How confident can you be?

Calibrating probabilities for AGN selection (and any other classification)

Rodrigo Carvajal
Instituto de Astrofísica e Ciências do Espaço - Universidade de Lisboa
Portugal

racarvajal@ciencias.ulisboa.pt

Astrophysical motivation

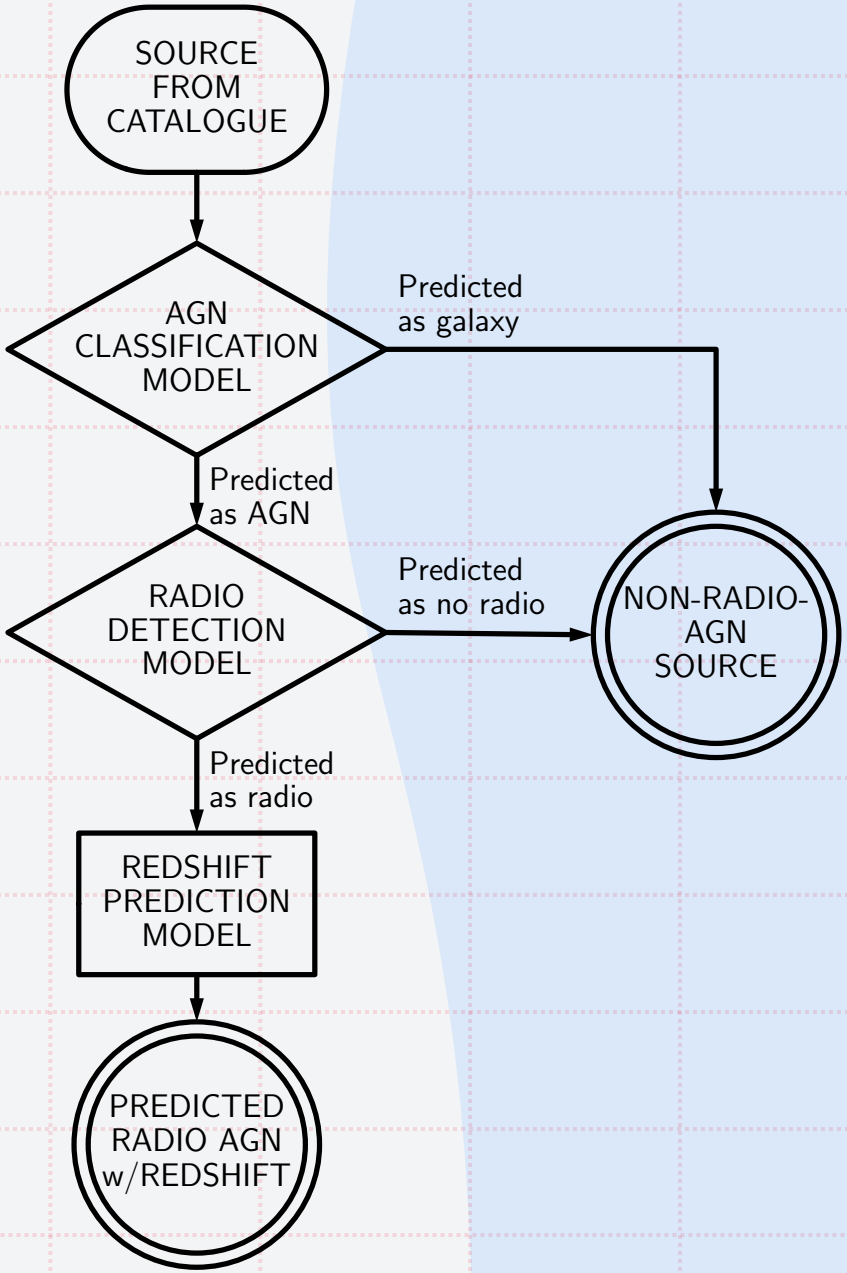
A&A 679, A101 (2023)
https://doi.org/10.1051/0004-6361/202245770
© The Authors 2023

Astronomy & Astrophysics

Selection of powerful radio galaxies with machine learning

R. Carvajal[1,2], I. Matute[1,2], J. Afonso[1,2], R. P. Norris[3,4], K. J. Luken[3,5], P. Sánchez-Sáez[6], P. A. C. Cunha[7,8], A. Humphrey[7,9], H. Messias[10,11], S. Amarantidis[12,1], D. Barbosa[1,2], H. A. Cruz[13], H. Miranda[1,2], A. Paulino-Afonso[7], and C. Pappalardo[1,2]

Pipeline of ML models



Astrophysical motivation

- Need accurate probabilities for source selection.
- Need to combine probabilities from separate models.

Machine Learning motivation

- Classifiers deliver classes and scores.
- No mathematical guarantee scores → probabilities.
- Combining results from separate models not trivial.

Probability calibration

- Make model scores representative of true likelihood.
- Allows one to use scores as proper probabilities.
- (!)Does not necessarily improve typical model metrics.

