Chandramouli, Suyog; Zhu, Yifan; Oulasvirta, Antti

Interactive Personalization of Classifiers for Explainability using Multi-Objective Bayesian Optimization

Published in:
UMAP 2023 - Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization

DOI:
10.1145/3565472.3592956

Published: 18/06/2023

Document Version
Publisher's PDF, also known as Version of record

Interactive Personalization of Classifiers for Explainability using Multi-Objective Bayesian Optimization

Suyog Chandramouli*†
Yifan Zhu*†
Antti Oulasvirta†
firstname.lastname@aalto.fi

ABSTRACT

Explainability is a crucial aspect of models which ensures their reliable use by both engineers and end-users. However, explainability depends on the user and the model's usage context, making it an important dimension for user personalization. In this article, we explore the personalization of opaque-box image classifiers using an interactive hyperparameter tuning approach, in which the user iteratively rates the quality of explanations for a selected set of query images. Using a multi-objective Bayesian optimization (MOBO) algorithm, we optimize for both, the classifier's accuracy and the perceived explainability ratings. In our user study, we found Pareto-optimal parameters for each participant, that could significantly improve explainability ratings of queried images while minimally impacting classifier accuracy. Furthermore, this improved explainability with tuned hyperparameters generalized to held-out validation images, with the extent of generalization being dependent on the variance within the queried images, and the similarity between the query and validation images. This MOBO-based method has the potential to be used in general to jointly optimize any machine learning objective along with any human-centric objective. The Pareto front produced after the interactive hyperparameter tuning can be useful during deployment, allowing for desired trade-offs between the objectives (if any) to be chosen by selecting the appropriate parameters. Additionally, user studies like ours can assess if commonly assumed trade-offs, such as accuracy versus explainability, exist in a given context.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies**;

KEYWORDS

Personalization, Explainable AI, Interactive AI, Bayesian Optimization, Multi-objective Optimization

*Both authors contributed equally to this work.
†Department of Information and Communications Engineering, Aalto University, Espoo, 02150, Finland.

UMAP '23, June 26–29, 2023, Limassol, Cyprus

ACM ISBN 978-1-4503-9932-6/23/06.
https://doi.org/10.1145/3565472.3592956

ACM Reference Format:
Suyog Chandramouli, Yifan Zhu, and Antti Oulasvirta. 2023. Interactive Personalization of Classifiers for Explainability using Multi-Objective Bayesian Optimization. In *UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26–29, 2023, Limassol, Cyprus.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3565472.3592956

1 INTRODUCTION

In this paper, we study a human-in-the-loop (HITL) technique for improving the explainability of deep learning-based image classifiers while also ensuring their accuracy. While deep learning models come with very powerful predictive capabilities, they are composed of millions of parameters and mathematically unintuitive constructs, which for all practical purposes make them uninterpretable [12]. The emergence of such high-performance yet opaque-box-like models has given rise to the nascent field of Explainable AI (sometimes referred to as xAI), which aims to increase users' understanding of opaque models and the basis on which they generate predictions [20].

Even though there are numerous methods for explainable AI [2], they frequently adopt a generic and one-size-fits-all approach that does not take into account the user or their preferences about explanations [17, 42]. After all, the explainees' personal attributes, knowledge, backgrounds, expertise, and context of model use affects the type and content of explanations they find useful [24, 46]. Consequently, implementations of xAI are one-way communication processes that do not engage the user in a dialogue. By interactively personalizing explanations to the user, we believe that the xAI goal of enhancing the user's comprehension of opaque-box models can be more effectively achieved. This comprehension can then be used in a variety of ways to influence human intervention on almost any step in the machine learning (ML) pipeline, from data verification and model training to model deployment.

In this paper, we use multi-objective Bayesian optimization (MOBO) to develop an interactive personalization approach for opaque-box classifiers, and their explanations. Our approach involves the user iteratively evaluating explanations of the model's predictions. These evaluations, in the form of ratings, are used by MOBO to jointly explore the hyperparameter space of both, the opaque-box model and the explainer. Approaches based on Bayesian Optimization (BO) are well suited for interactive applications as they explore the hyperparameter space more efficiently, leading to efficient usage of user inputs [5, 37] despite its inherent noisiness. The use of MOBO allows for joint optimization over the objectives of both, accuracy and explainability. The output of the interactions is a Pareto front of optimal trade-offs between the objectives such

that improving any one objective comes at the cost of the other objective. This Pareto front offers the best personalized solutions from which the user can choose the "Pareto-optimal" parameters that provide a desired accuracy vs. explainability trade-off. This approach with MOBO would thus allow one to cover both ML objectives and user-specific views of understandability, as well as apply it with the preferred trade-off.

We demonstrate the feasibility of our approach with a user study with human-in-the-loop MOBO. In the study which lasted over seven hours for each participant, we considered the case of images classified by a neural-network-based image classifier and explained with SHAP (Shapley Additive Explanations; [28]). We designed the user study to assess personalization to both, users and the context of use. Users may prefer explanations differently from each other, and the context of use would determine the accuracy/explainability trade-offs that are desired. We consider a setting in the user study with multiple images and image labels, to induce sufficient variance in users' perceptions of explainability. We examine context-dependence by varying the assessing the explanation quality of held-out test images from both within and outside the interactive training context.

We recommend that user studies such as ours should be carried out to (i) assess if an accuracy vs. explainability trade-off exists, (ii) assess the performance of personalization in different application contexts, and (iii) be the basis for judiciously choosing parameters from a Pareto-front that maximize the utility of personalization. The overarching approach we have developed is relevant for optimizing tany objective for which a user-in-the-loop can provide meaningful feedback, and where trade-offs between objectives potentially exist. However, in this article, our focus is on the objectives of user-evaluated explainability, and classifier accuracy. In our study, we found Pareto fronts for all participants for accuracy vs. explainability; we also found that Pareto-optimal parameters focusing on explainability produced better explanations than baselines on held-out images. MOBO hence holds significant promise for the interactive personalization of image classifiers for explainability; or even more generally for the interactive personalization of opaque models for any human-centric objectives in combination with ML objectives.

2 RELATED WORK

The need for personalization of explanations is becoming increasingly well recognized within the field of explainable AI, and it has been acknowledged that the interpretability of an explanation may vary considerably between explainees [24, 46]. There has been some effort to understand what makes for good explanations in explainable intelligent systems [9]. However, this has not yet been applied in widely used explainable AI systems or to account for individual variability. In the absence of stronger theories about explainability and how it trades off with other objectives in given settings, it becomes useful to empirically test these assumptions with user studies. In the "explanatory debugging" approach [26], explanations are used as the basis for alterations of prediction labels or parameter weights of features in order to improve the model. This approach highlights the common workflow of using explainers to interface with opaque-box models and personalize them. However, the explanations themselves are not personalized to the user.

While uncommon, there have been efforts in the field of recommendation systems to personalize explanations of the predictions to the user [49]. Approaches for this have included using users' demographic information [10], user logs [50], and user models [7]. In the BAndits for Recsplanations as Treatments [30], contextual bandits are proposed for joint personalization of recommendations and associated explanations that aim to maximize the reward signals related to the eventual user engagement with the system, and increased engagement is assumed to be indicative of user satisfaction. Along similar lines as contextual bandits, Reinforcement Learning (RL) methods such as Reinforcement Learning with Human Feedback (RLHF) have been central to applications such as InstructGPT [31] and ChatGPT, to finetune large language models with human feedback; RLHF however is not used for individual personalization, as it needs millions of responses across participants to finetune a single model.

