Lawrence Livermore National Laboratory

DSI AI in Healthcare Workshop

Building Reliable and Interpretable Clinical Models via Prediction Calibration

Jay Thiagarajan

Machine Intelligence Group | Center for Applied Scientific Computing

Collaborators


Vivek N (ASU)


Deepta Rajan (IBM AI)


Rushil Anirudh (LLNL)

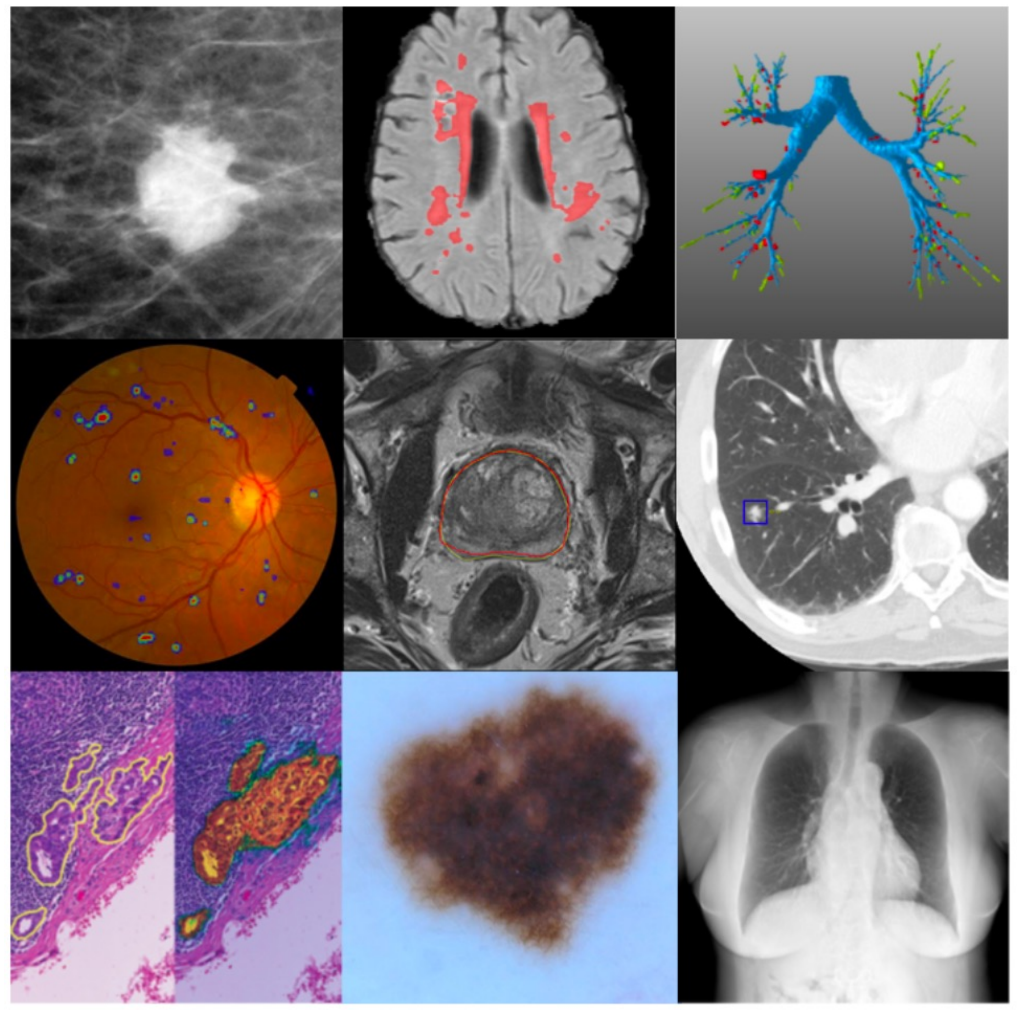

Akshay Chaudhari (Stanford)


Prasanna Sattigeri (IBM AI)

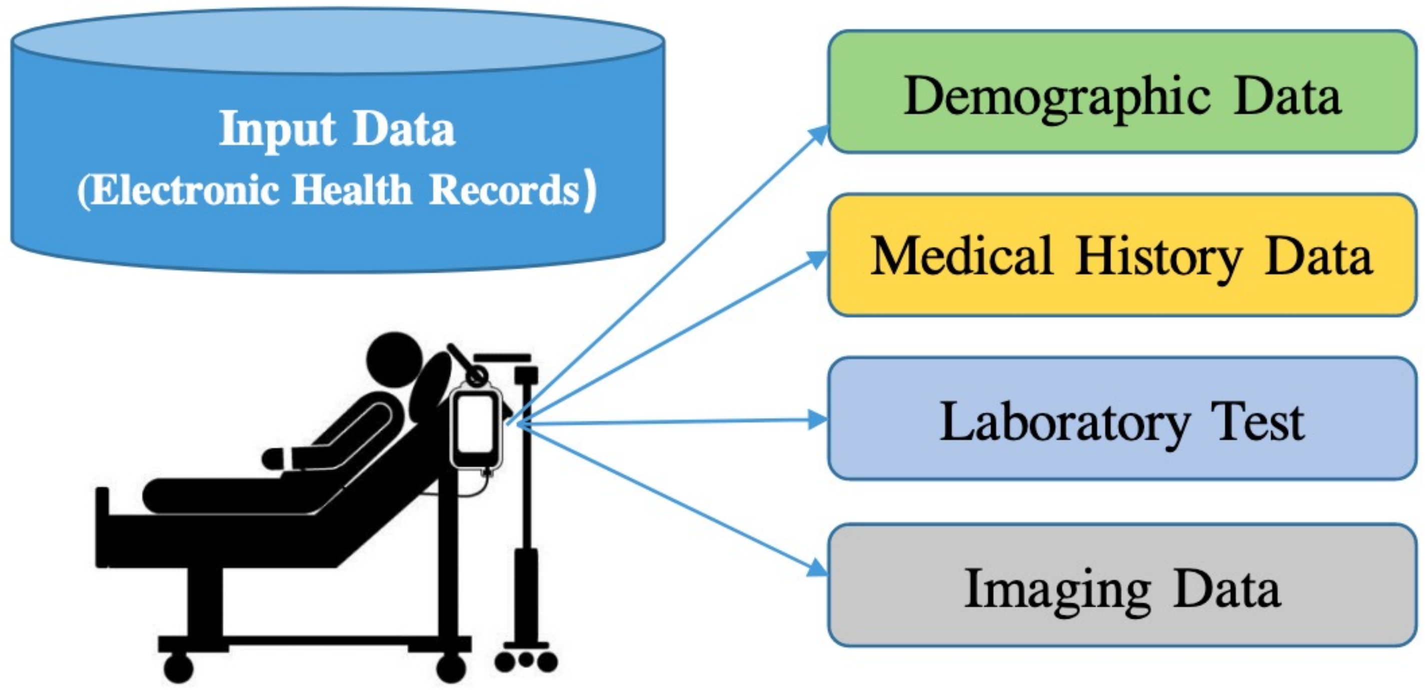

Andreas Spanias (ASU)

