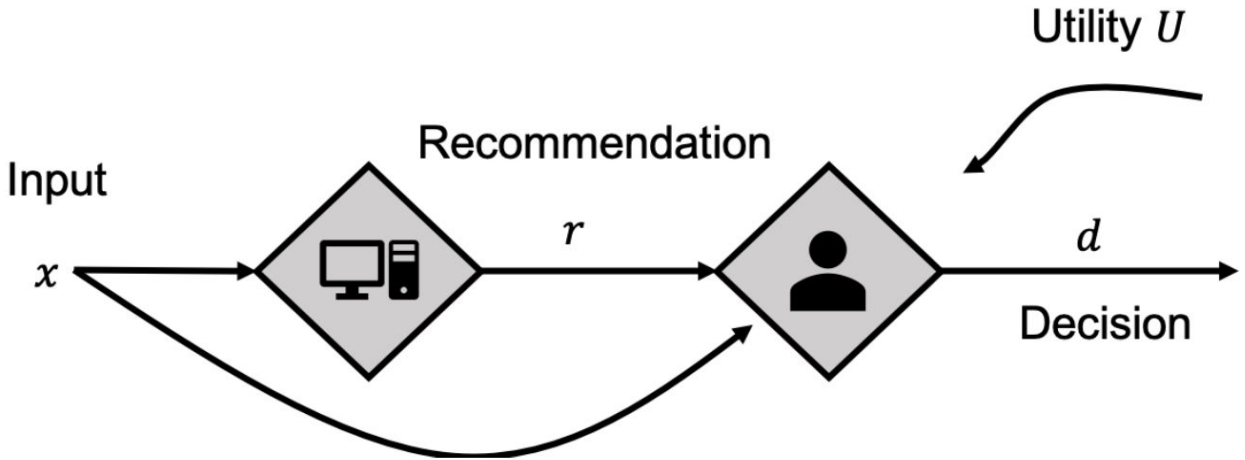
Human-AI Teams

Saumik Narayanan, Tatsuro Murakami, Will Wick

Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork

Gagan Bansal[1] **Besmira Nushi**[2] **Ece Kamar**[2] **Eric Horvitz**[2] **Daniel S. Weld**[1,3]
[1]University of Washington [2]Microsoft Research [3]Allen Institute for AI

Human Makes Final Decision



1. Human either accepts the recommendation or solves the task themselves

1. Human either accepts the recommendation or solves the task themselves
2. Human is more accurate than the AI

1. Human either accepts the recommendation or solves the task themselves
2. Human is more accurate than the AI
3. Solving the task is costly for humans

1. Human either accepts the recommendation or solves the task themselves
2. Human is more accurate than the AI
3. Solving the task is costly for humans
4. Human always accepts recommendation if confidence score exceeds some minimum threshold

Meta-decision/Decision	Correct	Incorrect
`Accept[A]`	1	$-\beta$
`Solve[S]`	$1-\lambda$	$-\beta-\lambda$

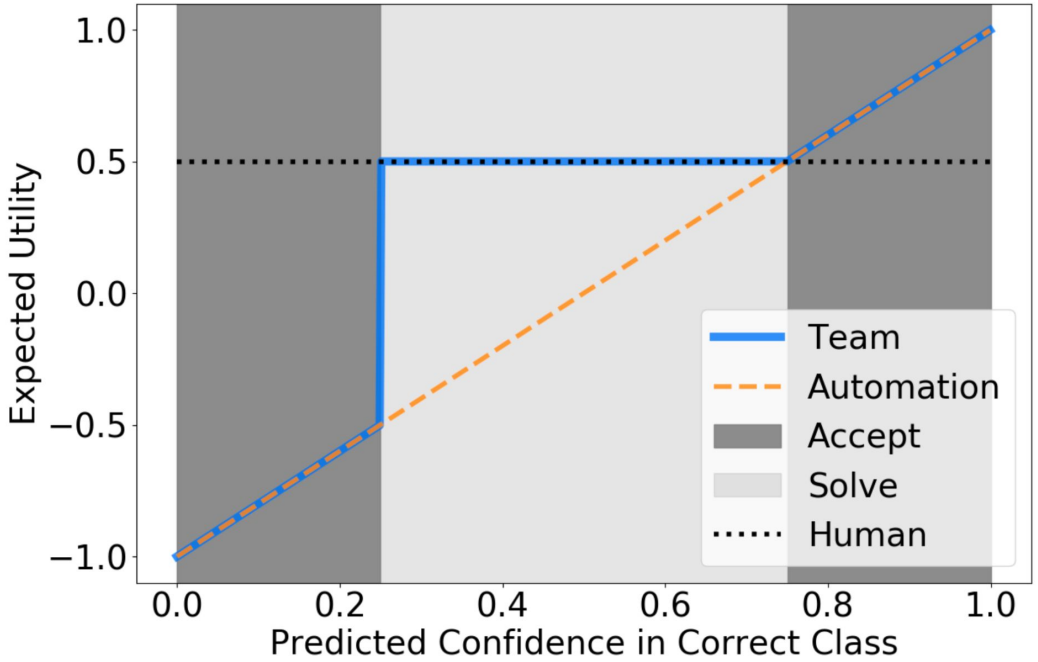
β - Cost for incorrect answer

λ - Cost to query human

$$\psi(x, y) = \begin{cases} (1+\beta)\cdot h(x)[y] - \beta & \text{if } h(x)[\hat{y}] \geq c(\beta, \lambda, a) \\ (1+\beta)\cdot a - \beta - \lambda & \text{otherwise} \end{cases}$$

Can we train a classifier with higher utility than the most accurate classifier?

Expected Utility Flat in Solve Region



Experiment Design

- 6 datasets
- 50 train/test splits
- Logistic regression (linear) and multi-layered perceptron (MLP)
- Pretrain on log-loss
- Measure changes in accuracy, expected utility, and empirical utility

Optimizing for expected utility generally results in increased expected utility

		Logloss			Expected Utility Loss		
Classifier	**Dataset**	**Accuracy**	**Expected Util.**	**Emp. Util.**	**Δ Accuracy**	**Δ Expected Util.**	**Δ Emp. Util.**
Linear	Fico	0.729	0.487	0.575	-0.247	**0.013**	-0.075
	German	0.754	0.529	0.594	-0.015	0	-0.019
	MIMIC	0.881	0.694	0.8	-0.004	**0.066**	-0.035
	Moons	0.885	0.687	0.79	-0.02	**0.079**	-0.006
	recidivism	0.669	0.485	0.52	-0.17	**0.015**	-0.02
	Scenario1	0.858	0.524	0.593	-0.165	**0.102**	**0.061**
MLP	Fico	0.725	0.472	0.574	-0.244	**0.028**	-0.074
	German	0.752	0.53	0.618	-0.036	-0.027	-0.056
	MIMIC	0.881	0.719	0.799	-0.001	**0.049**	-0.029
	Moons	1	0.944	0.989	0	**0.049**	**0.006**
	Recidivism	0.674	0.467	0.521	-0.168	**0.033**	-0.021
	Scenario1	1	0.826	0.854	-0.1	**0.08**	**0.057**

Optimizing for expected utility generally results in increased expected utility

		Logloss			Expected Utility Loss		
Classifier	Dataset	Accuracy	Expected Util.	Emp. Util.	Δ Accuracy	Δ Expected Util.	Δ Emp. Util.
Linear	Fico	0.729	0.487	0.575	-0.247	**0.013**	-0.075
	German	0.754	0.529	0.594	-0.015	0	-0.019
	MIMIC	0.881	0.694	0.8	-0.004	**0.066**	-0.035
	Moons	0.885	0.687	0.79	-0.02	**0.079**	-0.006
	recidivism	0.669	0.485	0.52	-0.17	**0.015**	-0.02
	Scenario1	0.858	0.524	0.593	-0.165	**0.102**	**0.061**
MLP	Fico	0.725	0.472	0.574	-0.244	**0.028**	-0.074
	German	0.752	0.53	0.618	-0.036	-0.027	-0.056
	MIMIC	0.881	0.719	0.799	-0.001	**0.049**	-0.029
	Moons	1	0.944	0.989	0	**0.049**	**0.006**
	Recidivism	0.674	0.467	0.521	-0.168	**0.033**	-0.021
	Scenario1	1	0.826	0.854	-0.1	**0.08**	**0.057**

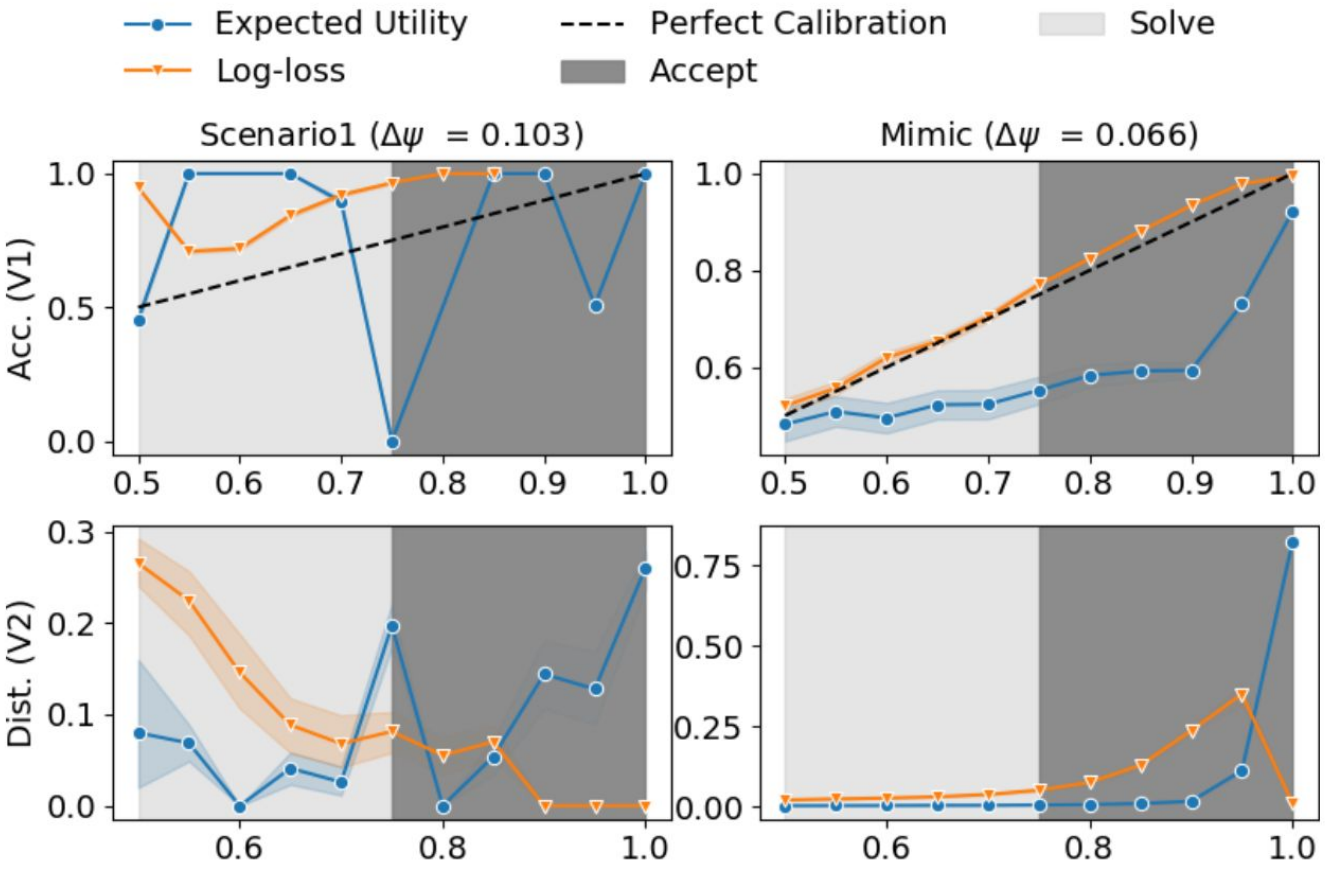
Expected Utility and Empirical Utility Fundamentally Different