BO is different in how it incorporates human input into ML models. BO and its multi-objective variants such as MOBO, treat the user as an opaque-box model to optimize, focusing on relevant objectives in a one-shot manner. Neither RL nor BO-based methods have thus far been used to incorporate human feedback to personalize for explainability. We use BO and MOBO in our setting due to their appropriateness and their various desirable properties. In recent years, BO [18, 41] has emerged as a popular candidate for hyperparameter tuning of opaque box models that are expensive to evaluate. BO works by probabilistically modeling the relationship between hyperparameter settings and their performance along a given objective using a surrogate model of this unknown function, and uses this to iteratively evaluate, choose and update to the most promising hyperparameter settings for meeting the objective. Due to (i) the flexibility it allows in defining an objective, (ii) its ability to deal with noisy inputs, and (iii) its superior sample efficiency, it has been used in a variety of settings. For instance, Automated Machine Learning (AutoML) uses BO [16, 45] to automate machine-learning pipelines with quantitative ML-related objectives, and ensures that the user remains out of the loop. At the same time, BO has been popular in human-in-the-loop ML settings as well, because of its sample efficiency with respect to the amount of user inputs which can also be noisy. Some examples of the use of BO for human-centric objectives include user preference learning for animation galleries [5, 14], tuning game mechanics to maximize a gamer's engagement [22], and an adaptive font generation system that optimizes for reading speed [21].

MOBO is a variant of BO that simultaneously optimizes a system for multiple objectives within a solution space. Different objectives here can often be in conflict with each other, and hence, the output of this approach is not a single optimal solution, but a Pareto frontier of optimal trade-offs between objectives where improving any objective, degrades the other. Multi-objective optimization has been used across fields including Human-Computer Interaction [3, 6, 13]. MOBO research continues to develop methods to discover the right frontiers in a computationally feasible and sample-efficient manner, for example, as the number of objectives increases [25].

Even though BO and MOBO have been used for hyperparameter tuning, it has been applied mainly to either ML-focused objectives

or human-centric objectives, and not as a glue through which to jointly personalize across both types of objectives. MOBO could be particularly useful for this case as we demonstrate in the following sections with our general framework and its application to personalizing to users jointly for accuracy and explainability.

3 A MULTI-OBJECTIVE INTERACTIVE APPROACH TO PERSONALIZATION FOR EXPLAINABILITY

We propose a generally applicable approach to personalize any opaque-box model along any number of human-centered objectives about which a user can provide meaningful feedback, in addition to optimization over machine-learning objectives. In this approach, the user iteratively considers predictions by the opaque-box model on a chosen set of inputs, and provides preference ratings along the desired objective(s); these ratings are used along with MOBO to jointly explore the hyperparameter spaces (i) θ_{opaque} of the opaque-box model, and (ii) θ_{xAI} of an intermediary user interface that exposes model performance to the user. The result of the iterative interactions is a Pareto frontier of optimal trade-offs between the objectives such that improving any one objective comes at the cost of the other objective(s). We illustrate the framework in Figure 1, for the scenario of an explainable AI (SHAP) interface, where the predictive accuracy of an opaque image classifier is the machine-learning objective, and the users' perceived explainability of image classifications is the human-centric objective. This special case of the general approach is the focus of our article.

In Figure 1, users evaluate explanations produced by a parameterized explainable AI method (e.g. SHAP). During each iteration, the user provides explainability ratings for the set of explanations presented to them; these ratings along with the classifier's current accuracy value are used by MOBO to explore the joint hyperparameter space of the classifier-explainer pair; the aim of MOBO is to propose hyperparameters that can simultaneously and efficiently optimize for both, the user's perceived explainability and the accuracy of the classifier which can potentially be in conflict with each other. The result of these iterations is a Pareto front; appropriate Pareto optimal parameters that can be chosen from the front, based on the desired accuracy vs. explainability trade-off. These parameters are effectively those that have been personalized to the user. One can make use of tests analogous to our user study (see section 5) to make further assessments regarding the suitability of the chosen parameters.

Delving deeper into Figure 1, our approach starts with partitioning the dataset into the training images and query images, $Image_t$ and $Image_q$, where $Image_t$ is used to train a new image classifier with the hyperparameters θ_{opaque} that have been proposed by MOBO, after every iteration of user feedback. For each training iteration, we calculate the accuracy of the classifier, which is one of the optimization objectives; this classifier is also used to classify a small set of query images, $Image_q$, whose explanations are evaluated by the user. Explainable AI methods such as SHAP provide visual explanations by "pixel-attribution", where pixels most influential for making a classification are highlighted. Under the hood, $Image_q$ is in fact first passed through the segmentator, which outputs segments (also known as super-pixels) for each image. These super-pixels are then sent to the classifier for classification. The xAI module then generates visual explanations for the classifications which we present to the user. Users provide ratings based on their evaluation of the explanations. These ratings and the current classifier accuracy are considered using MOBO which proposes new hyperparameters for the classifier and segmentator. This iterative process continues for N iterations. The data points collected from these iterations allow us to construct a Pareto frontier for accuracy vs. explainability.

An advantage of MOBO is that it is robust to noisy inputs. However, it would still be useful to have consistent rating scales on the objective for better personalization. Within-participant consistency ensures that interactive personalization with MOBO works as intended; between-participant consistency ensures that ratings across participants can be meaningfully compared in the user study. To enable this internal consistency, we recommend producing a standard scale along the objectives and using it consistently. This is the strategy we use in our implementation. Another strategy might be to invent an online task for the users, the performance on which will determine the implicit ratings provided to MOBO, but this is quite an unexplored territory in the field.

4 SYSTEM IMPLEMENTATION FOR USER STUDY

We conducted a user study to evaluate the feasibility and efficacy of the MOBO-based personalization approach for explainability illustrated in Figure 1. We also aimed to study the manner in which the quality of explanations generated by personalized hyperparameters carried over to held-out data, when the test and training contexts were shared, and when the contexts differed. The user study was designed to have 3 main phases: (i) a practise phase where users were familiarized with the experiment setup and rating schemes, (ii) an interactive hyperparameter tuning phase where image classifiers were trained based on user feedback, and (iii) a validation phase where the quality of explanations was assessed by users on held-out images. Following below are the details of the system we developed.

4.0.1 Image Classifier. The Image Classifier we constructed was governed by a few important desiderata. We would ideally be able to (i) work with a classifier that is powerful, (ii) conduct the user study in real time, (iii) evaluate the feasibility of our approach with a limited number of query images, and (iv) employ widely adopted frameworks with respect to image classification and explainability.

