Multiclass Calibration Assessment and Recalibration of Probability Predictions via the Linear Log Odds Calibration Function

Amy Vennos, Xin Xing, and Christopher T. Franck

DATAWorks 2026 Speed Session

April 21, 2026

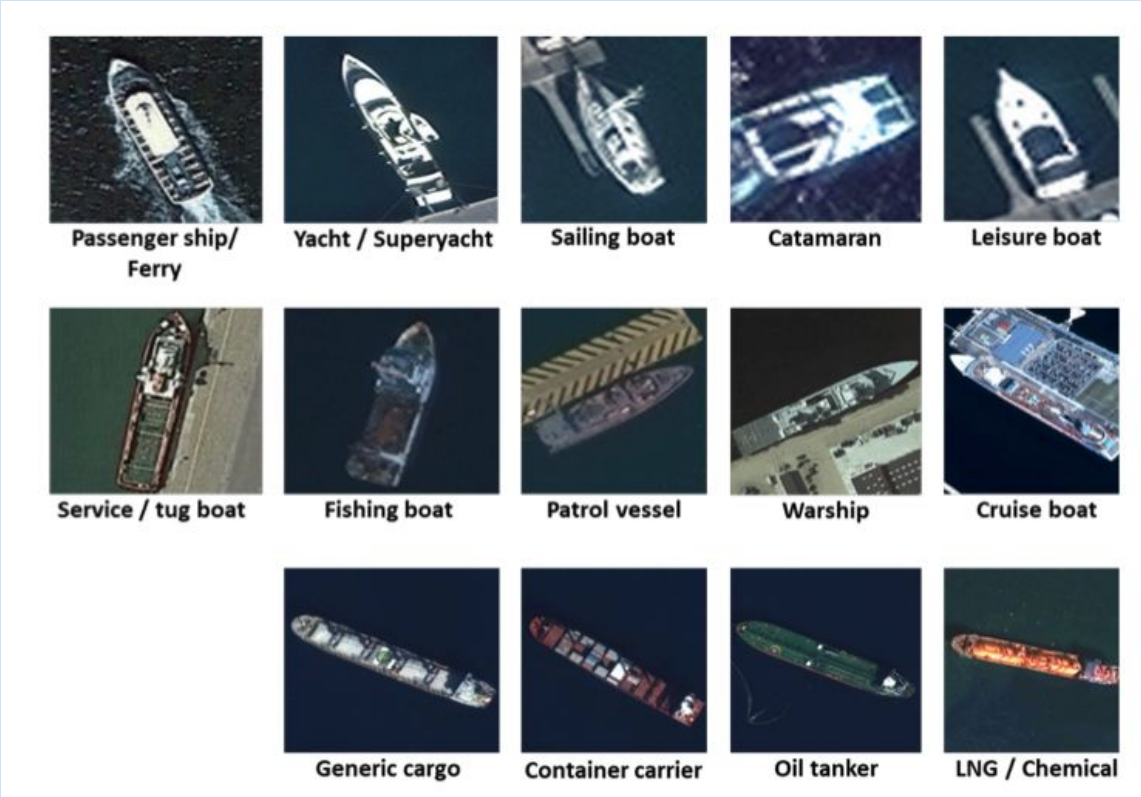
NATIONAL SECURITY INSTITUTE
VIRGINIA TECH

AIRC
ACQUISITION INNOVATION
RESEARCH CENTER

Multicategory classification problems are widespread in defense applications.

Examples:

- Satellite image classification of destroyer ships, cargo ships, and fishing vessels
- Natural Language Processing intercepted communication sentiments as hostile, neutral, cooperative or deceptive
- Classification of drone movement as surveillance, delivery, or attack.



Modern machine learning algorithms output probabilities for multicategory classification problems.

Visual Geometry Group Neural Net (VGG Net) calculated *confidence scores* of several labels for these images from the CIFAR-10 data set.

VGG Net Confidence Scores:

Plane: 1.00
Ship: 0.00
Bird: 0.00
Deer: 0.00
Cat: 0.00

VGG Net Confidence Scores:

Plane: 0.94
Cat: 0.04
Bird: 0.01
Ship: 0.01
Dog: 0.01

Modern machine learning algorithms output probabilities for multicategory classification problems.

Visual Geometry Group Neural Net (VGG Net) calculated *confidence scores* of several labels for these images from the CIFAR-10 data set.

