An Interactive Human-Machine Learning Interface for Collecting and Learning from Complex Annotations

Jonathan Erskine, Matt Clifford, Alexander Hepburn,
Raúl Santos-Rodríguez

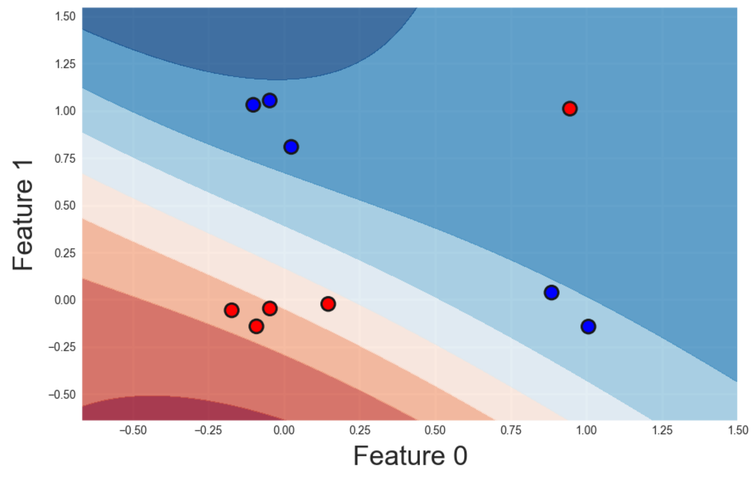
University of Bristol

1. Introduction

- We propose a graphical user interface for human annotators to evaluate machine learning systems beyond simple metrics in the form of *meta-evaluations*, and provide additional supervision during learning in *complex annotations* that go beyond simple class assignments.
- We show a potential loss function that utilises these complex annotations.
- As counterfactuals have been proven to be an effective learning signal [1, 2], we propose that a human can provide us with *counterfactual directions*.

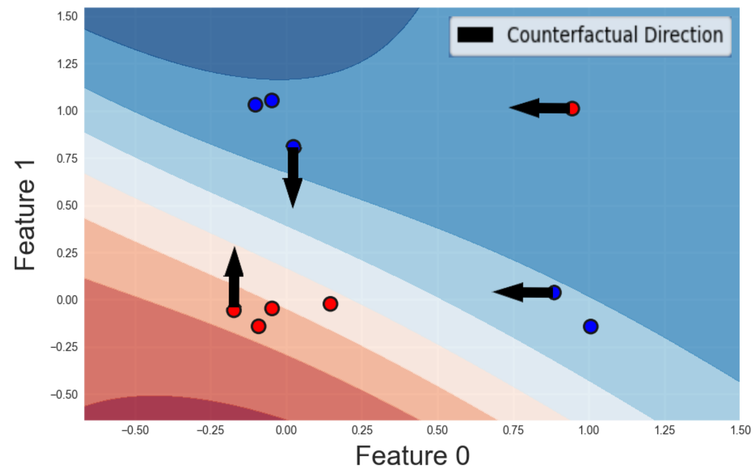
2. Meta-evaluation

- A *meta-evaluation* is an evaluation that goes beyond simple metrics. Here, we present the annotator with the model's decision surface.



3. Counterfactual Directions

- For a selected data point $\mathbf{x}_i$, humans can provide annotations as a direction $\mathbf{d}_i$ which indicates a change of class for $\mathbf{x}_i$.



4. Loss Function

- Given a counterfactual direction vector, we calculate the directional derivative of the model gradient in this direction.
- We want to enforce that along the counterfactual direction, are model gradients are negative and hence a decrease in the prediction probability

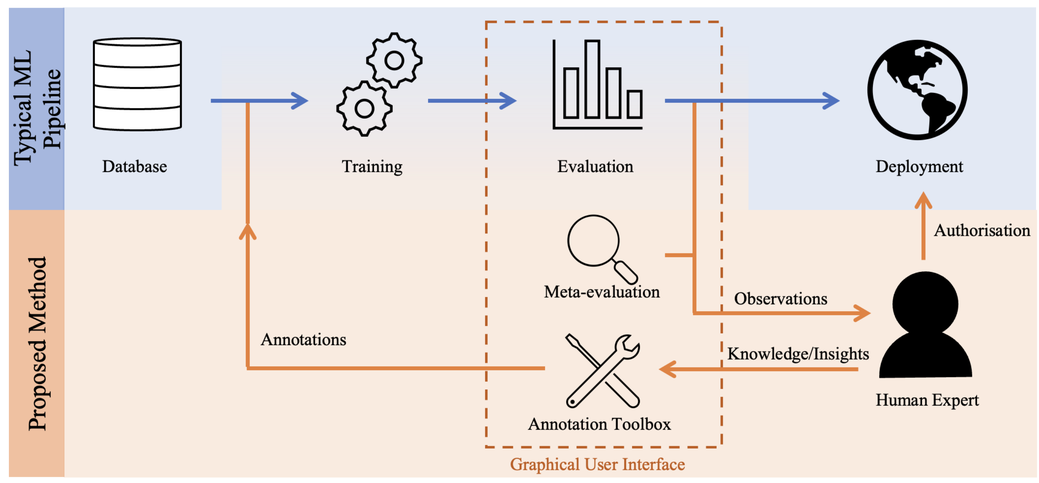
Our proposed loss function is defined as

$$\mathscr{L} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}\sum_{j=1}^{k}\left|\text{sign}(\nabla_{\mathbf{d}_{i,j}} f(\mathbf{x}_i)) + 1\right| \quad (1)$$

where $\nabla_{\mathbf{d}_i}$ is the gradient of the model in the direction of $\mathbf{d}_i$ and $f(\mathbf{x}_i)$ is the prediction for input $\mathbf{x}_i$ and N is the number of examples in the training set. The *sign* function is approximated with a steep *tanh*. This loss increases when the gradient of the model and the counterfactual direction is not aligned.

