Zhuang et al. *EPJ Data Science* (2023) 12:5
https://doi.org/10.1140/epjds/s13688-023-00380-y

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

A computational analysis of accessibility, readability, and explainability of figures in open access publications

Han Zhuang[1*], Tzu-Yang Huang[2] and Daniel E. Acuna[3]

*Correspondence:
hzhuang@syr.edu
[1]School of Information Studies, Syracuse University, Syracuse, USA
Full list of author information is available at the end of the article

Abstract

Figures are an essential part of scientific communication. Yet little is understood about how accessible (e.g., color-blind safe), readable (e.g., good contrast), and explainable (e.g., contain captions and legends) they are. We develop computational techniques to measure these features and analyze a large sample of them from open access publications. Our method combines computer and human vision research principles, achieving high accuracy in detecting problems. In our sample, we estimated that around 20.6% of publications contain either accessibility, readability, or explainability issues (around 2% of all figures contain accessibility issues, 3% of diagnostic figures contain readability issues, and 23% of line charts contain explainability issues). We release our analysis as a dataset and methods for further examination by the scientific community.

Keywords: Accessibility; Open Access; Computer Vision

1 Introduction

Figures are an essential part of scientific publications because they can present complex data relationships to readers in an efficient manner. However, figures can have several issues that reduce their communication quality. For example, when they contain color-blind issues, they preclude readers from understanding the underlying trends or even make them misinterpret results (Jambor et al. [10]; Jefferson & Harvey [11]). In some disciplines, editors and reviewers might not pay enough attention to figures (Helsloot [7]), partially because it is time-consuming. Using computational methods to help flag common patterns in figures could thus be important.

In particular, we create three classifiers that detect whether a panel within a figure has color-blind unsafe problems or low light, and whether a line chart has insufficient legend or captions. Our method is validated on a hand-annotated dataset and simulated image datasets, achieving high accuracy on these three tasks. We apply our classifiers to a large sample of 70,000+ publications and 300,000 figures from the PubMed Open Access Subset (see Fig. 1 for the process of our analysis). Our results show that around 2% of all figures contain accessibility issues, 3% of diagnostic figures contain readability issues, and 23% of

Springer

line charts contain explainability issues (see Fig. 2 for example figures with ARE issues). We analyze whether these issues are associated with bibliometric factors such as ranking of the journal, seniority of the researcher, country, and field. Finally, we discuss the applications of this technology to other aspects of good publication practices.

This study has the following two objectives:

1. Propose computational methods to detect accessibility issues (e.g., color-blind unsafe), readability issues (e.g., low light and contrast), and explainability issues (e.g., missing captions and legends, ARE issues).
2. Estimate the prevalence of ARE issues in open access publications and the statistical patterns between journals' bibliometric features and their proportion of articles with ARE issues.

2 Literature review

2.1 Principles of graphical presentation in science

Scientific figures are expected to convey precise information to readers. To achieve this goal, visualization researchers have proposed guidance on figures: figures should be understandable by readers, regardless of their disability (Jambor et al. [10]). Also, scientific figures should be understandable to color-blind and low-vision readers. Figures should contain enough information to be explainable to readers. For example, they should contain legends when necessary and captions. We now review three aspects: accessibility, readability, and explainability.

2.1.1 Accessibility issues

Accessibility in scientific articles is closely related to being "accessible information technology" artifacts, which is defined by the Americans with Disability Act (ADA) as "*technology that can be used by people with a wide range of abilities and disabilities. It incorporates the principles of universal design, whereby each user can interact with the technology in ways that work best for him or her.*" Accessibility issues is rooted in color combinations or color maps, such as the rainbow color map (Borland & Taylor Ii [3]) and can affect how various color-blind readers (such as red-green color-blindness and blue-yellow color-blindness) understand a figure and research findings. This issue should draw the attention of authors and publishers because, in some regions, 14% of middle-aged populations are color-blind (Jafarzadehpur et al. [9]), and from 1.69% to 8.73% of the population is color-blind (Shah et al. [20]). Some researchers have developed technologies to help color-blind readers view electronic documents. For example, researchers developed software to detect images with a rainbow color map (eLife Labs [19]) and a computer interface to assist color-blind computer users (Jefferson & Harvey [12]).

2.1.2 Readability issues

Readability issues prevent the reader from correctly parsing information presented in a figure. Among the factors affecting this ability are low light and contrast images and high complexity images. Low light and contrast issues are especially worrying among the elderly because they have difficulty reading low-contrast images (Owsley et al. [18]) because they are likely to make readers lose details of objects. High complexity images typically need more attention and effort for readers to understand, and this complexity can be quantified by computational methods (Hou & Zhang [8]). However, there is limited literature in vision science about the relationship between saliency map and their readability,

and vision science researchers have conducted multiple experiments on the contrast of images and their readability to viewers with different eye conditions. Interestingly, some researchers have developed a method to enhance low contrast images to improve their quality (Hasikin & Isa [5]). It is unclear whether scientific articles use low contrast and light figures. Since highly-cited scientists are relatively old compared to other scientists (Normile [17]), understanding this prevalence is essential.

2.1.3 Explainability issues

Explainability can be broadly understood as being able to produce an explanation of the trends and factors observed in a scientific figure. For example, some line charts or bar charts need legend or caption to assist readers to understand them because of their usage of colors and graphic design. Also, biological images can be hard to interpret if they do not have a scale bar. Explainability issues affect all the population because they relate to whether a figure contains a good caption or legend. A recent study found that between 5 to 17 percent of scientific figures do not provide enough information to explain the colors inside them (Jambor et al. [10]). Yet, this previously mentioned study has a relatively small sample of figures, and a large-scale analysis on this issue might estimate the prevalence of explainability issues more accurately.

3 Method and material

3.1 Materials

3.1.1 Data and image preprocessing

In this research, we first randomly selected 300,000 figures from 71,508 publications in PudMed Open Access, a subset of PubMed Central, containing millions of publications (see Fig. 1 for details of publications, figures, panels, and types of panels analyzed). We applied compound figure classification and separation to get 788,028 subplots to analyze each image or graph (we used the method described in (Zhuang et al. [26])). We estimated the accessibility of all subplots with computer vision techniques. We estimated the readability of diagnostic figures using a fine-tuned deep learning network called ResNet50 v2 (He et al. [6]). Finally, we estimated the explainability of line charts by fine-tuning a ResNet152 v2 classifier using our annotations.