VGG Net Confidence Scores:

Plane: 1.00
Ship: 0.00
Bird: 0.00
Deer: 0.00

VGG Net Confidence S

Plane: 0.94
Cat: 0.04
Bird: 0.01
Ship: 0.01

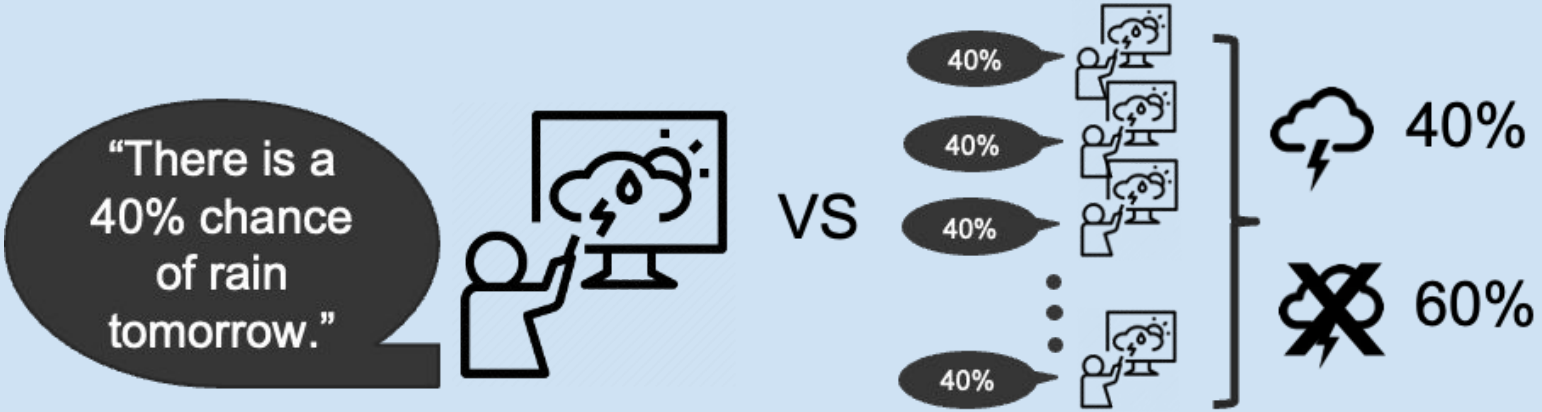
What if, historically, only 80% of images like these are actually planes?

Sometimes, confidence scores do not align with that happens in real life.

Calibration

- Confidence scores are **well-calibrated** if they agree with the long-run frequency of events.

Example:



- **Recalibration** maps probability predictions to those that are well calibrated.

Recalibration in Practice

VGG Net Confidence Scores:

Plane: 1.00
Ship: 0.00
Bird: 0.00
Deer: 0.00
Cat: 0.00

Recalibrated Confidence Scores:

Plane: 1.00
Ship: 0.00
Bird: 0.00
Deer: 0.00
Cat: 0.00

VGG Net Confidence Scores:

Plane: 0.94
Cat: 0.04
Bird: 0.01
Ship: 0.01
Dog: 0.01

Recalibrated Confidence Scores:

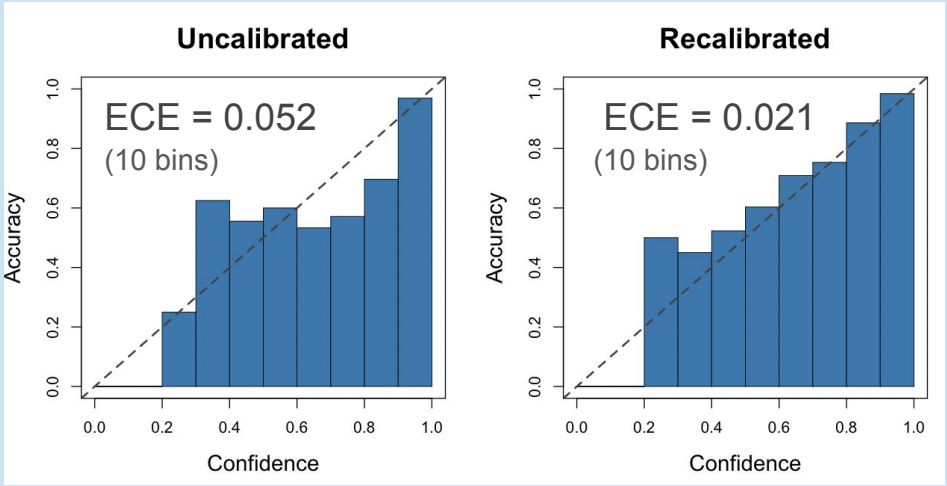
Plane: 0.80
Cat: 0.12
Bird: 0.03
Dog: 0.02
Ship: 0.02

Existing Measures of Multicategory Calibration

Visualizations:

- Reliability Diagrams

Bars following the x=y line represent data that are more well-calibrated.



