contains images that may be problematic to deuteranopes, we manually reviewed a "training set" of 4,964 images and a "test set" of 1,000 images. These images were published in biology-oriented articles in the *eLife* journal between 2012 and 2022. After identifying images that we deemed most likely to be problematic or not, we used machine-learning algorithms to identify patterns that could discriminate between these two categories of images and thus might be useful for automating the identification of problematic images. If successful, such an algorithm could be used to alert authors, presenters, and publishers that scientific images could be modified to improve visual accessibility and thus make biology more inclusive.

Methods

Image acquisition and summarization

We evaluated images in articles from *eLife*, an open-access journal that publishes research in "all areas of the life sciences and medicine"[8]. Article content from this journal is released under a Creative Commons Attribution license. On June 1, 2022, we downloaded all available images from an Amazon Web Services storage bucket provided by journal staff. We also cloned a GitHub repository that *eLife* provides (*https://github.com/elifesciences/elife-article-xml*); this repository contains text and metadata from all articles published in the journal since its inception. For each article, we parsed the article identifier, digital object identifier, article type, article subject, and publication date; we stored this information in a tab-delimited file. We excluded any article that was not published with the "Research article" type.

For each available image, we identified whether the image was either grayscale or contained colors. For each color image, we calculated a series of metrics to summarize the colors, contrasts, and distances between potentially problematic colors. These metrics have similarities to those used to assess recoloring algorithms, including global luminance error[9], local contrast error[10], and global chromatic diversity[11,12]. Before calculating the metrics, we sought to make the images more comparable to each other and to reduce the computational demands of analyzing the images. We scaled each image to a height of 300 pixels and generated a quantized version with a maximum of 256 colors. For each image, we then created a second version that simulated how a deuteranope would see the image. To facilitate these simulations, we used the *colorspace* package[13] and specified a "severity" value of 0.8. Severity values range between 0 and 1 (with 1 being the most severe). We chose this threshold under the assumption that a mild severity level might not be stringent enough to identify a lack of contrast in the images. However, because many people with deuteranomaly do not have complete deuteranopia, this threshold reflects more moderate cases.

Our approach and rationale for these metrics are described below. In these descriptions, we refer to the quantized, resized images as "original" images and their simulated counterparts as "simulated" images.

- *Mean, pixel-wise color distance between the original and simulated image*. Our rationale was that the most problematic images would show relatively large overall differences between the original and simulated versions. When calculating these differences, we used version

2000 of Hunt's distance[14], which quantifies red/green/blue (RGB) color differences in a three-dimensional space. This metric is symmetric, so the results are unaffected by the order in which the colors were specified; we used the absolute value of these distances.

- *Color-distance ratio between the original and simulated images for the color pair with the largest distance in the original image*. First, we excluded black, white, and gray colors. Second, we calculated the color distance (Hunt's method) between each unique pair of colors in the original image. Third, we calculated the color distance between the colors at the same locations in the simulated image. Fourth, we calculated the ratio between the original distance and the simulated distance. Our rationale was that problematic color pairs would have relatively high contrast (large distances) in the original images and relatively low contrast (small distances) in the simulated images. This approach is similar to that described by Gregor Aisch[15].
- *Number of color pairs that exhibited a high color-distance ratio between the original and simulated images*. This metric is similar to the previous one. However, instead of using the maximum ratio, we counted the number of color pairs with a ratio higher than five; this threshold was used by Gregor Aisch[15]. Our rationale was that even if one color pair did not have an extremely high ratio, the presence of many high-ratio pairs would indicate a potential problem.
- *Proportion of pixels in the original image that used a color from one of the high-ratio color pairs*. Again, using a threshold of five, we identified unique colors among the color-distance pairs and counted the number of pixels in the original image that used any of these colors. Our rationale was that a relatively large number of pixels with potentially problematic colors may make an image as difficult for a deuteranope to interpret as an image with a few extremely low-contrast pixels.
- *Mean Euclidean distance between pixels for high-ratio color pairs*. First, we identified color pairs with a ratio higher than five. For each color pair, we identified pixels in the original image that used the two colors and calculated the Euclidean distance between those pixels in the image's two-dimensional layout. Then, we calculated the mean of these distances. Our rationale was that potentially problematic color pairs close together in an image would be more likely to cause problems than color pairs that are distant within the image.

After calculating these metrics for each available image, we calculated a ranked-based score. First, we assigned a rank to each image based on each of the metrics separately. For the "Mean Euclidean distance between pixels for high-ratio color pairs," relatively large values were given relatively high ranks (indicating that they were less problematic). For the other metrics, relatively small values were given relatively high ranks. Finally, we averaged the ranks to calculate a combined score for each image.

When analyzing images, calculating metrics, and creating figures and tables, we used the R statistical software (version 4.2.1)[16] and the following packages:

- colorspace (version 2.0-3)[13]
- doParallel (1.0.17)[17]
- knitr (1.44)[18]
- magick (2.7.3)[19] - interfaces with the ImageMagick software (6.9.10.23)[20]
- pROC (1.17.0.1)[21]
- spacesXYZ (1.2-1)[22]
- tidyverse (2.0.0)[23]
- xml2 (1.3.3)[24]

Qualitative image evaluation

We manually reviewed images to assess qualitatively whether visual characteristics were likely to be problematic for deuteranopes. Our intent was to establish a reference standard for evaluating the quantitative metrics we had calculated. Initially, we randomly sampled 1,000 images from among those we had downloaded. Two authors of this paper (HPS and AFO) reviewed each of the original (non-quantized, non-resized) images and the corresponding image that was simulated to reflect deuteranopia (severity = 0.8). Neither of these authors has been diagnosed with deuteranopia. This ensured the reviewers could compare the images with and without deuteranopia simulation. To avoid confirmation bias, neither author played a role in defining the quantitative metrics described above. Both authors reviewed the images and recorded observations based on four criteria:

1. Did an image contain shades of red, green, and/or orange that might be problematic for deuteranopes?
2. When an image contained potentially problematic color shades, did the color contrasts negate the potential problem? (The reviewers examined the images in their original and simulated forms when evaluating the contrasts.)
3. When an image contained potentially problematic color shades, did within-image labels mitigate the potential problem?
4. When an image contained potentially problematic color shades, were the colors sufficiently spatially distant from each other so that the colors were unlikely to be problematic?

After discussing a given image, the reviewers recorded a joint conclusion about whether the image was "Definitely problematic," "Probably problematic," "Probably okay," or "Definitely okay." For images that had no visually detectable color, the reviewers recorded "Gray-scale."

After this preliminary phase, we randomly selected an additional 4,000 images and completed the same process. During the review process, we identified 36 cases where multiple versions of the same image had been sampled. We reviewed these versions manually and found that subsequent versions either had imperceptible differences or slight differences in the ways that sub-figures were laid out. None of these changes affected the colors used. Thus, we excluded the duplicate images and retained the earliest version of each image. The resulting 4,964 images constituted a "training set," which we used to evaluate our calculated metrics and to train classification models (see below). Later, we randomly selected an additional 1,000 images, which we used as a "hold-out test set." Again, we excluded duplicate images and those for which different versions were present in the training set and hold-out test set. The same authors (HPS and AFO) performed the manual review process for these images.

