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

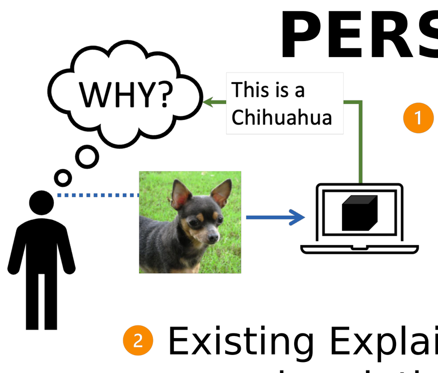
Aalto University

User Interfaces
Aalto University

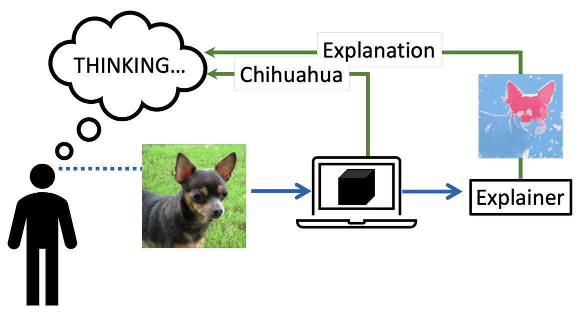
Suyog Chandramouli[*,†], Yifan Zhu[*,†,‡], Antti Oulasvirta[†]

[*]Equal Contribution
[†]Department of Information and Communications Engineering, Aalto University, Finland
[‡]Department of Computer Science, Aalto University, Finland

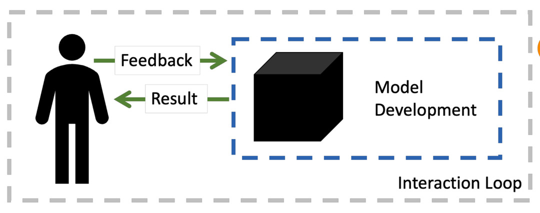
PUT HUMANS BACK IN THE LOOP FOR PERSONALIZING EXPLAINABILITY



1. Deep learning models have highly powerful predictive capabilities; however, the results are mostly **uninterpretable** due to the complicated model structure.



2. Existing Explainable ML techniques typically offer generic solutions to users, lacking consideration of **individual preferences or backgrounds**.



3. Human-in-the-loop ML integrates humans into model development, allowing for securing user **autonomy** and enhancing comprehension of the model, thereby fostering better **personalization** and greater **trust** for end-users.

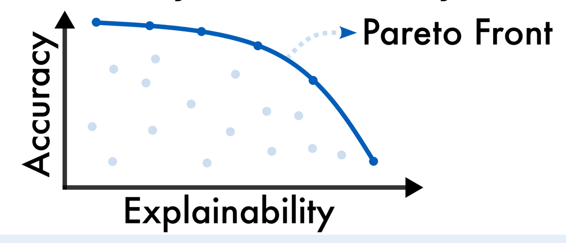
Can we offer personalized predictive models with improved expla-inability without compromising accuracy ?

CHALLENGES IN INTERACTIVE PERSONALIZATION

- Limited human feedback data
- Inherent noisiness in human feedback
- Expansive hyperparameter search space

