Generalizing Bayesian Human-AI Collaboration: Theory and Application in Data-Scarce Environments

Peng Liu[1], **Hailong Sun**[2*], **Chung-Piaw Teo**[3] and **Mabel Chou**[3]
[1]Singapore Management University
[2]Antai College of Economics and Management, Shanghai Jiao Tong University
[3]National University of Singapore
liupeng@smu.edu.sg, hailong.sun@sjtu.edu.cn, {bizteocp, mabelchou}@nus.edu.sg

Abstract

Combining predictions from heterogeneous classifiers—such as in-house deep learning models, human experts, and large language models (LLMs)—is a key challenge, especially in data-scarce environments such as humanitarian operations. We propose a flexible Bayesian framework to effectively fuse these diverse inputs. By integrating classifier logits with auxiliary human feedback (e.g., confidence, image clarity) using an ordered probit process, our model generalizes prior work to accommodate the real-world properties of these classifiers. We validate our framework on a challenging product recognition task in food bank operations, an environment defined by data scarcity and an inexperienced volunteer workforce. Our combined model significantly outperforms standalone ResNet, human, and LLM-based approaches, demonstrating the practical benefits of fusing these heterogeneous signal sources.

1 Introduction

Real-world AI systems increasingly rely on predictions formed from multiple heterogeneous sources rather than a single model. In many decision-making settings, predictions must be synthesized from components that differ fundamentally in nature, such as machine learning models trained on limited data, human judgments with variable expertise, and large pretrained foundation models. While combining such predictors is often empirically beneficial, the conditions under which collaboration improves accuracy—and the mechanisms by which it does so—remain poorly understood.

A central concept underlying successful collaboration is *complementarity*: performance gains arise when different predictors make distinct, weakly correlated errors. Prior theoretical work has shown that combining predictors can outperform any individual component when their decision processes are sufficiently diverse [Steyvers *et al.*, 2022]. However, existing analyses rely on assumptions that are difficult to justify in realistic deployments. In particular, most models assume equal uncertainty across predictors or focus solely on final predictions, ignoring auxiliary signals that accompany human judgments, such as confidence or perceived data quality.

This paper addresses the gap between idealized models of collaboration and the realities of heterogeneous decision-making. We study the fundamental problem of combining predictors whose latent decision variables exhibit unequal variance and correlation, and we characterize how these factors jointly determine whether collaboration is beneficial. By modeling predictor logits as correlated random variables with asymmetric uncertainty, we derive closed-form expressions for joint predictive accuracy. We also show how complementarity depends on an effective signal-to-noise ratio that captures both uncertainty imbalance and correlation structure.

Motivated by this theoretical perspective, we introduce a Bayesian combination framework that operationalizes collaborative prediction under heterogeneous uncertainty. The framework is designed to accommodate predictors that differ not only in accuracy but also in how their predictions are expressed. In addition to integrating model logits and categorical decisions, the framework incorporates auxiliary ordinal feedback—such as human confidence, perceived input clarity, and reflection—through an Ordered Probit process. These signals are treated as informative observations of an underlying latent belief state, thus exploiting information that is typically discarded in prediction-only aggregation schemes.

We demonstrate the practical value of this approach in a food bank setting, where donated items must be recognized and registered. For each donation, the system jointly predicts product type (70+ classes), halal status, and weight class by combining an in-house vision model, non-expert volunteer labels, and a large multimodal language model. Using volunteers rather than expert annotators reflects the NGO settings for which the framework is designed. Across different levels of data availability, the proposed framework consistently outperforms each standalone predictor. The results also match the theory: gains are largest when predictors have unequal uncertainty and weakly correlated errors, and auxiliary human feedback adds value beyond final labels alone.

By confronting unequal uncertainty and correlated errors directly, this work advances a more realistic understanding of collaborative prediction. The results clarify when combining heterogeneous predictors is beneficial, how auxiliary human signals can be formally integrated, and why collaboration succeeds in settings where individual components are unreliable.

*Corresponding author.

Together, these insights contribute toward principled design of human–AI systems that operate effectively under uncertainty. We position the contribution as a general framework, with the humanitarian-logistics study as a single realistic data-scarce instantiation, rather than a claim of universal validity.

The remainder of the paper is organized as follows. Section 2 summarizes prior work on human-AI collaboration, Bayesian aggregation, and learning from auxiliary supervision. Section 3 presents the Bayesian combination framework, followed up theoretical characterization of predictive complementarity under unequal uncertainty and correlated errors in Section 4. Section 5 reports empirical results in data-scarce settings and examines how the observed performance aligns with the theoretical analysis. Section 6 concludes.

2 Related Work

This work relates to research on human-AI collaboration, Bayesian aggregation of heterogeneous predictors, and learning from weak or auxiliary supervision. Rather than providing a broad survey, we focus on the modeling assumptions that limit existing approaches in realistic collaborative settings and motivate the need for frameworks that explicitly account for heterogeneous uncertainty and correlated errors.

Human-AI Collaboration and Complementarity. A substantial body of work studies how human judgment and algorithmic predictions can be combined to improve decision quality. Early research covers both benefits and challenges of such collaboration, including algorithm aversion, trust calibration, and delegation failures [Dietvorst *et al.*, 2015; Dietvorst *et al.*, 2018; Faraj *et al.*, 2018; Fügener *et al.*, 2022]. More recent work formalizes the notion of complementarity, showing that performance gains arise when human and machine errors are weakly correlated [Steyvers *et al.*, 2022].