Classification analyses

We used classification algorithms to discriminate between images that we had manually labeled as either "Definitely problematic" or "Definitely okay." Although it reduced our sample size, we excluded the "Probably problematic" and "Probably okay" images with the expectation that a smaller but more definitive set of examples would produce a more accurate model. Removing these images reduced our training set to 4,501 images.

First, we evaluated our ability to classify images as "Definitely problematic" or "Definitely okay" based on the five metrics we devised. For this task, we used the following classification algorithms, which are designed for one-dimensional data: Random Forests[25], k-nearest neighbors[26], and logistic regression[27]. We used implementations of these algorithms in scikit-learn (version 1.1.3)[28] with default hyperparameters, other than two exceptions; we used the "liblinear" solver for logistic regression, and we set the "class_weight" hyperparameter to "balanced" for Random Forests and Logistic Regression. For evaluation, we used three iterations of five-fold cross

validation. For the test samples in each fold, we calculated the area under the receiver operating characteristic curve (AUROC)[29,30] using the *yardstick* package (version 1.2.0)[31]. We calculated the median AUROC across the folds and then averaged them across the three iterations.

Second, we evaluated our ability to classify the images as "Definitely problematic" or "Definitely okay" based on the images themselves. We used a convolutional neural network (CNN) because they are capable of handling two-dimensional inputs and to account for spatial patterns and colors within images. To generate the CNN model, we used the Tensorflow (2.10.0) and Keras (2.10.0) frameworks[32,33]. To support transfer learning (described below), we scaled both dimensions of each image to 224. To perform hyperparameter optimization, we again used three iterations of five-fold cross validation (with the same assignments as the earlier classification analysis). Each hyperparameter combination extended a baseline model that had eight, two-dimensional, convolutional layers; each layer used batch normalization and the *relu* activation function. Subsequent layers increased in size, starting with 32 nodes and increasing to 64, 128, 256, 512, and 728. We trained for 30 epochs with an Adam optimization set, a learning rate of 1e-3, and the binary cross-entropy loss function. The output layer used a sigmoid activation function.

In addition to the baseline model, we tested 22 hyperparameter combinations based on the following techniques:

- *Class weighting* - To address class imbalance (most images were "Definitely okay" in the training set), we increased the weight of the minority class ("Definitely problematic") proportionally to its frequency in the training set.
- *Early stopping* - During model training, classification performance on the (internal) validation set is monitored to identify an epoch when the performance is no longer improving or has begun to degrade; the goal of this technique is to find a balance between underfitting and overfitting.
- *Random flipping and rotation* - In an attempt to prevent overfitting, we enabled random, horizontal flipping of training images and data augmentation via differing amounts of random image rotation[34]. We evaluated rotation thresholds of 0.2 and 0.3.
- *Dropout* - Again to prevent overfitting, we asked the model to temporarily remove a subset of neurons from the network. We evaluated dropout rates of 0.2 and 0.5.
- *Transfer learning* - This technique uses a corpus of ancillary images such that model building is informed by patterns observed previously. We evaluated two corpuses: MobileNetV2[35,36] and ResNet50[37]. MobileNetV2 is a 53-layer convolutional neural network trained on more than a million images from the ImageNet database to classify images with objects into 1,000 categories. ResNet50 is a 50-layer convolutional neural network, similarly trained. MobileNetV2 is designed for use on mobile devices, so it is optimized to be lightweight. As such, MobileNetV2 uses 3.4 million parameters, while ResNet50 uses over 25 million trainable parameters. When we applied transfer learning from either ResNet50 or MobileNetV2, we did not use the base model design referred to above. Instead, we added a global pooling function into a dense layer. Because our training dataset was relatively small, adding fewer layers may reduce the risk of overfitting.
- *Fine tuning* - When transfer learning was used, we sometimes performed fine tuning. With this approach, we allowed an initial model to be trained for 30 epochs (unless early stopping was specified). We then fine-tuned the model by unfreezing the ResNet50 (or MobileNetV2) layers and retraining the initial model with a lower learning rate of 1e-5.

To guide (internal) evaluation of the models, we specified the following metrics: true positives, false positives, true negatives, false negatives, accuracy, precision, recall, AUROC, and precision-recall curve. After identifying the best hyperparameter combination via cross validation, we created a hold-out test set, randomly selecting an additional 1,000 images from eLife articles published during the same timeframe as the training set. We did *not* use these images when developing the model or selecting hyperparameters.

Web application

We created a Web application using the *Node.js* framework[38]. The application enables researchers to evaluate uploaded images. First, users upload an image in PNG or JPEG format. The application displays the image alongside a deuteranopia-simulated version of the image. For simulation, we implemented the Machado et al.[39] matrix for deuteranopia in Javascript with a "severity" value of 0.8, the same parameter used in training. If the user requests it, the application uses our CNN model to predict whether the image is likely to be problematic for a deuteranope; the prediction includes a probabilistic score so that users can assess the model's confidence level. To facilitate execution of the CNN within the Web application, we used Tensorflow.js (version 4.0.0)[40].

Code and data availability

The images we used for evaluation and the trained TensorFlow model are stored in an Open Science Framework repository (*https://osf.io/8yrkb*). The code for processing and analyzing the images is available at *https://github.com/srp33/bio_image_colorblindness*. That repository also includes the calculated metrics, cross-validation assignments, results of image curation, and outputs of the classification algorithms. The Web application code is available at *https://github.com/srp33/colorblind_image_tester*.

Results

We downloaded images from research articles published in the eLife journal between 2012 and 2022. Not counting duplicate versions of the same image, we obtained 66,253 images. Of these images, 1,744 (2.6%) were grayscale (no color). Of these images, 56,816 (85.6%) included at least one color pair for which the amount of contrast might be problematic to people with moderate-to-severe deuteranopia ("deuteranopes"). To characterize potentially problematic aspects of each color-based image, we calculated five metrics based on color contrasts and distances; we also compared the color profiles against what deuteranopes might see (see Methods). The distributions of these metrics are depicted in Figures S1-S5.

