15th September

On the Calibration of Systems that Learn to Defer to Experts

Rajeev Verma

AI for Autonomous Systems

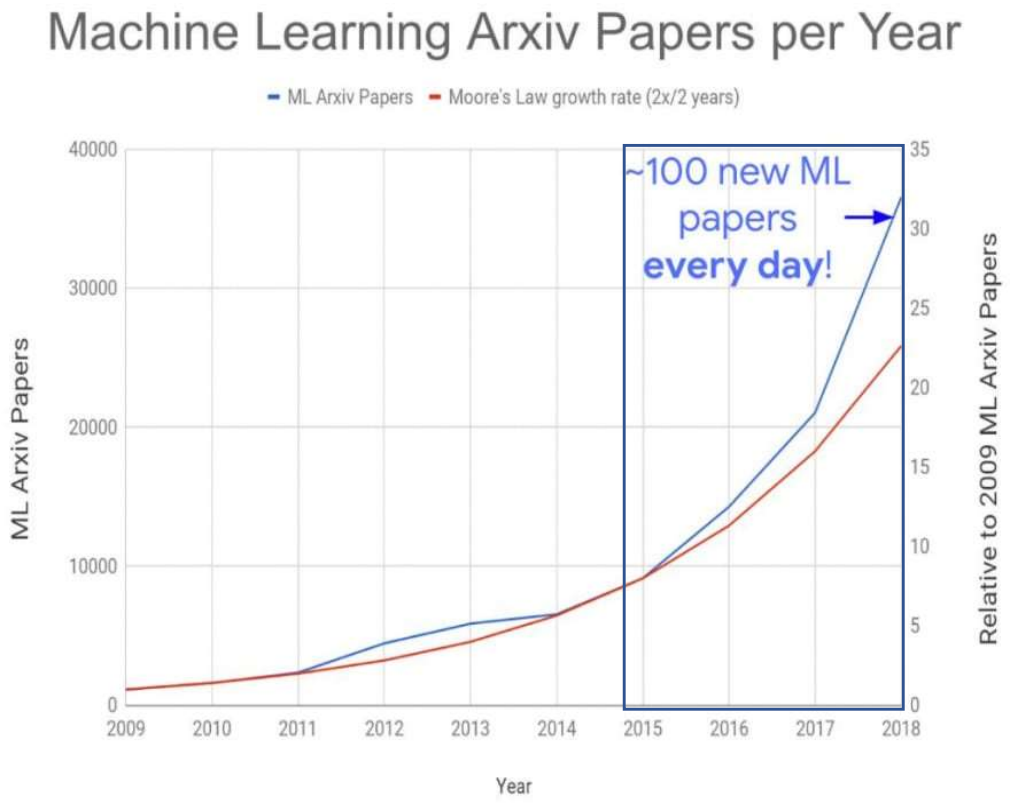
ICML 2015 DL Workshop

a long time ago

…A final thing that's really interesting is how do we ***interface*** *the results of the deep learning to humans. […] Humans don't have the same sense with the computer of sort of you know understanding its* ***uncertainty*** *how it expresses itself…*

Neil Lawrence
University of Cambridge

ICML 2015 DL Workshop
a long time ago

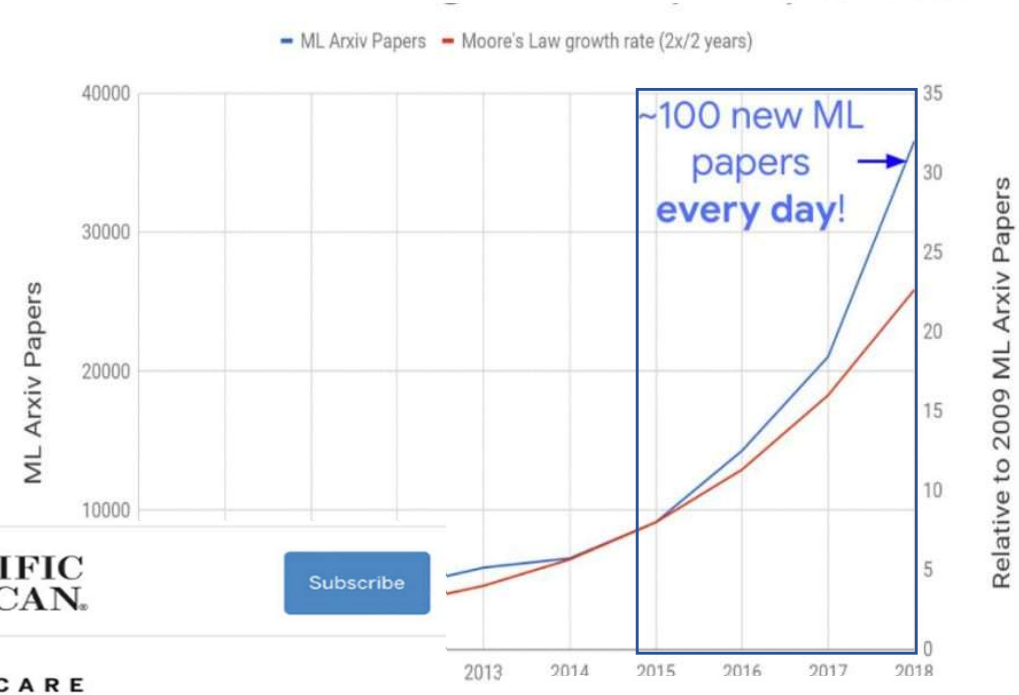

Machine Learning Arxiv Papers per Year
ML Arxiv Papers
Moore's Law growth rate (2x/2 years)
~100 new ML papers every day!
ML Arxiv Papers
Relative to 2009 ML Arxiv Papers
Year
40000
30000
20000
10000
0
35
30
25
20
15
10
5
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

Home / Cities / Gandhinagar / Use of AI can help judiciary dispose of pending cases: Gujarat HC CJ

Use of AI can help judiciary dispose of pending cases: Gujarat HC CJ

Suggesting similar use of AI algorithm to dispose of pending cases, Kumar said, “Consider Motor Vehicle (MV) cases that account for highest in any court... They can be broadly classified into death cases, grievous injuries cases, simple injuries and for insurance claims."

er Year



SCIENTIFIC AMERICAN.

Subscribe

HEALTH CARE

Algorithm That Detects Sepsis Cut Deaths by Nearly 20 Percent

Over two years, a machine-learning program warned thousands of health care providers about patients at high risk of sepsis, allowing them to begin treatments nearly two hours sooner

