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

Harlan P. Stevens[1], Carly V. Winegar[1], Arwen F. Oakley[1], Stephen R. Piccolo[1,*]

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

* - Please address correspondence to S.R.P. at stephen_piccolo@byu.edu.

Abstract

To help maximize the impact of scientific journal articles, authors must ensure that article figures are accessible to people with color-vision deficiencies (CVDs), which affect up to 8% of males and 0.5% of females. We evaluated images published in biology- and medicine-oriented research articles between 2012 and 2022. Most included at least one color contrast that could be problematic for people with deuteranopia ("deuteranopes"), the most common form of CVD. However, spatial distances and within-image labels frequently mitigated potential problems. Initially, we reviewed 4,964 images from *eLife*, comparing each against a simulated version that approximated how it might appear to deuteranopes. We identified 636 (12.8%) images that we determined would be difficult for deuteranopes to interpret. Our findings suggest that the frequency of this problem has decreased over time and that articles from cell-oriented disciplines were most often problematic. We used machine learning to automate the identification of problematic images. For hold-out test sets from *eLife* (n = 879) and PubMed Central (n = 1,191), a convolutional neural network classified the images with areas under the precision-recall curve of 0.75 and 0.38, respectively. We created a Web application (https://bioapps.byu.edu/colorblind_image_tester); users can upload images, view simulated versions, and

obtain predictions. Our findings shed new light on the frequency and nature of scientific images that may be problematic for deuteranopes and motivate additional efforts to increase accessibility.

Keywords: accessibility, color vision deficiency, deuteranopia, scientific publishing, image classification, convolutional neural network

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 and medical literature contains images that may be problematic to deuteranopes, we manually reviewed a "training set" of 4,964 images and two "test sets" of 879 and 1,191 images, respectively. These images were published in articles between the years 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 biological and medical fields more inclusive.

Methods

Image acquisition

We evaluated images in articles from *eLife*, an open-access journal that publishes research in "all areas of the life sciences and medicine". 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 excluded any article that was not published with the "Research article" type. These articles were published between the years 2012 and 2022.

On March 21, 2024, we downloaded a list of articles from the PubMed Central Open Access Subset[8]. We filtered the articles to those published between 2012 and 2022 that used a CC BY license (https://creativecommons.org and were categorized as research articles. This filtering resulted in 2,730,256 article candidates.

Image summarization metrics

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]
- diptest (0.77-1)[17]
- doParallel (1.0.17)[18]

- knitr (1.44)[19]
- magick (2.7.3)[20] - interfaces with the ImageMagick software (6.9.10.23)[21]
- pROC (1.17.0.1)[22]
- spacesXYZ (1.2-1)[23]
- tidyverse (2.0.0)[24]
- xml2 (1.3.3)[25]

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 *eLife* images from 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 from *eLife* 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 from *eLife*, 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.

From the candidate articles in the PubMed Central Open Access Subset, we randomly selected 2,000 articles. Two authors (HPS and AFO) manually reviewed 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

183 definitive set of examples would produce a more accurate model. Removing these images reduced our
184 training set to 4,501 images.

185 First, we evaluated our ability to classify images as "Definitely problematic" or "Definitely okay" based
186 on the five metrics we devised. For this task, we used the following classification algorithms, which are
187 designed for one-dimensional data: Random Forests[26], k-nearest neighbors[27], and logistic regression[28]. We
188 used implementations of these algorithms in scikit-learn (version 1.1.3)[29] with default hyperparameters,
189 other than two exceptions. We used the "liblinear" solver for logistic regression, and we set the
190 "class_weight" hyperparameter to "balanced" for Random Forests and Logistic Regression. For
191 evaluation, we used three iterations of five-fold cross validation. For the test samples in each fold, we
192 calculated the area under the receiver operating characteristic curve (AUROC)[30,31] using the *yardstick*
193 package (1.2.0)[32]; we calculated the area under the precision-recall curve (AUPRC) using the *PRROC*
194 package (1.3.1)[33]. We calculated the median AUROC and AUPRC across the folds and then averaged
195 them across the three iterations.

196 Second, we evaluated our ability to classify the images as "Definitely problematic" or "Definitely okay"
197 based on the images themselves. We used a convolutional neural network (CNN) because CNN models
198 are capable of handling two-dimensional inputs and accounting for spatial patterns and colors within
199 images. To generate the CNN models, we used the Tensorflow (2.10.0) and Keras (2.10.0) packages[34,35].
200 To support transfer learning (described below), we scaled both dimensions of each image to 224. To
201 select from different configurations, we again used three iterations of five-fold cross validation (with the
202 same assignments as the earlier classification analysis). Each model configuration extended a baseline
203 configuration that had eight, two-dimensional, convolutional layers; each layer used batch normalization
204 and the *ReLU* activation function. Subsequent layers increased in size, starting with 32 nodes and
205 increasing to 64, 128, 256, 512, and 728. We trained for 30 epochs with an Adam optimization set, a
206 learning rate of 1e-3, and the binary cross-entropy loss function. The output layer used a sigmoid
207 activation function.

In addition to the baseline configuration, we tested 22 model configurations based on combinations of 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[36]. We evaluated rotation thresholds of 0.2 and 0.3.
- *Dropout* - Again to prevent overfitting, we temporarily removed 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[37,38] and ResNet50[39]. 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 baseline configuration. Instead, we added a global pooling function into a dense layer. Because our training dataset was relatively small, we expected that adding fewer layers might reduce the risk of overfitting.

- *Fine tuning* - In combination with transfer learning, we sometimes employed a two-phase training process involving an initial training phase and a fine-tuning phase. In the initial phase, we used a pre-trained model (ResNet50 or MobileNetV2) as the base model. During this phase, we froze the layers of the base model to retain the learned weights and trained the combined model for 30 epochs with a learning rate of 0.001. In the fine-tuning phase, we unfroze the layers of the base model to allow the entire model to be retrained. To avoid large adjustments that could disrupt the pre-trained weights, we reduced the learning rate to 1e-5 and trained the model for an additional 15 epochs. This phase enabled the model to make subtle updates to the pre-trained weights. After fine-tuning, the layers of the base model were refrozen.