Dataset	Expected Util LL	Emp. Util LL	ΔExpected Util (A)	Δ Emp. Util (B)	Δ* Emp. Util (C)
Fico-2d	0.475	0.511	0.025	-0.011	-0.004
German-2d	0.514	0.6	0.076	-0.004	-0.016
MIMIC-2d	0.641	0.772	0.121	-0.009	0.005
Moons	0.767	0.813	0.016	-0.006	0.034
Recidivism-2d	0.478	0.518	0.022	-0.017	0.007
Scenario1	0.707	0.715	0.045	0.069	0.068

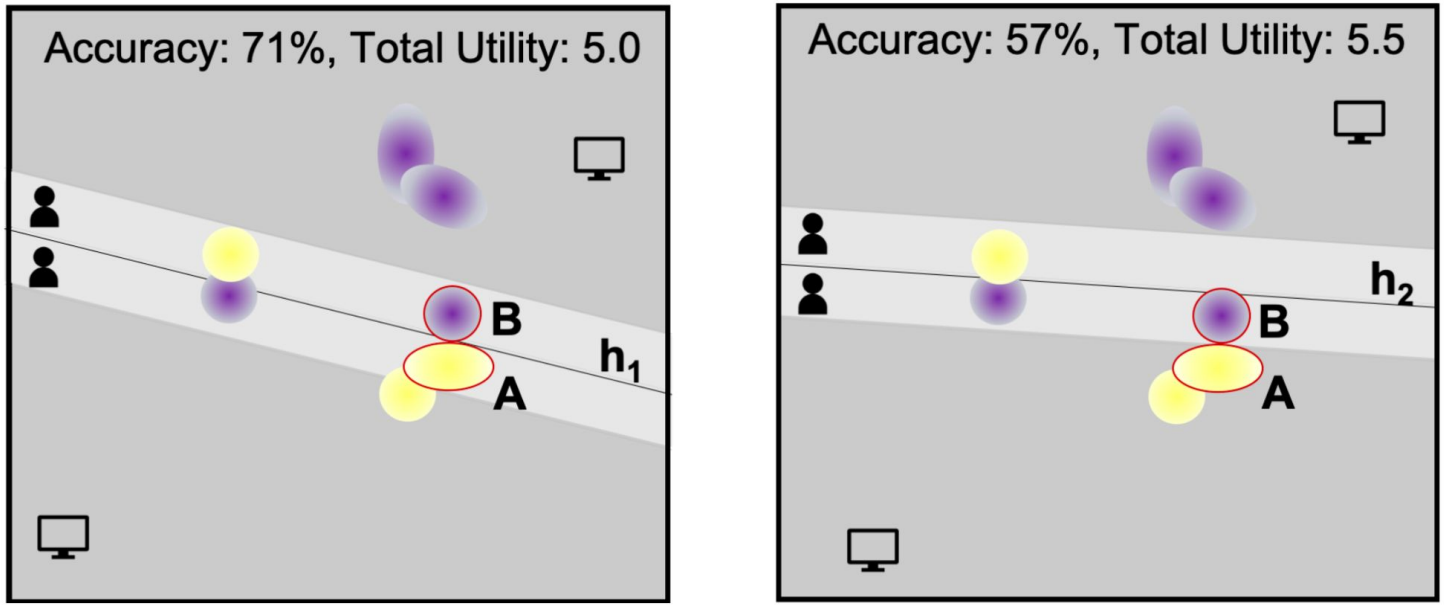
Expected Utility and Empirical Utility Fundamentally Different

Dataset	Expected Util LL	Emp. Util LL	ΔExpected Util (A)	Δ Emp. Util (B)	Δ* Emp. Util (C)
Fico-2d	0.475	0.511	0.025	-0.011	-0.004
German-2d	0.514	0.6	0.076	-0.004	-0.016
MIMIC-2d	0.641	0.772	0.121	-0.009	0.005
Moons	0.767	0.813	0.016	-0.006	0.034
Recidivism-2d	0.478	0.518	0.022	-0.017	0.007
Scenario1	0.707	0.715	0.045	0.069	0.068

How does the new model qualitatively differ from the most accurate model?


Expected Utility
Log-loss
Perfect Calibration
Accept
Solve
Scenario1 ($\Delta\psi$ = 0.103)
Mimic ($\Delta\psi$ = 0.066)
Acc. (V1)
Dist. (V2)
1.0
0.5
0.0
0.8
0.6
0.3
0.2
0.1
0.75
0.50
0.25
0.00
0.5
0.6
0.7
0.8
0.9
1.0

Machine willing to misclassify examples that it knows that human will solve anyway



Discussion Questions

What are potential issues with optimizing for something other than accuracy?

Does the model for human-AI interactions and team utility make sense? In what settings would you see it as applicable vs. not applicable?

Models trained on the expected utility loss function exhibited inaccurate confidence scores. Is this good or bad?

How do the properties of the task affect improvements in utility?

Increasing human accuracy decreases gains from optimizing for expected utility

dataset	a=0.8	a=0.9	a=1
Fico	0.257 (0.133)	0.337 (0.071)	0.487 (0.013)
German	0.397 (0.046)	0.444 (0.035)	0.529 (0)
MIMIC	0.625 (0.127)	0.644 (0.111)	0.694 (0.066)
Moons	0.582 (0.162)	0.616 (0.139)	0.687 (0.079)
Recidivism	0.155 (0.073)	0.292 (0)	0.485 (0.015)
Scenario1	0.224 (0.324)	0.364 (0.248)	0.524 (0.102)

Increasing human accuracy decreases gains from optimizing for expected utility

dataset	a=0.8	a=0.9	a=1
Fico	0.257 (0.133)	0.337 (0.071)	0.487 (0.013)
German	0.397 (0.046)	0.444 (0.035)	0.529 (0)
MIMIC	0.625 (0.127)	0.644 (0.111)	0.694 (0.066)
Moons	0.582 (0.162)	0.616 (0.139)	0.687 (0.079)
Recidivism	0.155 (0.073)	0.292 (0)	0.485 (0.015)
Scenario1	0.224 (0.324)	0.364 (0.248)	0.524 (0.102)

No conclusive trend from increasing cost of mistakes

dataset	β**=1**	β**=3**	β**=5**
Fico	0.487 (0.013)	0.474 (0.026)	0.481 (0.019)
German	0.529 (0)	0.427 (0.057)	0.367 (0.118)
MIMIC	0.694 (0.066)	0.58 (0.008)	0.543 (0)
Moons	0.687 (0.079)	0.637 (0.065)	0.594 (0.085)
Recidivism	0.485 (0.015)	0.495 (0.004)	0.498 (0.001)
Scenario1	0.524 (0.102)	0.501 (0.02)	0.5 (0)

No conclusive trend from increasing cost of mistakes

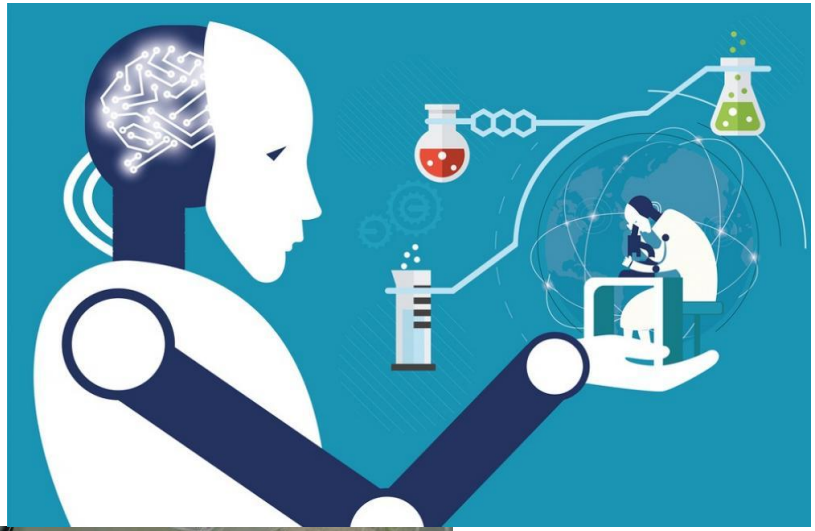
dataset	β=1	β=3	β=5
Fico	0.487 (0.013)	0.474 (0.026)	0.481 (0.019)
German	0.529 (0)	0.427 (0.057)	0.367 (0.118)
MIMIC	0.694 (0.066)	0.58 (0.008)	0.543 (0)
Moons	0.687 (0.079)	0.637 (0.065)	0.594 (0.085)
Recidivism	0.485 (0.015)	0.495 (0.004)	0.498 (0.001)
Scenario1	0.524 (0.102)	0.501 (0.02)	0.5 (0)

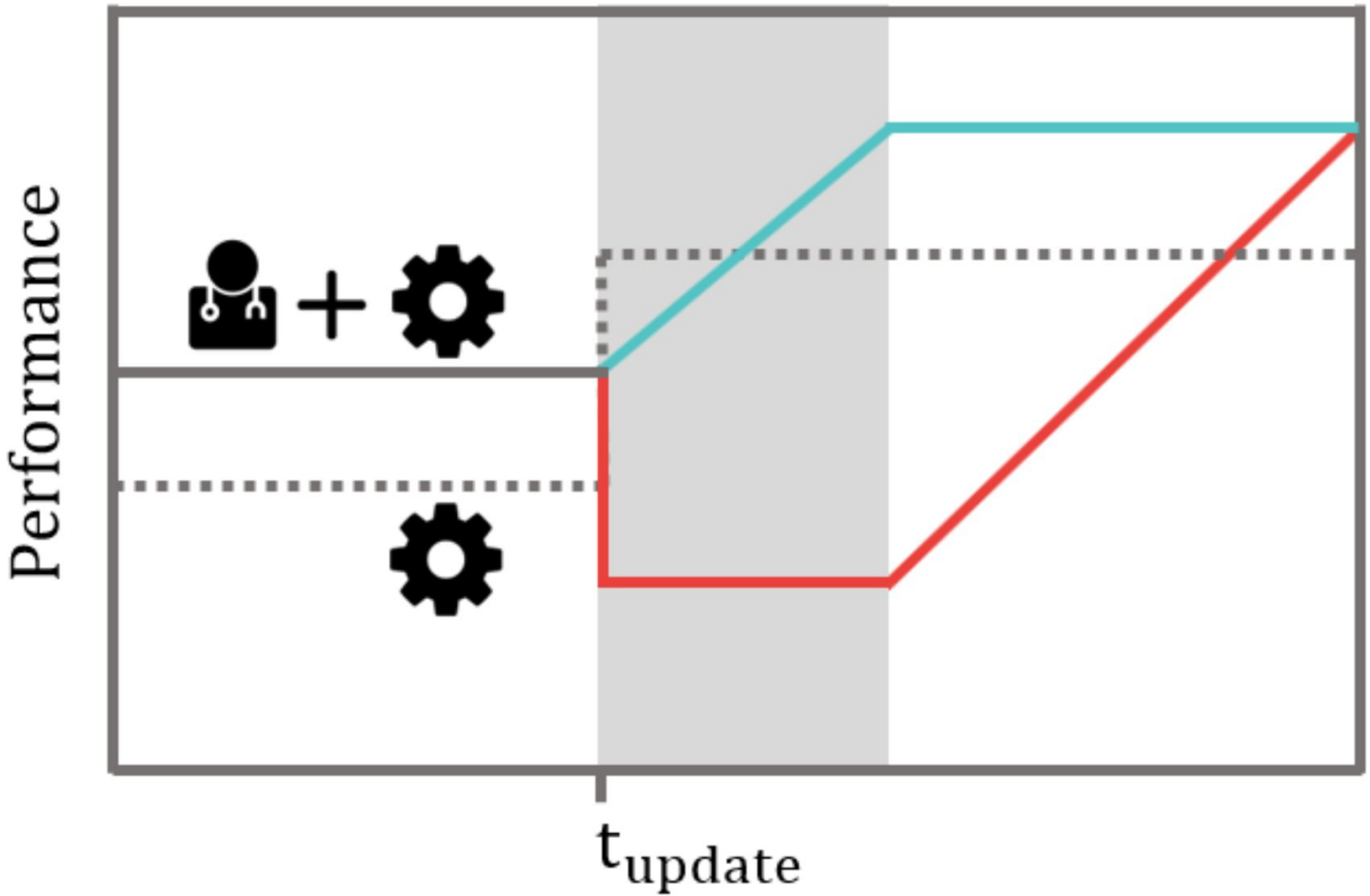
Summary

- Human makes final decision
- Increased team utility comes at the cost of lower accuracy
- Recommendations over-confident, greater number of high-confidence recommendations
- Gains diminish with increased human accuracy

Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff

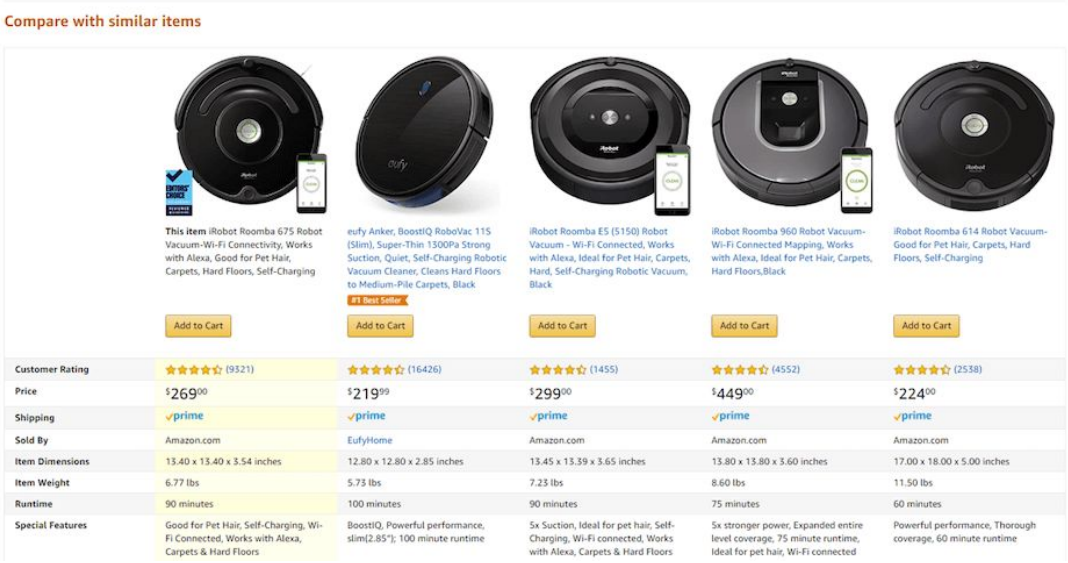
Gagan Bansal,[1] Besmira Nushi,[2] Ece Kamar,[2] Daniel S. Weld,[1] Walter S. Lasecki,[3] Eric Horvitz[2]
[1]University of Washington, [2]Microsoft Research, [3]University of Michigan




Performance
t_{update}

What is an update?

- The model is trained on additional data
- With new data, the model usually has better overall performance
- Earlier classifications might be changed though

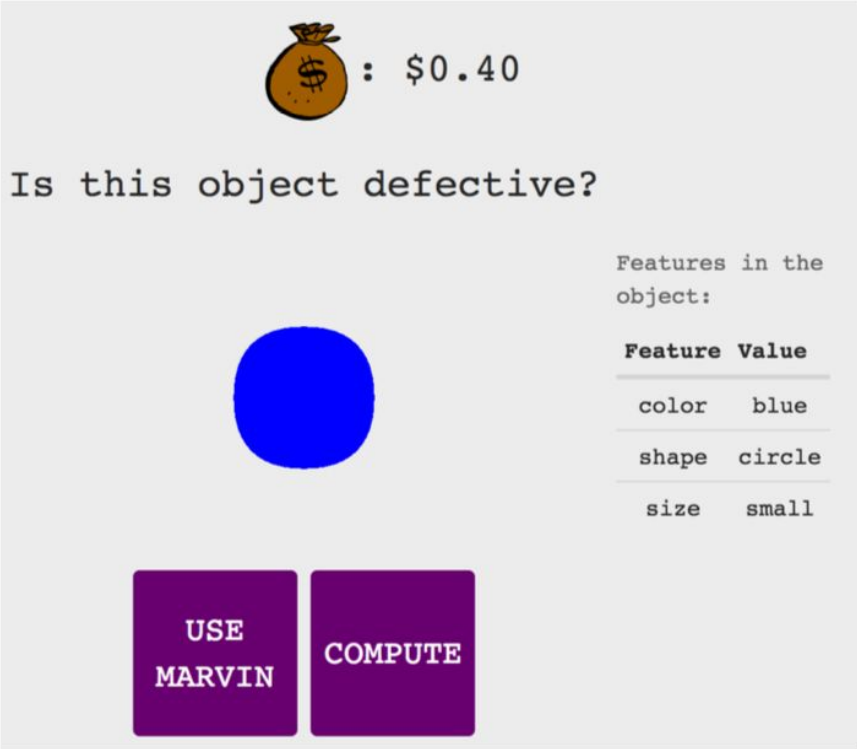


Discussion Questions

The paper claims that we sometimes change the way we interact with an AI as the AI is changed. Have you experienced such behavior yourself?

The authors also show that updating an AI with increased performance does not necessarily increase the performance of humans and may in fact hurt the performance. Can you come up with a real world example of such an instance?

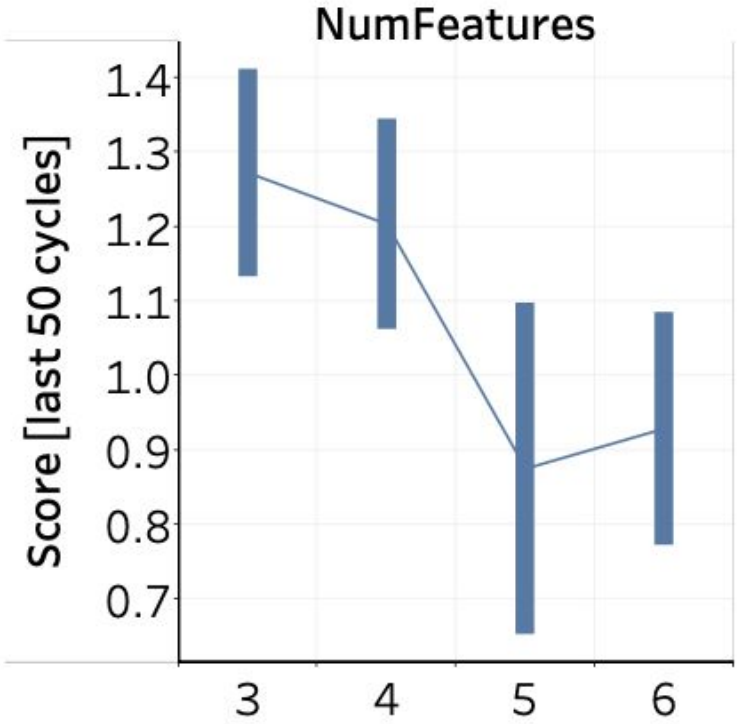
CAJA Platform



1. A new object appears
2. AI recommends if it's defective
3. Accept AI recommendation or compute
4. The reward is given as follows

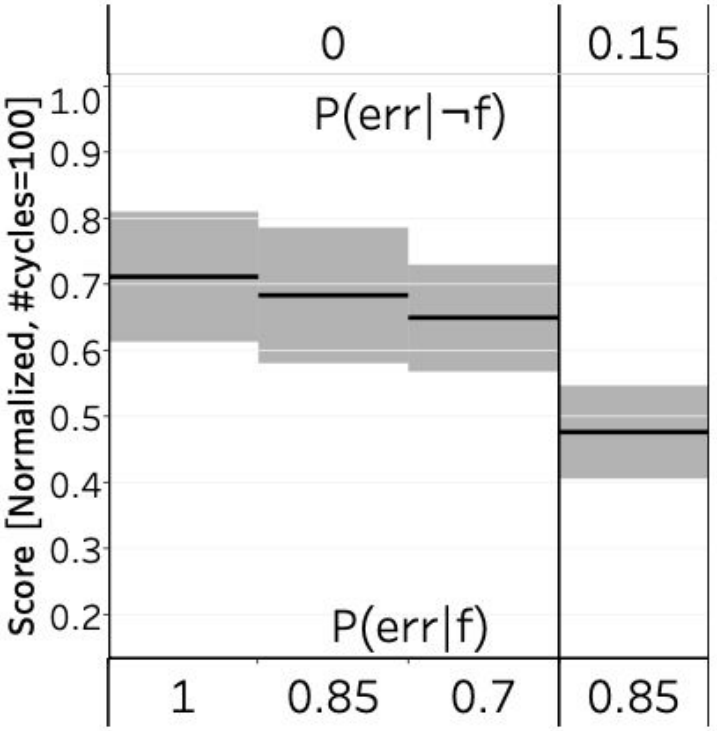
	Accept	Compute
AI right	$0.04	0
AI wrong	-$0.16	0

Do better mental models of AI lead to higher team performance?



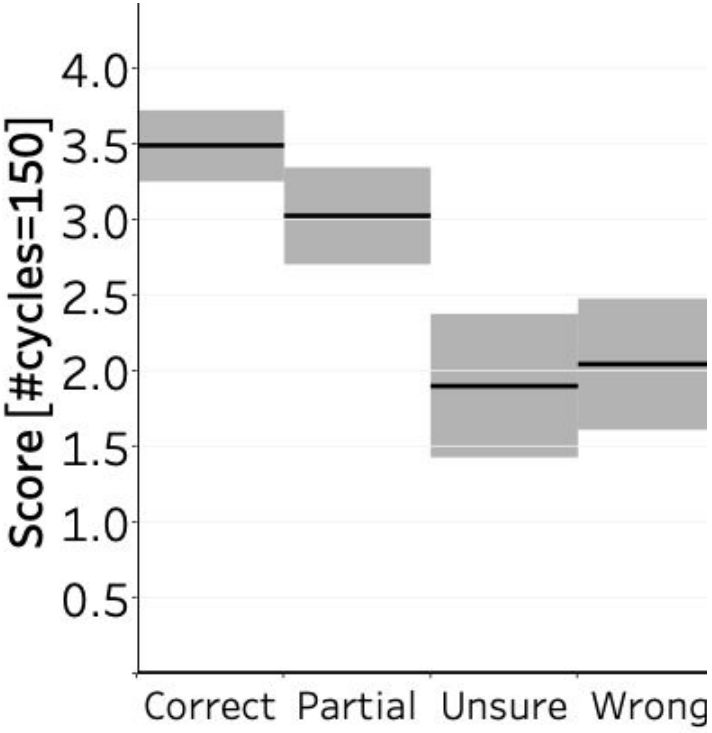
Team performance decrease as the number of features increases

Do better mental models of AI lead to higher team performance?



Team performance decrease as the stochasticity of a model increases

Do better mental models of AI lead to higher team performance?