Machine Learning Methods are Becoming an Integral Part of Clinical Workflows

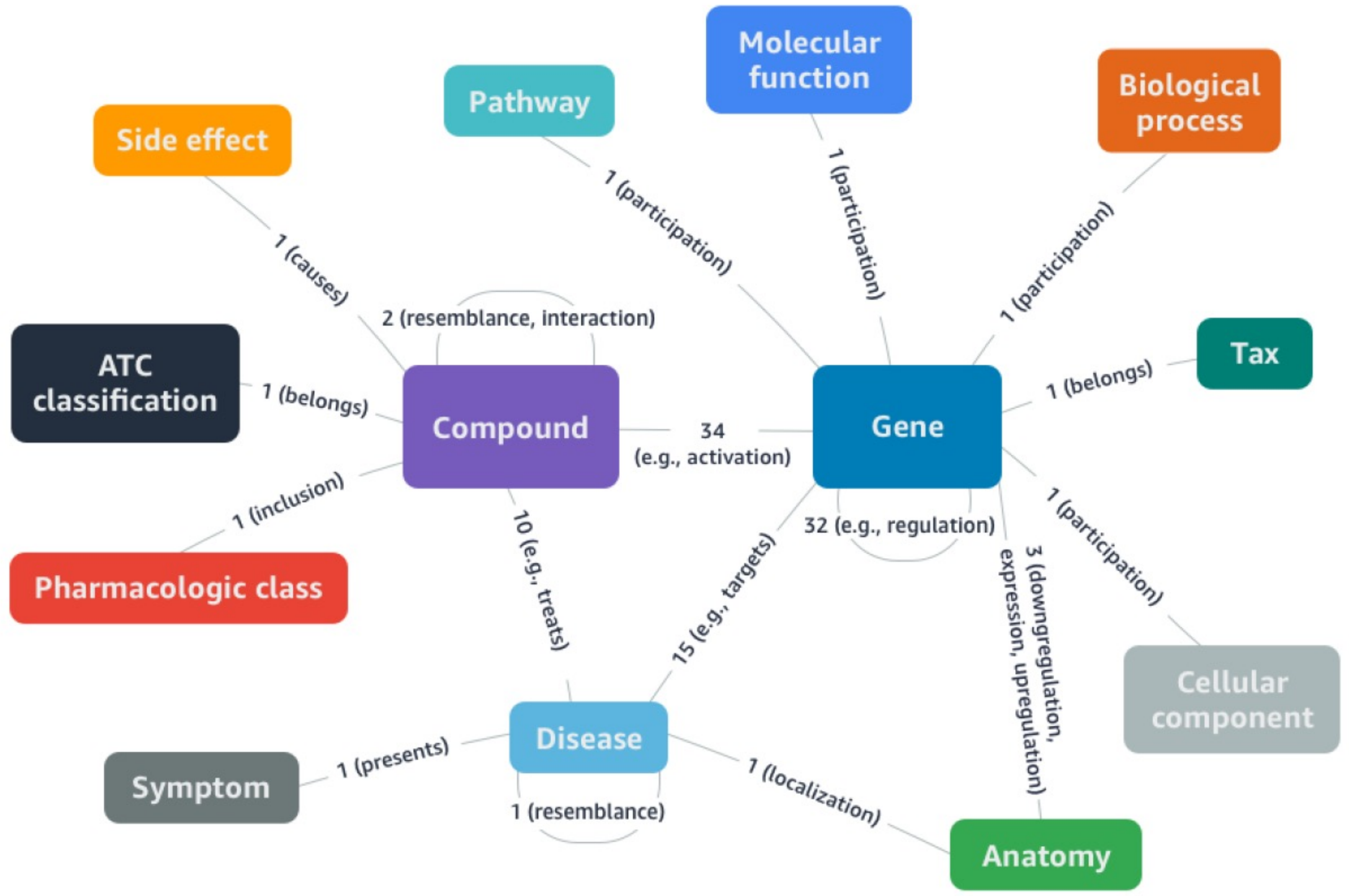
Diagnostic Tools



Patient Monitoring



Drug Discovery



Healthcare Management

Medical Data Security

Automated Workflow Assistance

Medical Risk Prediction

Virtual Nursing Assistants

AI in Healthcare: The Promise of Enabling Automation at Unprecedented Scales and Complexity

Building computational models for complex biological systems is extremely challenging – for example, disease evolution.

Big strides in adopting AI within clinical workflows:

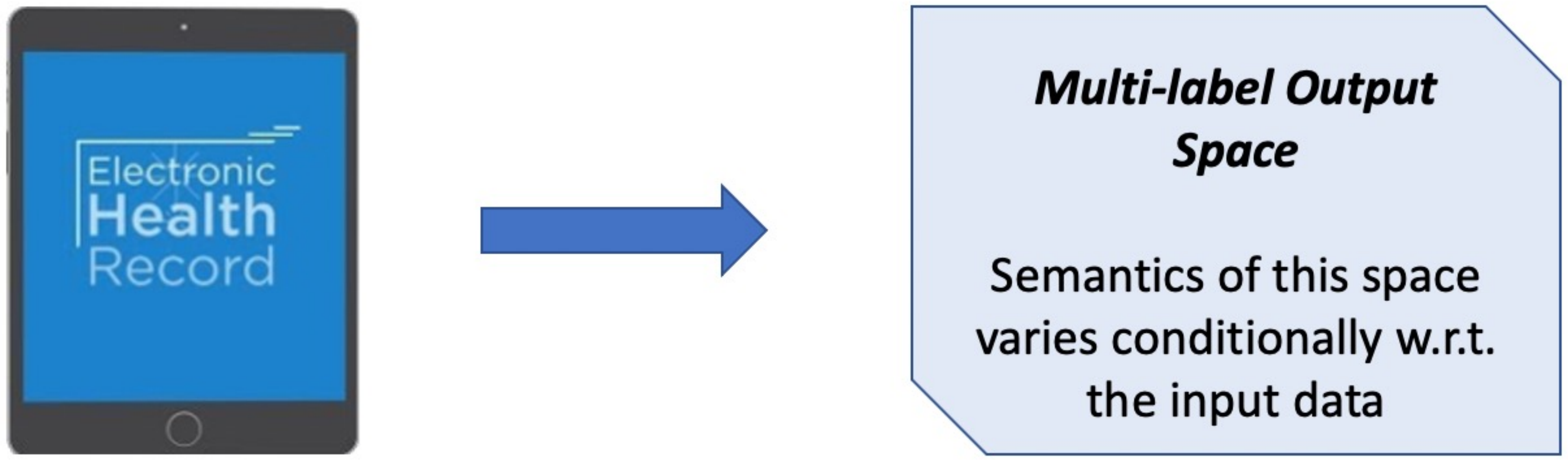
- Automate monotonous tasks.
- Digest heterogeneous data to make new hypotheses for improving patient care.
- Assist in clinical diagnosis and studying disease evolution.

A Case Study

Goal: Phenotyping from Electronic Health Records

Data: MIMIC-III benchmark with 76 measurements and 25 disease conditions

Model: Residual Networks with 1D-convolutions

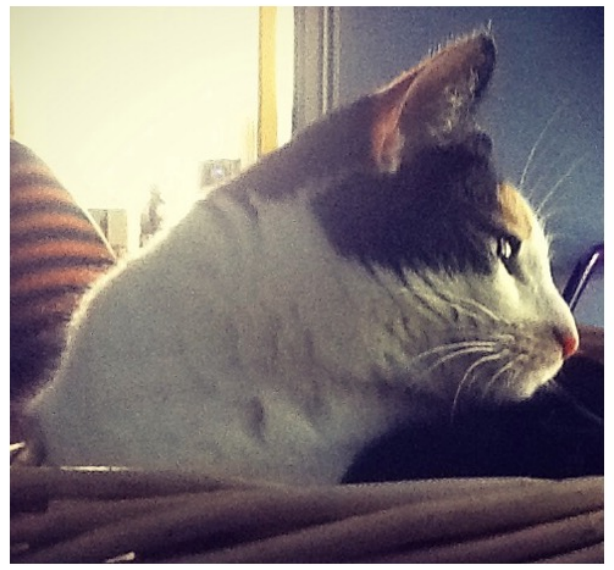


Time-varying measurements

By Design, Model Generalization is More Challenging in Clinical Diagnosis

$$\epsilon(\mathcal{D}_T; h) \leq \boxed{\mathcal{L}(\mathcal{D}_S; h)} + \boxed{\mathcal{L}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)} + \boxed{\mathcal{L}_{\delta}(h)}$$

Theoretical Limit on Expected Performance under Shifts



We Can Construct "Disease Landscapes" to Characterize the Complexity of the Task

Under different domain shifts, the task complexity changes!