Related strands investigate algorithmic strategies for allocating decisions between humans and models. Learning-to-defer and learning-to-complement frameworks optimize policies that decide when to rely on an expert versus automated system [Mozannar and Sontag, 2020; Wilder *et al.*, 2020]. Extensions consider selecting subsets of humans to maximize team performance [Singh and others, 2023]. While these approaches demonstrate that adaptive combination can improve outcomes, they primarily focus on policy optimization from data and do not provide analytic characterizations of when collaboration is fundamentally beneficial under heterogeneous uncertainty. In contrast, our work offers a theoretical analysis that explicitly characterizes how unequal uncertainty and error correlation determine the value of collaboration.

Bayesian Aggregation Models. Bayesian methods provide a natural framework for aggregating predictions under uncertainty. Classical Bayesian ensembles and hierarchical models typically assume homogeneous predictors or require calibrated probabilistic outputs. In human-AI collaboration, Bayesian latent variable models can represent shared structure between human and machine predictions [Steyvers *et al.*, 2022]. However, existing formulations generally rely on symmetric uncertainty assumptions and do not yield closed-form characterizations of joint predictive accuracy under heterogeneity.

Our approach builds on latent variable modeling by explicitly allowing predictors to differ in uncertainty while remaining correlated through shared latent decision variables. Related recent work studies collaborative uncertainty quantification in human-AI systems from a complementary perspective [Noorani *et al.*, 2025], but does not characterize predictive complementarity under unequal uncertainty and correlated errors. Other Bayesian ensemble methods emphasize robustness and dependence modeling [Surapunt and others, 2024], but do not address collaborative decision-making.

Learning from Weak and Auxiliary Supervision. Learning from weak or noisy supervision has been extensively studied in contexts such as crowd labeling, where multiple annotators provide imperfect labels that must be aggregated [Wang *et al.*, 2017; Yin *et al.*, 2021]. These methods typically focus on estimating annotator reliability from categorical labels and assume large annotator pools.

In collaborative decision-making settings, human judgments often include auxiliary signals beyond final label, such as confidence assessments or perceived input quality. While such signals can be informative, they are rarely integrated as first-class components of predictive models. The framework proposed here treats auxiliary ordinal feedback as informative observations of an underlying latent belief state shared with the human's categorical prediction, enabling principled inference even in small-sample, heterogeneous settings.

LLMs as Heterogeneous Predictors. Recent work explores the use of LLMs as proxies for human annotators or decision-makers [Ge *et al.*, 2021; Schanke *et al.*, 2021]. These studies primarily examine behavioral similarity or substitution effects. In this work, LLMs are treated as another instance of a heterogeneous, uncertain predictor rather than as a special case. This perspective allows domain-specific models and generalist foundation models to be combined within a single probabilistic framework, without relying on assumptions unique to LLMs.

Overall, existing literature provides important insights into collaboration, aggregation, and weak supervision, but leaves open how heterogeneous uncertainty and correlated errors jointly shape the benefits of collaboration. By addressing these factors directly, our work contributes a general theoretical and modeling framework for collaborative prediction under realistic conditions.

3 Bayesian Combination Model

We introduce a Bayesian combination model (cf. [Steyvers *et al.*, 2022]) designed to fuse predictions from heterogeneous classifiers and combine human and AI predictions to surpass either entity alone. This differs fundamentally from standard ensemble or Bayesian learning frameworks, which typically aggregate predictions from similar models (e.g., bagging or boosting). In traditional ensembles, predictions often come with probabilistic estimates and confidence intervals that reflect uncertainty. In contrast, our framework is designed to combine two inherently different types of predictors: a quantitative AI system and a human who often provides only a "best guess" without explicit, formal confidence metrics.

Our model leverages a generative process to account for both observed and hidden variables, using a Bayesian network

(Figure 1) to represent these variables as nodes within a probabilistic graphical model. This creates a framework where human intuition for ambiguous scenarios and AI-system consistency can complement each other. Complementarity arises when distinct characteristics of humans and AI combine to achieve higher predictive accuracy together.

3.1 AI Classifier: ResNet

Our primary AI component is an object recognition system built on a Convolutional Neural Network (CNN). We selected the Residual Network (ResNet) architecture [He *et al.*, 2016], which has demonstrated top-tier performance in object recognition tasks like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky *et al.*, 2015], while offering a good balance of predictive accuracy and model size. We fine-tune a pre-trained ResNet, harnessing the benefits of transfer learning to reduce training overhead.

Given the diversity of donations and operational requirements, we define a multilabel, multiclass framework that simultaneously predicts three labels for each item:

- Product Type: The most complex label, with over 70 classes (e.g., "Biscuits", "Sauce", "Rice"). This is crucial for basic inventory sorting.
- Halal Status: A critical label for serving diverse communities with specific religious dietary requirements, particularly in the Singapore context.
- Weight Class: Essential for food bank to accurately document the amount of food received and distributed for accountability to donors and stakeholders.

To handle this, we use a weighted loss function where the total loss consists of individual label losses, weighted according to the number of classes associated with each label so that more complex "product type" label is given priority. Furthermore, we adopt a conditional probability approach using a common backbone architecture to predict all three labels in a cascading structure. Specifically, the prediction of halal status is used as an input, together with the output from the backbone, to predict the weight class. Subsequently, the predictions for both are combined with the backbone to generate final prediction. For any class j, this model outputs logit scores λ_M which are converted to probabilities γ_M via a softmax function.