Better mental models result in higher team performance

How do we measure if updates are human friendly?

How do we measure if updates are human friendly?

We define **Compatibility Score** of two models

How do we measure if updates are human friendly?

We define **Compatibility Score** of two models

$$\mathcal{C}(h_1, h_2) = \frac{\sum_x A(x, h_1(x)) \cdot A(x, h_2(x))}{\sum_x A(x, h_1(x))}$$

Whether $h_1(x)$ is the correct action for x.
If so, returns 1, and 0 otherwise

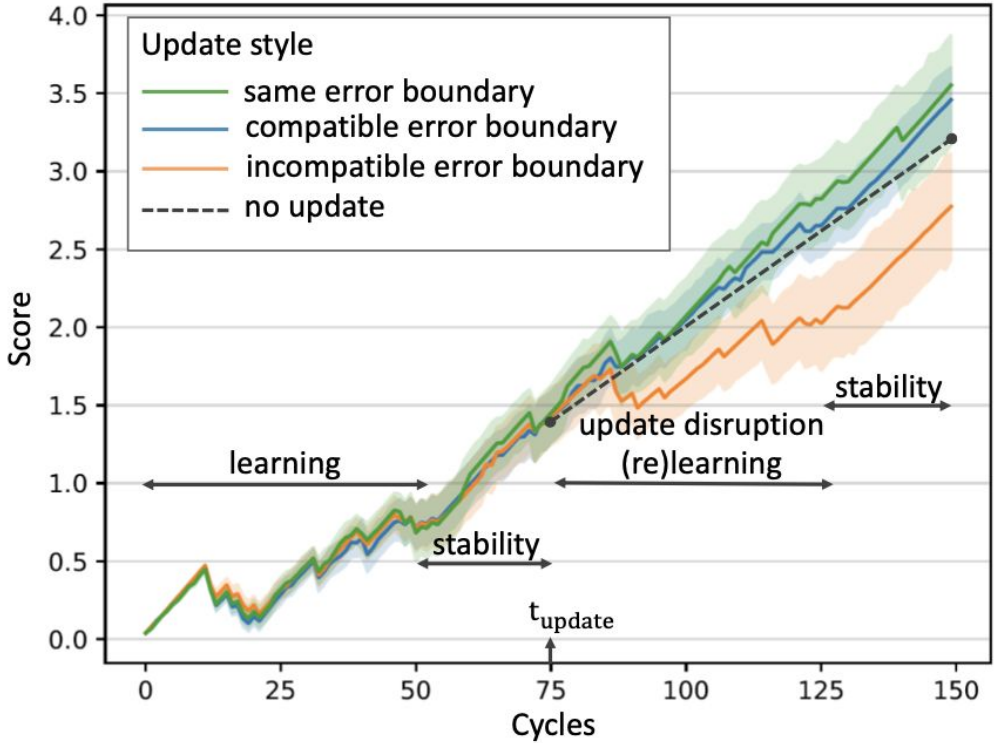
Old model

$$C(h_1, h_2) = \frac{\sum_x A(x, h_1(x)) \cdot A(x, h_2(x))}{\sum_x A(x, h_1(x))}$$

Updated model

Prediction made by model h_1 given the input x

Do more compatible updates lead to higher team performance than incompatible updates?



Compatible error:

AI makes mistake when blue∩square

After update, small∩blue∩square

←Mistake instance is in subset of original

Incompatible error:

AI makes mistake blue∩square

After update, red∩round

Do current ML classifiers produce compatible updates?

Trained on 200 samples

Trained on 5000 samples

How much h_2 honors h_1's predictions

Classifier	Dataset	ROC h_1	ROC h_2	$\mathcal{C}(h_1, h_2)$
LR	Recidivism	0.68	0.72	0.72
	Credit Risk	0.72	0.77	0.66
	Mortality	0.68	0.77	**0.40**
MLP	Recidivism	0.59	0.73	0.53
	Credit Risk	0.70	0.80	0.63
	Mortality	0.71	0.84	0.76

Do current ML classifiers produce compatible updates?

Trained on 200 samples

Trained on 5000 samples

How much h_2 honors h_1's predictions

Classifier	Dataset	ROC h_1	ROC h_2	$\mathcal{C}(h_1, h_2)$
LR	Recidivism	0.68	0.72	0.72
	Credit Risk	0.72	0.77	0.66
	Mortality	0.68	0.77	**0.40**
MLP	Recidivism	0.59	0.73	0.53
	Credit Risk	0.70	0.80	0.63
	Mortality	0.71	0.84	0.76

Do current ML classifiers produce compatible updates?

Trained on 200 samples

Trained on 5000 samples

How much h_2 honors h_1's predictions

Classifier	Dataset	ROC h_1	ROC h_2	$\mathcal{C}(h_1, h_2)$
LR	Recidivism	0.68	0.72	0.72
	Credit Risk	0.72	0.77	0.66
	Mortality	0.68	0.77	**0.40**
MLP	Recidivism	0.59	0.73	0.53
	Credit Risk	0.70	0.80	0.63
	Mortality	0.71	0.84	0.76

Do current ML classifiers produce compatible updates?

Trained on 200 samples

Trained on 5000 samples

How much h_2 honors h_1's predictions

Classifier	Dataset	ROC h_1	ROC h_2	$\mathcal{C}(h_1, h_2)$
LR	Recidivism	0.68	0.72	0.72
	Credit Risk	0.72	0.77	0.66
	Mortality	0.68	0.77	**0.40**
MLP	Recidivism	0.59	0.73	0.53
	Credit Risk	0.70	0.80	0.63
	Mortality	0.71	0.84	0.76

How do we train models with high compatibility score?

How do we train models with high compatibility score?

We penalize models with low compatibility score

How do we train models with high compatibility score?

We penalize models with low compatibility score

Dissonance (opposite of compatibility score)

$$\mathcal{D}(x, y, h_1, h_2) = \mathbb{1}(h_1(x) = y) \cdot L(x, y, h_2)$$

Indicator function: returns 1 if h_1's prediction for input x is correct

Log loss function

$$\mathcal{D}(x, y, h_1, h_2) = \mathbb{1}(h_1(x) = y) \cdot L(x, y, h_2)$$

How do we use dissonance to train models?

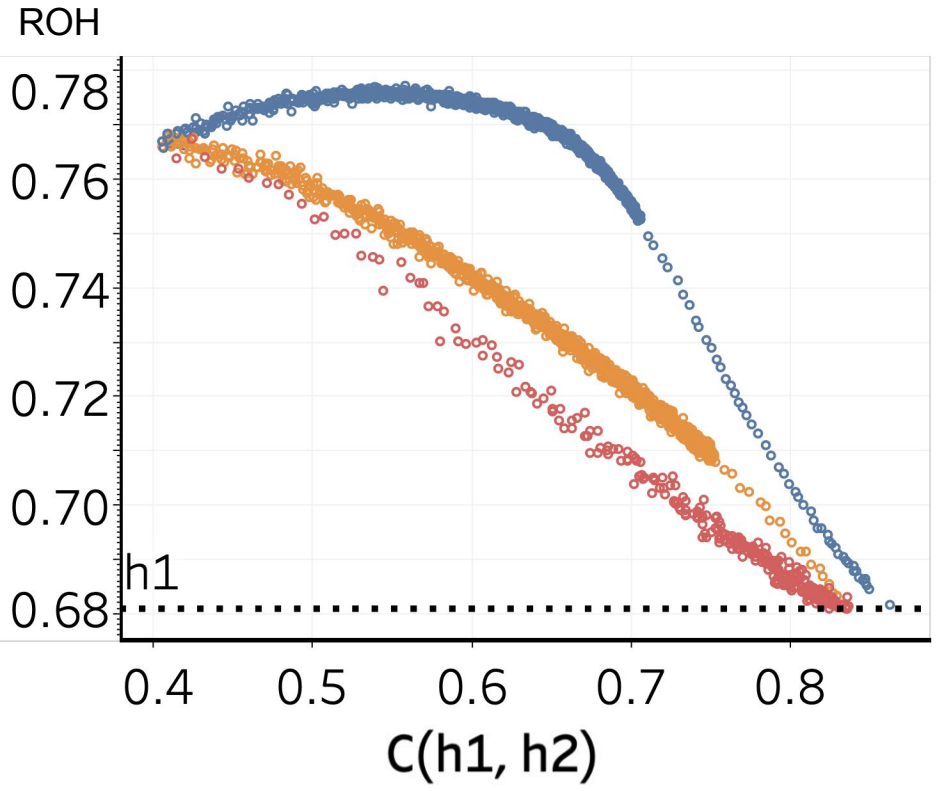
We can add it to tradition log loss function

$$L(x, y, h_2) = y \cdot \log p(h_2(x)) + (1 - y) \cdot \log(1 - p(h_2(x)))$$

$$L_c = L + \lambda_c \cdot \mathcal{D}$$

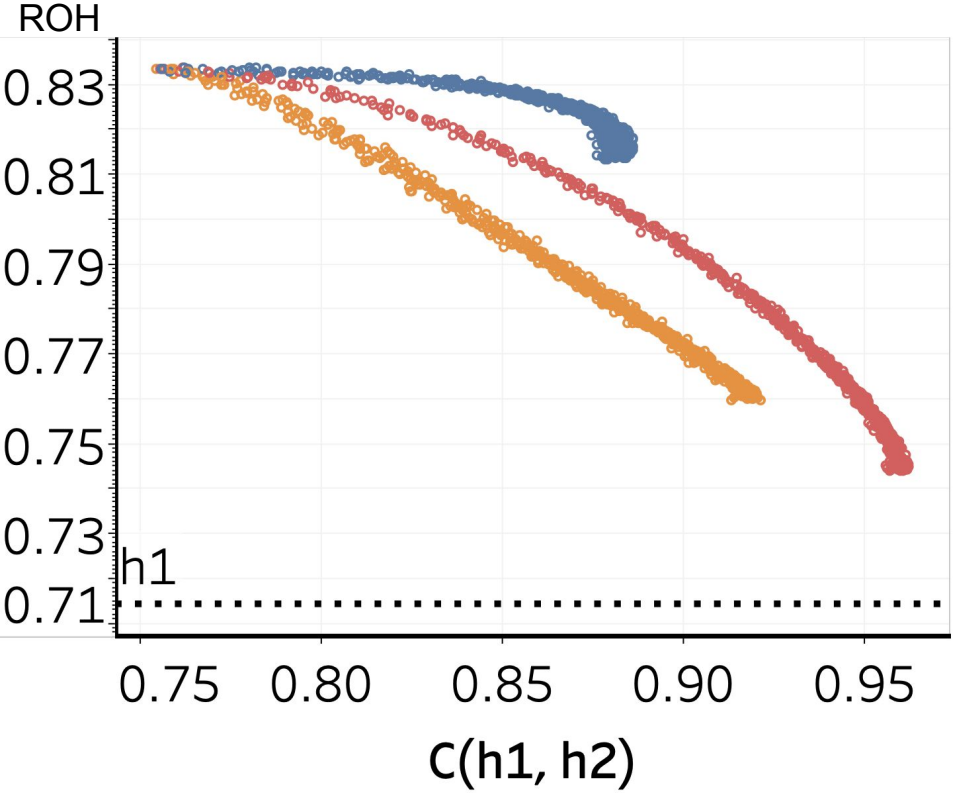
Dissonance

Does there exist a tradeoff between the performance and the compatibility of an update to AI?



The tradeoff between performance and compatibility of Linear Regression on Mortality dataset

Does there exist a tradeoff between the performance and the compatibility of an update to AI?