A brief evaluation suggested that many images with the highest scores (or lowest, as would be the case for the "Mean Euclidean distance between pixels for high-ration color pairs" metric) for these metrics were likely to be problematic for deuteranopes. However, we noted that certain color pairs were more problematic than others, and the use of effective labels and/or spacing between colors often mitigated potential problems. Thus, to better estimate the extent to which images in the biological literature are problematic for deuteranopes, we manually reviewed a sample of 4,964 images and judged whether it would be likely for deuteranopes to recognize the scientific message behind each image. Additional Data File S1 contains a record of these evaluations, along with comments that indicate either problematic aspects of the images or factors that mitigated potential problems. We concluded that 636 (12.8%) of the images were "Definitely problematic," whereas 3,865 of the images (77.9%) were "Definitely okay." The remaining images were grayscale (n = 179), or we were unable to reach a confident conclusion (n = 284). For the images that were "Definitely okay," we visually detected shades of green and red or orange in 2,348 (60.8%) images; however, in nearly all (99.3%) of these cases, we deemed that the contrasts between the shades were sufficient that a deuteranope could interpret the images. Furthermore, distance between the colors and/or labels within the images mitigated potential problems in 54.2% and 48.4% of cases, respectively.

For the images that were either "Definitely okay" or "Definitely problematic," we calculated the percentage of "Definitely problematic" images in a given year or associated with a given biology subdiscipline. The percentage of problematic images declined steadily between 2012 and 2021,

with a modest increase in 2022 (**Figure 1**). The subdisciplines with the highest percentages of problematic images were *Stem Cells and Regenerative Medicine*, *Developmental Biology*, and *Cell Biology* (**Figure 2**).

Despite the benefits of manual review, this process is infeasible on a large scale. Therefore, we evaluate techniques for automating classification. Firstly, we used the five metrics we calculated to classify the training images. We also combined these metrics into a single, ranked-based score for each image.**Figures 3**, S6-S10 illustrate differences for these metrics between the two classes. To quantify their predictive performance, we calculated the area under the receiver operating characteristic curve (AUROC)[29]. Values closer to 1.0 indicate relatively high predictive performance. A value of 0.5 indicates that predictions are no better than random guessing. The best-performing metric on the training set was the number of color pairs that exhibited a high color-distance ratio between the original and simulated images (AUROC: 0.75); the mean, pixel-wise color distance between the original and simulated image performed worse than random guessing (Table S1). Secondly, as an alternative to the combined rank score, we used classification algorithms to make predictions with the five metrics as inputs. In cross validation on the training set, the best-performing algorithm was Logistic Regression, attaining an AUROC of 0.82 (Table S2). Thirdly, we created a convolutional neural network (CNN) to make predictions according to visual and spatial patterns in the images. CNNs are highly configurable and often sensitive to model parameters. Accordingly, we performed multiple iterations of cross validation on the training set and identified a hyperparameter combination that attained an AUROC of 0.93, slightly outperforming other combinations (Table S3).

We manually reviewed a hold-out test set with an additional 1,000 images (Additional Data File S2). After removing images from the hold-out test set that were not “Definitely okay” or “Definitely problematic,” we were left with 879 images to test the classification algorithms. For the Logistic Regression algorithm and CNN, we trained models using the full training set and classified each hold-out image as “Definitely okay” or “Definitely problematic.” Logistic Regression classified the images with an AUROC of 0.82; the CNN classified the images with an AUROC of 0.89 (**Figures 4**-**5**). Upon reviewing the 92 misclassified images and comparing them against our manual annotations, we determined that in 13 cases, the reviewers had missed subtle patterns and that it would be justified to change the labels (Additional Data File S3). For 31 of the misclassified images, we visually identified patterns that might have confused the CNN; however, upon reevaluation, we maintain that the original labels were valid. For the remaining 48 misclassified images, we were unable to identify patterns that seemed likely to have confused the model.

Discussion

Significant prior work has been done to address and improve accessibility for individuals with CVD[41]. This work can be generally categorized into three types of studies: simulation methods, recolorization methods, and estimating the frequency of accessible images. Simulation methods have been developed to better understand how images appear to individuals with CVD. Brettel et al. first simulated CVDs using the long, medium, and short (LMS) colorspace[42]. For dichromacy, the colors in the LMS space are projected onto an axis that corresponds to the non-functional cone cell. Viénot et al. expanded on this work by applying a 3×3 transformation matrix to simulate images in the same LMS space[43]. Machado et al. created matrices to simulate CVDs based on the shift theory of cone cell sensitivity[39,44]. These algorithms allow individuals without CVD to qualitatively test how their images would appear to people with CVD. The simulation algorithms and matrices are freely available and accessible via websites and software packages[45–48].

CVD simulations have facilitated the creation of colorblind-friendly palettes[49], and they have led to algorithms that recolor images to become more accessible to people with CVD. Recolorization methods focus on enhancing color contrasts and preserving image naturalness[41]. Many

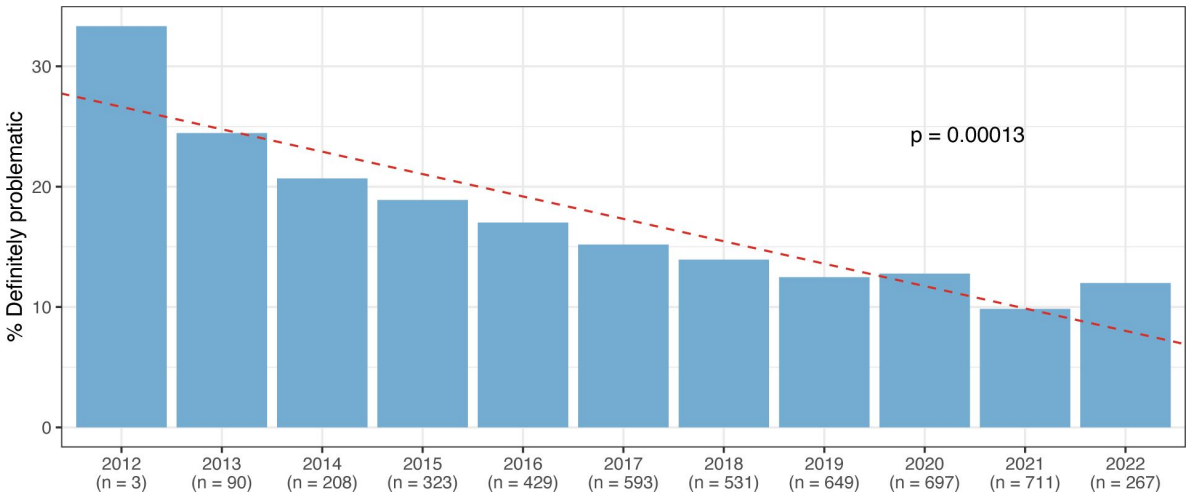

Figure 1

Longitudinal trends.

For the 4,964 images in the training set, we identified images that were "Definitely okay" or "Definitely problematic". This graph shows the percentage of these images that were "Definitely problematic" in a given year, over time. We used linear regression to calculate the P-value.