3.2 Bayesian Combination Framework

We propose a probabilistic graphical model in Figure 1 as an extension from [Steyvers *et al.*, 2022]. This framework is designed to fuse predictions from the human (y_H) and the AI (γ_M), where auxiliary human feedback are also incorporated.

Specifically, we add a set of new, observed nodes (the shaded nodes for "human confidence", "image clarity", and "surface reflection") to the human-generated side of the model. We formally model these ordinal ratings as draws from an OrderedProbit model. Crucially, these new nodes are parameterized by the same latent human probability simplex ($\gamma_{H,i,:}$) that generates the direct human prediction ($y_{H,i}$). This extension creates a richer, more comprehensive model of the human classifier. The framework no longer just learns from what the human predicted, but also from their qualitative assessment of the task's difficulty and data quality, providing a more nuanced signal for the subsequent MCMC inference.

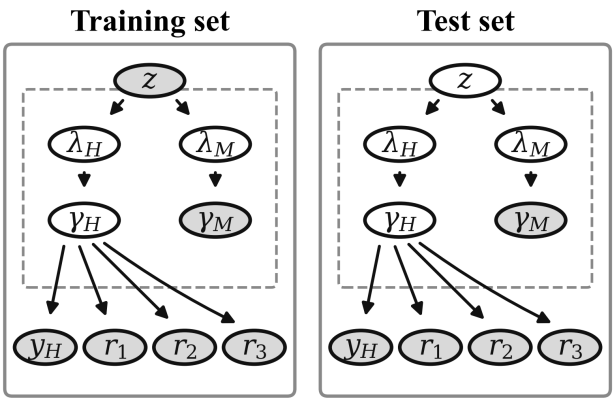

Figure 1: The Bayesian combination model for training and test regimes. Each outer plate is one data instance; the inner dashed plate ranges over classes $j = 1, \ldots, L$. Shaded nodes are observed, unshaded are latent. z is the true class label (observed in training, inferred at test). λ_H, λ_M are correlated human and machine logits and γ_H, γ_M corresponding class-probability simplices (AI simplex γ_M is observed). y_H is human's categorical label; r_1, r_2, r_3 are auxiliary ordinal ratings (confidence, image clarity, reflection).

Generative Process for Logits For each instance i and class $j \in \{1, \ldots, L\}$, we assume a class label $z_i \sim \text{Cat}(1:L, 1/L)$, where L is the total number of classes for a given label (e.g., $L = 70+$ for product type). The label z_i is observed for training instances and inferred for test instances.

Conditioned on z_i, the logit score $\lambda_{H,i,j}$ and $\lambda_{M,i,j}$ for each class j are simultaneously sampled from a bivariate normal distribution. We distinguish true-class and false-class means explicitly, such as using μ_H as logit-mean for human and μ_M for machine. We model the correlation between the human and AI classifiers as they both observe the same item:

$$\begin{pmatrix} \lambda_{H,i,j} \\ \lambda_{M,i,j} \end{pmatrix} \sim \begin{cases} \mathcal{N}\left(\begin{pmatrix} \mu_{H,j}^{\text{true}} \\ \mu_{M,j}^{\text{true}} \end{pmatrix}, \boldsymbol{\Sigma} \right) & \text{if } z_i = j \\ \mathcal{N}\left(\begin{pmatrix} \mu_{H,j}^{\text{false}} \\ \mu_{M,j}^{\text{false}} \end{pmatrix}, \boldsymbol{\Sigma} \right) & \text{if } z_i \neq j \end{cases}$$

The mean vectors distinguish the signal strength for the "true class" versus all "false classes". The covariance matrix $\boldsymbol{\Sigma}$ captures the individual classifier reliability and their correlation:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_H^2 & \rho_{HM}\sigma_H\sigma_M \\ \rho_{HM}\sigma_H\sigma_M & \sigma_M^2 \end{pmatrix}$$

Here, σ_H^2 and σ_M^2 represent the variance (unreliability) of the human and AI, respectively, and $\rho_{HM} \sim U(-1, 1)$ is their latent correlation.

Observed Variables and Softmax These latent logits are converted to class-probability simplexes $\gamma_{H,i,:}$ and $\gamma_{M,i,:}$ via a scaled softmax transformation with temperature parameter τ:

$$\gamma_{S,i,j} = \frac{\exp(\lambda_{S,i,j}/\tau_S)}{\sum_{k=1}^{L} \exp(\lambda_{S,i,k}/\tau_S)}, \quad S \in \{H, M\}.$$

The AI's full probability simplex $\gamma_{M,i,:}$ is directly observable from the ResNet model, with temperature $\tau_M = \tau$. However, it is impractical for a human to output a full probability distribution. We can only observe their single categorical prediction $y_{H,i}$. We therefore model this as a categorical draw from their latent probabilities, setting the human temperature $\tau_H = 1$:

$$y_{H,i} \sim \text{Cat}(1 : L, \gamma_{H,i,:})$$

Integrating Auxiliary Feedback via OrderedProbit In our experiment, human labeling involves six attributes: three direct predictions (product type, weight, halal) and three indirect descriptors (confidence, image clarity, and reflection), each categorized into three ordinal values ("Low", "Medium", "High"). We denote these ordinal ratings by $r_{H,i}^{\text{conf}}$, $r_{H,i}^{\text{clar}}$, and $r_{H,i}^{\text{refl}}$. We model these observed ordinal ratings as random draws from an OrderedProbit model[1]. The rating-prior parameters in Figure 1 are the cutoffs c_1, c_2 and sharpness δ. This model is parameterized by the same latent human probability simplex $\gamma_{H,i,:}$ that generated their direct prediction $y_{H,i}$. For example, the confidence rating is modeled as:

$$r_{H,i}^{\text{conf}} \sim \text{OrderedProbit}(\gamma_{H,i,:}, \mathbf{c}^{\text{conf}}, \delta^{\text{conf}})$$

This allows the model to learn the relationship between a human's latent (unobserved) probability and their stated confidence or assessment of image quality. This provides a much richer signal for inference, enabling a more refined and effective learning process within the Bayesian graphical model.