3.1.2 Annotations

To train the classification and detection models for explainability analysis, we manually annotated 1,407 line charts for legend detection and 1,454 line charts for legend necessary classification. In legend detection, we annotated *where* the legend was in the chart. For legend neediness, we annotated line charts as needed for a legend if there is more than one line or symbol.

3.2 Methods

The following methods are available in the GitHub repository https://github.com/sciosci/ARE-analysis.

3.2.1 Image preprocessing

To analyze these issues in each subplot, we applied the following preprocessing steps.

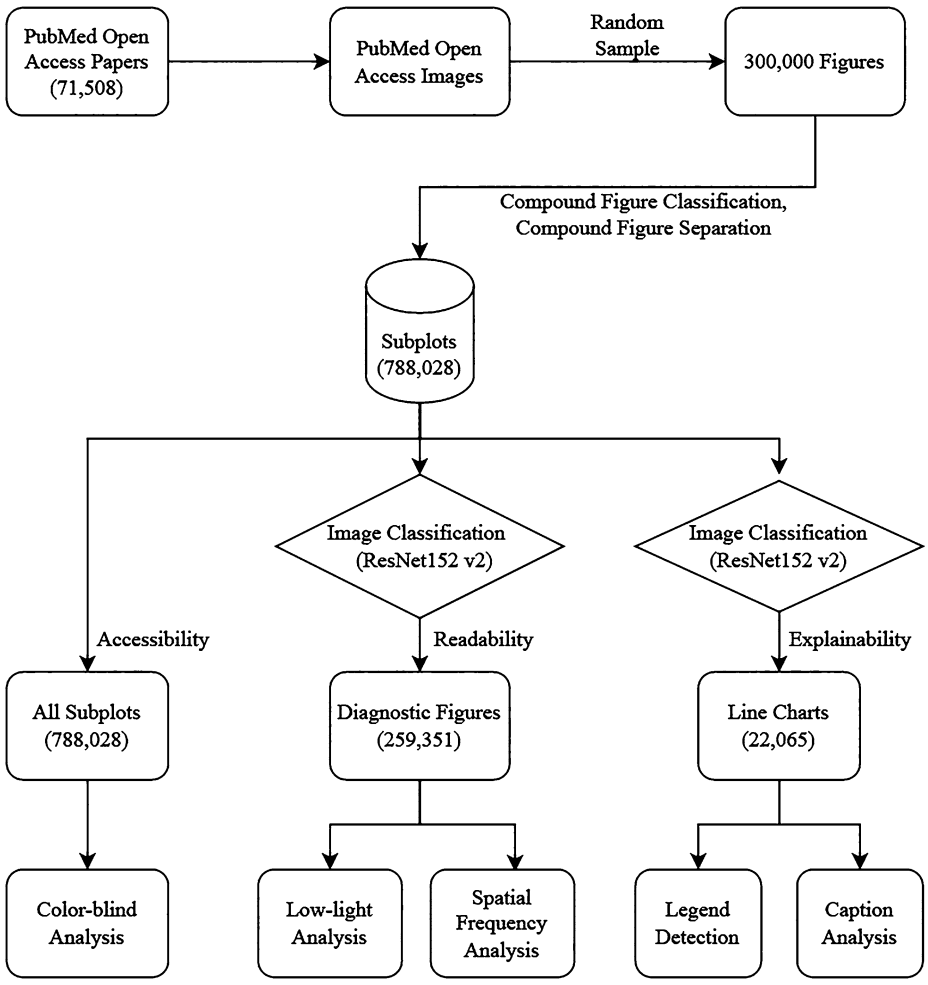

Figure 1 The data processing flow. We use deep learning techniques (ResNet152 v2 and YOLO v4) to detect compound figures, separate subplots, classify all subplots into image categories. Then we use computer vision, deep learning, and other techniques to detect color-blind unsafe figure, low light and contrast figures, and line charts with explainability issues

- *Compound figure classification*. Compound figures are common in publications to show relevant information and results together; nonetheless, we discuss issues on each figure in this work. Thus, we use a convolutional neural network-based model (Resnet-152 v2, pre-trained on ImageNet) as a feature extractor and train its top layer with a compound figure classification dataset from ImageCLEF 2016 (García Seco de Herrera et al. [4]).
- *Compound figure separation*. For compound figures, we separate each subplot from them by a fine-tuned convolutional neural network (YOLO v4 (Bochkovskiy et al. [2]), pre-trained on MS COCO dataset) with a subfigure separation dataset from ImageCLEF 2016 (García Seco de Herrera et al. [4]).
- *Image classification*. We apply a high-quality feature extractor based on a convolutional neural network (Resnet-152 v2, pre-trained on ImageNet) to classify figures: bar charts, line charts, scatter charts, heatmap charts, box charts, area charts,

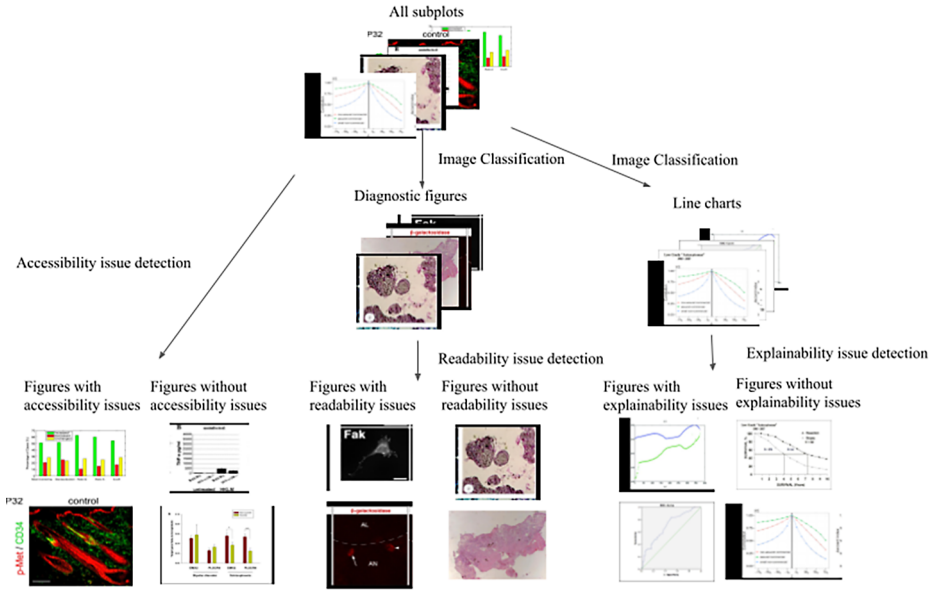

Figure 2 The illustration of results of our computational methods. After we obtain subplots from our image sample and separate subplots into different categories, our techniques detected examples above: color-blind unsafe figure, low light and contrast figures, and line charts with explainability issues

radar plots, maps, pie charts, tables, pareto charts, venn diagrams, violin charts, and diagnostic figures.