In order to meet these desiderata, we use a transfer learning approach on a pre-trained VGG16 model [39, 44] which is a powerful image classifier made with a convolutional neural network, and trained on the ILSVRC-2012 dataset (also known as Imagenet). We make the widely-prevalent assumption that a neural network's lower layers identify basic visual features such as edges and textures [36], which are agnostic to the image dataset. Hence, the classifier used in the user study is built by freezing and reusing the lower transferable layers of the pre-trained VGG-16 model, while adding a new dropout layer to it, to perform well on the image classification task without overfitting to the data. By selecting a subset of the ILSVRC-2012 dataset for our user study, consisting of ten labels from the original category set, we ensure that layers of the VGG16

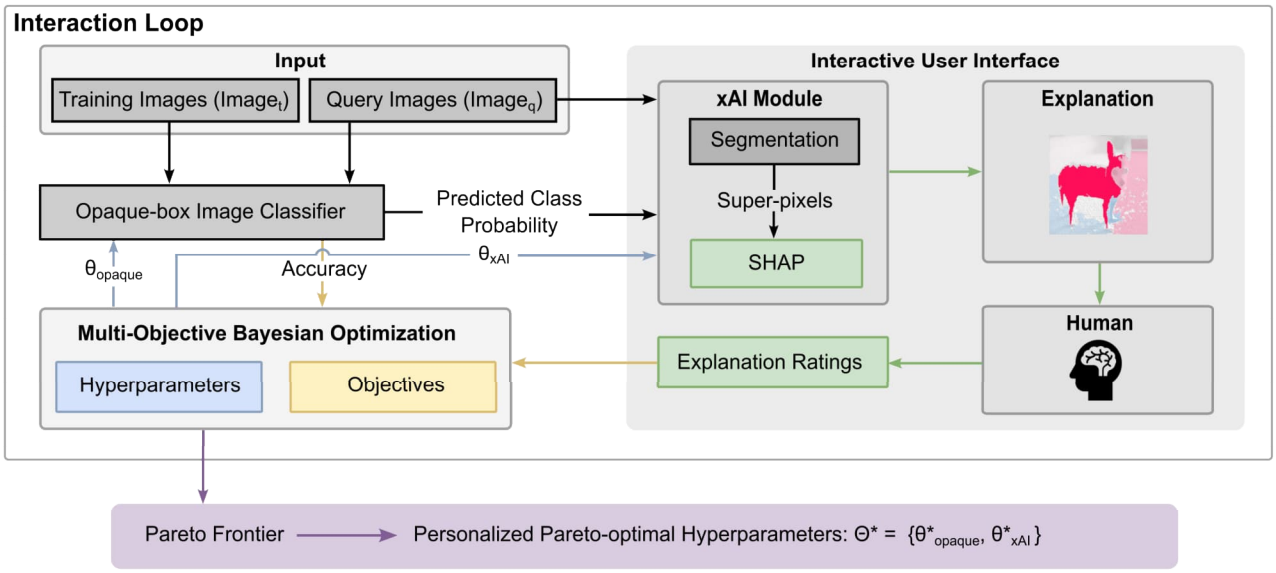

Figure 1: An application to jointly personalize for accuracy and explainability: $Image_t$ and $Image_q$ are partitions of the image dataset for classifier training and human explainability rating respectively; θ_{opaque} and θ_{xAI} are hyperparameters for the image classifier and the explainer module respectively; Θ^* is the selected Pareto optimal set of hyperparameters.

model have good transferability to the classification model in our study. The main reason to use this subset is that the original dataset has many thousands of labels which would result in an infeasible training time for an online interactive method. We built the model with Keras [8] and Tensorflow [1].

4.0.2 Segmentator. We used Felzenszwalb's algorithm for the segmentation [15] of classified images. This is a highly efficient graph-based image segmentation algorithm. For our implementation, we ran the algorithm with the help of scikit-image [47], a Python package for image processing algorithms.

4.0.3 SHAP Explainer. The segmented images are passed to an explainable AI module that uses SHAP [29], a popular explainable AI method for explaining image classifications. Visual explanation methods such as SHAP explain image classifications by shading the pixels of an image that were relevant for the image classification. We used Kernel SHAP as the method to estimate SHAP values for generating explanations, with the help of the shap package[1].

4.0.4 Multi-objective Bayesian Optimization. We chose parallel Noisy Expected Hypervolume Improvement (qNEHVI) [11] as the acquisition function. The MOBO optimizer was built using Ax[2], which is powered by BoTorch [4].

4.0.5 Hyperparameters. We chose three hyperparameters for interactive optimization – the dropout rate and learning rate from the image classifier, and the sigma of the segmentator.

The drop-out rate is the probability of dropping nodes during training. It was chosen because it is an important hyperparameter in deep-learning models that controls regularization and influences predictive performance. The probability for retaining nodes is typically set to values between 0.5 and 0.8 [43]; given this and the result of our pilot study, we set the range for dropout rate values to be between 0.01 and 0.6, which MOBO explores in its iterations.

The learning rate is another typical hyperparameter of deep learning models. By controlling the step size at which updates to model weights are made during training, it influences the model's predictive performance and training time. The optimal value for the learning rate is heavily dependent on model architecture [40] and is restrained by limited training epochs [48]. We performed a simple range test to observe how model performance changes with learning rates of different magnitudes for our setting, and observed that the accuracy fluctuated with the learning rate on 10^{-2} and was able to converge smoothly with the learning rate on 10^{-5}. Thus, we set its range to between 0.00005 and 0.02.

The third hyperparameter we set, Sigma, is relevant for image segmentation algorithms that use a Gaussian filter to preprocess images. It refers to the standard deviation of the Gaussian filter – large sigma values result in more smooth and blurry images compared to small ones. This is an important hyperparameter as our explainer contains a segmentator, meaning the value of sigma can determine the type of explanations that are produced. Since the appropriate range of sigma is highly dependent on the image dataset, we used a range test and considered commonly used defaults in segmentation algorithms, and picked the range for sigma to be between 0.2 to 0.9.

4.0.6 Objectives. Objectives refer to metrics we seek to optimize. In our case, we seek to jointly optimize the objectives of perceived explainability (as evaluated by the user), and classifier accuracy (calculated after each MOBO iteration). Accuracy is measured as the fraction of the validation set that is correctly classified. Explainability in our system is a metric quantified by the user; more specifically, it is the average of the 3 sub-ratings for background

[1]https://github.com/slundberg/shap
[2]https://ax.dev/

noise, object body, and main features, that are described in the next section. Both these metrics are opaque-box optimization problems, making them hard to optimize directly.

5 METHOD

For the user study, we use the implementation described in the previous section. The user study aimed to carry out the interactive loop (see illustrated in Figure 1) and examine if the study context would give rise to an accuracy vs. explainability trade-off for participants. Further, we aimed to examine the performance of the personalized classifier and explainer on held-out images as a function of image similarity across two training conditions with appropriate baselines. The user study accordingly had four parts: (i) the practice session, (ii) interactive hyperparameter-tuning of the classifiers, (iii) the validation study, and (iv) a short interview.

5.1 Participants

We recruited 12 adult participants (7 females, 5 males) through our group's directory of past participants and by posting on social media. All participants were students or staff at Aalto University, where the study took place. They were compensated 90€ for their participation in the user study which took around 7 hours in total. The study was conducted in accordance with the principles stated in the Helsinki Declaration as well as a local procedure for ethics approval. All participants volunteered under informed consent and agreed to the recording and anonymized publication of results.

5.2 Apparatus

The user study was carried out using desktop computers that ran the Ubuntu 20.04 64-bit operating system, with Intel(R) Xeon(R) CPU E5-1650 v4, and an NVIDIA Quadro P5000 graphics card. We provided the participants with a 23.8" screen and a standard optical mouse and keyboard. For the second phase of the study with human-in-the-loop hyperparameter tuning, we ran the two tuning conditions alternatively on two computers with the same configuration.

