Uncertainty Quantification for Fair and Transparent AI assisted Decision Making

—

Prasanna Sattigeri

Senior Research Scientist
IBM Research

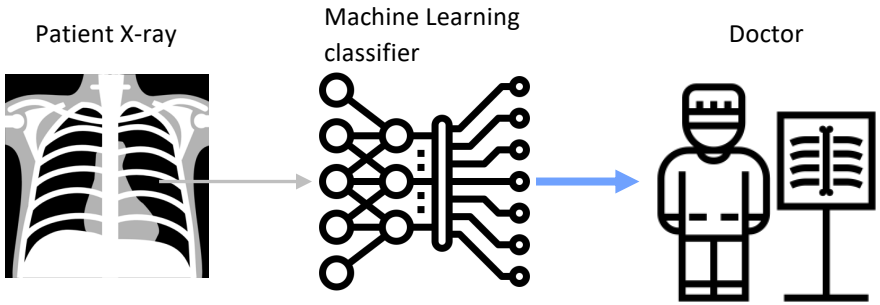
psattig@us.ibm.com

AI-assisted decision-making

one-way

AI makes recommendations

Human decision maker makes the final call

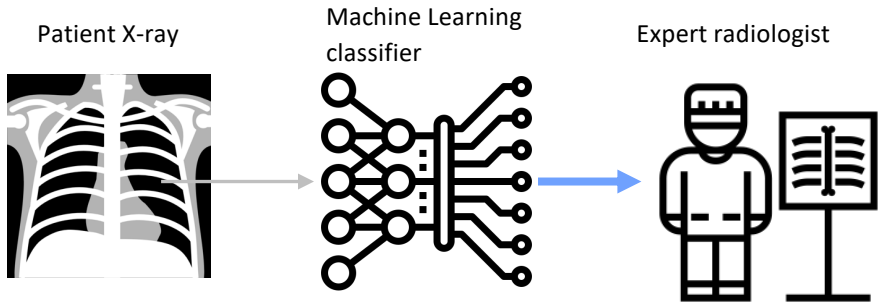


AI-assisted decision-making

one-way

AI makes recommendations

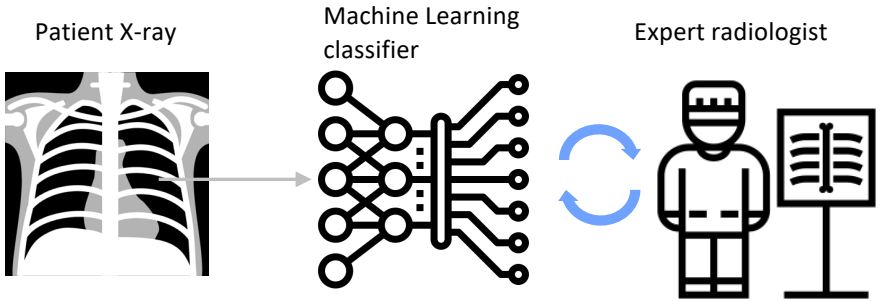
Human decision maker makes the final call



two-way

Human and AI “communicate” each others' strengths

Best “agent” decisions are accepted



Outline

Part 1: *one-way communication using uncertainty– Selective Prediction*

Part 2: *two-way communication – Learning to Defer*

Part 3: *richer communications*

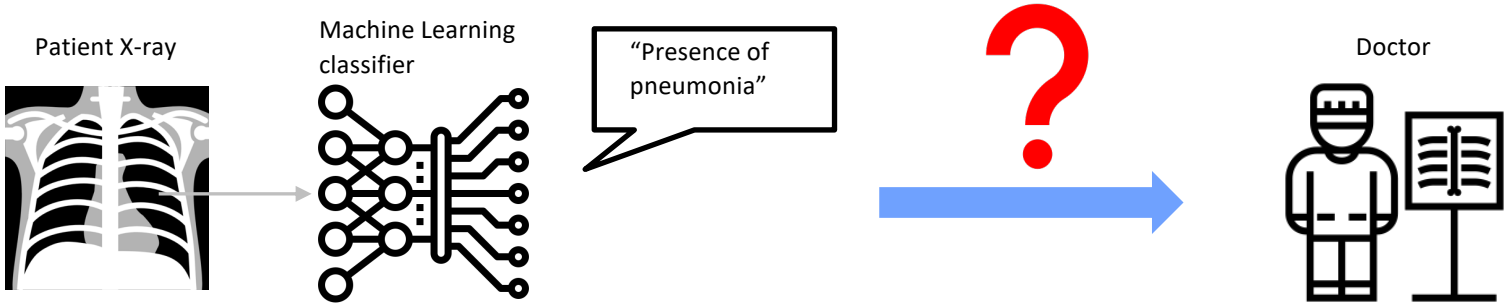
Part 4: *UQ360 toolkit*

Part 1: *one-way communication using uncertainty– Selective Prediction*

Joshua K. Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W. Wornell. "Fair Selective Classification via Sufficiency." In *International Conference on Machine Learning*, pp. 6076-6086. PMLR, 2021.

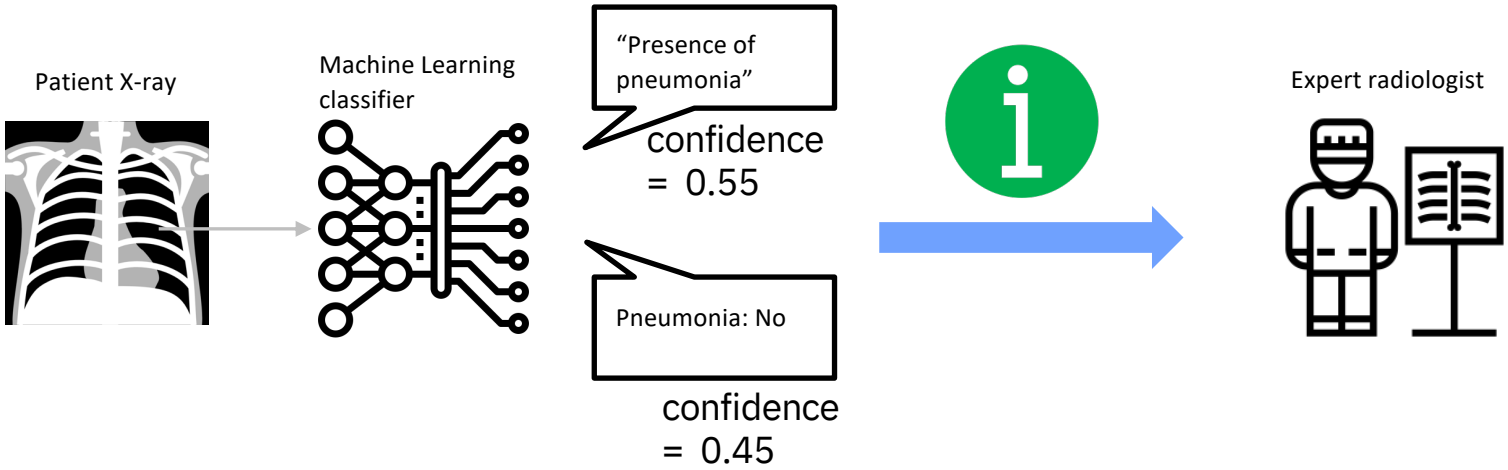
Abhin Shah, Yuheng Bu, Joshua K. Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W. Wornell. "Selective regression under fairness criteria." In *International Conference on Machine Learning*, pp. 19598-19615. PMLR, 2022.

