Coverage-Constrained Human-AI Cooperation with Multiple Experts

Zheng Zhang[1], Cuong C. Nguyen[1], Kevin Wells[1], Thanh-Toan Do[2], David Rosewarne[1], Gustavo Carneiro[1]*

[1]Univerisity of Surrey
[2]Monash University
zheng.zhang@surrey.ac.uk

Abstract

Human-AI cooperative classification (HAI-CC) aims to develop hybrid intelligent systems that enhance decision-making in various high-stakes real-world scenarios by leveraging both human expertise and AI capabilities. Current HAI-CC methods primarily focus on learning-to-defer (L2D), where decisions are deferred to human experts when AI is not confident, and learning-to-complement (L2C), where AI and human experts make predictions cooperatively. However, existing research in both L2D and L2C has not effectively been explored under diverse expert knowledge to improve decision-making, particularly when constrained by the operation cost of human involvement. In this paper, we address this research gap by proposing the **C**overage-constrained **L**earning to **D**efer and **C**omplement with Specific Experts (CL2DC) method. In particular, CL2DC assesses input data before making final decisions through either AI prediction alone or by deferring to or complementing a specific human expert. Furthermore, we propose a coverage-constrained optimisation to control the cooperation cost, ensuring it approximates a target probability for AI-only selection. This approach enables an effective assessment of system performance within a specified budget. Comprehensive evaluations on both synthetic and real-world datasets demonstrate that CL2DC achieves superior performance compared to state-of-the-art HAI-CC methods.

Code — https://github.com/zhengzhang37/CL2DC.git

Introduction

Machine learning models are becoming increasingly critical in real-world scenarios due to their high efficiency and accuracy. However, in high-stakes situations like risk assessment (Green and Chen 2019), breast cancer classification (Halling-Brown et al. 2020), and the detection of inaccurate or deceptive content produced by large language models (Ding et al. 2024), human experts often provide more reliable and safer predictions compared to AI models. To address the trade-off between human expertise and AI capabilities, *human-AI cooperative classification* (HAI-CC) methods have been developed (Dafoe et al. 2021). These approaches improve not only the accuracy, interpretability, and usability of AI models but also human efficiency and decision consistency over manual processes, significantly reducing human error (Dafoe et al. 2021).

HAI-CC approaches (Dafoe et al. 2021) aim to develop a *hybrid intelligence* system that maximises accuracy while minimising the cooperation costs with *learning-to-defer* (L2D) and *learning-to-complement* (L2C) strategies. Specifically, L2D (Mozannar and Sontag 2020) strategically decides when to classify with the AI model alone or defer to human, while L2C (Wilder, Horvitz, and Kamar 2021) aggregates the predictions of AI and human into a final decision. When facing with challenging or high-stake decisions, a single-expert HAI-CC (SEHAI-CC) system is able to defer to or complement with a fixed human expert (Mozannar et al. 2023). However, given the diverse range of expertise of different professionals, relying solely on a single expert for decisions across all input cases is impractical and potentially suboptimal. To address this, multiple-expert HAI-CC (MEHAI-CC) methods have been proposed to explore strategies for either complementing or deferring decisions to one or several experts simultaneously (Verma, Barrejon, and Nalisnick 2023; Zhang et al. 2025a), effectively leveraging diverse expert knowledge for more robust decision-making. Nevertheless, a remarkable gap in such MEHAI-CC approaches is that they rarely address L2D and L2C concomitantly, and even when they do consider both tasks in a single approach (Zhang et al. 2025a), they effectively disregard diverse expert knowledge by randomly selecting experts for the cooperative classification.

Another crucial issue in HAI-CC is the trade-off between accuracy and cooperation cost as it reflects the system's efficiency and effectiveness. Existing HAI-CC methods often analyse such trade-off through accuracy-coverage curves to evaluate performance at different coverage levels (Narasimhan et al. 2022; Mozannar et al. 2023; Zhang et al. 2025a). Coverage is defined as the *percentage of examples classified by the AI model alone*, in which 100% coverage indicates that all classifications are performed by the AI, and 0% coverage means that all classifications are handled exclusively by experts. However, existing HAI-CC methods (Narasimhan et al. 2022; Mozannar et al. 2023) do not consider cooperation cost into their optimisation functions. Even if some studies do consider the cost to balance prediction accuracy and human involvement cost (Zhang et al.

*Corresponding authors

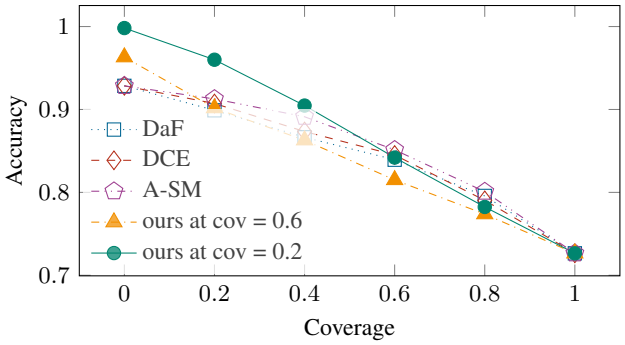


Figure 1: Applying the post-hoc analysis to construct the coverage - accuracy curves is unreliable because the same method trained on Chaoyang dataset (Zhu et al. 2021) with different coverage constraints produces different curves. When comparing to several HAI-CC methods (Charusaie, Fesharaki, and Samadi 2024; Wei, Cao, and Feng 2024; Cao et al. 2024) plotted with the same post-hoc approach, it is possible to select the curve showing the best coverage - accuracy result, which may present an overly optimistic assessment of the method's performance. For instance, our method trained for two different coverage constraints (i.e., 0.6 in orange and 0.2 in green) show quite different performances.