Bayesian Inference Let $\mathcal{D}_i$ collect the observed AI simplex, human label, and three ordinal ratings for instance i. The posterior $P(z_i \mid \mathcal{D}_i)$ is analytically intractable, so we estimate it using a Markov Chain Monte Carlo (MCMC) procedure, specifically Gibbs sampling [Casella and George, 1992]. The final prediction from the Bayesian combination model is the mode of the sampled marginal posterior distribution for z_i. Note that MCMC inference is well suited to the offline, batch-style analysis considered here; real-time or larger-scale deployment would call for approximate-inference variants such as variational Bayes or amortised inference, which we leave for future work.

4 Theoretical Analysis

In this section, we derive closed-form expressions for the joint accuracy and investigate its theoretical properties. Note that we allow unequal variance in the prior distribution of human and machine predictions, while earlier work, such as [Steyvers *et al.*, 2022], assumes equal uncertainty. In real-world scenarios, particularly in settings like non-profits, decision-making involves a mix of newly onboarded volunteers (high variance) and vision-based AI systems trained on heterogeneous datasets (variable variance). By allowing for unequal variance, our framework better captures these complexities and offers both theoretical and operational insights for designing more robust, human-centered AI collaboration strategies.

[1]The ordered probit model for a rating r_i (e.g., confidence) with three levels ("Low", "Medium", "High") is constructed as follows: r_i is "Low" with probability $\Phi(\delta(c_1 - \gamma_i))$, "Medium" with probability $\Phi(\delta(c_2 - \gamma_i)) - \Phi(\delta(c_1 - \gamma_i))$, and "High" with probability $1 - \Phi(\delta(c_2 - \gamma_i))$, where c_1, c_2 are cutoff parameters and δ is a sharpness parameter.

4.1 Single Classifier Accuracy

We start by assuming two independent and normally distributed random variables, $\lambda_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\lambda_2 \sim \mathcal{N}(0, \sigma_2^2)$, representing the predictive logit scores for the correct and incorrect classes, respectively. The classifier is correct if $\lambda_1 > \lambda_2$. The predictive accuracy A_C is:

$$A_C = P(\lambda_1 > \lambda_2) = \Phi\left(\frac{\mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \tag{1}$$

This generalizes the equal-variance case where $\sigma_1 = \sigma_2 = \sigma$, which simplifies to $A_C = \Phi(\frac{\mu_1}{\sqrt{2}\sigma})$.

This accuracy can also be understood through the lens of stochastic dominance.

Definition 1. *A random variable A has first-order stochastic dominance over another random variable B if, for any outcome x, the cumulative distribution function (CDF) of A is always below that of B, i.e., $F_A(x) \leq F_B(x)$ for all $x \in \mathbb{R}$, with strict inequality for at least one x.*

Definition 2. *Local stochastic dominance occurs when a random variable A stochastically dominates another random variable B at specific thresholds or within certain regions rather than across the entire range of possible outcomes.*

Given these definitions, Equation 1 implies that λ_1 has a local stochastic dominance over λ_2 after the crossover point of their respective CDFs, $x = \frac{\sigma_2 \mu_1}{\sigma_2 - \sigma_1}$ (assuming $\sigma_1 \neq \sigma_2$). We can also provide an alternative expression for A_C using the Wilcoxon–Mann–Whitney test.

Proposition 1. *The expression of the predictive accuracy A_C given in Equation (1) can be equivalently expressed as:*

$$P(X > Y) = \int_{-\infty}^{\infty} \left(1 - \Phi\left(\frac{y - \mu_1}{\sigma_1}\right)\right) \phi\left(\frac{y}{\sigma_2}\right) \cdot \frac{1}{\sigma_2}\, dy$$

where ϕ is the probability density function (PDF) of the standard normal distribution, with $X = \lambda_1$ and $Y = \lambda_2$.

4.2 Two-Classifier Joint Accuracy

We now turn to the case with two classifiers (C_1 and C_2) as described in Figure 1. We jointly model them using $(\lambda_{1,1}, \lambda_{2,1})$ for the correct class and $(\lambda_{1,i}, \lambda_{2,i}), i = \{2, \ldots, k\}$ for the incorrect class i, with a total of k classes. Each pair of predictive logits shares the same covariance structure with a correlation coefficient ρ as follows:

$$\begin{pmatrix}\lambda_{1,1}\\ \lambda_{2,1}\end{pmatrix} \sim N\Big[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \mathbf{\Sigma}\Big], \quad \begin{pmatrix}\lambda_{1,2}\\ \lambda_{2,2}\end{pmatrix} \sim N\Big[\begin{pmatrix}0\\ 0\end{pmatrix}, \mathbf{\Sigma}\Big] \tag{2}$$

where $\Sigma = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$.