Selective Prediction

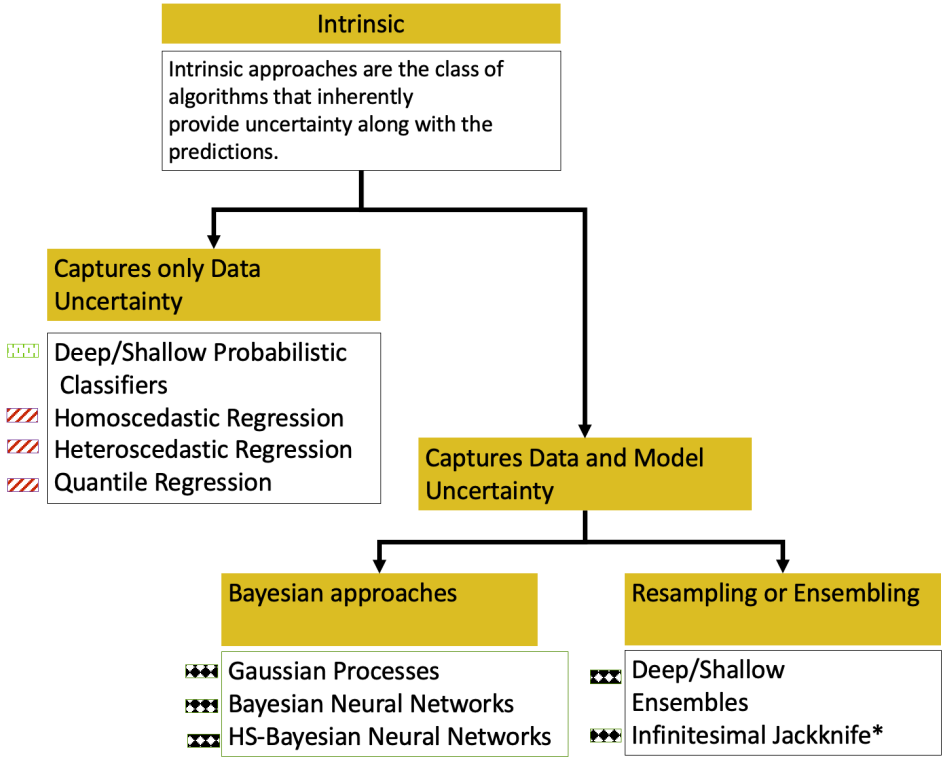


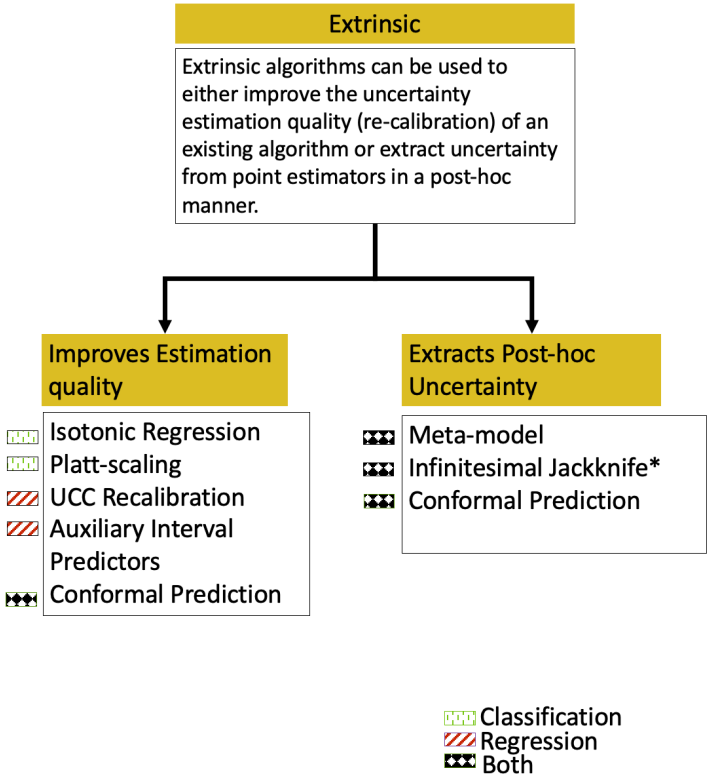
Uncertainty Quantification in AI

It is the ability of an AI model to convey the ***confidence*** in its predictions.

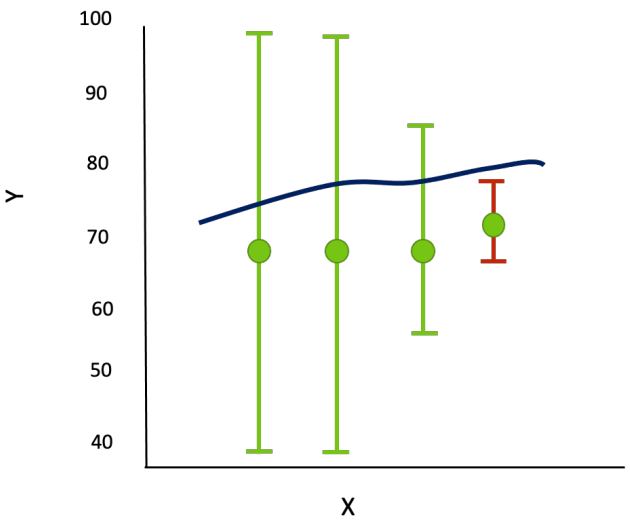


Ways to get uncertainty scores

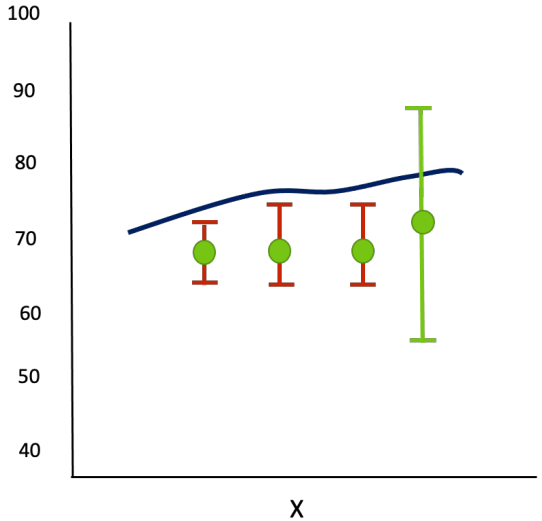




Good properties of uncertainty scores – Coverage



PICP: 0.75
MPIW: 40.0



PICP: 0.25
MPIW: 15.0

Prediction Interval Coverage Probability (PICP)
Mean Prediction Interval Width (MPIW)

Good properties of uncertainty scores – Enable Selective Prediction

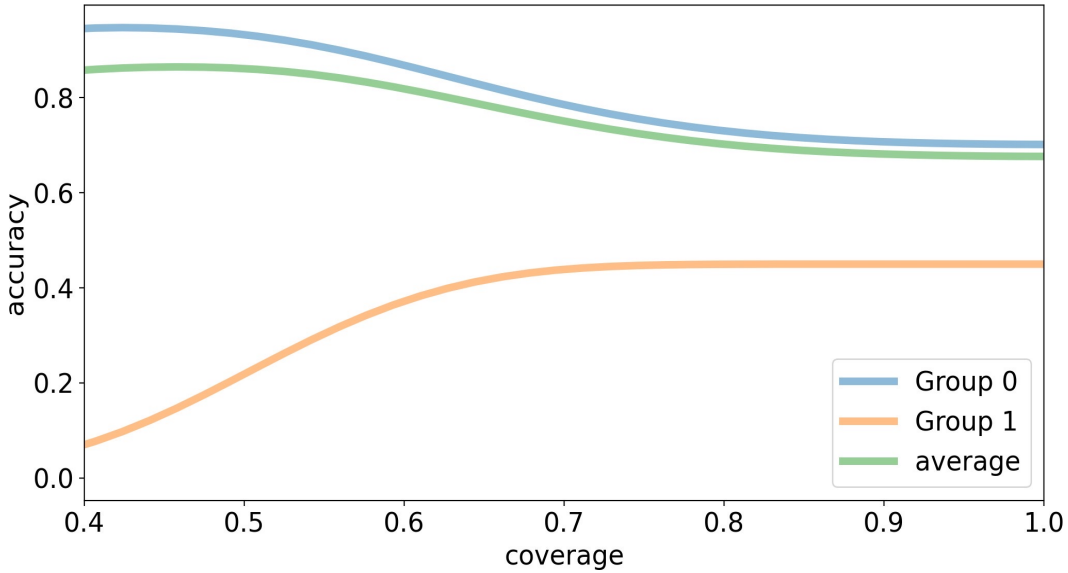
If we have uncertainty measure for each prediction, we can decide to defer decision making if uncertainty is above a certain threshold.

With a good uncertainty measure, reducing the threshold results in a better performance.

The tradeoff is that we have predictions for fewer samples.

Bias in Selective Prediction

Predictors can have ***good average selective prediction*** performance but perform poorly on certain groups, where reducing uncertainty threshold may result in a ***decrease in performance for protected group.***



Fair Selective Prediction

We want to construct a good predictor $\hat{y} = f(X)$ and uncertainty measure $g(X)$ under some loss functions (error rate or MSE) to perform selective prediction, while being "fair" with respect to D.

Fair Selective Prediction

We want to construct a good predictor $\hat{y} = f(X)$ and uncertainty measure $g(X)$ under some loss functions (error rate or MSE) to perform selective prediction, while being "fair" with respect to D.

There are two questions need to be addressed

- What is the most "useful" ***fairness criterion*** in selective prediction?

Fair Selective Prediction

We want to construct a good predictor $\hat{y} = f(X)$ and uncertainty measure $g(X)$ under some loss functions (error rate or MSE) to perform selective prediction, while being "fair" with respect to D.

There are two questions need to be addressed

- What is the most "useful" ***fairness criterion*** in selective prediction?
- How to ***construct fair selective predictor*** in a simple and efficient way?

Monotonic Selective Risk

What is the most “useful” ***fairness criterion*** in selective prediction?

We define the fairness notion called ***monotonic selective risk (MSR)***

> MSR requires our predictor and uncertainty measure to ensure that the subgroup error rate (or MSE for regression) decreases *monotonically* with a decrease in coverage for *every* subgroup.