2025a), the training process is brittle, meaning that small adjustments to the hyper-parameter controlling accuracy and cooperation cost often result in coverage values collapsing to either 0% or 100%. Importantly, this hyper-parameter does not set a specific coverage value but rather allows for only a rough adjustment of cost influence within the optimisation, making it challenging to achieve a precise coverage target. As a result, existing methods (Mozannar et al. 2023; Narasimhan et al. 2022) employ a *post-hoc* technique to construct accuracy - coverage curves by sorting deferral scores and adjusting the deferral threshold to obtain the accuracy at the expected coverage.

The post-hoc evaluation involves three main steps: (i) running inference on all test samples and storing the probabilities output by the rejector (a.k.a. *gating model*), (ii) sorting the test samples based on their rejection probabilities, and (iii) deferring samples that have highest rejection scores until meeting the targeted coverage value (see Algorithm 3 in the supplementary material for further details). Since this approach needs to pre-compute and sort the predictions on all test samples, it is unsuitable for real-world applications where each test sample must be decided to be analysed by AI or human in a sample-by-sample fashion, not to wait until gathering all possible testing samples. Additionally, using this method to analyse the coverage-accuracy trade-off of a model trained under "implicitly-specified" coverage setting is unreliable, as models trained with different coverage constraints yield different post-hoc curves, as shown in Fig. 1 (e.g., the orange and green curves). This variability makes it unreliable to assess the approach simply by selecting the best-performing method.

We, therefore, propose a novel HAI-CC method, called **C**overage-constrained **L**earning to **D**efer and **C**omplement with Specific Experts (CL2DC), to address both of the issues above. CL2DC integrates the strengths of L2D and L2C, particularly in training scenarios with multiple noisy-label annotations, enabling the system to either make final decisions autonomously or cooperatively with a specific expert. In particular, CL2DC not only determines when to defer to or complement with experts but also assesses the specific expertise of each expert, selecting the most suitable one for the decision-making process. We also introduce an innovative coverage constraint penalty in the loss function to effectively control how workload is distributed to AI or human experts. This penalty enables a reliable way to achieve the target coverage, while also allowing a consistent and meaningful analysis of efficiency and effectiveness between various methods using coverage-accuracy curves. Our main contributions are summarised as follows:

- we propose the CL2DC, which integrates both L2D and L2C strategies, enabling deferral to or complementation with specific experts in the presence of multiple noisy-label annotations; and
- we also introduce an innovative coverage constraint into the training process of L2D, targeting specific coverage values to effectively manage the trade-off between coverage and accuracy in HAI-CC.

We evaluate CL2DC against state-of-the-art (SOTA) HAI-CC methods (Verma, Barrejon, and Nalisnick 2023; Wei, Cao, and Feng 2024; Charusaie, Fesharaki, and Samadi 2024; Cao et al. 2024; Zhang et al. 2025a) on both synthetic (e.g., CIFAR-100 (Wei et al. 2021)) and real-world (e.g., Galaxy Zoo (Bamford et al. 2009), HAM10000 (Tschandl, Rosendahl, and Kittler 2018), NIH-ChestXray (Majkowska et al. 2020), MiceBone (Schmarje et al. 2022), and Chaoyang (Zhu et al. 2021)) datasets. Empirical results show that CL2DC consistently outperforms previous HAI-CC methods with higher accuracy for equivalent coverage values across all the evaluation benchmarks.

Related Work

Human-AI Cooperative Classification

HAI-CC approaches (Dafoe et al. 2021) seek to develop a *hybrid intelligent* system to maximise the cooperative accuracy beyond what either AI models or human experts can achieve independently, while simultaneously minimising the cooperation costs through *learning-to-defer* (L2D) and *learning-to-complement* (L2C) strategies.

Learning to Defer (L2D) aims to learn a classifier and a rejector to decide in which case the decision should be deferred to a human expert to make the final prediction (Madras, Pitassi, and Zemel 2018). Existing L2D approaches focus on the development of different surrogate loss functions to be consistent with the Bayes-optimal classifier (Narasimhan et al. 2022; Cao et al. 2024). Wei, Cao, and Feng (2024) explore the dependence between classifier and human, and propose a dependent Bayes optimality formulation. These methods, however, overlook practical settings, especially the ones with a wide diversity of multiple human

experts. Given such an issue, recent research in L2D shifts towards the multiple-expert setting (Mao et al. 2023; Verma, Barrejon, and Nalisnick 2023). For example, Verma, Barrejon, and Nalisnick (2023) proposed a L2D method to defer the decision to one of multiple experts. Mao, Mohri, and Zhong (2024) addressed both instance-dependent and label-dependent costs and proposed a novel regression surrogate loss function. Despite remarkable progress, current research in L2D lacks the ability to aggregate the predictions of human experts and classifier to make a joint decision.

Learning to Complement (L2C) methods aim to optimise the cooperation between human experts and classifier by combining their predictions (Wilder, Horvitz, and Kamar 2021; Hemmer et al. 2022; Zhang et al. 2025b,a). Charusaie, Fesharaki, and Samadi (2024) introduce a method to determine whether the classifier or a human expert should predict independently, or if they should collaborate on a joint prediction – this is effectively a combined L2D and L2C approach, but it is limited to single expert setting. Hemmer et al. (2022) introduce a model featuring an ensemble prediction involving both AI and human predictions per sample, yet it does not optimise the cooperation cost. Therefore, the system will always select human experts to engage in the decision process, preventing any evaluation of how the algorithm performs at different coverage values and making it impossible to study algorithmic behaviour under partial automation. Zhang et al. (2025a) propose to combine L2D-L2C, integrating AI predictions with multiple human experts, but overlooking expert specificity.Moreover, LECODU modulates coverage indirectly by adjusting the hiring cost, which only implicitly influences the deferral behaviour.

