

Metrics for Deploying Human-in-the-Loop Classification Models

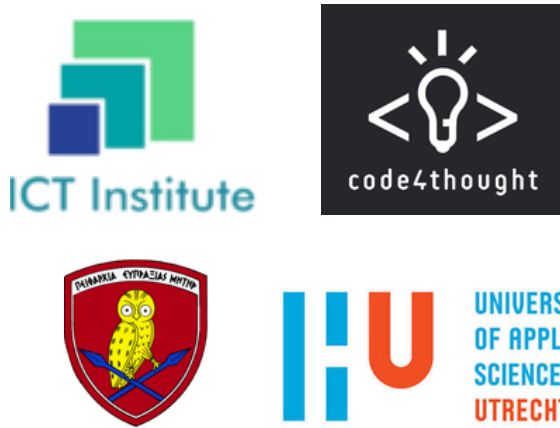
Due to new legislation and AI risk awareness, many AI decision models now defer uncertain predictions to human reviewers. Yet models are selected on metrics like accuracy and precision, which assume fully automated decisions. With humans in the pipeline, these metrics describe a system never deployed, can favour the wrong model, and obscure the cost of human intervention. Building on selective classification, we propose Deferral-Aware-F1 (DAF1), the F1 score over cases a classifier accepts at a given coverage, and Cost-Aware-Score (CAS), which weights DAF1 by relative human review cost. Both use confidence scores to defer decisions. Using UNSW-NB15 and a SOC alert-triage scenario requiring AI Act supervision, we show how our metrics change model rankings and error rates and support trade-off discussions between engineering, operations and business.

Authors

Pavlo Burda¹
Petros Ioannis Stavroulakis^{2,3}
Sieuwert van Otterloo^{1,4}

Affiliations

¹ICTInstitute, NL
²Code4Thought, GR
³HellenicMilitaryAcademy, GR
⁴HogeschoolUtrecht, NL



1 Why?

GDPR Article 22, AI Act Article 14 and everyday practice all place a human reviewer behind many deployed classifiers. Yet models are still selected on accuracy or F1, metrics that assume every decision is final. As a result, the wrong model can end up ranked first.

2 The metrics

Confidence = | $\hat{p}$ - t |. The least-confident share $1 - c$ is deferred to the human; c is the coverage.

DAF1: F1 over the accepted cases $A(c)$ only:

$$F_1(c) = \frac{2 TP_{A(c)}}{2 TP_{A(c)} + FP_{A(c)} + FN_{A(c)}}$$

CAS: DAF1 discounted by the cost α of a review relative to a misclassification; its maximum c^* is the operating point:

$$CAS(\alpha, c) = F_1(c) (1 - \alpha (1 - c))$$

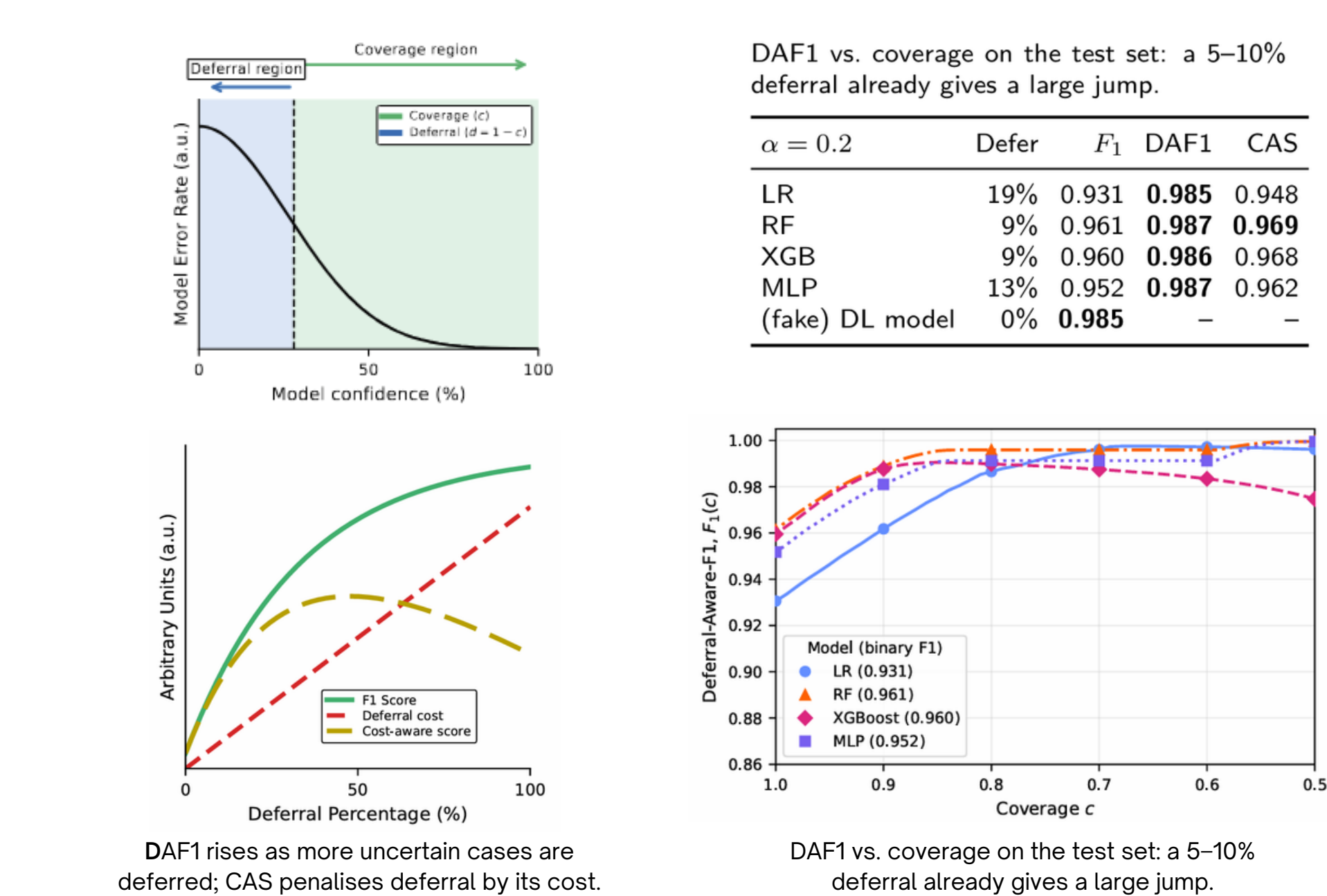
4 Example dataset

- UNSW-NB15 SOC intrusion data ($N = 257,673$); with LR, RF, XGBoost and MLP.
- Low-confidence alerts go to a SOC analyst; realistic budget $\approx 10\%$.

3 Example outcomes

- Deferring 10% increases LR's F1 from 0.93 to 0.96: about half of the FPs and FNs disappear.
- Under CAS, RF stays best but XGB needs the fewest deferrals; costlier reviews shift the optimum to less deferral.
- A fictitious deep model with $F1 = 0.985$ loses to all four simple models at $\alpha \leq 0.2$.

5 Figures



6 Takeaway

Reporting DAF1 next to F1 and showing CAS curves makes discussions between engineers, operations and business simpler.