Calibration diagram

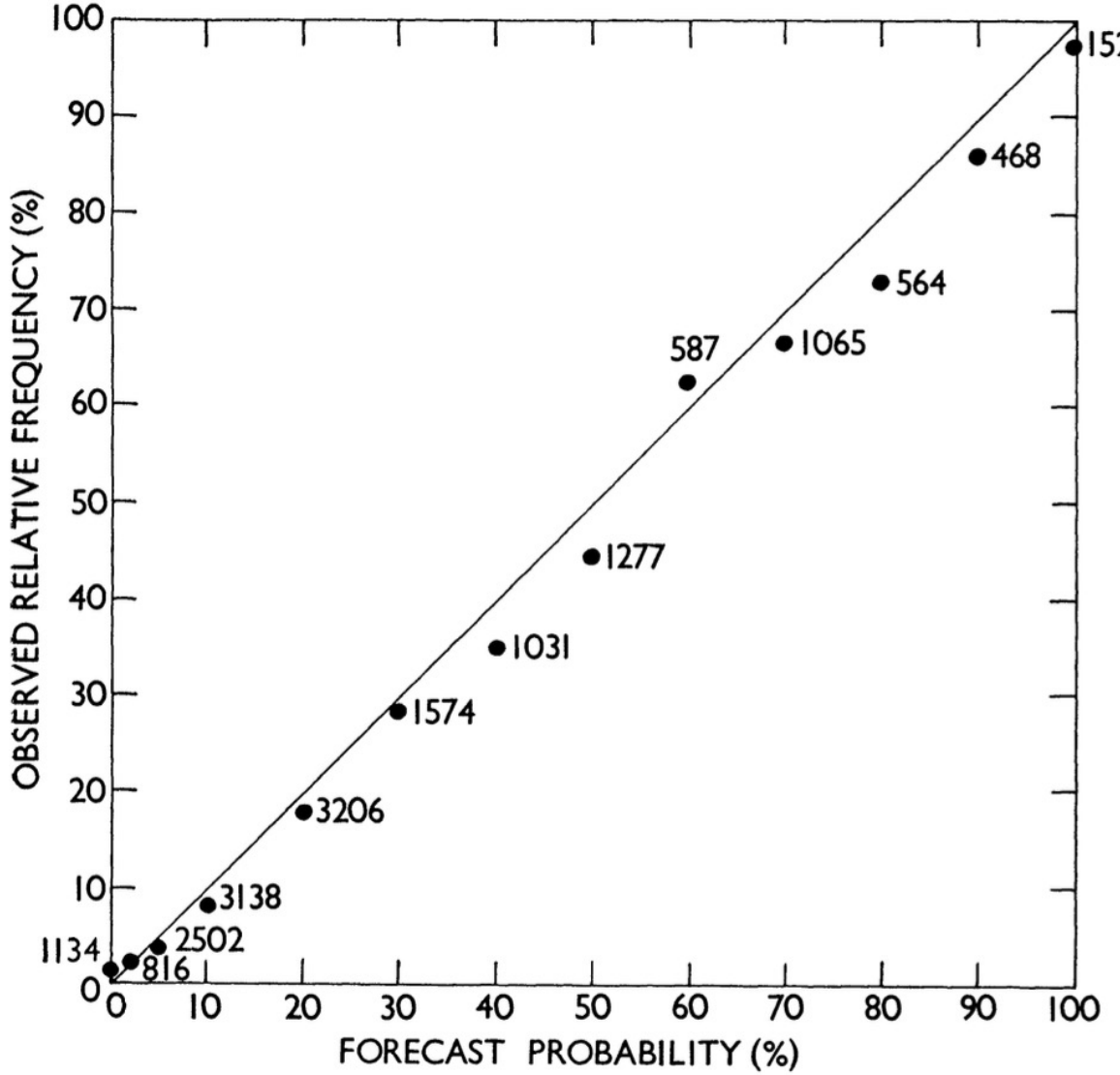


FIG. 1. The reliability diagram for all of the precipitation probability forecasts formulated by NWS forecasters at Chicago during the period from July 1972 to June 1976.

Murphy & Winkler (1977)

Niculescu-Mizil & Caruana (2005)

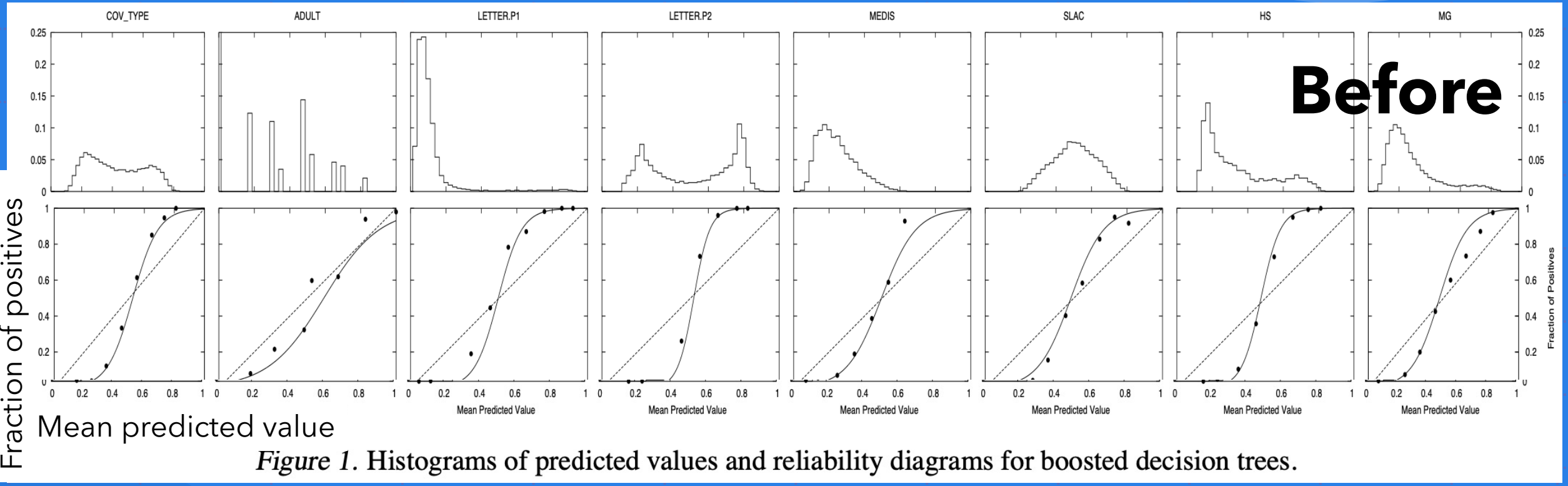
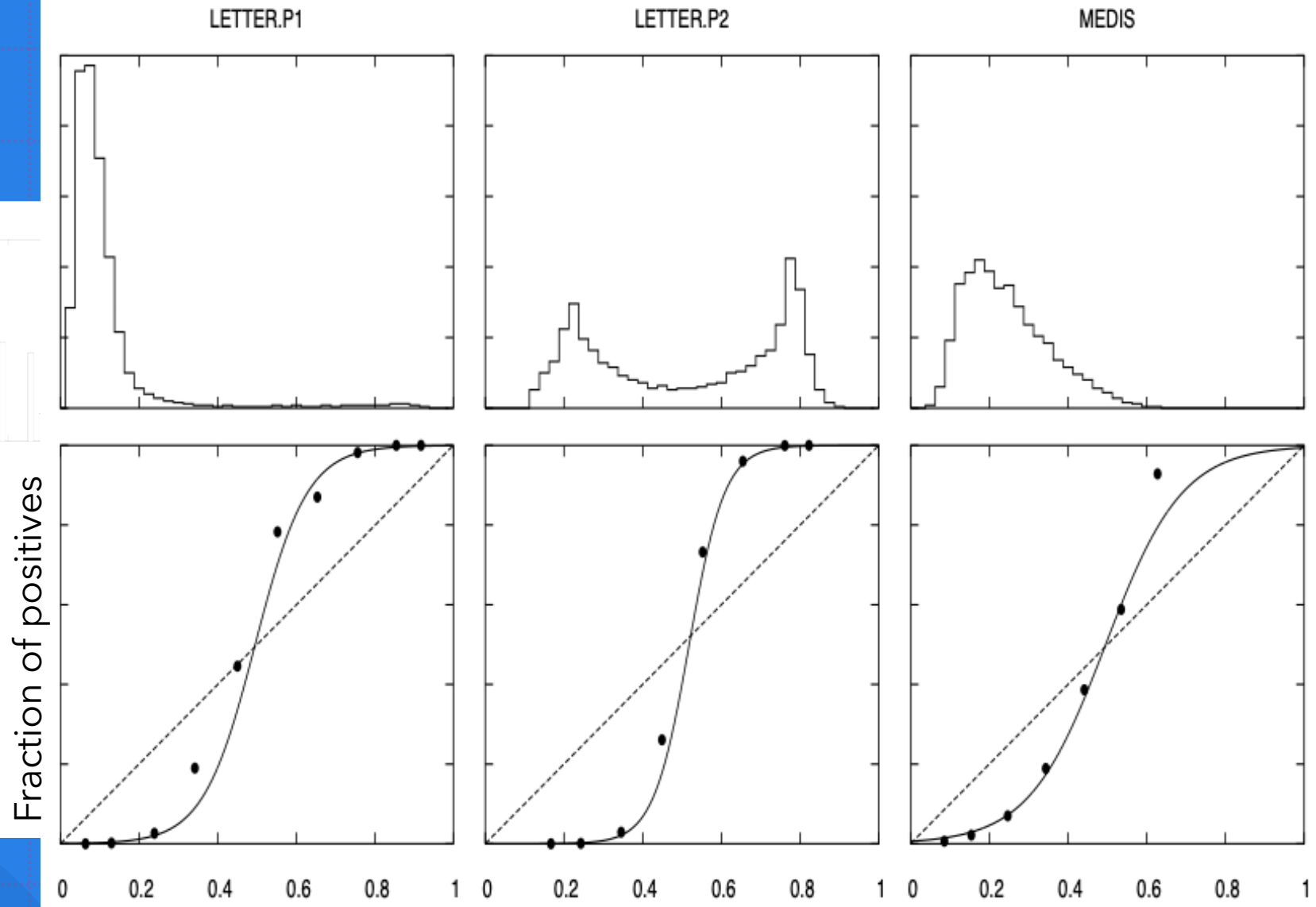


