Error Overlap in Human-AI Complementarity

Afifah Kashif
University of Cambridge
Centre for Human Inspired AI

Abstract

The default when deploying a model alongside a human decision-maker is to optimize for accuracy, but this work argues that this is insufficient. Across three binary classification tasks, we train five models per task, pair each with a simulated domain expert, and compare team performance under confidence-threshold deferral and learned deferral. The main finding is that error overlap between the model and the human, rather than each one's standalone accuracy, predicts team performance. Lower overlap is associated with stronger collaboration, most clearly on Heart Disease ($r = -0.86$), with the same direction on German Credit ($\rho = -0.60$) and Breast Cancer. On Heart Disease, the least accurate model forms the best team in expectation under learned deferral, beating the most accurate model's team by 2.36 points, though this gap is not significant at conventional levels ($p = 0.28$, Nadeau-Bengio corrected), a consequence of small test folds. Risk-coverage curves, the standard selective-prediction diagnostic, cannot detect this effect.

1 Introduction

The standard way to pick a model for deployment is to take the one with the best held-out accuracy. This stays the default even when there is a human in the loop, such as a doctor reviewing cases or a loan officer checking recommendations. The assumption is that the best model makes the best teammate.

However, in teamwork, what matters for a team is not how often each member is right, but also where each is right. If the model fails on the same cases the human finds hard, deferral cannot help. If the model fails on easy cases the human would catch, the team benefits even if the model is worse overall. This idea is formalized as the complementarity gap [10]: the reduction in risk achievable by optimally routing between model and human. Okati et al. [7] show that a model trained for standalone accuracy is generally suboptimal under triage. This work showcases that empirically in three ways.

First, models ranked by standalone accuracy produce a different ranking when evaluated as teammates, as on Heart Disease the accuracy ranking (1-5) becomes (3, 5, 3, 2, 1) under learned deferral. Second, error overlap is negatively correlated with team gain across all three datasets. Third, risk-coverage curves [2] are blind to this, because they say nothing about the human's accuracy on deferred inputs.

2 Setup

2.1 Datasets

We use three binary classification tasks with a "model + expert" workflow:

Heart Disease (UCI Cleveland, n=297, 13 features, 46% prevalence). Predict heart disease from clinical measurements.

German Credit (UCI Statlog, n=1,000, 20 features, 30% default rate). Predict bad credit risk from financial and personal attributes.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Breast Cancer Wisconsin (sklearn, n=569, 30 features, 37% malignant). Predict malignancy from fine-needle aspirate measurements.

2.2 Models

For each dataset we train five deliberately diverse models, all calibrated with isotonic regression [3]: Logistic Regression (L2, C=1.0), Logistic Regression (L1, C=0.3), Random Forest (150 trees, depth 6), Gradient Boosting (100 trees, depth 3, lr 0.1), and k-NN (k=15). The point of the diversity is to produce different error patterns, not to find the best single classifier. We repeat the train-test split 10 times and average the results.

2.3 Simulated Human Expert

Each dataset gets a simulated expert, which is a logistic regression (C=0.3) on a small, domain-plausible feature subset with Gaussian noise (σ=0.08) on its predicted probabilities. The cardiologist uses age, chest pain type, exercise-induced angina, ST depression, and number of major vessels (accuracy: 79.4%). The loan officer uses duration, amount, employment length, age, and housing (74.4%). The pathologist uses mean radius, texture, smoothness, compactness, and worst radius (94.2%).

This work does not claim that these are realistic clinicians. What matters is the structural property, the "human" uses fewer features and a simpler decision boundary, so its error pattern differs from the richer models in predictable ways.

2.4 Metrics and Deferral

Let $m(x)$ and $h(x)$ be the model's and human's predictions. We measure:

Error overlap: the fraction of total errors that are shared between the model and the human. This ranges from 0 (completely different errors) to 1 (identical errors).

Complementarity gap [10]: $\Delta_{\text{comp}} = \min(R_{\text{model}}, R_{\text{human}}) - R(\pi^*)$, where π^* always picks whichever agent is correct.

Two deferral strategies are compared, plus an oracle upper bound.