By Sophie Bushwick on August 1, 2022

Self-driving Vehicles with Human-like Perception

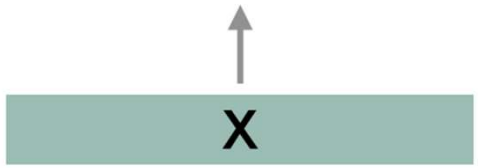
By **ELE Times** September 14, 2022


PRINSENGRACHT
bussia

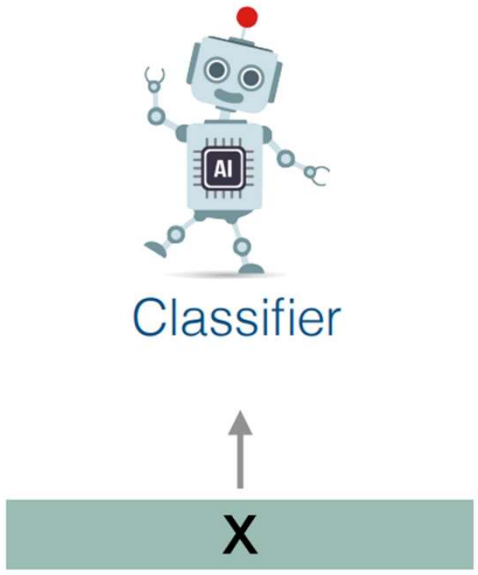
Gradual Automation

Gradual Automation

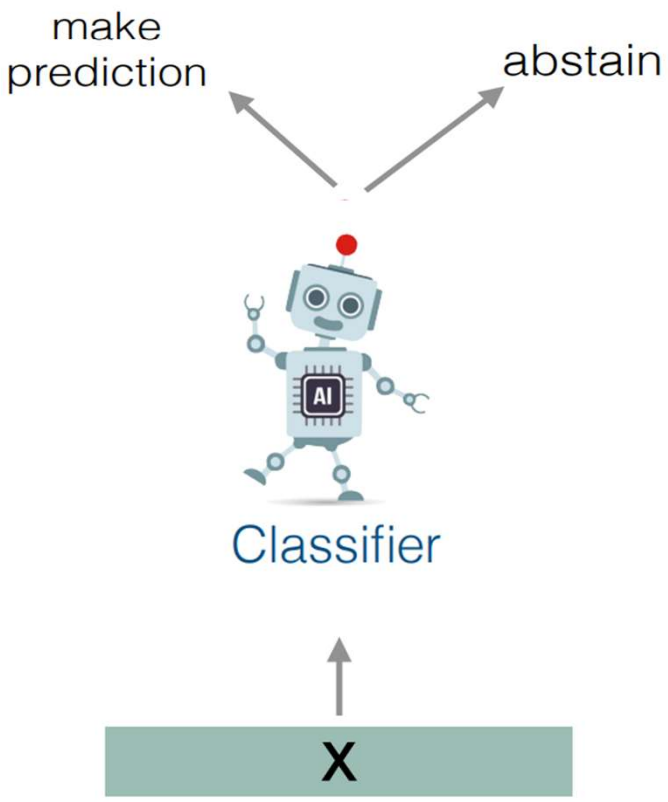
Learning to Defer



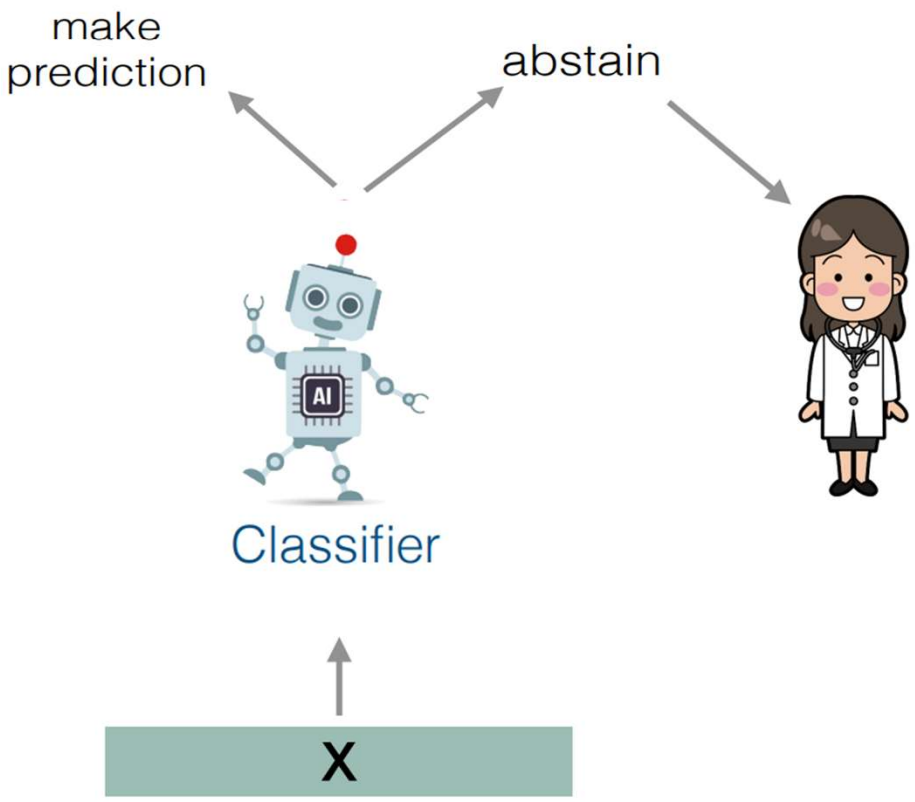
Learning to Defer



Learning to Defer

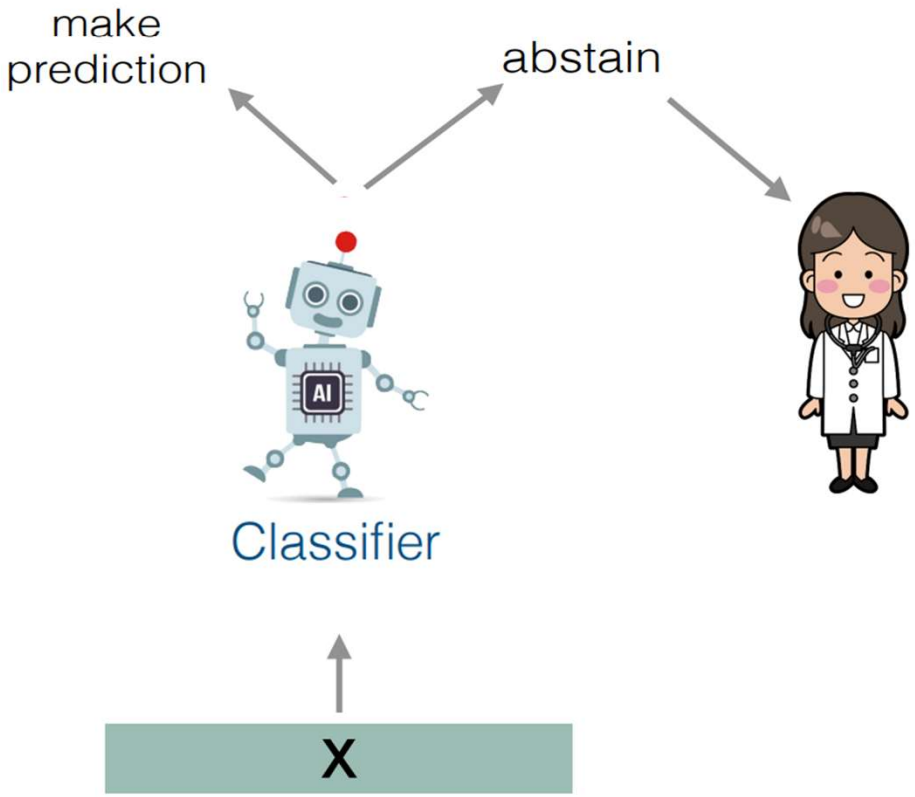


Learning to Defer



Learning to Defer

The decision to defer should also depend on the expertise of the expert.



Learning to Defer

Consistent Estimators for Learning to Defer to an Expert

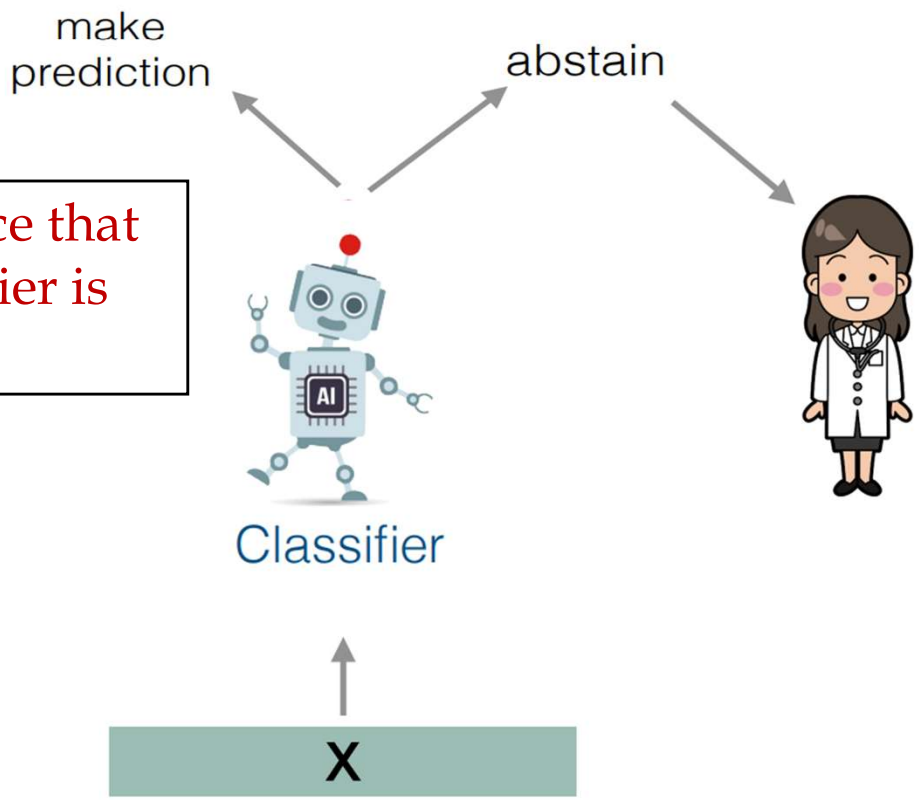
Hussein Mozannar * David Sontag †

Abstract

Learning algorithms are often used in conjunction with expert decision makers in practical scenarios, however this fact is largely ignored when designing these algorithms. In this paper we explore how to learn predictors that can either predict or choose to defer the decision to a downstream expert. Given only samples of the expert's decisions, we give a procedure based on learning a classifier and a rejector and analyze it theoretically. Our approach is based on a novel reduction to cost sensitive learning where we give a consistent surrogate loss for cost sensitive learning that generalizes the cross entropy loss. We show the effectiveness of our approach on a variety of experimental tasks.