When training each model configuration, we used AUROC to evaluate the predictive performance on the internal validation sets. After comparing the model configurations via cross validation, we used the full training set to train a model, which we used to make predictions for the hold-out test sets. We calculated the following: true positives, false positives, true negatives, false negatives, accuracy, precision (positive predictive value), recall (sensitivity), AUROC, and AUPRC.

Web application

We created a Web application using the *Node.js* framework[40]. 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.[41] 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 predicts 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. With the intent to maximize the generalizability of these predictions, we trained a model using images from the training set and both holdout test sets. To facilitate execution of the CNN within the Web application, we used Tensorflow.js (version 4.0.0)[42].

256 **Code and data availability**

257 The images we used for evaluation and the trained TensorFlow models are stored in an Open Science
258 Framework repository (https://osf.io/8yrkb). It also includes folders with images marked as "friendly" or
259 "unfriendly" to facilitate examination of images in either category. The code for processing and analyzing
260 the images is available at https://github.com/srp33/bio_image_colorblindness. That repository also
261 includes the calculated metrics, cross-validation assignments, results of image curation, and outputs of the
262 classification algorithms. The Web application code is available at
263 https://github.com/srp33/colorblind_image_tester.

264 Results

265 We downloaded images from research articles published in the *eLife* journal. Not counting duplicate
266 versions of the same image, we obtained 66,253 images. Of these images, 1,744 (2.6%) were grayscale
267 (no color). Of these images, 56,816 (85.6%) included at least one color pair for which the amount of
268 contrast might be problematic to people with moderate-to-severe deuteranopia ("deuteranopes"). To
269 characterize potentially problematic aspects of each color-based image, we calculated five metrics based
270 on color contrasts and distances; we also compared the color profiles against what deuteranopes might
271 see. The *mean, pixel-wise color distance between the original and simulated image* exhibited a bimodal
272 distribution, according to Hartigans' Dip Test for Unimodality ($p < 0.001$)[43]. Specifically, 4,708 images
273 (7.3%) had a difference smaller than 0.01, while the median difference for the remaining images was 0.05
274 (Figure S1). Most other metrics showed similar patterns, although bimodality was less apparent through
275 visual inspection (Figures S2-S4). The exception was the *proportion of pixels in the original image that*
276 *used a color from one of the high-ratio color pairs*, which was unimodal ($p = 1$; Figure S5).

277 We determined that many images with the highest (or lowest, as would be the case for the "Mean
278 Euclidean distance between pixels for high-ration color pairs") scores for these metrics would 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 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 1 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.

We evaluated longitudinal trends and differences among biology subdisciplines. In some cases for the *eLife* articles, multiple images came from the same journal article. Therefore, to avoid pseudoreplication, we categorized each *article* as either "Definitely okay" or "Definitely problematic." If an article included at least one "Definitely problematic" image, we categorized the entire article under this category. The percentage of "Definitely problematic" articles declined steadily between 2012 and 2021, with a modest increase in 2022 (Figure 1). (Fewer articles were available for 2022 than for prior years.) Using a generalized linear model with a binomial family to perform logistic regression, we found this decline to be statistically significant ($p < 0.001$). A χ^2 goodness-of-fit test revealed that the number of "Definitely problematic" articles differed significantly by subdiscipline ($p < 0.001$). The subdisciplines with the highest percentages of problematic articles were *Cell Biology*, *Developmental Biology*, and *Stem Cells and Regenerative Medicine* (Figure 2).

Despite the benefits of manual review, this process is infeasible on a large scale. Therefore, we evaluated techniques for automating image classification. As an initial test, we used the five image-quantification metrics. We also combined these into a single, ranked-based score for each image. In all cases, the metrics differed significantly between the "Definitely okay" and "Definitely problematic" images (Figures S6-S11). To estimate their predictive performance, we performed cross validation on the training set. Values relatively close to 1.0 indicate relatively high performance. A value of 0.5 indicates that predictions are no better than random guessing. The best-performing metric was *number of color pairs that exhibited a high color-distance ratio between the original and simulated images* (AUROC: 0.75; AUPRC: 0.34). All the other metrics—except *mean, pixel-wise color distance between the original and simulated image*—performed better than random guessing (Table S1). 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 (AUROC: 0.82; AUPRC: 0.43; Table S2).

Additionally, 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 and configurations. Accordingly, we performed multiple iterations of cross validation on the training set and compared a variety of hyperparameters and configurations. All non-default options performed better than Logistic Regression based on the image-quantification metrics (Table S3). The best-performing approach used class weighting; early stopping; random flipping and rotation (threshold: 0.2); a dropout rate of 0.5; the ResNet pre-trained model for transfer learning; and model fine tuning.

We manually reviewed a hold-out test set that consisted of 1,000 additional images from *eLife* (Additional Data File 2). After we removed images that were not "Definitely okay" or "Definitely problematic," 879 images remained. 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 (AUPRC: 0.49; Figures S12-S14). The CNN

classified the images with an AUROC of 0.89 (AUPRC: 0.77; Figures 3-4, S15). Additional Data File 3 indicates the performance of both models for a variety of classification metrics.

For the 92 hold-out images that were misclassified by the CNN model, we comparing them against our manual annotations and determined that in 13 cases, the reviewers had missed subtle patterns; we conclude that it would be justified to change these labels (Additional Data File 4). 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.

Lastly, we used the Logistic Regression and CNN models to predict "Definitely okay" or "Definitely problematic" status for a second hold-out set with images from PubMed Central (Additional Data File 5). Of the non-grayscale images, 104 were labeled as "Definitely problematic" and 1,087 as "Definitely okay." Again, the CNN model (AUROC: 0.78; AUPRC: 0.39; Figures 5-6, S16) outperformed the Logistic Regression model (AUROC: 0.73; AUPRC: 0.16; Figures S17-S19; Additional Data File 3).

Discussion

To learn more about the prevalence of images in the biology literature that may be inaccessible to deuteranopes, we examined thousands of articles from the *eLife* journal. *eLife* uses a content-licensing scheme that made it possible to perform this study in a transparent manner. Additionally, we selected articles from PubMed Central, as images from these articles represent life-science journals more broadly. After manual review, we estimate that 12.8% of the figures in *eLife* would be challenging to interpret for scientists with moderate-to-severe deuteranopia. The percentage of "Definitely problematic" figures in PubMed Central articles was considerably lower (5.2%). One reason is that a much higher percentage (38.5%) of the images from PubMed Central were grayscale compared to those from *eLife* (4.4%). The

findings for both sources indicate that color accessibility is a problem for thousands of journal articles per year.

