

Coverage-Constrained Human-AI Cooperation with Multiple Experts

Zheng Zhang¹, Cuong Nguyen¹, Kevin Wells¹, Toan Do², David Rosewarne^{1,3}, and Gustavo Carneiro¹

¹*Centre for Vision, Speech and Signal Processing, University of Surrey, United Kingdom*

²*Department of Data Science and AI, Monash University, Australia*

³*Royal Wolverhampton Hospitals NHS Trust, UK*

Abstract

Human-AI cooperative classification (HAI-CC) approaches aim 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, and learning-to-complement (L2C), where AI and human experts make predictions cooperatively. However, a notable research gap remains in effectively exploring both L2D and L2C under diverse expert knowledge to improve decision-making, particularly when constrained by the cooperation cost required to achieve a target probability for AI-only selection (i.e., coverage). In this paper, we address this research gap by proposing the Coverage-constrained Learning to Defer and Complement with Specific Experts (CL2DC) method. CL2DC makes final decisions through either AI prediction alone or by deferring to or complementing a specific expert, depending on the input data. 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. Also, CL2DC is designed to address scenarios where training sets contain multiple noisy-label annotations without any clean-label references. Comprehensive evaluations on both synthetic and real-world datasets demonstrate that CL2DC achieves superior performance compared to state-of-the-art HAI-CC methods.

1 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 [24], content moderation [43], breast cancer classification [26], and the detection of inaccurate or deceptive content produced by large language models [7, 19, 20, 55, 83], 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 [18, 62, 70]. These HAI-CC approaches enhance not only the accuracy, interpretability, and usability

of AI models but also improve user efficiency and decision consistency over manual processes, significantly reducing human error [18, 70].

HAI-CC approaches [18] aim to develop a *hybrid intelligent* system that maximises accuracy while minimising the cooperation costs with *learning-to-defer* (L2D) and *learning-to-complement* (L2C) strategies. In L2D [52], HAI-CC strategically decides when to classify with the AI model or defer to human experts, while L2C [86] combines the predictions of AI and human experts.

When facing with challenging or high-stakes decisions, single-expert HAI-CC (SEHAI-CC) systems allow the system to defer to or complement with a fixed expert [12–14, 49, 52, 54, 57, 59, 63, 74, 76, 86]. 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 [1, 28, 35, 47, 50, 72, 77, 78, 94, 95], 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 [95], they effectively disregard diverse expert knowledge by randomly selecting experts for the cooperative classification.

Another crucial issue in HAI-CC methods 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 [54, 57, 95]. Coverage is defined as the *percentage of examples classified by the AI model alone*, where 100% coverage indicates that all classifications are performed by the AI, and 0% coverage means that all classifications are handled exclusively by experts. HAI-CC methods [54, 57] are typically trained with optimisation functions that aim to balance accuracy and cooperation cost. However, in practice, the training process is brittle; small adjustments to the hyper-parameter controlling accuracy and cooperation cost often result in coverage values collapsing to either 0% or 100% [94, 95]. 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 HAI-CC methods [54, 57] often employ *post-hoc* technique to construct accuracy - coverage curves by sorting deferral scores and adjusting the prediction threshold to obtain the accuracy at the expected coverage. This post-hoc approach is impractical, as it requires access to all testing samples before making predictions and lacks the ability to set and evaluate workload control during training. In addition, using the post-hoc approach to analyse the coverage - accuracy of a model trained in one coverage setting is unreliable. For example, Fig. 1 illustrates that models of the same method, but trained under different levels of coverage constraints, yield different curves using the post-hoc method (e.g., the orange and green curves). Hence, reporting the result by simply selecting the best performing method does not represent a reliable assessment of the approach. Therefore, further research is needed to develop a principled mechanism for managing workload distribution in HAI-CC methods.