The tradeoff between performance and compatibility of MLP on Mortality dataset

Conclusion

- Trade off between performance and compatibility
- There are domains where human’s mental models are irrelevant
- Though with its difficulties, explaining updates to users is helpful to smoothly update mental models
- We need to learn more about how people create and update their mental models

Learning to Complement Humans

Bryan Wilder[1*], **Eric Horvitz**[2] and **Ece Kamar**[2]

[1] School of Engineering and Applied Sciences, Harvard University

[2] Microsoft Research

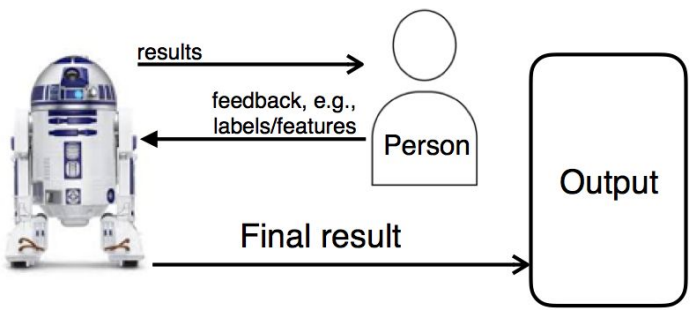
bwilder@g.harvard.edu, horvitz@microsoft.com, eckamar@microsoft.com

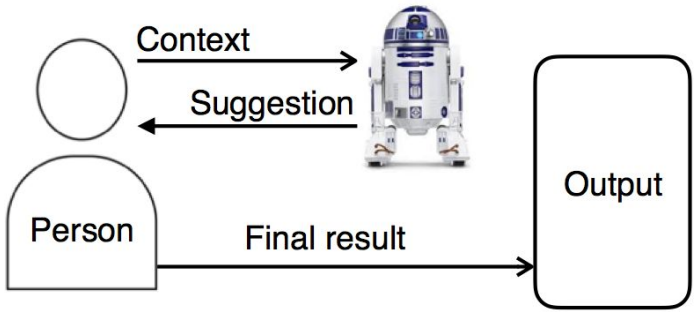
Human AI Teams:

Who makes the final decision?

Human-Centered Machine Learning

- From Human-In-The-Loop to Machine-In-The-Loop

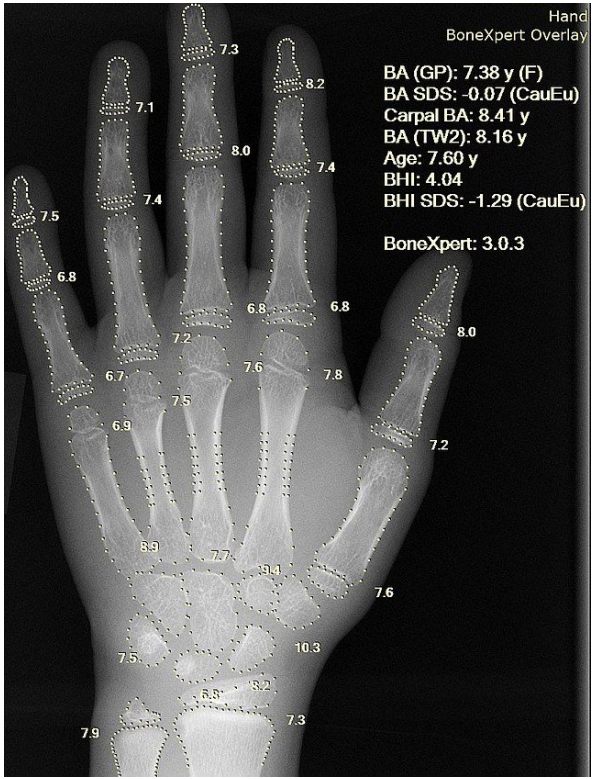




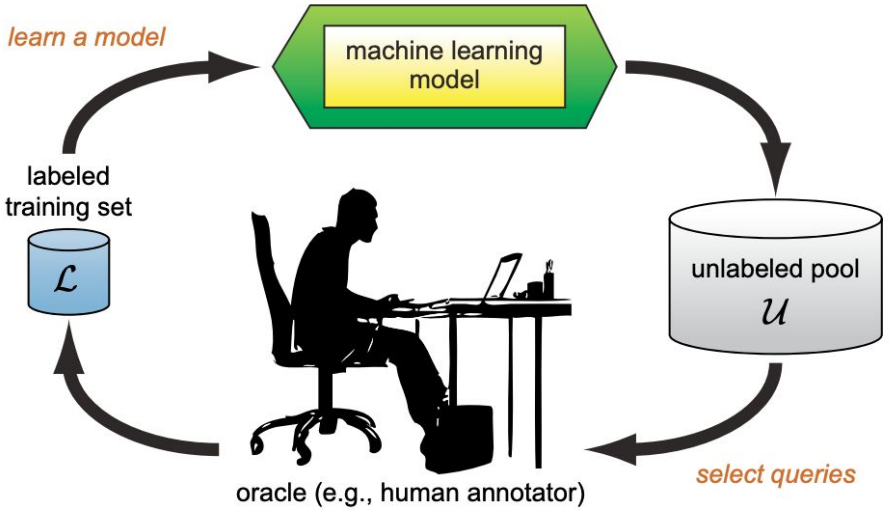
https://medium.com/@ChenhaoTan/human-centered-machine-learning-a-machine-in-the-loop-approach-ed024db34fe7

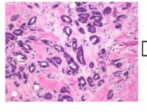
Who makes the final decision?

Human makes the final decision



AI makes the final decision



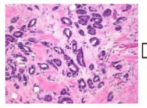


Input

AI Model

Decision to consult human expert (pay cost)

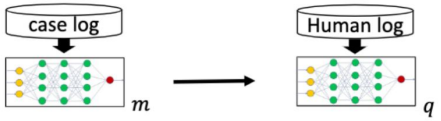
Classification decision



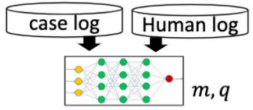
Input

Discriminative approaches

Fixed discriminative: Predictive model trained in isolation, followed by training of policy for querying

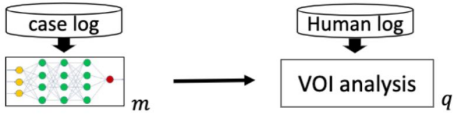


Joint discriminative: Predictive model and policy trained jointly

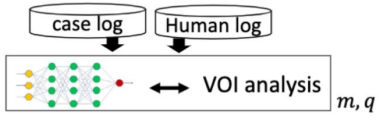


Decision-theoretic approaches

Fixed VOI: Predictive model trained in isolation, followed-up with principled VOI analysis

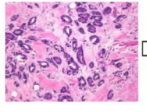


Joint VOI: Predictive model and policy trained jointly, with principled VOI analysis.



Decision to consult human expert (pay cost)

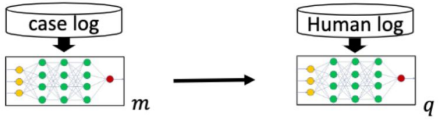
Classification decision



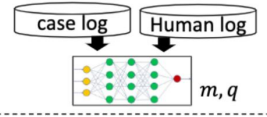
Input

Discriminative approaches

Fixed discriminative: Predictive model trained in isolation, followed by training of policy for querying

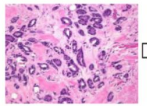


Joint discriminative: Predictive model and policy trained jointly



Decision to consult human expert (pay cost)

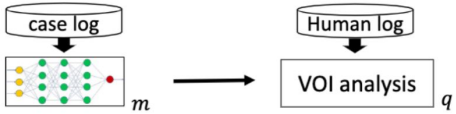
Classification decision



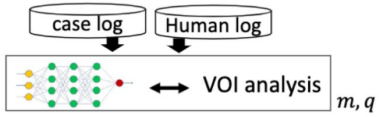
Input

Decision-theoretic approaches

Fixed VOI: Predictive model trained in isolation, followed-up with principled VOI analysis



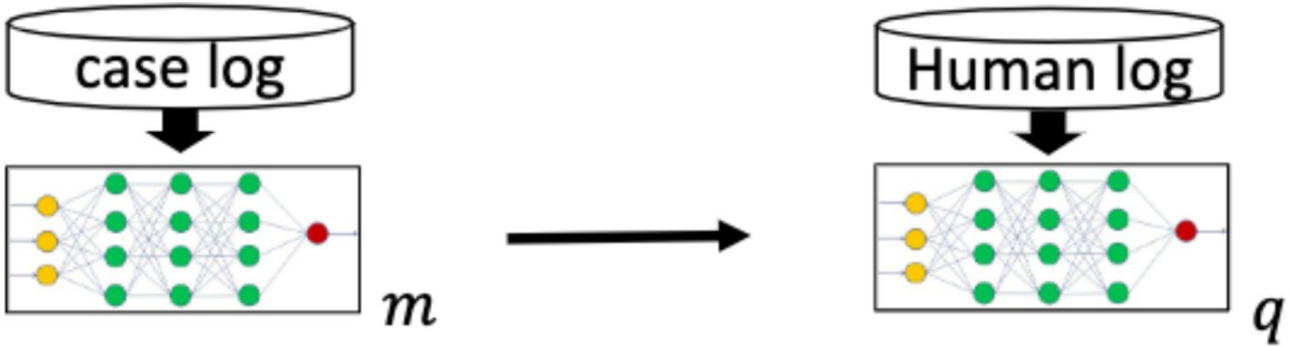
Joint VOI: Predictive model and policy trained jointly, with principled VOI analysis.



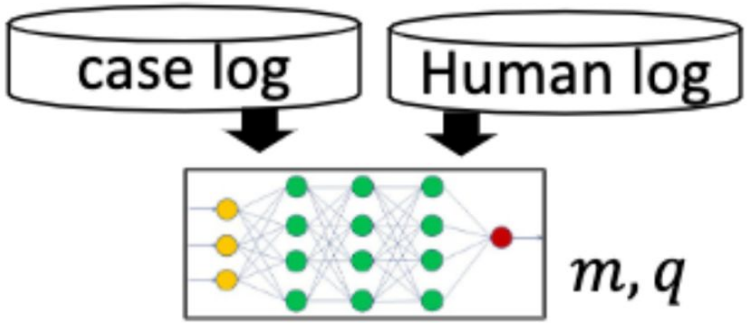
Decision to consult human expert (pay cost)

Classification decision

Fixed discriminative: Predictive model trained in isolation, followed by training of policy for querying

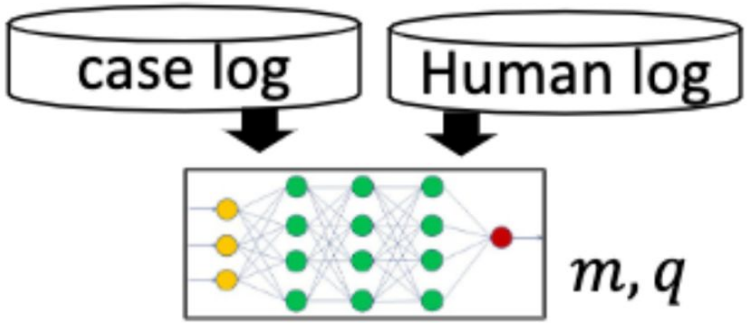


Joint discriminative: Predictive model and policy trained jointly



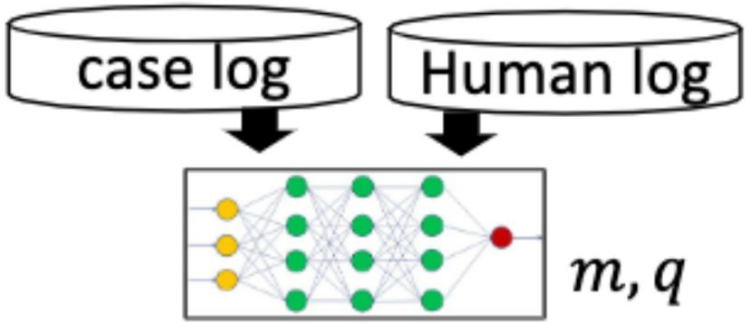
$$\ell\Big(y, q(x)m(x,h) + \big(1-q(x)\big)m(x)\Big) + cq(x)$$

