

What is Your Metric Telling You?

Evaluating Classifier Calibration under Different Definitions of Reliability

Eric Heim, Senior Machine Learning Researcher
AI Division, Software Engineering Institute, Carnegie Mellon University

Copyright 2021 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM21-0848

An Argument for Uncertainty in Machine Learned Models



An Argument for Uncertainty in Machine Learned Models

Predicted Class
Civilian Vehicle



An Argument for Uncertainty in Machine Learned Models

Predicted Class	Confidence
Civilian Vehicle	0.95



An Argument for Uncertainty in Machine Learned Models

Predicted Class	Confidence
Civilian Vehicle	0.6



An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
...	...



An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
...	...

Confidence/Uncertainty
can lead to more
informed decision-
making.



An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
...	...

if conf("Enemy Tank")
> 0.25

Alert!

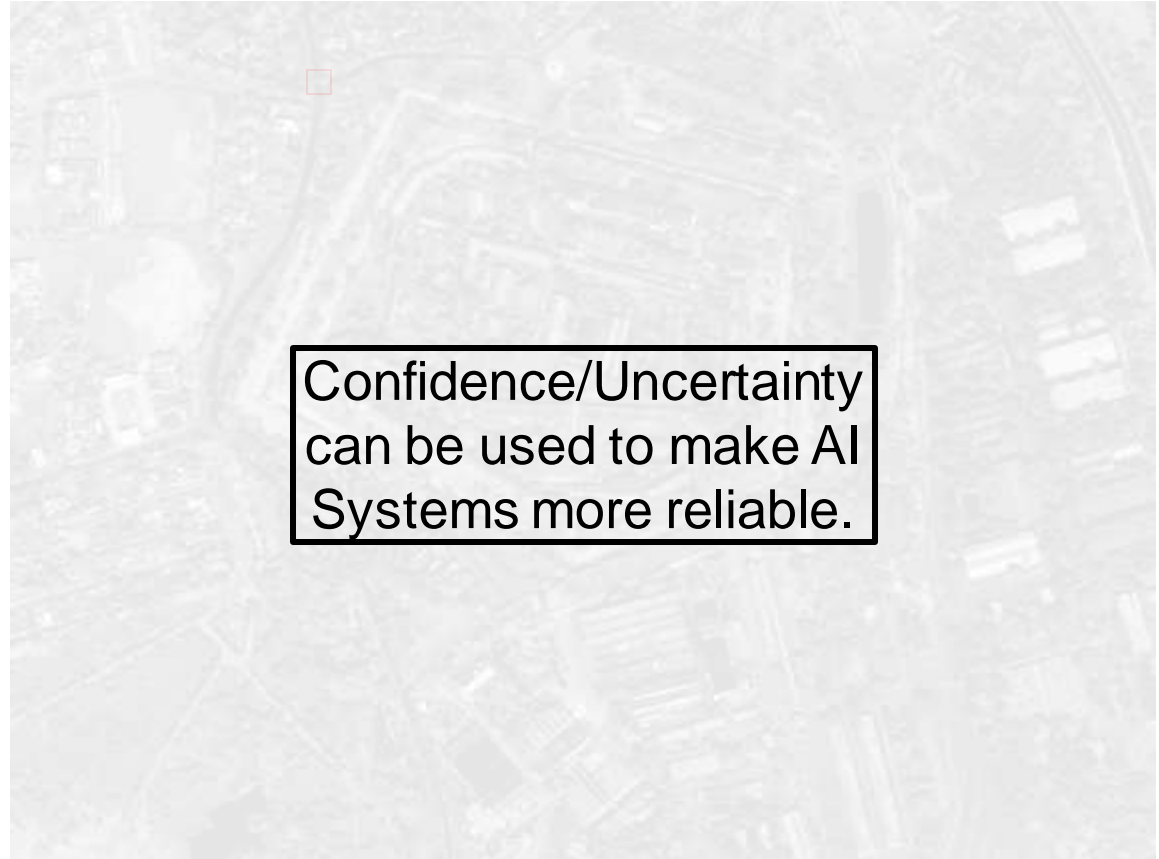


An Argument for Uncertainty in Machine Learned Models

Class	Confidence
Civilian Vehicle	0.6
Enemy Tank	0.35
...	...

if conf("Enemy Tank")
> 0.25

Alert!



This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.

This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.



This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for its ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.



How do we understand these values?

This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.



Classifier Calibration: Classifier outputs match the frequency of class labels.

This work: Evaluating classifiers for context-specific calibration

How do you evaluate your classifier for it's ability to accurately express uncertainty?

First, you need to define how to interpret the confidence outputs of classifiers.



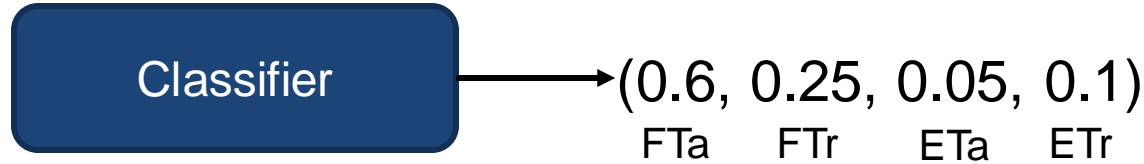
For **all possible inputs** that the **classifier** outputs $(0.6, 0.4)$...
60% of the inputs should be a Civilian Vehicle,
40% of the inputs should be an Enemy Tank.

Classifier Calibration: Classifier outputs match the frequency of class labels.