In this paper, we propose the novel **Coverage-constrained Learning to Defer and Complement with Specific Experts (CL2DC)** method. 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 cooperate with a

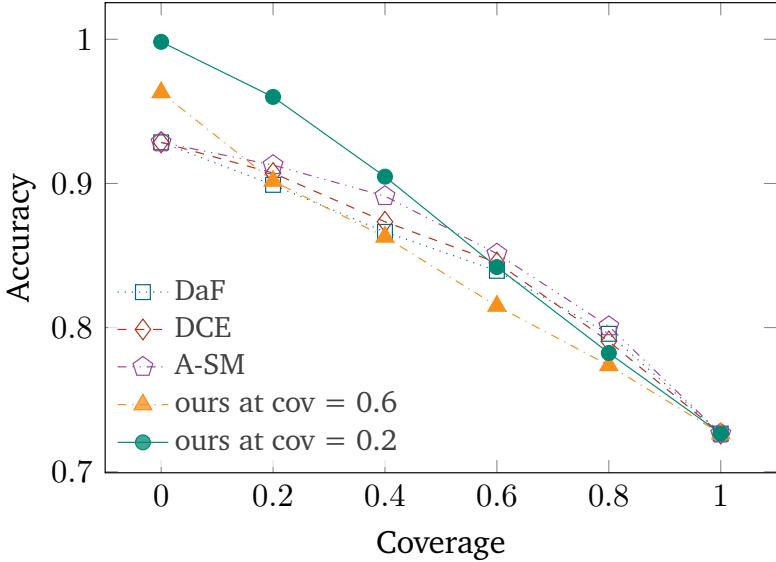


Figure 1: The post-hoc analysis to generate the coverage - accuracy curves of our proposed method on the Chaoyang dataset [96] is unreliable because the same method trained with different coverage constraints produces different curves. When comparing to several HAI-CC methods [8, 14, 85] plotted with 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 coverages (i.e., 0.6 in orange and 0.2 in green) show quite different performances.

specific expert. 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 coverage levels. This penalty enables a robust training process that reliably achieves the target coverage, allowing for a consistent and meaningful analysis of various methods using coverage-accuracy curves. Our main contributions are summarised as follows:

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

We evaluate our CL2DC method against state-of-the-art (SOTA) HAI-CC methods [8, 14, 78, 85, 95] using real-world and synthetic multi-rater noisy label benchmarks, such as CIFAR-100 [6, 82], Galaxy Zoo [4], HAM10000 [75], NIH-ChestXray [45, 80], Mice-Bone [66–68], and Chaoyang [96]. Results show that CL2DC consistently outperforms previous HAI-CC methods with higher accuracy for equivalent coverage values for all benchmarks.

2 Related work

2.1 Human-AI Cooperative Classification

HAI-CC approaches [18] 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) is an extension of *rejection learning* [17], which 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 [35, 44, 46, 57]. Existing L2D approaches focus on the development of different surrogate loss functions to be consistent with the Bayes-optimal classifier [8, 11–13, 41, 52–54, 57, 59, 63, 71, 76]. Wei et al. [85] explore the dependence between AI and human experts and propose a dependent Bayes optimality formulation for the L2D problem. However, these methods overlook practical settings in which there is a wide diversity of multiple human experts. Given such an issue, recent research in L2D shifts towards the multiple-expert setting [1, 3, 29, 35, 38, 46, 48, 50, 72, 77, 78, 94]. For example, Verma et al. [78] proposed a L2D method to defer the decision to one of multiple experts. Mao et al. [50] addressed both instance-dependent and label-dependent costs and proposed a novel regression surrogate loss function. Nevertheless, current research in L2D does not consider options that aggregate the predictions of human experts and AI model to make a joint decision.

Learning to Complement (L2C) methods aim to optimise the cooperation between human experts and the AI model by combining their predictions [5, 14, 28, 34, 35, 40, 70, 74, 86, 94, 95]. Liu et al. [40] leverage perceptual differences through post-hoc teaming, showing that human - machine collaboration can be more accurate than machine - machine collaboration. Recently, Charusaie et al. [14] introduce a method that determines whether the AI model 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. [28] introduce a model featuring an ensemble prediction involving both AI and human predictions, yet it does not optimise the cooperation cost. Zhang et al. [95] propose a combine L2D-L2C approach that integrates AI predictions with multiple random human experts, but overlooking the expert specificity.