Joint discriminative: Predictive model and policy trained jointly



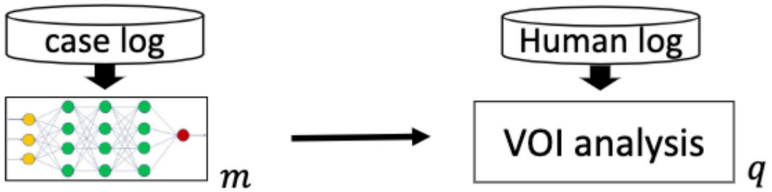
$$\ell\Big(y, q(x)m(x,h) + \big(1 - q(x)\big)m(x)\Big) + cq(x)$$

Joint discriminative: Predictive model and policy trained jointly



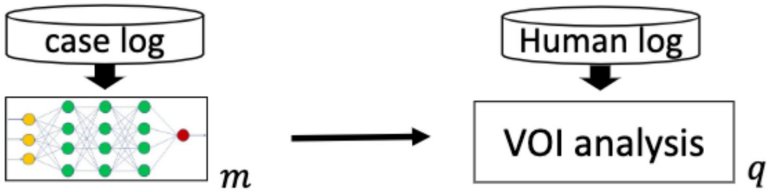
$$\ell\Big(y, q(x)m(x,h) + \big(1 - q(x)\big)m(x)\Big) + cq(x)$$

Fixed VOI: Predictive model trained in isolation, followed-up with principled VOI analysis



$$p_\alpha(y|x)$$
$$p_\beta(h|x)$$
$$p_\gamma(y|h,x)$$

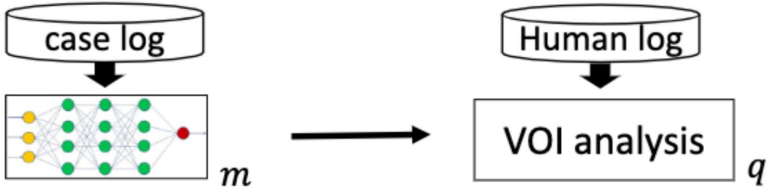
Fixed VOI: Predictive model trained in isolation, followed-up with principled VOI analysis



$$p_\alpha(y|x)$$
$$p_\beta(h|x)$$
$$p_\gamma(y|h,x)$$

$$u_{nq} = \max_{\hat{y} \in Y} \left(\sum_{y \in Y} p_\alpha(y|x) u(\hat{y}, y) \right)$$

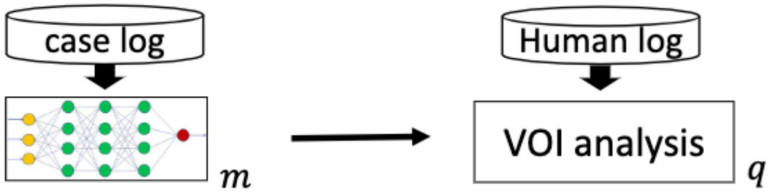
Fixed VOI: Predictive model trained in isolation, followed-up with principled VOI analysis



$$p_\alpha(y|x)$$
$$p_\beta(h|x)$$
$$p_\gamma(y|h,x)$$

$$u_q = \max_{\hat{y} \in Y} \left(\sum_{y \in Y} p_\gamma(y|x,h) u(\hat{y}, y) \right) - c$$

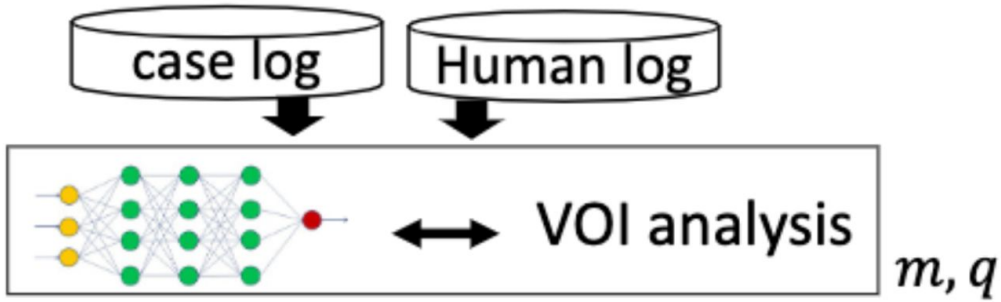
Fixed VOI: Predictive model trained in isolation, followed-up with principled VOI analysis



$$p_\alpha(y|x)$$
$$p_\beta(h|x)$$
$$p_\gamma(y|h,x)$$

$$u_q > u_{nq}$$

Joint VOI: Predictive model and policy trained jointly, with principled VOI analysis.

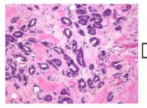


Algorithm 1 Joint VOI training

for T iterations **do**
  Sample a minibatch $B \subseteq [n]$
  for $i \in B$ **do**
    for $\hat{y} \in \mathcal{Y}$ **do**
      $u_{\text{nq}}(\hat{y}) = \sum_{y \in \mathcal{Y}} p_\alpha(y|x_i) u(\hat{y}, y)$
    end for
    $u_{\text{nq}} = \sum_{\hat{y} \in \mathcal{Y}} \frac{u_{\text{nq}}(\hat{y}) \exp(u_{\text{nq}}(\hat{y}))}{\sum_{y' \in \mathcal{Y}} \exp(u_{\text{nq}}(y'))}$
    for $\hat{y} \in \mathcal{Y}$ **do**
      $u_{\text{q}}(\hat{y}, h) = \sum_{y \in \mathcal{Y}} p_\gamma(y|x_i, h) u(\hat{y}, y)$
    end for
    $u_{\text{q}} = \sum_{h \in \mathcal{Y}} p_\beta(h|x) \sum_{\hat{y}} \frac{u_{\text{q}}(\hat{y},h) \exp(u_{\text{q}}(\hat{y},h))}{\sum_{y' \in \mathcal{Y}} \exp(u_{\text{q}}(y',h))}$
    $q = \frac{\exp(u_{\text{q}})}{\exp(u_{\text{q}}) + \exp(u_{\text{nq}})}$
    $\ell^i_{\text{combined}} = \ell(q\, p_\gamma(\cdot|x_i, h_i)$
      $+ (1-q) p_\alpha(\cdot|x_i)) + qc$
  end for
  Backpropagate $\frac{1}{|B|} \sum_{i \in B} \ell^i_{\text{combined}}$
  Every t iterations: update calibrators
end for

Algorithm 1 Joint VOI training

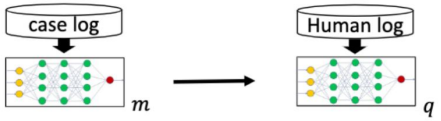
for T iterations **do**
  Sample a minibatch $B \subseteq [n]$
  for $i \in B$ **do**
    for $\hat{y} \in \mathcal{Y}$ **do**
      $u_{\text{nq}}(\hat{y}) = \sum_{y \in \mathcal{Y}} p_\alpha(y|x_i) u(\hat{y}, y)$
    end for
    $u_{\text{nq}} = \sum_{\hat{y} \in \mathcal{Y}} \frac{u_{\text{nq}}(\hat{y}) \exp(u_{\text{nq}}(\hat{y}))}{\sum_{y' \in \mathcal{Y}} \exp(u_{\text{nq}}(y'))}$
    for $\hat{y} \in \mathcal{Y}$ **do**
      $u_{\text{q}}(\hat{y}, h) = \sum_{y \in \mathcal{Y}} p_\gamma(y|x_i, h) u(\hat{y}, y)$
    end for
    $u_{\text{q}} = \sum_{h \in \mathcal{Y}} p_\beta(h|x) \sum_{\hat{y}} \frac{u_{\text{q}}(\hat{y},h) \exp(u_{\text{q}}(\hat{y},h))}{\sum_{y' \in \mathcal{Y}} \exp(u_{\text{q}}(y',h))}$
    $q = \frac{\exp(u_{\text{q}})}{\exp(u_{\text{q}}) + \exp(u_{\text{nq}})}$
    $\ell^i_{\text{combined}} = \ell(q\, p_\gamma(\cdot|x_i, h_i)$
      $+ (1-q) p_\alpha(\cdot|x_i)) + qc$
  end for
  Backpropagate $\frac{1}{|B|} \sum_{i \in B} \ell^i_{\text{combined}}$
  Every t iterations: update calibrators
end for



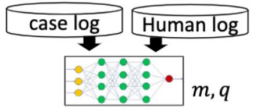
Input

Discriminative approaches

Fixed discriminative: Predictive model trained in isolation, followed by training of policy for querying

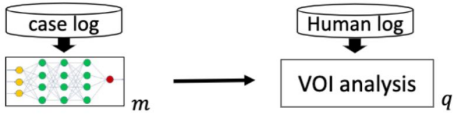


Joint discriminative: Predictive model and policy trained jointly

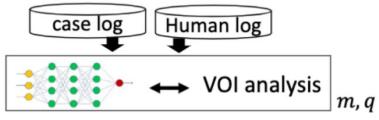


Decision-theoretic approaches

Fixed VOI: Predictive model trained in isolation, followed-up with principled VOI analysis



Joint VOI: Predictive model and policy trained jointly, with principled VOI analysis.

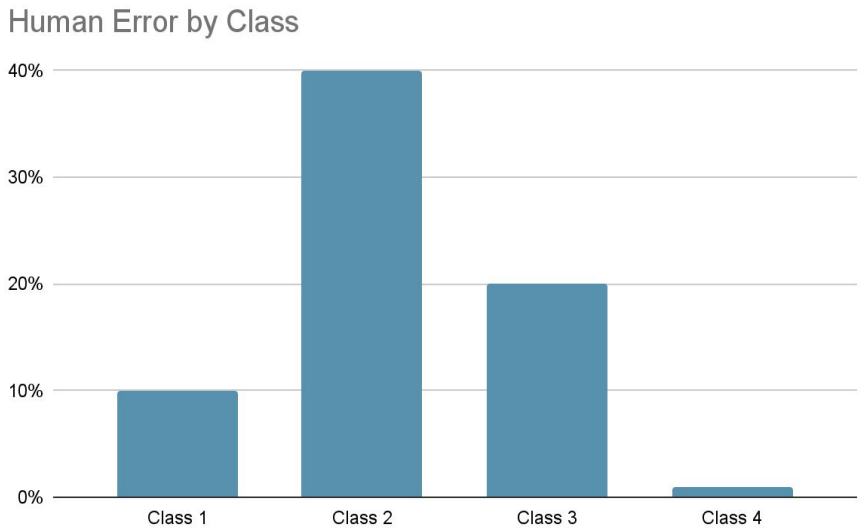


Decision to consult human expert (pay cost)

Classification decision

Discussion Question

What possible benefits could arise from joint training?

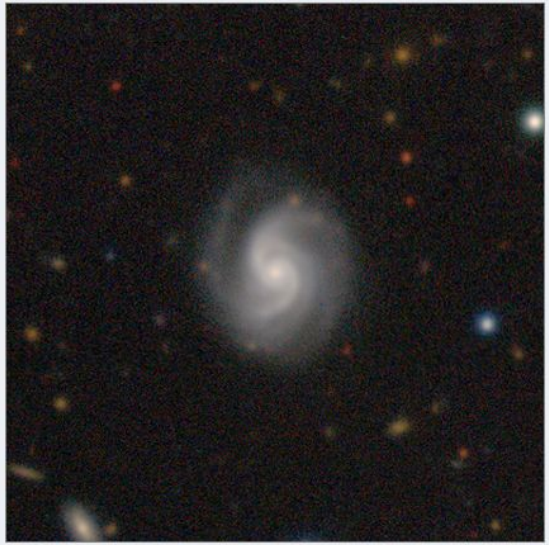


Experiments

- Galaxy Zoo

- Breast Cancer

Welcome to the new Galaxy Zoo!