This work: Evaluating classifiers for context-specific calibration

Modern machine learning literature has focused on evaluating classifier calibration according to their **Top-1 Expected Calibration Error (ECE)** (Guo et al; 2017)

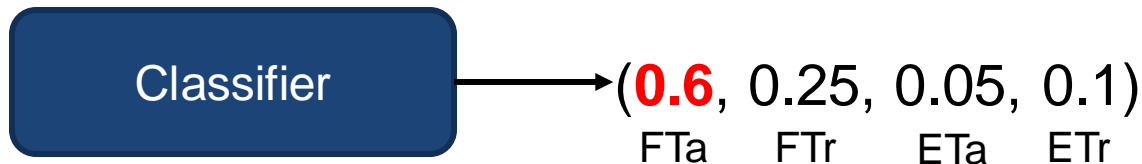
Classes = {Friendly Tank, Friendly Truck, Enemy Tank, Enemy Truck}



This work: Evaluating classifiers for context-specific calibration

Modern machine learning literature has focused on evaluating classifier calibration according to their **Top-1 Expected Calibration Error (ECE)** (Guo et al; 2017)

Classes = {Friendly Tank, Friendly Truck, Enemy Tank, Enemy Truck}



Top-1 Expected Calibration Error (ECE)

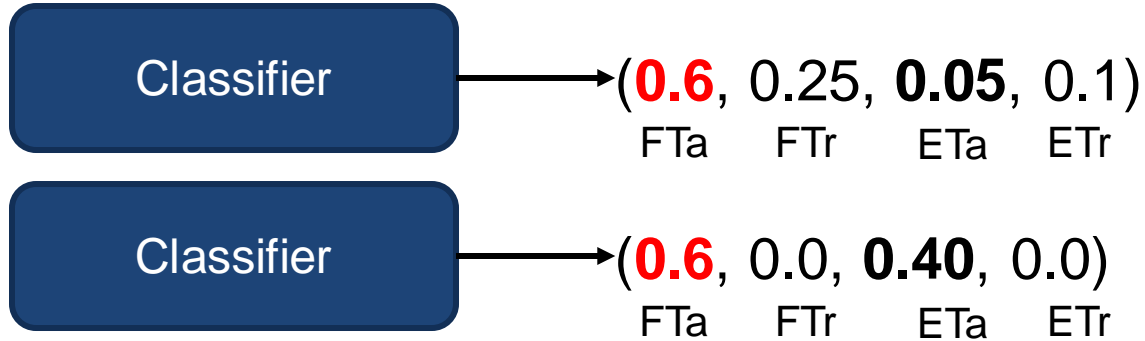
Considers only the most confident class in evaluating for calibration

For **all possible inputs** that the **classifier** outputs **0.6 as the most confident class**...
60% of the those inputs should be that class.

This work: Evaluating classifiers for context-specific calibration

Modern machine learning literature has focused on evaluating classifier calibration according to their **Top-1 Expected Calibration Error (ECE)** (Guo et al; 2017)

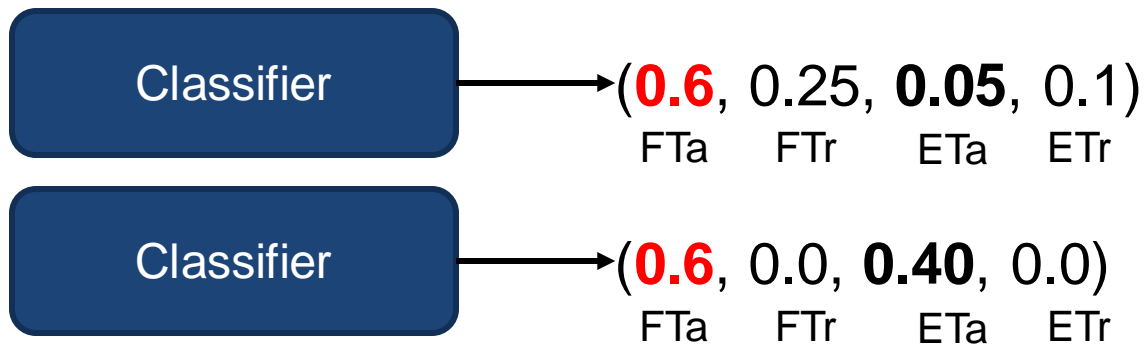
Classes = {Friendly Tank, Friendly Truck, Enemy Tank, Enemy Truck}



This work: Evaluating classifiers for context-specific calibration

Modern machine learning literature has focused on evaluating classifier calibration according to their **Top-1 Expected Calibration Error (ECE)** (Guo et al; 2017)

Classes = {Friendly Tank, Friendly Truck, Enemy Tank, Enemy Truck}



According to Top-1 ECE, these two classifiers are considered the same. However, the two outputs can mean very different things with mission context.

This motivates the need for calibration metrics that are able to map to contexts and use cases that match realistic usage of deployment settings.

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Canonical Calibration(Reliability) Condition

$$\mathbb{P}[Y \in \cdot \mid g(X)] = g(X)$$

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Canonical Calibration(Reliability) Condition

$$\mathbb{P}[Y \in \cdot \mid g(X)] = g(X)$$

$$\mathbb{P}[Y \in \cdot \mid g(X) = (0.7, 0.1, 0.2)] = (0.7, 0.1, 0.2)$$

Of all the times g outputs $(0.7, 0.1, 0.2)$...

... 70% of the time it was class A

... 10% of the time it was class B

... 20% of the time it was class C

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Canonical Calibration(Reliability) Condition

$$\mathbb{P}[Y \in \cdot \mid g(X)] = g(X)$$

“Lens”

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Top-1 Reliability Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X) = 0.7] = 0.7$$

Of all the times g 's highest confidence is 0.7

... 70% of the time it was the class with the highest confidence

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Top-1 Reliability Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X) = 0.7] = 0.7$$

Of all the times g 's highest confidence is 0.7

... 70% of the time it was the class with the highest confidence

Informally, a lens is a transformation of 1) The classifier's output 2) The probability simplex over classes that results in an induced classification problem.

