How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?

Supplementary Materials

A OVERVIEW

In this file, we provide the following appendices:

- **Appendix B**: The detailed description of the model and dataset we used as well as examples of rose-based explanations;
- **Appendix C**: The pre-experiment questionnaire;
- **Appendix D**: The post-experiment questionnaire;
- **Appendix E**: The additional results.

We also provide other files and materials:

- **examples.zip**: The file contains the examples of explanations we generated for the experiment.
- **alternative-representations.pdf**: The file contains the other alternative designs we considered for representing features.
- **sub_experiment-1-images.mp4**: The video file shows the flow of the first sub-experiment. It used image representations, but the first sub-experiment could use rose representations.
- **sub_experiment-2-roses.mp4**:The video file shows the flow of the second sub-experiment. It used rose representations, but the second sub-experiment could use image representations.

B GENERATING VISUAL EXPLANATIONS

B.1 The Classification Model

To train a supervised classification model, we first used stratified random sampling to split the modified data set into a training and a validation set. We followed a 3-step pipeline: (1) normalization, (2) principal components analysis, and (3) classification using a support vector machine. We searched a space of (2 SVM kernels × 10 SVM C-values × 10 SVM random seeds × 3 normalization techniques) = 600 unique parameter combinations, and selected the model with the best performance after 3-fold cross validation (parameters: $C = 0.4$, $kernel = linear$, and $normalization = MaxAbsScaler$). This best model had an F_1 score of 0.81 and 0.71 on the training and validation data respectively, which is fairly good given the low average number of training examples per class.

B.2 The Modified Dataset

In our modified dataset, each instances has one of the following 10 classes, with the number of instances noted in parentheses: Birch (14), Linden (13), Bougainvillea (13), Hazel (13), Spindle (12), Hackberry (12), Nettle (12), Primrose (12), Chestnut (12), or Saucer Magnolia (12); excluding English Oak, Cork Oak, Maple, and Boxwood because of their (in)distinctiveness.

B.3 Rosed-based Explanations

We show examples of rose-based explanations in Figure B.1.

C PRE-EXPERIMENT QUESTIONNAIRE

Do you currently do data analysis as part of your job at {withhold for review}?

C.1 Part 1a (if answering yes to the first question)

Please answer all the questions below.

- What is the current domain in which you conduct data analysis?
- How long have you been working in this domain?
- How would you classify yourself as a data analyst?
- How large is the typical dataset you work with?
- What is the dimensionality of the usual dataset you typically work with?
- Are you familiar with supervised machine learning?
- What systems do you use to analyze your data?
- Do these systems include automatic analysis (e.g., machine learning, artificial intelligence)?
  - (if yes to the previous question)
  - Which system(s) provide automatic analysis?
  - Pick one of the above systems that you like the most or you are most familiar with.
  - Regarding your pick, to what extent do you trust the automated analysis of this system?
  - Regarding your pick, do you understand that how the automated analysis of this system works?
- Are you interested in a system that provides automatic analysis for you?
- Is it important for you to understand how the (above) system works?

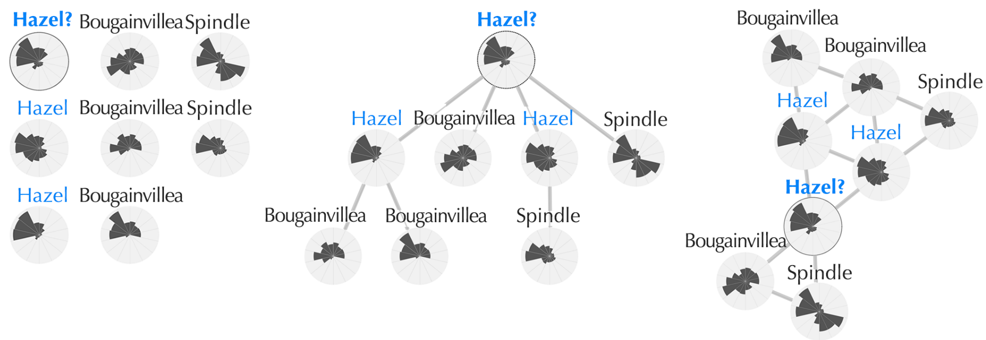
C.2 Part 1b (if answering no to the first question)

Please answer all the questions below.