5.3 Rating Scheme

In order to have a consistent scale across the hyperparameter tuning iterations, and to meaningfully aggregate experimental data across participants, we devised standardized scales for evaluating explainability. The rating scheme uses the perceived (red) coverage of the explanations along three dimensions: the object body, the main features of the object, and the (lack of) background noise. Participants correspondingly rate (on a scale of 1 to 10), the percentages of (i) the object's body captured by the explanation, (ii) the object's main features (as assessed by the participant; Figure 2(A) assumes that the face is the main feature), and (iii) the irrelevant background features not captured by the explanations.

Figure 2 (A) provides two examples of the rating scheme. Separating the ratings into three separate dimensions enables participants to focus on only one aspect of the explanation at a time, and put numbers to potentially complicated-looking explanations such as those in Figure 2(B). The average of these three ratings is sent to MOBO after each iteration. Even while we standardize ratings to enable meaningful comparison or aggregation across users, the personalization element is preserved due to each participant needing to subjectively determine the key distinguishing features of the presented image. We used what we thought was a reasonable way to rate explanations. However, further research is required in the field to determine the most effective methods for users to assess visual explanations [23, 38].

5.4 Dataset

To shorten the time needed for image classifier training and image classification, we chose 10 animal categories in total. 6 of these are types of dogs (namely American bulldog, American pit bull terrier, Basset hound, Beagle, Boxer, and Chihuahua), and 4 of these are types of cats (namely Abyssinian, Bengal, Birman, and Bombay). We formed the main dataset largely based on the Oxford-IIIT Pet dataset [32][3], which has 37 pet categories with about 200 images in each class. To prevent overfitting in the CNN model due to the relatively small 10-category dataset, we enlarged it through data augmentation and through adding images with these labels from the ILSVRC2012 dataset [35].

In the user study, we split the main dataset into 2 sub-datasets, $Image_t$ and $Image_q$, where the former is for model training and the latter is for generating explanations. $Image_t$ had 4050 images of 10 categories, which were then split into training set, validation set and test set by random stratified sampling, with the proportion being 7:2:1. For explanation rating, we manually picked images for $Image_q$ to make sure that there was only one animal object of the 10 categories in each image. Apart from that selection criteria, images were picked at random. Since we have two different training conditions and one validation study (see section 5.5), $Image_q$ was further split into 3 sub-datasets randomly: $Image_{q-narrow}$, $Image_{q-broad}$, and $Image_{val}$, where $Image_{q-narrow}$ had 12 images of one dog category Chihuahua; $Image_{q-broad}$ had 12 images of 3 dog categories Chihuahua, Basset hound and Beagle; $image_{val}$ had 40 images of 10 categories, with 8 from Chihuahua, Basset hound, Beagle respectively and 4 from American bulldog, American pit bull terrier, Boxer and cats respectively. $Image_{q-narrow}$, and $Image_{q-broad}$ are representative of narrow or broad variety of images for use in the training conditions, while $Image_{val}$ is a dataset held out for validating personalized explanations across category labels.

5.5 Experimental Design

Details for the four different parts of the user study follow.

5.5.1 Practice. The study begins with a practice session to ensure that participants understood the instructions, and are familiarized with the rating scales.

5.5.2 Interactive hyperparameter tuning. In our method, interactive hyperparameter tuning is the basis on which the classifier-explainer pair for each participant is personalized. In 20 iterations, participants provide feedback on a set of 12 fixed query images, culminating in a Pareto-frontier that reflects any trade-offs between accuracy and explainability. We considered two experimental conditions to examine the influence of qualitatively different training images $Image_q$ on the resulting Pareto-frontier and carry-over of

[3] https://www.robots.ox.ac.uk/~vgg/data/pets/

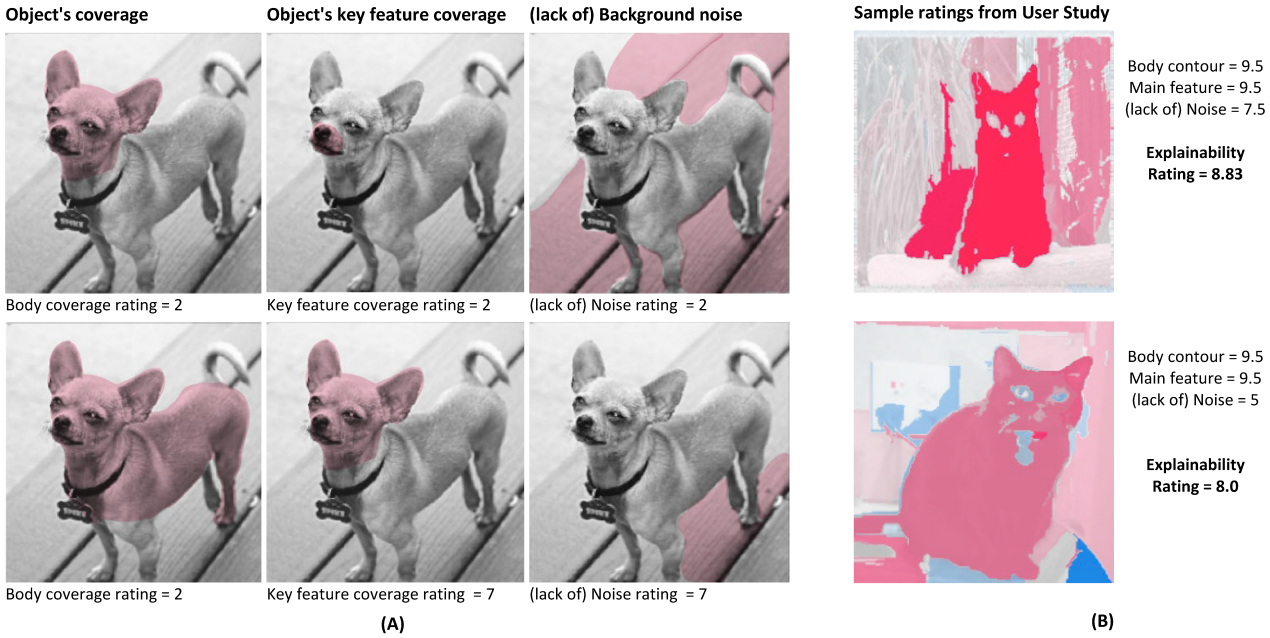

Figure 2: (A) The rating scheme provided a standard rating scale (from 1 to 10) for participants based on (i) coverage of the explained object, (ii) coverage of the key features for identifying the explained object (as determined by each user), and (iii) (lack of) coverage of irrelevant background features. (B) Two samples of ratings provided by participants in the user study.

explainability to new settings. In the first condition (Narrow-HITL), the images presented to participants, $Image_q$, were sampled from a single narrow category, while the second condition (Broad-HITL) consisted of $Image_q$ from a broader semantic category as detailed in the earlier description of the dataset.