Learning with Noisy Labels

The vast majority of HAI-CC methods assume that the ground-truth *clean* annotations are available in the training set. However, such an assumption is not warranted in practice, particularly in applications like medical imaging, where a definitive *clean* label may not be available due to the absence of final pathology. Therefore, we can only access expert opinions, which means multiple noisy annotations per training sample. Only recently, a few HAI-CC systems have been designed to handle noisy-label problems.

Among the top-performing learning with noisy labels (LNL) methods (Carneiro 2024), ProMix (Wang et al. 2023) introduces an optimisation based on a matched high-confidence selection technique. DEFT (Wei et al. 2024) utilises the alignment of textual and visual features, pre-trained on auxiliary image-text pairs to sieve out noisy labels. Despite achieving remarkable results, LNL with a single noisy label suffers from the identifiability issue (Liu, Cheng, and Zhang 2023), meaning that a robust training may require multiple noisy labels per samples. Methods that can deal with multiple noisy labels per sample are generally known as multi-rater learning (MRL) approaches.

MRL aims to train a robust classifier with multiple noisy labels from multiple human annotators. Recently, UnionNet (Wei et al. 2022) has been developed to integrate all labelling information from multiple annotators as a union and maximise the likelihood of this union through a parametric transition matrix. CROWDLAB (Goh, Tkachenko, and Mueller 2022) is a state-of-the-art (SOTA) MRL method that produces consensus labels using a combination of multiple noisy labels and the predictions by an external classifier.

Methodology

For a C-way classification task, let $\mathcal{D} = \{\mathbf{x}_i, \mathcal{M}_i\}_{i=1}^N$ be the noisy-label multi-rater training set of size N, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ denotes a d-dimensional input sample, and $\mathcal{M}_i = \{m_{i,j} : m_{i,j} \in \mathcal{Y} = \{1, \dots, C\}\}_{j=1}^M$ denotes the noisy annotations of M human experts for the input image $\mathbf{x}_i$. Let's also denote $\Delta^{K-1} = \{\mathbf{p} : \mathbf{p} \in [0,1]^K \wedge \mathbf{1}^\top \mathbf{p} = 1\}$ as the $(K-1)$ probability simplex, and $[M] = \{1, , \dots, M\}$.

The proposed system is inspired by the mixture of experts, which consists of three components as follows:

1. an *AI classifier*, denoted as $f_\theta : \mathcal{X} \to \Delta^{C-1}$,
2. a *gating model*, denoted as $g_\phi : \mathcal{X} \to \Delta^{2M+1}$ produces the probability to decide if the decision is made by the classifier alone (i.e., $g_\phi^{(\text{AI})}(.)$), or deferring the decision to one of the M human experts (i.e., $g_\phi^{(\text{L2D}_j)}(.)|_{j=1}^M$ denote the probability of selecting expert 1 through expert M), or performing a complementary classification between the AI classifier and one human expert (i.e., $g_\phi^{(\text{L2C}_j)}(.)|_{j=1}^M$ represent the probability of selecting AI + expert 1, through AI + expert M), and
3. a *complementary module* $h_\psi : \Delta^{C-1} \times \mathcal{Y} \times [M] \to \Delta^{C-1}$ aggregates the predictions made by the AI model and a selected human expert to produce a final prediction.

The three models form an adaptive decision system (see Fig. 2) that leverages the efficiency of machine learning models while maintaining human oversight, ensuring a balance between performance and trustworthiness in complex decision environments.

Consensus labels In standard HAI-CC, ground truth labels are required for training, while in our setting, they are unavailable. Following LECODU (Zhang et al. 2025a), which also assumes the unavailability of ground truth labels, we use the SOTA MRL method CROWDLAB (Goh, Tkachenko, and Mueller 2022) to produce the *consensus* labels to be used as the ground truth in our training. Specifically, CROWDLAB takes training samples and experts' annotations $(\mathbf{x}_i, \mathcal{M}_i) \in \mathcal{D}$, together with the classifier's predictions $f_\theta(\mathbf{x}_i)$ to produce a consensus label $\hat{y}_i \in \mathcal{Y}$ associated with a quality (or confidence) score α_i as follows:

$$\hat{y}_i, \alpha_i = \mathsf{CrowdLab}(\mathbf{x}_i, f_\theta(\mathbf{x}_i), \mathcal{M}_i) \cap \{\alpha_i > 0.5\}. \quad (1)$$

Objective function We propose a loss function that minimises the weighted-average loss across all available decision options (AI alone, deferral to specific expert, or human-AI complementary classification), with the weights produced by the gating model $g_\phi(.)$ representing the probability produced by the gating model:

$$\min_{\theta,\phi,\psi} {}^1\!/_N \sum_{i=1}^N g_\phi^\top(\mathbf{x}_i) \boldsymbol{\ell}(\mathbf{x}_i, \hat{y}_i, \mathcal{M}_i, \theta, \psi), \quad (2)$$