2.2 Learning with Noisy Labels

The vast majority of HAI-CC methods assume that the ground-truth *clean* annotations are available in the training set. The vast majority of HAI-CC methods assume that the ground-truth clean annotations are available in the training set. However, such 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, several HAI-CC systems have been designed to handle noisy-label problems. Here, we provide a short review of learning with noisy labels (LNL) methods that can be used in the HAI-CC context [10, 69].

LNL approaches have explored many techniques, including: robust loss functions [22, 92], co-teaching [27, 33], label cleaning [31, 90], semi-supervised learning (SSL) [39,

[60], iterative label correction [2, 16], meta-learning [65, 89, 91, 93], diffusion models [15], and graphical models [21]. Among the top-performing LNL methods, we have ProMix [79] that introduces an optimisation based on a matched high-confidence selection technique. Another notable LNL approach is DEFT [84] that utilises the robust alignment of textual and visual features, pre-trained on millions of auxiliary image-text pairs, to sieve out noisy labels. Despite achieving remarkable results, LNL with a single noisy label per sample suffers from the identifiability issue [42], 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, which can be divided into inter- and intra-annotator approaches. The inter-annotator methods focus on characterising the variabilities between annotators [25, 32, 51, 64], while the intra-annotator approaches focus on estimating the variability of each annotator [9, 36, 73, 87]. Recently, UnionNet [81] 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. Annot-Mix [30] extend mixup and estimates each annotator’s performance during the training process. CROWDLAB [23] 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.

3 Methodology

Let $\mathcal{D} = \{\mathbf{x}_i, \mathcal{M}_i\}_{i=1}^N$ be the noisy-label multi-rater training set of size N , where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H \times W \times R}$ denotes an input image of size $H \times W$ with R channels, 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 , with C being the number of classes.

Our proposed method contains three components, an AI classifier, a gating model, and a complementary module, which form an adaptive decision system that leverages AI strengths while maintaining human oversight, ensuring a balance between performance and trustworthiness in complex decision environments. More specifically, we have: 1) an *AI classifier* $f_\theta : \mathcal{X} \rightarrow \Delta^{C-1}$, parametrised by $\theta \in \Theta$, where $\Delta^{K-1} = \{\mathbf{p} : \mathbf{p} \in [0, 1]^K \wedge \mathbf{1}^\top \mathbf{p} = 1\}$ denotes the $(K - 1)$ -dimensional probability simplex; 2) a *gating model* $g_\phi : \mathcal{X} \rightarrow \Delta^{2M+1}$, parameterised by $\phi \in \Phi$, which produces a categorical distribution reflecting the probability of selecting the prediction made by the AI model alone (i.e., $g_\phi^{(\text{AI})}(\cdot)$), or deferring the decision to one of the M human experts (i.e., $g_\phi^{(\text{L2D}_j)}(\cdot)|_{j=1}^M$ denote the probability of selecting expert 1 through expert M), or performing a complementary classification between the AI model and one human expert (i.e., $g_\phi^{(\text{L2C}_j)}(\cdot)|_{j=1}^M$ represent the probability of selecting AI + expert 1, through AI + expert M); and 3) a *complementary module* $h_\psi : \Delta^{C-1} \times \mathcal{Y} \rightarrow \Delta^{C-1}$, parametrised by $\psi \in \Psi$, that aggregates the predictions made by the AI model and a selected human expert to produce a final prediction if the gating model decides to complement AI with one human expert. These components can be visualised in Fig. 2.