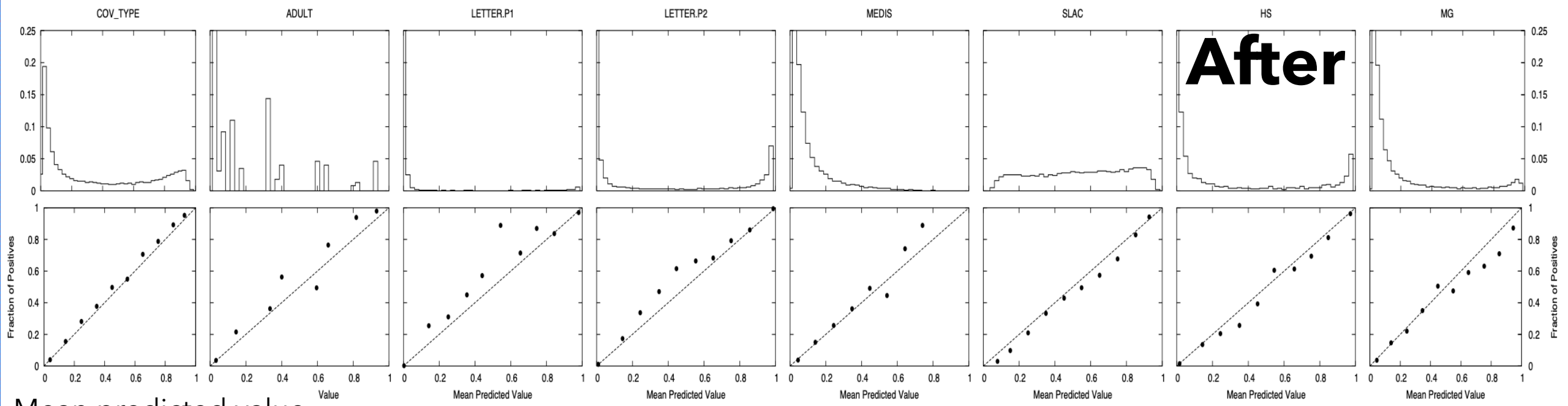
Figure 1. Histograms of predicted values and reliability diagrams for boosted decision trees.

l & Caruana (2005)

Before



Niculescu-Mizil & Caruana (2005)

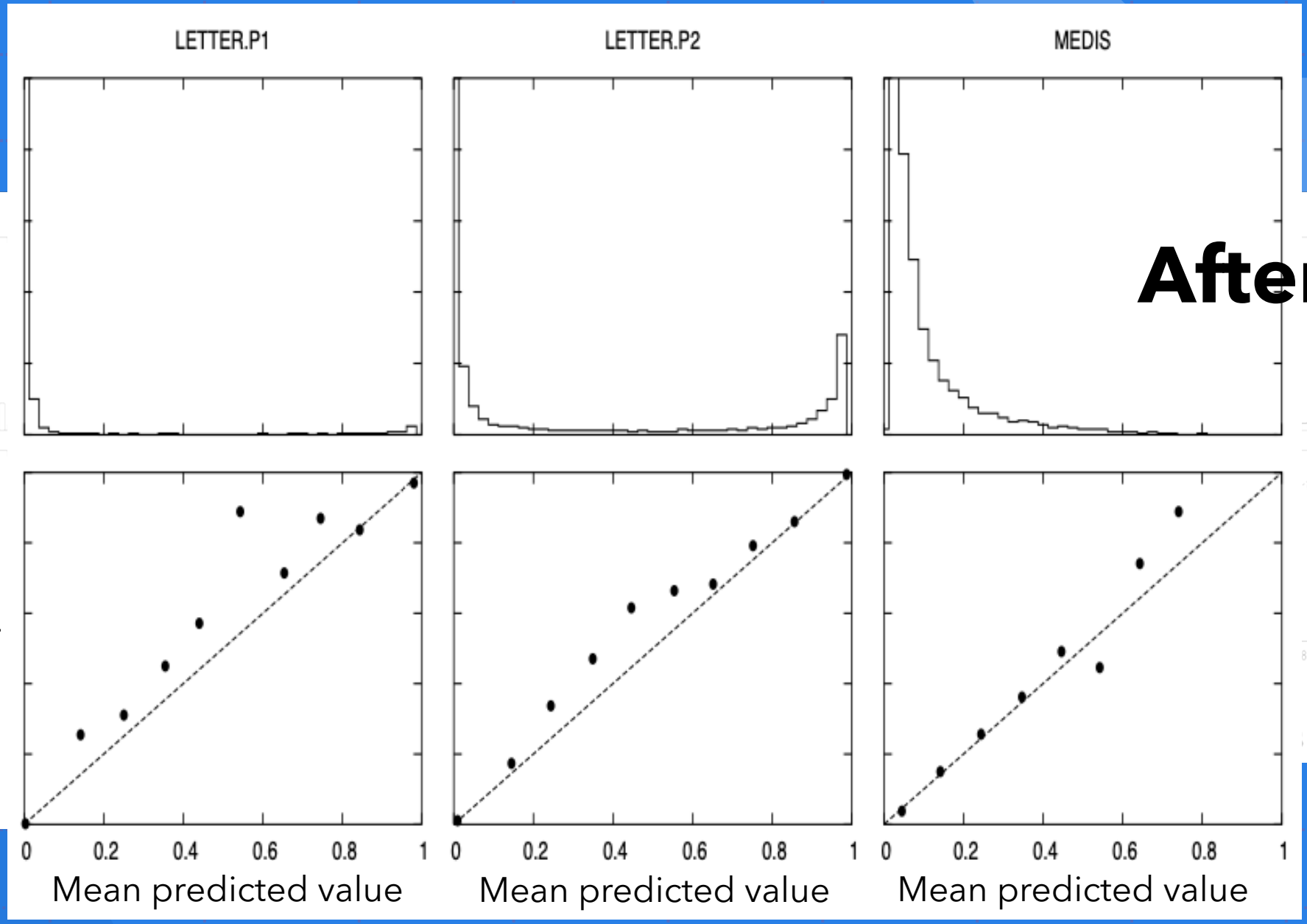


Mean predicted value

Figure 2. Histograms of predicted values and reliability diagrams for boosted trees calibrated with Platt's method.

Caruana (2005)

After



How to measure calibration?

- Brier Score (Brier loss)

$$\mathrm{BS} = \frac{1}{n_{\mathrm{samples}}} \sum_{i=0}^{n_{\mathrm{samples}}-1} (y_i - p_i)^2$$

- Brier Skill Score

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}$$

Probability calibration in Python

- One (additional) step to train a “mapper” between scores and calibrated probabilities.
- Re-distributes scores in [0, 1] range.
- Need to use separate subset (or cross-validation) to avoid data leakage.