For more details, please see (Zhuang et al. [26]).

3.2.2 Accessibility issue detection

There are various kinds of accessibility issues because deuteranomaly color-blind readers are red-green colorblind, tritanopia color-blind readers are blue-yellow color-blind, and other kinds of color-blind issues. In this article, we focus on red-green color-blind issue as a representative of accessibility issues because it is the most common color-blindness. To detect figures with this issue, we obtain figures in color-blind vision systems through a simulation method (Machado et al. [15]). Then, we examine if an original figure contains red and green areas simultaneously, and if the red area disappears in the simulated figure, this figure can be colorblind unsafe. More specially, we first denoise the image and then resize the image to save computational load. Then for each pixel, we compute the distance between its color (RGB value) with colors (blue, green, red, cyan, magenta, yellow, black, white) in lab color space and consider the pixel has one specific color when it is closest to the color. After identifying the color for each pixel, we can compute the size of each color by dividing the total number of pixels in each color by the total number of pixels in the image. Then if both red and green areas are present in an image (more than 2% of the image) and the red area in the simulated image disappears; thus we consider it might be color-blind unsafe.

3.2.3 Readability issue detection

We fine-tuned a convolutional neural network with a low light image dataset (Low Light dataset (Wei et al. [23])) as our classifier to detect figures that have readability issues. Because some scientific figures can be different from natural scenes in our training dataset,

we focus on diagnostic figures in detecting readability issues (see Fig. 1 for our data processing flow, because figures of table or texts can be misclassified as diagnostic figures, we removed them from our analysis). We also applied principles from vision science as a constraint to improve the accuracy of our predictions. More specifically, when the spatial frequency of the object in images is too high (e.g., greater than 30 of spatial frequency. In this article, we take the extreme, 60 of spatial frequency), details of some objects can be hard for some readers to view even with high contrast (Kara et al. [13]; Watson & Ahumada [22]). In our image analysis, the spatial frequency of one image was estimated by transforming the image with Fast Fourier Transform and getting the spatial frequency of pixels. Then we can measure if one image has a large area (greater than half of the image) with high spatial frequency. Taking our low-light image classifier and spatial frequency analysis together, if an image is classified as a low-light image by our classifier and also contains a large area of high spatial frequency, we consider such images with readability issues.

3.2.4 Explainability issue detection

In this work, we focus on line charts as legend is usually necessary when there are two or more groups of data. To estimate the explainability, we have two parts of analysis to estimate the explainability: legend detection and caption analysis. The following are the models and methods to detect legend and analyze caption.

Legend detection we fine-tuned a deep neural network model (YOLO-v4, pre-trained on MS COCO dataset) on human-annotated figures to identify legends on scientific figures (0.88 precision, 0.84 recall, and 0.936 mean average precision on testing set). Some compound figures may only contain one legend and apply to all subplots, so we also detect if a legend exists in original compound figures of corresponding subplots.

Legend neediness classification Sometimes, the legend is not always necessary when there is only one line in the graph. To filter this situation, we find-tune a convolution neural network (ResNet152v2, pre-trained on ImageNet) on human-annotated charts to classify if figures need legend or not (0.73 precision, 0.81 recall on testing set).

Caption analysis Scientific figures might have explanations in corresponding captions; therefore, we used Pubmed Parser to extract captions from the Pubmed dataset and check if color or symbol explanations exist (e.g. blue, red, green, dashed line, solid line, triangle, square, etc.).

Lastly, we classify a line chart to have explainability issues if we could not detect a legend and explanation in the caption for legend needed charts.

4 Result

4.1 Methods performance

Accessibility issue detection: to examine the performance of method, we generated two set of color-blind unsafe and safe figures: a set of images with random shapes and a set of academic charts, such as bar charts, box charts, and other charts to validate our method. For the test images with random shapes (1,120 images), our method shows a precision of 1.0 and a recall of 0.56. For the test academic charts (990 images), our method has a precision of 0.98 and a recall of 0.42.

Readability issue detection: we validated our method against the test split (15 pairs of low-light images and normal images) in the low light image (LOL) dataset and our method achieved a precision of 1.0 and a recall of 1.0 in detecting low light images. As mentioned in the method section, we applied one vision principle as a constraint to the detected images to only keep images of low light and contrast.

Explainability issue detection: To understand how the models work generally, we randomly sampled 100 line charts to evaluate errors in prediction against human experts. In this evaluation, we examine the performance of our two main steps in this task (see method section for more details): legend detection in line chars and legend neediness in line chart. Our legend detection has a precision of 0.94 and recall of 0.86. Also, our legend neediness classification has a precision of 0.85 and a recall of 0.85.

4.2 Bibliometric analysis of ARE issues in open access publications

All of our results use the journal as the unit of analysis. We predict ARE issues in figures first, and we conducted linear regression for ARE issues separately, in the level of journals. More specifically, we aggregate images into publication level first: one publication is considered with ARE issues if any of its figures have ARE issues. Then we get the percentage of publications with ARE issues by aggregating the publications (we exclude journals in our sample with publications of fewer than 3). So each data point in the linear regression is a journal. We measure country, the field of study, journal's rank, number of publications in a journal, journal's age, h-index of authors, and author's academic age. We use a threshold of 5% for all regression analyses to assess statistical significance for all regression analyses. We confirm with a GVIF analysis whether the independent variables are excessively correlated, using a maximum of GVIF 5 (Montgomery [16]), which all the regression analyses passed.

We first want to understand how the bibliometric factors (see Table 1) are distributed by themselves (Fig. 3). After matching publications in our sample to Microsoft Academic Graph and removing outliers (see Materials and Methods), we analyze 57,837 publications from 1,818 journals from 1966 until 2018. The three most popular publication countries are the United States, China, and Great Britain. The three most popular fields are Biology, Medicine, and Chemistry.