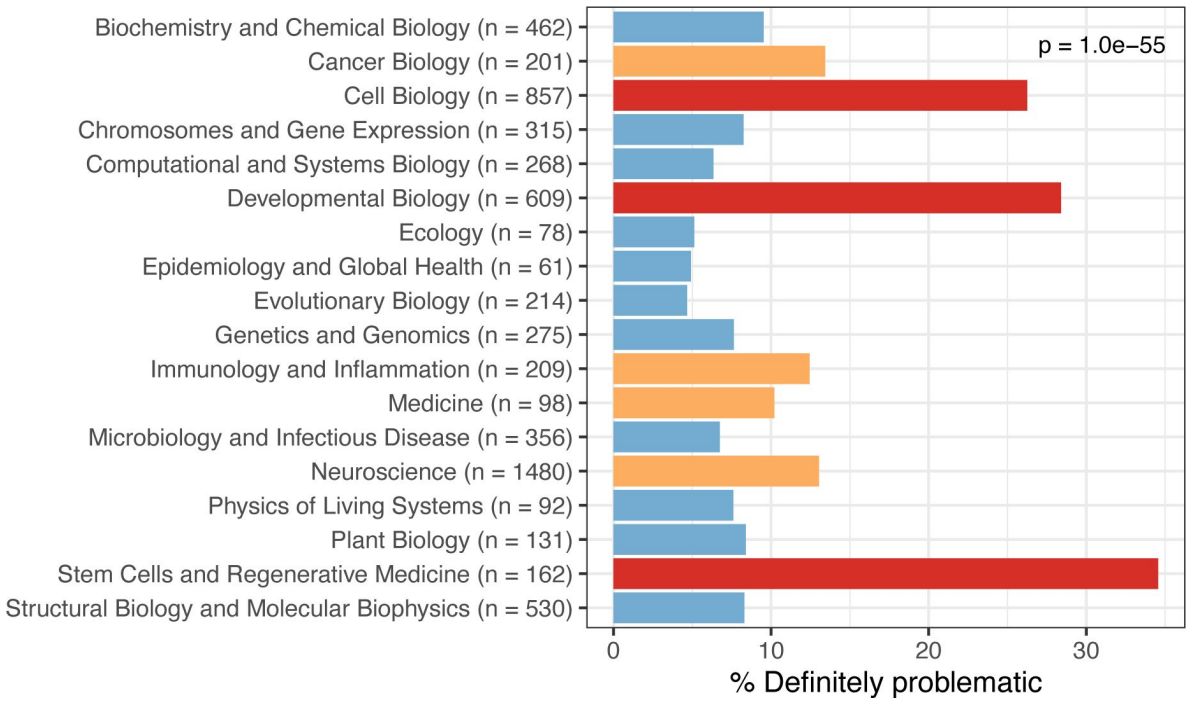

Figure 2

Trends by biology subdiscipline.

For the 4,964 images in the training set, we identified images that were "Definitely okay" or "Definitely problematic". This graph shows the percentage of these images that were "Definitely problematic" for a given subdiscipline, as indicated in the article metadata. In many cases, a single image was associated with multiple subdisciplines; these images are counted separately for each subdiscipline. We used a chi-squared goodness-of-fit test to calculate the P-value.

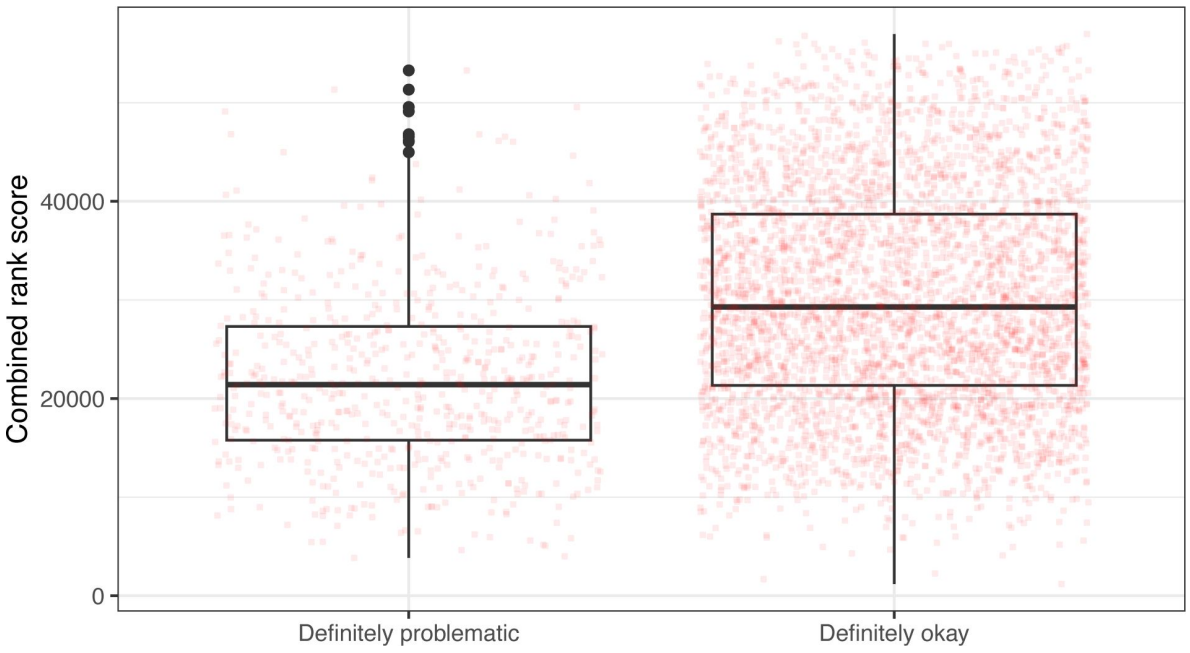

Figure 3

Rank-based metric score for images categorized as “Definitely okay” or “Definitely problematic”.