As with [Steyvers *et al.*, 2022], we can express the joint accuracy for the binary case as the probability of observing the correct draw:

$$\begin{aligned} A_{C_1,C_2} = P\big\{&\phi(\boldsymbol{\ell}_1 \mid \boldsymbol{\mu}, \Sigma)\phi(\boldsymbol{\ell}_2 \mid \mathbf{0}, \Sigma) \\ &> \phi(\boldsymbol{\ell}_1 \mid \mathbf{0}, \Sigma)\phi(\boldsymbol{\ell}_2 \mid \boldsymbol{\mu}, \Sigma)\big\} \end{aligned}$$

where $\boldsymbol{\ell}_1 = (\lambda_{1,1}, \lambda_{2,1})^T$, $\boldsymbol{\ell}_2 = (\lambda_{1,2}, \lambda_{2,2})^T$, and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$.

This leads to our first result on joint accuracy under the general unequal variance setting.

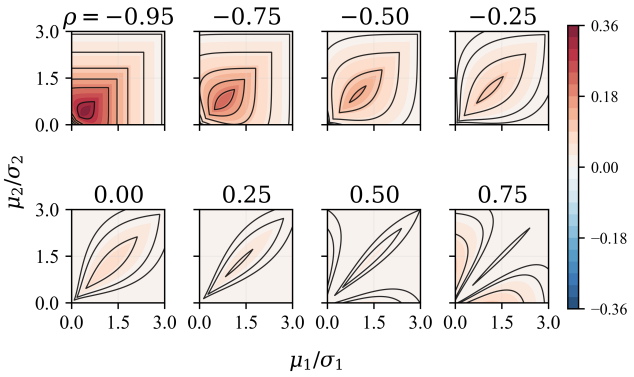

Figure 2: Difference in performance, $A_{C_1,C_2} - \max(A_{C_1}, A_{C_2})$, of the combined classifier over the better of the two individual classifiers. Complementarity is strongest when the correlation ρ is negative.

Theorem 1 (Joint Accuracy: Binary Case)**.** *Assume a combined set of classifiers C_1 and C_2 whose predictive logits are jointly modeled as bivariate normal random variables using Equation 2 for correct and incorrect classes, respectively. The joint predictive accuracy A_{C_1,C_2}, which allows for unequal variances $\sigma_1^2 \neq \sigma_2^2$, is:*

$$A_{C_1,C_2} = \Phi\left(\frac{\mu^*}{\sqrt{2}\sigma^*}\right)$$

where $\mu^* = \frac{\mu_1^2}{\sigma_1^2} + \frac{\mu_2^2}{\sigma_2^2} - \frac{2\rho\mu_1\mu_2}{\sigma_1\sigma_2}$ *and* $(\sigma^*)^2 = (\frac{\mu_1}{\sigma_1} - \frac{\rho\mu_2}{\sigma_2})^2 + (\frac{\mu_2}{\sigma_2} - \frac{\rho\mu_1}{\sigma_1})^2 + 2\rho(\frac{\mu_1}{\sigma_1} - \frac{\rho\mu_2}{\sigma_2})(\frac{\mu_2}{\sigma_2} - \frac{\rho\mu_1}{\sigma_1})$.

Relative to [Steyvers *et al.*, 2022], this result corrects and generalizes the derivation. Figure 2 plots the improvement of this Bayesian combination method over the accuracy of the best individual classifier. We observe the strongest improvement (complementarity) when the correlation ρ is highly negative (e.g., -0.95), and minimal improvement when ρ is highly positive, as the predictions become redundant.

We further generalize this to multi-class (k-class) setting.

Theorem 2 (Joint Accuracy: Multi-class Case)**.** *Under the same assumptions as Theorem 1, the joint predictive accuracy for a k-class task is:*

$$A_{C_1,C_2} = \int_{-\infty}^{\infty} \left[\Phi\left(y + r_{H,M}\right)\right]^{k-1} \phi(y) dy$$

where $r_{H,M} = \frac{\mu^*}{\sigma^*}$ is the effective "relative signal-to-noise ratio" of the combined classifier, with μ^* and σ^* as defined in Theorem 1.

4.3 Properties of Complementarity

The term $r_{H,M}$ from Theorem 2 determines the predictive accuracy of the joint model. Let H be the Human and M be the AI. Let $a_H = \mu_H/\sigma_H$ and $a_M = \mu_M/\sigma_M$ represent the individual predictive power, i.e., the signal-to-noise ratio (SNR), of each classifier. We can then express $r_{H,M}$ as:

$$r_{H,M} = \frac{a_H^2 + a_M^2 - 2\rho a_H a_M}{\sqrt{\begin{array}{c}(a_H - \rho a_M)^2 + (a_M - \rho a_H)^2 \\ + 2\rho(a_H - \rho a_M)(a_M - \rho a_H)\end{array}}}$$

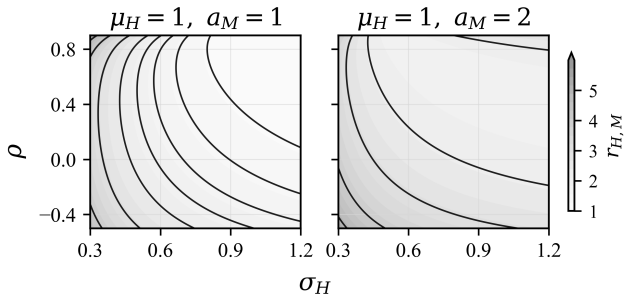

Figure 3: Contour plot of the joint signal-to-noise ratio $r_{H,M}$ as a function of human uncertainty (σ_H) and correlation (ρ). Joint performance $r_{H,M}$ generally increases as ρ or σ_H decreases.

