Evaluating User Trust in Active Learning Systems Through Query Policy and Uncertainty Visualization

IAN THOMAS and DANIELLE SZAFIR, The University of North Carolina at Chapel Hill, USA

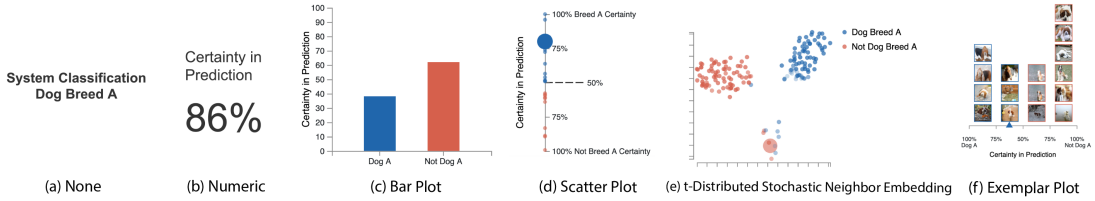


Fig. 1. We evaluate the role of classification uncertainty on analyst trust in active learning systems. We visualize classification uncertainties using six different techniques (a)-(f) designed to reveal increasing amounts of information about an algorithm's classification decisions and pair these visualizations with five different query policies.

Interactive data-labeling systems have become increasingly popular for various applications in machine learning. However, ensuring appropriate analyst trust in these systems remains a significant obstacle. We investigate how different active learning (AL) query policies coupled with classification uncertainty visualizations affect analyst trust in automated image classification systems. A common AL strategy is to query a human oracle to refine labels for datapoints where the classifier has the highest uncertainty, which yields maximal information gain. Model-centric policies do not consider potential priming effects on the human analyst and the consequent manner in which the human will interact with the system post-training. We present an empirical study (N = 468) evaluating how AL query policies and visualizations lending transparency to their classification certainty influence trust in automated image classifiers. We found that while participants did not adjust their level of agreement with the classifier according to querying policy or accompanying visualization, participants self-reported higher levels of confidence in the classifier when queried for uncertain datapoints.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Human-AI collaboration, Human-AI teams, Autonomous agent

ACM Reference Format:
Ian Thomas and Danielle Szafir. 2023. Evaluating User Trust in Active Learning Systems Through Query Policy and Uncertainty Visualization. In *Hamburg '23: ACM CHI April 23 - 28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 11 pages.

1 INTRODUCTION

Automated systems powered by supervised learning algorithms are becoming ubiquitous in various domains of human activity, from social media [11, 29] to military intelligence [8, 9, 17]. However, these systems often encounter challenges such as insufficient labeled data or the inability to capture nuanced expert knowledge. Systems can address these

Manuscript submitted to ACM

limitations by integrating human analysts into the classifier training process using techniques like active learning. Active learning (AL) is a form of semi-supervised machine learning that incorporates an analyst's knowledge into the algorithm's training, typically for classification tasks. This process entails a human analyst being queried to provide knowledge in the form of labels for instances the system is most uncertain of.

AL approaches often focus on optimal candidate selection strategies, where the best strategies are those that optimize algorithm performance [18]. These strategies may identify candidate data points selected from large collections of unlabeled data that yield optimal information gain. However, these approaches focus exclusively on best strategies for the automation system and do not account for effects on the analyst and their confidence in the system. For example, presenting analysts with only low confidence examples may bias analysts' impressions of the system's capabilities, priming analysts to mistrust the system [1, 25]. In this study, we investigate how these query policies and the methods used to communicate classifier confidence influence analyst trust in an automated image classifier's decisions through a human subject experiment.

Contributions: Our main contributions are empirical results measuring how query policies and visualization influence analyst trust in autonomous systems using active learning. Our empirical findings reveal that there are few significant differences in the number of label changes when using different uncertainty visualizations or query policies, indicating that people do not significantly adjust their level of trust towards the classifier depending on query policy or level of transparency into the uncertainty of a classification. However, participants self-reported significantly higher levels of confidence when asked to label instances that the classifier is most uncertain of, indicating that participants have more confidence in classifiers that query for uncertain points.