Metrics:

- Expected Calibration Error (ECE)
- Maximum Calibration Error (MCE)

Take values from 0 to 1

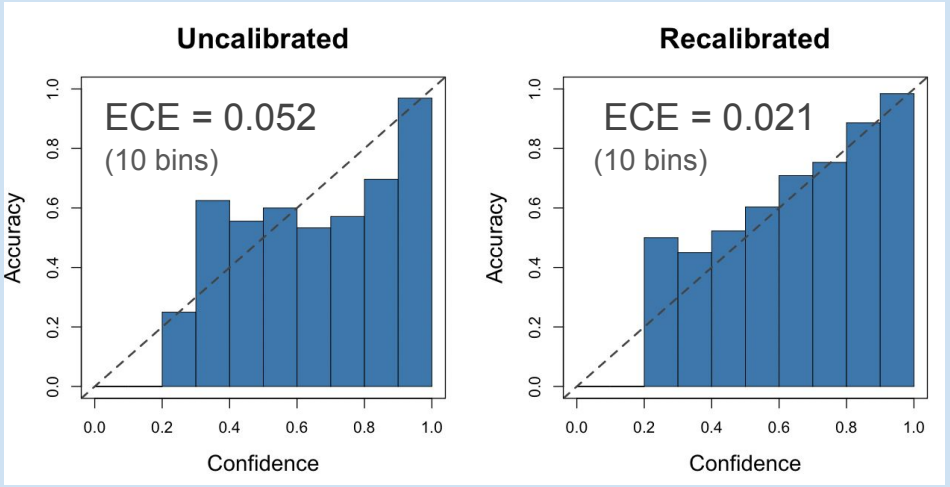
Lower scores indicate better calibration.

Existing Measures of Multicategory Calibration

Visualizations:

- Reliability Diagrams

Bars following the x=y line represent data that are more well-calibrated.



Metrics:

- Expected Calibration Error (ECE)
- Maximum Calibration Error (MCE)

Take values from 0 to 1

Lower scores indicate better calibration.

Limitations of current methods:

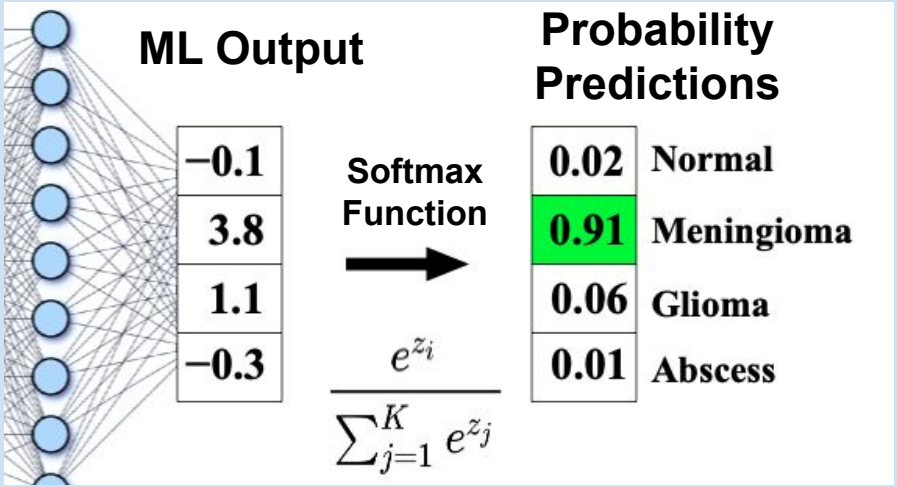
- Depends on a user-specified number of bins
- Offers little insight to how well a single model performs.

Existing Multicategory Recalibrators

- Temperature scaling
- Vector scaling
- Binning methods

Existing Multicategory Recalibrators

- Temperature scaling
- Vector scaling
- Binning methods



Not Allowed!

Limitations of current methods:

- Temperature scaling and vector scaling require internal ML model access.
- Recalibrators that are not likelihood-based and cannot test if a set of probabilities is well calibrated.
- Some methods do not greatly improve calibration of uncalibrated predictions.

Our Approach: Multicategory Linear Log Odds (MCLLO) Recalibration

c equations with 2(c-1) parameters:

$$log\left(\frac{g_{ij}}{g_{ic}}\right) = \log(\delta_j) + \gamma_j log\left(\frac{x_{ij}}{x_{ic}}\right)$$

$$\sum_{j=1}^{c} g_{ij} = 1 \qquad i = \{1, \dots n\} \qquad j = \{1, \dots c-1\}$$