Disease Landscape: An information-theoretic modeling of semantic dependencies in the outcome space

Each output is a random variable

$$TC(\tilde{X}) = \sum_{i=1}^{d} H(\tilde{X}_i) - H(\tilde{X})$$

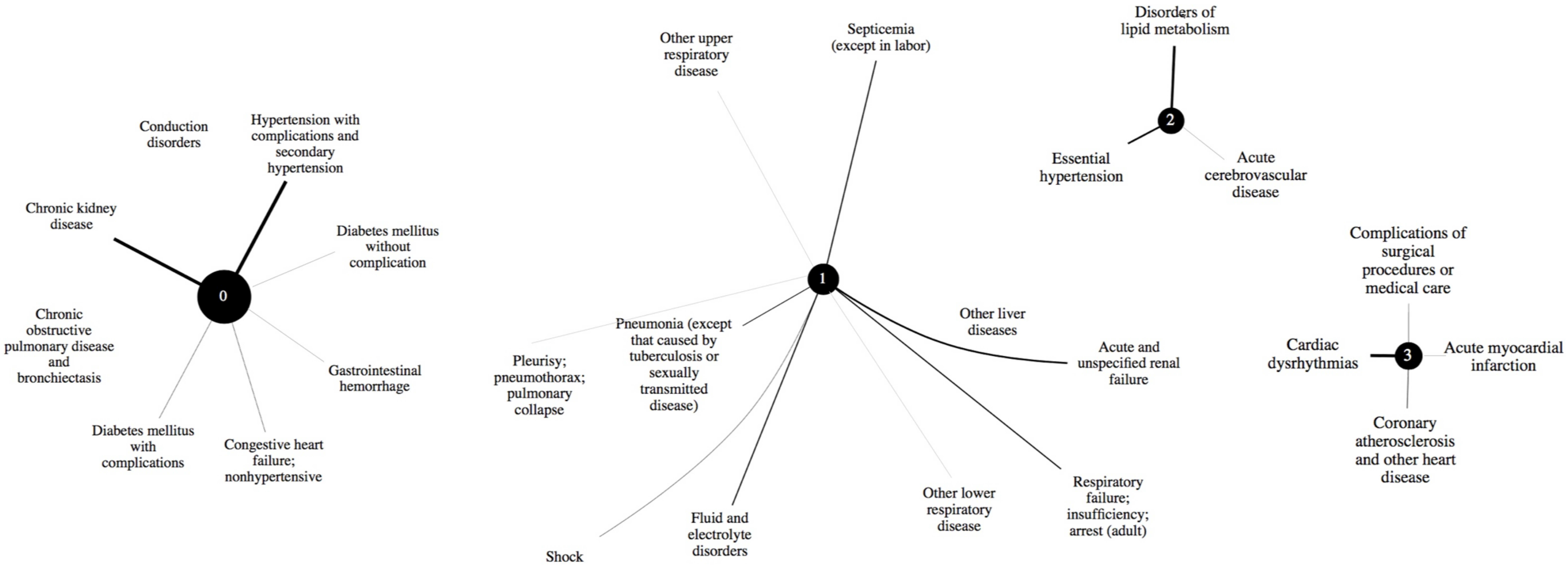
Marginal Entropy

Goal: Minimize $\boxed{TC(\tilde{X};\tilde{Z}) = TC(\tilde{X}) - TC(\tilde{X}|\tilde{Z})}$

Find latent factors that maximally describe total correlation between all disease conditions

* KDD DS Health 2019

We Can Construct "Disease Landscapes" to Characterize the Complexity of the Task



The Effect of Shifts on Clinical Models Can be Better Studied through Changes in Disease Landscapes

Population Biases

Age: a. Older (60+) to younger (<=60); b. Younger to older

Gender: a. 90%M-10%F to 10%M-90%F; b. 10%M-90%F to 90%M-10%F

Race: Whites to Minority

Label Distribution Shifts

Novel Diseases: a. [Resp] to [Resp + Renal + Cardiac]; b. [Cerebro] to [Cerebro + Renal + Cardiac]

Dual to Single: [Cardiac + Renal]

Single to Dual: [Cardiac], [Renal]

Measurement Discrepancies

Noisy Labels: [Resp] to [Resp + Renal + Cardiac], with 10% or 20% label flips

Sampling Rate Change: 96h to 48h window

Missing Meas.: pH, Temperature, Height, Weight, and all Verbal Response GCS

What Do We Find?

Deep clinical models can handle measurement discrepancies – No major changes in the disease landscape

Using markers learned for detecting certain abnormalities are descriptive enough to "extend" to other abnormalities – Extending landscapes with new latent factors

Learned markers do no generalize from patients presenting individual conditions to those with combinations – Changes to associations in the landscape

Population biases are the most challenging to handle – Landscapes with large degrees of change (different latent factors)

Moving towards the Design of "Reliable" Predictive Models

Architectures: Better priors on learnable functions (e.g., DDxNet for time-varying data)

Objectives: Suitable loss functions, leveraging priors, explainability by design

Training: Self-supervision, outlier exposure, consistency, adversarial training

Characterization: Uncertainty quantification, OOD detection, robustness under shifts

Uncertainty-Driven Characterization of Clinical Diagnosis Models

Epistemic Uncertainty attempts to answer the question – "Where in the data space is the model most likely to gain knowledge?"

Key question: Does the model "fake" knowledge when it should not know (unintended) and shows "lack" of knowledge when it should know (intended)?

For a well-calibrated uncertainty estimator, the *total uncertainty* of a model at a given input is the expected loss of the model

$$U(f;\mathrm{x}) = \int \ell(f(\mathrm{x},\mathrm{y}))dP(\mathrm{y}|\mathrm{x})$$

Uncertainty-Driven Characterization of Clinical Diagnosis Models

Epistemic Uncertainty attempts to answer the question – "Where in the data space is the model most likely to gain knowledge?"

Key question: Does the model "fake" knowledge when it should not know (unintended) and shows "lack" of knowledge when it should know (intended)?

Aleatoric uncertainty corresponds to the irreducible error – expected loss of a Bayes optimal predictor

$$A(f; \mathrm{x}) = U(f^*; \mathrm{x})$$

Uncertainty-Driven Characterization of Clinical Diagnosis Models

Epistemic Uncertainty attempts to answer the question – “Where in the data space is the model most likely to gain knowledge?”

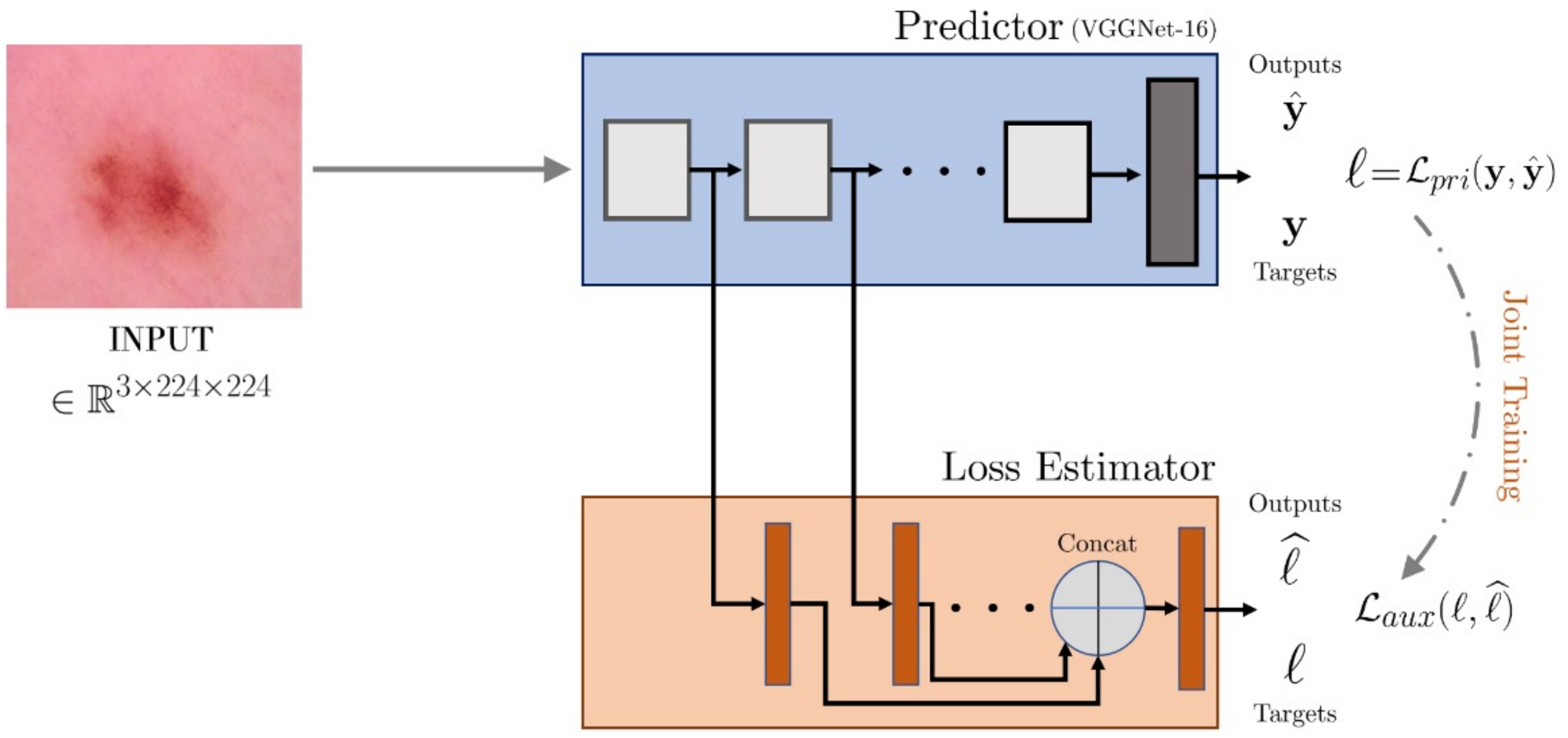
Key question: Does the model “fake” knowledge when it should not know (unintended) and shows “lack” of knowledge when it should know (intended)?