A statistical framework for context-specific ECE

We adopt the framing of ECE from (Vaicenavicius et al; 2019)

Top-1 Reliability Condition

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$$

$$\mathbb{P}[Y = \operatorname{argmax} g(X) \mid \max g(X) = 0.7] = 0.7$$

Of all the times g' 's highest confidence is 0.7

... 70% of the time it was the class with the highest confidence

Informally, a lens is a transformation of 1) The classifier's output 2) The probability simplex over classes that results in an induced classification problem.

By designing lensing functions that result induced classification problems that map to specific mission contexts, you can evaluate classifiers in more application-specific ways.

A statistical framework for context-specific ECE

Key considerations for designing a context-specific ECE metrics:

1. Lensing functions: o, t

A statistical framework for context-specific ECE

Key considerations for designing a context-specific ECE metrics:

1. Lensing functions: o, t
2. Distance metric to compare lenses classifier outputs and class-probabilities.

$$d(o(g(x)), t(y))$$

A statistical framework for context-specific ECE

Key considerations for designing a context-specific ECE metrics:

1. Lensing functions: o, t
2. Distance metric to compare lenses classifier outputs and class-probabilities.

$$d(o(g(x)), t(y))$$

3. Selection operators

$$\lambda(\mathcal{D}) \subseteq \mathcal{D}$$

A statistical framework for context-specific ECE

Key considerations for designing a context-specific ECE metrics:

1. Lensing functions: o, t
2. Distance metric to compare lenses classifier outputs and class-probabilities.

$$d(o(g(x)), t(y))$$

3. Selection operators

$$\lambda(\mathcal{D}) \subseteq \mathcal{D}$$

4. Estimation of expectation over data domain

$$\mathbb{E}_{x,y \sim \lambda(\mathcal{D})} d(o(g(x)), t(y))$$

A statistical framework for context-specific ECE

Key considerations for designing a context-specific ECE metrics:

1. Lensing functions: o, t
2. Distance metric to compare lenses classifier outputs and class-probabilities.

$$d(o(g(x)), t(y))$$

3. Selection operators

$$\lambda(\mathcal{D}) \subseteq \mathcal{D}$$

4. Estimation of expectation over data domain

$$\mathbb{E}_{x,y \sim \lambda(\mathcal{D})} d(o(g(x)), t(y))$$

Different choices for each of these result in a different, context-specific ECE metric

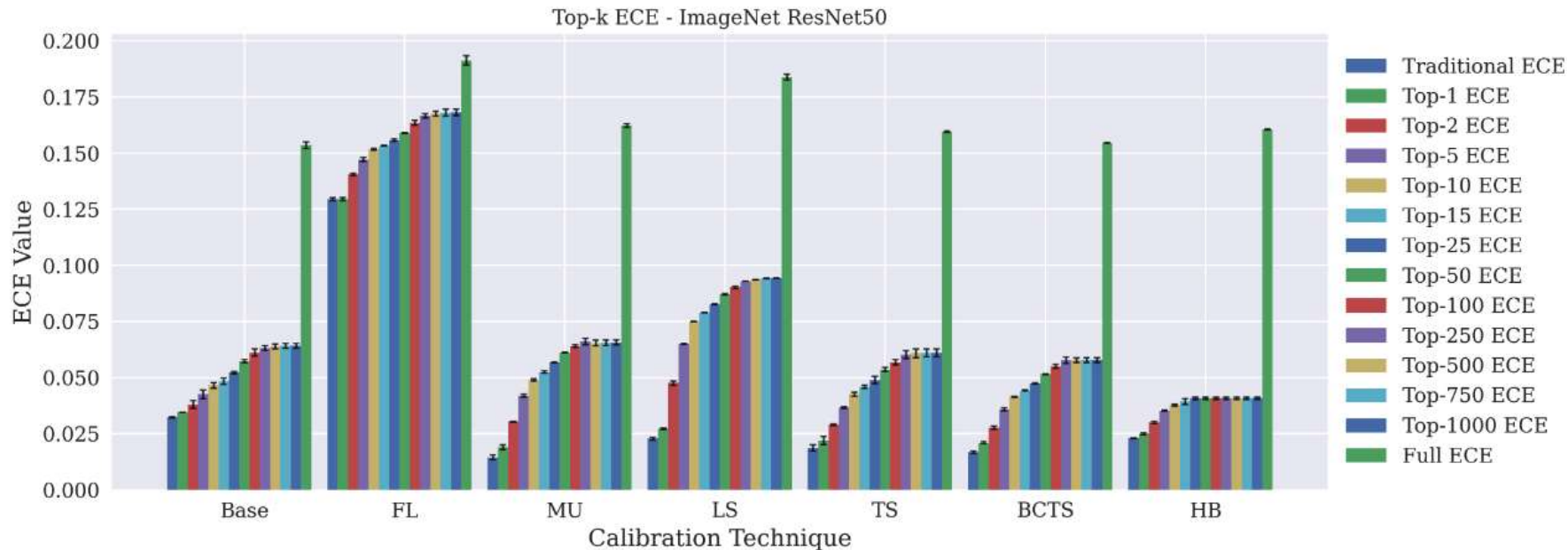
Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