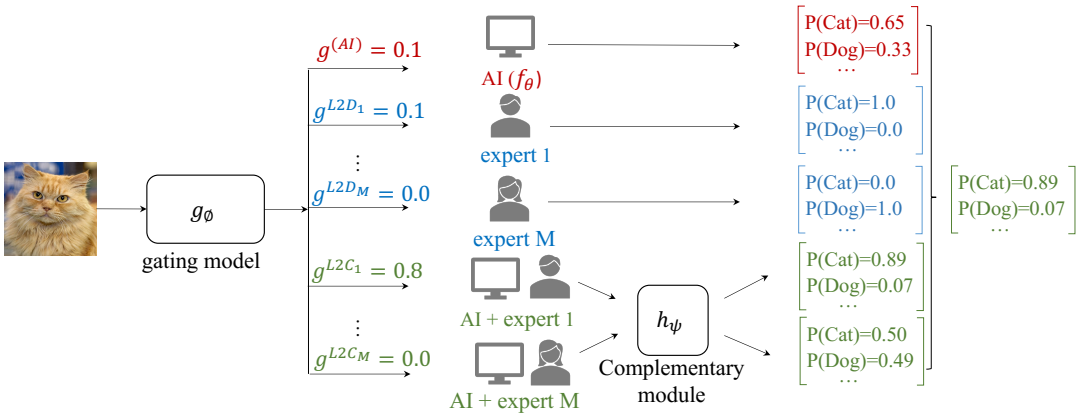


Figure 2: CL2DC contains a gating model $g_\phi(.)$, a complementary module $h_\psi(.)$, and an AI model $f_\theta(.)$. The gating model aims to decide whether we use LNL-trained AI model $f_\theta(.)$ alone (i.e., when $g_\phi^{(\text{AI})}(.)$ has the largest probability), defer the decision to one of the M experts (i.e., when one of the $g_\phi^{(\text{L2D}_j)}(.)|_{j=1}^M$ has the largest probability), or complement the LNL AI model's prediction, through the complementary module, with one of the M experts (i.e., when one of $g_\phi^{(\text{L2C}_j)}(.)|_{j=1}^M$ has the largest probability). In the figure, the gating model selects L2C between AI and human expert 1, given its largest probability of 0.8, to make the final prediction, on the right.

where $\boldsymbol{\ell}(.)$ is a $(2M+1)$-dimensional vector concatenating the cross-entropy loss induced by the classifier (i.e., $\ell_{\text{CE}}(\hat{y}_i, f_\theta(\mathbf{x}_i))$, the M human experts (i.e., $\ell_{\text{CE}}(\hat{y}_i, m_j), j \in [M]$ and the complementary module (i.e., $\ell_{\text{CE}}(\hat{y}_i, h_\psi(f_\theta(\mathbf{x}_i), m_M, M)), j \in [M]$.

In the standard assumption of L2D, **human experts perform better than the classifier** in general. Hence, naively minimising the loss in Eq. (2) results in approximately zero coverage, meaning that the gating model will most likely defer to or complement with a human expert without letting the classifier solely making decision at all. We, therefore, propose to integrate a constraint into the optimisation in Eq. (2) to control the coverage as follows:

$$1/N \sum_{i=1}^N g_\phi^{(\text{AI})}(\mathbf{x}_i) \geq \varepsilon, \tag{3}$$

where: $g_\phi^{(\text{AI})}(.)$ is the probability of selecting AI alone produced by the gating model, and ε is the targeted coverage.

Optimising the loss in Eq. (2) with a constraint in Eq. (3) is, however, difficult. We, therefore, employ the *penalty method* (Nocedal and Wright 1999, Chapter 17), which is common in optimisation to convert a constrained optimisation into an unconstrained one by introducing a penalty term into the main optimisation. That new positive loss term is large if constraints are violated and reaches zero when constraints are met. In general, the constrained optimisation can be rewritten into a penalty program as follows:

$$\min_{\theta,\phi,\psi} 1/N \sum_{i=1}^N g_\phi^\top(\mathbf{x}_i)\boldsymbol{\ell}(\mathbf{x}_i, \hat{y}_i, \mathcal{M}_i, \theta, \psi) + \beta_k c(\phi, \varepsilon), \tag{4}$$

where k indexes the training iteration, and $\beta_{k+1} = \lambda(\beta_k + k)$, with $\beta_1, \lambda > 0$ being hyper-parameters, and $c(\phi, \varepsilon)$ is the penalty function of the constraint in (3) defined as:

$$c(\phi, \varepsilon) = \left[\max\left(0, \varepsilon - \tfrac{1}{N}\sum_{i=1}^N g_\phi^{(\text{AI})}(\mathbf{x}_i)\right)\right]^2. \tag{5}$$

The hyper-parameters λ and β_1 in Eq. (4) convert a constrained optimisation into a non-constrained one. Specifically, when the coverage constraint is violated, the penalty loss term in Eq. (4) becomes larger (as $c(\phi, \varepsilon) > 0$ and β_k increases w.r.t the training iteration k), forcing the training of the gating model to satisfy the constraint as training progresses. Otherwise, $c(\phi, \varepsilon)$ is approximately zero, allowing the gating model to focus on minimising the misselection loss. It is important to emphasise that the collaboration cost is addressed by the input parameter ϵ, while the hyper-parameters λ and β_1 are defined to enable a progressive increase of β_k during training. In our experiments, we fix $\beta_1 = 1$ and provide a study on different values for λ in the ablation studies.

The train and test procedures of CL2DC are summarised in Algorithms 1 and 2 of the supplementary material.

Experiments

We evaluate the performance of CL2DC on a variety of datasets including ones with synthetic experts (e.g., CIFAR-100 (Krizhevsky 2009), HAM10000 (Tschandl, Rosendahl, and Kittler 2018) and Galaxy Zoo (Bamford et al. 2009)) and real-world ones with human's annotations (e.g., Chaoyang (Zhu et al. 2021), MiceBone (Schmarje et al. 2022) and NIH-ChestXray (Majkowska et al. 2020)).