One example: Beta Calibration

Kull et al. (2017)

https://betacal.github.io

$$\mu_{beta}(s;a,b,c) = \frac{1}{1+1\Big/\left(e^{c}\frac{s^{a}}{(1-s)^{b}}\right)}$$

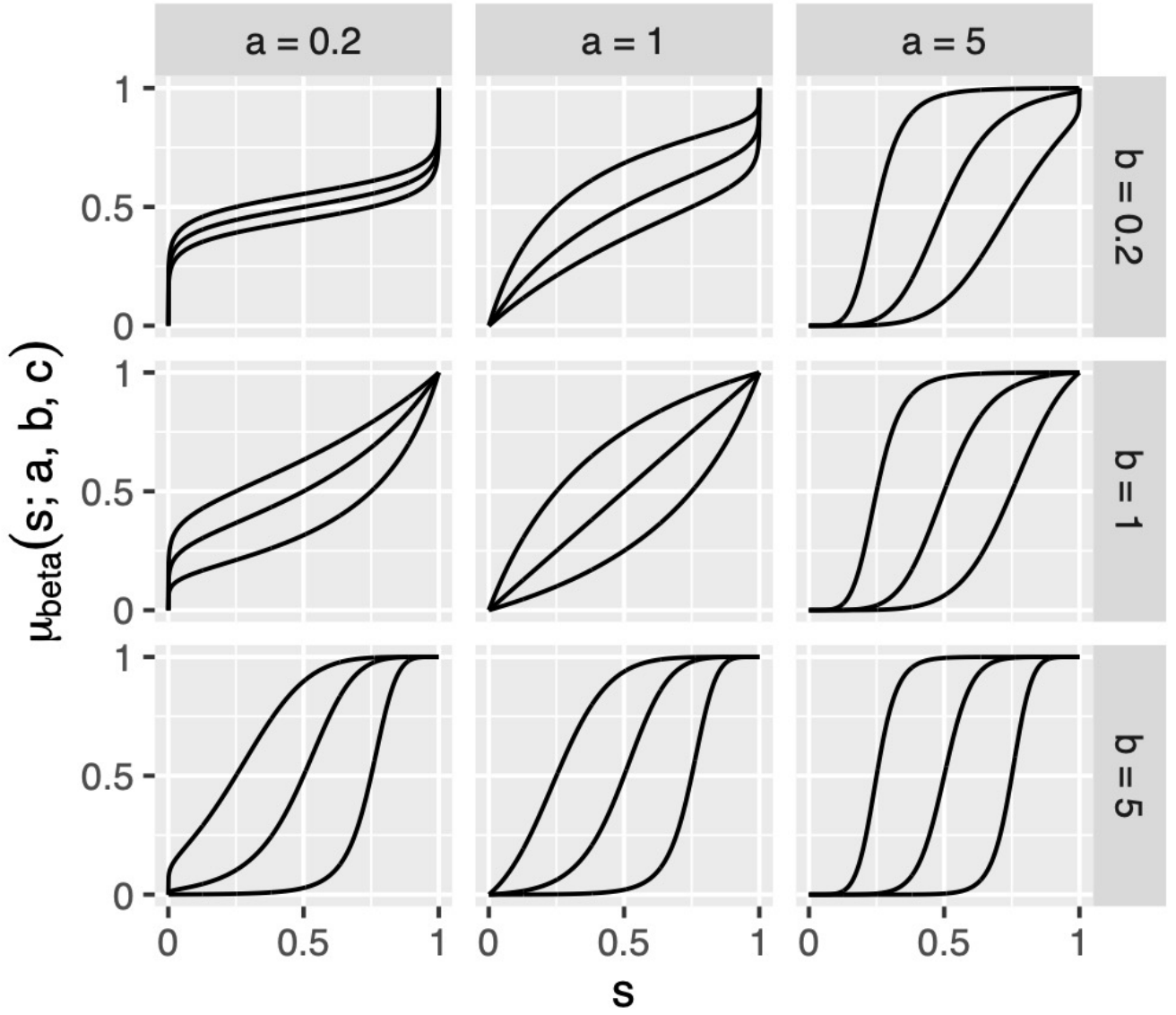


Figure 3: Examples of beta curves with parameters $a, b \in \{0.2, 1, 5\}$, $m \in \{0.25, 0.5, 0.75\}$ and $c = b\ln(1-m) - a\ln m$.

Kull et al. (2017)

Beta Calibration

```
from betacal import BetaCalibration
from sklearn.ensemble import RandomForestClassifier

# With no calibration
clf                 = RandomForestClassifier(n_estimators=10)
clf.fit(X_train, y_train)
prob_pos_clf        = clf.predict_proba(X_cal)[:, 1]
prob_pos_clf_test = clf.predict_proba(X_test)[:, 1]

# With Beta calibration
# Fit three-parameter beta calibration
bc                  = BetaCalibration(parameters="abm")
bc.fit(prob_pos_clf.reshape(-1, 1), y_cal)
prob_pos_beta   = lr.predict(prob_pos_clf_test)
```

Calibration diagram (again)