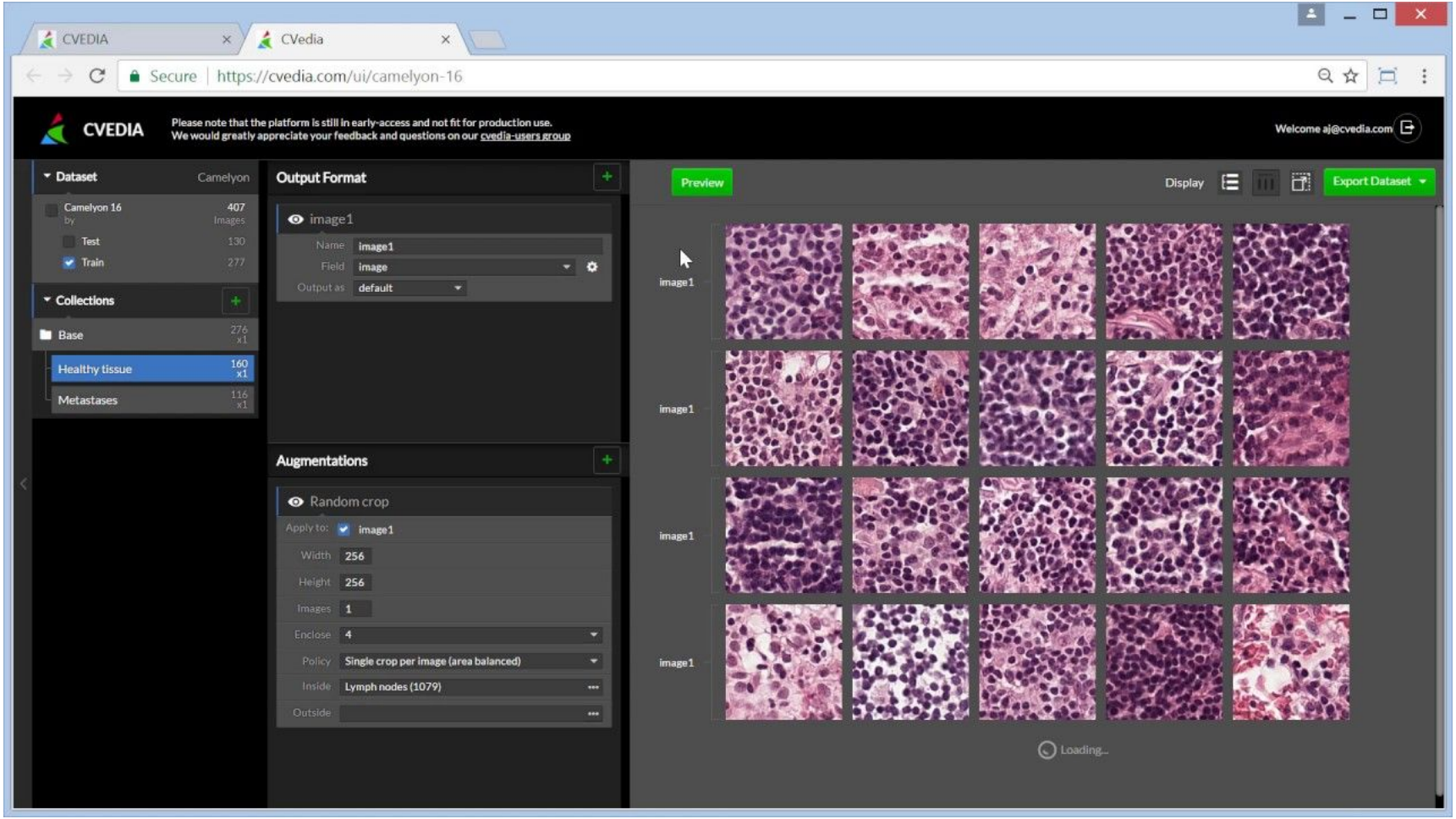
Is the galaxy simply smooth and rounded, with no sign of a disk?

- Smooth
- Features or Disk
- Star or Artifact

Need some help with this task?

Back | Done & Talk | Done

Show the project tutorial


CVEDIA
CVedia
Secure
https://cvedia.com/ui/camelyon-16
CVEDIA
Please note that the platform is still in early-access and not fit for production use.
We would greatly appreciate your feedback and questions on our cvedia-users group
Welcome aj@cvedia.com
Dataset
Camelyon
Camelyon 16
by
407
Images
Test
130
Train
277
Collections
Base
276
x1
Healthy tissue
160
x1
Metastases
116
x1
Output Format
image1
Name
image1
Field
image
Output as
default
Augmentations
Random crop
Apply to:
image1
Width
256
Height
256
Images
1
Enclose
4
Policy
Single crop per image (area balanced)
Inside
Lymph nodes (1079)
Outside
Preview
Display
Export Dataset
image1
image1
image1
image1
Loading...

Experiment Results

Task	Layers	Hidden	% diff. (min / avg / max)
GZ	1	-	21.8 / 38.9 / 73.3
GZ	2	50	2.13 / 9.02 / 14.0
GZ	2	100	-1.05 / 8.89 / 13.5
CAM.	1	-	-3.10 / 4.51 /10.4
CAM. (asym.)	1	-	-1.26 / 5.13 / 15.2
CAM.	2	20	0.30 / 1.82 / 2.65
CAM. (asym.)	2	20	-0.80 / 1.91 / 4.85
CAM.	2	50	0.00 / 0.03 / 2.31
CAM. (asym.)	2	50	-0.67 / 1.70 / 2.28

Experiment Results

Task	Layers	Hidden	% diff. (min / avg / max)
GZ	1	-	21.8 / 38.9 / 73.3
GZ	2	50	2.13 / 9.02 / 14.0
GZ	2	100	-1.05 / 8.89 / 13.5
CAM.	1	-	-3.10 / 4.51 /10.4
CAM. (asym.)	1	-	-1.26 / 5.13 / 15.2
CAM.	2	20	0.30 / 1.82 / 2.65
CAM. (asym.)	2	20	-0.80 / 1.91 / 4.85
CAM.	2	50	0.00 / 0.03 / 2.31
CAM. (asym.)	2	50	-0.67 / 1.70 / 2.28

Experiment Results

Task	Layers	Hidden	% diff. (min / avg / max)
GZ	1	-	21.8 / 38.9 / 73.3
GZ	2	50	2.13 / 9.02 / 14.0
GZ	2	100	-1.05 / 8.89 / 13.5
CAM.	1	-	-3.10 / 4.51 /10.4
CAM. (asym.)	1	-	-1.26 / 5.13 / 15.2
CAM.	2	20	0.30 / 1.82 / 2.65
CAM. (asym.)	2	20	-0.80 / 1.91 / 4.85
CAM.	2	50	0.00 / 0.03 / 2.31
CAM. (asym.)	2	50	-0.67 / 1.70 / 2.28

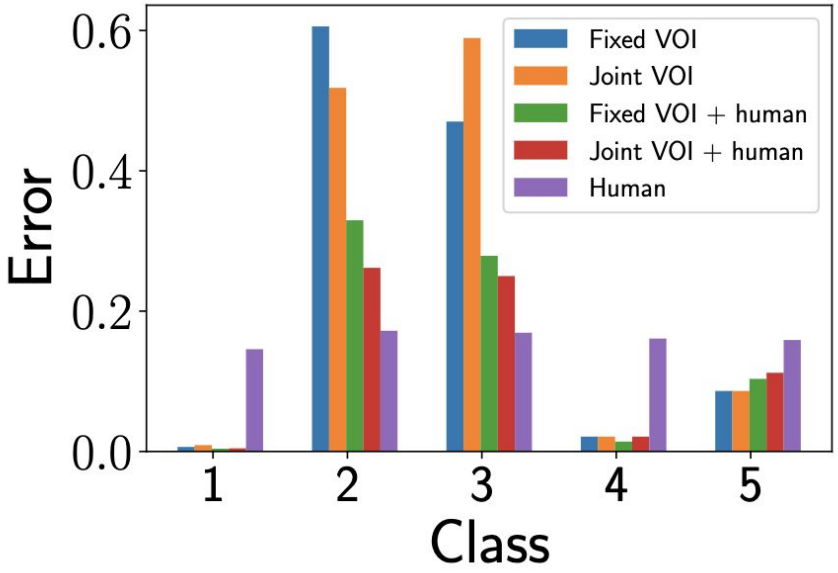
Experiment Results

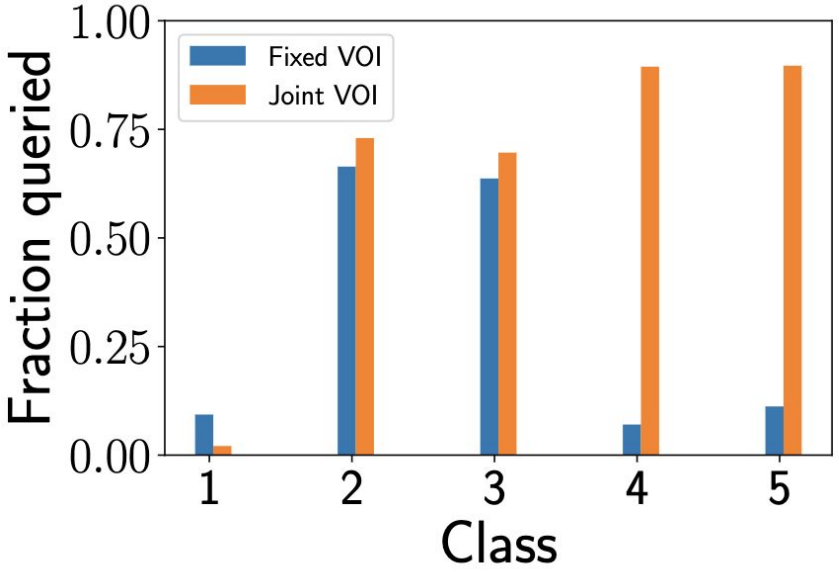
Task	Layers	Hidden	% diff. (min / avg / max)
GZ	1	-	21.8 / 38.9 / 73.3
GZ	2	50	2.13 / 9.02 / 14.0
GZ	2	100	-1.05 / 8.89 / 13.5
CAM.	1	-	-3.10 / 4.51 /10.4
CAM. (asym.)	1	-	-1.26 / 5.13 / 15.2
CAM.	2	20	0.30 / 1.82 / 2.65
CAM. (asym.)	2	20	-0.80 / 1.91 / 4.85
CAM.	2	50	0.00 / 0.03 / 2.31
CAM. (asym.)	2	50	-0.67 / 1.70 / 2.28