- **Confidence thresholding**: the model handles inputs where $\max_y \hat{p}(y|x) \geq \tau$, the rest go to the human; we sweep τ and report the best. This asks "is the model uncertain?"
- **Learned deferral** [5]: the label space is augmented with a deferral class $\perp$ following Mozannar and Sontag's surrogate, training on cross-validated model errors. This asks "is the human better here?" [9]
- **Oracle deferral**: always picks whichever agent is right, not achievable but bounds what is available

3 Results

3.1 Accuracy Rankings Flip Under Collaboration

Table 1 shows the most important study takeaways. On Heart Disease, Gradient Boosting is **last** by standalone accuracy ($80.8 \pm 6.8\%$) but **first** by learned-deferral team accuracy ($84.9 \pm 6.4\%$). It manages this because it has the lowest error overlap with the cardiologist (0.474), giving it the highest complementarity gap (5.1%). The most accurate model (Logistic L1, 82.2%) forms a worse team. This was an unexpectedly obvious result, as the accuracy ranking nearly inverts for the best and worst models.

On German Credit, Gradient Boosting (best by accuracy) drops to third under learned deferral, while Random Forest (second by accuracy) takes first. The effect is present but less dramatic than on Heart Disease. On the Breast Cancer dataset, where performance exceeds 95% across the board, the effect largely disappears because there is little room for complementarity when all predictors are already highly accurate.

Table 1: Standalone accuracy, error overlap with human, and team accuracy under three deferral strategies. Mean ± std over 10-fold CV. **Acc.R**: rank by accuracy. **Lrn.R**: rank by learned-deferral team accuracy. Bold: best team per dataset. The first row per dataset (italic) shows the human expert alone. No pairwise differences in learned-deferral team accuracy are significant after Nadeau-Bengio correction [6].

	Model	Acc.	Overlap ↓	Δ_{comp}	Oracle	Thresh.	Learned	Acc.R	Lrn.R
Heart	*Human*	*.794*	—	—	—	—	—	—	—
	Log. L1	.822±.080	.506±.190	.037	.865±.083	.849±.086	.825±.044	1	3
	Log. L2	.818±.082	.544±.185	.040	.859±.082	.842±.082	.825±.050	2	5
	k-NN	.815±.076	.462±.170	.054	.872±.077	.849±.082	.825±.047	3	3
	Rand. F.	.812±.084	.513±.210	.037	.862±.087	.845±.102	.832±.047	4	2
	Grad. B.	.808±.068	**.474**±.160	**.051**	**.872**±.061	.831±.074	**.849**±.064	5	**1**
Credit	*Human*	*.702*	—	—	—	—	—	—	—
	Grad. B.	.779±.036	.474±.095	.054	**.833**±.038	.785±.034	.736±.040	1	3
	Rand. F.	.765±.045	.488±.075	.059	.825±.035	.777±.035	**.754**±.052	2	**1**
	Log. L2	.763±.032	.461±.069	.068	.831±.030	.772±.026	.724±.037	3	5
	Log. L1	.761±.029	**.451**±.086	**.072**	.833±.036	.770±.025	.730±.042	4	4
	k-NN	.755±.026	.513±.068	.061	.816±.027	.766±.030	.747±.029	5	2
Breast	*Human*	*.942*	—	—	—	—	—	—	—
	Log. L2	.982±.012	.215±.206	.005	**.988**±.012	.968±.020	.951±.018	1	2
	Log. L1	.968±.022	.346±.231	.005	.977±.017	.963±.015	.949±.023	2	4
	k-NN	.956±.029	.351±.136	.012	.974±.017	**.970**±.017	.951±.022	3	2
	Grad. B.	.953±.035	**.332**±.218	**.014**	.974±.021	.963±.024	**.954**±.026	4	**1**
	Rand. F.	.951±.028	.320±.209	.019	.974±.017	.963±.024	.949±.030	5	4

3.2 Error Overlap Predicts Team Gain

Figure 1 plots error overlap against oracle team gain (team accuracy minus the better individual). On Heart Disease, the negative correlation is clear: $r = -0.86$ ($p = 0.059$). On German Credit the direction is the same ($\rho = -0.60$) but does not reach significance ($p = 0.28$), partly because the spread in overlap is narrower. On Breast Cancer the correlation weakens further ($r = 0.59$, $p = 0.30$), which makes sense, since baseline accuracy is above 95% and there is a low amount of variance in both overlap or gain.

Risk-coverage curves for all models were nearly identical despite different team performance, confirming that selective prediction diagnostics are blind to complementarity. For instance, on Heart Disease, k-NN and Logistic L1 have nearly identical curves across most coverage levels, but their team performance differs by over 2 points under learned deferral.