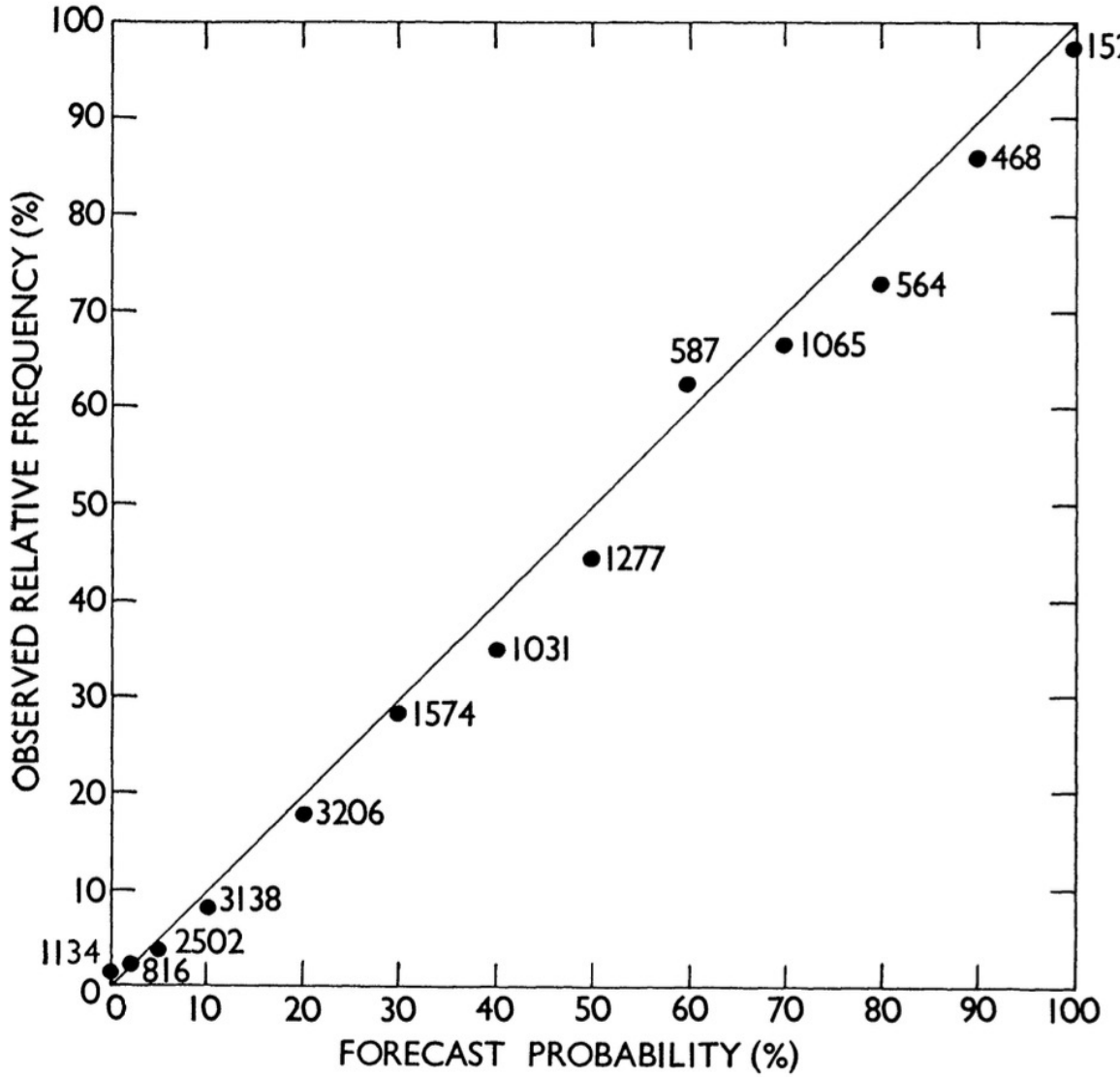
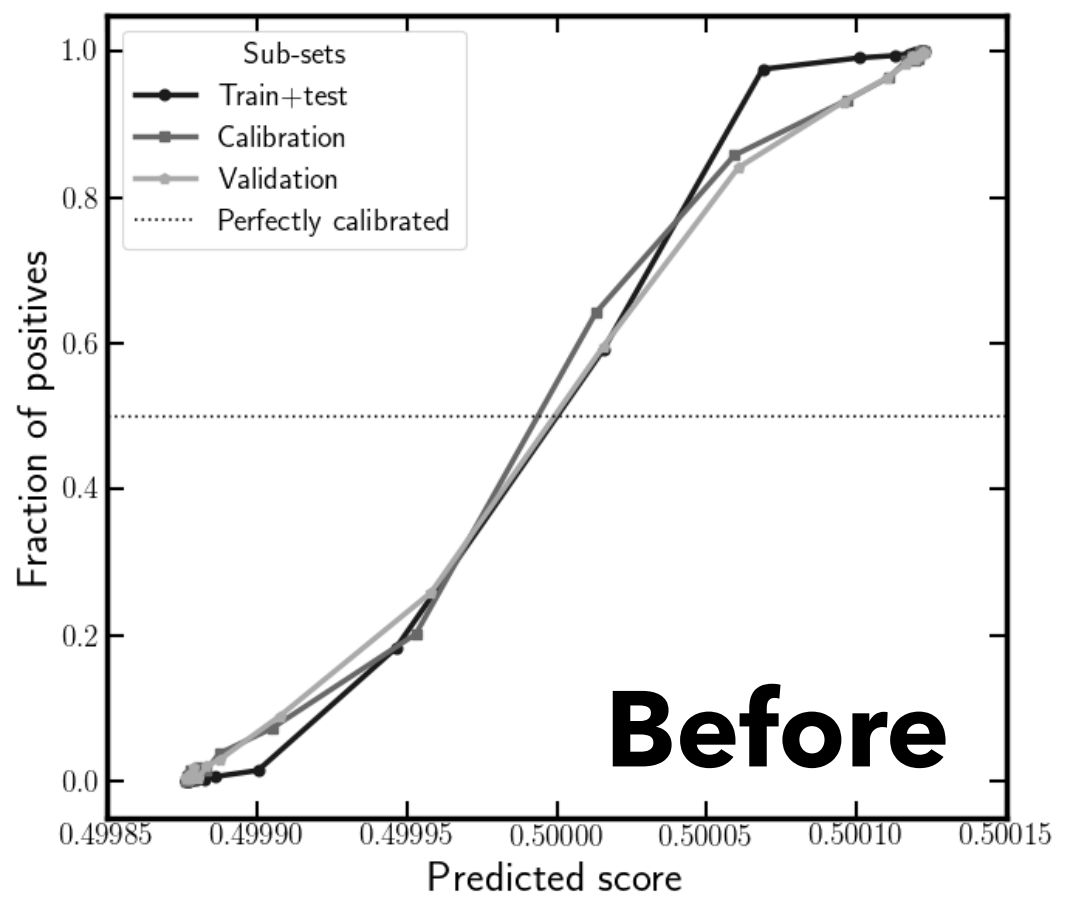
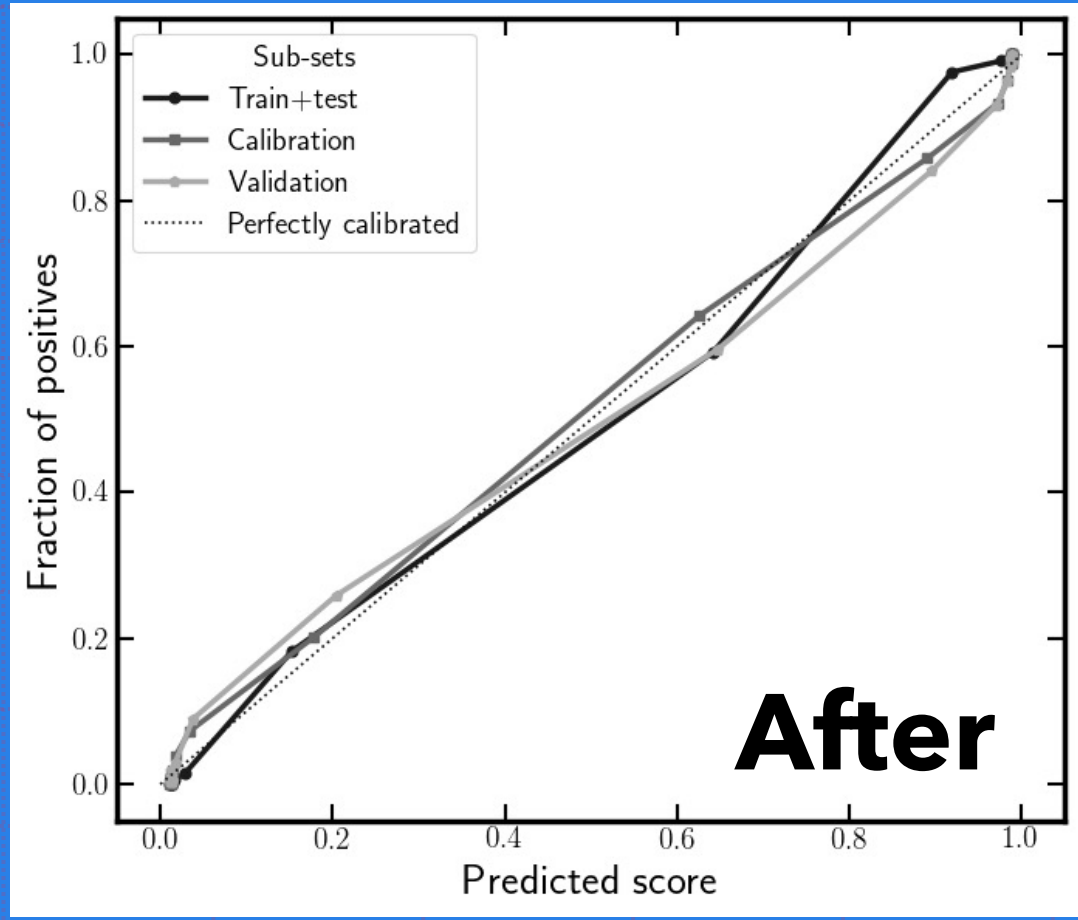


FIG. 1. The reliability diagram for all of the precipitation probability forecasts formulated by NWS forecasters at Chicago during the period from July 1972 to June 1976.