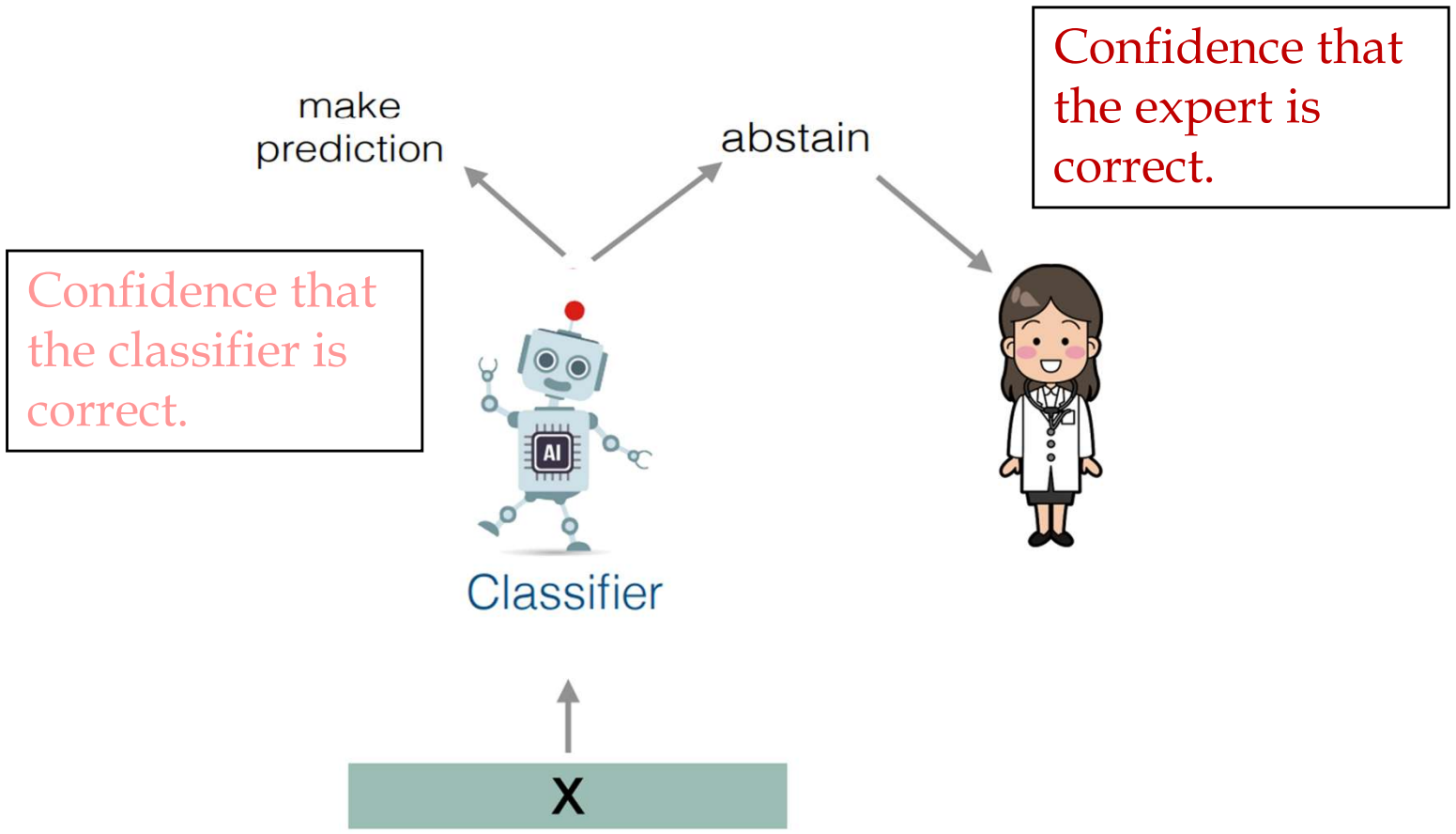
ICML 2020

Learning to Defer

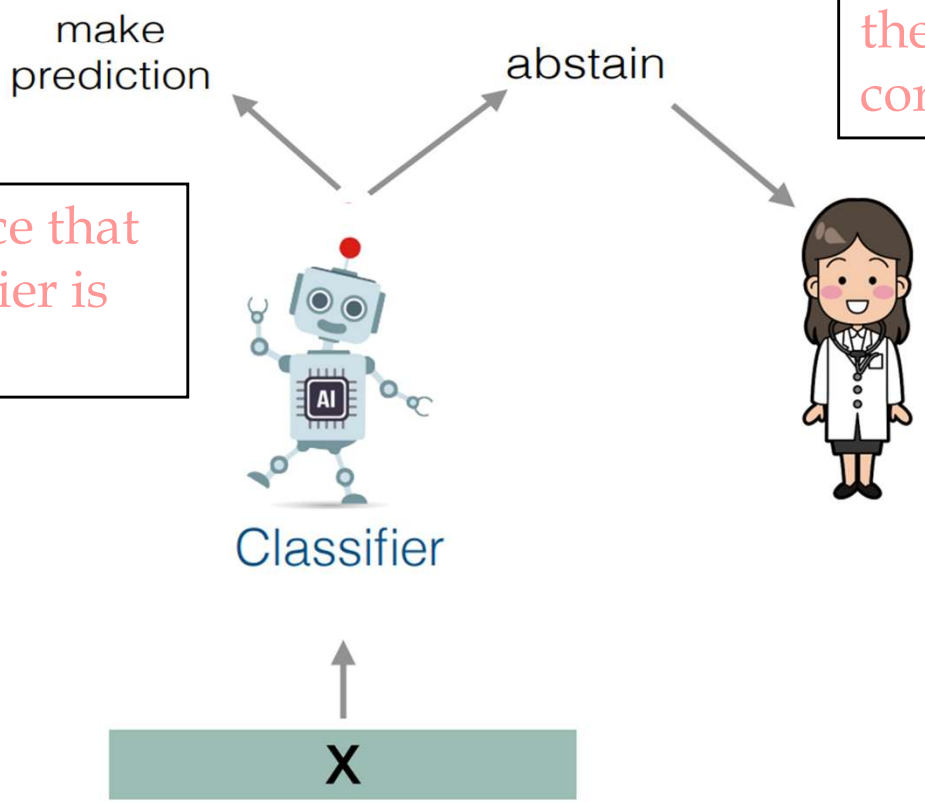


Confidence that the classifier is correct.

Learning to Defer



Learning to Defer



Confidence that the classifier is correct.

Confidence that the expert is correct.

Defer to the expert if the expert is more confident than the classifier.