- Please briefly describe your work at {withhold for review}.
- Do these systems include automatic analysis (e.g., machine learning, artificial intelligence)?
  - (if yes to the previous question)
  - Which system(s) provide automatic analysis?
  - Pick one of the above systems that you like the most or you are most familiar with.
  - Regarding your pick, to what extent do you trust the automated analysis of this system?
  - Regarding your pick, do you understand that how the automated analysis of this system works?
- Are you interested in a system that provides automatic analysis for you?
- Is it important for you to understand how the (above) system works?

C.3 Part 2

Consider the recommendation systems you might use in your daily life: your email spam blocker, Netflix recommendations for movies and TV shows, Facebook recommendations for friends, Amazon recommendations for products, etc.



(a) grid **(b) tree** **(c) graph**

Figure B.1: The examples of rose-based explanations, corresponding to the three visual explanations in Figure 1.

Please rate the statements below on a 7 point scale from *Strongly disagree* (1) to *Strongly agree* (7).

- I usually trust automated systems until there is a reason not to.
- For the most part, I distrust automated systems.
- In general, I would rely on an automated system to assist me.
- My tendency to trust automated systems is high.
- It is easy for me to trust automated systems to do their job.
- I am likely to trust an automated system even when I know little about it.

D POST-EXPERIMENT QUESTIONNAIRE

Please answer all the questions below. If you don't understand a specific question or answer, feel free to ask the experimenter(s). Please consider the experiment today.

- To what extent are you comfortable with identifying these instances?
- Which classifier do you like the most for the tasks today?
- To what extent were you comfortable with each classifier in the experiment today?
- To what extent did you understand each explanation in the experiment today?
- How much do you think you learned about classifying these instances by yourself?
- How did you adjust trust meter (when did you decide to decrease/increase your trust)?
- How did you compare different instances (what patterns did you look at)?
- Any additional comments?

E ADDITIONAL RESULTS

We provide additional results (Cohen's d) from the experiment.

- Table E.1: the 95% bootstrap confidence intervals of Cohen's d for RQ1, corresponding to Figure 5;
- Table E.2: the 95% bootstrap confidence intervals of Cohen's d for RQ2, corresponding to Figure 6;
- Table E.3: the 95% bootstrap confidence intervals of Cohen's d for RQ3, corresponding to Figure 7;
- Table E.4: the 95% bootstrap confidence intervals of Cohen's d for RQ5, corresponding to Figure 9.

Table E.1: Cohen's d and 95% bootstrap CIs for RQ1 (Figure 5)

	Appropriate trust	Overtrust	Undertrust	Self-confidence	
grid - none	0.65 [0.44, 0.86]	-0.66 [-0.95, -0.33]	-0.38 [-0.54, -0.19]	0.74 [0.30, 1.07]	**images**
tree - none	0.70 [0.46, 0.93]	-0.59 [-0.92, -0.24]	-0.41 [-0.58, -0.22]	0.75 [0.37, 1.07]	
graph - none	0.65 [0.43, 0.87]	-0.63 [-0.94, -0.25]	-0.37 [-0.54, -0.13]	0.56 [0.23, 0.84]	
g/t/g - none	0.70 [0.47, 0.89]	-0.65 [-0.95, -0.30]	-0.39 [-0.55, -0.21]	0.70 [0.37, 1.02]	
grid - none	0.46 [0.13, 0.75]	-0.63 [-1.11, -0.11]	-0.075 [-0.41, 0.28]	0.83 [0.54, 1.12]	**roses**
tree - none	0.84 [0.58, 1.11]	-1.01 [-1.66, -0.42]	-0.24 [-0.54, 0.083]	0.84 [0.54, 1.11]	
graph - none	0.57 [0.23, 0.91]	-0.62 [-1.07, -0.14]	-0.18 [-0.48, 0.17]	0.80 [0.46, 1.08]	
g/t/g - none	0.67 [0.40, 0.95]	-0.81 [-1.39, -0.21]	-0.17 [-0.48, 0.16]	0.85 [0.54, 1.12]	