Monotonic Selective Risk

What is the most “useful” ***fairness criterion*** in selective prediction?

We define the fairness notion called ***monotonic selective risk (MSR)***

> MSR requires our predictor and uncertainty measure to ensure that the subgroup error rate (or MSE for regression) decreases *monotonically* with a decrease in coverage for *every* subgroup.

Another possible fairness notion is to require same performance and coverage curve for all subgroups.

But it is still possible that the error rate of both subgroups increases when coverage decreases, which violates the primary goal of selective prediction!

Monotonic Selective Risk

How to ***construct fair selective predictor*** in a simple and efficient way?

Suppose the representation $\Phi(X)$ *is sufficient, i.e.,* $Y \perp D|\Phi(X)$. *Let* $f(\Phi(X)) = \mathbb{E}[Y|\Phi(X)]$ *and* $g(\Phi(X)) = \mathrm{Var}[Y|\Phi(X)]$. *Then,* $f(\Phi(X))$ *and* $g(\Phi(X))$ *satisfy the Monotonic Selective Risk condition.*

Monotonic Selective Risk

How to ***construct fair selective predictor*** in a simple and efficient way?

> *Suppose the representation* $\Phi(X)$ *is sufficient, i.e.,* $Y \perp D|\Phi(X)$*. Let* $f(\Phi(X)) = \mathbb{E}[Y|\Phi(X)]$ *and* $g(\Phi(X)) = \mathrm{Var}[Y|\Phi(X)]$*. Then,* $f(\Phi(X))$ *and* $g(\Phi(X))$ *satisfy the Monotonic Selective Risk condition.*

Calibrated mean & variance ensures same behavior for all groups

Using conditional mean as prediction and conditional variance as uncertainty measure is optimal under both zero-one and squared loss

Imposing Sufficiency Criteria

The fair selective prediction objective can be expressed as

$$\min_{\theta} \quad L(\hat{y}, y) \quad \text{s.t.} \quad Y \perp D | \Phi(X),$$

where L is cross-entropy loss and θ are the model parameters.

This hard constraint can be relaxed into the following soft constraint:

$$\min_{\theta} L(\hat{y}, y) + \lambda I(Y; D | \Phi(X)).$$

where λ is a regularizer.

Imposing Sufficiency Criteria

We use the following upper bound of mutual information in our algorithm

For random variables X, Y and Z, we have

$$I(X;Y|Z) \leq \mathbb{E}_{P_{XYZ}}[\log P(Y|X,Z)] - \mathbb{E}_{P_X}[\mathbb{E}_{P_{YZ}}[\log P(Y|X,Z)]],$$

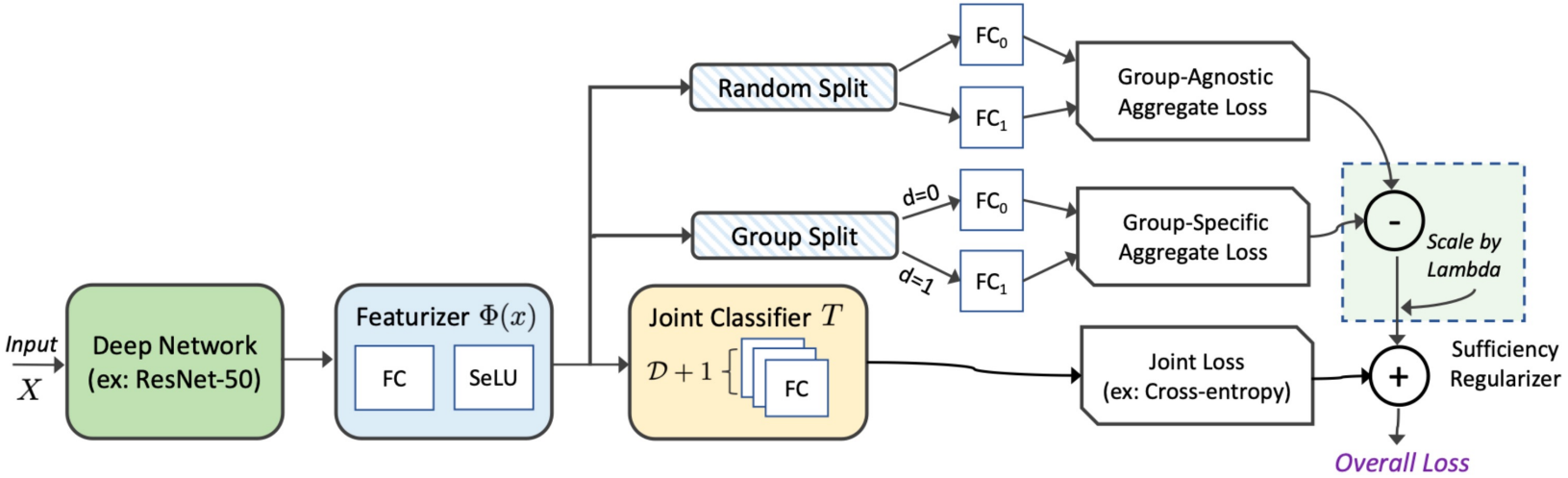
where equality is achieved if and only if $X \perp Y \mid Z$.

In our case, we have

$$I(Y;D|\Phi(X)) \leq \mathbb{E}_{P_{XYD}}[\log P(Y|\Phi(X),D)] - \mathbb{E}_{P_D}[\mathbb{E}_{P_{YX}}[\log P(Y|\Phi(X),D)]]$$

Imposing Sufficiency Criteria

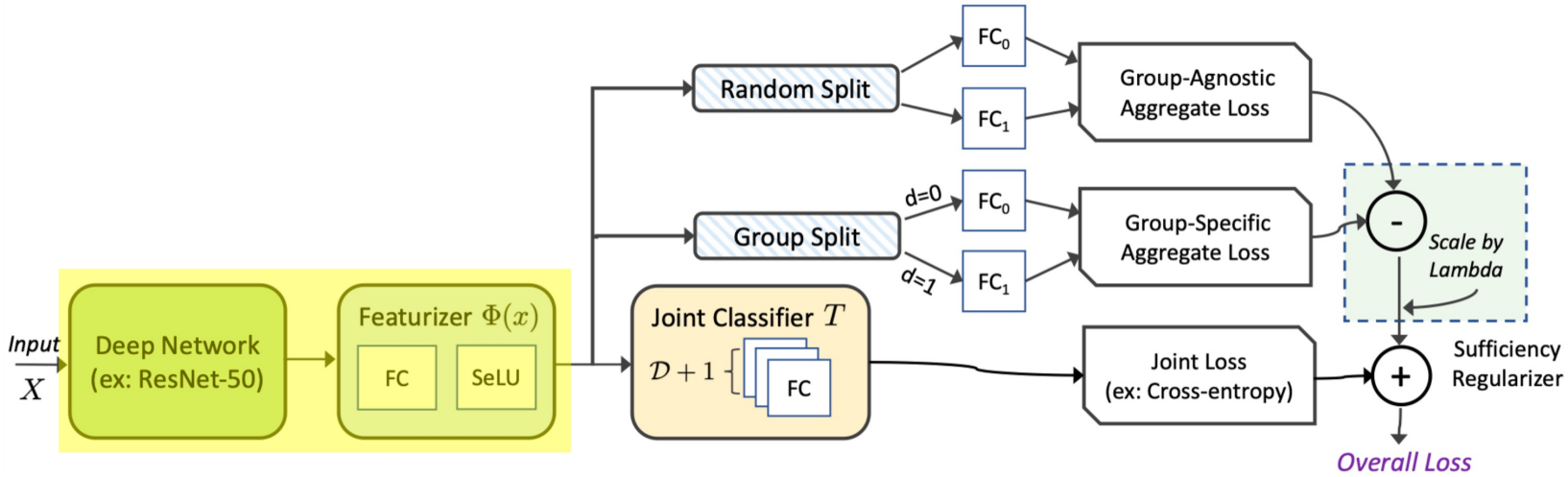
$$\min L\big(T(\Phi(X)), Y\big)$$
$$+ \lambda\big(\mathbb{E}_{P_{XYD}}[\log P(Y|\Phi(X), D)] - \mathbb{E}_{P_D}[\mathbb{E}_{P_{YX}}[\log P(Y|\Phi(X), D)]]\big)$$