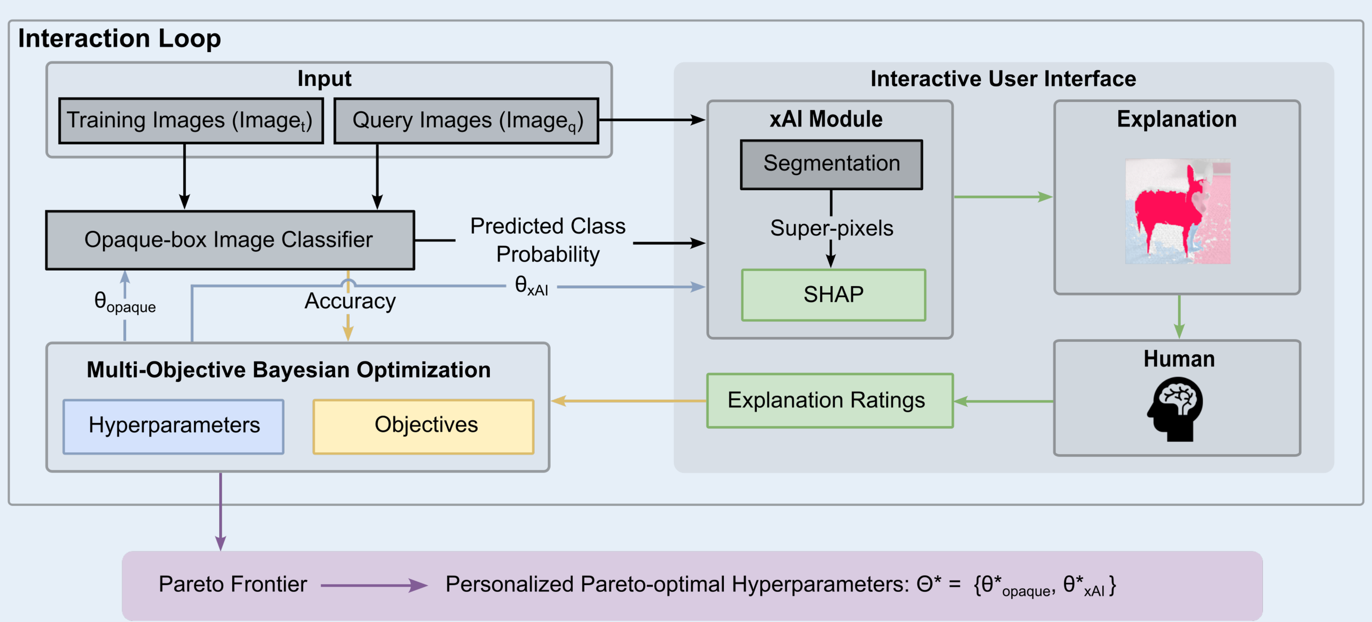
MAIN HYPOTHESIS

There exisits an inherent **trade-off** between model explainability and accuracy.



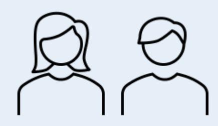
METHOD: HUMAN-IN-THE-LOOP MULTI-OBJECTIVE BAYESIAN OPTIMIZATION

Interactive loop for hyperparameter optimization with iterative evaluations from both the model and the human



1. **Multi-objective Bayesian Optimization (MOBO)**
   - Efficiently sample the hyperparameter space
   - Maximize both model accuracy and explainability rating
   - Identify a user-specific accuracy-explainability trade-off
2. **Classifier Training**
   - Transfer learning based on the pre-trained VGG16 model
3. **Explanation Generation**
   - Generate visual explanations with SHAP
4. **Human Evaluation**
   - Quantify model explainability with human ratings of visual explanations for the classification
   - Design a rating scheme to ensure consistent rating standards