Significant work has been done to address and improve accessibility for individuals with CVD[44]. This work can be categorized into four types of studies: simulation methods, recolorization methods, estimating the frequency of accessible images, and educational. 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[45]. 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 3x3 transformation matrix to simulate images in the same LMS space[46]. Machado et al. created matrices to simulate CVDs based on the shift theory of cone cell sensitivity[41,47]. These algorithms allow individuals without CVD to qualitatively test how their images might appear to people with CVD. The simulation algorithms and matrices are freely available and accessible via websites and software packages[48–51].

CVD simulations have facilitated the creation of colorblind-friendly palettes[52], 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[44]. Many algorithms have been developed to compensate for dichromacy[53–64]. These algorithms apply a variety of techniques including hue rotation, customized difference addition, node mapping, and generative adversarial networks[44,60]. Many of these methods have been tested for efficacy, both qualitatively and quantitatively[44]. Recolorization algorithms have been applied to PC displays, websites, and smart glasses[65]. 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 publishing scientific figures with problematic colors. Frane examined the prevalence of images in psychology journals that could be confusing to people with CVD[66]. 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[67]. 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.

Educational resources are available to researchers looking to make their figures suitable for people with CVD. For example, Jambor, et al. provide guidelines and examples to help researchers avoid common problems[68]. *JetFighter* scans preprints from bioRxiv and searches for rainbow-based color schemes[69]. When these are identified, *JetFighter* notifies the authors about page(s) that might need to be adjusted. However, as we have shown, the presence of particular color combinations does not necessarily indicate that an image is problematic to people with deuteranopia. Frequently, a more nuanced evaluation is necessary.

The *seaborn* Python package includes a "colorblind" palette[70]. The *colorBlindness* package for R provides simulation tools and CVD-friendly palettes[71]. The *scatterHatch* package facilitates creation of CVD-friendly scatter plots for single-cell data[72]. 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 an 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

399 distinguishable to people with CVD. Many of the problematic figures that we identified in this study
400 originated from fluorescence microscopy experiments, where red and green dyes were used. Choosing
401 alternative color dyes could reduce this problem and improve the interpretability of microscopy images
402 for people in all fields.

403 Our analysis has limitations. Firstly, it relied on deuteranopia simulations rather than the experiences of
404 deuteranopes. However, by using simulations, the reviewers were capable of seeing two versions of each
405 image: the original and a simulated version. We believe this is important in assessing the extent to which
406 deuteranopia could confound image interpretations. Conceivably, this could be done with deuteranopes
407 after recoloration, but it is difficult to know whether deuteranopes would see the recolored images in the
408 same way that non-deuteranopes see the original images. Secondly, because we used a single, relatively
409 high severity threshold, our simulations do not represent the full spectrum of experiences that scientists
410 with deuteranopia have. Thus, our findings and tools should be relevant to some (but not all) people with
411 deuteranopia. Furthermore, recent evidence suggests that commonly used mathematical representations of
412 color differences are unlikely to reflect human perceptions perfectly[73]. As methods evolve for more
413 accurately simulating color perception, we will be more capable of estimating the extent to which
414 scientific figures are problematic for deuteranopes. Thirdly, our evaluations focused on deuteranopia, the
415 most common form of CVD. It will be important to address other forms of CVD, such as protanopia, in
416 future work. Fourthly, we identified some images as "Probably problematic" or "Probably okay." Using
417 our review process, we were unable to draw firm conclusions about these images. To avoid adding noise
418 to the classification analyses—we excluded these images and provided notes reflecting our reasoning.
419 Future work may help to clarify these labels. Finally, our CNN model performed well at differentiating
420 between "Definitely okay" and "Definitely problematic" images in the *eLife* hold-out test set; however,
421 the model's predictive performance dropped considerably when applied to the PubMed Central hold-out
422 test set. Many of the *eLife* images are from cell-related research, and we labeled many of these as
423 problematic. Many other image types were also identified as unfriendly, including heat maps, line charts,

maps, three-dimensional structural representations of proteins, photographs, network diagrams, etc. Our model may have developed a bias toward patterns specific to image types that are over-represented in *eLife*, affecting its performance for other journals. The PubMed Central Open Access Subset contains articles for thousands of journals, spanning diverse subdisciplines of biology and medicine. It seems likely that this diversity is a factor behind the drop in performance. Future efforts to review larger collections of PubMed Central articles could help to overcome this limitation.

By summarizing color patterns in more than 66,000 images and manually reviewing 8,000 images, we have created an open data resource that other researchers can use to develop their own methods. Using all of these images, we trained a machine-learning model that predicts whether images are friendly to deuteranopes. It is available as a Web application (https://bioapps.byu.edu/colorblind_image_tester). Scientists and others can use it to obtain insights into whether individual images are accessible to scientists with deuteranopia. However, this tool should be used as a starting point only. Human judgment remains essential.

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[74].