Table 1 bibliometric features explanations and computation methods. We use MAG to produce the features in this table

Bibliometric feature	Computational method
Field of study	We estimate this feature for each journal by aggregating their publications' field of study, which is predicted by Microsoft Academic Graph (MAG) with hierarchical topic modeling
The number of publications	We estimate this feature for each journal by counting unique publications in these journals from MAG
Average h-index of Authors	We estimate this feature for each journal by computing h-index of authors in MAG and then take the average of the authors in each journal
Journal Rank	The rank of a journal is computed by MAG based on the citation network of publications. The more top-ranked journals (the smaller the rank is) get more citations from the citation network
Author academic age	We estimate this feature for each journal by computing the number of years to 2022 since the authors' year of first publications in MAG and then take the average of the authors for each journal
Journal age	The number of years to 2022 since the year of the journal's first publication in MAG

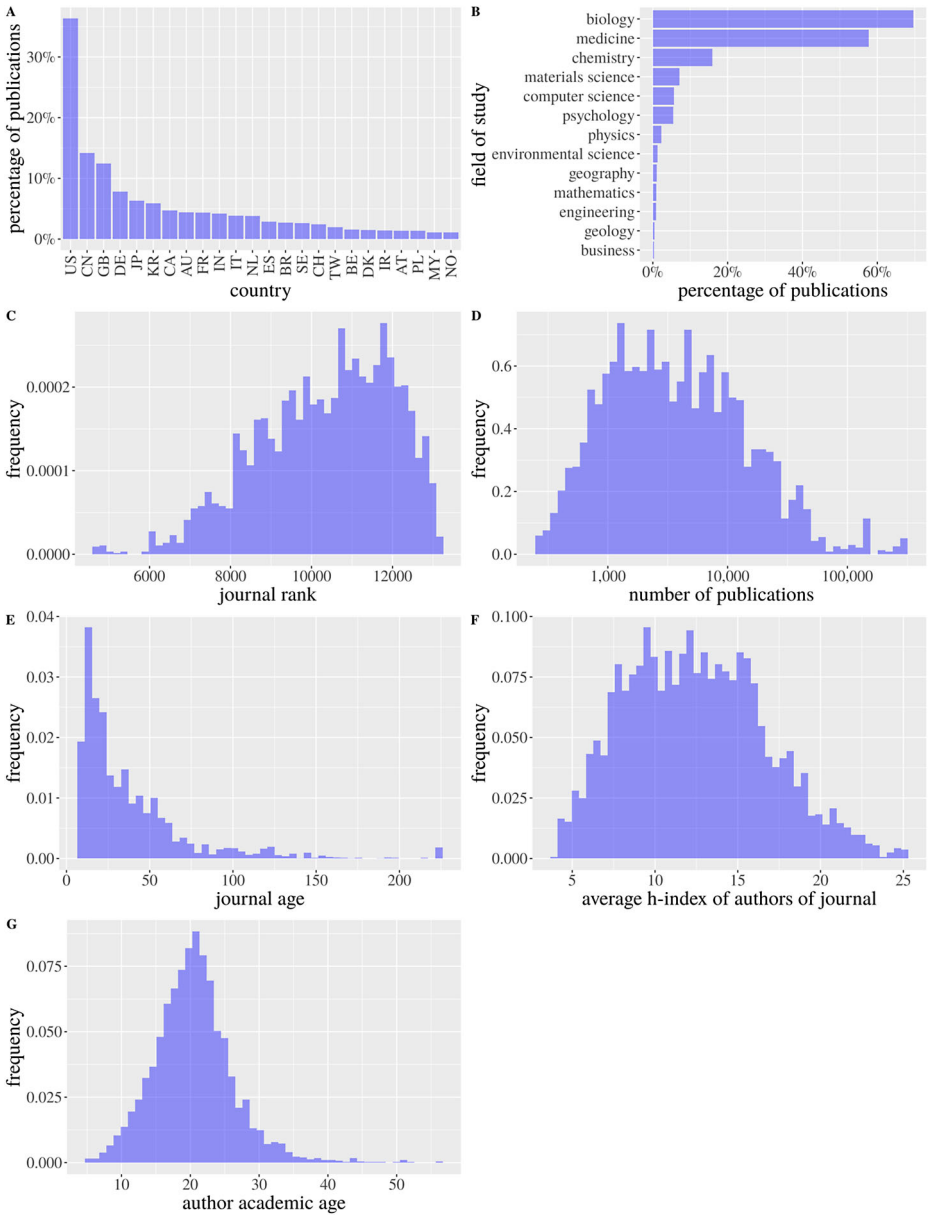

Figure 3 Distribution of journals' affiliation country, the field of study, rank, number of publications, age, average author h-index, and author academic ag

The journal rank is based on the PageRank of the citations to the journal, as calculated by the Microsoft Academic Graph (see Materials and Methods). The average rank of journals is 10,353, with an SD = 1,638. The average number of publications by a journal is 10,193.22, with a minimum of 248 and a maximum of 276,186. The average age of a journal is 37.41, with the newest being seven years old and the oldest being 222 years old. The average *h*-index of authors in journals is 12.47, with an SD of 4.19. Finally, the average academic age is 20.48 years, with an SD of 5.79 years.

Before delving into analyses of journals and ARE issues, we wanted to understand how the variables are related to each other through a correlation analysis (Fig. 4). We codify