…A final thing that's really interesting is how do we ***interface*** *the results of the deep learning to humans. […] Humans don't have the same sense with the computer of sort of you know understanding its* ***uncertainty*** *how it expresses itself…*

Neil Lawrence
University of Cambridge

ICML 2015 DL Workshop
a long time ago

nature medicine

LETTERS

https://doi.org/10.1038/s41591-020-0942-0

Check for updates

Human–computer collaboration for skin cancer recognition

Philipp Tschandl[1,17], Christoph Rinner[2,17], Zoe Apalla[3], Giuseppe Argenziano[4], Noel Codella[5], Allan Halpern[6], Monika Janda[7], Aimilios Lallas[3], Caterina Longo[8,9], Josep Malvehy[10,11], John Paoli[12,13], Susana Puig[10,11], Cliff Rosendahl[14], H. Peter Soyer[15], Iris Zalaudek[16] and Harald Kittler[1]✉

The rapid increase in telemedicine coupled with recent advances in diagnostic artificial intelligence (AI) create the imperative to consider the opportunities and risks of inserting AI-based support into new paradigms of care. Here we

competitive view of AI is evolving based on studies suggesting that a more promising approach is human–AI cooperation[10–15]. The role of human–computer collaboration in health-care delivery, the appropriate settings in which it can be applied and its impact on the

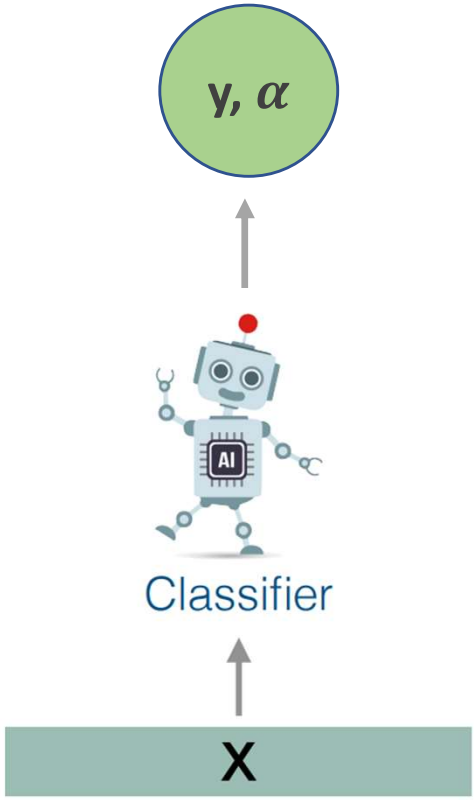
"The least experienced [physicians] tended to accept AI-based support that contradicted their initial diagnosis even if they were confident."

…A final thing that's really interesting is how do we ***interface*** *the results of the deep learning to humans.* ***I think we're in a difficult area people sort of when they look at what a computer says to them and outputs something they just assume it's right. Humans don't have the same sense with the computer of sort of you know understanding its uncertainty*** *how it expresses itself…*

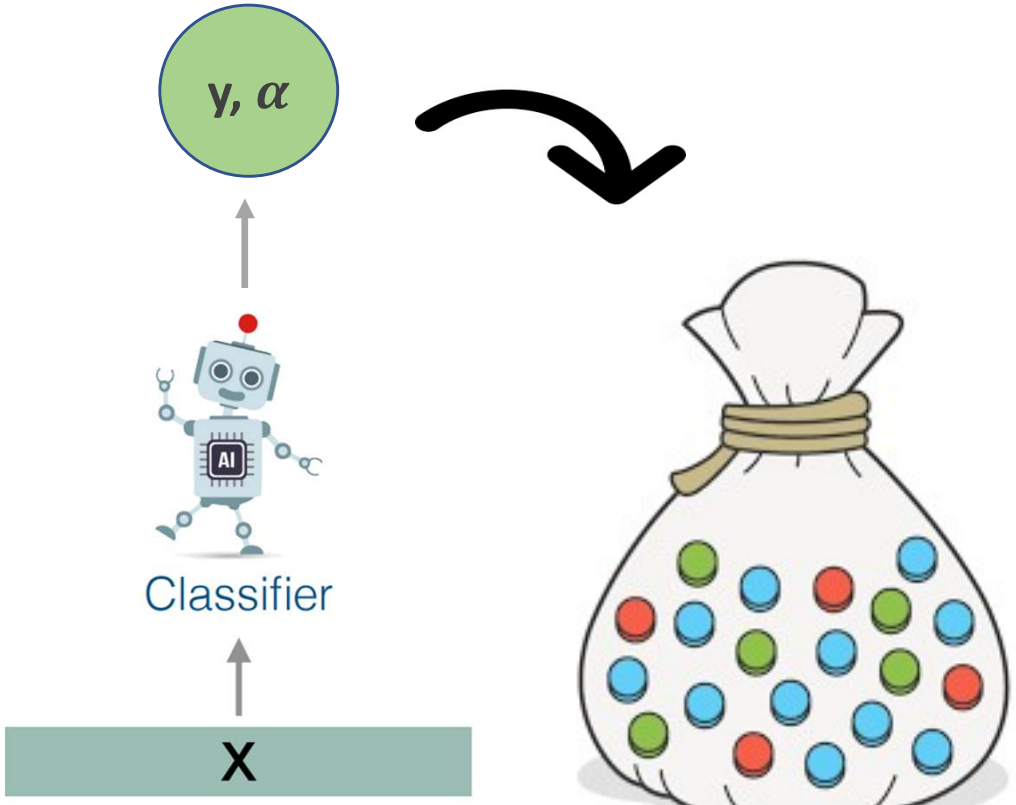
Neil Lawrence
University of Cambridge

ICML 2015 DL Workshop
a long time ago

Calibration



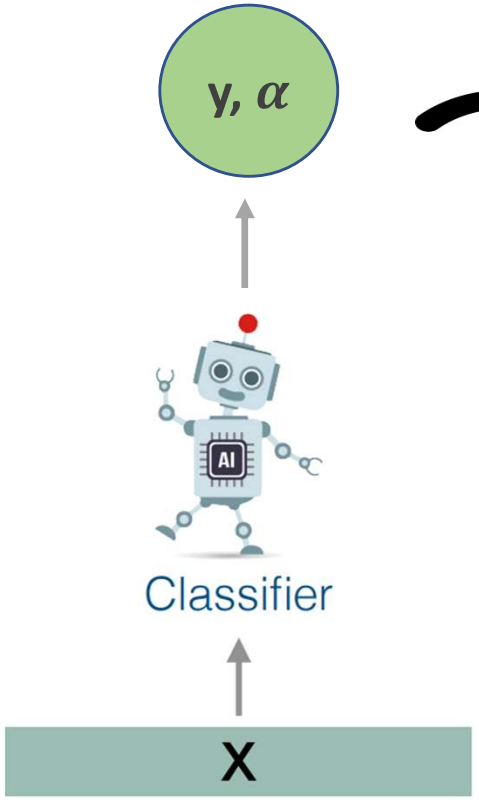
Calibration

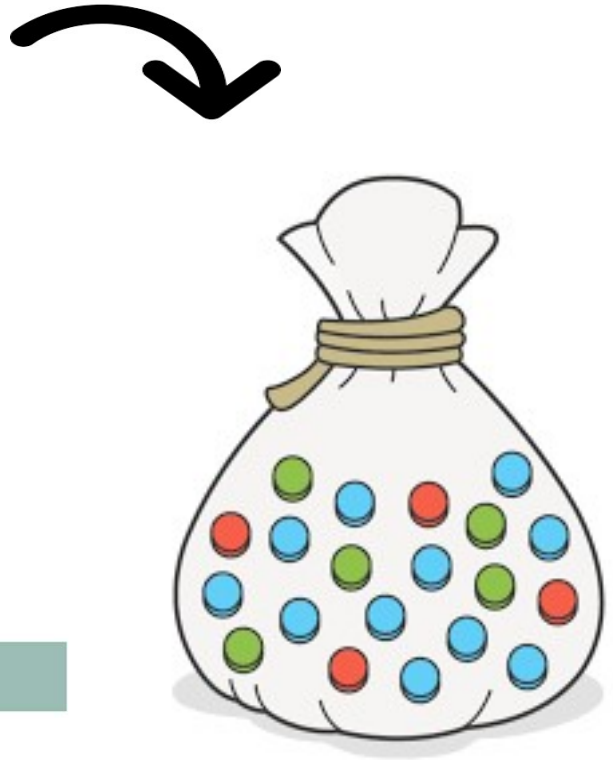


α confidence bag

Calibration

Confidence calibration means that the proportion of samples for which the classifier makes *correct* prediction must be α.





α confidence bag

Our Work

Is the learning to defer system calibrated? Is the system a good forecaster?