Datasets For CIFAR-100, we follow the setting of (Hemmer et al. 2023) to generate synthetic labels representing synthetic experts. We generate 3 experts, each one labelling correctly on 6 or 7 different super-classes, while making 50% labelling mistakes on the remaining 13 or 14 super-classes using asymmetric label noise (i.e., labels can be randomly flipped to other classes within the same super-class). For HAM10000 and Galaxy-zoo, we follow the set-

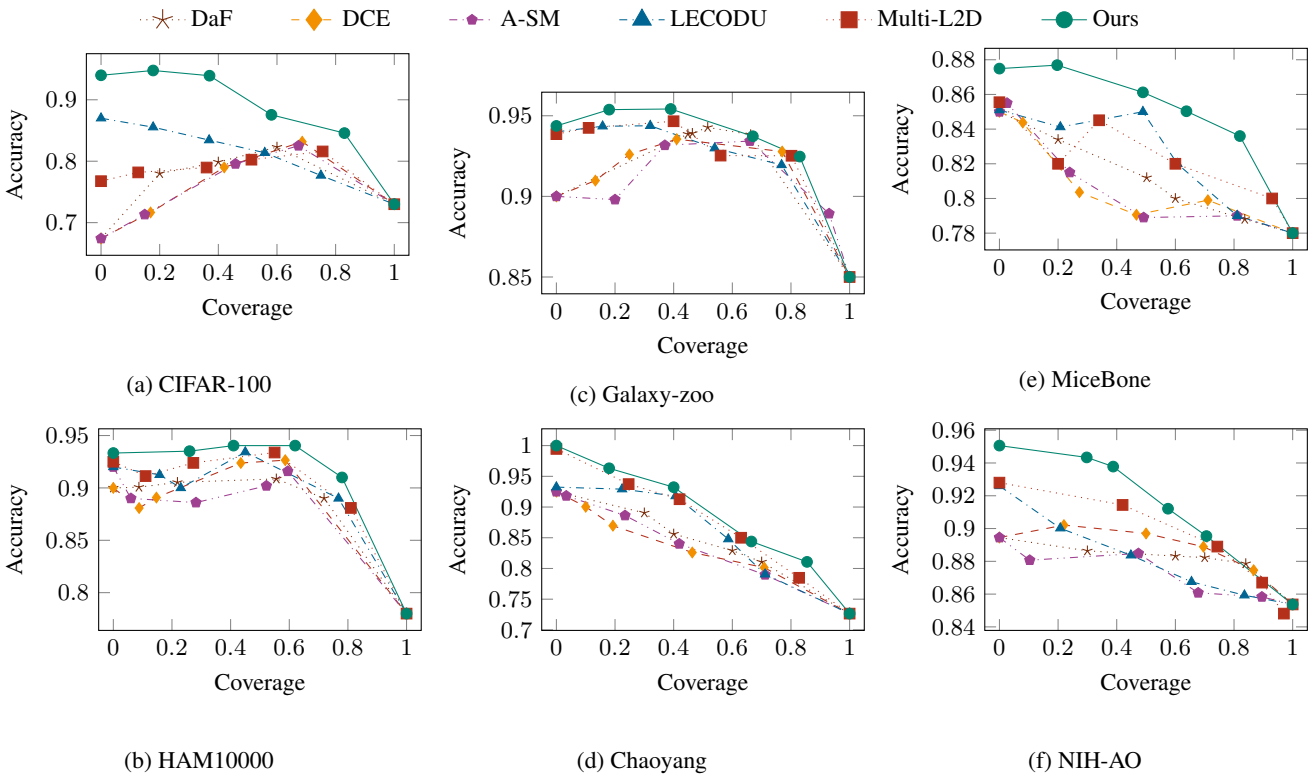


Figure 3: Accuracy-coverage curves of our method and competing SEHAI-CC (Mozannar et al. 2023; Charusaie, Fesharaki, and Samadi 2024; Wei, Cao, and Feng 2024; Cao et al. 2024) and MEHAI-CC (Zhang et al. 2025a; Verma, Barrejon, and Nalisnick 2023) methods.

ting in (Verma, Barrejon, and Nalisnick 2023) to simulate two experts based on two super-classes, each following an asymmetric label noise, similarly to CIFAR-100.

For real-world datasets, we utilise the annotations made by real-world human experts. In Chaoyang dataset, there are three human experts with accuracies 91%, 88% and 99%, and we consider two setups: 1) "Chaoyang2Exp" consists of the two pathologists with accuracies 88% and 91%, respectively; and 2) "Chaoyang3Exp" consists of all three pathologists. In MiceBone, we consider eight annotators who label the whole dataset to represent human experts, where each has an accuracy varying from 84% to 86%. In NIH-ChestXray, three human experts who label four radiographic findings are used. Similar to previous studies (Mozannar et al. 2023; Hemmer et al. 2022), we focus on the classification of airspace opacity (NIH-AO) because of the balanced prevalence of this finding. The prediction accuracies of the three experts in the NIH-AO dataset are approximately 89%, 94%, 80% both in training and testing. Please refer to the supplementary material for more details on the datasets, training parameters and training time.

Evaluation The evaluation is based on the prediction accuracy as a function of coverage measured on the testing sets. Coverage denotes the percentage of samples classified by the AI model alone, with 100% coverage representing the classification performed exclusively by the AI model, while 0% coverage denoting a classification exclusively done by experts. We report the mean result computed from 3 runs, each is evaluated at the last training epoch.

Baselines We assess our method in both single and multiple expert human-AI cooperation classification (SEHAI-CC and MEHAI-CC) settings. For the SEHAI-CC setting, we consider several SOTA methods, such as Asymmetric SoftMax (A-SM) (Cao et al. 2024), Dependent Cross-Entropy (DCE) (Wei, Cao, and Feng 2024), and defer-and-fusion (DaF) (Charusaie, Fesharaki, and Samadi 2024) as baselines. For a fair comparison, we randomly sample a single annotation for each image as a way to simulate a single expert from the human annotators to train those SEHAI-CC methods. For the MEHAI-CC settings, we consider the L2D to multiple experts (MultiL2D) (Verma, Barrejon, and Nalisnick 2023) and learning to complement and to defer to multiple experts (LECODU) (Zhang et al. 2025a). All classification for the $\{$SE,ME$\}$HAI-CC methods have the same backbone, and all hyper-parameters are set as previously reported in (Mozannar et al. 2023; Zhang et al. 2025a; Verma, Barrejon, and Nalisnick 2023). To maintain fairness in the accuracy-coverage comparisons, we integrate the coverage constraint in Eq. (3) into all baseline methods. In particular, we set the hyper-parameter ε to control the coverage lower bound, to $\{0, 0.2, 0.4, 0.6, 0.8\}$, so we can train all methods and plot their corresponding accuracy-coverage curves.