2 RELATED WORK

Recent work in visual analytics has explored how visualization can support analysts in more effectively leveraging machine learning for data analysis. For example, visualization can communicate aspects of a classifier's internal state [20, 26], classification performance [13, 21], and aid in model debugging [15]. While these approaches help bridge human analysts and automated classification, the impact that automated classification systems may have on the way analysts use classification outputs is still subject to ongoing research. For example, recent work regarding AI-human teams explore potential effects of accompanying explanations for AI-assisted classification decisions. Bansal et al. showed that even when the human is capable of demonstrating higher accuracy than the assisting model on a given task, humans are more susceptible to accepting a model's decision when explanations are provided regardless of whether the model's decision is correct [2].

As datasets increase in size and complexity, active learning can help attenuate the exhaustive efforts required for generating labeled datasets by querying analysts to provide relevant information for important subsets of data. Work in this field has focused on devising model-centric approaches that optimize information gain from the fewest number of instances by querying an oracle according to a prescribed query policy [10]. For example, the system may ask analysts to label data points the classifier is least certain about [12, 14] or candidate data points representative of a cluster of similar instances [31].

Recent work has explored model- versus human-centric modeling approaches. Tam et al. provide an in-depth analysis of machine learning versus visual analytics approaches to classification model building and found that the human-centric approaches provide significant improvements to the model building task, especially with sparse datasets [22]. They argue that the incorporation of an analysts' "soft knowledge", i.e., additional knowledge neither well-defined nor available to the machine-centric approach, fosters improved performance and conclude that a combination of machine

Table 1. Query policies used in the training phase of our primary study.

Policy	Description
Random	Confidence randomly interspersed
ALC	All low confidence
AHC	All high confidence
HtL	Low to high confidence
LtH	High to low confidence

learning and visual analytics, where the strengths of each approach are leveraged, will provide the best performance. We expand upon this notion by exploring the counterbalancing of machine- and analyst-centric optimizations to better understand the role of instance selection on analyst trust.

3 MOTIVATION & RESEARCH QUESTIONS

In this study, we seek to address two research questions relevant to analyst trust in active learning systems:

(1) Will query policies formulated on system confidence affect the analyst's confidence in the system during autonomous classification?
(2) How will visualizations of classifier confidence influence an analyst's trust in autonomous classification?

We explore these questions in two phases. In the first phase, we conducted a preliminary data collection study to formulate a ground truth confidence based on participants' performance on an image classification task. In the second phase, we used these confidences to simulate a synthetic classifier aligned with anticipated human confidence in classification. We asked participants to actively train the classifier based on a set of queried examples selected using one of five tested query policies paired with six visualization types the provide increasing amounts of information about classifier confidence regarding the queried image. We then gauge participant trust in the resulting algorithm using a series of objective (percentage agreement) and subjective (self-reported perceptions of confidence, trust) metrics.

4 DATA COLLECTION STUDY

We first conducted a data collection study to generate baseline confidence scores for use in the primary research study. We utilized these baseline confidence measurements to simulate classification in our primary study as opposed to a machine learning classifier's confidence in order to mirror participants' certainty in the classification and to avoid classifier-dependent results. We conducted a 4 (image class) x 4 (downsampling rate) mixed factors experiment measuring human accuracy at an image classification task to establish this ground-truth confidence dataset. We recruited 178 participants (average age of 37.7, 85 males, 91 females, 2 no replies) via Amazon Mechanical Turk. 38 participants were excluded for labeling all images with the same label or failing four attention checks.

4.1 Stimuli

We selected our stimuli from a collection of dog breed images drawn from the Oxford-IIIT Pet Dataset [16]—a 37 category pet dataset with roughly 200 images for each category (see Figure 2). Two categories of dog breeds were chosen heuristically for each set such that they shared a sufficiently large subset of common features such that they may prove challenging to visually discriminate. Twenty-five images from each of the six dog breeds were sampled. Four new images were generated from each of the 25 images via down-sampling using maxpooling operations at four different down-sampling rates: 2, 4, 7, and 10. For example, for a down-sampling rate of 2, 2x2 pixel grids would be

Fig. 2. Representative images of dog breeds used to construct the datasets used in our study.

pooled into a 1x1 pixel grid for the next image, where the 1x1 pixel is the maximum pixel value from the 2x2 pixel grid of the original, full-resolution image. This generated a dataset of 2,400 images (3 breed pairs × 200 images per breed pair × 4 downsampling rates).