5.5.3 Validation study. In the validation study, we are interested in investigating: (i) the evaluations of personalized explanations for held-out images, (ii) the relationship between explained images and queried images (examined in terms of category membership of the held-out validation images), (iii) the difference between using personalization over the classifier-explainer hyperparameters versus only the personalized explainer hyperparameters, and (iv) the performance of the personalization approach against plausible baseline approaches. With these aims in mind, two validation studies were conducted to evaluate the performance of the personalized Pareto-optimal model against the baseline models.

To define notation, we index each participant by an integer i (ranging from 1 to 12 in our study), so that $\theta_i^* = \{\theta^*_{opaque_i}, \theta^*_{xAI_i}\}$ represents their personalized Pareto-optimal parameters for the accuracy vs. explainability trade-off chosen after the interactive hyperparameter tuning (in either Narrow-HITL or Broad-HITL conditions). Each of the models that use personalized hyperparameters and the baselines, make predictions and generate explanations for held-out data $image_{val}$. It is these explanations of $image_{val}$ that are evaluated by participants in the blinded studies.

Since we propose a novel method for joint personalization over multiple objectives, we lacked standard baselines or other methods for comparison. Hence we devised two plausible baselines (producing two sets of baseline parameters) and conducted a blinded validation study for each. Apart from employing distinct baseline and personalized models for classification and explanation, both studies shared the same design. The ordering of the studies should have minimal impact due to its blinded nature; participants were only asked to rate different explanations of varying quality through the two studies, and in each trial order them based on what seemed to be a better explanation.

Validation Study 1: Baseline (BO) vs. HITL. In the first validation study, we set the baseline by emulating a practitioner who has access to standard BO but is not considering personalized explanations. They would have plausibly carried out hyperparameter-tuning separately for each objective – once for accuracy with BO, and once for the quality of segmentation with Human-in-the-loop BO, to generate a θ_{opaque} and a θ_{xAI} (sigma for the segmentator). Hence the baseline parameter θ_{opaque} was set by using BO to optimize the classifier for accuracy. The quality of segmentation was used as the objective to obtain θ_{xAI} for human-in-the-loop BO carried out by the experimenter. BO for each of the objectives was carried out here for twenty iterations, to match the number of iterations in the experimental condition.

Validation Study 2: Baseline (default sigma) vs. HITL sigma. In the second validation study, we investigated whether solely employing the explainer parameter from the HITL phase would yield satisfactory performance. Consequently, the classifiers were fixed to be the same across the different conditions and the baseline (with θ_{opaque} that maximized accuracy). However, the θ_{xAI} values were informed by the interactive personalization phase that had been carried out. The baseline for this study used the commonly used segmentator sigma default value $\theta_{xAI} = 0.8$.

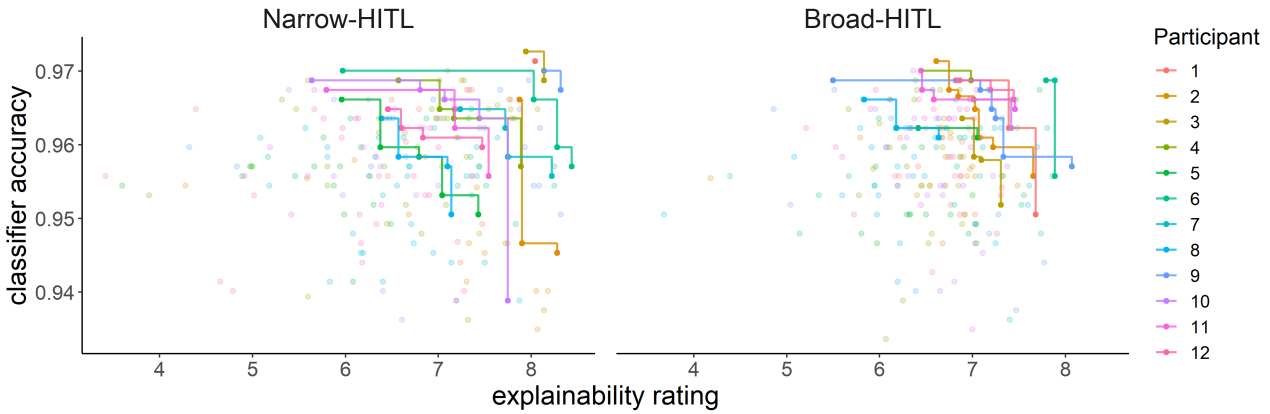

Figure 3: Pareto fronts between accuracy and explainability for each participant, obtained using interactive hyperparameter tuning with MOBO, when the queried images are from a narrow semantic category (Left) and a broad semantic category (Right). Light dots represent other parameterizations explored by the participant during the hyperparameter tuning phase.

5.5.4 Interview. At the end of the study, participants were interviewed about their experience. We asked about (i) the main feature(s) they chose during the study to understand the variability in this factor; (ii) their perceived consistency while rating explanations along the scale to understand noisiness in their input, and any other feedback they had.

5.6 Procedure

In each phase of the user study, the main task for the participants in the study was to rate explanations of image classifications based on the rating scheme that we provided. In the practice session, we provided participants with oral and printed instructions as well as the printed rating scheme. For the practice phase, 30 example explanation images with a variety of expected ratings along the three rating dimensions were provided (via a program created using PsychoPy [33]). Each entered rating was followed by feedback on the ideal target rating. None of these examples were used again in the user study. In the next part of the study, interactive hyperparameter tuning was carried out by participants for the two conditions, Narrow_HITL, and Broad_HITL on two computers in parallel[4]. Due to the time taken to gather ratings on the two conditions over twenty iterations as well as the time taken for classifier training, this part of the procedure resulted in an average of 6.5 hours per participant. The interface for the participants constituted a program running within a Jupyter notebook. Participants would see an image from $Image_q$ and its visual explanation adjacent to each other, and thereafter input the three ratings below the images. Participants would then enter the number between 1 and 10 that represented their ratings and scroll down to the next image .

In the validation phase, the two validation studies took place sequentially. The interface for this phase also constituted a Jupyter notebook, but now with 4 images adjacent to each other. The first image was the original image from $image_{val}$, and the next three were randomly positioned explanations generated by the participants' personalized parameters for (i) Narrow-HITL, (ii) Broad HITL, and (iii) the Baseline for the study. The personalized Pareto-optimal parameters chosen were those associated with the highest explainability for each of the participants[5]. Participants provided the three ratings (on the rating scale) for each of the explanations as before, and also provided a ranking for the explanations based on how sensible the explanations seemed to them, before moving on to the next image from $image_{val}$ and their corresponding explanations. These rankings among the three explanations were a way for us to test if participants' general preferences for the explanations aligned with those predicted by the rating scale. At the end of the study, we conducted a short interview with each participant as described in the previous section.

6 RESULTS AND DISCUSSION

Interactive Personalization phase. The initial interactive hyperparameter tuning stage of the user study resulted in the Pareto-fronts for participants shown in Figure 3 for the Narrow-HITL and Broad-HITL training conditions. These Pareto fronts demonstrate that there is an accuracy vs. explainability trade-off in the training setting. Qualitatively, participants mostly had broad Pareto fronts and obtained significant gains in explainability (around 2 points on a 10-point scale) while trading off up to 2% in accuracy. Although a 2% trade-off in accuracy might appear to be minor, it must be noted that deep learning classifiers' baseline accuracies often tend to be quite substantial, here at around 97%. In a couple of cases, the discovered Pareto front consisted of only one or two points - while this still captures personalization as it is based on subjective ratings, it implies reduced freedom in terms of personalization for the desired trade-off during deployment, unless even more iterations are carried out to discover new Pareto-optimal parameterizations.