USER STUDY



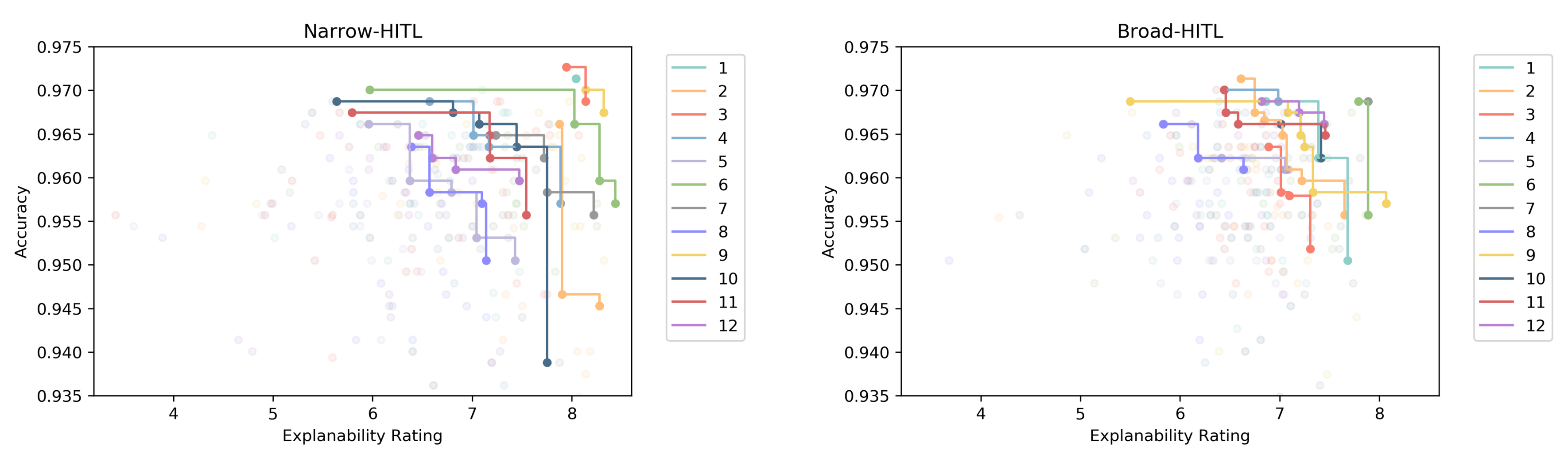
12 participants, 3 phases

Instructions and Practice

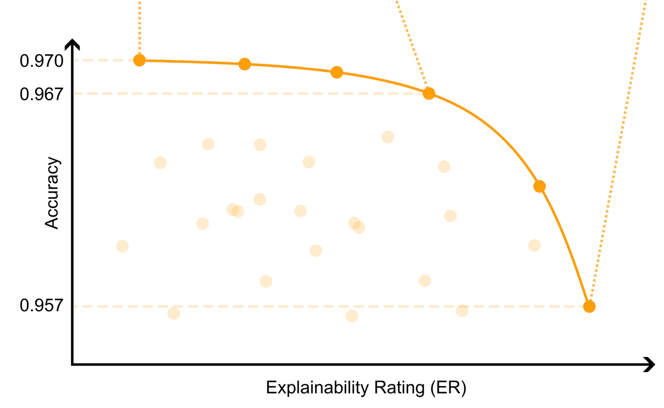
Interactive Hyperparameter Tuning

Validation Study

ACCURACY-EXPLAINABILITY TRADE-OFF

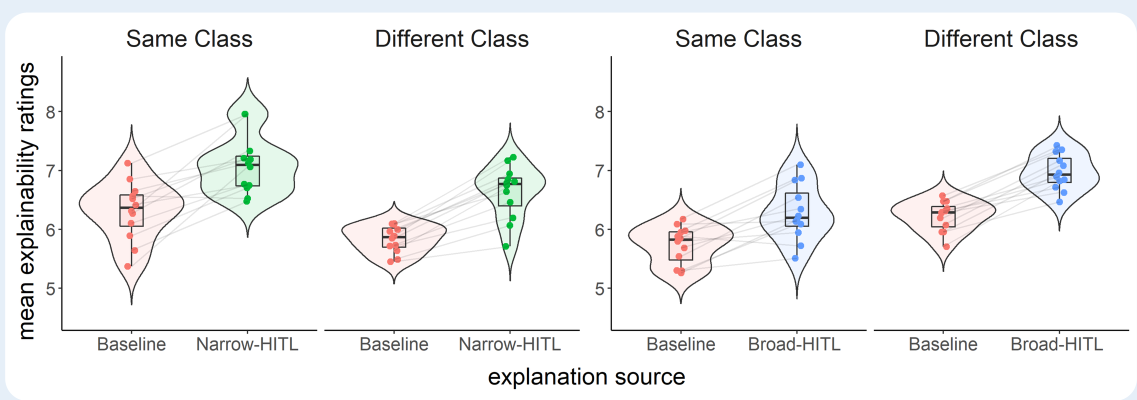


What does the trade-off mean in visual explanations?

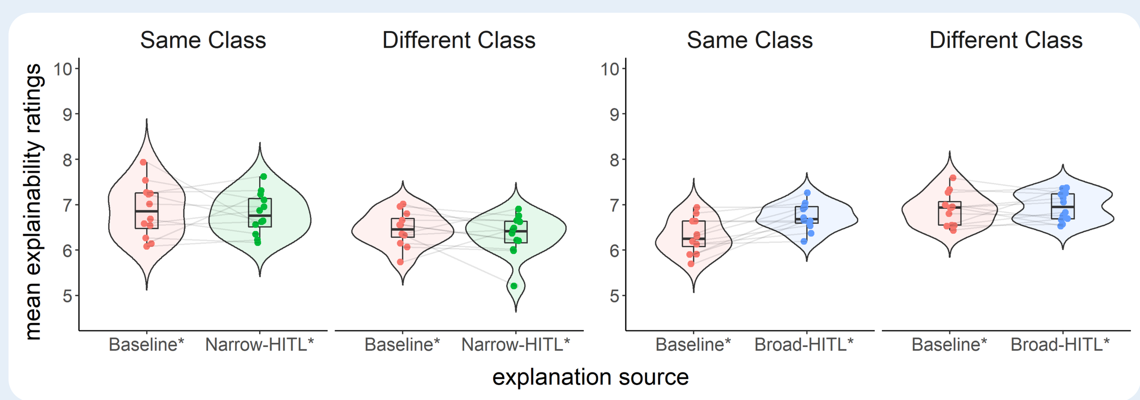


- Within the context of our specific problem setting and study design, a trade-off exists between the accuracy and human-perceived explainability of deep-learning-based image classifiers.
- Pareto-fronts were identified for 11 participants in both experimental conditions on training context.
- Pareto fronts from Narrow-HITL have a floating range of 2 points for explainability rating and around 2% for model accuracy on average.
- Pareto fronts from Broad-HITL have a floating range of around 1 point for explainability rating and 1% for model accuracy.
- With the same number of optimization iterations, contexts with lower variance in image category and features are more likely to identify Pareto optimal solutions with better performance and also higher variance.

PERSONALIZED EXPLAINABILITY QUALITY



Compared to the standard baseline optimized with BO only, personalized models in the condition Narrow-HITL obtained around 13% improvement in explainability rating at a cost of 1% accuracy on average; In Broad-HITL, the improvement in explainability rating was around 11% with a small 1% compromise in accuracy.



Compared to the default baseline set with default explainer hyperparameter, the personalized explainer obtained similar explainability ratings as the baseline for held-out images from both the same and different context.

TAKEAWAY

- Our method is capable of identifying optimal accuracy-explainability trade-offs for individuals.
- Personalized models obtained higher explainability ratings compared to a standard Bayesian optimization based baseline without human factor.
- Our method is capable for suggesting optimal default hyperparameters with only a few iterations.
- Personalization achieved via 2 routes
  - User's subjective evaluations of explainability
  - Enabling identifying a user-specific accuracy-explainability trade-off

Presenter: Yifan Zhu Incoming ELLIS PhD Student @AaltoUniversity @UniversityofManchester

Or check my Master's thesis!

The presenter is away?
Check the presentation recording for our UMAP paper!

Still interested?
Great!
Check our UMAP paper!

UMAP 2023