This expression can also be interpreted as the signal-to-noise ratio of a linear combination of two standardized normal variates (X_H, X_M) with mean (a_H, a_M) and correlation ρ:

$$r_{H,M} = \frac{\mathbb{E}\left[(a_H - \rho a_M)X_H + (a_M - \rho a_H)X_M\right]}{\sqrt{\mathbf{Var}[(a_H - \rho a_M)X_H + (a_M - \rho a_H)X_M]}}$$

This formulation leads to several important insights (proofs are provided in the supplementary material).

Proposition 2. *The joint accuracy $A_{H,M}$ is monotonically increasing in the relative signal-to-noise ratio $r_{H,M}$.*

Proposition 3. *If both classifiers are equally strong (i.e., $a_H = a_M = a$), then $r_{H,M} = \frac{a\sqrt{2}}{\sqrt{1+\rho}}$, which is monotonically decreasing in ρ. This confirms that complementarity is greatest when classifiers are diverse (low ρ).*

Proposition 4. *If the human classifier provides no discriminatory signal (i.e., $a_H = 0$), the joint classifier's SNR becomes $r_{H,M} = \frac{a_M}{\sqrt{1-\rho^2}}$. Since $\sqrt{1-\rho^2} \leq 1$, $r_{H,M} \geq a_M$, and therefore the joint classifier still performs at least as well as the AI classifier alone ($A_{H,M} \geq A_M$).*

This surprising result shows that even a "noisy" human, if their noise is imperfectly correlated with the AI, can provide a signal that improves the combined system. Figure 3 visualizes the relationship between σ_H, ρ, and the resulting joint signal-to-noise ratio $r_{H,M}$, confirming that $r_{H,M}$ generally increases as correlation ρ or human uncertainty σ_H decreases.

5 Experiments and Results

We validate our framework using a case study at Food Bank Singapore (FBS), a real-world environment characterized by data scarcity and classifier unreliability. We conduct two main sets of experiments: (1) combining our in-house ResNet with human volunteer predictions, and (2) combining the ResNet with ChatGPT-4O as a human proxy. The full data and code are available at https://github.com/jackliu333/object_detection_using_tensorflow.

5.1 Experimental Setup

Classifier Characteristics We first summarize the qualitative characteristics of our three heterogeneous classifiers in Table 1. The in-house ResNet is data-hungry but accurate on

Type	Name	Data	Accuracy	Sens.
Solo	ResNet	High	High (In-Sample)	High
	Human	Low	High (Familiar)	Low
	ChatGPT	Low	High (General)	Med
Ens.	+Human	Med	Robust (All)	Low-Med
	+ChatGPT	Med	Robust (All)	Low-Med

Table 1: Summary of model characteristics.

in-sample products, while being sensitive to environmental changes. Humans require minimal data but are prone to errors on unfamiliar items. ChatGPT shows high accuracy comparable to humans but has moderate sensitivity. Our goal is to create an ensemble model that combines these strengths.

Data Collection For the Human+ResNet experiment, we collected 306 product images. We developed a prototype to randomly display these images to five volunteers, simulating the task of registering a new, unrecognized product. For each image, volunteers provided six attributes: three direct predictions (product type, weight class, halal status) and three auxiliary ordinal ratings (confidence, image clarity, reflection).

For the ChatGPT+ResNet experiment, we collected a separate set of 292 products, capturing 360-degree video footage for each. To simulate the human process of rotating an item, we extracted 10 equally-spaced frames from each video and sent them to the GPT-4o API along with a pre-defined prompt outlining the classification task and rules.

Training and Evaluation Design To test the model under varying levels of data scarcity, we trained the ResNet and Bayesian models using four different train-test splits for the 306-image dataset: 10%, 20%, 50%, and 80% test set ratios. A higher ratio (e.g., 80%) simulates a data-scarce environment where the AI has little training data.

We fine-tuned the ResNet models via transfer learning for 15,000 epochs using images collected under varying environmental conditions (lighting, angle, etc.) and repeated the procedure with ten random seeds to mitigate sampling bias, yielding 40 trained models whose logits serve as inputs to the Bayesian framework. For the Bayesian models, we trained a separate model for each participant (unlike the shared model in [Steyvers *et al.*, 2022]), yielding 200 total models (4 ratios $\times$ 5 volunteers $\times$ 10 seeds); each was trained using the "rjags" package for 5000 iterations with a 2000-iteration burn-in and three simultaneous MCMC chains, with AI predictions substituted for incomplete human submissions.