In standard HAI-CC, ground truth labels are often required for training, while in our setting, ground truth labels are unavailable. Following LECODU [95], which also assumes that ground truth labels are unavailable, we use the SOTA MRL method CROWDLAB [23] to produce the *consensus* labels to be used as the ground truth in our training. CROWDLAB

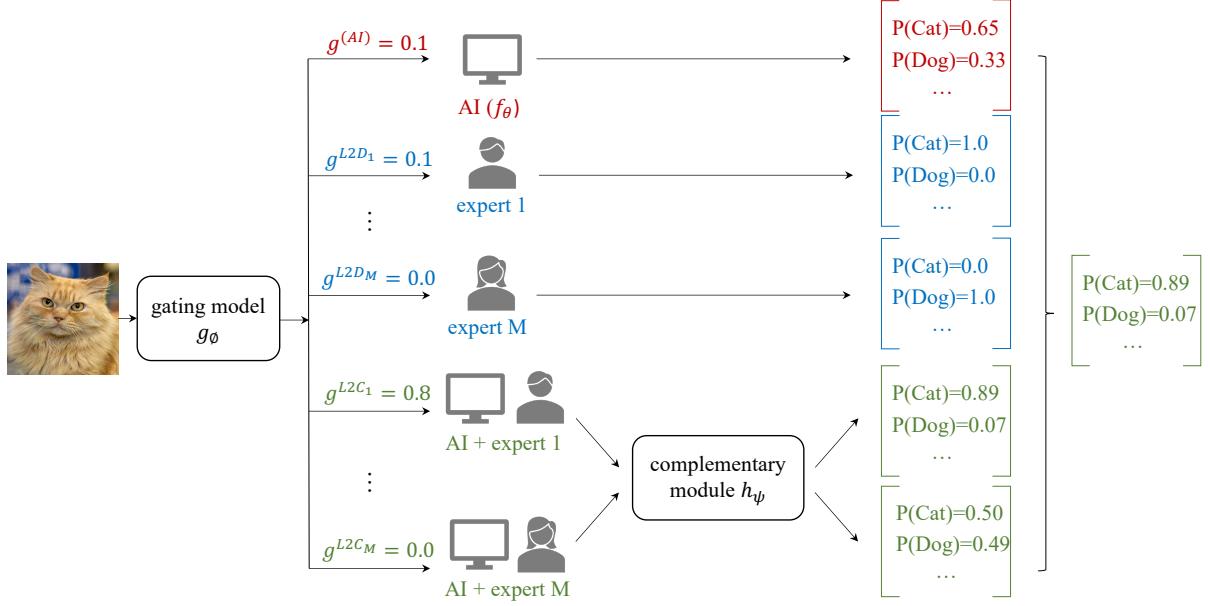


Figure 2: CL2DC contains a gating model $g_\phi(\cdot)$, a complementary module $h_\psi(\cdot)$, and an AI model $f_\theta(\cdot)$. The gating model aims to decide whether we use LNL-trained AI model $f_\theta(\cdot)$ alone (i.e., when $g_\phi^{(AI)}(\cdot)$ has the largest probability), defer the decision to one of the M experts $\{1, \dots, M\}$ (i.e., when one of the $g_\phi^{(L2D_j)}(\cdot)|_{j=1}^M$ 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 $g_\phi^{(L2C_j)}(\cdot)|_{j=1}^M$ has the largest probability). In the figure, the gating model selects L2C between AI and user 1, given its largest probability of 0.8, to make the final prediction, on the right.

takes the training images and experts’ annotations $(\mathbf{x}_i, \mathcal{M}_i) \in \mathcal{D}$, together with the AI classifier’s predictions $f_\theta(\mathbf{x}_i)$ to produce a consensus label $\hat{y}_i \in \mathcal{Y} = \{1, \dots, C\}$ associated with a quality (or confidence) score α_i . Formally, the pseudo clean data set produced by CROWDLAB can be written as follows:

$$\begin{aligned} \hat{\mathcal{D}} = & \{(\mathbf{x}_i, \hat{y}_i, \mathcal{M}_i) : i \in \{1, \dots, N\} \wedge (\mathbf{x}_i, \mathcal{M}_i) \in \mathcal{D} \wedge \hat{y}_i \in \\ & \{\hat{y}_i : (\hat{y}_i, \alpha_i) = \text{CrowdLab}(\mathbf{x}_i, f_\theta(\mathbf{x}_i), \mathcal{M}_i) \wedge \alpha_i > 0.5\}\}. \end{aligned} \quad (1)$$