The variety of Pareto fronts in Figure 3 reflects the differences between participants' trade-offs between accuracy and explainability, which captures the first element of personalization. Narrow-HITL

[4]Since training the classifier after every iteration took about 10 minutes, participants alternated between iterations of Narrow_HITL and Broad_HITL so that they provided ratings on one of the systems while the other was busy with training the classifier.

[5]Occasionally the second-highest was chosen if the explainability ratings were within 0.1 of each other and the accuracy values were much higher.

had broader Pareto-fronts than Broad-HITL – the larger variance in labels and features in the Broad-HITL condition may have made it harder to obtain a wider range of Pareto-optimal parameters. However, it is possible that with more hyperparameter turning iterations, broader Pareto fronts can be obtained for both conditions. Figure 4 shows a sample of explanations generated from different regions of the Pareto front for a participant – the parameters that optimize for accuracy have lower perceived explanation quality than the ones that optimize for explainability.

Validation phase. The first part of the validation study examined the quality of the explanations generated by the personalized classifier and explainer $\{\theta^*_{opaque_i}, \theta^*_{xAI_i}\}$ for held-out images. The explanation quality was assessed across the training conditions and category memberships of the held-out images. Figure 5 shows that participants rated explanations generated by the personalized models to be better than those generated by their corresponding baselines for the validation images – this was the case irrespective of whether the held-out images were from within the same or different class as the queried images, and supported by statistical analyses.

Statistical tests comparing baselines and the HITL conditions were separately carried out with one-way repeated measures ANOVA, with the mean explainability ratings set as the dependent variable. For the Narrow-HITL evaluation, significant main effects were found for both the explanation source, $F(1,11) = 172.82$, $p < .01$, and Image Class, $F(1,11) = 24.25$, $p < .01$. The interaction of Explanation Source × Image Class was not significant, $F(1,11) = 0.01$, $p = .941$. For the Broad-HITL setting significant main effects were found for both, the explanation source, $F(1,11) = 153.64$, $p < 0.01$, and Image Class, $F(1,11) = 40.83$, $p < .01$. The interaction of Explanation Source × Image Class was not significant, $F(1,11) = 1.55$, $p = .238$. These observations indicate that the mean explainability ratings for the Narrow-HITL and Broad-HITL conditions were higher than those of their respective baselines.

The second part of the validation study and its analysis were the same in all respects as the first validation study, except the explanations were generated across the conditions by a fixed classifier optimized only for accuracy θ_{opaque}, but with a personalized explainer $\theta^*_{xAI_i}$, for each participant from the training phase. The results for this part of the study are presented in Figure 6. From the two separate one-way repeated measures ANOVA, significant main effects were found for both the explanation source, $F(1,11) = 5.21$, $p = .043$, and Image Class, $F(1,11) = 26.03$, $p < .001$. The interaction of Explanation Source × Image Class was not significant, $F(1,11) = 1.94$, $p = .191$. For the right panel, significant main effects were found for both the explanation source, $F(1,11) = 25.04$, $p < .001$, and Image Class, $F(1,11) = 34.26$, $p < 0.001$. The interaction of Explanation Source × Image Class was significant, $F(1,11) = 5.06$, $p = .046$. These analyses indicate that for both Narrow-HITL and Broad-HITL, the mean explainability ratings across the two image classes were different with the Same Class having higher ratings than the Different Class and no significant interactions between the explanation source and membership class.

Figure 6 shows that the individual personalization approach for the explainer produced explanations better than the default explainer sigma for images from the same class, and showed almost equivalent explanations for images from a different class. The key takeaway here is that with just a few iterations of MOBO, we were able to obtain results similar to those that took the field a lot of trial and error to get to. In novel multiple objective explainability scenarios, there may not be defaults available; our results suggest that human-in-the-loop MOBO might be suitable for such situations.

In addition to using the rating scheme, participants also ranked explanations presented to them based on their preferences with respect to explainability. The results of this step are presented in Figure 7. We used four repeated measures ANOVA analyses to examine the differences in participants' preferences for explanations, for the combinations of the two different types of personalization (joint vs. explainer only), and the two different types of training (Narrow HITL vs. Broad HITL). In each, we examined the effect of explanation source and class membership of the held-out images with respect to the queried images, compared to the respective baselines.

In the first repeated measures ANOVA (relevant to the top left of Figure 7), there were significant main effects of explanation source, $F(1,11) = 23.3$, $p < .001$, and class membership, $F(1,11) = 82369$, $p < .001$. Post-hoc Bonferroni tests revealed that the preferences for explanations within the same class as queried class was significantly higher than when held-out images were from a different class ($p < .001$); preferences were also significantly higher for Narrow HITL compared to its corresponding Baseline ($p < .001$). Similarly, for the second repeated measures ANOVA (relevant to the bottom left of Figure 7), there were significant main effects of explanation source, $F(1,11) = 15.02$, $p = .003$, and class membership, $F(1,11) = 9409$, $p < .001$. Post-hoc Bonferroni tests again showed preferences for explanations within the same class compared to a different class ($p < .001$); Broad HITL was preferred compared to its Baseline ($p = .003$).

The next analyses were for the personalization that was applied only to the explainer while keeping neural network hyperparameters fixed. There were no significant main effects of explanation source or class membership in the data, even though a slight preference for personalized explanations seems to be present within the same class on visual examination of Figure 7. The reason for this could be the high quality of the baseline making the preference data less sensitive than the rating data along three dimensions.

Interview. All participants reported being consistent in their ratings through the study. 11 participants reported choosing the same "main feature" while rating explanations ("head"), and one participant reported choosing the face ("head minus the ears"). We did not anticipate this; even though we validate our method with held-out images, a future study would benefit from ensuring sufficient variance in the subjective aspects of evaluation across participants.

7 SUMMARY AND OUTLOOK

In this article, we discuss and demonstrate with a user study, an interactive method for personalizing opaque image classifiers for explainability while also optimizing its predictive accuracy. Our method is based on MOBO which is data-efficient and robust to noisy inputs. The key takeaways of the approach and the user study are the following:

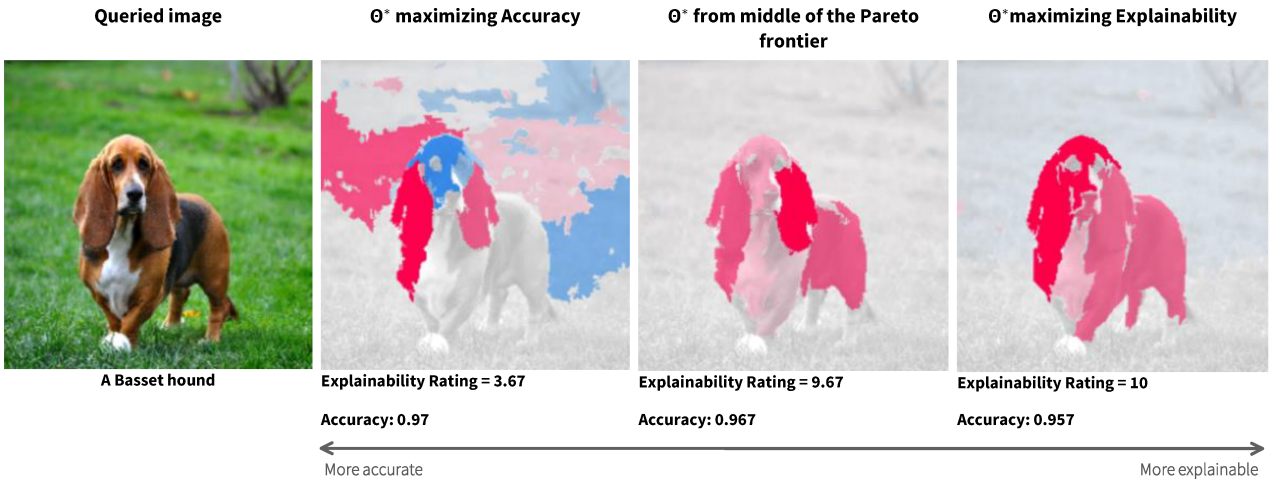

Figure 4: Example of explanations generated by Pareto-optimal parameters in the study which capture different accuracy vs. explainability trade-offs for a particular user.

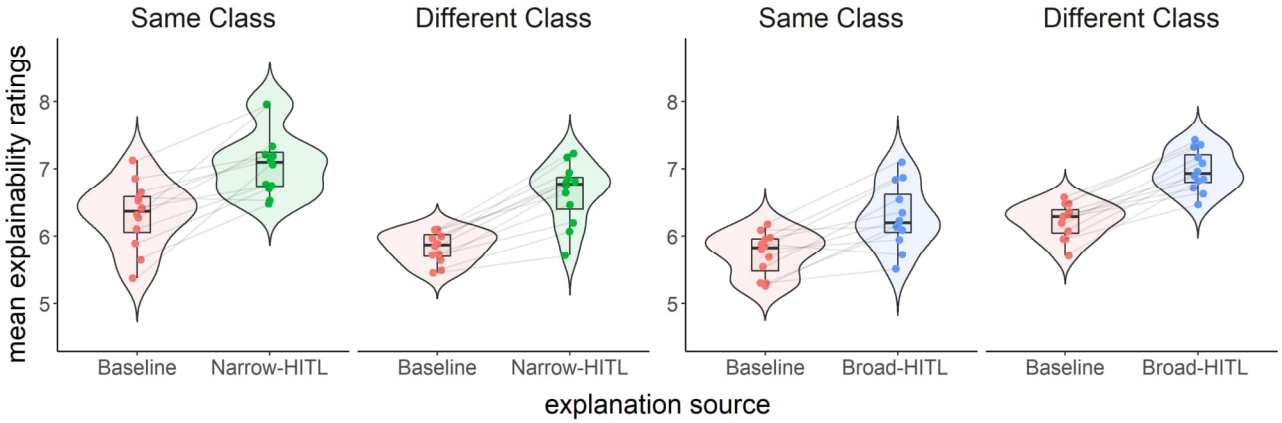

Figure 5: Validation study 1: The mean explainability ratings on held-out images of each participant across different explanation sources and different image classes. Personalized models resulted in higher ratings than their corresponding baselines irrespective of whether the held-out images belonged to the categories seen during training. Red represents the Baselines, green represents Narrow-HITL, and Blue represents Broad-HITL. Each dot connected by a line represents the mean ratings of the participant. Predictions and explanations here are generated by the personalized $\{\theta^*_{opaque_i}, \theta^*_{xAI_i}\}$ from the HITL phase, for each participant i.

- We achieve personalization in our interactive method through two routes: (i) by taking into account the user's subjective evaluations of explainability, and (ii) by enabling users to pick a personally preferred accuracy vs. explainability trade-off from their Pareto fronts.
- The main component of the approach involves interactive hyperparameter tuning. The hyperparameter space of the classifier and the explainer are jointly explored by MOBO with the predictive accuracy of the classifier and user evaluation of explainability as the objectives. During each iteration, accuracy is calculated by re-training the classifier, and evaluation of explanations is provided by a human in the loop.
- In a blinded validation study with held-out images, users consistently rated personalized explanations to be higher than explanations generated by plausible baseline techniques for improving general explainability.
- User evaluations of personalized explanations of new images depended on the similarity of these images with the query images used to identify their Pareto fronts – higher explanation ratings were observed when the target images were from the queried image class.

In this study, we were limited to twelve participants due to the length of time and resources taken for each run of the experiment. However, the within-subjects study design is not only more suited to assess personalization but also enables higher statistical power with a smaller sample size [19]. We were also limited in this study by the amount of computation required for classifier training after each iteration which made it necessary for us to limit ourselves to

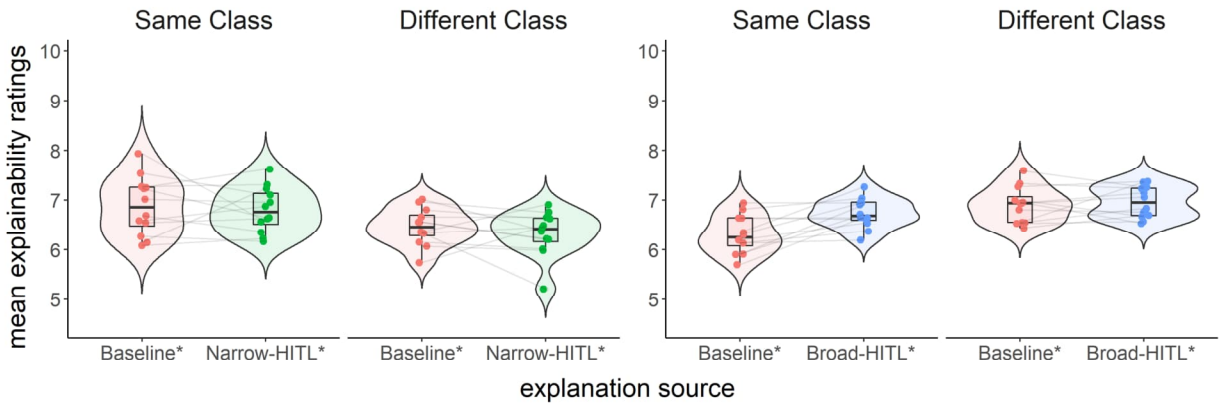

Figure 6: Validation study 2: The mean explainability ratings on held-out images of each participant across different explanation sources and different image classes. Red represents the Baseline*, green represents Narrow-HITL*, and Blue represents Broad-HITL*. Predictions here are generated by the same classifier optimized in the usual way for accuracy while borrowing θ^*_{xAI} from the HITL phase, optimizing for explainability.