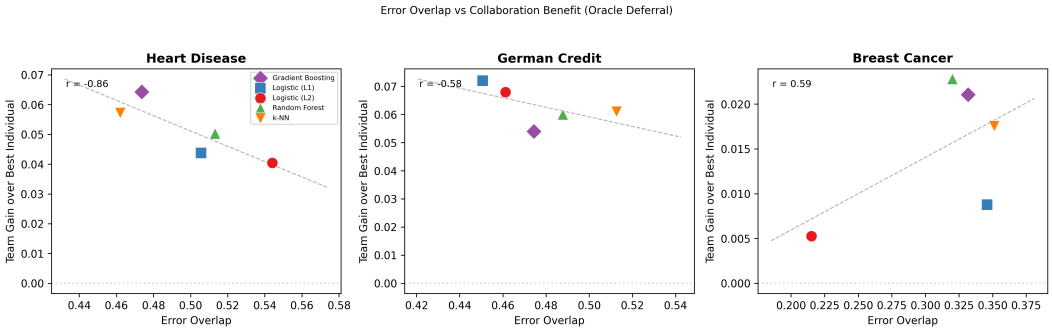

Figure 1: Error overlap vs. team gain over the better individual (oracle deferral). Each point is one model-human pair. Lower overlap = better team. Pearson r annotated per panel.

3.3 Statistical Caveats

Unfortunately, the individual rank flips are not statistically significant. Gradient Boosting beats Logistic L1 in 5 of 10 folds with 3 ties and 2 reversals. Wilcoxon gives $p = 0.078$ and Nadeau-

Bengio corrected t-test gives $p = 0.28$, reflecting the reality that $\sim$30-sample test folds are noisy. On German Credit the gap between best-by-accuracy and best-by-team is 1.8 points ($p = 0.24$, Nadeau-Bengio corrected). We do not correct for multiple comparisons across datasets as the correlation tests are exploratory rather than confirmatory.

However, the direction is reliable, as lower overlap is associated with better teams on both Heart Disease and German Credit. Confirming specific rank flips would need bigger datasets or a repeated-trial design.

3.4 Deferral Rates Across Groups

The German Credit dataset includes a `foreign_worker` attribute (96% foreign, 4% non-foreign). The model is less accurate on foreign workers (77.4% for Gradient Boosting) than non-foreign workers (92.5%), while the human expert shows the same pattern (70.5% vs. 92.5%). Under confidence thresholding, foreign workers are deferred at higher rates than non-foreign workers (e.g. 12.0% vs. 0.0% for Logistic L2), because the model is less confident on the majority group. Under learned deferral the pattern reverses: non-foreign workers are deferred more (e.g. 15.8% vs. 5.1% for Gradient Boosting), because the human expert is more accurate on that subgroup and learned deferral correctly identifies this. The deferral strategy thus changes not just team accuracy but *who experiences human review*, a concern raised by Madras et al. [4].

4 Discussion

4.1 Conditional Risk vs. Marginal Accuracy

Optimal deferral depends on the conditional risk of each agent given the specific input, not on marginal accuracy averaged over the whole dataset. Two models with the same average error rate can have very different error *locations* relative to the human. Gradient Boosting's sequential error-correction creates a distinctive failure pattern, as it tends to fail on globally ambiguous cases near the decision boundary while succeeding on locally unusual ones. The limited-feature human fails on cases needing feature interactions but handles simple threshold cases fine. These complementary failures are invisible to standalone accuracy, but exactly what the complementarity gap [10] measures.

4.2 What the Deferral Strategy Changes

The gap between these two strategies is most interesting for models with moderate overlap. Confidence thresholding treats all uncertain cases the same regardless of the human's ability. Learned deferral can identify specific regions where the human excels, which is why it helps low-overlap models like Gradient Boosting more (84.9% vs. 83.1% for thresholding).

4.3 Model Selection

Before choosing a model, compute error overlap and Δ_{comp} on a calibration set with expert labels. This requires the same data needed to train a deferral policy anyway, so there is no significant extra cost. In capacity-constrained settings where human review slots are limited, overlap-based selection becomes more important, since the system should spend scarce human attention on cases where the human can actually help which requires knowing where errors diverge.

4.4 Limitations

The largest limitation is that the experts are simulated. This work claims similarity to structural properties of fewer features, different errors but not to any real clinician. A user study would be much stronger. The datasets are also small, which is why pairwise differences cannot be resolved. Additionally, the learned deferral always uses logistic regression as the extended classifier, a joint training approach [10] that redirects model capacity toward the complementary region could amplify the effect. Finally, the human is treated as fixed, but in practice routing changes future human capability through skill atrophy and over-reliance [1, 8]. A model picked for complementarity today might lose its edge as the human's errors shift in response to the routing policy.

References

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[2] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[4] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[5] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[6] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003.

[7] Negin Okati, Abhishek De, and Manuel Rodriguez. Differentiable learning under triage. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[8] Manish Raghavan. *The Societal Impacts of Algorithmic Decision-Making*. PhD thesis, Cornell University, 2021.

[9] Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[10] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.