Our independent variables were dog breed and downsampling rate. Dog breed consisted of four levels, one for each breed (Basset Hound, Saint Bernard, Chihuahua, Miniature Pinscher), tested as a between-subjects factor, with each participant asked to classify whether or not an image is a member of one breed. Distractor breeds for each primary breed are shown in Figure 3. Down-sampling rate consisted of four levels (2, 4, 7, and 10), tested as a within-subjects factor. Our dependent variable is accuracy, which is measured as the proportion of correct labels applied by the population.

4.2 Experimental Task

Our study consistent of four phases: (1) informed consent, (2) training, (3) labeling, and (4) demographics. Participants first provides informed consent in accordance with our IRB protocol. After accepting the terms, the participant proceeds to Phase 2, where they are presented with an example image for each of the two dog breeds at the top of the screen, a set of instructions, and a single test image of either the tested breed ("dog breed A") or its complement within the tested set ("not dog breed A"). Note, actual dog breed names are not used to provide consistency across datasets. The test image is presented in the middle of the screen, and the participant is instructed to label it as either dog breed A or not dog breed A by pressing either key F (dog breed A) or J (not dog breed A). These keys align with natural pointer finger positioning on a QWERTY keyboard.

Each participant saw 50 non-downsampled images randomly from each breed pair (25 per breed) with images counterbalanced between participants to ensure an equal distribution of responses. The images in the tutorial segment are randomly selected without replacement from this set, resulting in 25 tutorial images. The participant must correctly label 5 images from each breed before moving onto Phase 3, indicating they have sufficiently learned to discriminate between the two breeds. The remaining 25 images are used in Phase 3.

In Phase 3 we removed the breed examples and asked participants to label the currently displayed image as either dog breed A or not dog breed A using the same keypress inputs as in the tutorial. Participants complete this identification task for each of the remaining 25 images at each of the four tested downsampling rates, with images presented in random order to mitigate potential transfer effects. Four full resolution images were inserted at trial numbers 20, 40, 60,

and 80 to serve as attention checks, resulting in the participant labeling a total of 104 images. Lastly, in Phase 4 the participant is asked to provide demographic information.

4.3 Results

We computed confidence scores for each breed as the proportion of times each breed was labeled as "Breed A" for each of the six target breeds. We elected to use confidence rather than correctness as our primary metric to ensure our simulated classifier replicated the anticipated behaviors of the human oracle. Participants had a mean confidence for images in the breed pair for dataset A with 82% mean confidence (σ = 15%), and dataset B with 71% mean confidence (σ = 14%). The performance distribution within these datasets allows us to draw from a broad dataset of low- and high-confidence images for each breed pair in order to apply different query policies to the training phases of our primary experiment. We use the statistical mean of each dataset to divide our corpus into low- and high-confidence scores. This division created 104 and 108 low-confidence images and 96 and 92 high-confidence images for datasets A and B respectively.

5 PRIMARY STUDY

We use the results of our data collection study to simulate an active learning classifier leveraging different query policies and classification uncertainty visualizations in order to understand the role of query policy and visualization on analyst trust in a system trained using active learning. We conducted this analysis using a 5 (query policy) × 6 (visualization method) mixed factors experiment conducted on Amazon Mechanical Turk.

5.1 Stimuli

We used the ground truth accuracy from our data collection study to generate the stimuli for our second experiment, with high confidence defined as being above the statistical mean for the respective data set, and low confidence being below the statistical mean, as described in 4.3. We implemented each of the five policies by sampling images within each breed based on these thresholds. The data generated from our data collection study was utilized to generate the visualizations. In all visualizations, blue marks correspond to the classification label dog breed A and red corresponds to not dog breed A.