Our aim is to learn the parameters of the AI model, the gating model and the complementary module to produce an accurate final prediction, while satisfying the coverage constraint with the following optimisation:

$$\begin{aligned} & \min_{\theta, \phi, \psi} \frac{1}{N} \sum_{i=1}^N g_\phi^\top(\mathbf{x}_i) \ell(\mathbf{x}_i, \hat{y}_i, \mathcal{M}_i, \theta, \psi) \\ & \text{s.t.: } \frac{1}{N} \sum_{i=1}^N g_\phi^{(AI)}(\mathbf{x}_i) \geq \varepsilon, \end{aligned} \quad (2)$$

where $g_\phi^{(AI)}(\cdot)$ is the probability of selecting AI alone produced by the gating model, and

$\ell(\cdot)$ is a $(2M + 1)$ -dimensional vector defined as:

$$\ell(\mathbf{x}_i, \hat{y}_i, \mathcal{M}_i, \theta, \psi) = \begin{bmatrix} \ell_{\text{CE}}(\hat{y}_i, f_\theta(\mathbf{x}_i)) \\ \ell_{\text{CE}}(\hat{y}_i, m_1) \\ \vdots \\ \ell_{\text{CE}}(\hat{y}_i, m_M) \\ \ell_{\text{CE}}(\hat{y}_i, h_\psi(f_\theta(\mathbf{x}_i), m_1, 1)) \\ \vdots \\ \ell_{\text{CE}}(\hat{y}_i, h_\psi(f_\theta(\mathbf{x}_i), m_M, M)) \end{bmatrix}, \quad (3)$$

with ℓ_{CE} denoting the cross-entropy loss.

Intuitively, the optimisation in Eq. (2) minimises the weighted-average loss across all available deferral options, with the weights representing the probability produced by the gating model, while the constraint enforces the average probability of selecting the AI classifier alone to be above a certain threshold ε , representing the target coverage. The lower bounded constraint is used in Eq. (2) due to the standard assumption in HAI-CC, where human experts generally perform better than the AI classifier. Thus, without imposing such a constraint, the trained gating model will most likely defer to or complement with a human expert without selecting the AI classifier to make the decision alone. In other words, $g_\phi^{(\text{AI})}(\mathbf{x}_i) \approx 0$ if there is no constraint.

To optimise the constrained objective in Eq. (2), we use the *penalty method* [58, Chapter 17]. In particular, the penalty function of the constraint in Eq. (2) is defined by

$$c(\phi, \varepsilon) = \left[\max \left(0, \varepsilon - \frac{1}{N} \sum_{i=1}^N g_\phi^{(\text{AI})}(\mathbf{x}_i) \right) \right]^2, \quad (4)$$

which approximates to zero when the constraint is satisfied and becomes positive as the constraint is violated. The loss function in Eq. (2) can then be rewritten into a penalty program as follows:

$$\min_{\theta, \phi, \psi} \frac{1}{N} \sum_{i=1}^N g_\phi^\top(\mathbf{x}_i) \ell(\mathbf{x}_i, \hat{y}_i, \mathcal{M}_i, \theta, \psi) + \beta_k c(\phi, \varepsilon), \quad (5)$$

where k indexes the training iteration, $\beta_1 > 0$, and $\beta_{k+1} = \lambda(\beta_k + k)$, with $\lambda > 0$ being a hyper-parameter. Hence, $\beta_{k+1} > \beta_k$ for all training iterations k , meaning the training process initially prioritizes minimizing classification loss and gradually shifts its focus toward achieving the target coverage.

The training and testing procedures of CL2DC are summarised in Algorithms 1 and 2 in the supplementary material. We also depict the inference flow of CL2DC in Fig. 2.

4 Experiments

We evaluate the performance of the proposed method on a variety of datasets including ones with synthetic experts (e.g., CIFAR-100 [6, 37], HAM10000 [75] and Galaxy Zoo [4]) and real-world ones with labels provided by human experts (e.g., Chaoyang [96], MiceBone [66–68] and NIH-ChestXray [45, 80]).