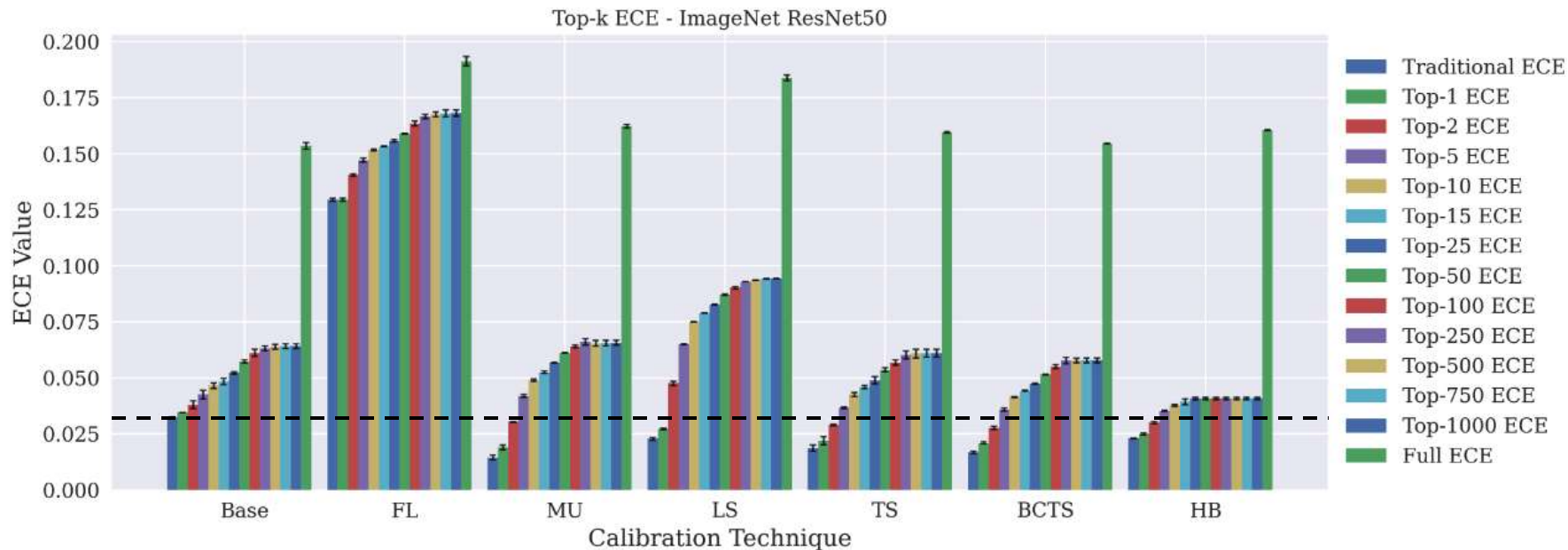
Experiment #1: Top- k ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

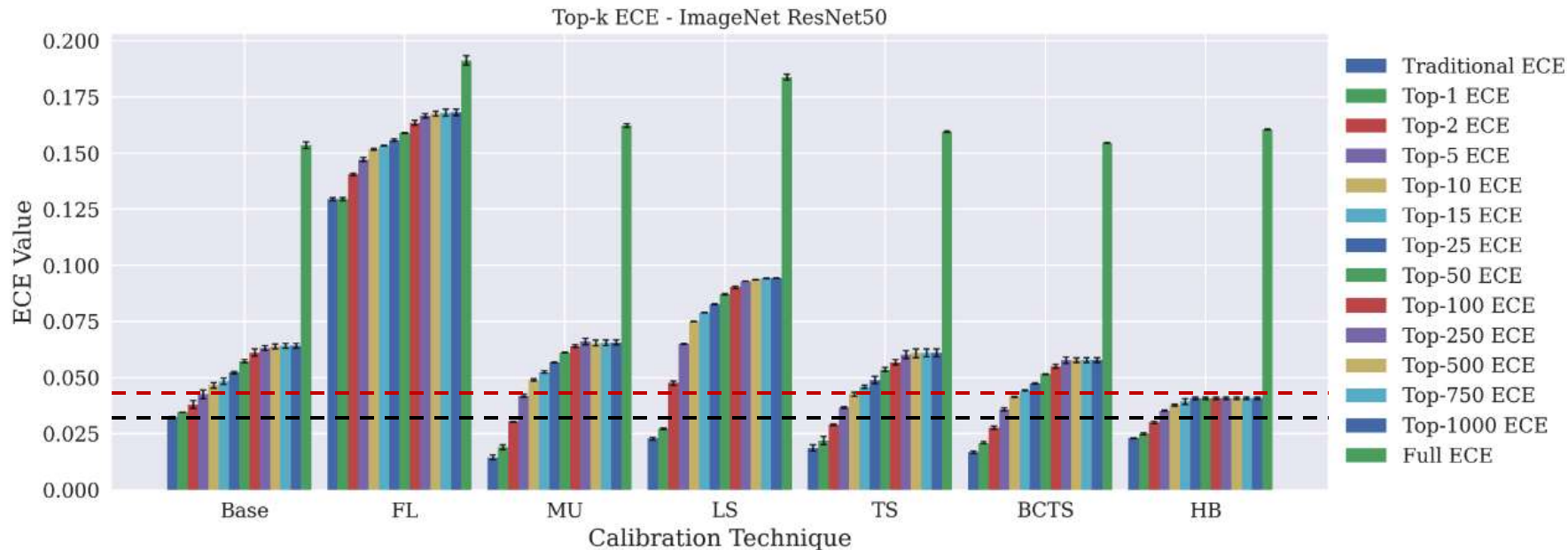
Experiment #1: Top- k ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