5.2 Query Policies

We developed five query policies, as shown in Table 1, to examine the potential impacts of query policy based on prior work related to learning [4–6] and cognitive science [23, 28]. The first policy serves as our random-baseline condition, which randomly samples images from the full dataset, resulting in a sequence of images with a random distribution of confidences. Carvalho & Goldstone indicate that interspersed versus blocking policies influence learning capabilities [6]. Consequently, we can explore these interspersed (Random) versus blocking effects (ALC, AHC) by developing policies that randomly sample images that are exclusively low or high confidences, respectively. The last two policies, HtL and LtH, randomly sample images from the set and then arrange them from high to low and low to high confidences, respectively. Past work on priming and anchoring effects suggests that the first instances of learning may overwhelmingly govern the decision-making tasks in an evaluation phase, despite the information provided at the trailing end of a learning phase [23].

5.3 Visualization

We tested six conditions for visualization (see Figure 1) that span a spectrum of classification transparency. The first visualization provides no information about classification uncertainty and therefore no transparency into the system's certainty or decision-making processes, offering a baseline performance metric. The numeric visualization provides a raw confidence score, offering a basic level of transparency into the systems' classification certainty. The bar chart visualizes this same confidence for the classification decision alongside the confidence of the contrary classification. The scatterplot visualization provides a higher level of transparency by offering the analyst contextual information about the current image's classification and degree of certainty within the population of images. The t-SNE embedding plot, an approach commonly used in modern visualization tools [3, 24], visualizes the population of images as a scatter plot in an embedded feature space, with coordinates representing images determined by image similarity. The exemplar plot shows selected examples at different confidence levels (separated into quartiles) and the respective quartile for the current instance. This plot gives participants comparable examples of each breed while making their decision to agree or disagree with the system. Images are pulled at random from each confidence quartile for use in the plot.

5.4 Experimental Design

This experiment was designed as a two-stage 5 (query policy) × 6 (visualization type) full factorial between-participants experiment. We tested two independent variables: query policy and visualization method. The query policy consisted of five levels summarized in Table 1. Visualization methods consist of six levels summarized in Figure 1.

Our objective metric of trust consisted of label disagreement—the number of times an analyst flips the classification of a system-applied label—computed over a set of 15 images with a uniform distribution of high and low confidence and correctness randomly sampled from a binomial distribution weighted according to the confidence score of the image (e.g., if the image confidence is 75%, the system will provide the correct label in 75% of instances). Prior work suggests that percentage agreement is a meaningful dependent measure of trust [27], the more an analyst changes a label, especially when levels of self-certainty are high, the less the analyst is willing to rely upon the system.

Subjective measures were taken per-image, consisting of two 5-point Likert-type questions. We analyzed our objective measures and per-question subjective measures using a two-factor (visualization method, query policy) full factorial ANCOVA. We then used Tukey's Honest Significant Difference test (HSD) wtih Bonferroni correction for all post-hoc analysis ($\alpha = .05$).

5.5 Experimental Task

The study consists of five phases: (1) informed consent, (2) tutorial, (3) training, (4) testing, and (5) survey and demographics.

In Phase 1, the participant is provides informed consent and proceeds to Phase 2, which is the tutorial for the study. In the tutorial, the participant is presented with an example image for each of the two dog breeds at the top of the screen, a set of instructions describing both the task and how to interpret the tested visualization, and a single image of either dog breed A or not dog breed A with an accompanying visualization (one of the six possible visualizations). Note, as in the first study, actual dog breed names are not used. A single image is presented in the middle of the screen that the user must label as either dog breed A or not dog breed A. To label an image, the participant is instructed to press key F to label the image as dog breed A and to press key J to label the image as not dog breed A. The images in the tutorial segment come from the set of 50 images per breed generated for the tutorial set, as described in 4.1. The

participant must correctly label 5 images from each breed before moving onto Phase 3, indicating they can discriminate between the two breeds. After completing the tutorial, participants receive a brief explanation of the accompanying visualization before proceeding to Phase 3.