Imposing Sufficiency Criteria

$$\min L\big(T(\Phi(X)), Y\big)$$

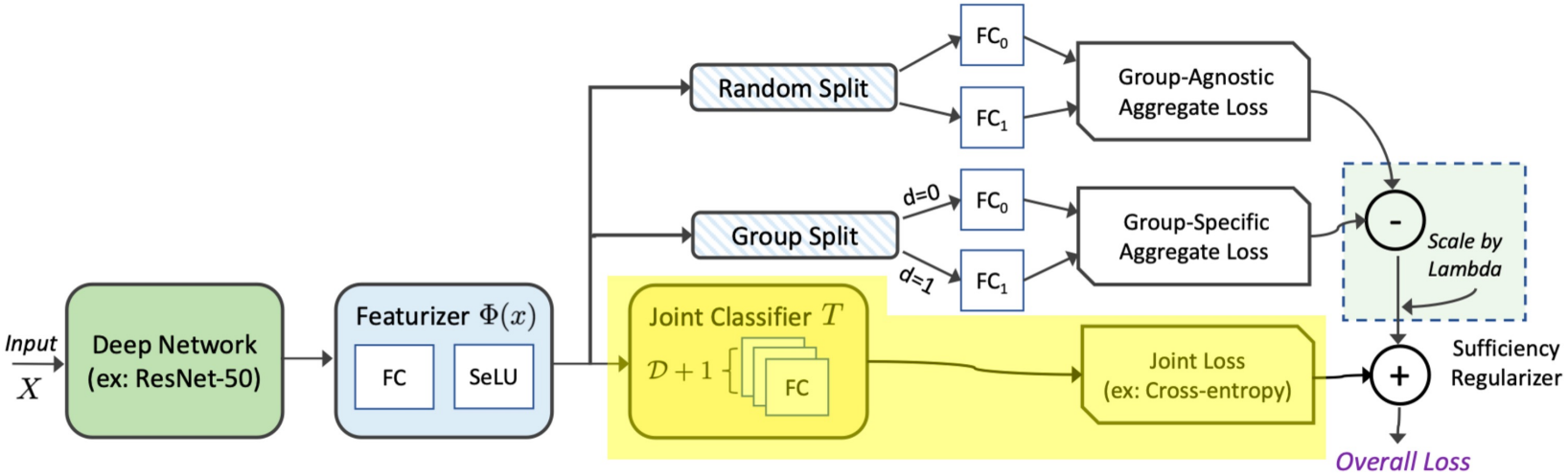
$$+ \lambda\big(\mathbb{E}_{P_{XYD}}[\log P(Y|\Phi(X), D)] - \mathbb{E}_{P_D}[\mathbb{E}_{P_{YX}}[\log P(Y|\Phi(X), D)]]\big)$$



Imposing Sufficiency Criteria

$$\min \; L\big(T(\Phi(X)), Y\big)$$

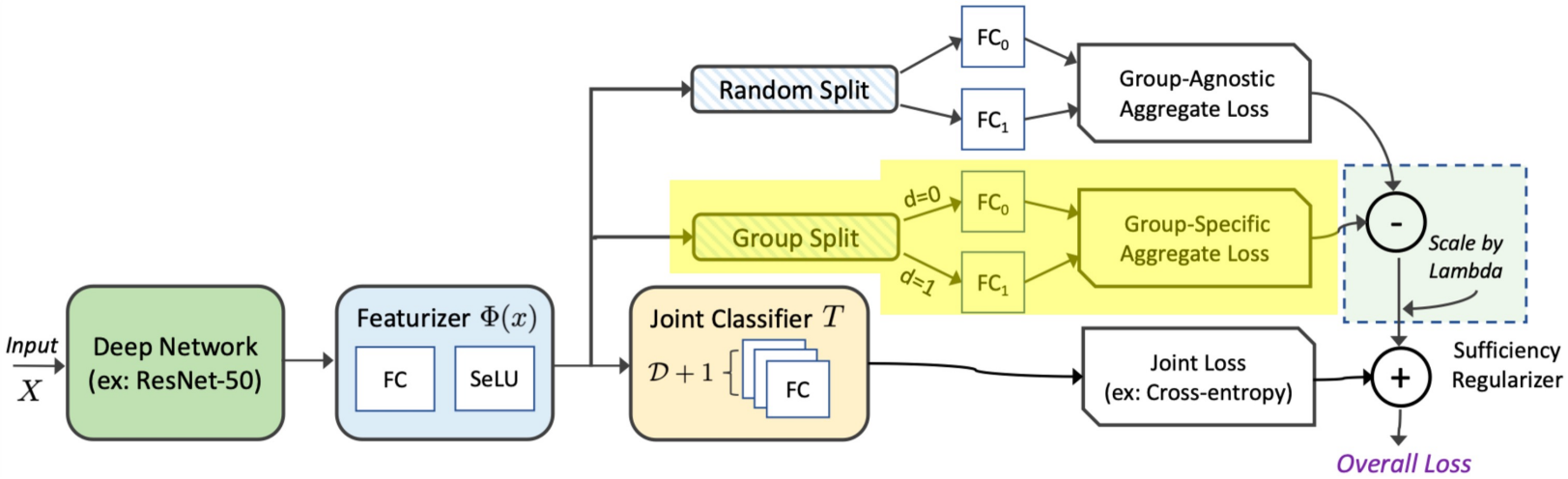
$$+ \lambda\Big(\mathbb{E}_{P_{XYD}}[\log P(Y|\Phi(X), D)] - \mathbb{E}_{P_D}[\mathbb{E}_{P_{YX}}[\log P(Y|\Phi(X), D)]]\Big)$$



Imposing Sufficiency Criteria

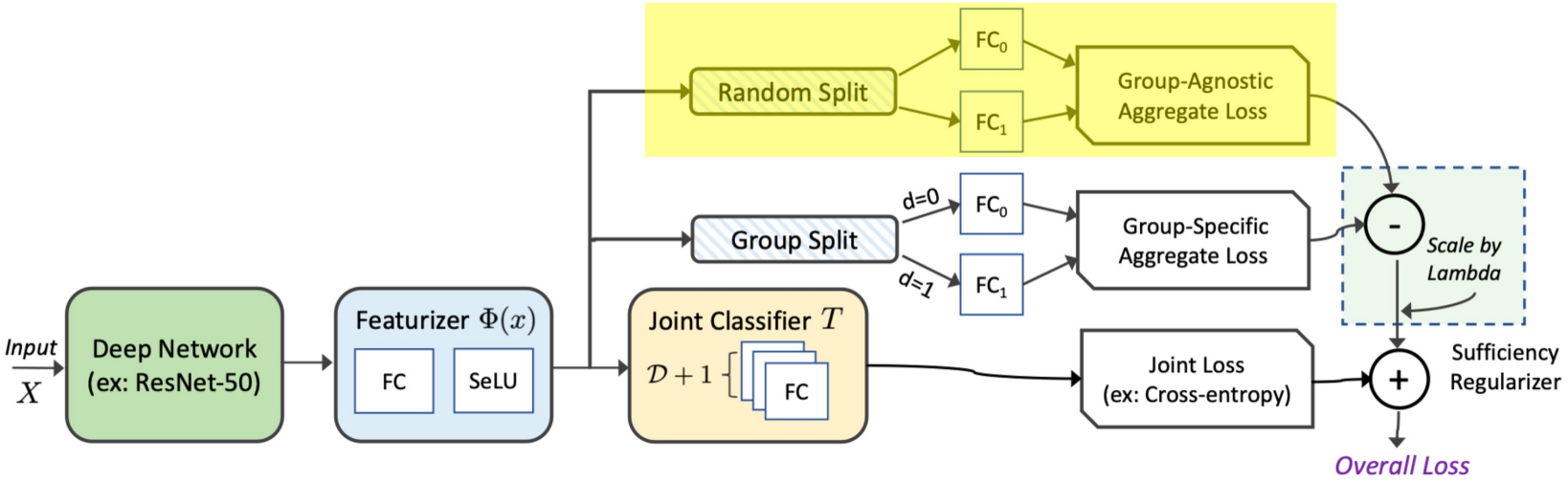
$$\min L\big(T(\Phi(X)), Y\big)$$

$$+ \lambda\big(\mathbb{E}_{P_{XYD}}[\log P(Y|\Phi(X), D)] - \mathbb{E}_{P_D}[\mathbb{E}_{P_{YX}}[\log P(Y|\Phi(X), D)]]\big)$$



Imposing Sufficiency Criteria

$$\min L\big(T(\Phi(X)), Y\big)$$
$$+ \lambda\big(\mathbb{E}_{P_{XYD}}[\log P(Y|\Phi(X), D)] - \mathbb{E}_{P_D}[\mathbb{E}_{P_{YX}}[\log P(Y|\Phi(X), D)]]\big)$$



Fair Selective Regression

In selective regression, there is no direct method to extract the conditional variance from an existing regressor designed to predict only the conditional mean.