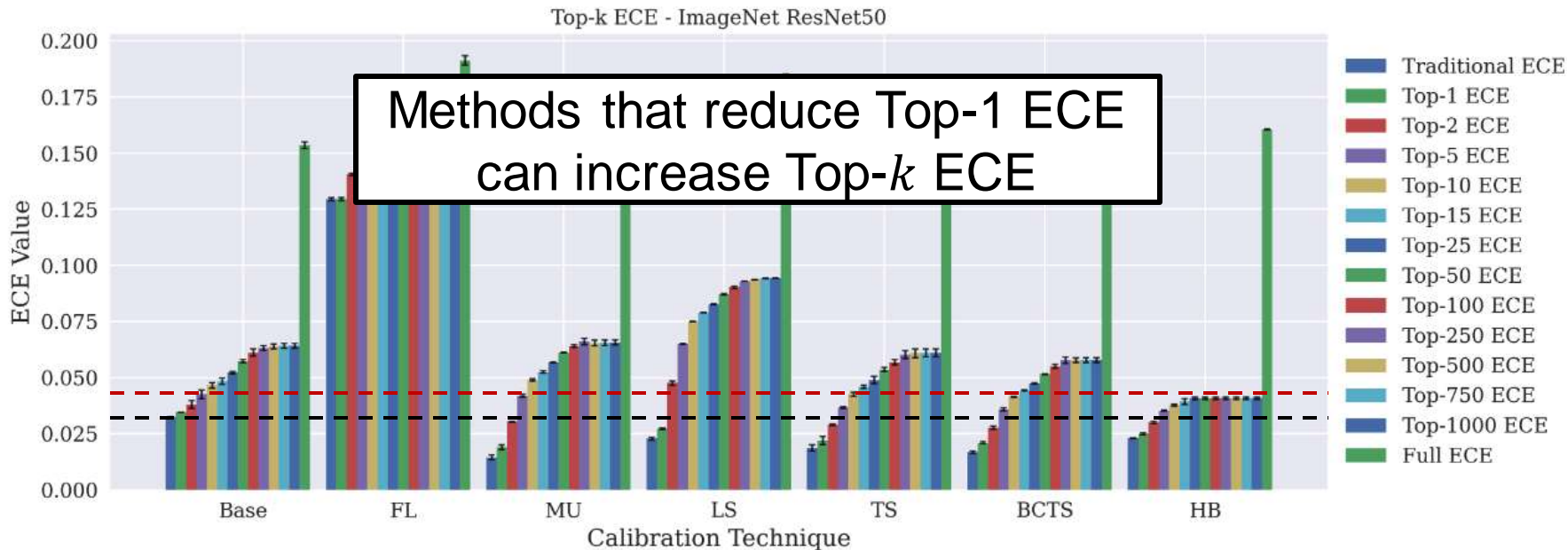
Experiment #1: Top- k ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

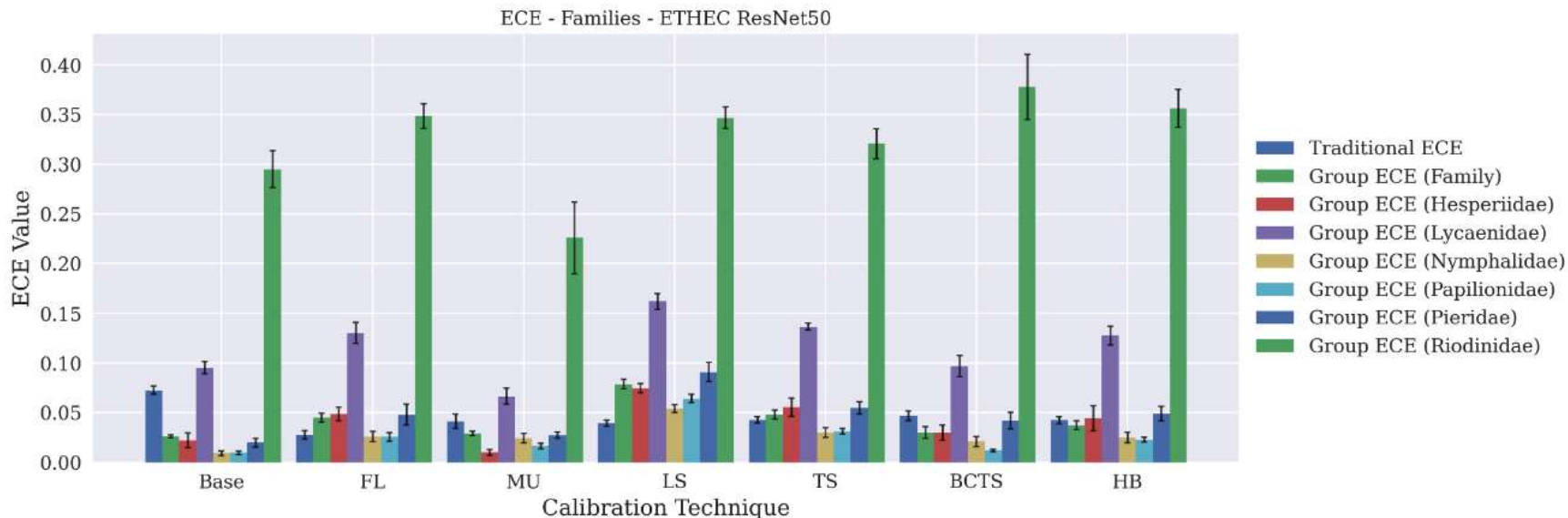
Experiment #1: Top- k ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