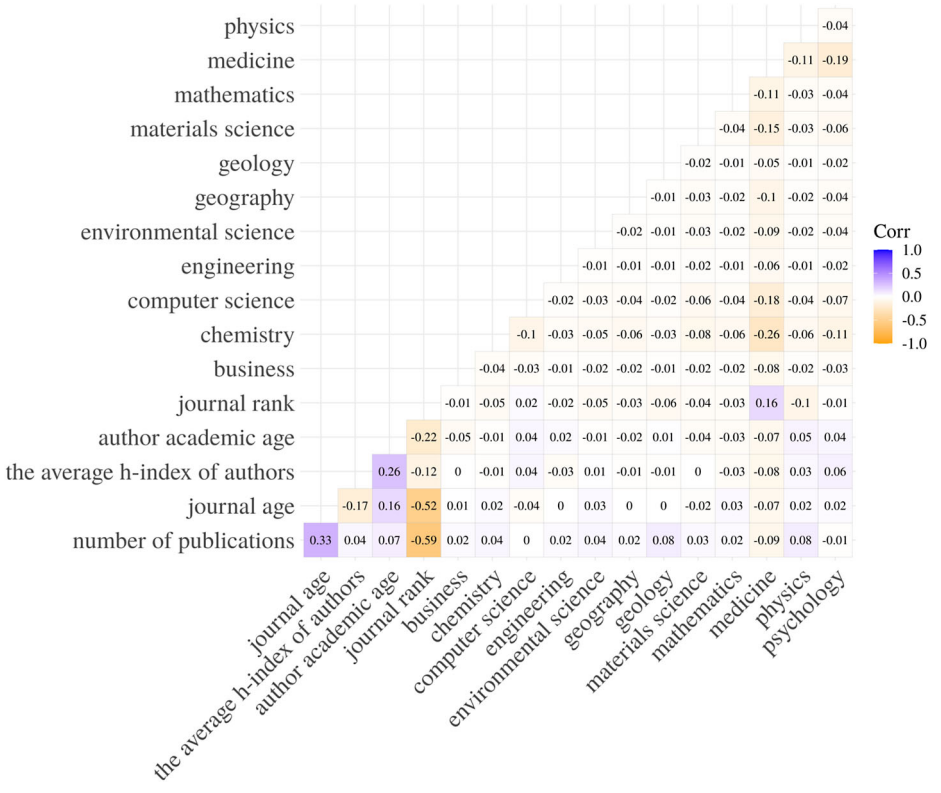

Figure 4 Correlation matrix of predictors. Only a few of predictors have a relatively high correlation between each other, such as journal age and journal rank

the field as dummy variables. The most correlated features are the journal's rank, number of publications, and journal age. Journal rank is numerically high if the journal is not cited as often as a journal with a rank numerically smaller. These correlations are negative, meaning that top-cited journal cites produce more papers and are older.

4.3 Accessibility issues

We analyzed accessibility issues in scientific publications and found that 2% of scientific figures (788,028 figures in our sample, and see Fig. 1 for more details) contain color-blind unsafe figures. We use multiple linear regression analysis to examine if journals' bibliometric factors associated with the percentage of their publications with accessibility issues. We found that the journals in business and in engineering fields published the least and most color-blind unsafe figures, respectively (Table 2). Interestingly, we found that journals' average author h-index has the biggest coefficient (standardized coef = 0.20, $t(3895) = 11.43$, $p < 0.001$), suggesting that journals with highly-cited authors have a higher proportion of articles with accessibility issues. This is somewhat counteracted by the negative effect of authors' academic age (standardized coef = −0.057, $t(3895) = -3.41$, $p < 0.001$). Older journals have a lower proportion of articles with accessibility issues (standardized coef = −0.0493, $t(3895) = -2.6251$, $p = 0.0087$) (see Fig. 5).

Table 2 Linear regression (standardized coefficients) on journals' proportion of publications with accessibility issues

Term	Estimate (Standardize*)	Std error	*t* value	p-value
(Intercept)	0.030	0.031	0.957	0.339
Field:				
Business	−0.482	0.146	−3.308	0.001
Chemistry	0.087	0.054	1.599	0.110
Computer science	0.010	0.069	0.142	0.887
Engineering	0.523	0.182	2.876	0.004
Environmental science	−0.250	0.128	−1.948	0.052
Geography	−0.394	0.112	−3.505	0.001
Geology	0.090	0.244	0.370	0.711
Materials science	0.464	0.082	5.631	<0.001
Mathematics	−0.030	0.104	−0.289	0.772
Medicine	−0.112	0.041	−2.746	0.006
Physics	0.376	0.114	3.299	0.001
Psychology	−0.204	0.068	−2.995	0.003
Number of publications (log)	−0.048	0.047	−1.021	0.307
Author average h-index	0.200	0.018	11.431	<0.001
Journal rank (log)	−0.100	0.050	−2.004	0.045
Author academic age	−0.058	0.017	−3.406	0.001
Journal age	−0.049	0.019	−2.625	0.009

$F(17, 3895) = 18.547$, $p < 0.001$.

$R^2 = 0.075$, Adj. $R^2 = 0.070$.

*: only the independent variables are standardized.

4.4 Readability

Readability issues (low light and contrast images) can also affect older readers. The elderly population is likely to have difficulty reading images with low contrast (Owsley et al. [18]). Our research community must be aware of the prevalence of low contrast and light scientific figures and reduce these issues. We develop a model to automatically assess low contrast and low light in images (see Materials and Methods). We find that 3% of medical diagnostic figures (259,351 diagnostic figures, see Fig. 1 for details) have low light and contrast. We use multiple linear regression analysis to examine if journals' bibliometric factors associated with the percentage of their publications with readability issues. We found that journals in physics and business fields published the most and least figures with readability issues, respectively (Table 3). We found that journals' average author h-index has the biggest coefficient (standardized coef = 0.20, $t(3228) = 10.20$, $p < 0.001$), suggesting that journals with highly-cited authors have a higher proportion of articles with readability issues. Also, older journals have a higher proportion of articles with readability issues (standardized coef = 0.066, $t(3228) = 3.16$, $p = 0.002$) (see Fig. 6).

4.5 Explainability

Another issue in scientific figures is explainability: some figures have no legend even if they have multiple colors, symbols, or lines or have no caption. In this study, we focus this analysis on line charts only because they typically need an explanation. We develop a method that splits the classification into detecting legend and whether the caption exists and contains legend-related information (see Methods). We found that the explainability issues with line charts are very high: around 23% of them (22,065 line charts, see Fig. 1 for more details) lacking legends and color explanations in their captions. While our method is not perfect, in the worst-case scenario (see Methods for performance evaluation), we would