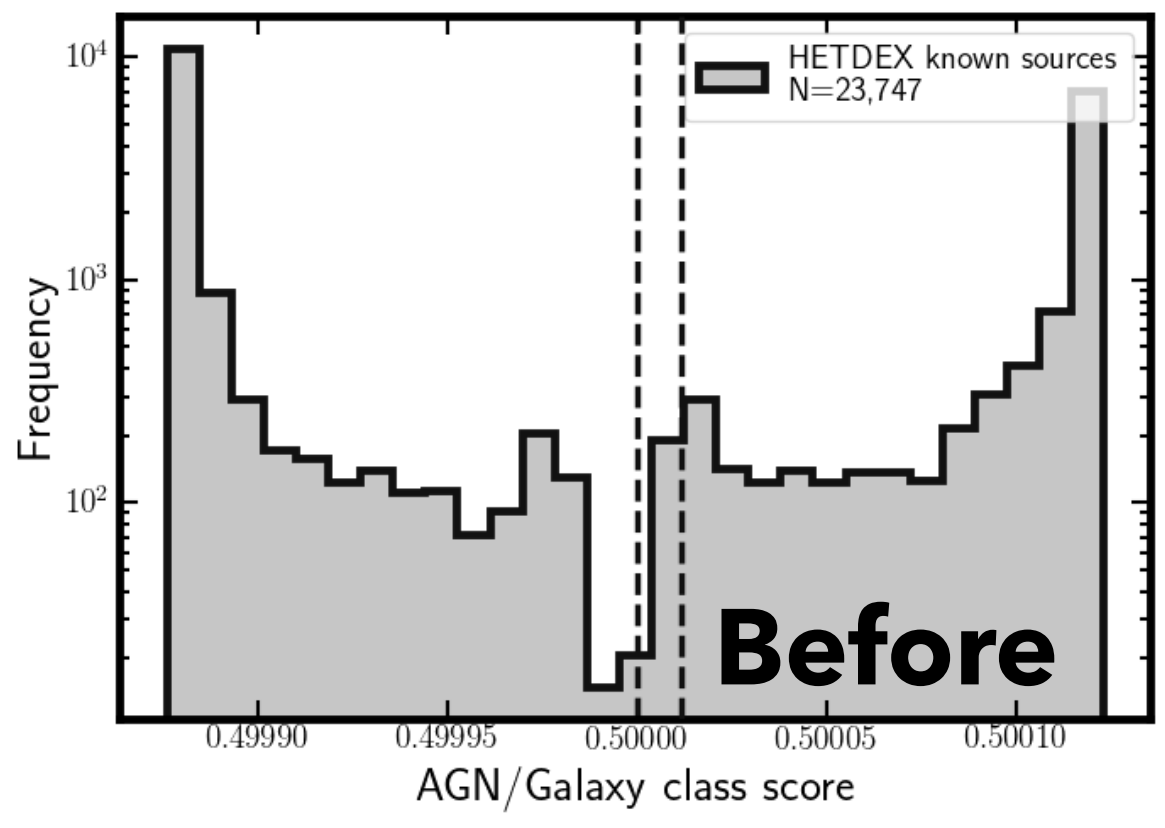
Murphy & Winkler (1977)

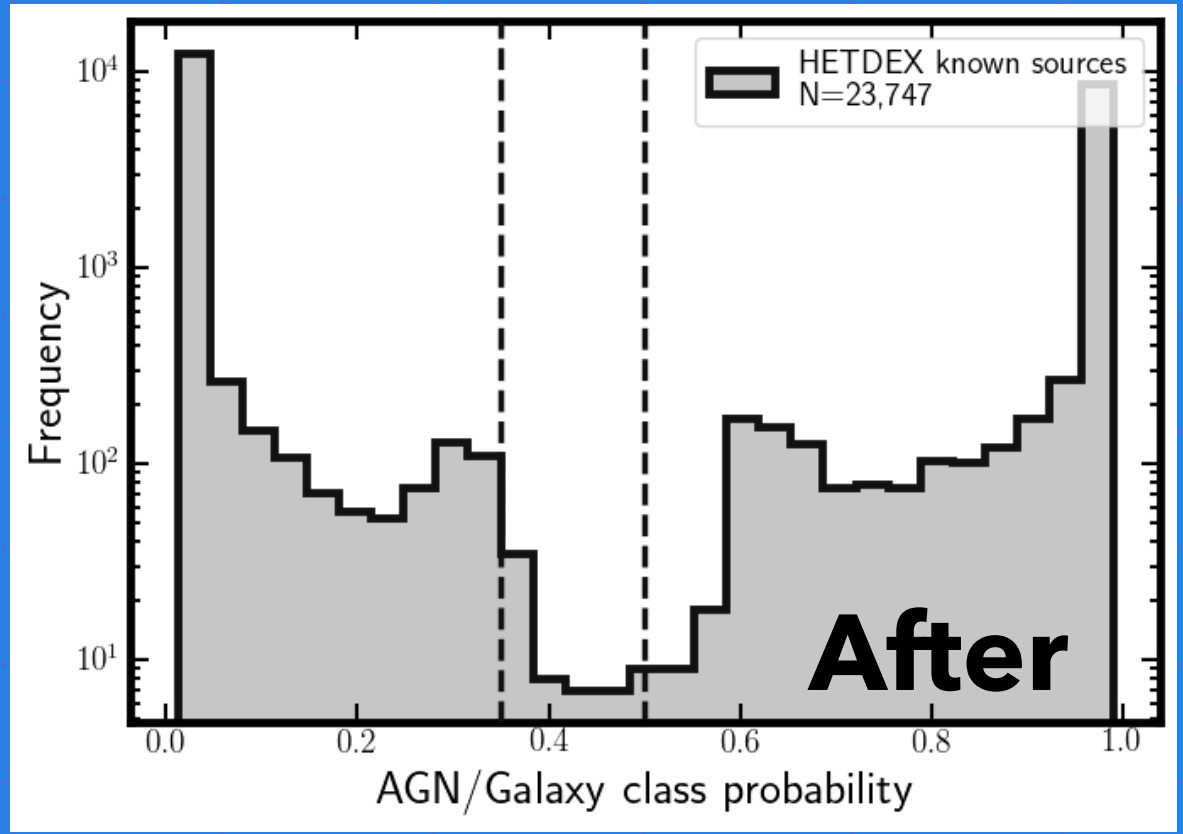
AGN / SFG classifier (Carvajal+2023)