**g/t/g stands for "grid/tree/graph."*

Table E.2: Cohen's d and 95% bootstrap CIs for RQ2 (Figure 6)

	Appropriate trust	Overtrust	Undertrust	Self-confidence	Helpfulness	
grid - tree	0.092 [-0.28, 0.45]	-0.085 [-0.48, 0.27]	-0.077 [-0.39, 0.30]	0.21 [-0.20, 0.64]	0.41 [0.01, 0.74]	images
tree - graph	-0.058 [-0.39, 0.31]	0.063 [-0.31, 0.44]	0.032 [-0.34, 0.35]	0.16 [-0.20, 0.49]	0.36 [0.01, 0.72]	
graph - grid	-0.057 [-0.44, 0.30]	0.037 [-0.33, 0.37]	0.044 [-0.31, 0.38]	-0.34 [-0.68, 0.065]	-0.62 [-0.96, -0.22]	
grid - tree	-0.59 [-0.93, -0.24]	0.40 [0.028, 0.77]	0.34 [0.00, 0.66]	0.045 [-0.31, 0.39]	-0.044 [-0.38, 0.32]	roses
tree - graph	0.22 [-0.13, 0.52]	-0.30 [-0.67, 0.063]	-0.047 [-0.39, 0.30]	0.037 [-0.32, 0.39]	0.41 [0.066, 0.71]	
graph - grid	0.24 [-0.11, 0.62]	-0.10 [-0.46, 0.24]	-0.21 [-0.60, 0.17]	-0.075 [-0.42, 0.28]	-0.22 [-0.54, 0.17]	

Table E.3: Cohen's d and 95% bootstrap CIs for RQ3 (Figure 7)

	Appropriate trust	Overtrust	Undertrust	Self-confidence	Helpfulness
none	1.35 [0.85, 1.85]	-1.56 [-2.12, -1.10]	-0.73 [-1.07, -0.31]	1.14 [0.82, 1.45]	na
g/t/g	1.90 [1.60, 2.20]	-1.63 [-1.88, -1.40]	-1.13 [-1.33, -0.90]	1.43 [1.23, 1.64]	0.64 [0.40 , 0.84]
grid	2.15 [1.51, 2.93]	-1.63 [-2.14, -1.12]	-1.38 [-1.77, -1.01]	1.71 [1.19, 2.21]	0.90 [0.44 , 1.32]
tree	2.52 [1.76, 3.29]	-1.57 [-2.02, -1.16]	-1.35 [-1.82, -0.94]	1.57 [1.10, 2.09]	0.58 [0.043, 0.95]
graph	2.44 [1.81, 2.99]	-1.92 [-2.61, -1.36]	-1.34 [-1.66, -0.96]	1.67 [1.21, 1.97]	0.45 [0.085, 0.75]

na means Cohen's d does not apply because it is not a comparison.

Table E.4: Cohen's d and 95% bootstrap CIs for RQ5 (Figure 9)

	Before feedback		After feedback		Differences (After - Before)		
	correct	incorrect	correct	incorrect	correct	incorrect	
g/t/g	na	na	na	na	0.72 [0.38, 1.05]	-0.36 [-0.72, 0.011]	images
none	na	na	na	na	-0.13 [-0.48, 0.25]	0.15 [-0.21, 0.48]	
g/t/g - none	0.49 [0.22, 0.72]	-0.33 [-0.67, 0.008]	0.32 [-0.010, 0.57]	-0.080 [-0.45, 0.27]	0.73 [0.36, 1.02]	-0.45 [-0.81, -0.082]	
g/t/g	na	na	na	na	1.03 [0.72, 1.36]	-1.02 [-1.35, -0.65]	roses
none	na	na	na	na	-0.045 [-0.39, 0.33]	0.16 [-0.19, 0.50]	
g/t/g - none	0.34 [-0.11, 0.65]	-0.37 [-0.64, -0.083]	0.18 [-0.21, 0.55]	-0.14 [-0.43, 0.25]	0.95 [0.64, 1.26]	-1.02 [-1.37, -0.72]	