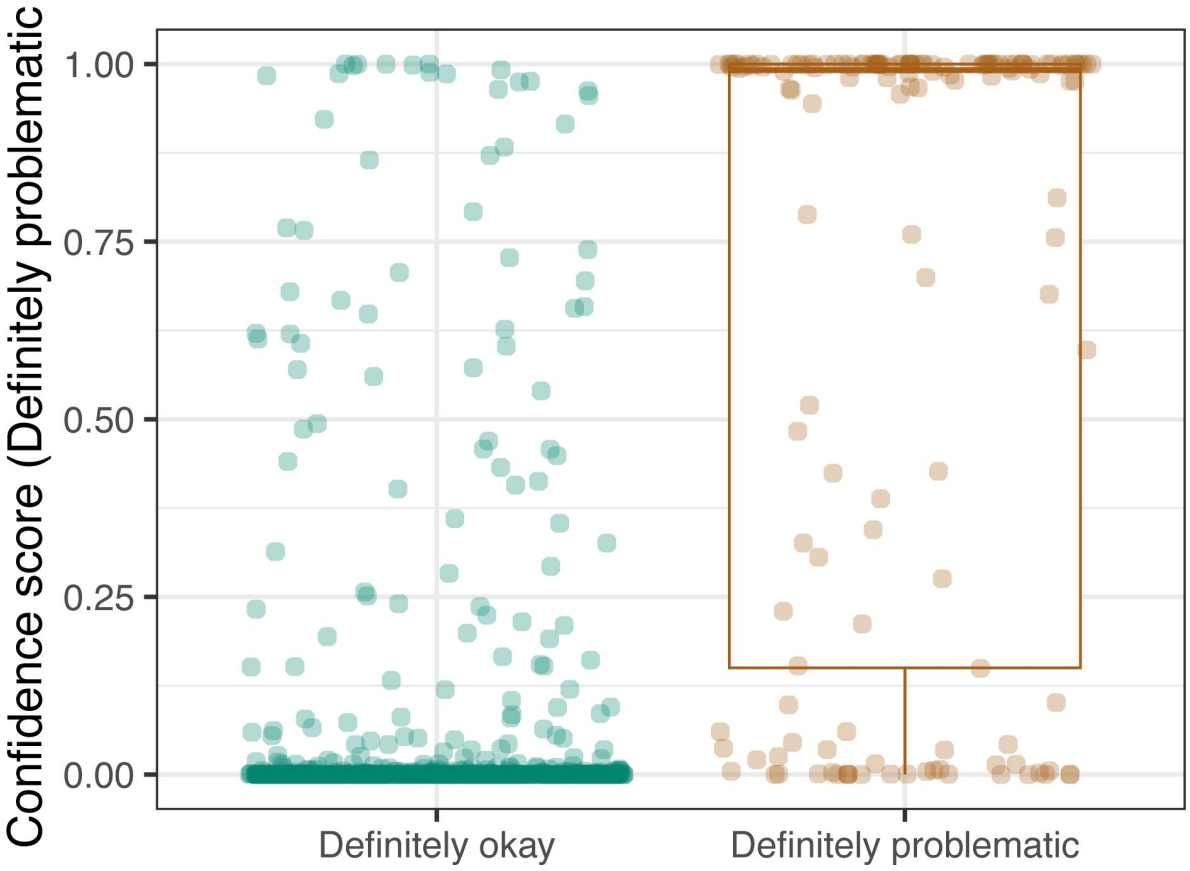

Figure 4

Convolutional Neural Network predictions for the hold-out test set.

Each point represents the prediction for an image from the hold-out test set. Relatively high confidence scores indicate that the model had more confidence that a given image was “Definitely problematic” for a person with deuteranopia.

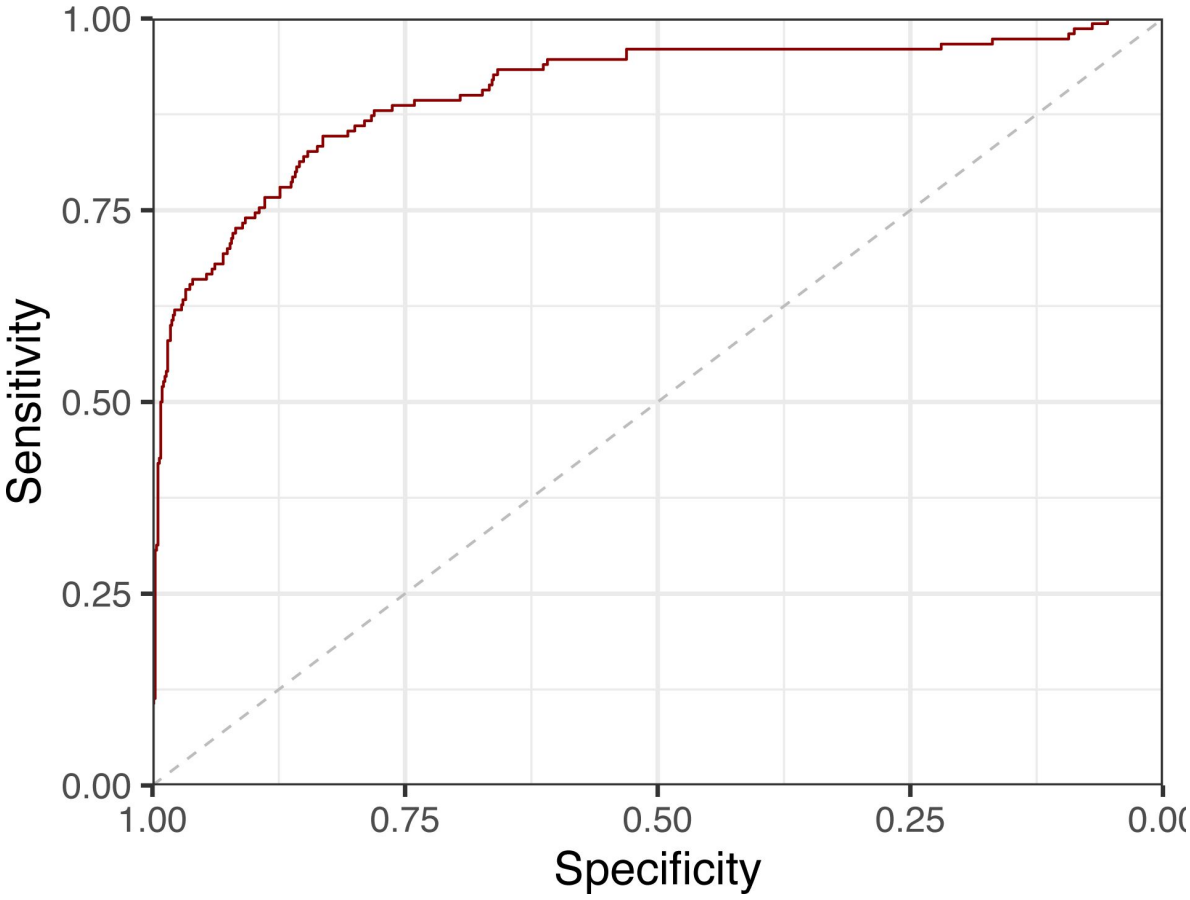

Figure 5

Receiver operating characteristic curve for Convolutional Neural Network predictions on the hold-out test set.

This curve illustrates tradeoffs between sensitivity and specificity for the Convolutional Neural Network on the hold-out test set. The area under the curve is 0.89.

algorithms have been developed to compensate for dichromacy[50–61]. These algorithms apply a variety of techniques including hue rotation, customized difference addition, node mapping, and generative adversarial networks[41,57]. Many of these methods have been tested for efficacy, both qualitatively and quantitatively[41]. Recoloring algorithms have been applied to PC displays, websites, and smart glasses[62]. Despite the prevalence of these algorithms, current techniques have not been systematically compared and may sacrifice image naturalness to increase contrast. Additionally, recoloring may not improve the accessibility of some scientific figures because papers often reference colors in figure descriptions. Recoloring the image could interfere with matching colors between the text and images.