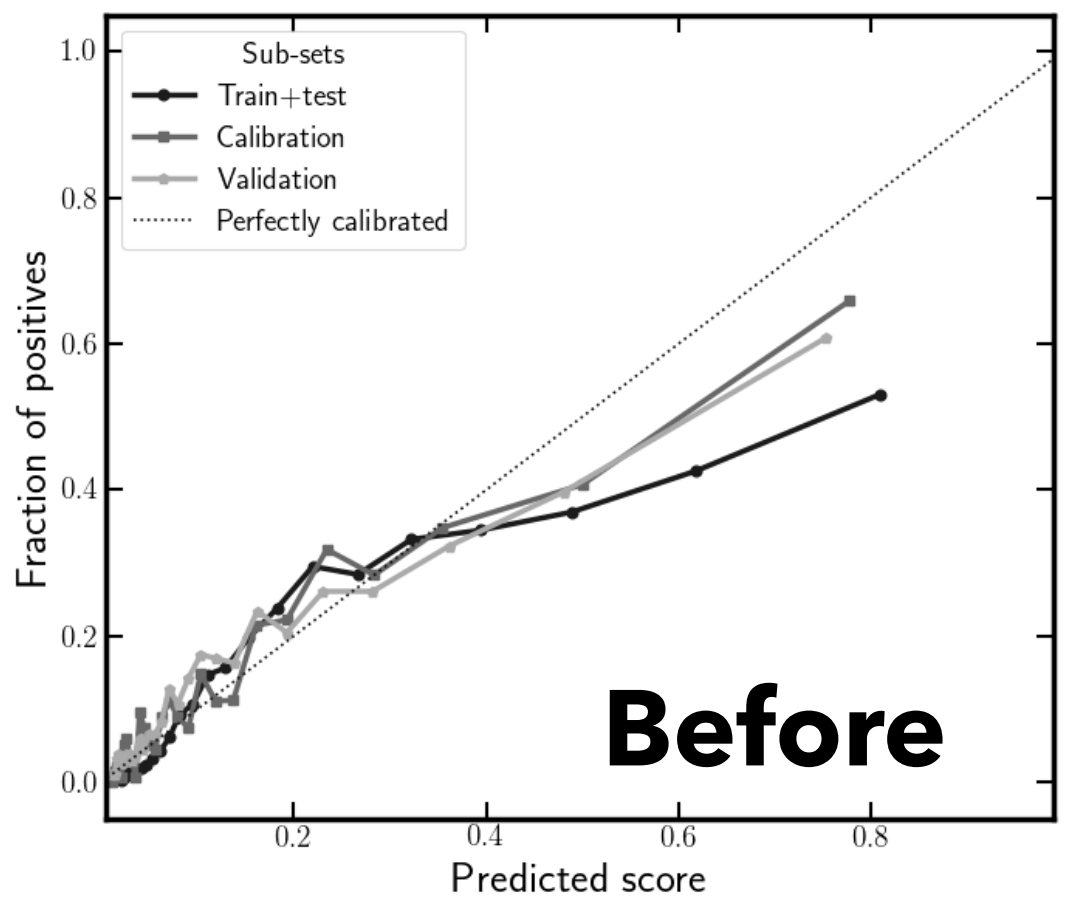


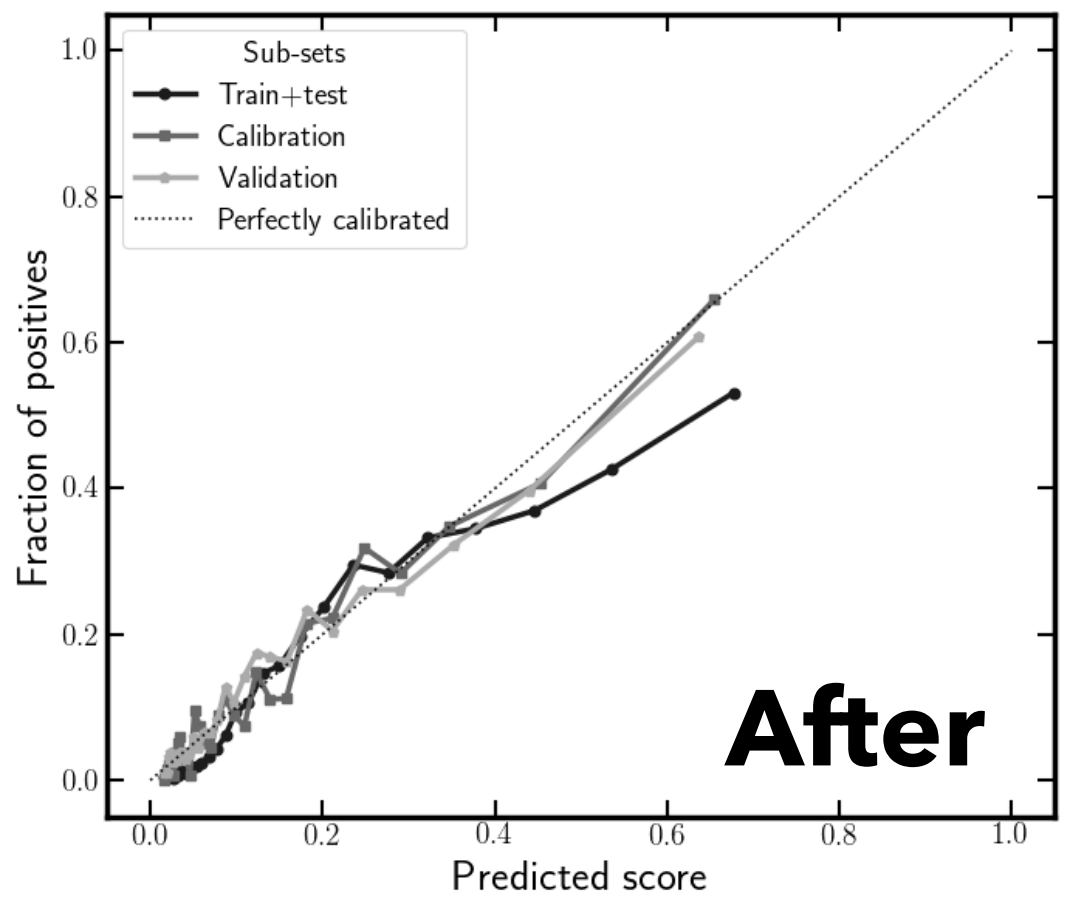
AGN / SFG classifier (Carvajal+2023)