Results We report the *accuracy-coverage curves* of the HAI-CC strategies and our proposed method across various datasets in Fig. 3. These curves illustrate the trade-off between accuracy and cooperation cost as coverage varies from 0% to 100%, where 0% coverage indicates complete reliance on human experts, and 100% coverage implies classification solely by the AI model. We also provide a concise quantitative analysis by calculating the *area under the accuracy-coverage curve* (AUACC) of Fig. 3 results and showing in Tab. 1 of the appendix. The higher AUACC values, the more optimal the accuracy - coverage trade-offs.

In general, our method outperforms all competing HAI-CC methods at every coverage level in all benchmarks. Compared with MEHAI-CC methods, the accuracy of SEHAI-CC methods is limited by the lack of specific expert labelling. Consequently, MEHAI-CC methods generally surpass SEHAI-CC approaches, particularly at lower coverage values. However, even in this scenario they still do not match the performance of our method.

In synthetic datasets (i.e., CIFAR-100, Galaxy-zoo and HAM10000), we focus on the setting that different experts have relatively high accuracy on specific categories , as explained in the supplementary material. The proposed method CL2DC excels by effectively identifying and cooperating with specific experts for their relevant tasks, thereby optimising decision-making. In the CIFAR-100 dataset, LECODU achieves higher accuracy than SEHAI-CC at low and intermediate coverage levels but shows lower accuracy than our approach. Another baseline is MultiL2D that performs comparatively to or even outperforms SEHAI-CC. This is due to the ability to identify the best labeller in MultiL2D, as opposed to SEHAI-CC methods that pick one of the labellers randomly. In the Galaxy-zoo dataset, other MEHAI-CC methods (i.e., LECODU and MultiL2D) show lower accuracy than ours, but they become relatively competitive at higher coverage levels, which underscores our method's superior adaptability and efficiency in optimising human-AI cooperation. Compared with them, our method excels by effectively identifying and collaborating with specific experts for their relevant tasks, thereby optimising decision-making.

In real-world scenarios (i.e., Chaoyang, Micebone, and NIH-AO), our method consistently outperforms other strategies. Notably, in Chaoyang, where one of the pathologists has an accuracy close to 100%, our method adeptly selects this most accurate pathologist, surpassing the performance of LECODU, which randomly selects an expert rather than identifying the optimal one. Although the performance of MultiL2D is competitive in Chaoyang, it is worse than our method in the Micebone and NIH-AO datasets.

Tab. 2 shows a few examples of the inference of CL2DC at a coverage rate of 40% on test images of Galaxyzoo. Each example includes a test image, the human-provided labels ($\mathcal{M}$), classifier's prediction ($f_\theta(.)$), complementary module prediction $h_\psi(.)$, prediction probability vector by the gating model ($g_\phi(.)$), final prediction of CL2DC, and ground truth (GT) label. Notably, when the classifier or the human experts make individual mistakes, the final prediction tends to be correct, highlighting system robustness. When the classifier is correct, the probability selecting the classifier alone, $g_\phi^{\mathrm{AI}}(.)$, tends to be high, suggesting that the classifier can be trusted. When the L2D options are selected, that usually happens with very high probability for one of the options in $g_\phi^{\mathrm{L2D}_j}(.)|_{j=1}^{M}$ and quite low value for $g_\phi^{\mathrm{AI}}(.)$, suggesting a complete lack of trust in the classifier. On the other hand, when one of the L2C options are selected, notice that both $g_\phi^{\mathrm{AI}}(.)$ and one of the options in $g_\phi^{\mathrm{L2C}_j}(.)|_{j=1}^{M}$ show high values, indicating that the classifier can be partially trusted.