Heteroskedastic NN: We train a single a neural network with two heads — one to predict the conditional mean and the other to predict the conditional variance

Under the assumption $P(Y|\Phi(X), D)$ is Gaussian, we have similar implementation as selective classification

Experimental Results: Adult Dataset

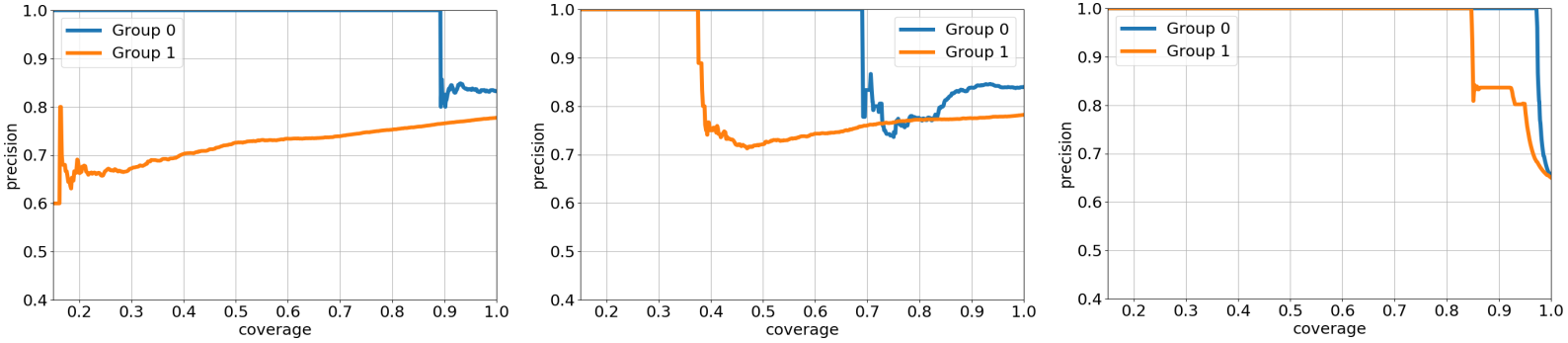
Adult Dataset: census data drawn from the 1994 Census database

Using Demographics to predict individual Income high/low with Sex as sensitive attribute.

We plot coverage-precision graph for each group, where coverage is the fraction of samples we make decisions on.

Compare our method to a baseline trained using only the cross-entropy loss, and to the Distributionally Robust Optimization (DRO) method of (Sagawa et al., 2019), which has been shown to mitigate this disparity in prior works.

Experimental Results: Adult Dataset



(a) Baseline (b) DRO (c) Ours

Group-specific precision-coverage curves for Adult dataset

Experimental Results: Communities and Crimes (C&C)[1]

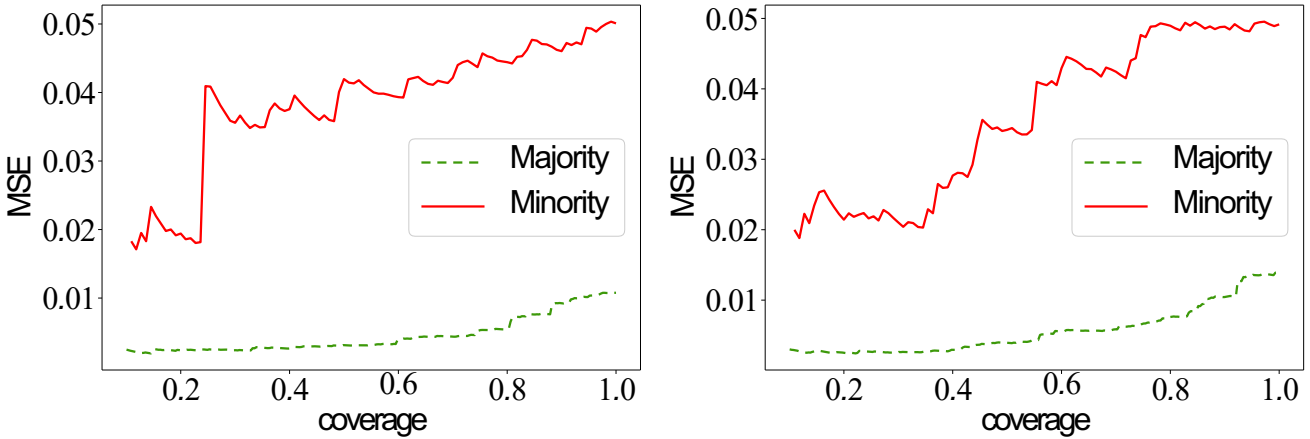
Predict the crime rate of a community

Using a set of 121 real-valued statistics (distributions of income, age, urban/rural, etc.)

Race is set to be sensitive attribute

[1] http://archive.ics.uci.edu/ml/datasets/communities+and+crime

Experimental Results: Communities and Crimes (C&C)



(a) Baseline

(b) Ours

Group-specific MSE-coverage curves for C&C dataset

Fair Selective Prediction - Summary

Monotonic selective risk (MSR) is a natural choice of fairness criteria in selective prediction

Sufficiency criterion can be used ensure MSR, which mitigate disparities between different groups

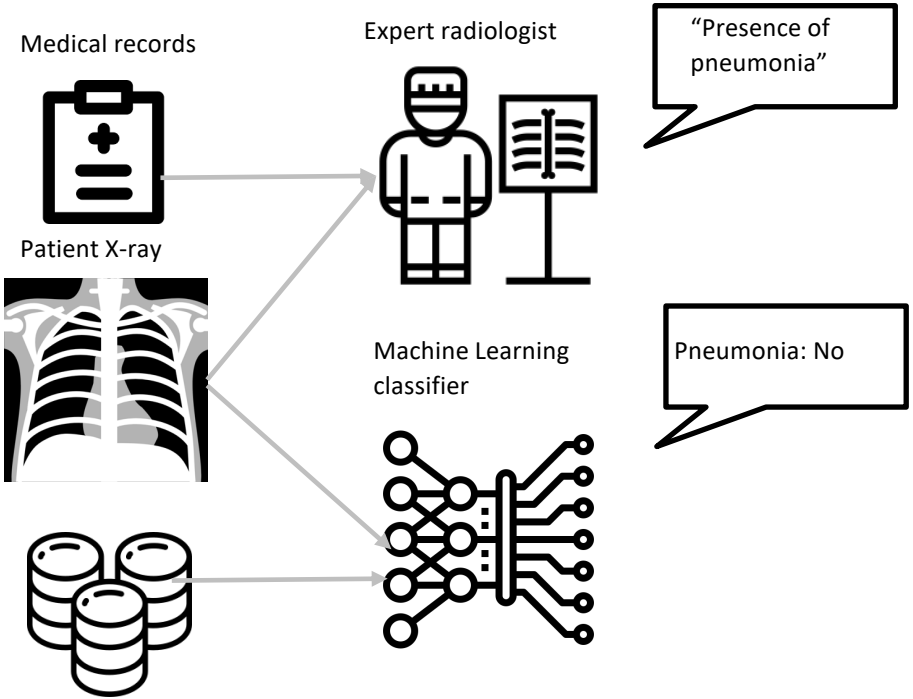
Due to formulation of sufficiency condition, a contrastive upper bound of the mutual information can be used, coupled with group-specific classifiers in implementation.

Part 2: *two-way communication – Learning to Defer*

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, David Sontag, “Algorithms For Learning to Defer”, On-going work.