An increase in available resources for making figures accessible to individuals with CVD has prompted some researchers to investigate whether these resources have been impactful in decreasing the frequency of published scientific figures with problematic color pairings. Frane examined the prevalence of images in psychology journals that could be confusing to people with CVD[63]. A group of panelists with CVD qualitatively evaluated 246 images and found that 13.8% of color figures caused difficulty for at least one panelist; this percentage is similar to our findings. They also found that in instructions to authors, journals rarely mentioned the importance of designing figures for CVD accessibility. Angerbauer et al. recruited crowdworkers to analyze a sample of 1,710 published images and to identify issues with the use of color[64]. On average, 60% of the sampled images were given a rating of "accessible" across CVD types. From 2000 to 2019, they observed a slight increase in CVD accessibility for published figures.

Our study focuses on the biology literature and uses simulations to estimate the prevalence of problematic figures. Our work suggests that approximately 13% of figures in the *eLife* journal are challenging for scientists with moderate-to-severe deuteranopia to interpret. Although this journal may not represent biology-related journals more broadly, *eLife* does publish articles across diverse biology subdisciplines, and its content-licensing scheme made it possible to perform this study in a transparent manner. We hope that providing examples of figures that may cause issues for individuals with CVD will raise awareness of this problem and encourage authors and publishers to be more proactive about publishing CVD-friendly figures.

By summarizing color patterns in more than 66,000 images and manually reviewing 6,000 images, we have created an open data resource that other researchers can use to develop their own methods. Furthermore, we have created an automated process for assigning predictions to other scientific figures. We have made this model accessible to other scientists by deploying it in a Web application (*https://bioapps.byu.edu/colorblind_image_tester*); we anticipate that scientists and journal editors will find this tool useful when preparing or evaluating figures for publication. However, it is infeasible to perfectly automate the process of determining whether an image is CVD friendly, so we encourage biologists to use this tool as a starting point only.

Our analysis has limitations. Firstly, it relied on deuteranopia simulations rather than the experiences of deuteranopes. However, by using simulations, the reviewers were capable of seeing two versions of each image: the original and a simulated version. Secondly, because we used a single, relatively high severity threshold, our simulations do not represent the full spectrum of experiences that scientists with deuteranopia have. Furthermore, recent evidence suggests that commonly used mathematical representations of color differences are unlikely to reflect human perceptions perfectly[65]. As methods evolve for more accurately simulating color perception, we will be more capable of estimating the extent to which scientific figures are problematic for deuteranopes. Thirdly, our evaluations focused on deuteranopia, the most common form of CVD. It will be important to address other forms of CVD, such as protanopia, in future work. Finally, our CNN model was trained using biology-oriented images and thus might be less effective for other image types. Future studies could expand on this work, using images from other disciplines and/or non-science sources.

Diverse resources are available to researchers looking to make their figures suitable for people with CVD. For example, the *seaborn* Python package includes a "colorblind" palette[66]. The *colorBlindness* package for R provides simulation tools and CVD-friendly palettes[67]. When designing figures, researchers may find it useful to first design them so that key elements are distinguishable in grayscale. Then, color can be added—if necessary—to enhance the image. Color should not be used for the sole purpose of making the image aesthetically pleasing. Using minimal color avoids problems that arise from color pairing issues. Rainbow color maps, in particular, should be avoided. If a researcher finds it necessary to include problematic color pairings in figures, they can vary the saturation and intensity of the colors so they are more distinguishable to people with CVD. Many of the problematic figures that we identified in this study originated from fluorescence microscopy experiments, where red and green dyes are commonly used. Accordingly, cell biology might be a biology subdiscipline that is disproportionately inaccessible to deuteranopes. Choosing alternative color dyes could reduce this problem and improve the interpretability of microscopy images for people in all fields.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Financial disclosure

The authors received no specific funding for this work.

Author contributions

The following contributions are described using the CRediT Taxonomy[68].

HPS: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Original Draft, Writing – Review and Editing

CVW: Conceptualization, Investigation, Methodology, Software, Writing – Original Draft, Writing – Review and Editing

AFO: Data Curation, Formal Analysis, Writing – Review & Editing

SRP: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Visualization, Writing – Original Draft, Writing – Review & Editing

References

1. Delpero W. T., O'Neill H., Casson E., Hovis J. (2005) **Aviation-relevent epidemiology of color vision deficiency** *Aviat. Space Environ. Med* **76**:127–133

2. Nathans J., Thomas D., Hogness D. S. (1986) **Molecular genetics of human color vision: The genes encoding blue, green, and red pigments** *Science* **232**:193–202

3. Simunovic M. P. (2010) **Colour vision deficiency** *Eye* **24**:747–755

4. Flatla D. R. (2011) **Accessibility for individuals with color vision deficiency** :31–34 https://doi.org/10.1145/2046396.2046412

5. Lin H.-Y., Chen L.-Q., Wang M.-L. (2019) **Improving Discrimination in Color Vision Deficiency by Image Re-Coloring** *Sensors* **19**

6. Tsekouras G. E., et al. (2021) **A Novel Approach to Image Recoloring for Color Vision Deficiency** *Sensors* **21**

7. Crameri F., Shephard G. E., Heron P. J. (2020) **The misuse of colour in science communication** *Nat Commun* **11**

8. About · eLife **About · eLife.**

9. Kuhn G. R., Oliveira M. M., Fernandes L. A. F. (2008) **An improved contrast enhancing approach for color-to-grayscale mappings** *Visual Comput* **24**:505–514

10. Zhu Z., et al. (2019) **Naturalness-and information-preserving image recoloring for redgreen dichromats** *Signal Process. Image Commun* **76**:68–80

11. Chen W., Chen W., Bao H. (2011) **An Efficient Direct Volume Rendering Approach for Dichromats** *IEEE Trans. Vis. Comput. Graph* **17**:2144–2152

12. Ma Y., Gu X., Wang Y. (2009) **Color discrimination enhancement for dichromats using self-organizing color transformation** *Inf. Sci* **179**:830–843

13. Stauffer R., Mayr G. J., Dabernig M., Zeileis A. (2009) **Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations** *Bull. Am. Meteorol. Soc* **96**:203–216

14. Hunt R. W. G. (2005) **The reproduction of colour**

15. Aisch G. (2018) **I wrote some code that automatically checks visualizations for non-colorblind safe colors. Here's how it works** *vis4*

16. R Core Team (2022) **R: A language and environment for statistical computing**

17. Corporation M., Weston S. (2022) **doParallel: Foreach parallel adaptor for the 'parallel' package**

18. Xie Y., Stodden V., Leisch F., Peng R. D. (2014) **Knitr: A comprehensive tool for reproducible research in R** *Implementing reproducible computational research*

19. Ooms J. (2021) **Magick: Advanced graphics and image-processing in R**

20. Still M. (2006) **The Definitive Guide to ImageMagick**

21. Robin X., et al. (2011) **pROC: An open-source package for R and S+ to analyze and compare ROC curves** *BMC bioinformatics* **12**