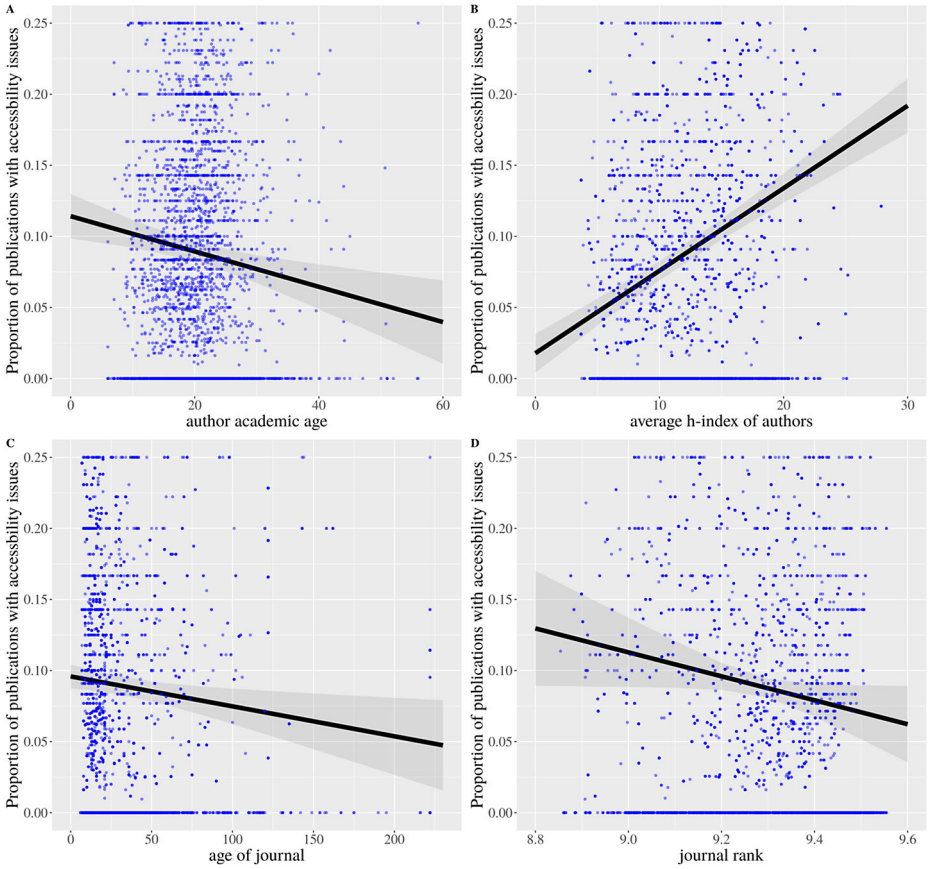

Figure 5 Statistically significant marginal effects on journals' proportion of publications with accessibility issues. (**A**) journal average author academic age's marginal effect (**B**) journal average author h-index's marginal effect (**C**) journal age's marginal effect (**D**) journal rank's marginal effect

predict the prevention of explainability issues to be surprisingly high still. We use multiple linear regression analysis to examine if journals' bibliometric factors associate with the percentage of their publications with explainability issues. We found that journals' average author h-index has the biggest coefficient (standardized coef = −0.20, $t(2191) = -8.77$, $p < 0.001$), suggesting that journals with highly-cited scientists have a lower proportion of articles with explainability issues (Table 4 and Fig. 7).

5 Discussion

This paper developed a method to detect accessibility in broad terms. We focused on color-blind unsafe (Accessibility), low contrast and light (Readability), and missing legend and caption (Explainability). Our method combines computer vision, machine learning and human vision principles in our methods and achieves high accuracy. We further investigate bibliometric trends about scientific figures with ARE issues. At the journal level, a multiple regression analysis revealed that the author's h-index and academic age, journal rank, and journal age are all significantly predictive of one of the ARE factors. We found that a surprisingly large proportion of line charts lack a legend and proper caption and journals with higher average author h-index are less likely to have publications with this issue. However, journals with higher average author's h-index are more likely to have

Table 3 Linear regression (standardized coefficients) on journals' proportion of publications with readability issues

Term	Estimate (Standardize*)	std error	*t* value	p-value
(Intercept)	0.044	0.034	1.3	0.194
Fields:				
Business	−0.489	0.160	−3.065	0.002
Chemistry	0.034	0.058	0.595	0.552
Computer science	−0.114	0.073	−1.571	0.116
Engineering	0.111	0.194	0.570	0.569
Environmental science	−0.422	0.139	−3.036	0.002
Geography	−0.372	0.124	−3.006	0.003
Geology	−0.401	0.254	−1.579	0.114
Materials science	0.043	0.085	0.498	0.618
Mathematics	−0.091	0.109	−0.831	0.406
Medicine	−0.058	0.046	−1.277	0.202
Physics	0.180	0.118	1.534	0.125
Psychology	−0.068	0.079	−0.866	0.387
Number of publications (log)	0.037	0.051	0.743	0.458
Author average h-index	0.198	0.019	10.196	<0.001
Journal rank (log)	−0.021	0.054	−0.400	0.689
Author academic age	−0.005	0.019	−0.270	0.788
Journal age	0.066	0.021	3.157	0.002

$F(17, 3228) = 12.4870$, $p < 0.001$.

$R^2 = 0.0617$, Adj. $R^2 = 0.0568$.

*: only the independent variables are standardized.

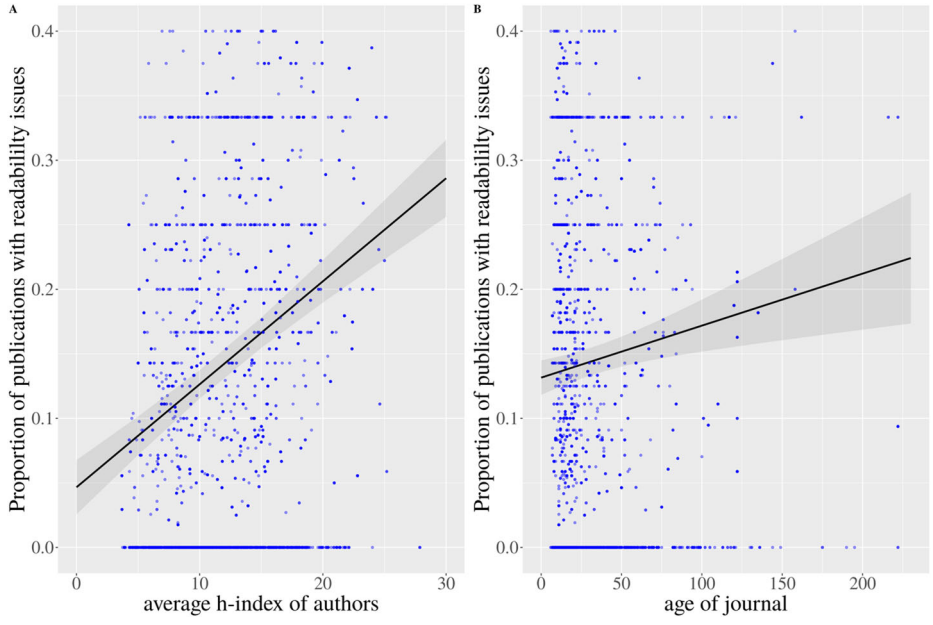

Figure 6 Statistically significant marginal effects on journals' proportion of publications with readability issues. (**A**) journal average author h-index's marginal effect. (**B**) journal age's marginal effect