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

Figures

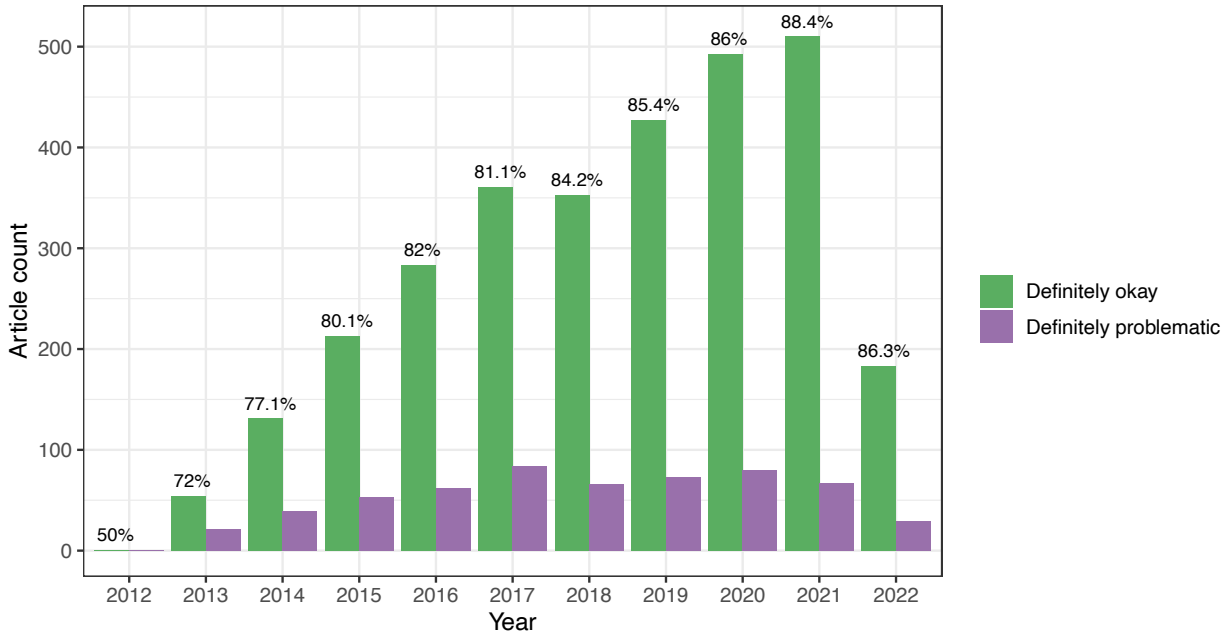


Figure 1: Longitudinal trends for the *eLife* articles. For the training set, we summarized our findings per article. This graph shows article counts for the "Definitely okay" and "Definitely problematic" categories for each year evaluated.

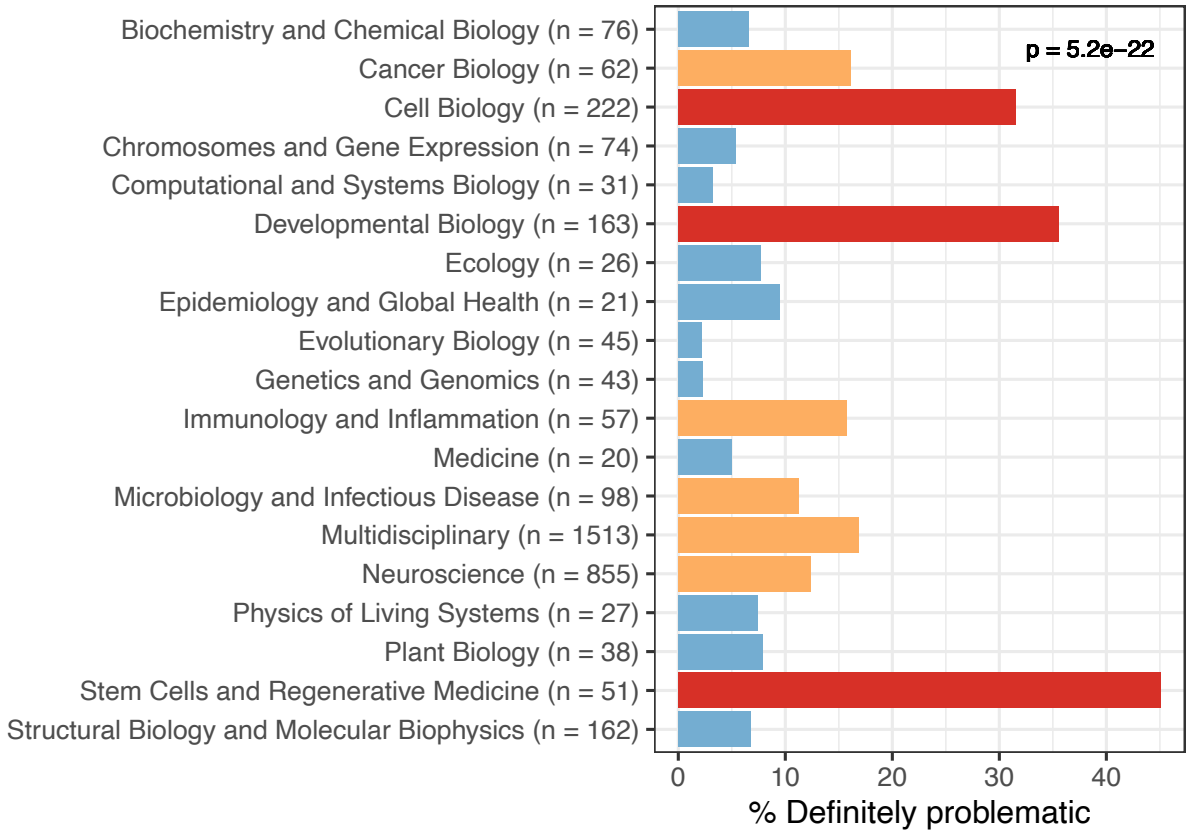
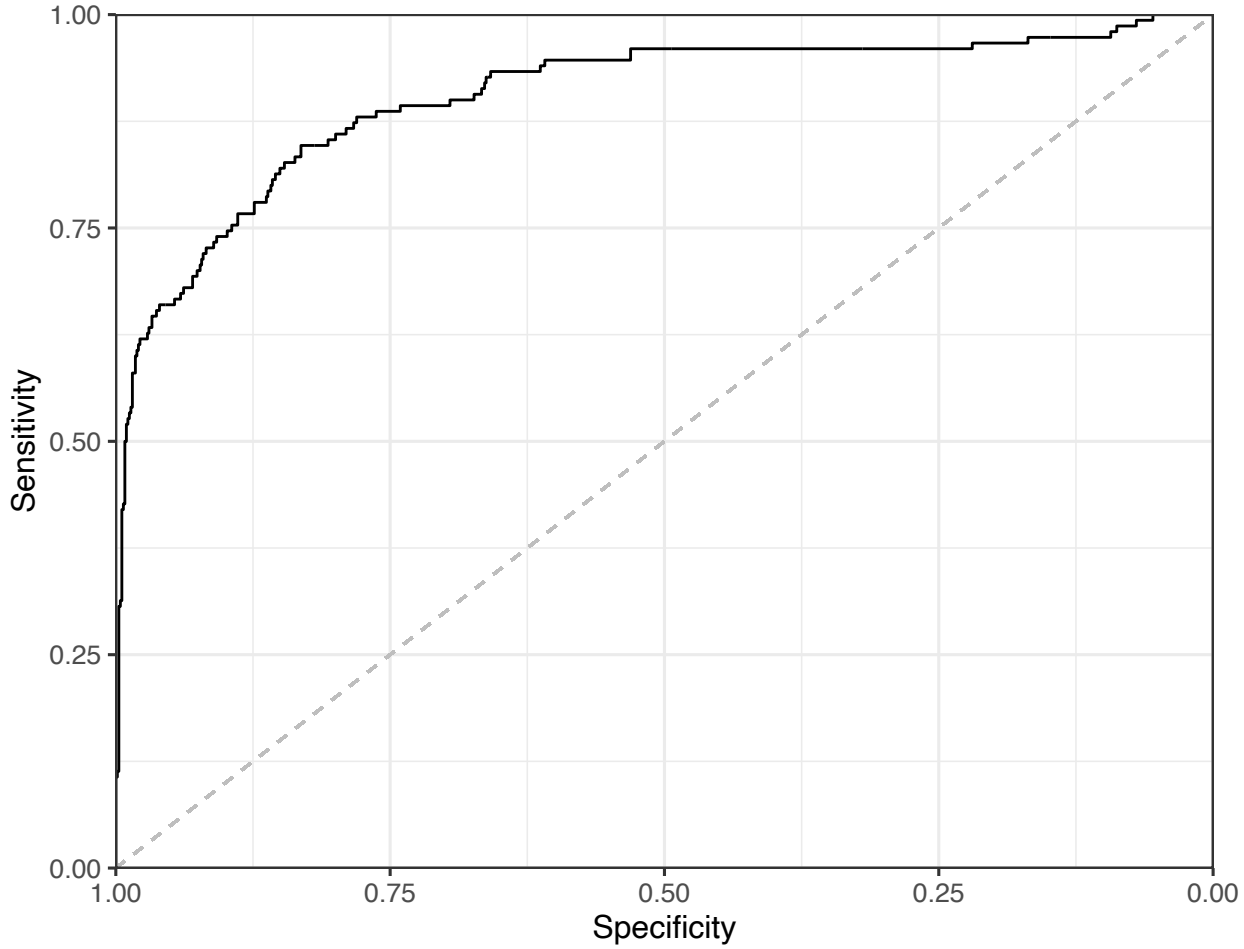