Epistemic uncertainty can be defined as the gap between generalization and the irreducible error of a model

$$E(f; \mathrm{x}) = U(f; \mathrm{x}) - U(f^*; \mathrm{x})$$

A Well-Calibrated Uncertainty Predictor Can be Designed by Learning to Directly Predict the Generalization Error

Build an auxiliary uncertainty predictor that directly matches the generalization error – this estimator can be used even for test data.

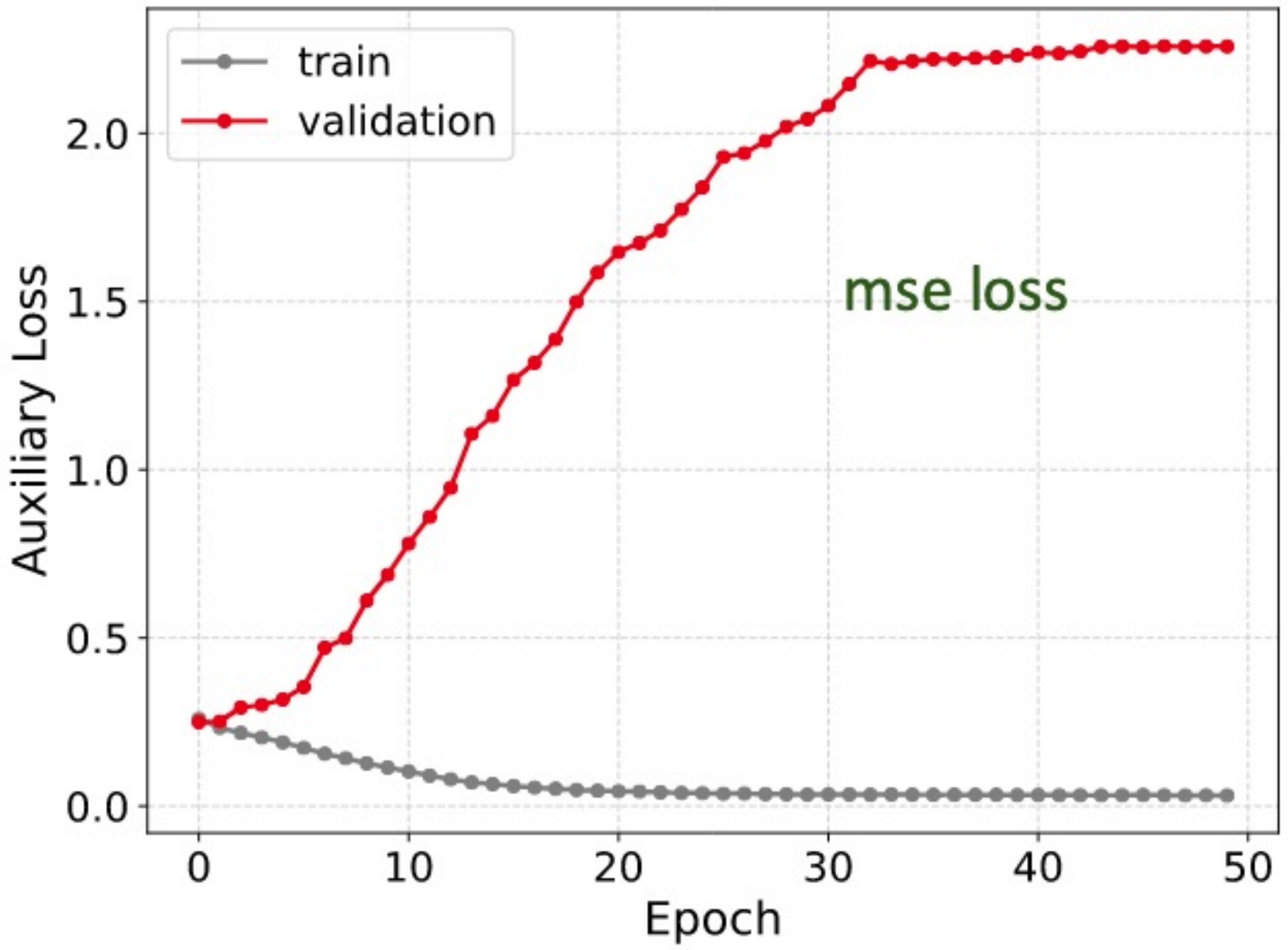


Joint training of an uncertainty estimator inherently regularizes the predictive model

* AAAI 2021, arXiv:2103.03788

How Do We Train the Uncertainty Predictor?

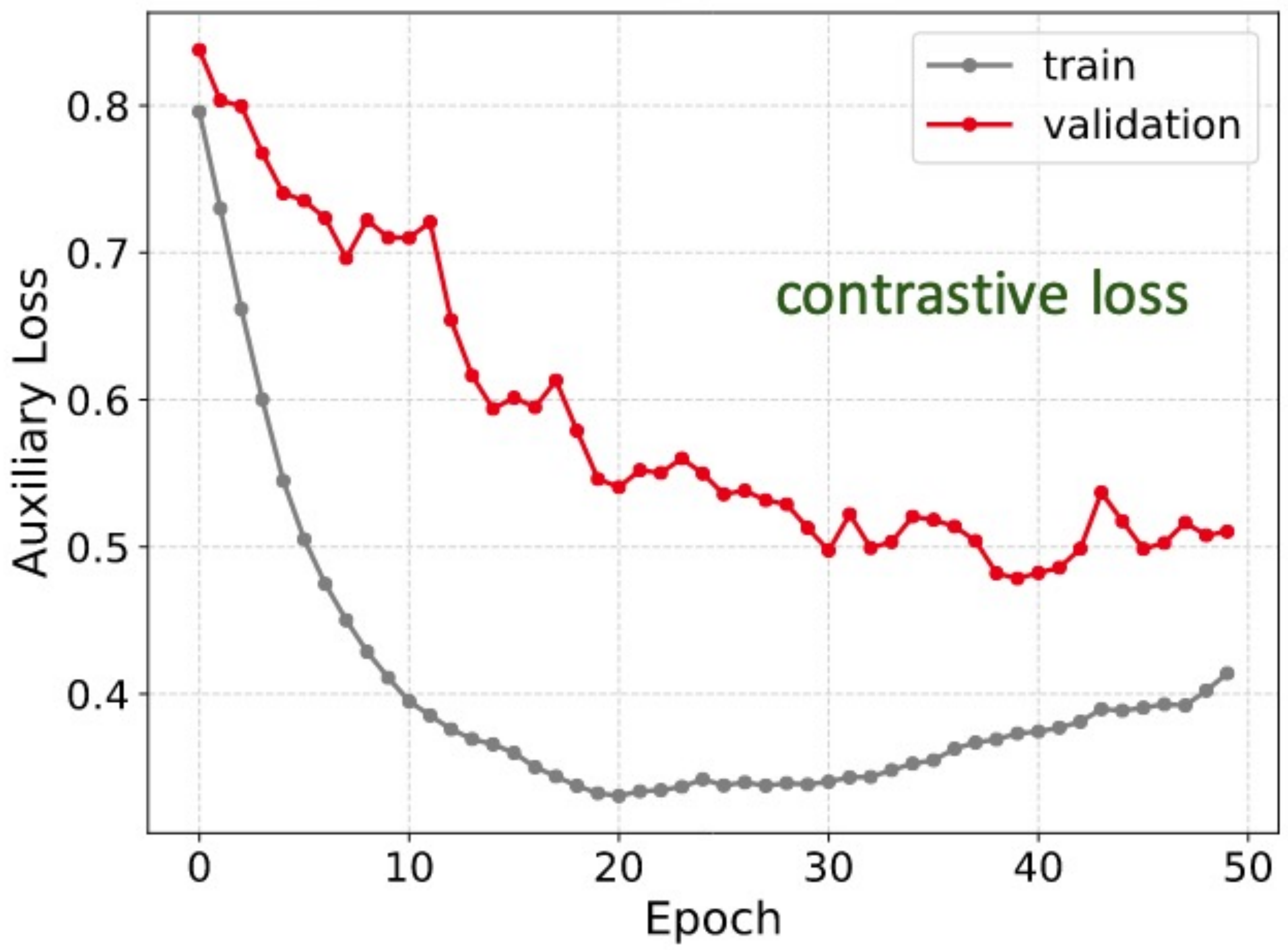
Can we directly match the loss value for each sample in the batch?