publications with accessibility and readability problems. A plausible explanation is that these journals tend to be more interdisciplinary and accept standards from many more disciplines, making them more vulnerable to these kinds of issues. Given the limitation of our sample size, some journals cannot be well represented. Thus, our linear regression might not predict specific journals accurately. In sum, our results show that AI promises

Table 4 Linear regression (standardized coefficients) on journals' proportion of publications with explainability issues

Term	Estimate (Standardize*)	std error	t value	p-value
(Intercept)	−0.055	0.041	−1.329	0.184
Fields				
Business	0.243	0.175	1.391	0.164
Chemistry	−0.003	0.065	−0.052	0.959
Computer science	0.103	0.084	1.234	0.217
Engineering	−0.054	0.214	−0.251	0.802
Environmental science	0.249	0.147	1.686	0.092
Geography	0.123	0.136	0.905	0.365
Geology	0.463	0.281	1.65	0.099
Materials science	0.104	0.094	1.102	0.27
Mathematics	−0.001	0.118	−0.004	0.997
Medicine	0.069	0.058	1.185	0.236
Physics	0.027	0.122	0.22	0.826
Psychology	0.145	0.091	1.597	0.11
Number of publications (log)	−0.092	0.055	−1.67	0.095
Average author h-index	−0.201	0.023	−8.77	<0.001
Journal rank (log)	0.096	0.058	1.649	0.099
Author academic age	0.001	0.022	0.026	0.979
Journal age	−0.023	0.024	−0.958	0.338

$F(17, 2191) = 12.4265$, $p < 0.001$.

$R^2 = 0.0879$, Adj. $R^2 = 0.0809$.

*: only the independent variables are standardized.

Figure 7 Journal average author h-index's significant marginal effects on journals' proportion of publications with explainability issues

to be a useful tool for analyzing ARE factors at scale and that we can use these findings to understand patterns of problems in scientific figures.

Previous research has mostly used manual analysis to understand ARE issues. Although human experts can detect these issues accurately, they can be slow. Our results open the door to conducting a much larger scale analysis and increase the power of statistical anal-

yses. The method we propose is general enough to accommodate other types of analyses. Previous research has shown how to detect proportional ink issues (Zhuang et al. [26]), figure manipulation (Xiang & Acuna [24]), figure duplication (Acuna et al. [1]), and citation missingness (Zeng & Acuna [25]). Other researchers have also developed many other analyses that journals could automate (e.g. (Kroon et al. [14])), and they have a big opportunity for automation.

Older researchers are less likely to produce publications with color-blind safety issues. Researchers with a higher impact are less likely to produce publications that need legends or explanations in captions. Interestingly, we did not find that either senior researchers or high-impact journals are less likely to produce publications with low contrast and light. These analyses are somewhat limited in that we focused on the journal level. In the future, we will analyze whether co-authorship, temporal, and training factors are at play.

We believe our method and AI techniques have the potential to help people with vision disabilities. For example, one of our methods can measure the light of an image, and AI techniques could automatically improve the light of images (Wang et al. [21]). Similarly, our method measures whether a line chart needs a legend, and new AI techniques might automatically create such annotation based on the text citing the figure. There is a potential to collaborate with researchers in Human-Computer Interaction.

Our analysis shows that some disciplines have more accessibility and readability issues, which can be hard to solve because of the nature of the figures. However, there is reasonable room for research communities to remedy such issues and help viewers, especially with color blindness or other vision deficiencies. We recommend authors highlight the name of colors in figure captions if their figures can be color-blind unsafe. We also suggest publishers consider implementing an editorial policy to encourage authors to underline their use of color-unsafe figures in their manuscripts.

Our techniques make mistakes, which can introduce errors to our estimation of the prevalence of accessibility, readability, and explainability issues. Also, some images in our sample might with low quality and could not be analyzed by our method. These limitations in our detection might affect our regression analysis because our dependent variable, the percentage of ARE issues in a journal, is affected by these errors.

Our analysis only focuses on a handful of accessibility, readability, and explainability issues. The choices that we made to analyze low light, lack of legends, and so on were because of computational convenience. The ultimate goal would be to analyze a manuscript automatically and measure all the missing pieces that make it inaccessible to the many limitations in human perception and cognition. However, we believe the choices provide a playground to test our ideas and show that it is possible to use AI to measure these factors. In the future, we will continue expanding these analyses.

Some of our automated judgment was based on principles of human vision, while we did not test whether human annotators would agree with these predictions. For example, we computed spatial frequency and taking thresholds that, on average, make those frequencies hard to detect for humans. However, the field of human vision is much more complex than these functions, and low contrast depends on many factors beyond our control. For example, we do not know how far readers will read the article or under which lighting conditions. We think this opens opportunities for collaborating with vision researchers and device designers.

6 Conclusions

In this article, we use AI to analyze accessibility, readability, and explainability issues. Our method is based on a combination of classifiers, and it works relatively well. We found several bibliometric factors at the journal level are predictive of these ARE issues.

We believe one of the obvious next steps is to fix figures based on our results automatically. Our method detects ARE factors, but there is no reason to prevent us from touching up and generating images that make all these issues disappear. Perhaps a tool like this could be automatically incorporated into PDF viewers or screen readers for low vision scientists. Or they could be part of statistical packages that could warn the user if a figure contains an ARE issue.

Our method is unique enough to highlight how journals and reviewers can use AI to analyze figures for accessibility factors. Our method is accurate, fast, and allows us to understand patterns of ARE issues. While there is still much room for improvement, our method and results pave the way for future research. And to the best of our knowledge, this research is one of the first ones to be done.

Acknowledgements
Not applicable.