Our Work

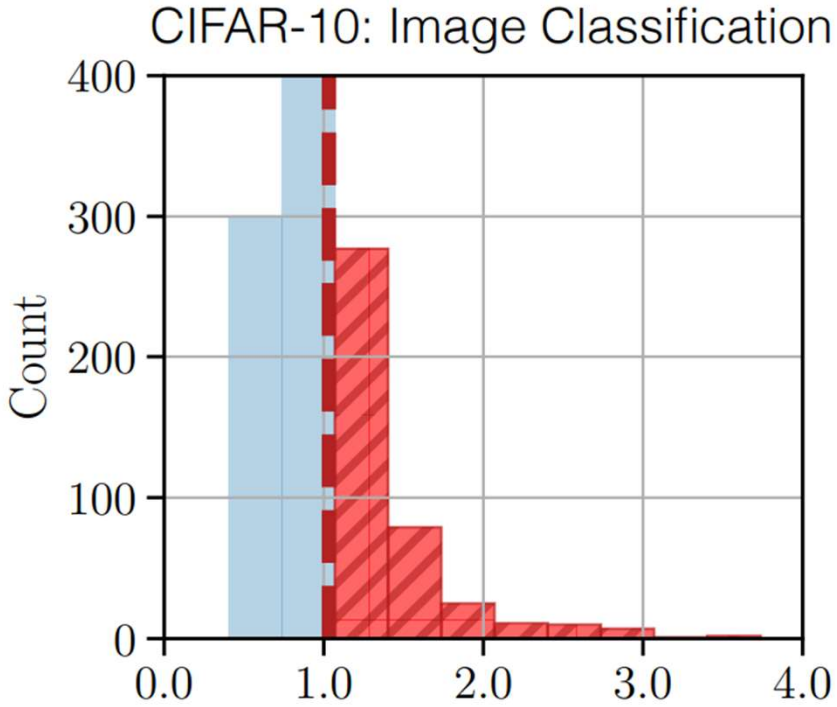
1. **Does the system correctly estimate the classifier's confidence in its prediction?**

2. **Does the system correctly estimate the expert's correctness confidence?**

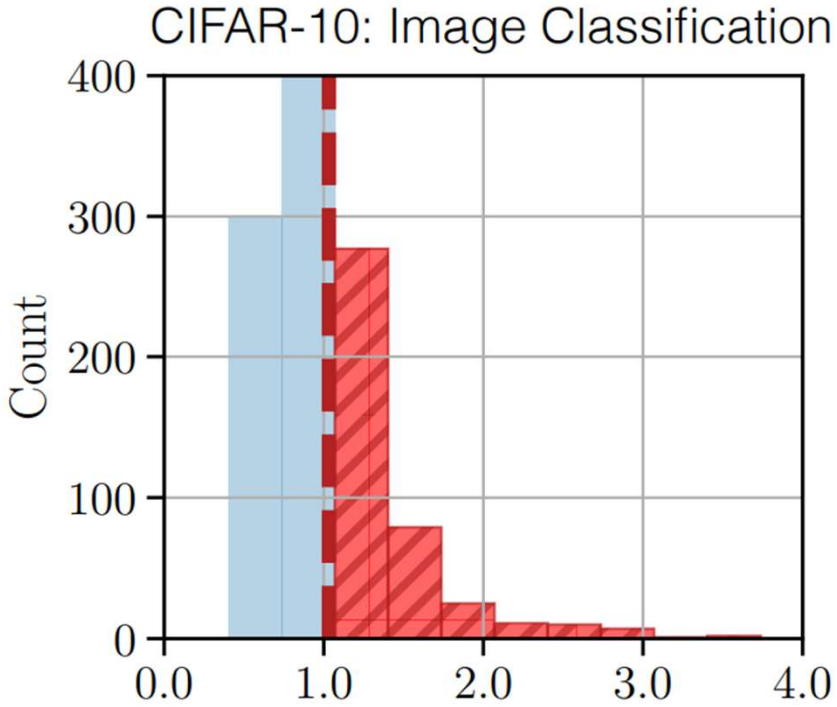
Our Work

1. **Does the system correctly estimate the classifier's confidence in its prediction?**

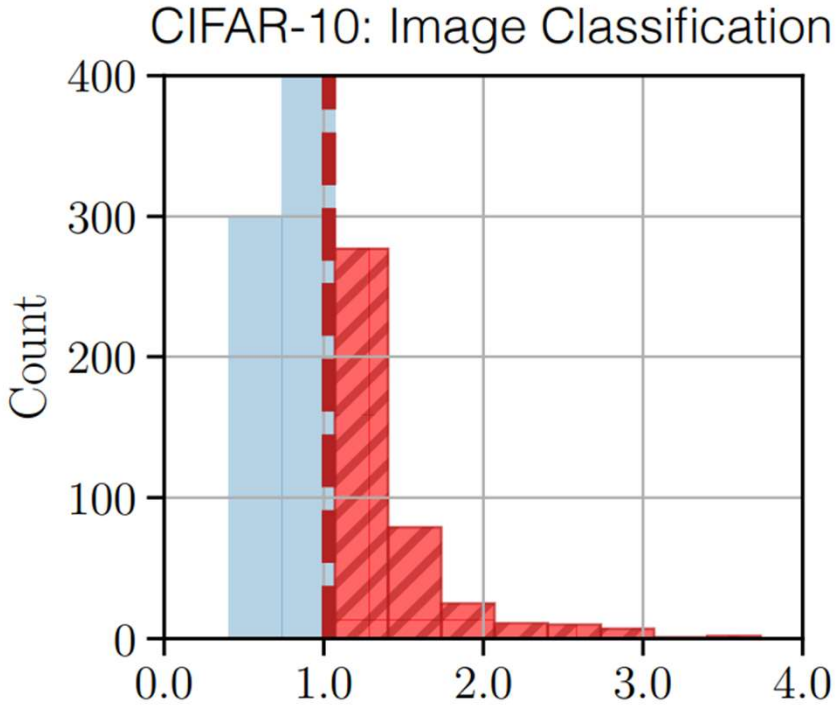
2. **Does the system correctly estimate the expert's correctness confidence?**



Expert's correctness confidence



The system provides degenerate confidences for expert's correctness.

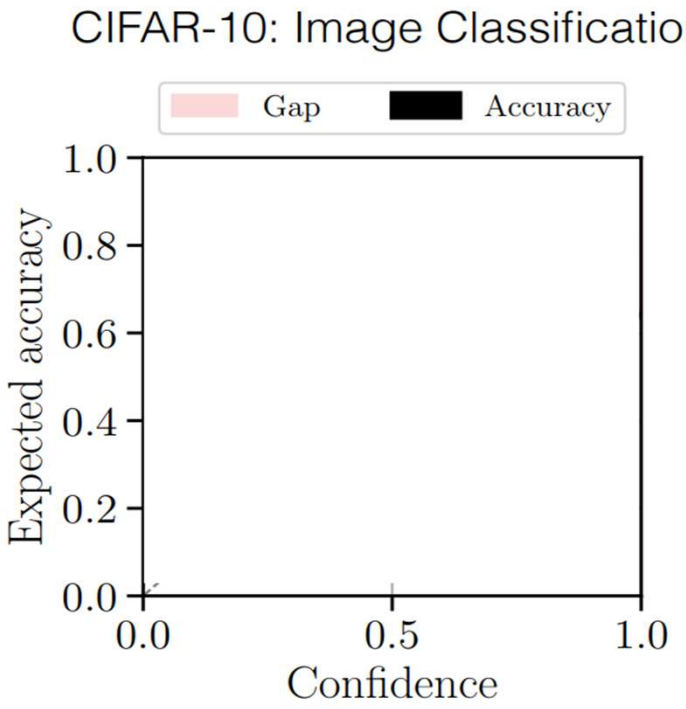


The system provides degenerate confidences for expert's correctness.