No. the resulting uncertainty predictor does not generalize

How Do We Train the Uncertainty Predictor?

A contrastive training strategy to produce generalizable uncertainty predictors



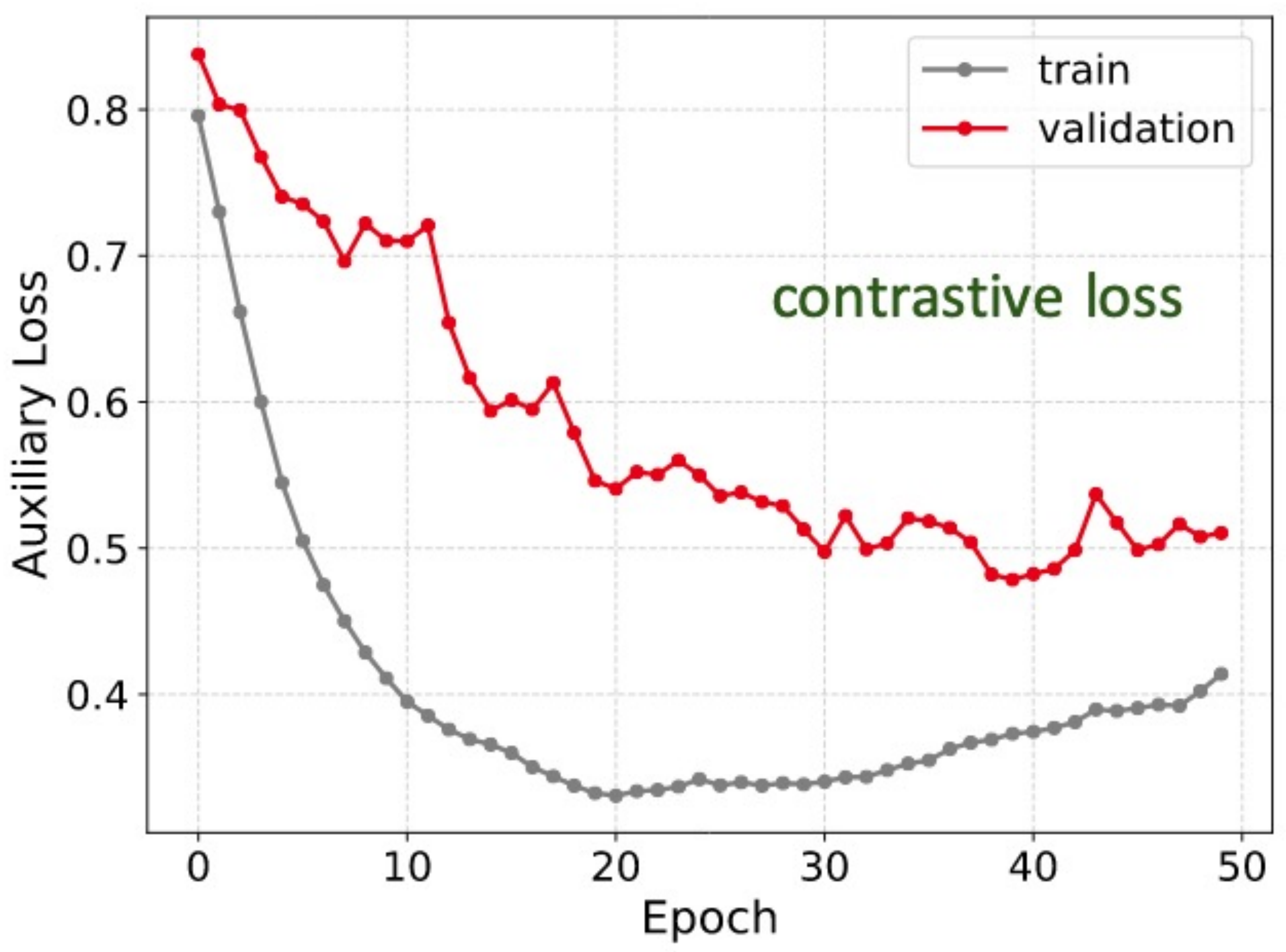
Preserve the order of samples based on their loss values

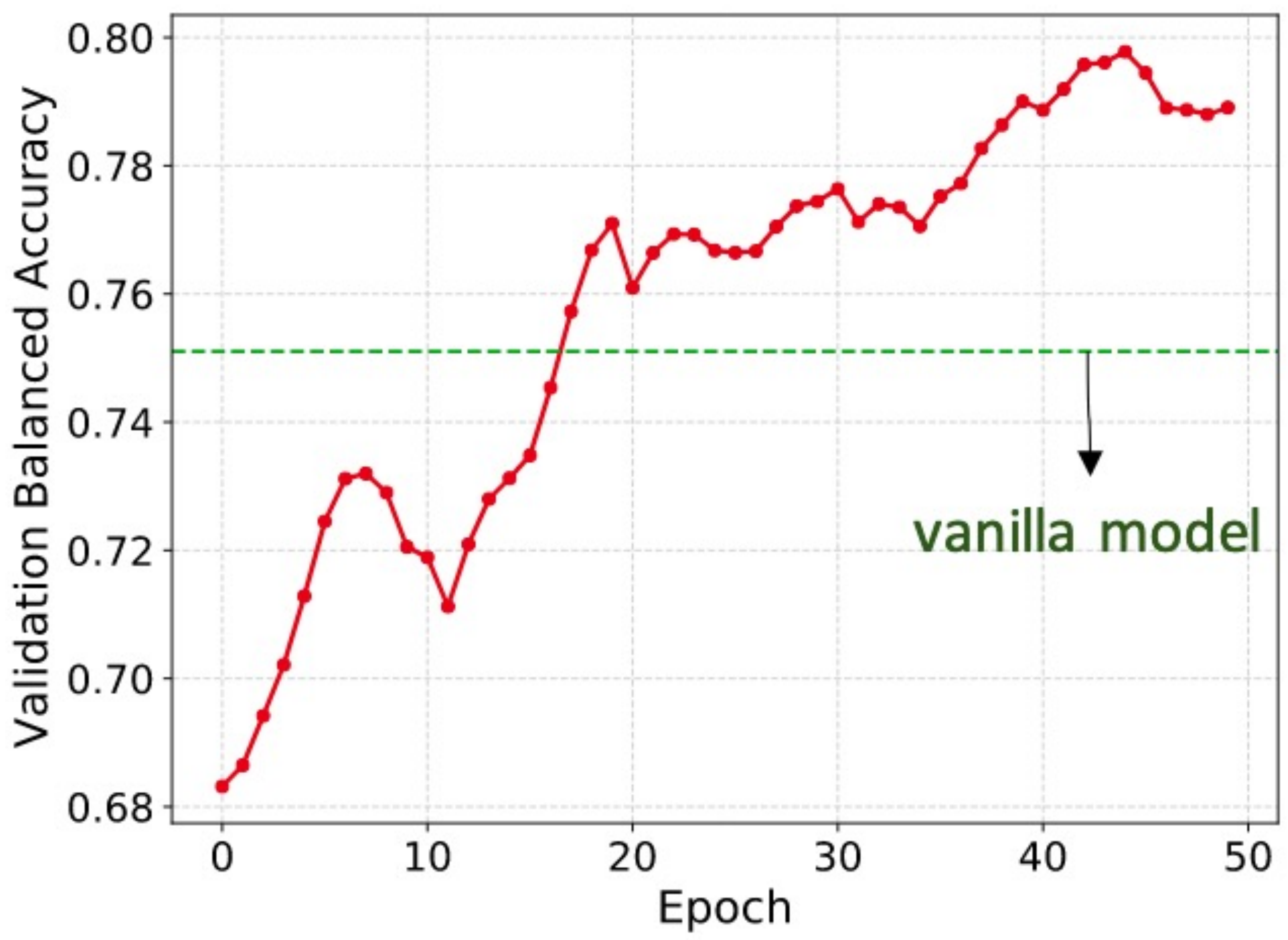
$$\sum_{(i,j)} \max\left(0, -\mathbb{I}(\ell_i, \ell_j).(\hat{\ell}_i - \hat{\ell}_j) + \gamma\right),$$

$$\text{where } \mathbb{I}(\ell_i, \ell_j) = \begin{cases} 1, & \text{if } \ell_i > \ell_j, \\ -1, & \text{otherwise.} \end{cases}$$