Experiment Results

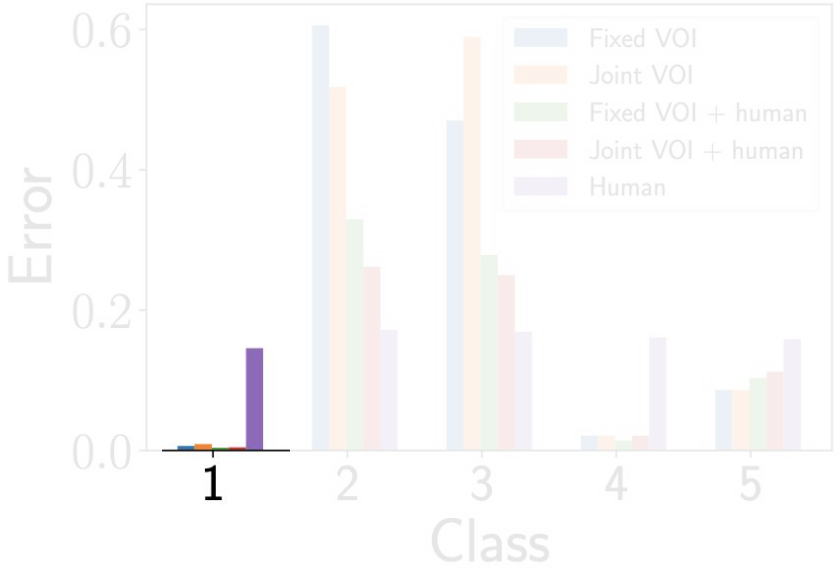
Task	Layers	Hidden	% diff. (min / avg / max)
GZ	1	-	21.8 / 38.9 / 73.3
GZ	2	50	2.13 / 9.02 / 14.0
GZ	2	100	-1.05 / 8.89 / 13.5
CAM.	1	-	-3.10 / 4.51 /10.4
CAM. (asym.)	1	-	-1.26 / 5.13 / 15.2
CAM.	2	20	0.30 / 1.82 / 2.65
CAM. (asym.)	2	20	-0.80 / 1.91 / 4.85
CAM.	2	50	0.00 / 0.03 / 2.31
CAM. (asym.)	2	50	-0.67 / 1.70 / 2.28

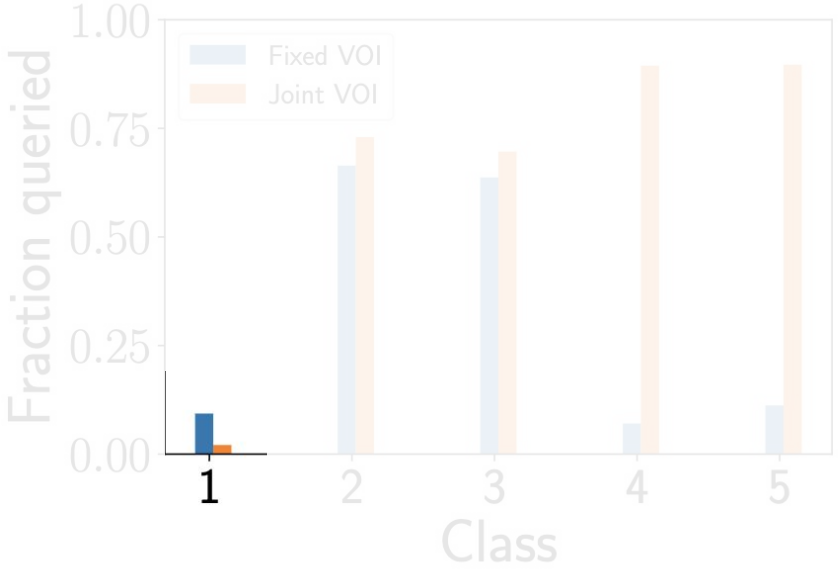
Experiment Results



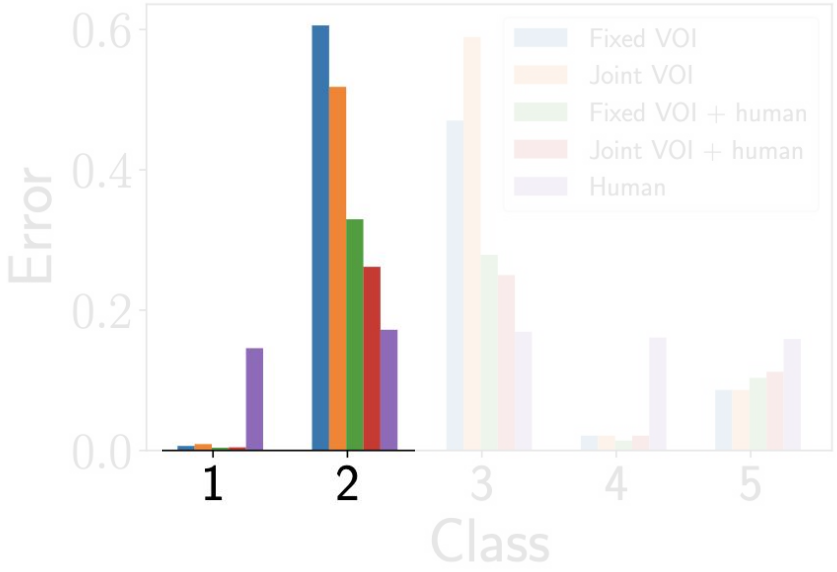


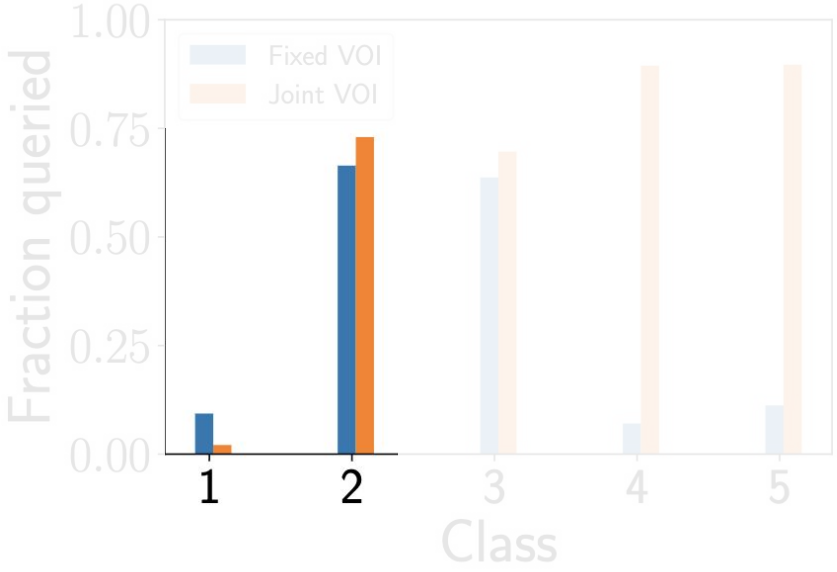
Experiment Results



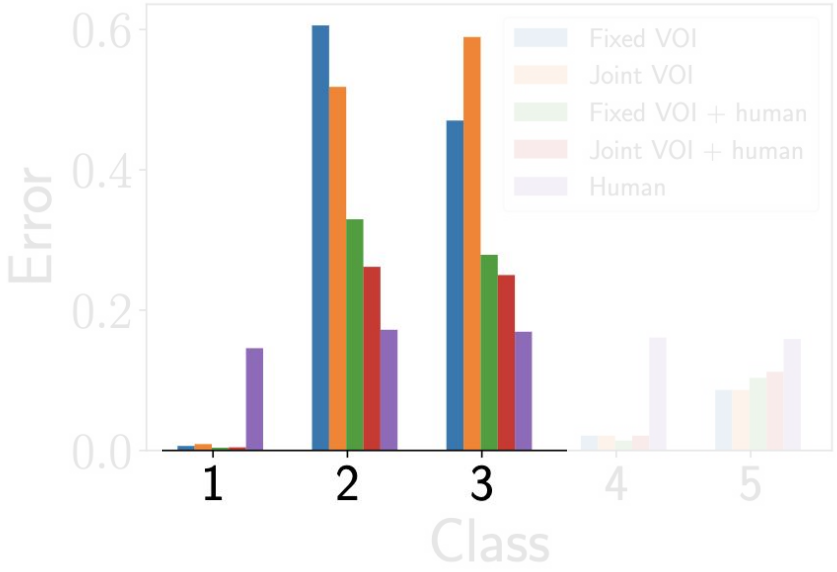


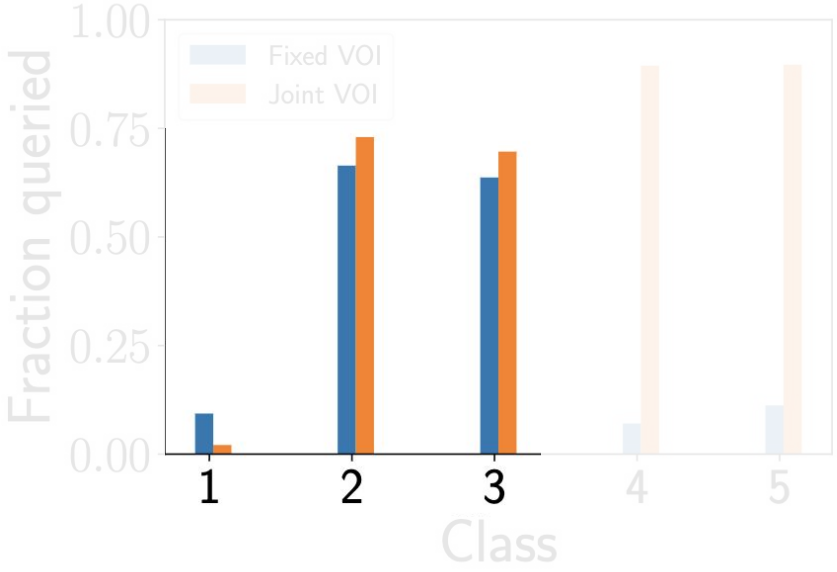
Experiment Results



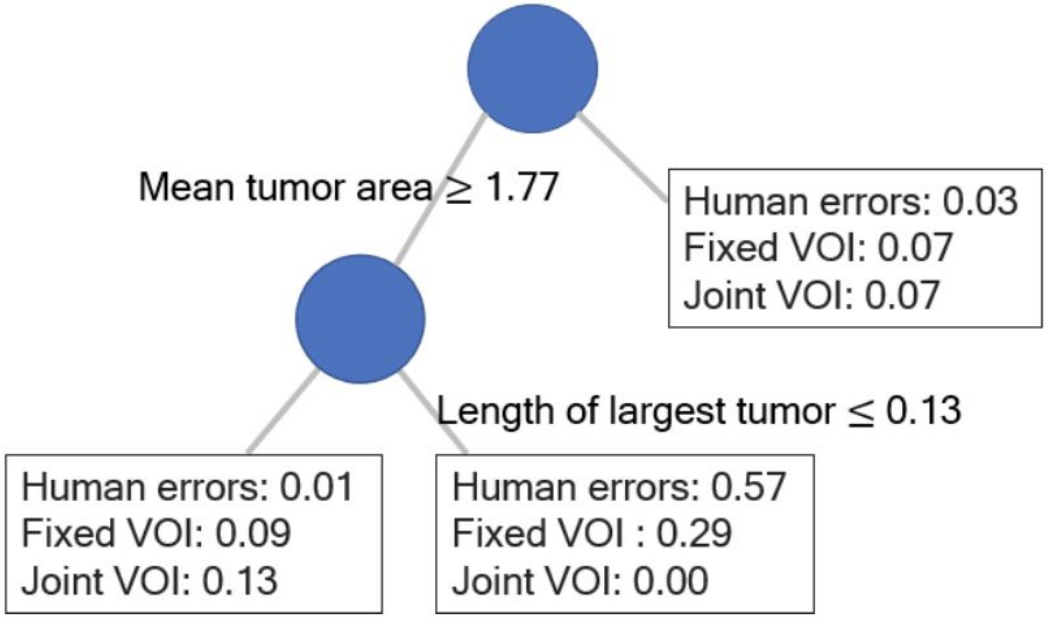


Experiment Results





Experiment Results



Conclusion

- Discussed how ML algorithms can be optimized to complement humans
- Evaluated approaches on two real-world datasets
- Explored how joint training improves optimization