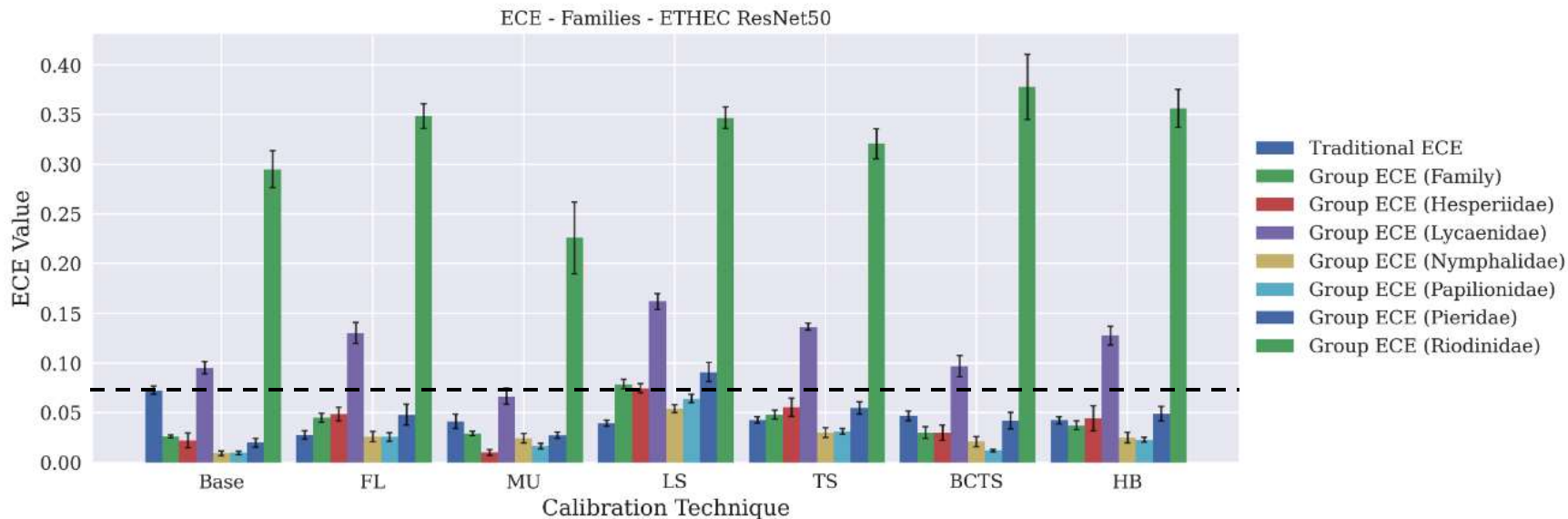
Experiment #2: Group-wise ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

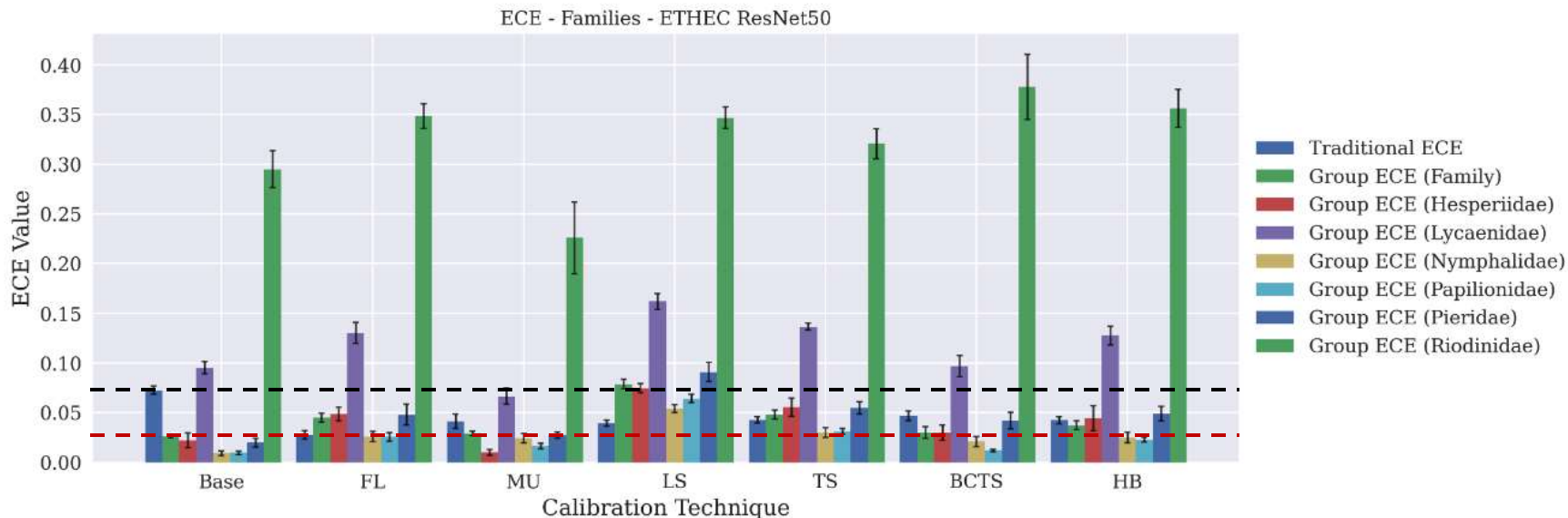
Experiment #2: Group-wise ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