In Phase 3, participants are instructed that they will be training a classifier to better discriminate between the two dog breeds the participant was trained on, simulating an active labeling protocol. The participant will be presented 18 images, one image at a time, along with the system's binary classification of that object—dog breed A or not dog breed A stemming from the confidence measure generated by sampling from a binomial distribution structured according to the confidence score for each image from the data collection study in 4—and an accompanying visualization. 18 training images are selected without replacement according to the query policy sampling described in 5.1.

Once the participant has finished training the classifier on the 18 images, they proceed to a holding page that instructs them that the classifier is updating based on their feedback the participant provided the system. While the system trains, which takes approximately 10 seconds with simulated progress indicated using a horizontal progress bar. The participant is instructed they will now evaluate the system's performance, and they proceed to Phase 4. In Phase 4, the participant will be sequentially presented with 15 images based on the same procedure as Phase 3.

The 15 images correspond to the test set images described in 5.4 with images randomly sampled and presented in a random order. Unlike in the previous segment the participant is not provided with an accompanying visualization. This enables us to focus exclusively on trust constructed during active labeling. Showing classification confidence post-labeling integrates confidence into the testing phase and could interfere with primed perceptions. Therefore, the participant must choose to accept the classifier's labeling of the given image as either correct or incorrect without the visualization. Additionally, for each image, the participant must fill out two 5-point Likert-type questions pertaining to their confidence in their own decision ("*How confident are you in your decision?*") and their confidence in the classifier's decision ("*How confident are you in the system's decision?*").

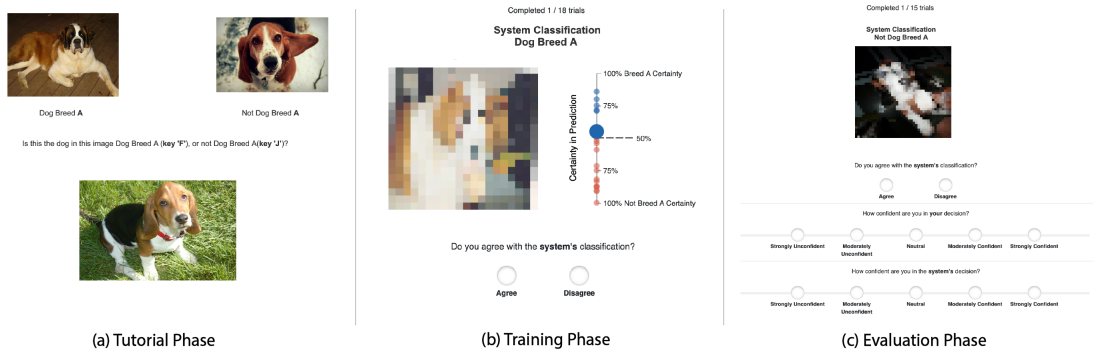
After completing the system evaluation stage, the participant proceeds to Phase 5, where the participant fills out the post-survey questionnaire and a short demographics survey including self-reported experience with machine learning and data analytics tools.

5.6 Participants

For the second study, 535 participants with an average age of 39.72 (σ = 11.19) and gender distribution of 291 males, 238 females, 3 non-binary, and 4 no replies participated in the experiment via Amazon Mechanical Turk. We excluded 67 participants for failing at least two of the three attention checks in the evaluation phase, or labeling all images with the same label. An attention check consisted of three full resolution images inserted in the evaluation phase.

6 RESULTS

Figure 4 shows the frequency of label changes per query policy and visualization type. All label-change frequencies were close to our random baseline. The AHC condition had a slightly higher label change frequency than ALC, while HtL was slightly lower than LtH. Our two-factor ANCOVA indicated no statistically significant difference between any conditions for label change frequency. Analysis of visualization conditions paired with varying query policies followed a similar pattern, showing no statistically different proportion of label changes across conditions. Though, the Exemplar plot showed a slightly higher label-change frequency than other types when paired with other query policies. The pairings of query policy and visualization type with the highest rate of disagreement and agreement were



(a) Tutorial Phase (b) Training Phase (c) Evaluation Phase

Fig. 3. Example interfaces for three of the five phases of the primary study, corresponding to phases 2 - 4 as described in 5.5. (a) Tutorial phase, where participants must correctly label five instances of each dog breed to proceed. (b) Training phase, where participants must either agree or disagree with 18 per-image classifications from the system, presented according to the query policy and accompanied with an uncertainty visualization. (c) Evaluation phase, where participants respond to 15 randomly drawn instances and self-evaluate their level of confidence in the system's decision and their own decision.