Funding
HZ and DEA were funded by NSF award #1800956, and the US Office of Research Integrity (ORI) awards ORIIR180041, ORIIIR190049, ORIIIR200052, and ORIIIR210062. HZ was funded by the Syracuse University Dissertation Fellowship and the 2022 Summer Fellowship of the School of Information Studies at Syracuse University. TYH was partially funded by the DHHS-ORI grants ORIIR180041 and ORIIIR190049. DEA was funded by the Alfred P. Sloan Foundation grant #G-2020-12618.

Abbreviations
ARE issues, accessibility, readability and explainability issues; ADA, Americans with Disability Act; MAG, Microsoft Academic Graph; LOL dataset, LOw-Light image dataset.

Availability of data and materials
Code is available at the GitHub repository https://github.com/sciosci/ARE-analysis.

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
HZ designed and performed all analyses and prepared the initial manuscript. TYH conducted analysis on figure legend and caption. DEA envisioned the project, acquired funding and designed all analyses. All authors contributed to editing the paper. All authors read and approved the final manuscript.

Authors' information
HZ is a Ph.D. Candidate at Syracuse University. TYH worked on this study as a graduate student at Syracuse University and is an analytic at Amazon Inc. DEA is an associate professor at University of Colorado at Boulder.

Author details
[1]School of Information Studies, Syracuse University, Syracuse, USA. [2]Amazon, Inc., New York, USA. [3]Department of Computer Science, University of Colorado at Boulder, Boulder, USA.

Publisher's Note

Received: 28 May 2022 Accepted: 14 February 2023 Published online: 02 March 2023

References

1. Acuna DE, Brookes PS, Kording KP (2018) Bioscience-scale automated detection of figure element reuse. bioRxiv. https://doi.org/10.1101/269415
2. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) YOLOv4: optimal Speed and Accuracy of Object Detection. http://arxiv.org/abs/2004.10934

3. Borland D, Ii R (2007) Rainbow color map (still) considered harmful. IEEE Comput Graph Appl. 27(2):14–17
4. De Herrera A, Bromuri S, Schaer R, Müller H (2016) Overview of the medical tasks in ImageCLEF 2016. CLEF working notes. Evora, Portugal
5. Hasikin K, Isa NAM (2012) Enhancement of the low contrast image using fuzzy set theory. In: 2012 UKSim 14th international conference on computer modelling and simulation, pp 371–376. https://doi.org/10.1109/UKSim.2012.60
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90
7. Helsloot A (2019) In the eye of the beholder. Elsevier connect. https://www.elsevier.com/connect/authors-update/in-the-eye-of-the-beholder
8. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. 2007 IEEE Conference On Computer Vision And Pattern Recognition
9. Jafarzadehpur E, Hashemi H, Emamian MH, Khabazkhoob M, Mehravaran S, Shariati M, Fotouhi A (2014) Color vision deficiency in a middle-aged population: the shahroud eye study. Int Ophthalmol 34(5):1067–1074. https://doi.org/10.1007/s10792-014-9911-2
10. Jambor H, Antonietti A, Alicea B, Audisio TL, Auer S, Bhardwaj V, Burgess SJ, Ferling I, Gazda MA, Hoeppner LH (2021) Creating clear and informative image-based figures for scientific publications. PLoS Biol 19(3):e3001161
11. Jefferson L, Harvey R (2006) Accommodating color blind computer users. In: Proceedings of the 8th international ACM SIGACCESS conference on computers and accessibility, pp 40–47. https://doi.org/10.1145/1168987.1168996
12. Jefferson L, Harvey R (2007) An interface to support color blind computer users. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1535–1538. https://doi.org/10.1145/1240624.1240855
13. Kara S, Gencer B, Ersan I, Arikan S, Kocabiyik O, Tufan HA, Comez A (2016) Repeatability of contrast sensitivity testing in patients with age-related macular degeneration, glaucoma, and cataract. Arq Bras Oftalmol 79:323–327. https://doi.org/10.5935/0004-2749.20160092
14. Kroon C, Breuer L, Jones L, An J, Akan A, Ali EAM, Busch F, Fislage M, Ghosh B, Hellrigel-Holderbaum M, Kazezian V, Koppold A, Restrepo CAM, Riedel N, Scherschinski L, Gonzalez FRU, Weissgerber T (2022) Blind spots on western blots: a meta-research study highlighting opportunities to improve figures and methods reporting. bioRxiv. https://doi.org/10.1101/2022.05.13.491761
15. Machado G, Oliveir M, Fernandes L (2009) A physiologically-based model for simulation of color vision deficiency. IEEE Trans Vis Comput Graph 15(6):1291–1298
16. Montgomery D (2012) Introduction to linear regression analysis. https://people.duke.edu/~rnau/regintro.htm
17. Normile D (2004) Older scientists win majority of funding. Science 303(5665):1746–1746. https://doi.org/10.1126/science.303.5665.1746a
18. Owsley C, Sekuler R, Boldt C (1981) Aging and low-contrast vision: face perception. Investig Ophthalmol Vis Sci 21(2):362–365
19. Saladi S (2019) JetFighter: towards figure accuracy and accessibility. eLife
20. Shah A, Hussain R, Fareed M, Afzal M (2013) Prevalence of red-green color vision defects among muslim males and females of manipur, India. Iran J Pub Health 42(1):16–24
21. Wang Y, Wan R, Yang W, Li H, Chau L, Kot A (2022) Low-light image enhancement with normalizing flow. Proc AAAI Conf Artif Intell
22. Watson AB, Ahumada AJ Jr (2005) A standard model for foveal detection of spatial contrast. J Vis 5(9):6. https://doi.org/10.1167/5.9.6
23. Wei C, Wang W, Yang W, Liu J (2018) Deep retinex decomposition for low-light enhancement. ArXiv Preprint arXiv:1808.04560
24. Xiang Z, Acuna DE (2020) Scientific image tampering detection based on noise inconsistencies: a method and datasets. http://arxiv.org/abs/2001.07799
25. Zeng T, Acuna DE (2020) Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. Scientometrics 124:399–428
26. Zhuang H, Huang T-Y, Acuna DE (2021) Graphical integrity issues in open access publications: detection and patterns of proportional ink violations. PLoS Comput Biol 17(12):e1009650. https://doi.org/10.1371/journal.pcbi.1009650