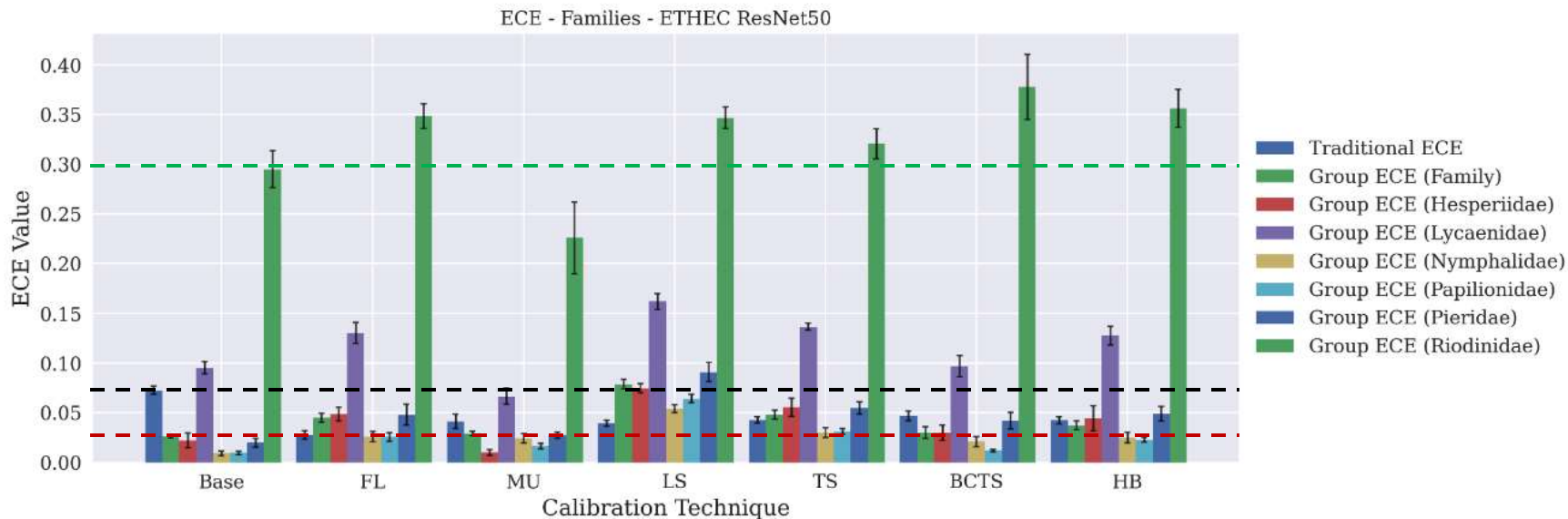
Experiment #2: Group-wise ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

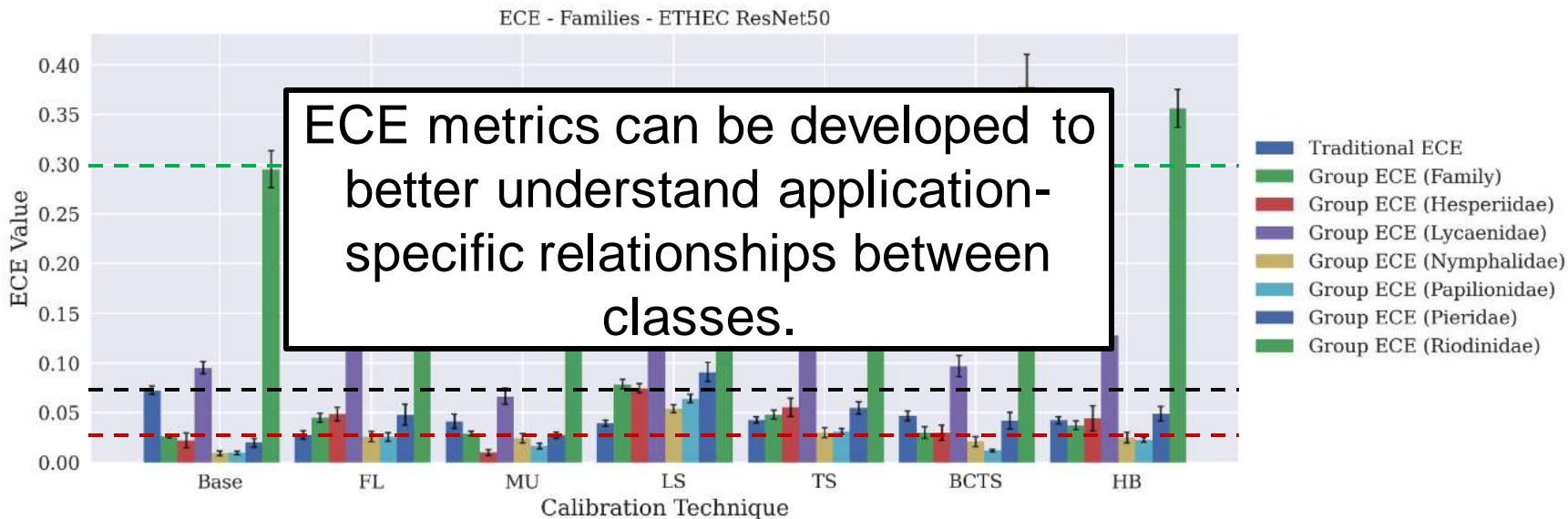
Experiment #2: Group-wise ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