22. Davis G. (2022) **spacesXYZ: CIE XYZ and some of its derived color spaces**

23. Wickham H., et al. (2019) **Welcome to the tidyverse** *J. Open Source Softw* **4**

24. Wickham H., Hester J., Ooms J. (2021) **Xml2: Parse XML**

25. Breiman L. (2001) **Random forests** *Mach. Learn* **45**:5–32

26. Fix E., Hodges J. L. (1989) **Discriminatory analysis. Nonparametric discrimination: Consistency properties** *Int. Stat. Rev. Int. Stat* **57**:238–247

27. Nelder J. A., Wedderburn R. W. (1972) **Generalized linear models** *J. R. Stat. Soc. Ser. A Stat. Soc* **135**:370–384

28. Pedregosa F., et al. (2011) **Scikit-learn: Machine learning in Python** *J. Mach. Learn. Res* **12**:2825–2830

29. Tanner W. P., Swets J. A. (1954) **A decision-making theory of visual detection** *Psychol. Rev* **61**

30. Swets J. A. (1988) **Measuring the accuracy of diagnostic systems** *Science* **240**:1285–1293

31. Kuhn M., Vaughan D., Hvitfeldt E. (2023) **Yardstick: Tidy characterizations of model performance**

32. Abadi M., et al. (2016) **{}TensorFlow{}: A system for {large-scale} machine learning** :265–283

33. Géron A. (2022) **Hands-on machine learning with scikit-learn, keras, and TensorFlow**

34. Wong S. C., Gatt A., Stamatescu V., McDonnell M. D. (2016) **Understanding Data Augmentation for Classification: When to Warp?** :1–6 https://doi.org/10.1109/DICTA.2016.7797091

35. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C. (2019) **Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. (2019) doi:10.48550/arXiv.1801.04381.** https://doi.org/10.48550/arXiv.1801.04381

36. Krizhevsky A., Sutskever I., Hinton G. E. (2012) **Imagenet classification with deep convolutional neural networks** *Adv. Neural Inf. Process. Syst* **25**

37. He K., Zhang X., Ren S., Sun J. (2016) **Deep residual learning for image recognition** :770–778

38. Node.js **Node.js. Node.js.**

39. Machado G. M., Oliveira M. M., Fernandes L. A. F. (2009) **A physiologically-based model for simulation of color vision deficiency** *IEEE Trans Vis Comput Graph* **15**:1291–1298

40. TensorFlow.js | Machine Learning for JavaScript Developers. TensorFlow **TensorFlow.js | Machine Learning for JavaScript Developers. TensorFlow.**

41. Zhu Z., Mao X. (2021) **Image recoloring for color vision deficiency compensation: A survey** *Vis Comput* **37**:2999–3018

42. Brettel H., Viénot F., Mollon J. D. (1997) **Computerized simulation of color appearance for dichromats** *J. Opt. Soc. Am. A, JOSAA* **14**:2647–2655

43. Viénot F., Brettel H., Mollon J. D. (1999) **Digital video colourmaps for checking the legibility of displays by dichromats** *Color Res. Appl* **24**:243–252

44. Stockman A., Sharpe L. T. (2000) **The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype** *Vision Research* **40**:1711–1737

45. Coblis Color Blindness Simulator Colblindor **Coblis Color Blindness Simulator Colblindor.**

46. Color blind safe colors on color wheel | Adobe Color **Color blind safe colors on color wheel | Adobe Color.**

47. DaltonLens-Python (2023) **DaltonLens-Python. (2023).**

48. Wilke, C. Colorblindr (2023) **Wilke, C. Colorblindr. (2023).**

49. Olson J. M., Brewer C. A. (1997) **An Evaluation of Color Selections to Accommodate Map Users with Color-Vision Impairments** *Ann. Assoc. Am. Geogr* **87**:103–134

50. Jefferson L., Harvey R. (2007) **An interface to support color blind computer users** :1535–1538 https://doi.org/10.1145/1240624.1240855

51. Huang J.-B., Tseng Y.-C., Wu S.-I., Wang S.-J. (2007) **Information Preserving Color Transformation for Protanopia and Deuteranopia** *IEEE Signal Process. Lett* **14**:711–714

52. Rumiński J., et al. (2010) **Color transformation methods for dichromats** :634–641 https://doi.org/10.1109/HSI.2010.5514503

53. Rasche K., Geist R., Westall J. (2005) **Re-coloring Images for Gamuts of Lower Dimension** *Comput. Graph. Forum* **24**:423–432

54. Machado G. M., Oliveira M. M. (2010) **Real-Time Temporal-Coherent Color Contrast Enhancement for Dichromats** *Comput. Graph. Forum* **29**:933–942

55. Ching S.-L., Sabudin M. (2010) **Website image colour transformation for the colour blind** :255–259 https://doi.org/10.1109/ICCTD.2010.5645874

56. Ribeiro M., Gomes A. J. P. (2019) **Recoloring Algorithms for Colorblind People: A Survey** *ACM Comput. Surv* **52**:72–72

57. Li H., et al. (2020) **Color vision deficiency datasets & recoloring evaluation using GANs** *Multimed Tools Appl* **79**:27583–27614

58. Wang X., et al. (2021) **Fast contrast and naturalness preserving image recolouring for dichromats** *Computers & Graphics* **98**:19–28

59. Nakauchi S., Onouchi T. (2008) **Detection and modification of confusing color combinations for red-green dichromats to achieve a color universal design** *Color Res. Appl* **33**:203–211

60. Zhu Z., et al. (2019) **Processing images for redgreen dichromats compensation via naturalness and information-preservation considered recoloring** *Vis Comput* **35**:1053–1066

61. Ma Y., Gu X., Wang Y. (2009) **Color discrimination enhancement for dichromats using self-organizing color transformation** *Information Sciences* **179**:830–843

62. Tanuwidjaja E., et al. (2014) **Chroma: A wearable augmented-reality solution for color blindness** :799–810 https://doi.org/10.1145/2632048.2632091

63. Frane A. (2015) **A Call for Considering Color Vision Deficiency When Creating Graphics for Psychology Reports** *J. Gen. Psychol* **142**:194–211

64. Angerbauer K., et al. (2022) **Accessibility for Color Vision Deficiencies: Challenges and Findings of a Large Scale Study on Paper Figures** :1–23 https://doi.org/10.1145/3491102.3502133

65. Bujack R., Teti E., Miller J., Caffrey E., Turton T. L. (2022) **The non-Riemannian nature of perceptual color space** *Proc. Natl. Acad. Sci. U.S.A* **119**

66. Waskom M. L. (2021) **Seaborn: Statistical data visualization** *J. Open Source Softw* **6**

67. Ou J. (2021) **colorBlindness: Safe color set for color blindness**

68. Brand A., Allen L., Altman M., Hlava M., Scott J. (2015) **Beyond authorship: Attribution, contribution, collaboration, and credit** *Learn. Publ* **28**:151–155

Article and author information

Harlan P. Stevens
Department of Biology, Brigham Young University, Provo, UT, USA

