

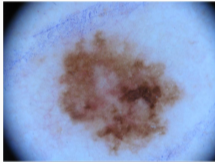
Calibration tests in multi-class classification: A unifying framework

David Widmann* Fredrik Lindsten[‡] Dave Zachariah*

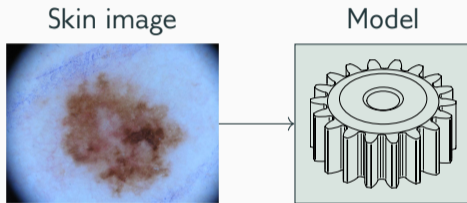
*Department of Information Technology, Uppsala University, Sweden

[‡]Division of Statistics and Machine Learning, Linköping University, Sweden

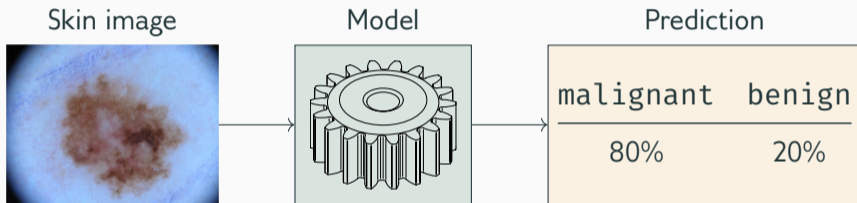
Skin image



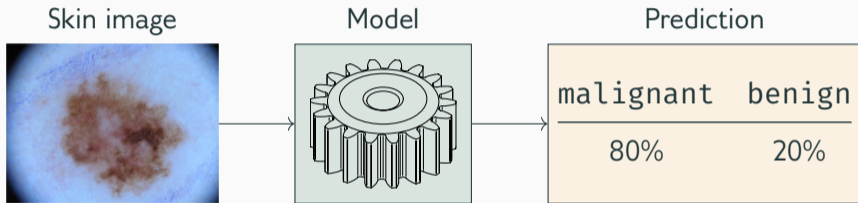
Motivation



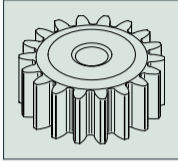
Motivation



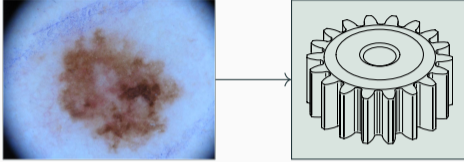
Motivation



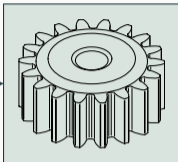
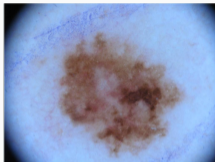
How can we ensure that the predicted confidence scores are “meaningful”?