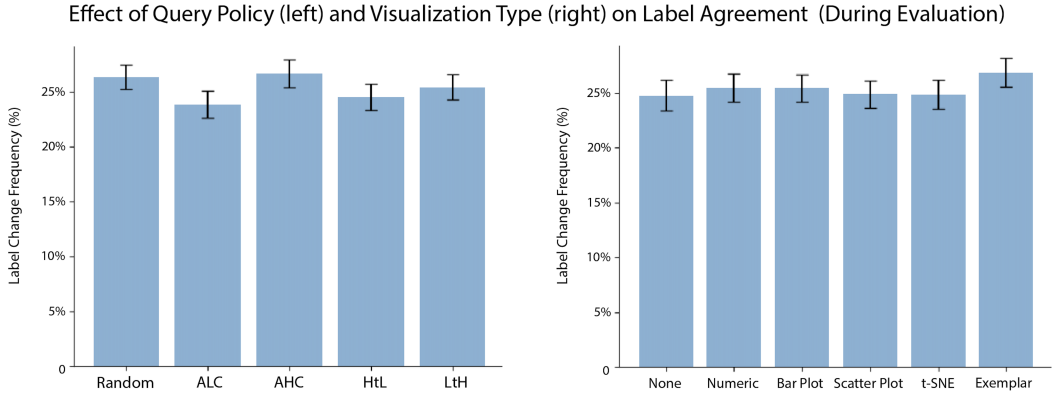


Fig. 4. Percent disagreement in the evaluation phase of the primary study, by query policy (left) and uncertainty visualization type (right).

AHC accompanied by an exemplar plot (μ = .316, σ = 0.18) and HtL accompanied by a t-SNE plot (μ = 0.195, σ = 0.11) respectively.

All participants demonstrated a similar level of accuracy in their agreements with the ground-truth label for each instance of the evaluation phase, indicating that their responses closely aligned with the actual label regardless of query policy (See Table 2). These findings indicate that users did not exhibit automation bias in their decision-making process.

We observed a significant effect of query policy on participants' self-reported confidence in the system's decisions ($F(4, 463) = 3.912$, $p < .005$). Participants in the ALC exhibited the greatest levels of confidence in their self-reported responses (μ = 3.901, σ = 0.51) while participants responded with the lowest levels of confidence in the AHC condition (μ = 3.776, σ = 0.556) during evaluation. Tukey's HSD reveals that the ALC condition significantly increases the participants' perceptions of system confidence compared to AHC. We found a statistically significant difference (F(5,

Table 2. Evaluation phase results by query policy used in training. The last two columns (Confidence in Decision, Confidence in System) correspond to the two per-image subjective measures.

Policy	Accuracy	Label Disagreement	Confidence in Decision	Confidence in System
Random	$\mu = 0.76$, $\sigma = 0.19$	$\mu = 0.263$, $\sigma = 0.10$	$\mu = 4.04$, $\sigma = 0.76$	$\mu = 3.788$, $\sigma = 0.51$
ALC	$\mu = 0.738$, $\sigma = 0.13$	$\mu = 0.238$, $\sigma = 0.09$	$\mu = 3.97$, $\sigma = 0.54$	$\mu = 3.901$, $\sigma = 0.51$
AHC	$\mu = 0.736$, $\sigma = 0.17$	$\mu = 0.266$, $\sigma = 0.12$	$\mu = 3.95$, $\sigma = 0.58$	$\mu = 3.776$, $\sigma = 0.56$
HtL	$\mu = 0.734$, $\sigma = 0.16$	$\mu = 0.245$, $\sigma = 0.11$	$\mu = 3.98$, $\sigma = 0.56$	$\mu = 3.892$, $\sigma = 0.51$
LtH	$\mu = 0.734$, $\sigma = 0.17$	$\mu = 0.254$, $\sigma = 0.12$	$\mu = 3.98$, $\sigma = 0.53$	$\mu = 3.829$, $\sigma = 0.58$