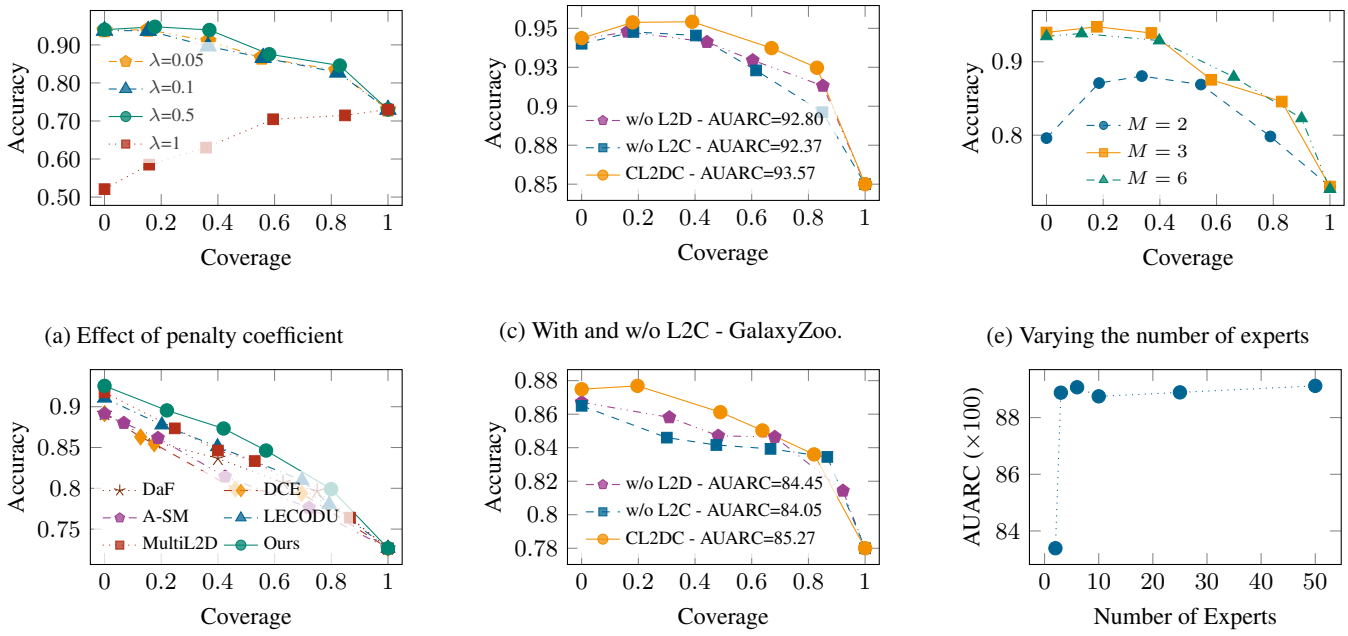
Ablation Studies

Penalty coefficient We study the effect of the penalty coefficient β_k (with $\beta_1 = 1$ fixed at the first epoch) in Eq. (4) via the hyper-parameter λ on CIFAR-100. As shown in Fig. 4a, the setup with $\beta_1 = 1$ consistently demonstrates that when $\lambda = 1$, accuracy is notably low at lower coverage levels due to the strong influence of the constraint in Eq. 5. Reducing λ, however, yields higher accuracy across nearly all coverage levels. We did not observe any perceptible changes in results when using other values for β_1 that are close to 1, indicating that the outcomes are robust to minor variations around this value.

Diverged expertise of human experts We investigate the effect of altering the number and quality of experts in the experimental setting. Focusing on the "Chaoyang2Exp" setup, the outcomes, displayed in Fig. 4b and Tab. **??** (appendix), show that our method outperforms other HAI-CC methods more distinctly when compared with the original results in Fig. 3d that use all three pathologists for the "Chaoyang3Exp" setup. This highlights the robustness of our proposed method when varying the expertise of human experts.

Effectiveness of L2D and L2C We study the influence of L2D and L2C by modifying the deferral options of the gating model on Galaxyzoo and Micebone datasets in Fig. 4c and Fig. 4d. CL2DC w/o L2D denotes that the decision process only contains the prediction of the classifier and L2C options, while CL2DC w/o L2C represents that the decision is made by the classifier or an expert without any complement. In Fig. 4c, CL2DC w/o L2D outperforms CL2DC w/o L2C at large coverage values, meaning that when the expert's accuracy is high, L2C can leverage the accurate expert's prediction, while mitigating the influence of weak experts by combining their predictions with the prediction made by the classifier. CL2DC outperforms CL2DC w/o L2C and CL2DC w/o L2D at all coverage values, showing the advantage of integrating both L2D and L2D into HAI-CC. In Fig. 4d, CL2DC w/o L2C performs better than CL2DC w/o L2D, when coverage is larger than 0.6. At a large coverage, L2C may combine a weak expert especially when the expert pool contains a large number of experts who have relatively low accuracies (from 84% to 86%). In general, CL2DC tends to work better than CL2DC w/o L2C and CL2DC w/o L2D for most coverage values by leveraging advantages of both L2D and L2C.

Number of human experts We further study the scalability of CL2DC w.r.t. the number of experts on CIFAR-100 by



(a) Effect of penalty coefficient

(c) With and w/o L2C - GalaxyZoo.

(e) Varying the number of experts

(b) Effect of human expertise

(d) With and w/o L2C - MiceBone.

(f) Adding more experts

Figure 4: Ablation studies on CL2DC in various settings: (a) different penalty coefficients on CIFAR-100, (b) varying human's expertise on Chaoyang with 2 experts, (c) and (d) the influence of L2D and L2C on Galaxyzoo and Micebone, respectively, and (e) and (f) varying the number of experts on CIFAR-100.

generating additional 47 synthetic experts, each performing similarly to the ones described in the "Datasets" paragraph above, i.e., predicting correctly on 6/7 random super-classes, while making 50% mistakes via instance-dependent noise on the remaining super-classes. The results of accuracy-coverage curves and AUACC in Figs. 4e and 4f show a significant improvement when increasing from 2 to 3 human experts and quickly stabilises with more experts. Such an observation is due to the "complement" of expertise. Specifically, the first two synthetic experts can complementarily make correct predictions on only two-third of all super-classes, while adding another expert could cover the whole sets of the super-classes. Hence, it is unsurprising that the AUACC tends to stabilise as the number of experts continues to increase, indicating that additional redundant experts results in diminishing returns.

Coverage values during inference We show the gating model's averaged probability output on the Galaxy-Zoo test set at different coverage values in Fig. 5 of the Appendix (accuracy result for the same dataset is in Fig. 3c). These results show that the coverage levels (on the hor. axis) match the gating output for the AI model (blue bar), while the distribution for the other gating outputs shows a slight preference for experts alone, particularly in high-coverage scenarios.

Conclusion

In this paper, we propose the novel Coverage-constrained Learning to Defer and Complement with Specific Experts (CL2DC) method. CL2DC integrates the strengths of learning-to-defer and learning-to-complement, particularly in training scenarios with multiple noisy-label annotations, enabling the system to either make final decisions autonomously or cooperate with a specific expert. We also introduce and integrate coverage-constraint through an innovative penalty method into the loss function to control the coverage. This penalty allows us to run a robust training procedure where the target coverage can be reached. Such an approach enables a reliable analysis of different methods through the coverage - accuracy curves. Comprehensive evaluations across real-world and synthetic multiple noisy label datasets demonstrate CL2DC's superior accuracy to SOTA HAI-CC methods.

Acknowledgements

This work has been funded by the Engineering and Physical Sciences Research Council (EPSRC) through grant EP/Y018036/1.