Interestingly, this Self-Calibration Process Regularizes the Predictive Model

A contrastive training strategy to produce generalizable uncertainty predictors

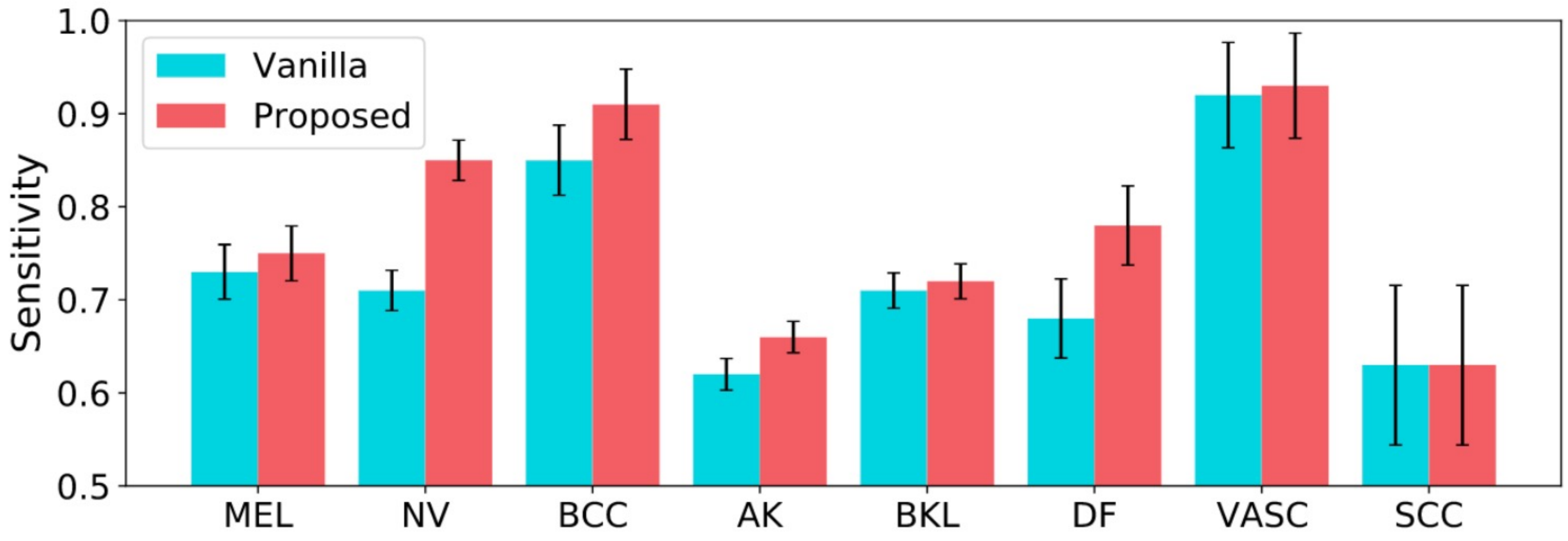




Improved Generalization in "Intended Regimes"

Goal: Skin Lesion Type Detection using Dermoscopy images

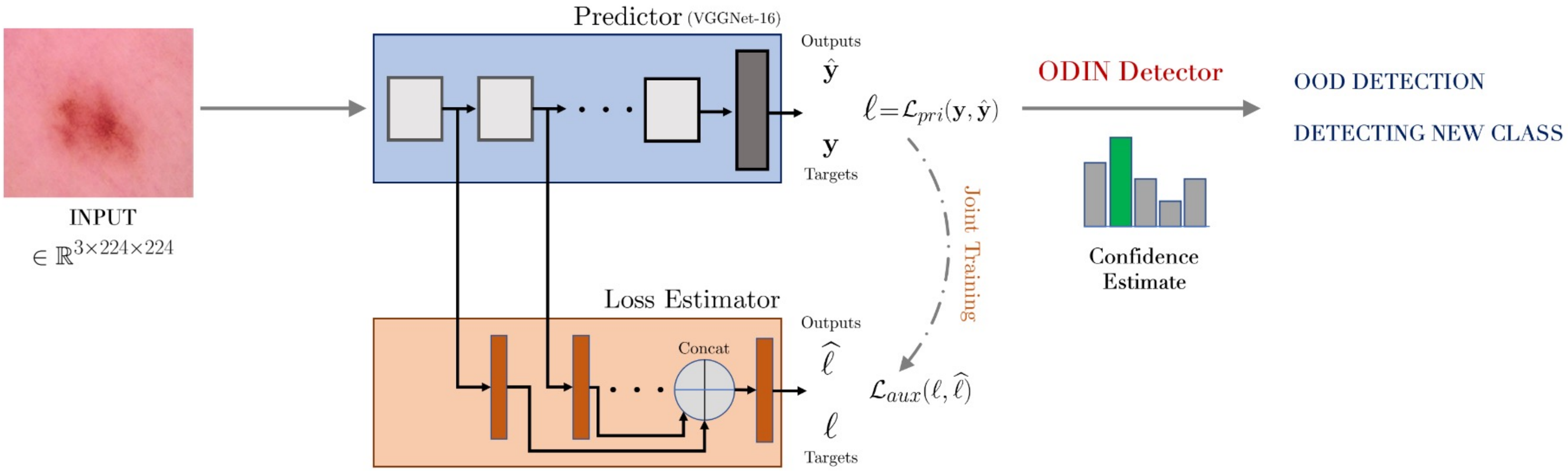
Data: ISIC 2019 Challenge Dataset



Balanced Accuracy (%)

Vanilla	Proposed
73.1 +/- 0.6	**77.9 +/- 1.5**

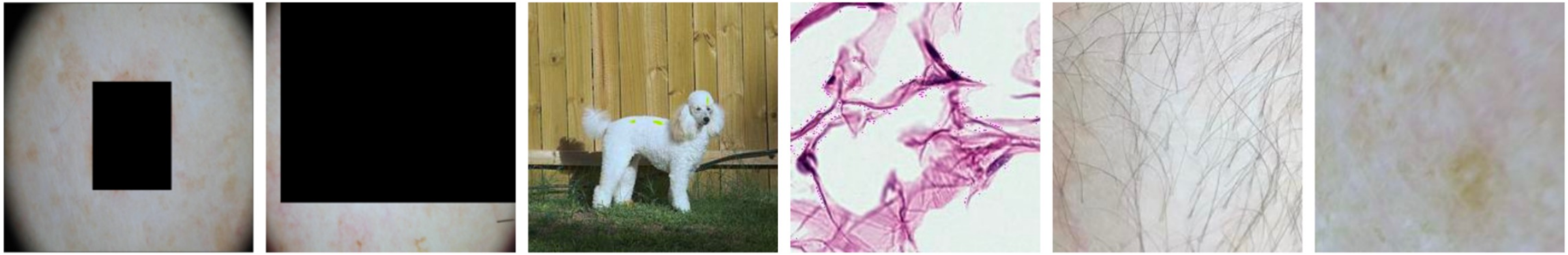
Controlled Generalization in "Unintended Regimes"



Controlled Generalization in "Unintended Regimes"