Truncation: $p_m(x) \in (0,1]$

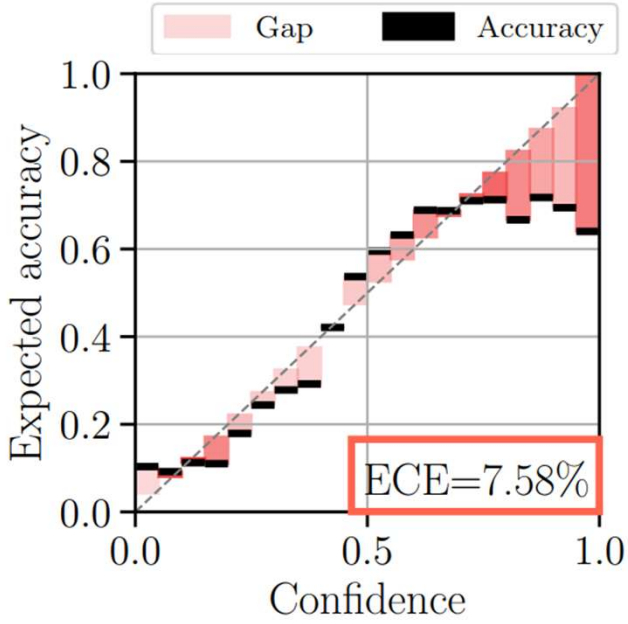
CIFAR-10: Image Classification

Truncation: $p_m(x) \in (0,1]$



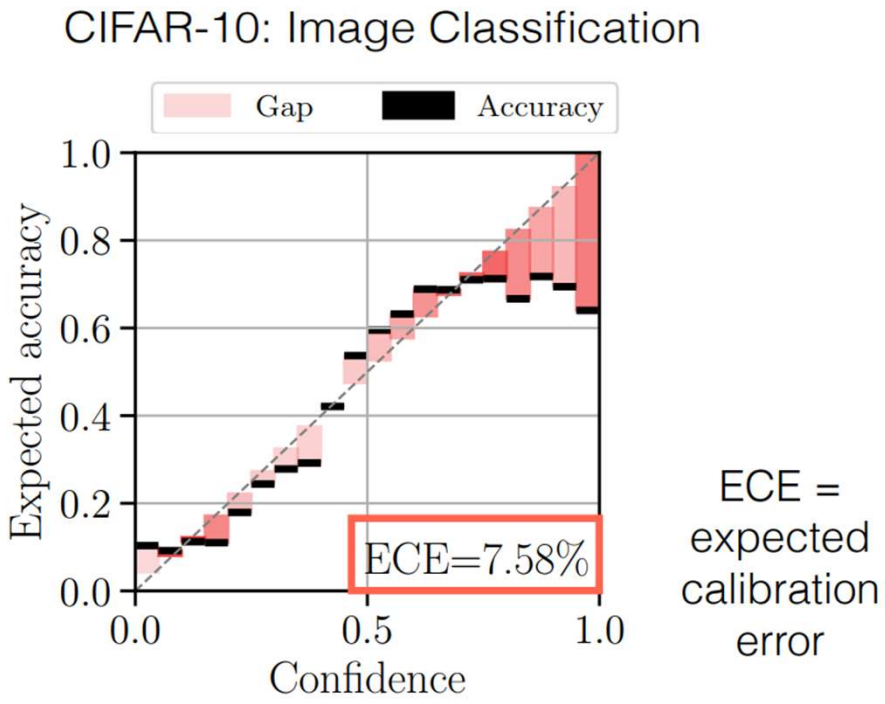
Truncation: $p_m(x) \in (0,1]$

CIFAR-10: Image Classification



ECE = expected calibration error

Truncation: $p_m(x) \in (0,1]$



ECE = expected calibration error

The system is overconfident.

Our Work

Calibrated Learning to Defer with One-vs-All Classifiers

Rajeev Verma [1] Eric Nalisnick [1]

Abstract

The *learning to defer* (L2D) framework has the potential to make AI systems safer. For a given input, the system can defer the decision to a human if the human is more likely than the model to take the correct action. We study the calibration not is usually derived from the model's confidence. For a self-driving car, a winding stretch of road could make the system unconfident in its abilities. The system would then refuse to drive and forces the human to take control. When the system becomes confident again (e.g. on a straight road), it can then take back control from the human.

2022

ICML 2022

Our Work

Calibrated Learning to Defer with One-vs-All Classifiers

Rajeev Verma [1] Eric Nalisnick [1]

Abstract

The *learning to defer* (L2D) framework has the potential to make AI systems safer. For a given input, the system can defer the decision to a human if the human is more likely than the model to take the correct action. We study the calibration

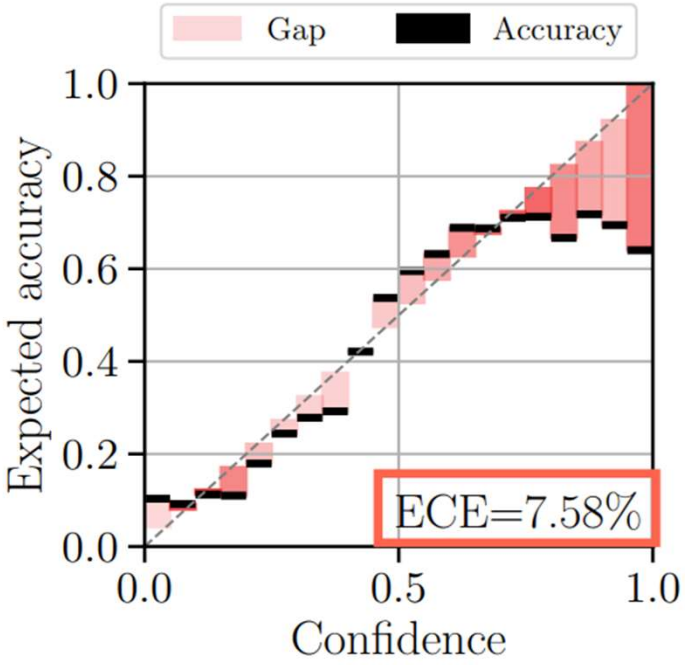
not is usually derived from the model's confidence. For a self-driving car, a winding stretch of road could make the system unconfident in its abilities. The system would then refuse to drive and forces the human to take control. When the system becomes confident again (e.g. on a straight road), it can then take back control from the human.

ICML 2022

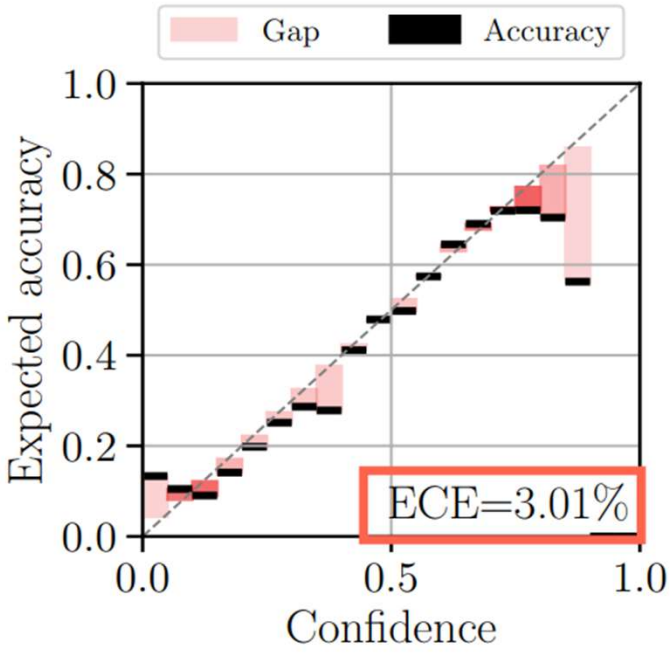
We propose a new surrogate loss for learning to defer that is provably consistent and provides well-calibrated confidence estimates.