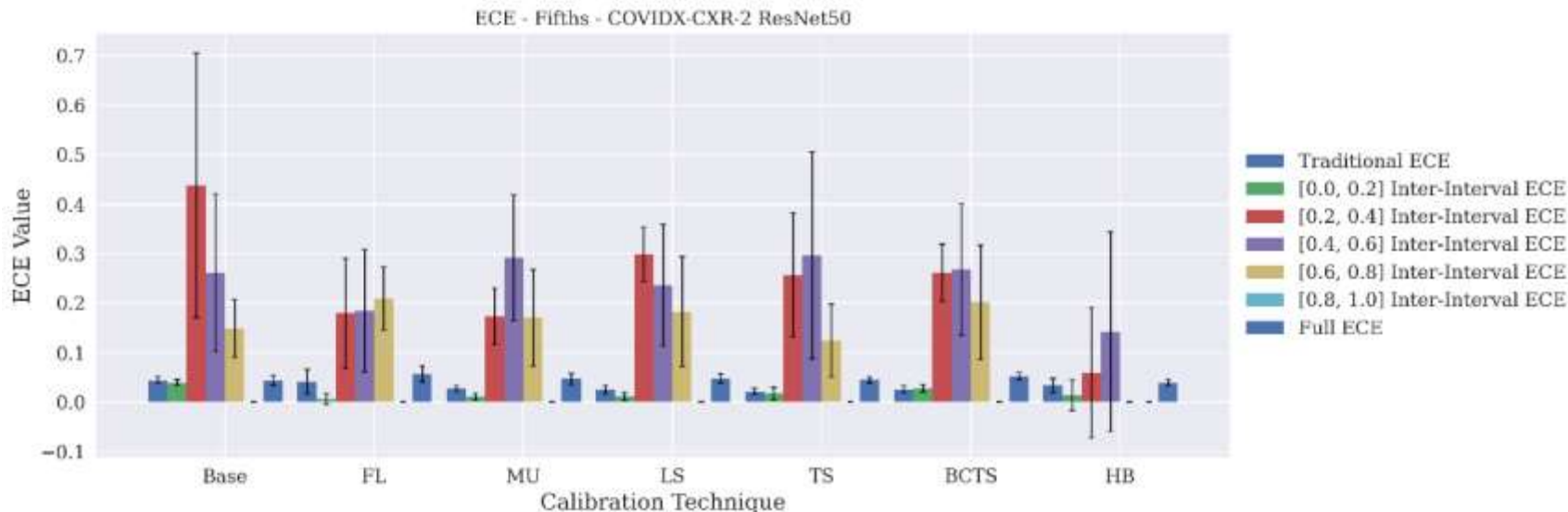
Experiment #2: Group-wise ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

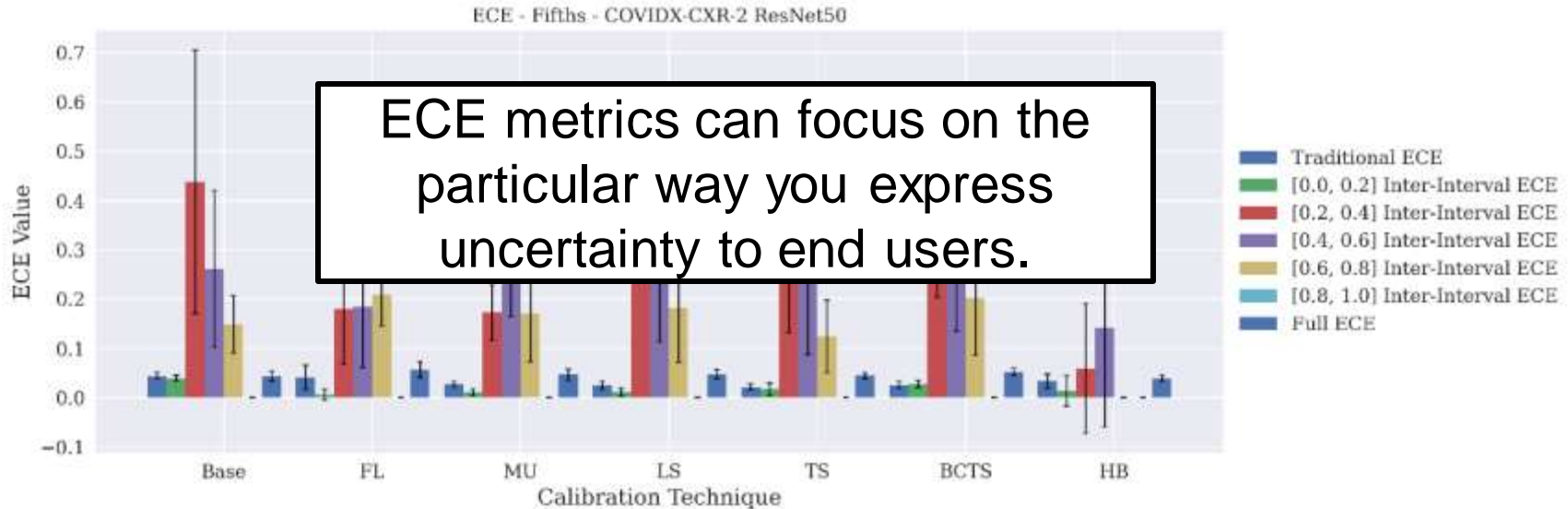
Experiment #3: Likert Category ECE



Empirical Results

Goal: Evaluate a sampling of the current state-of-the-art in classifier calibration with respect to traditional Top-1 ECE and ECE metrics tailored to particular contexts.

Experiment #3: Likert Category ECE



Summary

Accurate uncertainty estimates of ML model predictions can play a key role in better informed decision-making, as well as more reliable AI systems.

Our work focuses on evaluating ML models for their ability to estimate uncertainty in ways that map to important contexts in which the model will be used.

This work is part of a larger project on quantifying, reasoning about, and rectifying uncertainty in ML models with collaborations across government and academia.

We are very interested in hearing about mission relevant settings where reasoning under uncertainty is important.

If you have interest in this work or on a related topic, please contact me:

etheim@sei.cmu.edu