463) = 2.317, $p < .01$) between visualization types on self-reported confidence in the system between the None-type control visualization condition ($\mu = 3.76$, $\sigma = 0.50$) and the t-SNE plot condition ($\mu = 3.89$, $\sigma = 0.51$). We also found a significant effect (F(5, 463) = 2.724, $p < .01$) on self-reported confidence in participants' own decision between accompanying t-SNE plot ($\mu = 4.0$, $\sigma = 0.53$) and exemplar plot ($\mu = 3.81$, $\sigma = 0.61$) conditions.

7 DISCUSSION

This study adds to a growing body of work investigating human-AI teaming through interaction with an image classifier using active learning to query participants for labels. Our findings suggest that individuals may be relatively robust to querying strategies employed by AL systems. This is consistent with previous research which indicates that humans are highly skilled at categorizing common images and may not be susceptible to the influence of automation bias when performing a task at which humans excel [19].

Our results suggest that participants' self-reported confidence in the system was noticeably impacted by query policy, with substantially lower confidence reported in cases where the system requested labels for high-confidence examples. One possible explanation for this phenomena is that when users witness the classifier seeking labels for easily identifiable high-confidence examples, they are less inclined to trust the system, suggesting that the system requires assistance with tasks that should be straightforward.

We observed fewer effects dependent on the type of accompanying visualization, with the only significant differences being a similar effect for participants' self-reported confidence in the system between the None-type control group that had no accompanying visualization, and participants that had a t-SNE plot accompanying their decision. We also found an effect on participants' confidence in their own decision depending on whether their accompanying visualization was a t-SNE plot or exemplar plot. This discrepancy may exist due to the fact that t-SNE embeds position based on similarity, which lends further insight into the classifier in contrast to the none-type baseline condition and the exemplar plot, which lend little insight into the classifier's decision.

A key limitation of this work is the lack of interaction between the user and classifier beyond label-changes. In a real-world scenario, interaction may consist of many other elements, such as interaction with a visualization of multiple instances [7], or even as natural language dialogue between the analyst and the model [30]. Future work includes exploration of other methods of interaction between the analyst and the automation in active learning contexts, and through other types of interactive labeling systems. This could also include more extensive testing methodology through longer-term interactions with automated systems rather than shorter interactions as tested in this work.

REFERENCES

[1] Hasmik Atoyan, Jean-Rémi Duquet, and Jean-Marc Robert. 2006. Trust in New Decision Aid Systems. In *Proceedings of the 18th Conference on l'Interaction Homme-Machine* (Montreal, Canada) *(IHM '06)*. Association for Computing Machinery, New York, NY, USA, 115–122. https:

//doi.org/10.1145/1132736.1132751

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[3] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. 2018. Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 298–308. https://doi.org/10.1109/TVCG.2017.2744818

[4] Paulo Carvalho and Robert Goldstone. 2014. Effects of Interleaved and Blocked Study on Delayed Test of Category Learning Generalization. *Frontiers in Psychology* 5 (08 2014). https://doi.org/10.3389/fpsyg.2014.00936

[5] Paulo F. Carvalho and Robert L. Goldstone. 2015. The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review* 22 (2015), 281–288.

[6] Paulo F. Carvalho and Robert L. Goldstone. 2017. The Sequence of Study Changes What Information Is Attended to, Encoded, and Remembered During Category Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43 (2017), 1699–1719.

[7] Mohammad Chegini, Jürgen Bernard, Jian Cui, Fatemeh Chegini, Alexei Sourin, Keith Andrews, and Tobias Schreck. 2020. Interactive visual labelling versus active learning: an experimental comparison. *Frontiers of Information Technology & Electronic Engineering* 21 (2020), 524–535.

[8] Mary Cummings, Sylvain Bruni, S. Mercier, and P. Mitchell. 2008. Automation Architecture for Single Operator Multiple UAV Command and Control. *The International C2 Journal* 1 (01 2008).