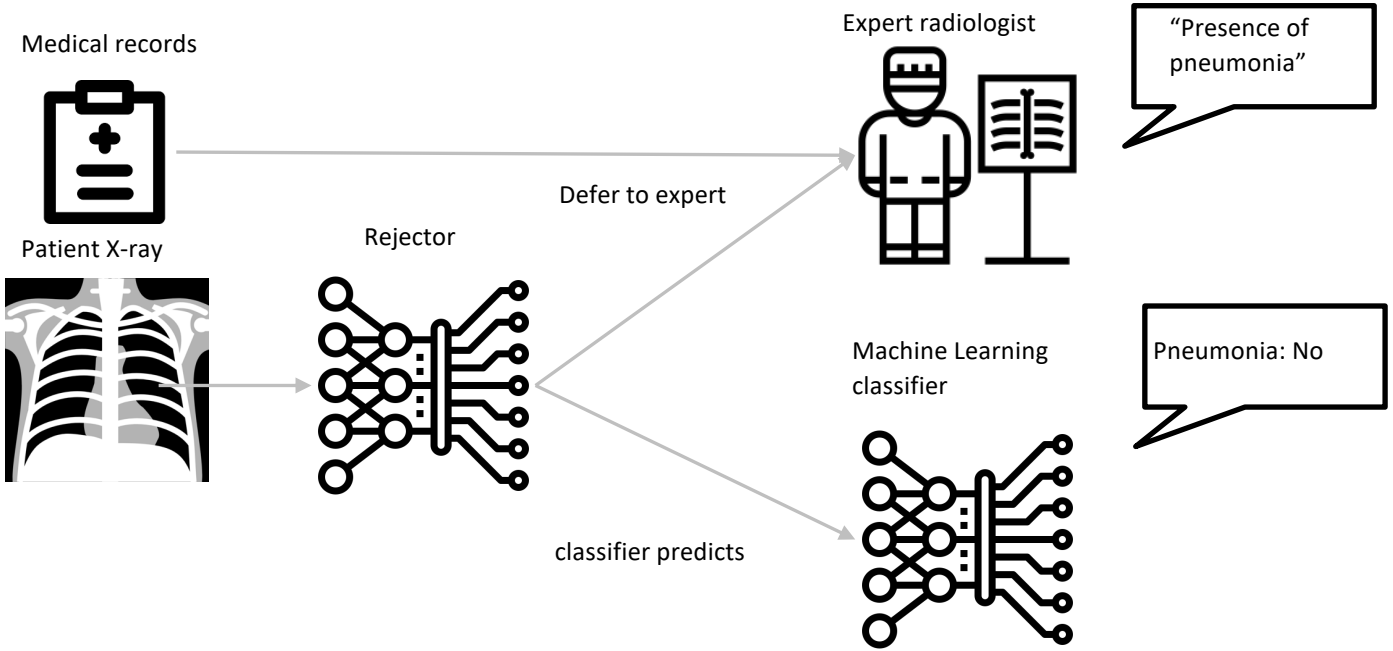
Learning to Defer

two-way communication



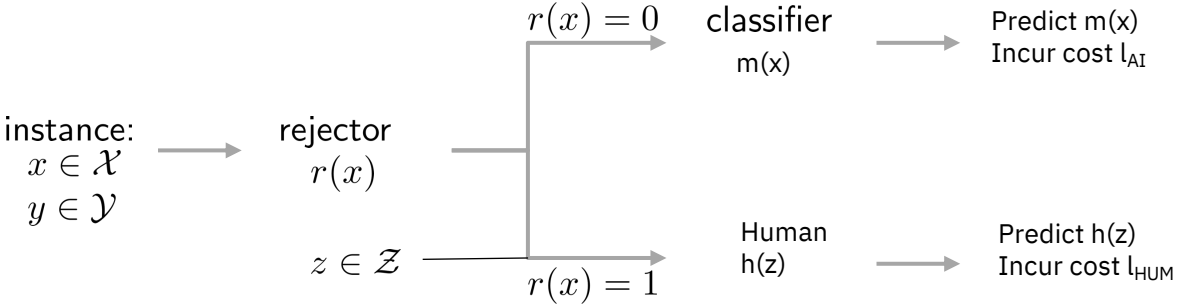
Learning to Defer

two-way communication



Learning to Defer

Problem Formulation



Jointly learn a classifier *m(x)* and rejector *r(x)* to minimize system loss:

$$L_{\text{def}}(m, r) = \mathbb{E}_{X,Y,Z}\ [\ \ell_{\text{AI}}\big(X, Y, m(X)\big) \cdot \mathbb{I}_{r(X)=0} + \ell_{\text{HUM}}(X, Y, h(Z)) \cdot \mathbb{I}_{r(X)=1}\ ].$$

Solving Learning to Defer Exactly

Mixed integer linear program (MILP) that can exactly solve the learning to defer problem in the linear setting:

- MILP can handle binary and multiclass problems
- MILP can easily incorporate many constraints such as fairness, coverage constraints and more.
- Practically: use pre-trained domain representations and learn linear classifiers and rejectors

$$
\begin{aligned}
&M^*, R^*, \cdots = \\
&\underset{M,R,\{r_i\},\{t_i\},\{\phi_i\}}{\arg\min} \sum_i \phi_i + r_i \mathbb{I}_{h_i \neq y_i}, \ \ s.t. \\
&\phi_i \geq t_i - r_i, \qquad \phi_i \geq 0 \quad \forall i \in [n] \\
&K_m t_i \geq \gamma_m - y_i M^\top x_i \quad \forall i \in [n] \\
&R^\top x_i \leq K_r r_i + \gamma_r (r_i - 1), \\
&R^\top x_i \geq K_r (r_i - 1) + \gamma_r r_i \quad \forall i \in [n] \\
&r_i \in \{0, 1\}, t_i \in \{0, 1\}, \\
&\phi_i \in \mathbb{R}^+ \quad \forall i \in [n], \ M, R \in \mathbb{R}^d
\end{aligned}
$$

Differentiable Surrogate with Provable Guarantees

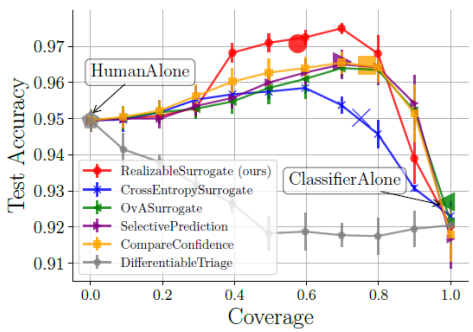
The MILP main limitations are 1) expensive run-time, 2) restriction to linear setting

let $g_i : \mathcal{X} \to \mathbb{R}$ for $i \in [K+1]$

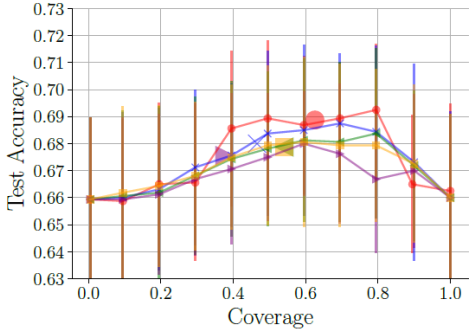
$$L_{CE}(\mathbf{g}, \cdot) = -\log\left(\frac{\exp(g_y(x))}{\sum_{y' \in \mathcal{Y} \cup \perp} \exp(g_{y'}(x))}\right) - \mathbb{I}_{h=y} \log\left(\frac{\exp(g_\perp(x))}{\sum_{y' \in \mathcal{Y} \cup \perp} \exp(g_{y'}(x))}\right)$$

$$L_{RS}(\mathbf{g}, \cdot) = -2\log\left(\frac{\exp(g_y(x)) + \mathbb{I}_{h=y} \exp(g_\perp(x))}{\sum_{y' \in \mathcal{Y} \cup \perp} \exp(g_{y'}(x))}\right)$$