Introduction

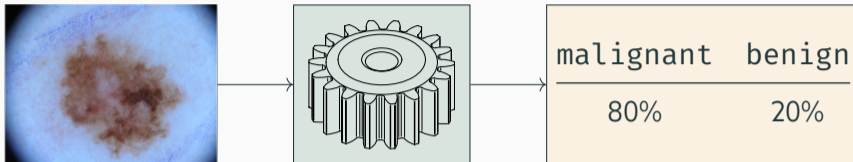


Introduction




malignant	benign
80%	20%

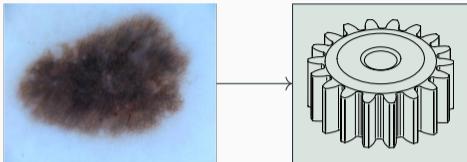
Introduction



Empirical frequency

	malignant	benign
		

Introduction



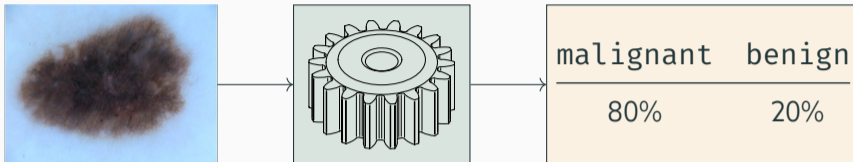
Empirical frequency

malignant


benign



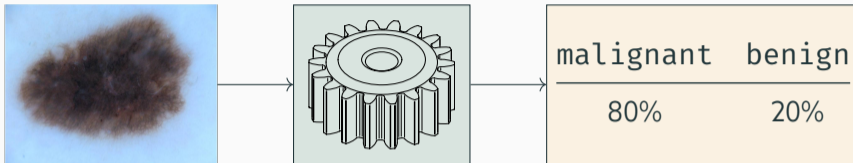
Introduction





Empirical frequency

	malignant	benign
		

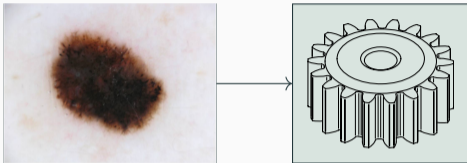
Introduction





Empirical frequency

	malignant	benign
		

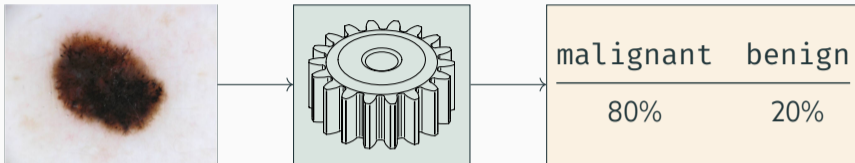
Introduction





Empirical frequency

	malignant	benign
		

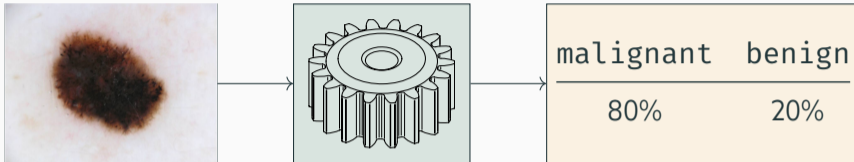
Introduction






Empirical frequency

	malignant	benign
		

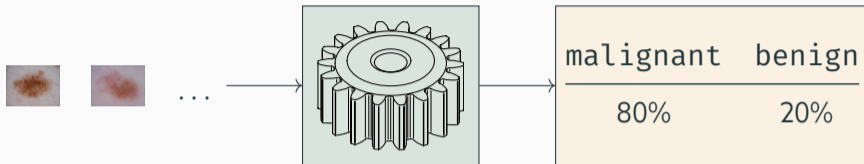
Introduction







Empirical frequency

malignant		benign
		

Introduction



Empirical frequency

malignant		benign	
			
⋮	⋮	⋮	⋮

Calibrated model

A **calibrated model** reports **predictions consistent with empirically observed frequencies** of outcomes.






Prediction

malignant	benign
80%	20%

?

=

Empirical frequency

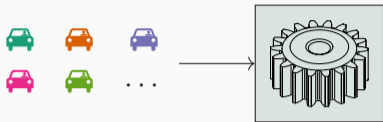
malignant				benign
				
⋮	⋮	⋮	⋮	⋮

Multi-class classification: All scores matter!



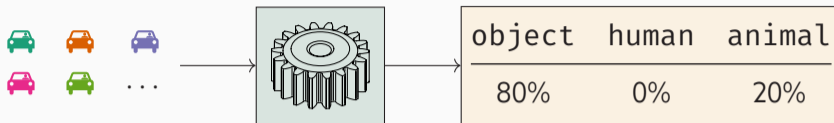
 Juozas Vaicenavicius et al. "Evaluating model calibration in classification". In: *Proceedings of Machine Learning Research* (2019)

Multi-class classification: All scores matter!



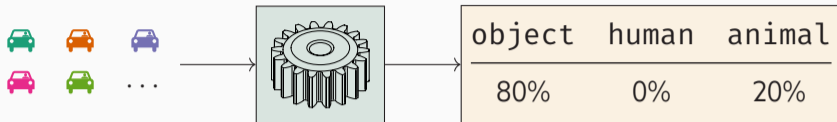
 Juozas Vaicenavicius et al. "Evaluating model calibration in classification". In: *Proceedings of Machine Learning Research* (2019)

Multi-class classification: All scores matter!



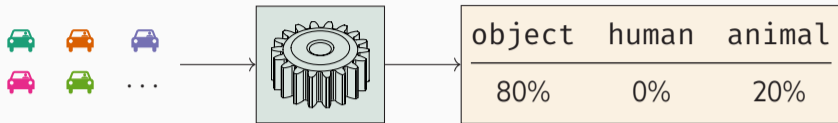
 Juozas Vaicenavicius et al. "Evaluating model calibration in classification". In: *Proceedings of Machine Learning Research* (2019)

Multi-class classification: All scores matter!



Common calibration evaluation techniques consider only the most-confident score

Multi-class classification: All scores matter!