Figure 2: Trends by biology subdiscipline for the *eLife* articles. For the training set, this graph shows the percentage of articles categorized as "Definitely problematic" for a given subdiscipline, as indicated in the article metadata. In many cases, a single article was associated with multiple subdisciplines; these articles are shown as "Multidisciplinary." We used a χ^2 goodness-of-fit test to calculate the p-value, with the overall proportion of each discipline as the expected probability.



466

467 **Figure 3: Receiver operating characteristic curve for the Convolutional Neural Network**

468 **predictions for the images in the *eLife* hold-out test set.** This curve illustrates tradeoffs between

469 sensitivity and specificity. The area under the curve is 0.89.

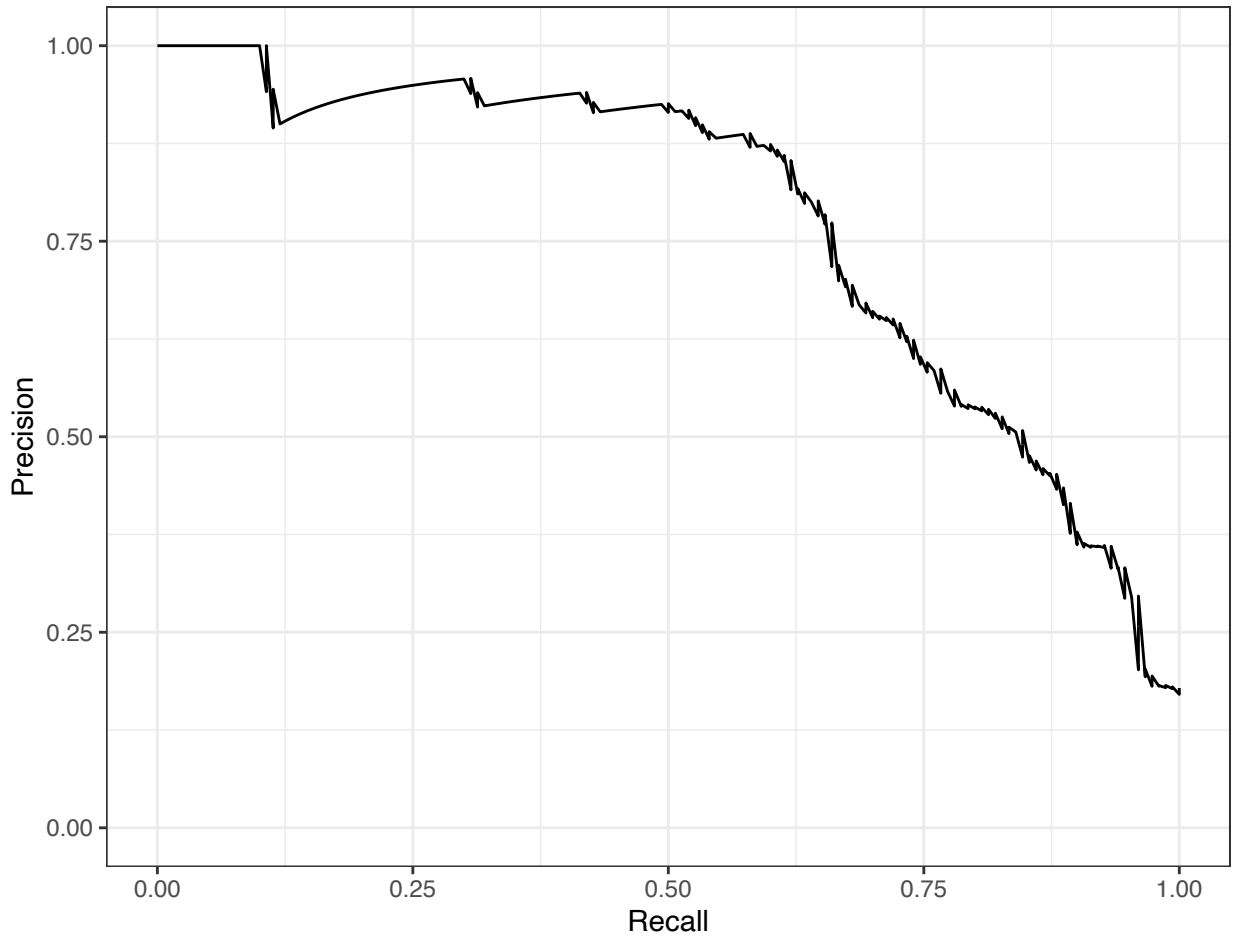


Figure 4: Precision-recall curve for the Convolutional Neural Network predictions for the images in the *eLife* hold-out test set. This curve illustrates tradeoffs between precision and recall. The area under the curve is 0.75.