Carly V. Winegar
Department of Biology, Brigham Young University, Provo, UT, USA

Arwen F. Oakley
Department of Biology, Brigham Young University, Provo, UT, USA

Stephen R. Piccolo
Department of Biology, Brigham Young University, Provo, UT, USA
For correspondence: stephen_piccolo@byu.edu

Copyright

Editors

Reviewing Editor
Tracey Weissgerber
Berlin Institute of Health (BIH) at Charité, Berlin, Germany

Senior Editor
Peter Rodgers
eLife, Cambridge, United Kingdom

Reviewer #1 (Public Review):

Summary:
The authors of this study developed a software application, which aims to identify images as either "friendly" or "unfriendly" for readers with deuteranopia, the most common color-vision deficiency. Using previously published algorithms that recolor images to approximate how they would appear to a deuteranope (someone with deuteranopia), authors first manually assessed a set of images from biology-oriented research articles published in eLife between 2012 and 2022. The researchers identified 636 out of 4964 images as difficult to interpret ("unfriendly") for deuteranopes. They claim that there was a decrease in "unfriendly" images over time and that articles from cell-oriented research fields were most likely to contain "unfriendly" images.
The researchers used the manually classified images to develop, train, and validate an automated screening tool. They also created a user-friendly web application of the tool, where users can upload images and be informed about the status of each image as "friendly" or "unfriendly" for deuteranopes.

Strengths:
The authors have identified an important accessibility issue in the scientific literature: the use of color combinations that make figures difficult to interpret for people with color-vision deficiency. The metrics proposed and evaluated in the study are a valuable theoretical contribution. The automated screening tool they provide is well-documented, open source, and relatively easy to install and use. It has the potential to provide a useful service to the scientists who want to make their figures more accessible. The data are open and freely accessible, well documented, and a valuable resource for further research. The manuscript is well written, logically structured, and easy to follow.

Weaknesses:
(1) The authors themselves acknowledge the limitations that arise from the way they defined what constitutes an "unfriendly" image. There is a missed chance here to have engaged deuteranopes as stakeholders earlier in the experimental design. This would have allowed to determine to what extent spatial separation and labelling of problematic color combinations responds to their needs and whether setting the bar at a simulated severity of 80% is inclusive enough. A slightly lowered barrier is still a barrier to accessibility.

(2) The use of images from a single journal strongly limits the generalizability of the empirical findings as well as of the automated screening tool itself. Machine-learning algorithms are highly configurable but also notorious for their lack of transparency and for being easily biased by the training data set. A quick and unsystematic test of the web application shows that the classifier works well for electron microscopy images but fails at

recognizing red-green scatter plots and even the classical diagnostic images for color-vision deficiency (Ishihara test images) as "unfriendly". A future iteration of the tool should be trained on a wider variety of images from different journals.

(3) Focusing the statistical analyses on individual images rather than articles (e.g. in figures 1 and 2) leads to pseudoreplication. Multiple images from the same article should not be treated as statistically independent measures, because they are produced by the same authors. A simple alternative is to instead use articles as the unit of analysis and score an article as "unfriendly" when it contains at least one "unfriendly" image. In addition, collapsing the counts of "unfriendly" images to proportions loses important information about the sample size. For example, the current analysis presented in Fig. 1 gives undue weight to the three images from 2012, two of which came from the same article. If we perform a logistic regression on articles coded as "friendly" and "unfriendly" (rather than the reported linear regression on the proportion of "unfriendly" images), there is still evidence for a decrease in the frequency of "unfriendly" eLife articles over time. Another issue concerns the large number of articles (>40%) that are classified as belonging to two subdisciplines, which further compounds the image pseudoreplication. Two alternatives are to either group articles with two subdisciplines into a "multidisciplinary" group or recode them to include both disciplines in the category name.

(4.)The low frequency of "unfriendly" images in the data (under 15%) calls for a different performance measure than the AUROC used by the authors. In such imbalanced classification cases the recommended performance measure is precision-recall area under the curve (PR AUC: https://doi.org/10.1371%2Fjournal.pone.0118432) that gives more weight to the classification of the rare class ("unfriendly" images).

https://doi.org/10.7554/eLife.95524.1.sa2

Reviewer #2 (Public Review):

Summary:
An analysis of images in the biology literature that are problematic for people with a color-vision deficiency (CVD) is presented, along with a machine learning-based model to identify such images and a web application that uses the model to flag problematic images. Their analysis reveals that about 13% of the images could be problematic for people with CVD and that the frequency of such images decreased over time. Their model yields 0.89 AUC score. It is proposed that their approach could help making biology literature accessible to diverse audiences.

Strengths:
The manuscript focuses on an important yet mostly overlooked problem, and makes contributions both in expanding our understanding of the extent of the problem and in developing solutions to mitigate the problem. The paper is generally well-written and clearly organized. Their CVD simulation combines five different metrics. The dataset has been assessed by two researchers and is likely to be of high-quality. Machine learning algorithm used (convolutional neural network, CNN) is an appropriate choice for the problem. The evaluation of various hyperparameters for the CNN model is extensive.

Weaknesses:
The focus seems to be on one type of CVD (deuteranopia) and it is unclear whether this would generalize to other types. The dataset consists of images from eLife articles. While this is a reasonable starting point, whether this can generalize to other biology/biomedical articles is not assessed. "Probably problematic" and "probably okay" classes are excluded from the analysis and classification, and the effect of this exclusion is not discussed. Machine learning aspects can be explained better, in a more standard way. The evaluation metrics used for

validating the machine learning models seem lacking (e.g., precision, recall, F1 are not reported). The web application is not discussed in any depth.

https://doi.org/10.7554/eLife.95524.1.sa1

Reviewer #3 (Public Review):

Summary:
This work focuses on accessibility of scientific images for individuals with color vision deficiencies, particularly deuteranopia. The research involved an analysis of images from eLife published in 2012-2022. The authors manually reviewed nearly 5,000 images, comparing them with simulated versions representing the perspective of individuals with deuteranopia, and also evaluated several methods to automatically detect such images including training a machine-learning algorithm to do so, which performed the best. The authors found that nearly 13% of the images could be challenging for people with deuteranopia to interpret. There was a trend toward a decrease in problematic images over time, which is encouraging.

Strengths:
The manuscript is well organized and written. It addresses inclusivity and accessibility in scientific communication, and reinforces that there is a problem and that in part technological solutions have potential to assist with this problem.

The number of manually assessed images for evaluation and training an algorithm is, to my knowledge, much larger than any existing survey. This is a valuable open source dataset beyond the work herein.

The sequential steps used to classify articles follow best practices for evaluation and training sets.

Weaknesses:
I do not see any major issues with the methods. The authors were transparent with the limitations (the need to rely on simulations instead of what deuteranopes see), only capturing a subset of issues related to color vision deficiency, and the focus on one journal that may not be representative of images in other journals and disciplines.

https://doi.org/10.7554/eLife.95524.1.sa0