Common calibration evaluation techniques consider only the most-confident score

Common approaches do not distinguish between the two predictions even though the control actions based on these might be very different!

object	human	animal
80%	0%	20%
80%	20%	0%

Unifying framework of calibration errors

- ▶ Based on the full predictions with all scores

Unifying framework of calibration errors

- ▶ Based on the full predictions with all scores
- ▶ Encompasses existing measures such as the expected calibration error (ECE)

Unifying framework of calibration errors

- ▶ Based on the full predictions with all scores
- ▶ Encompasses existing measures such as the expected calibration error (ECE)
- ▶ Enables derivation of a **kernel calibration error (KCE)**

Unifying framework of calibration errors

- ▶ Based on the full predictions with all scores
- ▶ Encompasses existing measures such as the expected calibration error (ECE)
- ▶ Enables derivation of a **kernel calibration error (KCE)**

Estimating calibration errors

- ▶ The standard ECE estimator is usually biased and inconsistent

Unifying framework of calibration errors

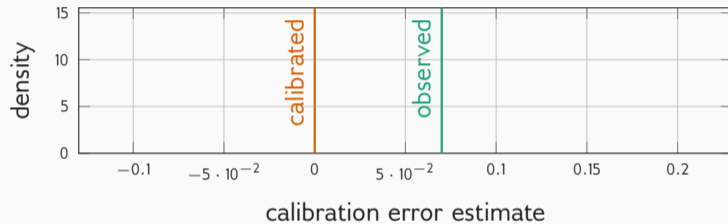
- ▶ Based on the full predictions with all scores
- ▶ Encompasses existing measures such as the expected calibration error (ECE)
- ▶ Enables derivation of a **kernel calibration error (KCE)**

Estimating calibration errors

- ▶ The standard ECE estimator is usually biased and inconsistent
- ▶ The KCE yields **unbiased** and **consistent** estimators

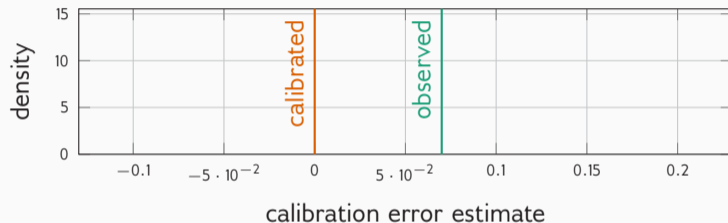
Calibration errors have no meaningful unit or scale

Our contribution: Calibration tests in multi-class classification



Our contribution: Calibration tests in multi-class classification

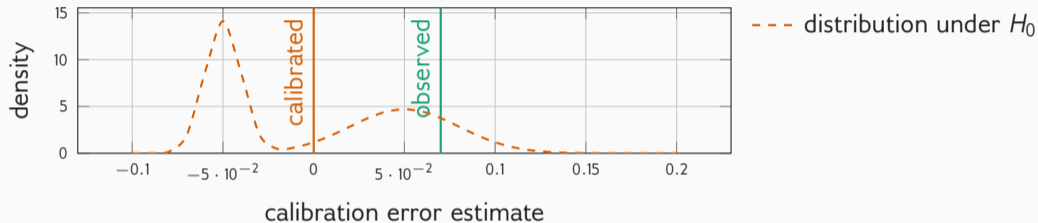
Test the null hypothesis $H_0 :=$ “model is calibrated”



- Jochen Bröcker and Leonard A. Smith. “Increasing the reliability of reliability diagrams”. In: *Weather and Forecasting* (2007)
- Juozas Vaicenavicius et al. “Evaluating model calibration in classification”. In: *Proceedings of Machine Learning Research* (2019)

Our contribution: Calibration tests in multi-class classification

Test the null hypothesis $H_0 :=$ “model is calibrated”

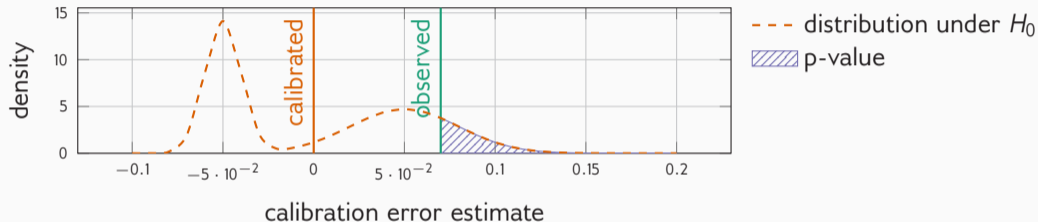


 Jochen Bröcker and Leonard A. Smith. “Increasing the reliability of reliability diagrams”. In: *Weather and Forecasting* (2007)