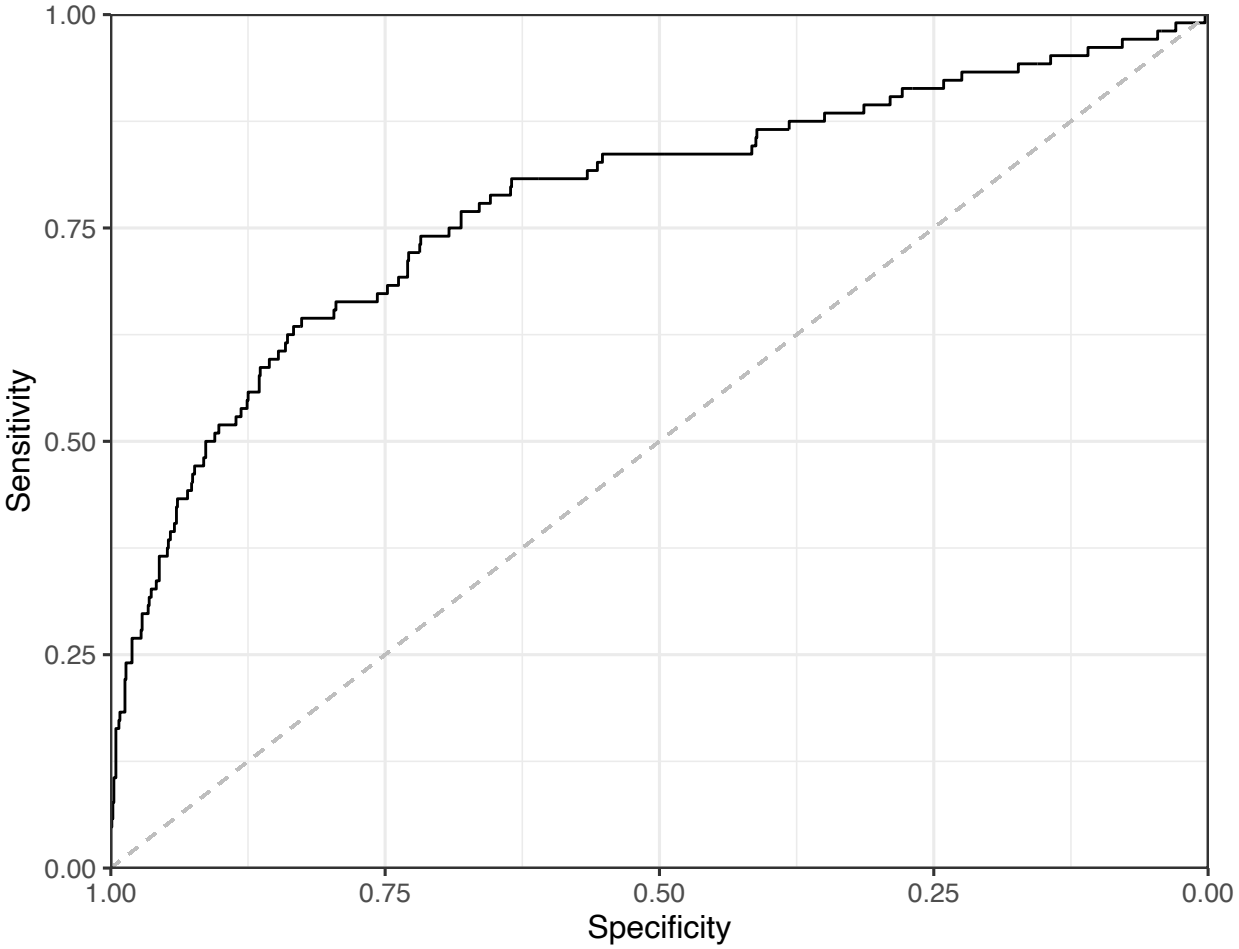


Figure 5: Receiver operating characteristic curve for the Convolutional Neural Network predictions for the images in the PubMed Central hold-out test set. This curve illustrates tradeoffs between sensitivity and specificity. The area under the curve is 0.78.