Our Work

Earlier method

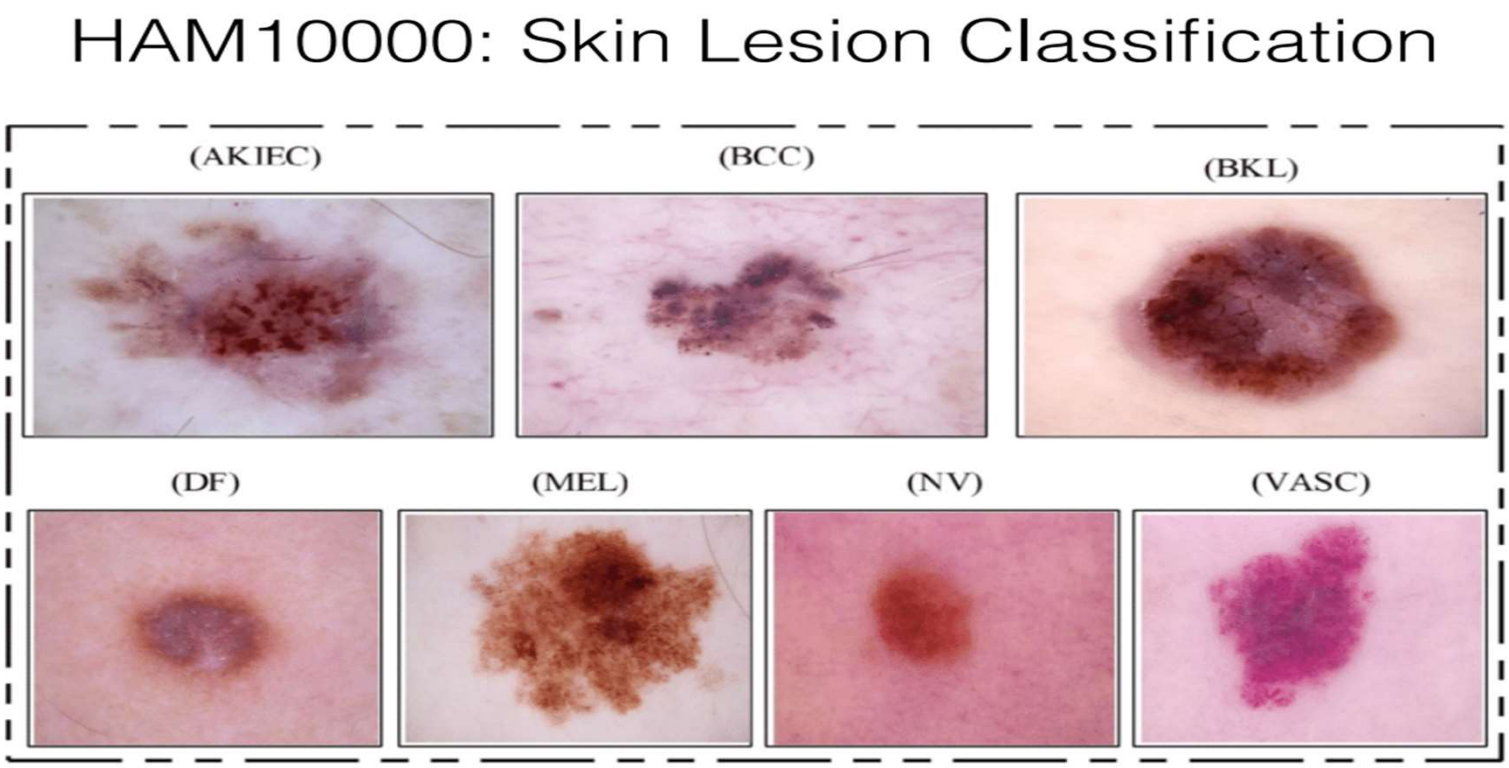


Our method

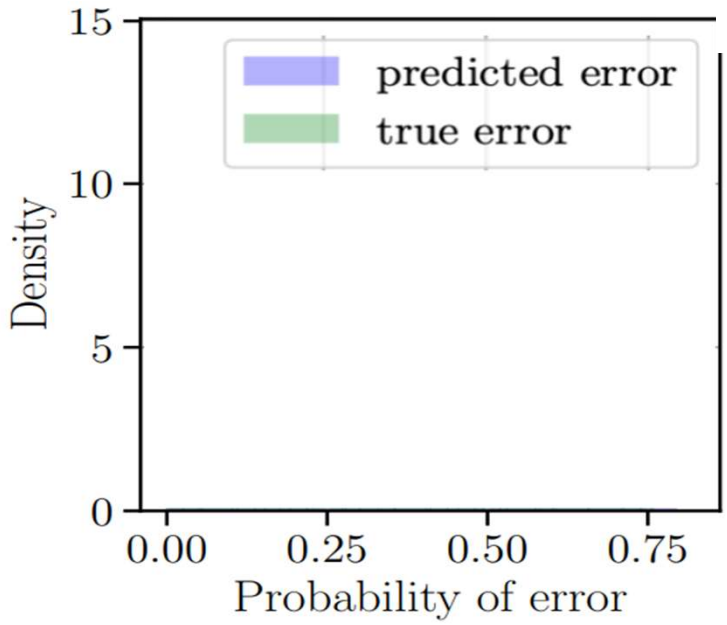


We propose a new surrogate loss for learning to defer that is provably consistent and provides well-calibrated confidence estimates.

Our Work


HAM10000: Skin Lesion Classification
(AKIEC)
(BCC)
(BKL)
(DF)
(MEL)
(NV)
(VASC)

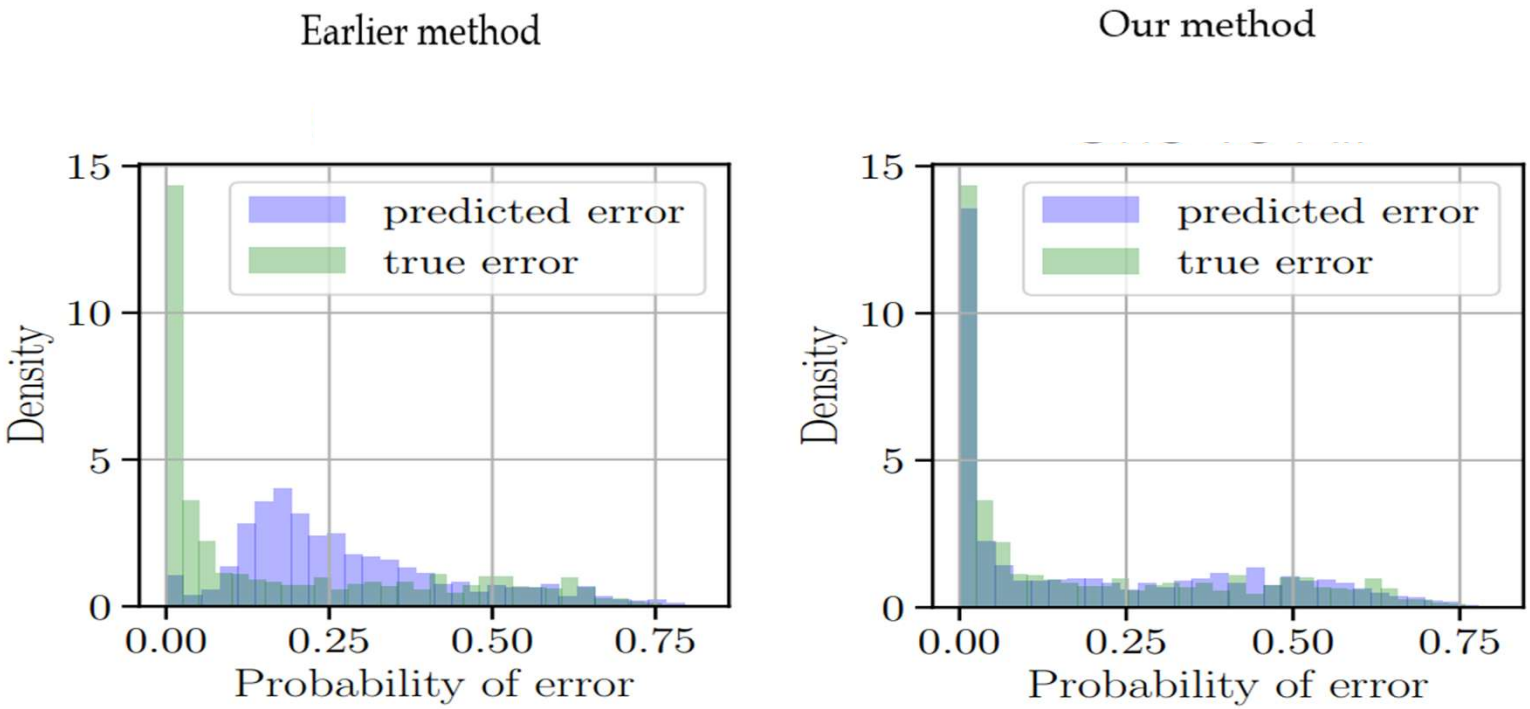
Our Work



We propose a new surrogate loss for learning to defer that is provably consistent and provides well-calibrated confidence estimates.