 Juozas Vaicenavicius et al. “Evaluating model calibration in classification”. In: *Proceedings of Machine Learning Research* (2019)

Our contribution: Calibration tests in multi-class classification

Test the null hypothesis $H_0 :=$ “model is calibrated”

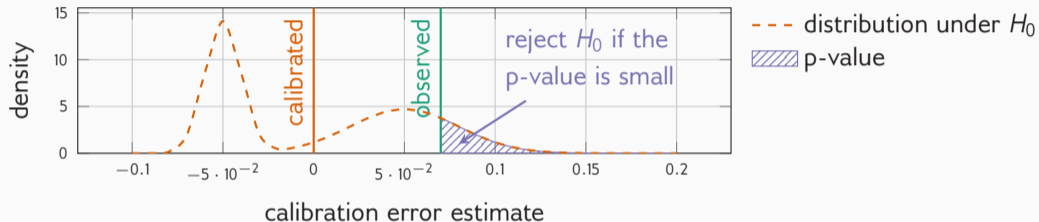


 Jochen Bröcker and Leonard A. Smith. “Increasing the reliability of reliability diagrams”. In: *Weather and Forecasting* (2007)

 Juozas Vaicenavicius et al. “Evaluating model calibration in classification”. In: *Proceedings of Machine Learning Research* (2019)

Our contribution: Calibration tests in multi-class classification

Test the null hypothesis $H_0 :=$ “model is calibrated”

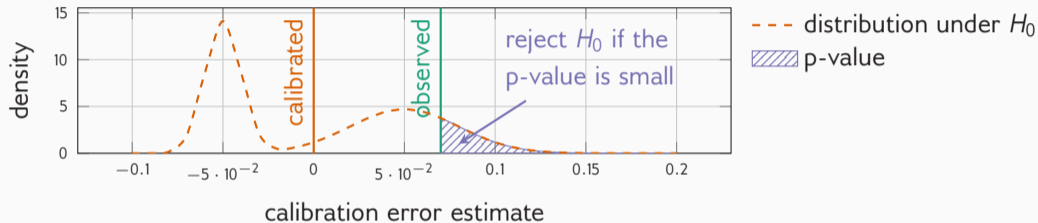


Jochen Bröcker and Leonard A. Smith. “Increasing the reliability of reliability diagrams”. In: *Weather and Forecasting* (2007)

Juozas Vaicenavicius et al. “Evaluating model calibration in classification”. In: *Proceedings of Machine Learning Research* (2019)

Our contribution: Calibration tests in multi-class classification

Test the null hypothesis $H_0 :=$ “model is calibrated”



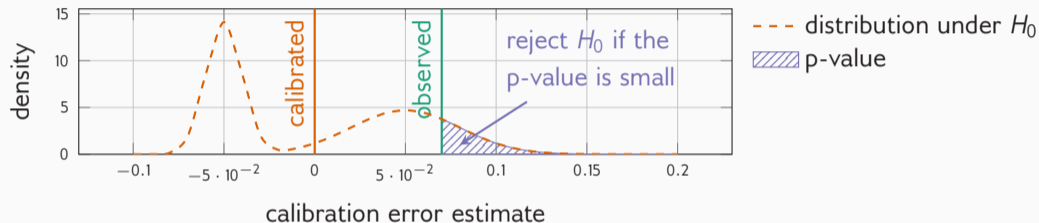
- Existing ECE-based approach seems prone to underestimating the p-value

 Jochen Bröcker and Leonard A. Smith. “Increasing the reliability of reliability diagrams”. In: *Weather and Forecasting* (2007)

 Juozas Vaicenavicius et al. “Evaluating model calibration in classification”. In: *Proceedings of Machine Learning Research* (2019)

Our contribution: Calibration tests in multi-class classification

Test the null hypothesis $H_0 :=$ “model is calibrated”



- ▶ Existing ECE-based approach seems prone to underestimating the p-value
- ▶ **Well-founded bounds and approximations** of the p-value for the KCE

📖 Jochen Bröcker and Leonard A. Smith. “Increasing the reliability of reliability diagrams”. In: *Weather and Forecasting* (2007)

📖 Juozas Vaicenavicius et al. “Evaluating model calibration in classification”. In: *Proceedings of Machine Learning Research* (2019)

Thank you for listening!
Come see our poster #39

Code available at:

<https://github.com/devmotion/CalibrationPaper>