References

Bamford, S. P.; Nichol, R. C.; Baldry, I. K.; Land, K.; Lintott, C. J.; Schawinski, K.; Slosar, A.; Szalay, A. S.; Thomas, D.; Torki, M.; Andreescu, D.; Edmondson, E. M.; Miller, C. J.; Murray, P.; Raddick, M. J.; and Vandenberg, J. 2009. Galaxy Zoo: The dependence of morphology and colour on environment. *Monthly Notices of the Royal Astronomical Society*, 393(4): 1324–1352.

Cao, Y.; Mozannar, H.; Feng, L.; Wei, H.; and An, B. 2024. In Defense of Softmax Parametrization for Calibrated and Consistent Learning to Defer. In *NeurIPS*, volume 36.

Carneiro, G. 2024. *Machine Learning with Noisy Labels: Definitions, Theory, Techniques and Solutions*. Elsevier Science.

Charusaie, M.-A.; Fesharaki, A. J.; and Samadi, S. 2024. Defer-and-Fusion: Optimal Predictors that Incorporate Human Decisions. In *5th Workshop on practical ML for limited/low resource settings*.

Dafoe, A.; Bachrach, Y.; Hadfield, G.; Horvitz, E.; Larson, K.; and Graepel, T. 2021. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857): 33–36.

Ding, D.; Mallick, A.; Wang, C.; Sim, R.; Mukherjee, S.; Ruhle, V.; Lakshmanan, L. V.; and Awadallah, A. H. 2024. Hybrid LLM: Cost-efficient and quality-aware query routing. In *ICLR*.

Goh, H. W.; Tkachenko, U.; and Mueller, J. 2022. CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. In *NeurIPS Human in the Loop Learning Workshop*.

Green, B.; and Chen, Y. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *FAccT*, 90–99.

Halling-Brown, M. D.; Warren, L. M.; Ward, D.; Lewis, E.; Mackenzie, A.; Wallis, M. G.; Wilkinson, L. S.; Given-Wilson, R. M.; McAvinchey, R.; and Young, K. C. 2020. OPTIMAM mammography image database: A large-scale resource of mammography images and clinical data. *Radiology: Artificial Intelligence*, 3(1): e200103.

Hemmer, P.; Schellhammer, S.; Vössing, M.; Jakubik, J.; and Satzger, G. 2022. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. In *IJCAI*, 2478–2484.

Hemmer, P.; Thede, L.; Vössing, M.; Jakubik, J.; and Kühl, N. 2023. Learning to defer with limited expert predictions. In *AAAI*, volume 37, 6002–6011.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Liu, Y.; Cheng, H.; and Zhang, K. 2023. Identifiability of label noise transition matrix. In *ICML*, 21475–21496.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. In *NeurIPS*, volume 31.

Majkowska, A.; Mittal, S.; Steiner, D. F.; Reicher, J. J.; McKinney, S. M.; Duggan, G. E.; Eswaran, K.; Cameron Chen, P.-H.; Liu, Y.; Kalidindi, S. R.; Ding, A.; Corrado, G. S.; Tse, D.; and Shetty, S. 2020. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294: 421–431.

Mao, A.; Mohri, C.; Mohri, M.; and Zhong, Y. 2023. Two-stage learning to defer with multiple experts. In *NeurIPS*.

Mao, A.; Mohri, M.; and Zhong, Y. 2024. Regression with Multi-Expert Deferral. In *ICML*.

Mozannar, H.; Lang, H.; Wei, D.; Sattigeri, P.; Das, S.; and Sontag, D. 2023. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. In *AISTATS*.

Mozannar, H.; and Sontag, D. 2020. Consistent estimators for learning to defer to an expert. In *ICML*, 7076–7087.

Narasimhan, H.; Jitkrittum, W.; Menon, A. K.; Rawat, A.; and Kumar, S. 2022. Post-hoc estimators for learning to defer to an expert. In *NeurIPS*, volume 35, 29292–29304.

Nocedal, J.; and Wright, S. J. 1999. *Numerical optimization*. Springer.

Schmarje, L.; Santarossa, M.; Schröder, S.-M.; Zelenka, C.; Kiko, R.; Stracke, J.; Volkmann, N.; and Koch, R. 2022. A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. In *ECCV*, 363–380. Springer.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.

Verma, R.; Barrejon, D.; and Nalisnick, E. 2023. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. In *AISTATS*, 11415–11434. PMLR.

Wang, H.; Xiao, R.; Dong, Y.; Feng, L.; and Zhao, J. 2023. ProMix: combating label noise via maximizing clean sample utility. In *IJCAI*.

Wei, H.; Xie, R.; Feng, L.; Han, B.; and An, B. 2022. Deep learning from multiple noisy annotators as a union. *IEEE TNNLS*.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2021. Learning with noisy labels revisited: A study using real-world human annotations. In *ICLR*.

Wei, T.; Li, H.-T.; Li, C.-S.; Shi, J.-X.; Li, Y.-F.; and Zhang, M.-L. 2024. Vision-Language Models are Strong Noisy Label Detectors. In *NeurIPS*.

Wei, Z.; Cao, Y.; and Feng, L. 2024. Exploiting Human-AI Dependence for Learning to Defer. In *ICML*.

Wilder, B.; Horvitz, E.; and Kamar, E. 2021. Learning to Complement Humans. In *IJCAI*.

Zhang, Z.; Ai, W.; Wells, K.; Rosewarne, D.; Do, T.-T.; and Carneiro, G. 2025a. Learning to Complement and to Defer to Multiple Users. In *ECCV*, 144–162. Springer.

Zhang, Z.; Nguyen, C.; Wells, K.; Do, T.-T.; and Carneiro, G. 2025b. Learning to complement with multiple humans. *Pattern Recognition*, 112376.

Zhu, C.; Chen, W.; Peng, T.; Wang, Y.; and Jin, M. 2021. Hard sample aware noise robust learning for histopathology image classification. *IEEE TMI*, 41(4): 881–894.