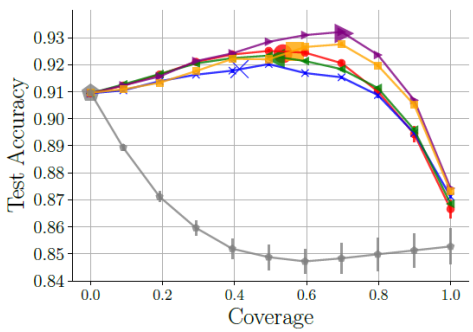
Evaluation on Real Datasets



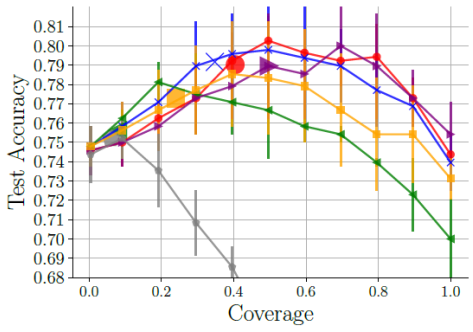
(a) CIFAR-10H



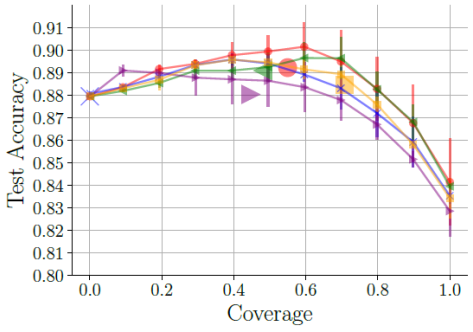
(b) COMPASS



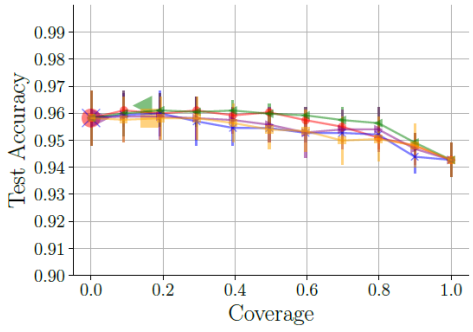
(c) HateSpeech



(d) ImageNet-16H



(e) Chest X-ray - Airspace Opacity



(f) Chest X-ray - Pneuomathorax

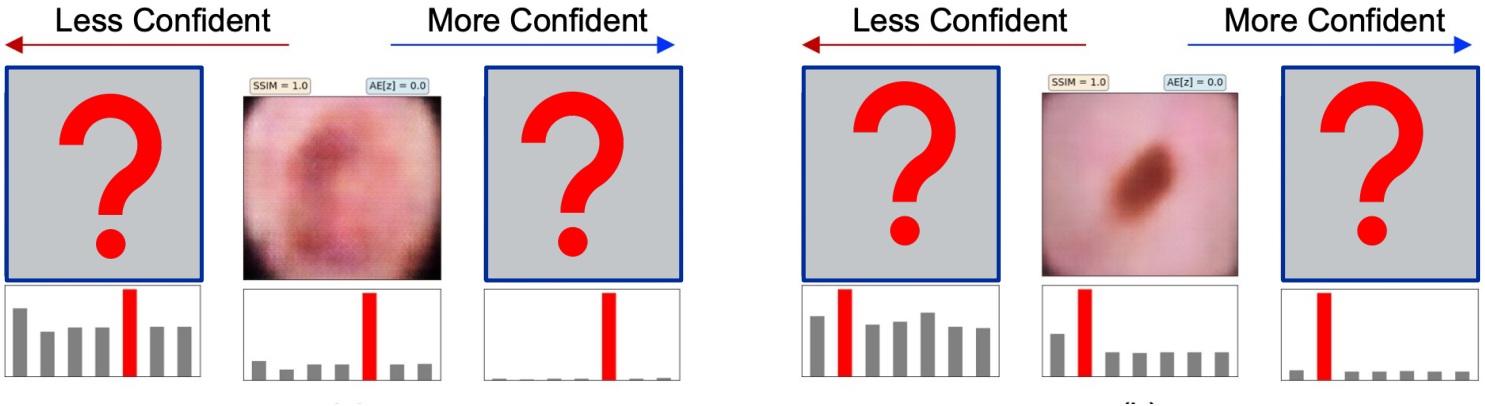
Part 3: *richer communications*

Jayaraman J. Thiagarajan, Bindya Venkatesh, Deepta Rajan, and Prasanna Sattigeri. "Improving reliability of clinical models using prediction calibration." In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pp. 71-80. Springer, Cham, 2020.

Bhatt, Umang, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon et al. "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401-413. 2021.

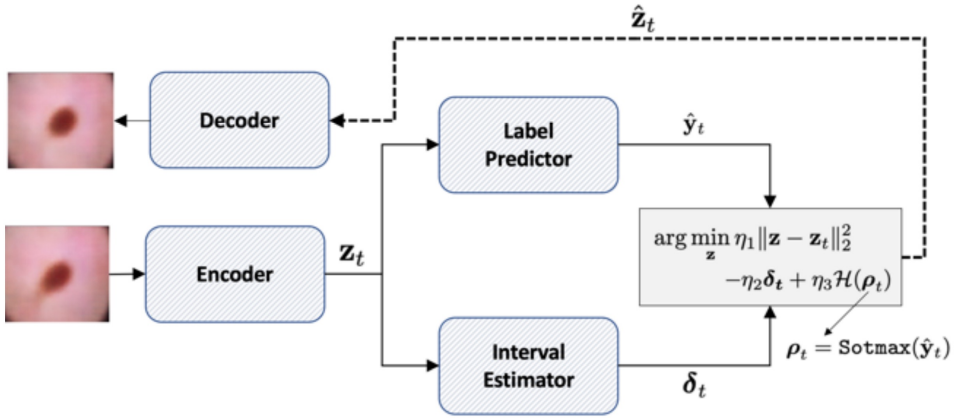
Uncertainty based Introspection

Decision makers want to know what makes the model confident and vice versa?



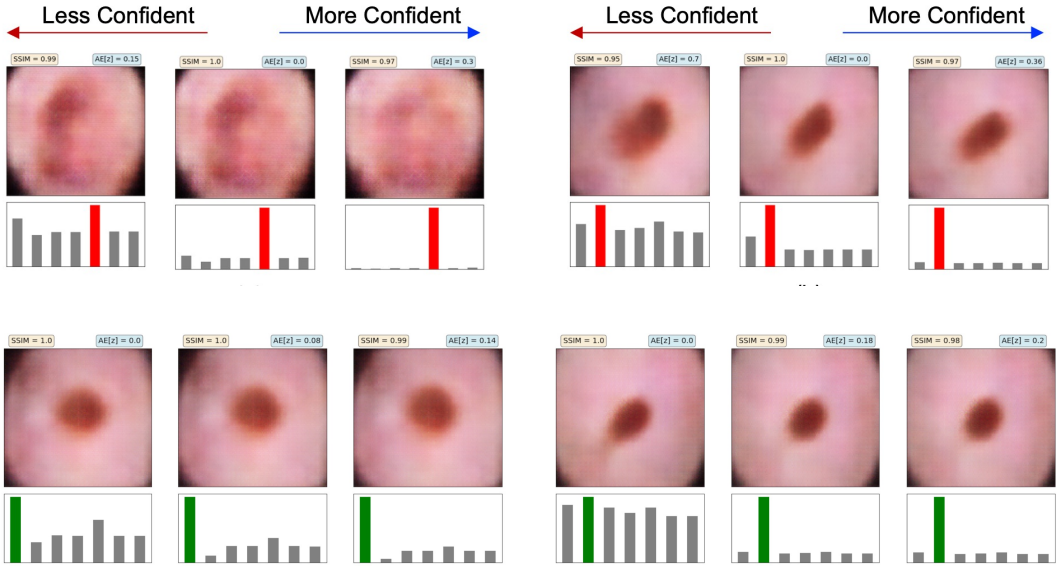
Uncertainty based Introspection

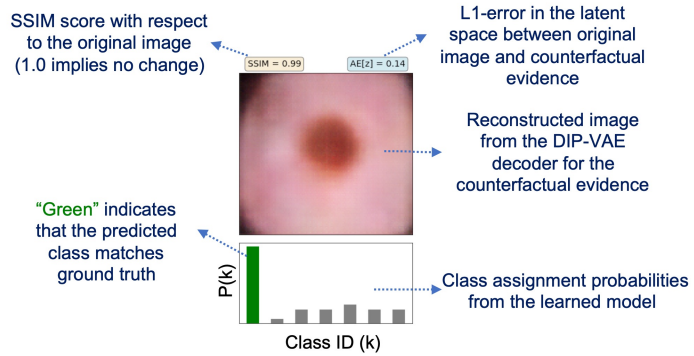
Better Uncertainty scores can lead to holistic Model Introspection

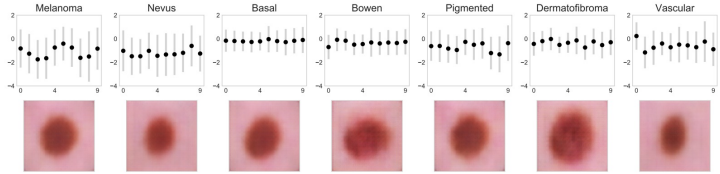


Optimization strategy for generating uncertainty based contrastive evidences

Uncertainty based Explanations





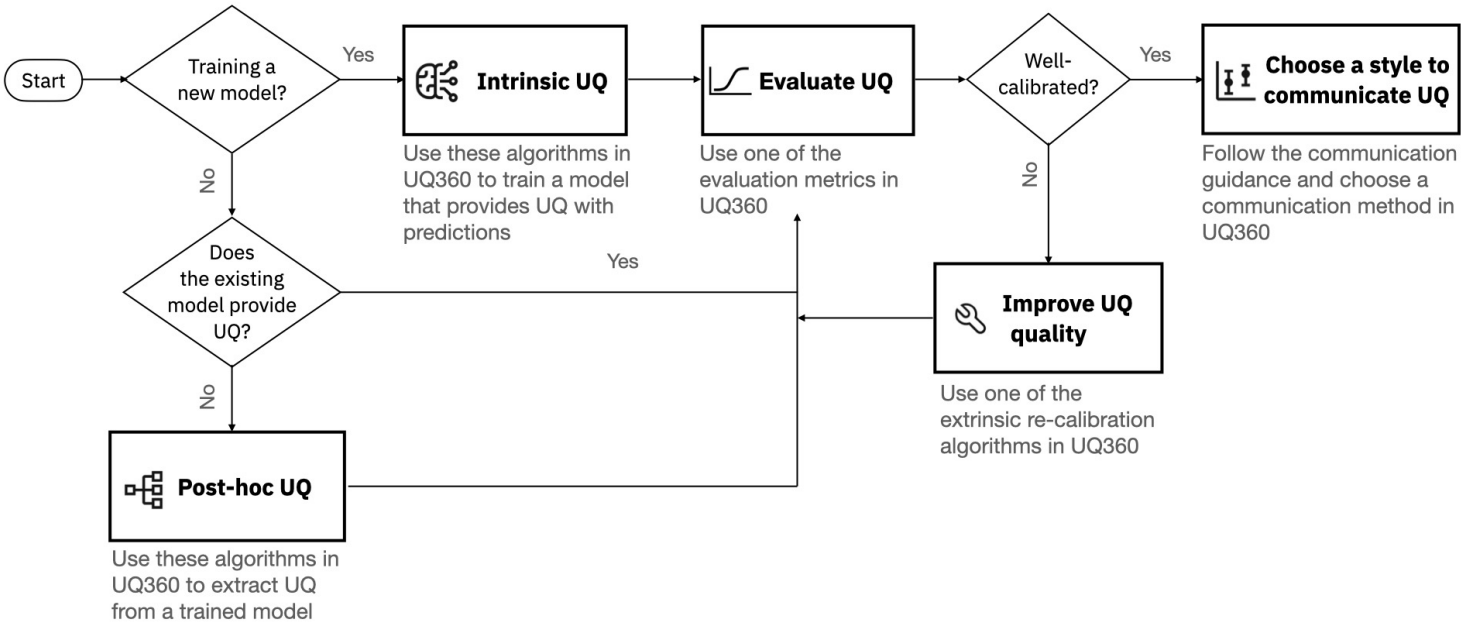


By simultaneously viewing the evidences in different confidence regimes, one can obtain a holistic understanding of the model.