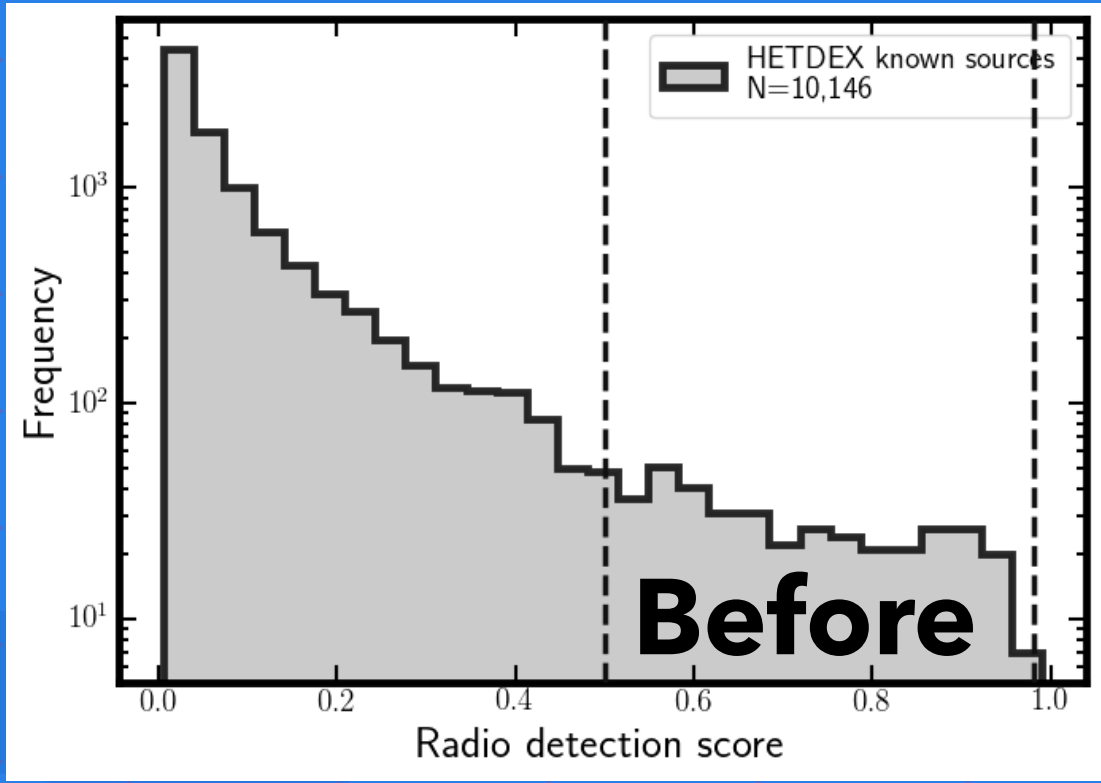


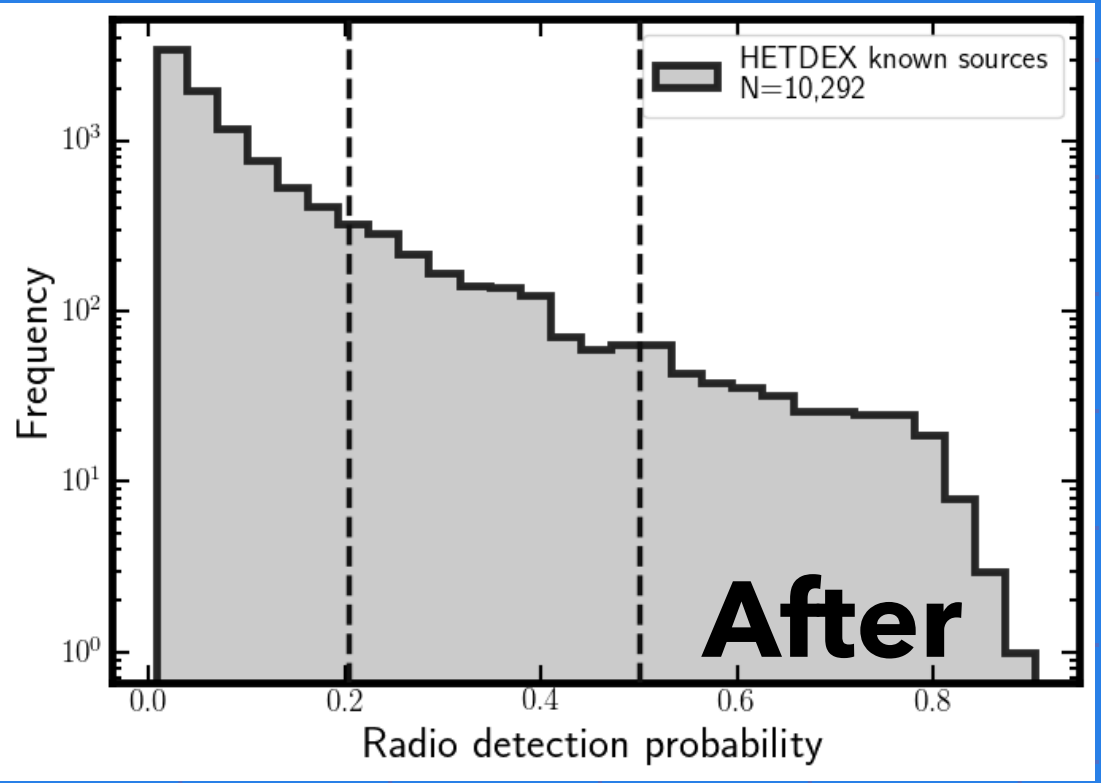
Radio detection classifier (Carvajal+2023)





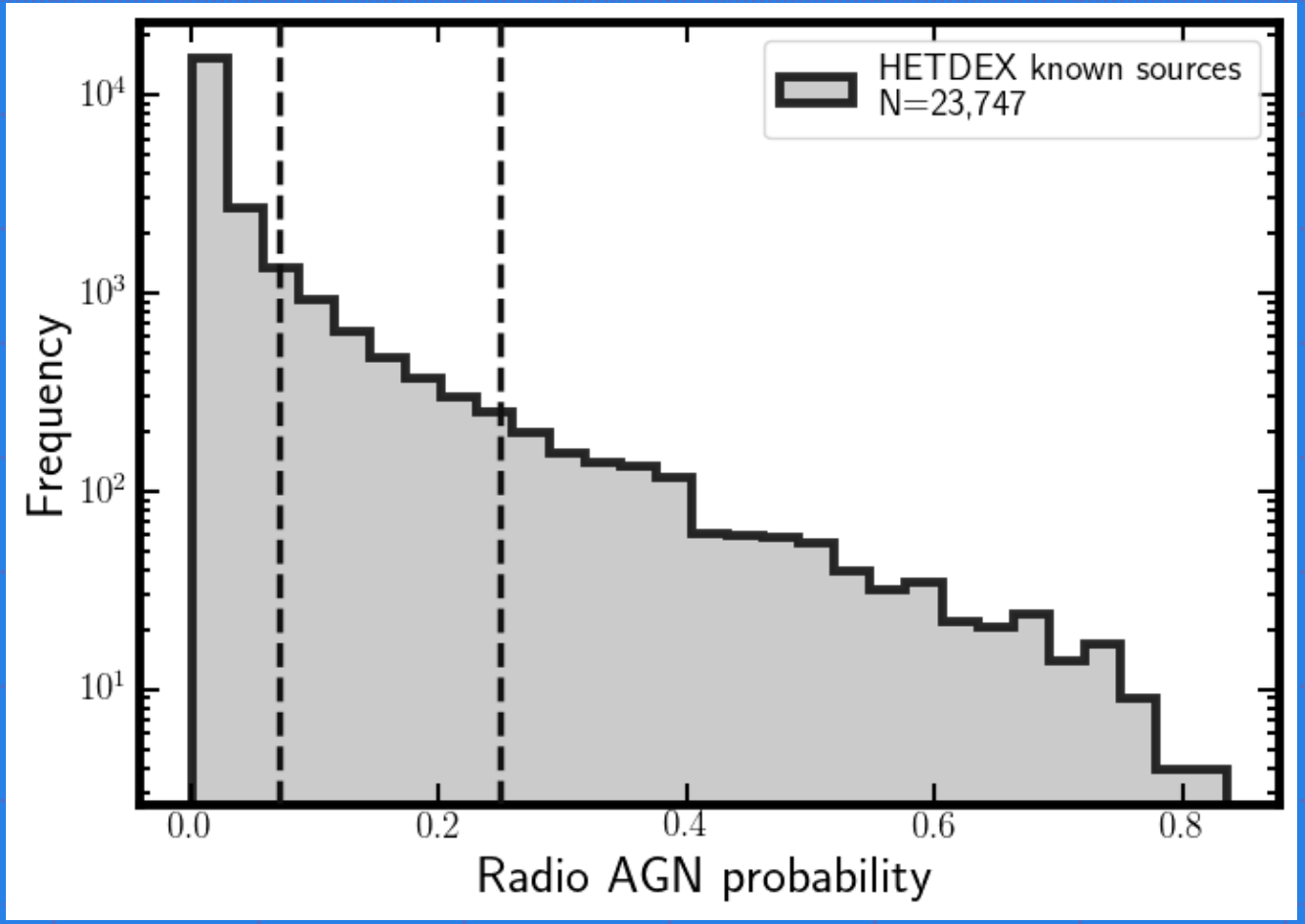
Radio detection classifier (Carvajal+2023)





Combined probabilities (Carvajal+2023)

$$P_{\mathrm{rAGN}} = P_{\mathrm{AGN}} \times P_{\mathrm{radio}}$$



To conclude...

- Classification scores are not always probabilities.
- Scores can be calibrated.
- Calibration does not have improving metrics as a goal.
- In Python, calibration is one additional model to be trained.

How confident can you be?

Calibrating probabilities for AGN selection (and any other classification)

Rodrigo Carvajal
Instituto de Astrofísica e Ciências do Espaço - Universidade de Lisboa
Portugal

racarvajal@ciencias.ulisboa.pt

What about Neural Networks?

Neural Networks

- Early NNs were very close to calibrated.
- Modern NNs (and intermediate steps) deteriorate reliability.
- Special care needed if goal are probabilities.
- ArXiv: 1706.04599 (Guo+2017)