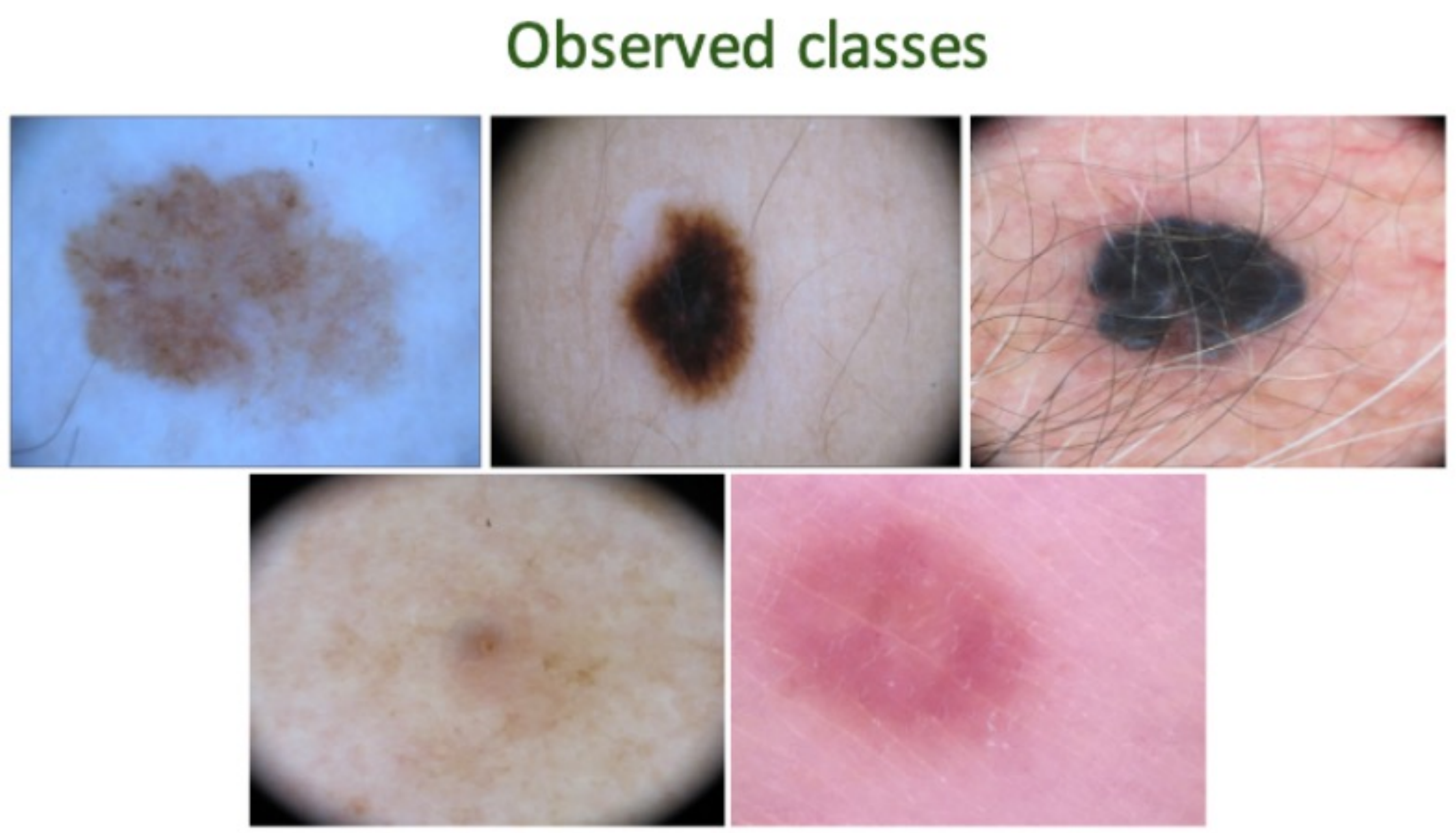
AUROC Metric

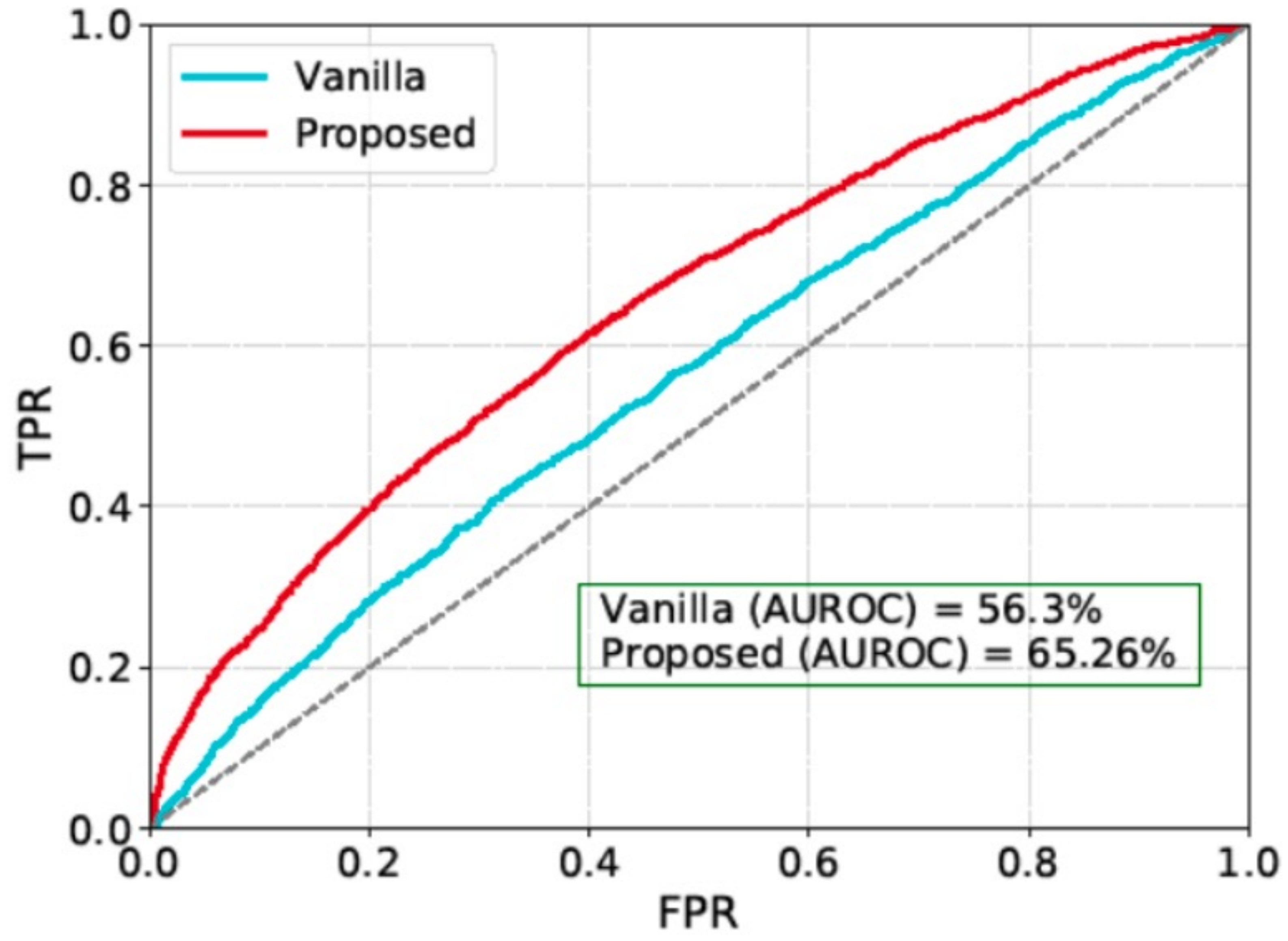
(a)	(b)	(c)	(d)	(e)	(f)
90.74 / **98.74**	**100 / 100**	79.03 / **87.57**	87.27 / **96.43**	82.58 / **92.15**	87.14 / **95.69**



(a) B-Box (b) B-Box-70 (c) Imagenet (d) NCT (e) Clin Skin (f) Derm Skin

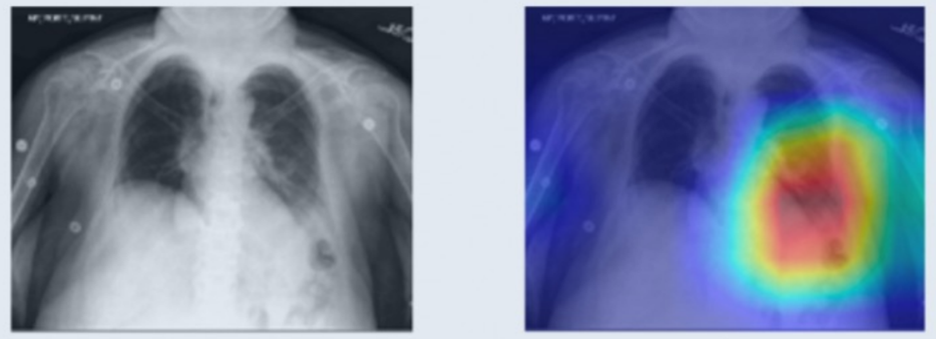
Controlled Generalization in "Unintended Regimes"





Explainable AI Methods are Routinely used to Validate Model Behavior and Shed Light into its Vulnerabilities

Local Explanations



What features in the given image explain the prediction?