Our Work

HAM10000: Skin Lesion Classification



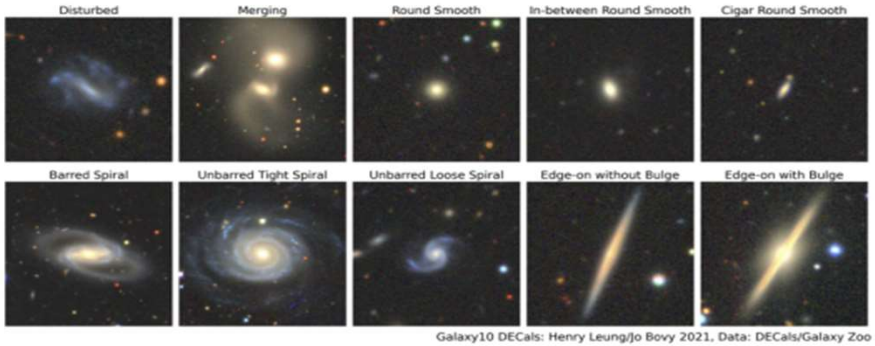
We propose a new surrogate loss for learning to defer that is provably consistent and provides well-calibrated confidence estimates.

Our Work

one-vs-all (ours) softmax confidence

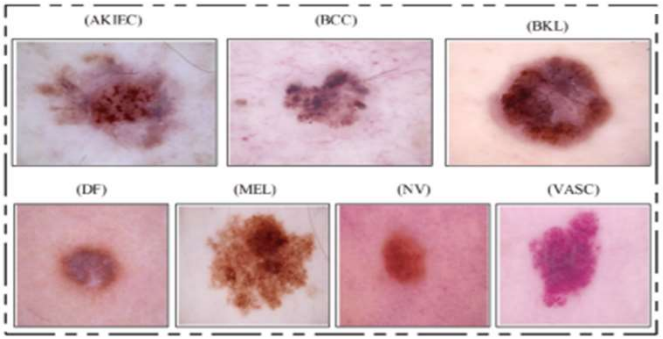
score differentiable triage

Galaxy Zoo



Galaxy10 DECals: Henry Leung/Jo Bovy 2021, Data: DECals/Galaxy Zoo

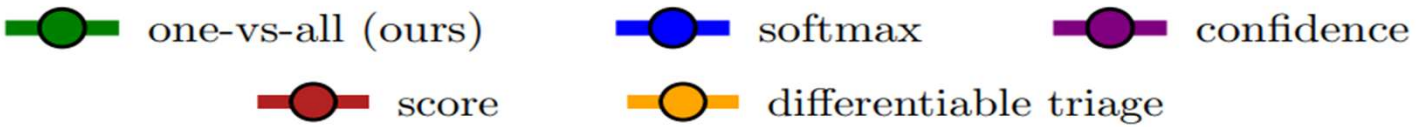
HAM10000

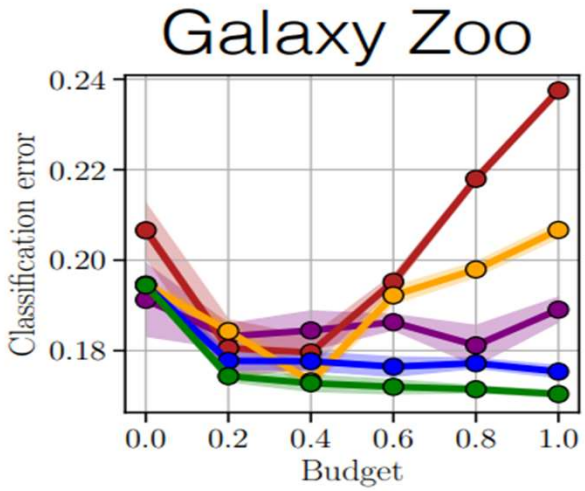


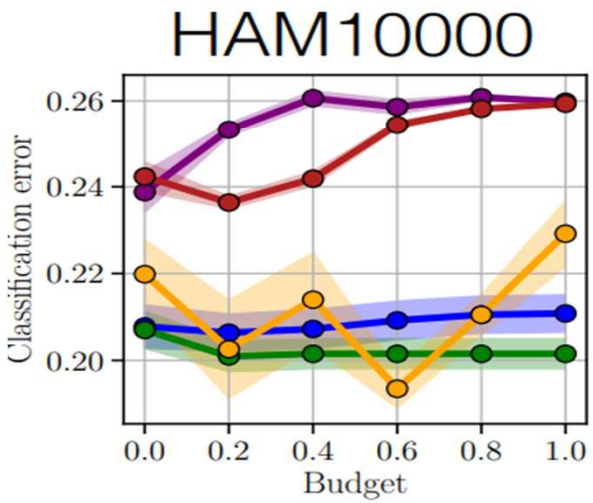
Hate Speech

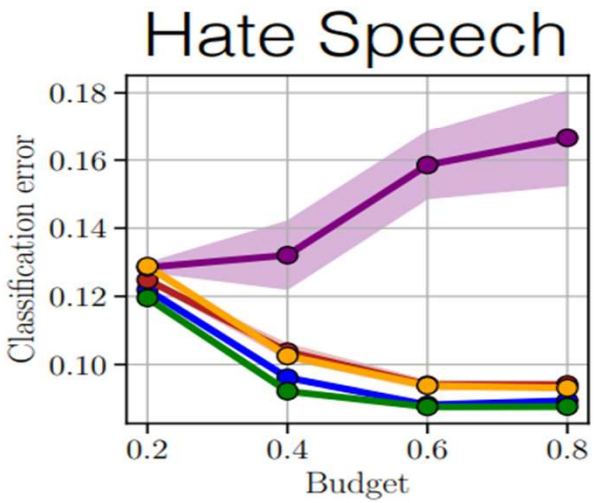
[Davidson et al., ICWSM 2017]

Our Work


one-vs-all (ours)
softmax
confidence
score
differentiable triage


Galaxy Zoo
Classification error
0.24
0.22
0.20
0.18
0.0
0.2
0.4
0.6
0.8
1.0
Budget


HAM10000
Classification error
0.26
0.24
0.22
0.20
0.0
0.2
0.4
0.6
0.8
1.0
Budget


Hate Speech
Classification error
0.18
0.16
0.14
0.12
0.10
0.2
0.4
0.6
0.8
Budget

Our Work

Calibrated Learning to Defer with One-vs-All Classifiers

Rajeev Verma [1] Eric Nalisnick [1]

Abstract

The *learning to defer* (L2D) framework has the potential to make AI systems safer. For a given input, the system can defer the decision to a human if the human is more likely than the model to take the correct action. We study the calibration not is usually derived from the model's confidence. For a self-driving car, a winding stretch of road could make the system unconfident in its abilities. The system would then refuse to drive and forces the human to take control. When the system becomes confident again (e.g. on a straight road), it can then take back control from the human.

2022

ICML 2022

We propose a new surrogate loss for learning to defer that is provably consistent and provides well-calibrated confidence estimates, and doesn't compromise on accuracy while doing so.

Our Work

On the Calibration of Learning to Defer to Multiple Experts

Rajeev Verma [1] Daniel Barrejón [2] Eric Nalisnick [1]

Abstract

We study the calibration properties of multi-expert *learning to defer* (L2D). In particular, we study the framework's ability to estimate $\mathbb{P}(\mathrm{m}_j = \mathrm{y}|\boldsymbol{x})$, the probability that the jth expert

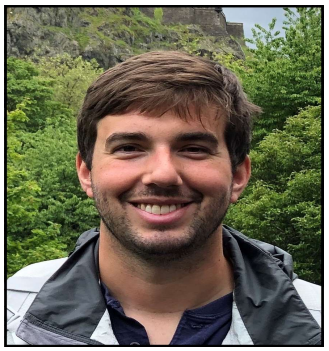
2. Learning To Defer to Multiple Experts

Data We first define the data for multi-class, multi-expert *learning to defer* (L2D). Let $\mathcal{X}$ denote the feature space, and let $\mathcal{Y}$ denote the output space, which we will always assume to be a categorical encoding of multiple (K) classes. We assume that we have samples from the true generative

ICML 2022 HMCaT Worksop

We also analysed calibration for learning to defer in multi-experts setting and found similar patterns to hold.

Collaborators



Eric Nalisnick

Daniel Barrejón

*Some slides taken from Eric Nalisnick.
*Some images taken from Google. Copyright: their respective owners.