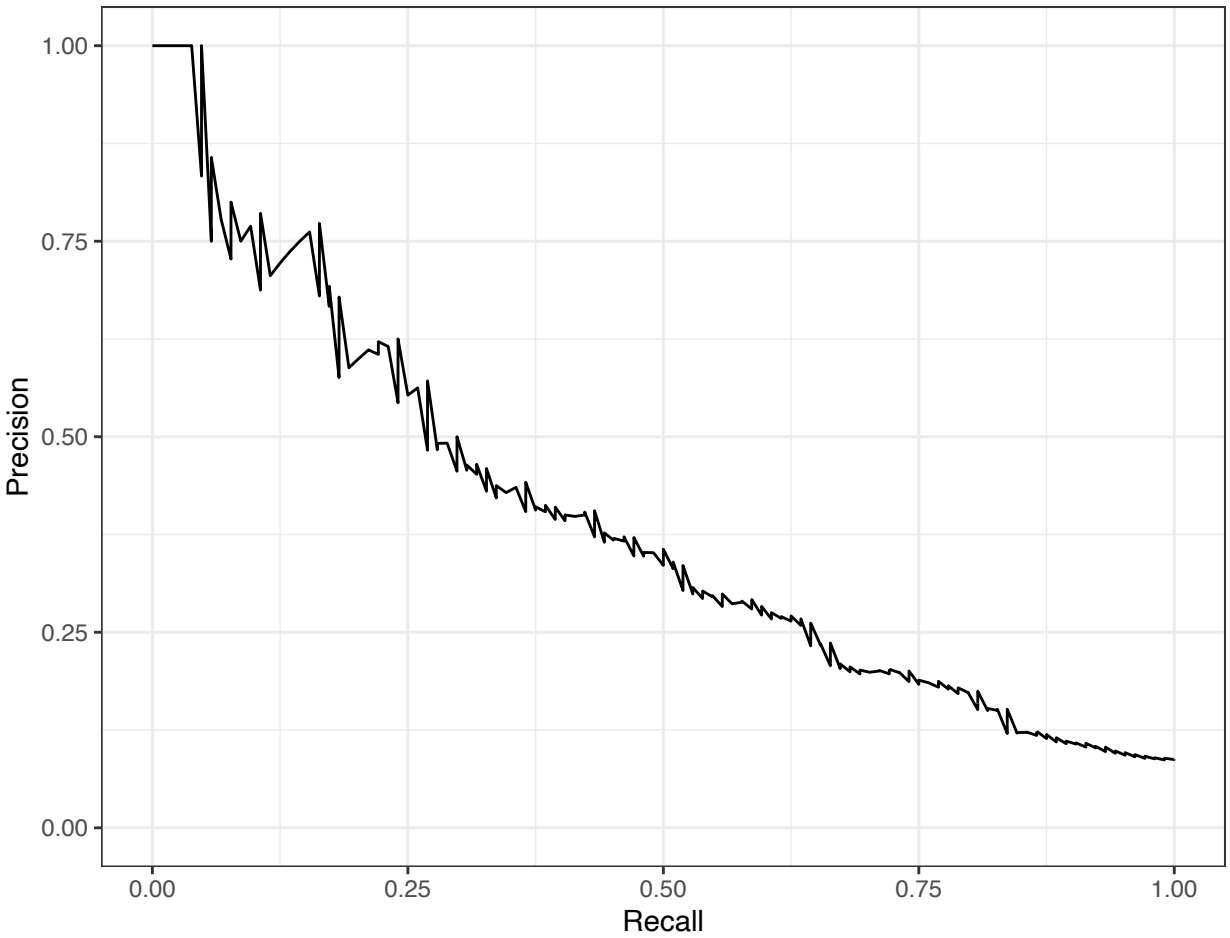


Figure 6: Precision-recall curve for the Convolutional Neural Network predictions for the images in the PubMed Central hold-out test set. This curve illustrates tradeoffs between precision and recall. The area under the curve is 0.39.

Additional Data Files

Additional Data File 1: Results of manual curation for 5,000 images from the training set.

Additional Data File 2: Results of manual curation for 1,000 images from the *eLife* hold-out test set.

Additional Data File 3: Performance metrics for predictions made on hold-out test sets. This file provides a variety of classification metrics for the two hold-out test sets. AUROC = area under receiver operating characteristic curve. AUPRC = area under precision-recall curve.

Additional Data File 4: Results of manual review of misclassified images from the hold-out test set. We manually reviewed each image that we had previously classified as "Definitely okay" but that the model predicted as "Definitely problematic" (or vice versa). The Conclusion column indicates our categorical reevaluation of each image. "Unclear" = We continue to conclude that our manual label was correct, and it is unclear what confused the model. "Understandable" = We continue to conclude that our manual label was correct, and we think we understand what confused the model. "Agree" = We acknowledge that the manual label was incorrect, and the model helped us identify that.

Additional Data File 5: Results of manual curation for 2,000 images used from the PubMed Central hold-out test set.

References

1. Delpero, W. T., O'Neill, H., Casson, E. & Hovis, J. Aviation-relevent epidemiology of color vision deficiency. *Aviation, space, and environmental medicine* **76**, 127–133 (2005).
2. Nathans, J., Thomas, D. & Hogness, D. S. Molecular genetics of human color vision: The genes encoding blue, green, and red pigments. *Science* **232**, 193–202 (1986).
3. Simunovic, M. P. Colour vision deficiency. *Eye* **24**, 747–755 (2010).
4. Flatla, D. R. Accessibility for individuals with color vision deficiency. in *Proceedings of the 24th annual ACM symposium adjunct on User interface software and technology* 31–34 (Association for Computing Machinery, New York, NY, USA, 2011). doi:10.1145/2046396.2046412.
5. Lin, H.-Y., Chen, L.-Q. & Wang, M.-L. Improving Discrimination in Color Vision Deficiency by Image Re-Coloring. *Sensors* **19**, 2250 (2019).

6. Tsekouras, G. E. *et al.* A Novel Approach to Image Recoloring for Color Vision Deficiency. *Sensors* **21**, 2740 (2021).

7. Crameri, F., Shephard, G. E. & Heron, P. J. The misuse of colour in science communication. *Nat Commun* **11**, 5444 (2020).

8. PMC Open Access Subset. *PubMed Central (PMC)*.

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

10. Zhu, Z. *et al.* Naturalness-and information-preserving image recoloring for red–green dichromats. *Signal Processing: Image Communication* **76**, 68–80 (2019).

11. Chen, W., Chen, W. & Bao, H. An Efficient Direct Volume Rendering Approach for Dichromats. *IEEE Transactions on Visualization and Computer Graphics* **17**, 2144–2152 (2011).

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

13. Stauffer, R., Mayr, G. J., Dabernig, M. & Zeileis, A. Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bulletin of the American Meteorological Society* **96**, 203–216 (2009).

14. Hunt, R. W. G. *The Reproduction of Colour*. (John Wiley & Sons, 2005).

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

16. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2022).

17. Maechler, M. *Diptest: Hartigan's Dip Test Statistic for Unimodality - Corrected*. (2024).

18. Corporation, M. & Weston, S. *doParallel: Foreach Parallel Adaptor for the 'Parallel' Package*. (2022).

19. Xie, Y. Knitr: A comprehensive tool for reproducible research in R. in *Implementing reproducible computational research* (eds. Stodden, V., Leisch, F. & Peng, R. D.) (Chapman and Hall/CRC, 2014).