Complementarity (Human + ResNet) We first analyze the marginal improvement of our Bayesian model over the standalone Human and ResNet (Machine) models. We define the improvement metric I as the percentage difference in correct predictions, e.g.: $I_{\text{bayes\_over\_human}} = \frac{\text{correct\_Bayes} - \text{correct\_human}}{\text{correct\_human}}$. As shown in Figure 4, the Bayesian combination model generates a positive marginal improvement over both human and AI predictions by a sizable amount in most cases. The improvement over the Machine (ResNet) is highest with a large test set (0.8). When the AI is poorly trained (lacks sufficient data), human expertise is highly complementary and provides a significant boost. Moreover, improvement over Human is highest with a

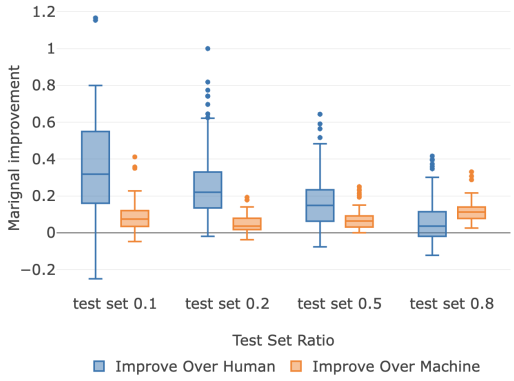

Figure 4: Box plots of marginal improvements of the Bayes model over Human and AI predictions for each test set ratio.

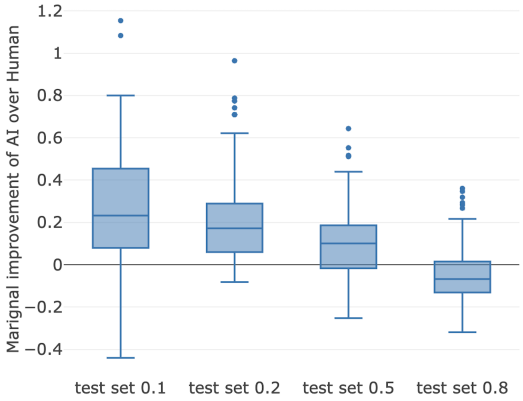

Figure 5: Box plots of marginal improvements of the AI model over Human for each test set ratio. AI dominates in data-rich settings, while Human dominates in data-scarce settings.

small test set (0.1). When the AI is well-trained, it provides reliable predictions that effectively corrects human errors and boosts the performance of the human.

To understand why this fusion is so effective, we first compare the standalone AI and Human performance in Figure 5. ResNet (AI) generally performs better than humans when the test set ratio is 0.5 or less (i.e., when the AI has 50% or more of the data for training). However, in the most data-scarce setting (0.8 test set, 20% training), the human predictions are more accurate. This performance crossover demonstrates that neither classifier is dominant in all settings, motivating the need for our Bayesian fusion model.

We next analyze the latent correlation ρ (Figure 6), which our theory identifies as a key driver of complementarity. We found ρ is consistently low (0 to 0.2). This low correlation reveals that humans and the ResNet use different decision-making mechanisms, empirically corroborating the theory: the Bayesian model delivers larger complementarity when predictions are diverse and less correlated [Steyvers *et al.*, 2022]. Interestingly, we observe an increasing trend in ρ as the test set ratio increases (i.e., in data-scarce settings). This suggests that as the AI model becomes weaker due to

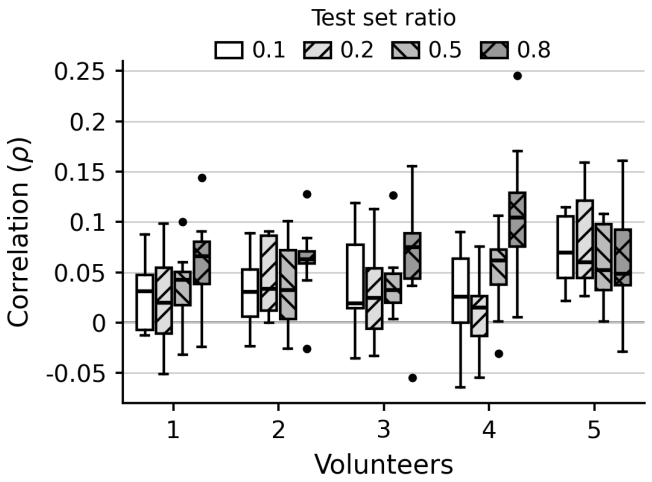

Figure 6: Correlation factor ρ across different test set ratios for each volunteer. The low correlation demonstrates high complementarity.

insufficient training, its predictions begin to behave "more like" the human's, and the two become more correlated.

Value of Auxiliary Features We test the value of our model's novel component: the integration of auxiliary features. We trained one set of models with this data (clarity, reflection, confidence) and another 200 models without it (using only direct predictions, as in [Steyvers *et al.*, 2022]). As shown in Figure 7, the model with auxiliary information (green) consistently achieves a higher number of correct predictions. This confirms that qualitative human assessments can provide a valuable, independent signal.

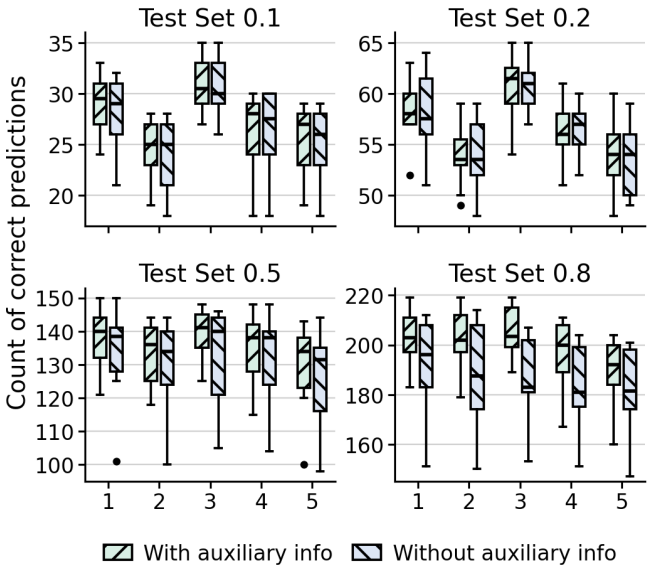

Figure 7: Comparing Bayesian models with and without additional features (clarity and reflection) across different test sets. The model with auxiliary info (green) consistently outperforms.