Part 4: *UQ360 toolkit*

Soumya Ghosh, Q. Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush Varshney, and Yunfeng Zhang. "Uncertainty Quantification 360: A Hands-on Tutorial." In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pp. 333-335. 2022.

UQ360 - an Open-Source Toolkit for UQ workflows



API and Usage

```
from uq360.algorithms.quantile_regression import QuantileRegression
```

Train Quantile Regression

```
config = {
    "alpha":0.95,
    "n_estimators":20,
    "max_depth":3,
    "learning_rate":0.1,
    "min_samples_leaf":20,
    "min_samples_split":20
}

uq_model = QuantileRegression(model_type='gbr', config=config)
```

```
uq_model = uq_model.fit(X_train, y_train.squeeze())
```

```
y_mean, y_lower, y_upper = uq_model.predict(X_test)
y_mean, y_lower, y_upper = scaler_y.inverse_transform(y_mean), scaler_y.inverse_transform(y_lower)
, scaler_y.inverse_transform(y_upper)
```

`uq360` models with sklearn's `GridsearchCV`

```
sklearn_picp = make_sklearn_compatible_scorer(task_type="regression", metric="picp", greater_is_be
tter=True)
```

```
clf = GridSearchCV(QuantileRegression(config=base_config), configs, scoring=sklearn_picp)
```

```
clf.fit(X_train, y_train)
```

UQ Communication Methods

Guidance on Communicating Uncertainty

Overview | Communicate for Regression | Communicate for Classification

Overview: What to consider when choosing communication methods

Communicating UQ means presenting the output of UQ estimates to stakeholders, assuming you have chosen the right UQ algorithm to generate the right type of UQ estimates (see our UQ algorithm selection guidance). This is a crucial step because even a well-calibrated UQ estimates could be misunderstood by people if they have difficulty or biases in interpreting the numbers or statistics. In this guide, we will introduce you to some key considerations for communicating UQ and example methods. In practice, it is necessary to conduct tests with your target users or stakeholders to make sure the chosen UQ communication method is understood.

Let's start with a few key questions that should guide your choice of UQ communication methods.

What is the form of the UQ output?

The first step is to identify the form of the UQ to be communicated, i.e. whether it is a single confidence score or a distribution of possible outcomes. In current ML tasks, the former is what UQ estimates of a classification model looks like, and the latter is the form of UQ estimates of a regression model.

Note here we focus on the form instead of the source of UQ. For example, for a regression model, UQ of different sources, whether it is data uncertainty, model uncertainty, or overall predictive uncertainty, are all distributions of possible outcomes. They could be communicated in the same way but it is possible that users would perceive them or act on them differently.

Communicating UQ of a single instance or a group of instances?

Then select a communication method below:

Range of interval (verbal)
Easy to read at a glance, but could miss the details of how possible values are distributed in the range

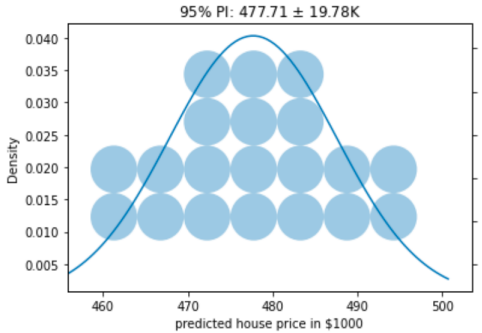
Probability density plot
Gives detailed information with a visualization about how possible values are distributed in the prediction interval

Quantile dot plot
Shows distribution with a visualization that makes it easier to judge the relative likelihood of where possible values can fall

Recommended price:

478K

The quantile plot below shows the probability of the right price.

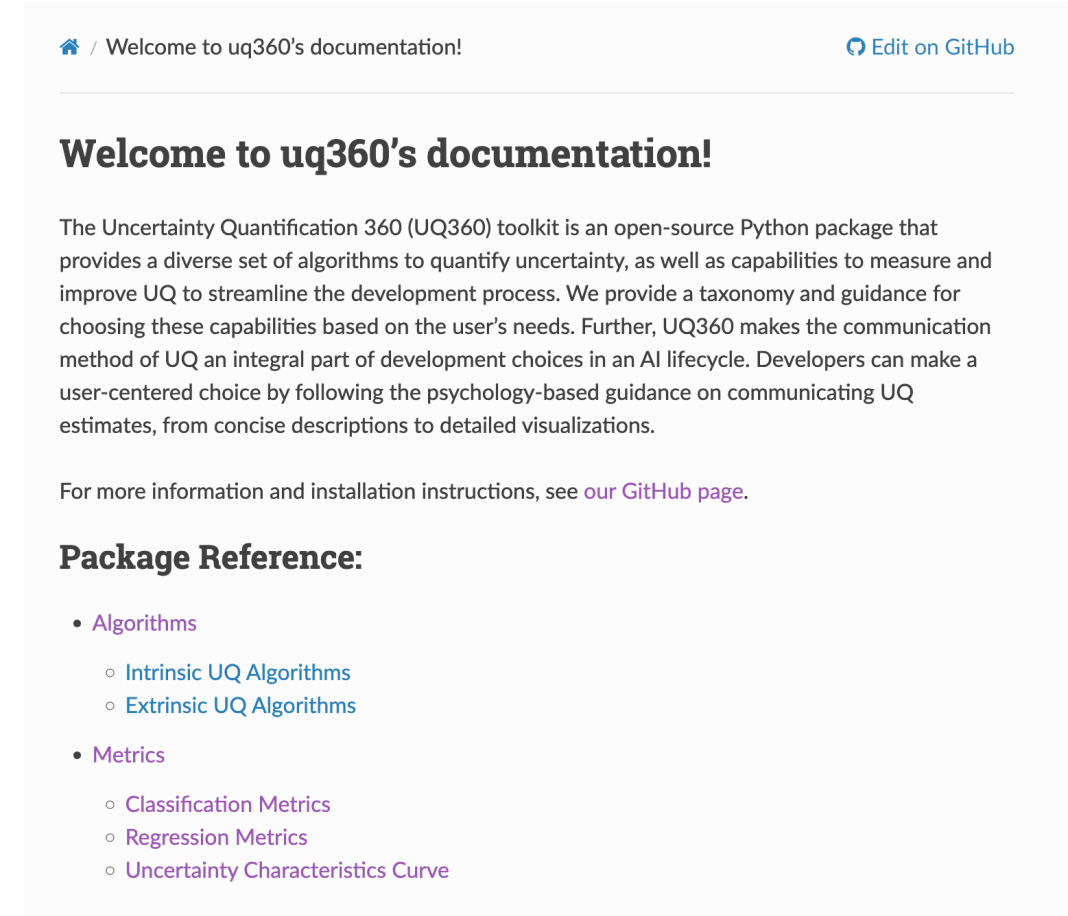


Ella understands that while the model prediction output is 478K, the recommended range says the price can fall anywhere between 458K and 498K.

This is a somewhat large range. Ella has to also take into consideration what she knows about the housing market:

- The buying demand is on the rise in the neighborhood
- It is common to set the first price slightly higher

We welcome contributions and feedback!



http://uq360.mybluemix.net
https://github.com/ibm/uq360
https://pypi.org/project/uq360

Please join the Slack community
http://aif360.mybluemix.net/community

Channel: #uq360-users

Trustworthy AI toolkits

AI Fairness 360 http://aif360.mybluemix.net/

AI Explainability 360 http://aix360.mybluemix.net/

Adversarial Robustness 360 http://art360.mybluemix.net/

Uncertainty Quantification 360 http://uq360.mybluemix.net/

AI Privacy 360 http://aip360.mybluemix.net/

Causal Inference 360 http://ci360.mybluemix.net/

AI FactSheets 360 http://aifs360.mybluemix.net/

Thank you

Prasanna Sattigeri
Research Staff Member
—
psattig@us.ibm.com