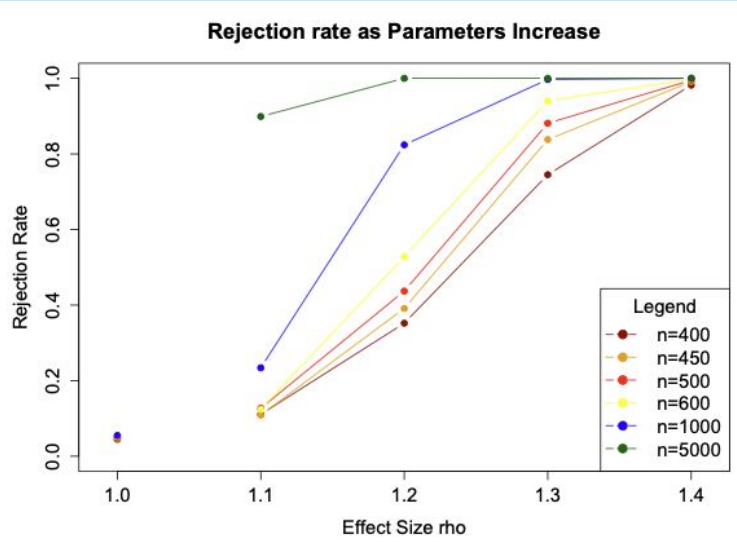
Transforms the original log odds of confidence scores, to the log odds of the recalibrated probability forecasts.

Parameters: two length - $(c-1)$ vectors

$$\boldsymbol{\delta} = (\delta_1, \dots \delta_{c-1})$$

$$\boldsymbol{\gamma} = (\gamma_1, \dots \gamma_{c-1})$$

MCLLO Likelihood Ratio Test (LRT)



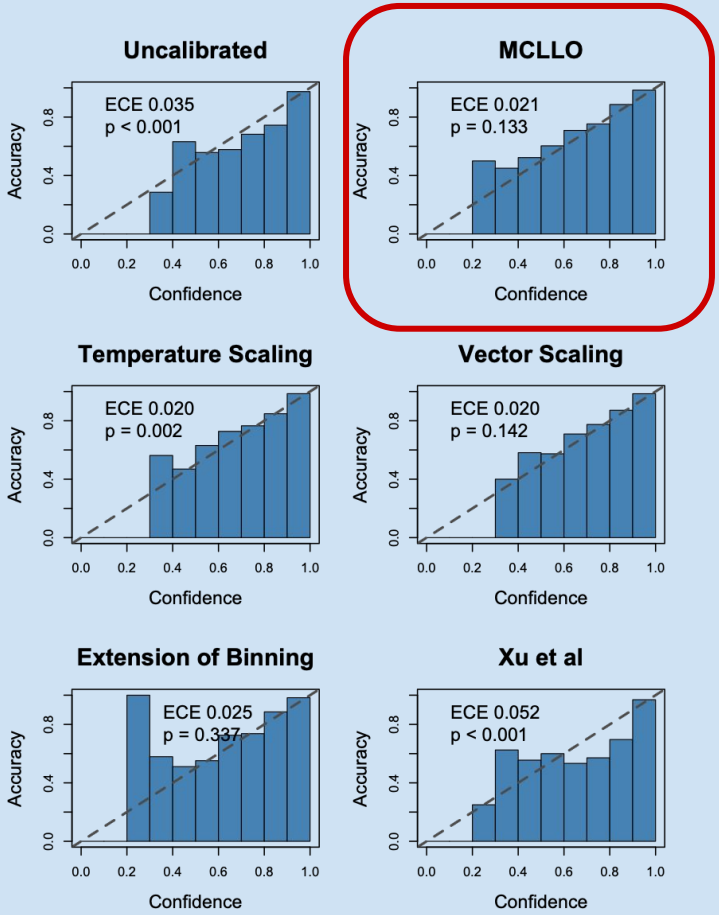
MCLLO has a built-in LRT to test for well-calibrated confidence scores.

- H_0: Confidence scores are well-calibrated.
- If rejected, apply recalibration.

The LRT test statistic:

- Does not depend on a user-specified bin size
- Controls the type one (false positive) error rate
- Gains power as sample size or departure from calibration grows.
- Does not require internal access to the ML model

MCLLO Recalibration



- Improves calibration of uncalibrated probability predictions more than competitor methods.
- Applies to a wider range of classification problems than the popular temperature scaling and vector scaling.

This is demonstrated by a case study on CIFAR-10 neural net confidence scores.

Note: Reported p values are from the MCLLO Likelihood Ratio Test

Summary

We contribute:

- The MCLLO recalibration function.
- The MCLLO likelihood and likelihood ratio test.

The MCLLO framework expands upon previous research in the following ways:

- Flexibility of applications
- Does not require under-the-hood access to ML architecture
- Built-in hypothesis testing for calibration
- Effective recalibration
- Direct measure of the calibration of one model
- No dependence on user-specified number of bins.

Thank you!

Arxiv paper link:

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense (DoD) through the Office of the Under Secretary of Defense for Acquisition and Sustainment (OUSD(A&S)) and the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under Contract HQ003424D0023. The Acquisition Innovation Research Center (AIRC) is a multi-university partnership led and managed by the Stevens Institute of Technology through the Systems Engineering Research Center (SERC) – a federally funded University Affiliated Research Center. Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Government (including the DoD and any government personnel).