20. Ooms, J. *Magick: Advanced Graphics and Image-Processing in R*. (2021).

21. Still, M. *The Definitive Guide to ImageMagick*. (Apress, 2006).

22. Robin, X. *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* **12**, 77 (2011).

23. Davis, G. *spacesXYZ: CIE XYZ and Some of Its Derived Color Spaces*. (2022).

24. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

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

26. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

27. Fix, E. & Hodges, J. L. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique* **57**, 238–247 (1989).

28. Nelder, J. A. & Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society* **135**, 370–384 (1972).

29. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

30. Tanner Jr, W. P. & Swets, J. A. A decision-making theory of visual detection. *Psychological review* **61**, 401 (1954).

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

32. Kuhn, M., Vaughan, D. & Hvitfeldt, E. *Yardstick: Tidy Characterizations of Model Performance*. (2023).

33. Grau, J., Grosse, I. & Keilwagen, J. PRROC: Computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).

34. Abadi, M. *et al.* {}TensorFlow{}: A system for {large-scale} machine learning. in *12th USENIX symposium on operating systems design and implementation (OSDI 16)* 265–283 (2016).

35. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. (" O'Reilly Media, Inc.", 2022).

36. Wong, S. C., Gatt, A., Stamatescu, V. & McDonnell, M. D. Understanding Data Augmentation for Classification: When to Warp? in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* 1–6 (2016). doi:10.1109/DICTA.2016.7797091.

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

38. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**, (2012).

39. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).

40. Node.js. *Node.js*.

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

42. TensorFlow.js Machine Learning for JavaScript Developers. *TensorFlow*.

43. Hartigan, J. A. & Hartigan, P. M. The dip test of unimodality. *The annals of Statistics* 70–84 (1985).

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

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

46. Viénot, F., Brettel, H. & Mollon, J. D. Digital video colourmaps for checking the legibility of displays by dichromats. *Color Research & Application* **24**, 243–252 (1999).

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

48. Coblis — Color Blindness Simulator – Colblindor.

49. Color blind safe colors on color wheel Adobe Color.

50. DaltonLens-Python. (2023).

51. Wilke, C. Colorblindr. (2023).

52. Olson, J. M. & Brewer, C. A. An Evaluation of Color Selections to Accommodate Map Users with Color-Vision Impairments. *Annals of the Association of American Geographers* **87**, 103–134 (1997).

53. Jefferson, L. & Harvey, R. An interface to support color blind computer users. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1535–1538 (Association for Computing Machinery, New York, NY, USA, 2007). doi:10.1145/1240624.1240855.

54. Huang, J.-B., Tseng, Y.-C., Wu, S.-I. & Wang, S.-J. Information Preserving Color Transformation for Protanopia and Deuteranopia. *IEEE Signal Processing Letters* **14**, 711–714 (2007).

55. Rumiński, J. *et al.* Color transformation methods for dichromats. in *3rd International Conference on Human System Interaction* 634–641 (2010). doi:10.1109/HSI.2010.5514503.

56. Rasche, K., Geist, R. & Westall, J. Re-coloring Images for Gamuts of Lower Dimension. *Computer Graphics Forum* **24**, 423–432 (2005).

57. Machado, G. M. & Oliveira, M. M. Real-Time Temporal-Coherent Color Contrast Enhancement for Dichromats. *Computer Graphics Forum* **29**, 933–942 (2010).

58. Ching, S.-L. & Sabudin, M. Website image colour transformation for the colour blind. in *2010 2nd International Conference on Computer Technology and Development* 255–259 (2010). doi:10.1109/ICCTD.2010.5645874.

59. Ribeiro, M. & Gomes, A. J. P. Recoloring Algorithms for Colorblind People: A Survey. *ACM Comput. Surv.* **52**, 72:1–72:37 (2019).

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

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

610 62. Nakauchi, S. & Onouchi, T. Detection and modification of confusing color combinations for red-
611 green dichromats to achieve a color universal design. *Color Research & Application* **33**, 203–211 (2008).
612 63. Zhu, Z. *et al.* Processing images for red–green dichromats compensation via naturalness and
613 information-preservation considered recoloring. *Vis Comput* **35**, 1053–1066 (2019).
614 64. Ma, Y., Gu, X. & Wang, Y. Color discrimination enhancement for dichromats using self-
615 organizing color transformation. *Information Sciences* **179**, 830–843 (2009).
616 65. Tanuwidjaja, E. *et al.* Chroma: A wearable augmented-reality solution for color blindness. in
617 *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*
618 799–810 (Association for Computing Machinery, New York, NY, USA, 2014).
619 doi:10.1145/2632048.2632091.
620 66. Frane, A. A Call for Considering Color Vision Deficiency When Creating Graphics for
621 Psychology Reports. *The Journal of General Psychology* **142**, 194–211 (2015).
622 67. Angerbauer, K. *et al.* Accessibility for Color Vision Deficiencies: Challenges and Findings of a
623 Large Scale Study on Paper Figures. in *Proceedings of the 2022 CHI Conference on Human Factors in*
624 *Computing Systems* 1–23 (Association for Computing Machinery, New York, NY, USA, 2022).
625 doi:10.1145/3491102.3502133.
626 68. Jambor, H. *et al.* Creating clear and informative image-based figures for scientific publications.
627 *PLoS Biol* **19**, e3001161 (2021).
628 69. JetFighter: Towards figure accuracy and accessibility. *eLife* (2019).
629 70. Waskom, M. L. Seaborn: Statistical data visualization. *Journal of Open Source Software* **6**, 3021
630 (2021).
631 71. Ou, J. *colorBlindness: Safe Color Set for Color Blindness*. (2021).
632 72. Guha, T., Fertig, E. J. & Deshpande, A. Generating colorblind-friendly scatter plots for single-
633 cell data. *eLife* **11**, e82128 (2022).
634 73. Bujack, R., Teti, E., Miller, J., Caffrey, E. & Turton, T. L. The non-Riemannian nature of
635 perceptual color space. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2119753119 (2022).

636 74. Brand, A., Allen, L., Altman, M., Hlava, M. & Scott, J. Beyond authorship: Attribution,
637 contribution, collaboration, and credit. *Learned Publishing* **28**, 151–155 (2015).