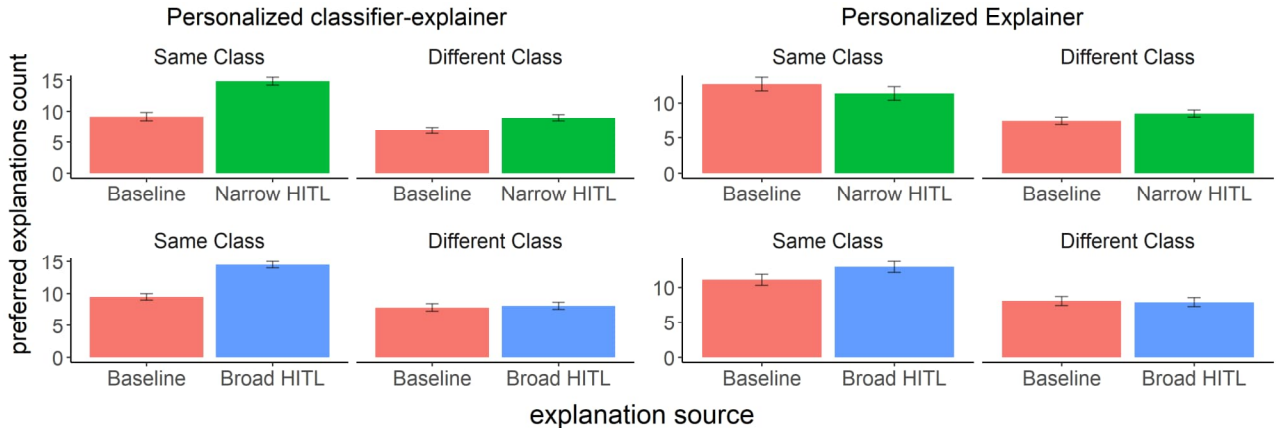

Figure 7: In the validation studies, participants provide preferences for explanations generated by baselines and models personalized to them. The top panel shows performance of Narrow-HITL against baselines, based on the relationship between queried and target images. The bottom panel compares Broad-HITL against baselines. Personalized explanations are preferred to baselines for personalized classifier-explainer (Left); however, no significant differences are found for the personalized explainer (Right).

ten labels from the original dataset. Even with the simplification, the study provides evidence that personalization for explanations is feasible and that methods like MOBO would be promising. It would be worth exploring quicker methods for classifier training and MOBO [27] in a future study for practical applications. We employed category membership as a proxy for the similarity between new images and queried images; however, a continuous measure of pairwise similarity between individual images (e.g. [34]) may be better suited to assess how well explainability transfers to new images. Ethically, the field of explainable AI contributes to transparency, understanding, and fairness. Even though personalization for explainability has the potential to enhance these attributes, it would generally be a good practice to validate explanations and understand the opaque model better before using it to guide societally impactful decisions.

ACKNOWLEDGMENTS

This work was supported by the Finnish Center for Artificial Intelligence (FCAI), the Academy of Finland through the projects Human Automata (grant 328813) and BAD (grant 318559), and the Aalto Science-IT Project. We thank Calvin Suarez for discussions stemming from his work in a similar direction. We thank the reviewers for their feedback.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[3] Gilles Bailly, Antti Oulasvirta, Timo Kötzing, and Sabrina Hoppe. 2013. Menuoptimizer: Interactive optimization of menu systems. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 331–342.

[4] Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. 2020. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*. http://arxiv.org/abs/1910.06403

[5] Eric Brochu, Tyson Brochu, and Nando De Freitas. 2010. A Bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 103–112.

[6] Liwei Chan, Yi-Chi Liao, George B Mo, John J Dudley, Chun-Lien Cheng, Per Ola Kristensson, and Antti Oulasvirta. 2022. Investigating Positive and Negative Qualities of Human-in-the-Loop Optimization for Designing Interaction Techniques. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[7] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[8] François Chollet et al. 2015. Keras. https://keras.io.

[9] Roberto Confalonieri, Tarek R Besold, Tillman Weyde, Kathleen Creel, Tania Lombrozo, Shane T Mueller, and Patrick Shafto. 2019. What makes a good explanation? Cognitive dimensions of explaining intelligent machines.. In *CogSci*. 25–26.

[10] Yae Dai, HongWu Ye, and SongJie Gong. 2009. Personalized recommendation algorithm using user demography information. In *2009 Second International Workshop on Knowledge Discovery and Data Mining*. IEEE, 100–103.

[11] Sam Daulton, Maximilian Balandat, and Eytan Bakshy. 2021. Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement. In *NeurIPS*.

[12] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[13] Mark Dunlop and John Levine. 2012. Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2669–2678.

[14] Brochu Eric, Nando Freitas, and Abhijeet Ghosh. 2007. Active preference learning with discrete choice data. *Advances in neural information processing systems* 20 (2007).

[15] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. 2004. Efficient Graph-Based Image Segmentation. *Int. J. Comput. Vis.* 59, 2 (2004), 167–181. https://doi.org/10.1023/B:VISI.0000022288.19776.77

[16] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2020. Auto-sklearn 2.0: The next generation. *arXiv preprint arXiv:2007.04074* 24 (2020).

[17] Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. 2020. Fostering human agency: a process for the design of user-centric XAI systems. (2020).

[18] Roman Garnett. 2022. Bayesian Optimization.

[19] Anthony G Greenwald. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin* 83, 2 (1976), 314.

[20] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web* 2, 2 (2017), 1.

[21] Florian Kadner, Yannik Keller, and Constantin Rothkopf. 2021. Adaptifont: Increasing individuals' reading speed with a generative font model and bayesian optimization. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[22] Mohammad M Khajah, Brett D Roads, Robert V Lindsey, Yun-En Liu, and Michael C Mozer. 2016. Designing engaging games using Bayesian optimization. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5571–5582.

[23] Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. Hive: evaluating the human interpretability of visual explanations. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*. Springer, 280–298.

[24] Alexandra Kirsch. 2017. Explain to whom? Putting the user in the center of explainable AI. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2017)*.

[25] Joshua Knowles. 2006. ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation* 10, 1 (2006), 50–66.

[26] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[27] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qingfu Zhang, and Sam Kwong. 2018. A batched scalable multi-objective bayesian optimization algorithm. *arXiv preprint arXiv:1811.01323* (2018).

[28] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[29] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[30] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. 2018. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM conference on recommender systems*. 31–39.

[31] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[32] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), 3498–3505.

[33] Jonathan Peirce, Jeremy Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51 (02 2019). https://doi.org/10.3758/s13428-018-01193-y

[34] Brett D Roads and Bradley C Love. 2021. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 3547–3557.

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[36] Manali Shaha and Meenakshi Pawar. 2018. Transfer Learning for Image Classification. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 656–660. https://doi.org/10.1109/ICECA.2018.8474802

[37] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. 2015. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* 104, 1 (2015), 148–175.

[38] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. 2023. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction* 39, 7 (2023), 1390–1404.

[39] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556* (09 2014).

[40] Leslie N. Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *CoRR* abs/1803.09820 (2018). arXiv:1803.09820 http://arxiv.org/abs/1803.09820

[41] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* 25 (2012).

[42] Kacper Sokol and Peter Flach. 2020. One explanation does not fit all. *KI-Künstliche Intelligenz* 34, 2 (2020), 235–250.

[43] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1 (jan 2014), 1929–1958.

[44] Srikanth Tammina. 2019. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)* 9, 10 (2019), 143–150.

[45] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2013. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 847–855.

[46] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing

interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).

[47] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. 2014. scikit-image: image processing in Python. *PeerJ* 2 (2014), e453.

[48] D.R. Wilson and T.R. Martinez. 2001. The need for small learning rates on large problems. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, Vol. 1. 115–119 vol.1. https://doi.org/10.1109/IJCNN.2001.939002

[49] Yongfeng Zhang, Xu Chen, et al. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.

[50] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.