[9] Mary L. Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems.

[10] Yifan Fu, Xingquan Zhu, and Bin Li. 2012. A survey on instance selection for active learning. *Knowledge and Information Systems* 35 (2012), 249–283.

[11] Tushar Gaikwad. 2018. Artificial Intelligence based Chat-Bot. *International Journal for Research in Applied Science and Engineering Technology* 6 (2018), 2305–2306.

[12] Alex Holub, Pietro Perona, and Michael C. Burl. 2008. Entropy-based active learning for object recognition. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2008), 1–8.

[13] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5686–5697. https://doi.org/10.1145/2858036.2858529

[14] David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '94)*. Springer-Verlag, Berlin, Heidelberg, 3–12.

[15] Shixia Liu, Jiannan Xiao, Junlin Liu, Xiting Wang, Jing Wu, and Jun Zhu. 2018. Visual Diagnosis of Tree Boosting Methods. *IEEE Transactions on Visualization and Computer Graphics* 24 (2018), 163–173.

[16] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and Dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[17] Steven J. Pereira, Grady Dale Lee, and Jeffrey Howard. 2006. A System-Theoretic Hazard Analysis Methodology for a Non-advocate Safety Assessment of the Ballistic Missile Defense System.

[18] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A Survey of Deep Active Learning. *ACM Comput. Surv.* 54, 9, Article 180 (oct 2021), 40 pages. https://doi.org/10.1145/3472291

[19] Victor A. Riley. 1996. Operator reliance on automation: Theory and data.

[20] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. 2018. LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan 2018), 667–676. https://doi.org/10.1109/TVCG.2017.2744158

[21] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S. Tan. 2009. EnsembleMatrix: Interactive Visualization to Support Machine Learning with Multiple Classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1283–1292. https://doi.org/10.1145/1518701.1518895

[22] Gary K. L. Tam, Vivek Kothari, and Min Chen. 2017. An Analysis of Machine- and Human-Analytics in Classification. *IEEE Transactions on Visualization and Computer Graphics* 23 (2017), 71–80.

[23] André Calero Valdez, Martina Ziefle, and Michael Sedlmair. 2018. Priming and Anchoring Effects in Visualization. *IEEE Transactions on Visualization and Computer Graphics* 24 (2018), 584–594.

[24] Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing data using t-SNE. *Journal of Machine Learning Research* 9 (11 2008), 2579–2605.

[25] Timothy D. Wilson, Christopher E. Houston, Kathy Etling, and Nancy Brekke. 1996. A new look at anchoring effects: basic anchoring and its antecedents. *Journal of experimental psychology. General* 125 4 (1996), 387–402.

[26] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mané, Doug Fritz, Dilip Krishnan, Fernanda B. Viégas, and Martin Wattenberg. 2018. Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics* 24 (2018), 1–12.

[27] Jessie Yang, Christopher D. Wickens, and Katja Hölttä-Otto. 2016. How users adjust trust in automation: Contrast effect and hindsight bias. In *Proceedings of the HFES 60th Annual Meeting (Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 1)*. Human Factors and Ergonomics Society, 196–200. https://doi.org/10.1177/1541931213601044 Annual Meeting of Human Factors and Ergonomics Society, HFES ;

Conference date: 19-09-2016 Through 23-09-2016.

[28] X. Jessie Yang, Christopher D. Wickens, and Katja Hölttä-Otto. 2016. How users adjust trust in automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60 (2016), 196 – 200.

[29] Daniel Zeng, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. 2010. Social Media Analytics and Intelligence. *IEEE Intelligent Systems* 25, 6 (2010), 13–16. https://doi.org/10.1109/MIS.2010.151

[30] Ye Zhang, Matthew Lease, and Byron C. Wallace. 2017. Active Discriminative Text Representation Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI'17)*. AAAI Press, 3386–3392.

[31] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. 2009. Multi-Instance Learning by Treating Instances as Non-I.I.D. Samples *(ICML '09)*. Association for Computing Machinery, New York, NY, USA, 1249–1256. https://doi.org/10.1145/1553374.1553534