TASK	RESNET	CHATGPT	**BAYESIAN**
PRDTYPE	0.8134	0.6585	**0.8310**
WEIGHT	0.7432	0.7466	**0.9212**
HALAL	0.8219	0.7637	**0.9247**

Table 2: Predictive accuracy of all models across different prediction tasks. The Bayesian combination model generates a higher predictive accuracy than ChatGPT or ResNet alone.

COMPARISON	PRDTYPE	WEIGHT	HALAL
VERSUS RESNET	0.10 (0.09)	0.33 (0.09)	0.15 (0.07)
VERSUS CHATGPT	0.26 (0.00)	0.21 (0.02)	0.18 (0.04)

Table 3: Average percentage improvement (standard deviation in brackets) of the Bayesian model over baselines.

Combining ChatGPT and ResNet In our final experiment, we test the model's flexibility by replacing the human with LLM. We used 292 product videos, allocating half for training and half for testing. We acknowledge these products might already be in ChatGPT training data; however, our in-house ResNet is trained only on our own data, making it a small-scale, specialized counterpart to the generalist LLM.

Table 2 shows that the Bayesian combination model (Bayesian) achieves the highest accuracy across all three prediction tasks. It significantly outperforms both standalone models. This result shows that our framework is not limited to human-AI collaboration but can also serve as a robust method for combining different classes of AI models (a specialized vision model and a generalist multi-modal LLM).

To further quantify these gains, we repeated the experiment five times and report average percentage improvement in Table 3. We note that marginal improvement for ResNet in product type is lower (9.5%) than in the other two tasks, possibly due to high weight assigned in the ResNet's loss function. Conversely, improvements over ChatGPT are more consistent.

6 Conclusion

We proposed a generalized Bayesian combination model for human-AI collaboration. Our primary contributions are twofold. Methodologically, we introduced a novel framework that enriches the combination model by integrating auxiliary human feedback (confidence, clarity) using an OrderedProbit process. Empirically, we demonstrated the high practical value of this framework by successfully fusing three heterogeneous classifier types (in-house ResNet, Human, and ChatGPT) in a challenging, real-world humanitarian setting. We also provided theoretical grounding for this model by generalizing the closed-form expression for the joint predictive accuracy of the Bayesian combination model. Our results show that this approach is a robust and flexible way to achieve complementarity in complex, data-scarce environments. Future work could explore active learning policies based on this framework, to dynamically decide "when to ask" a human versus "when to trust" the AI based on model-inferred uncertainty.

Acknowledgments

Peng Liu acknowledges funding from Singapore MOE Tier 1 grant MSS25B001. Hailong Sun is supported by the National Natural Science Foundation of China [Grants 72301168 and 72472098] and the Shanghai Pujiang Programme [Grant 23PJC062]. Mabel Chou is supported by the National Natural Science Foundation of China Grant No. 72374036.

References

[Casella and George, 1992] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[Dietvorst *et al.*, 2015] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015.

[Dietvorst *et al.*, 2018] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018.

[Faraj *et al.*, 2018] Samer Faraj, Stelios Pachidi, and Karim Sayegh. Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1):62–70, 2018.

[Fügener *et al.*, 2022] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696, 2022.

[Ge *et al.*, 2021] Ruyi Ge, Zhiqiang Zheng, Xuan Tian, and Li Liao. Human–robot interaction: when investors adjust the usage of robo-advisors in peer-to-peer lending. *Information Systems Research*, 32(3):774–785, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Mozannar and Sontag, 2020] H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. In *Proc. International Conference on Machine Learning (ICML)*, 2020.

[Noorani *et al.*, 2025] Sima Noorani, Shayan Kiyani, George Pappas, and Hamed Hassani. Human–ai collaborative uncertainty quantification. *arXiv preprint*, 2025. https://arxiv.org/abs/2510.23476.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[Schanke *et al.*, 2021] Scott Schanke, Gordon Burtch, and Gautam Ray. Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, 32(3):736–751, 2021.

[Singh and others, 2023] S. Singh et al. On subset selection of multiple humans to improve accuracy with ai models. In *Proc. of AAMAS*, 2023.

[Steyvers *et al.*, 2022] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022.

[Surapunt and others, 2024] T. Surapunt et al. Ensemble modeling with a bayesian maximal information coefficient. *Information*, 15(4):228, 2024.

[Wang *et al.*, 2017] Jing Wang, Panagiotis G Ipeirotis, and Foster Provost. Cost-effective quality assurance in crowd labeling. *Information Systems Research*, 28(1):137–158, 2017.

[Wilder *et al.*, 2020] B. Wilder, E. Horvitz, and E. Kamar. Learning to complement humans. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[Yin *et al.*, 2021] Junming Yin, Jerry Luo, and Susan A Brown. Learning from crowdsourced multi-labeling: A variational bayesian approach. *Information Systems Research*, 32(3):752–773, 2021.