Approximation (LIME, ANCHORS):
Sample in the vicinity of the input and fit an explainer

Example-Based:
Search in the vicinity such that prediction remains the same or **<u>changes</u>**

Why did the model make a specific diagnosis for a given subject ?

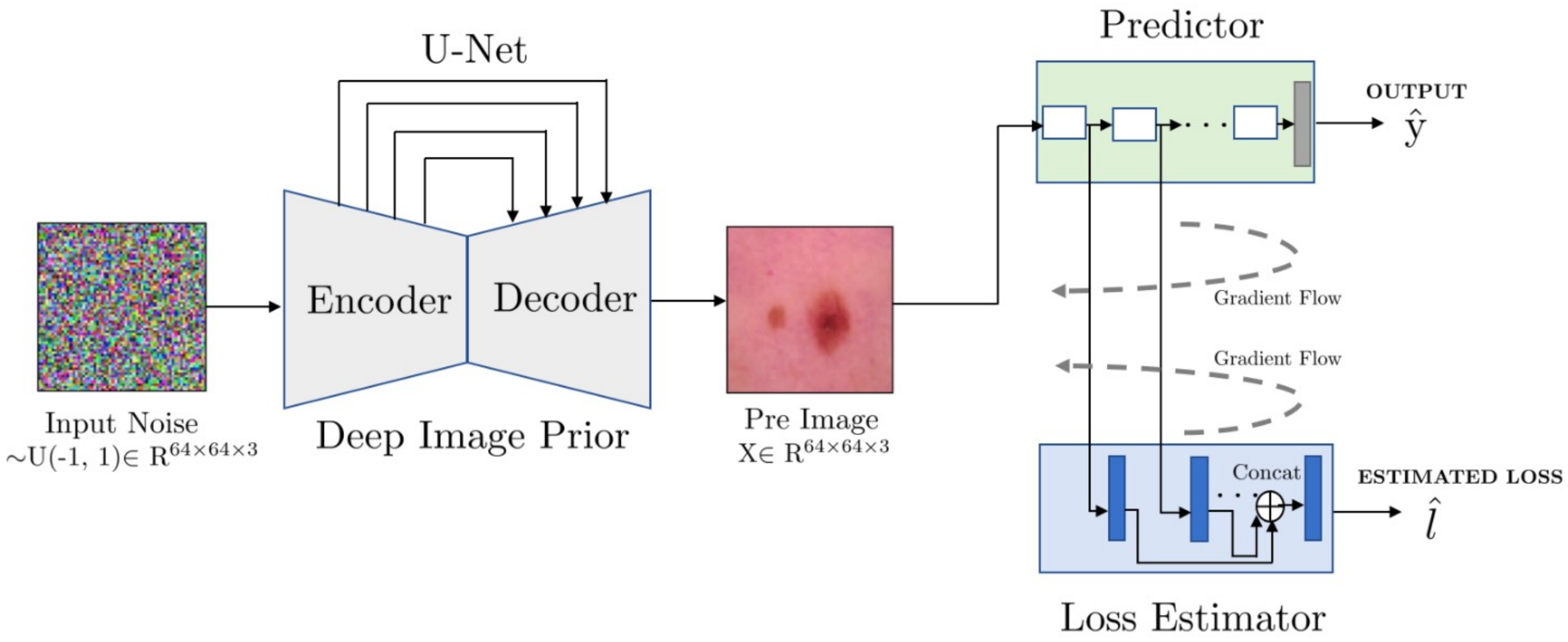
How should the data signatures change for a different prediction?

Counterfactual Analysis Allows for Exploratory Investigation of Learned Models

$$\arg\min_{\bar{\mathrm{x}}} \underbrace{d(\mathrm{x}, \bar{\mathrm{x}})}_{\text{Amount of change}} \quad \text{s.t.} \quad \mathcal{F}(\underbrace{\bar{\mathrm{x}}}_{\text{Counterfactual}}) = \underbrace{\bar{\mathrm{y}}}_{\text{Desired target}}$$

Key requirement: Synthesized counterfactuals must belong to the true data distribution

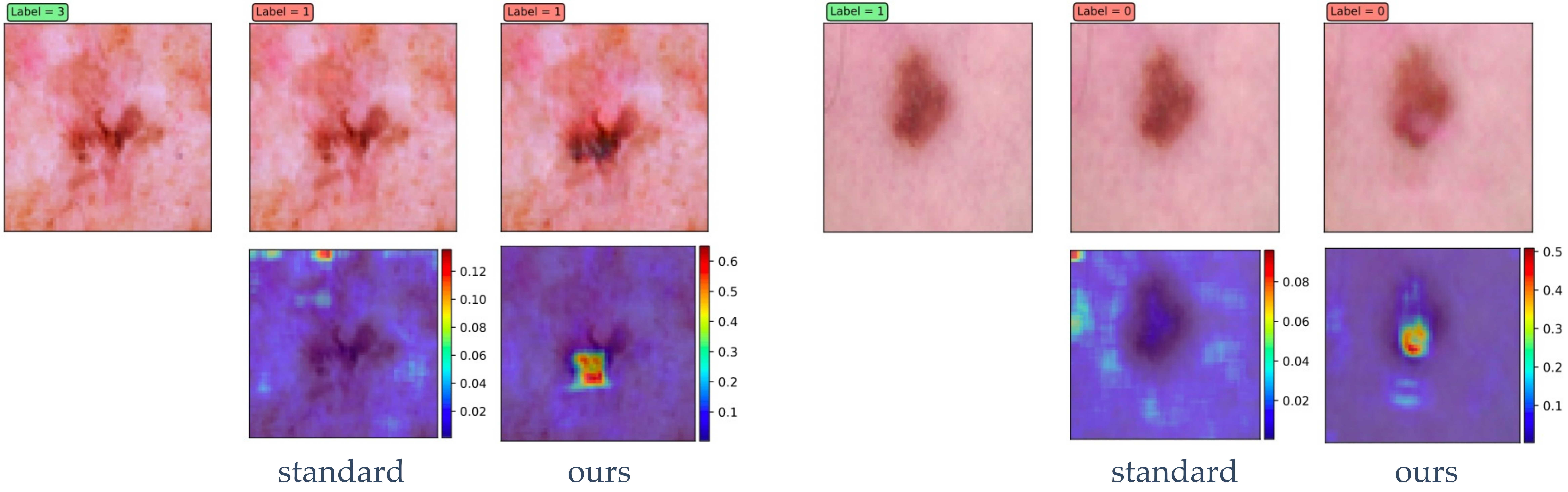
Loss Estimators can be used to Construct a Hypothesis Test for Data Consistency



$$\mathbf{x}^* = arg \min_{\mathbf{x}\in\mathbb{R}^{H\times W\times C}} \mathcal{L}(\Psi(\mathbf{x}), \Psi(\mathbf{x}_0)) + \lambda_1 \mathcal{M}(\mathbf{x})$$

$$+\lambda_2 \mathcal{L}_{CE}(\mathrm{y}, \mathrm{y}_t).$$

With the Ability to Better Characterize "In-Distribution" Data, We Can Effectively Explore the Data Space



* arXiv:2010.12046

Moving towards the Design of "Reliable" Predictive Models

Architectures: Better priors on learnable functions, ease of training, efficiency

Objectives: Suitable loss functions, leveraging priors, explainability by design

Training: Self-supervision, outlier exposure, semi-supervised learning, adversarial training

Characterization: Uncertainty quantification, OOD detection, robustness under shifts

Iterative Design with Benchmarks, Experts-in-the-loop and Rigorous Evaluation Methodologies

Related Publications

[1] Loss Estimators Improve Model Generalization, arXiv:2103.03788 (preprint).

[2] Using Deep Image Priors to Generate Counterfactual Explanations, arXiv:2010.12046 (preprint).

[3] Accurate and Robust Feature Importance Estimation under Distribution Shifts, *AAAI* 2021.

[4] DDxNet: a deep learning model for automatic interpretation of electronic health records, electrocardiograms and electroencephalograms, *Nature Scientific Reports* 2020.

[5] Understanding Behavior of Clinical Models under Domain Shifts, *KDD DS-Health Workshop*, 2019.

jjayaram@llnl.gov https://jjthiagarajan.com/ jjayaram7