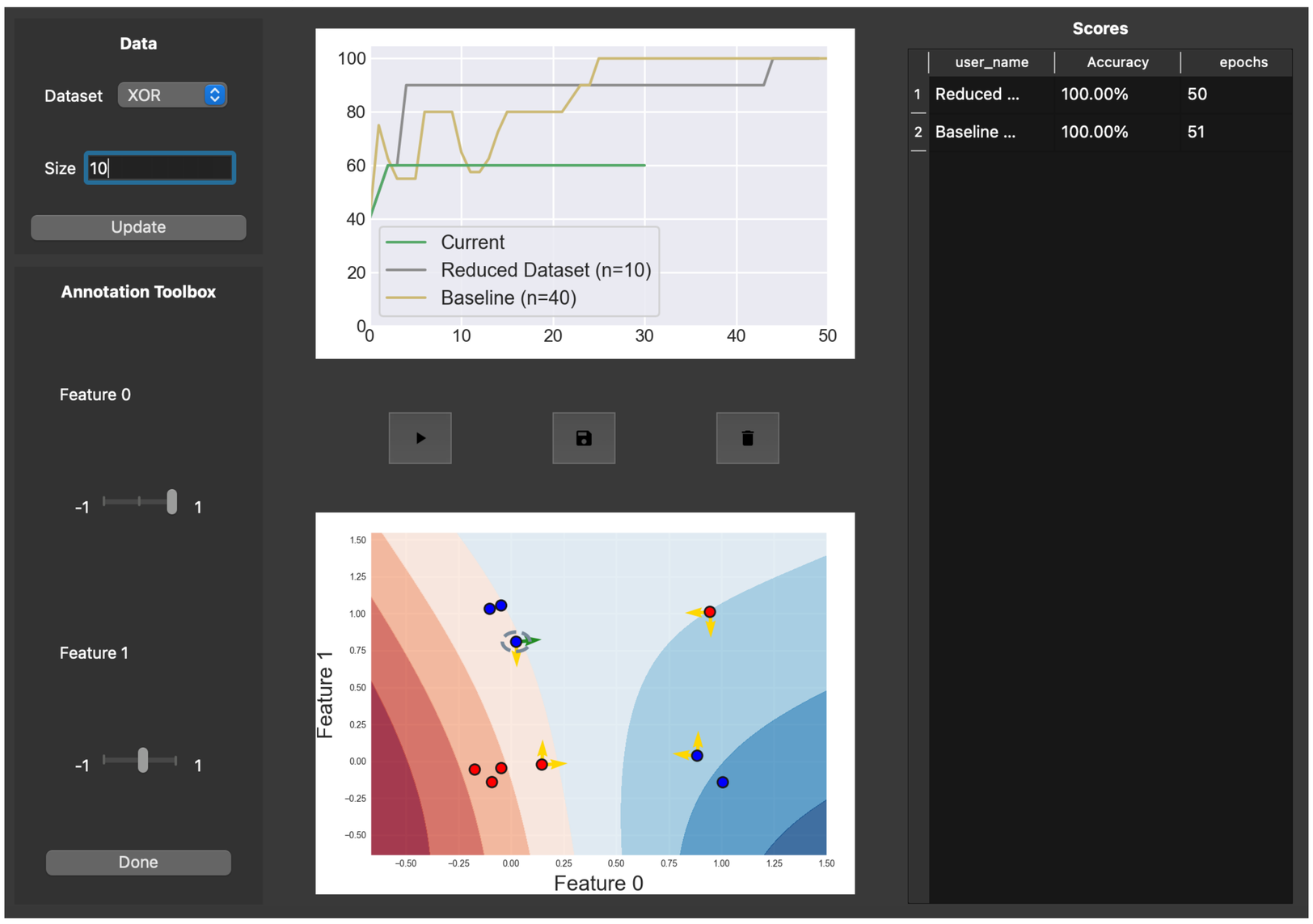
5. Pipeline

We include the meta-evaluations and labelling of complex annotations during the optimisation procedure.



6. Graphical User Interface

Users can select training points from the *meta-evaluation* and provide the counterfactual direction.



7. Future Work

- Extending this solution to more dimensions is important. We hope to use dimensionality reduction techniques or a subset of features in order to present meta-evaluations and gather the complex annotations.
- Potential applications include healthcare, where a clinician can provide directions in which they think the patient would deteriorate. We also aim to apply this to image processing and natural language processing, as counterfactuals have been proven to be useful in these scenarios.

References

[1] D. Kaushik *et al.*, "Learning the difference that makes a difference with counterfactually-augmented data," in *ICLR*, 2020.

[2] D. Teney *et al.*, "Learning what makes a difference from counterfactual examples and gradient supervision," in *ECCV*, pp. 580–599, Springer, 2020.

This work partially funded by Thales Training & Simulation Ltd, UKRI CDT in Interactive AI EP/S022937/1 and UKRI Turing AI Fellowship EP/V024817/1.