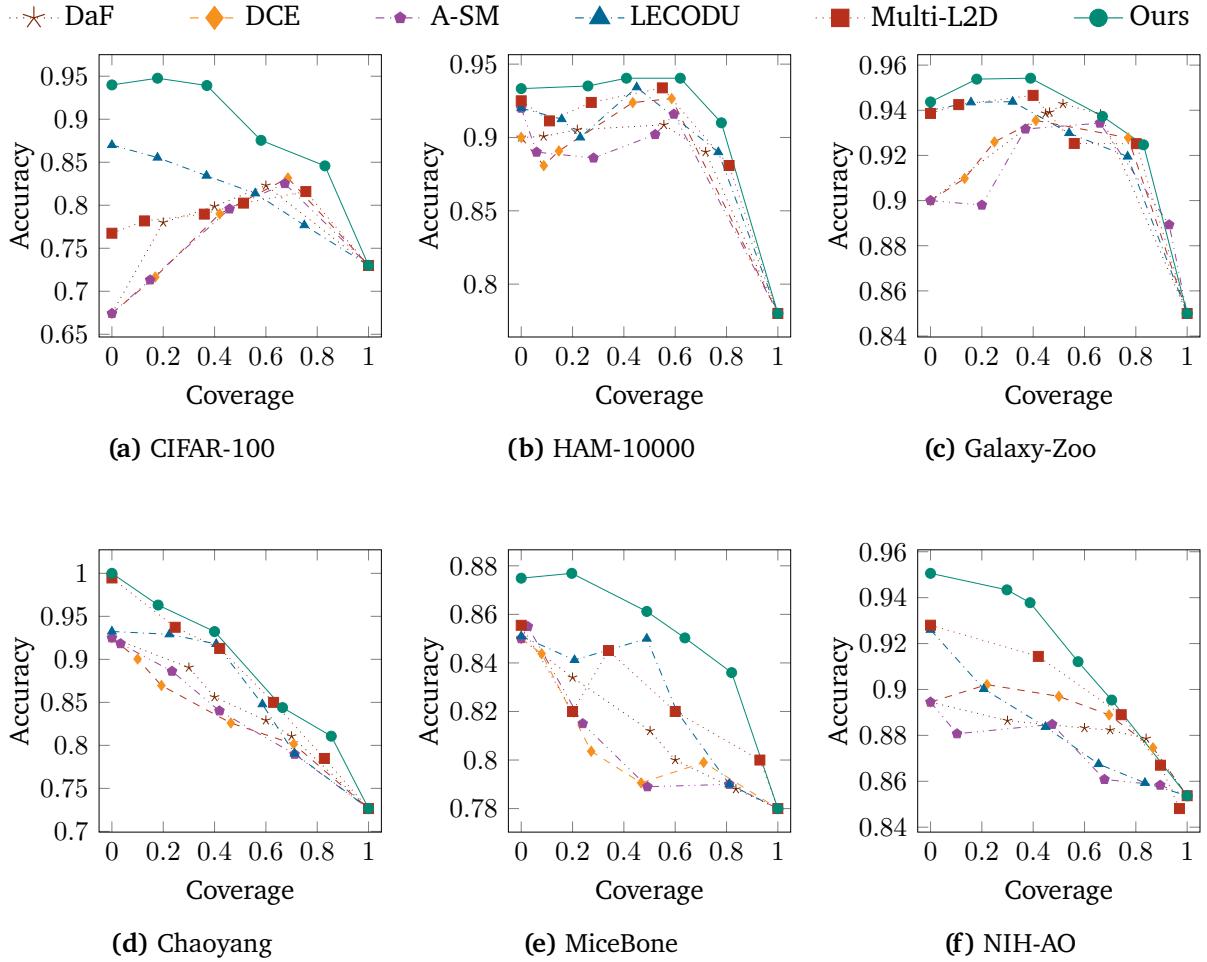


Figure 3: Accuracy-coverage curves of our method and competing SEHAI-CC [8, 14, 54, 85] and MEHAI-CC [78, 95] methods.

4.1 Implementation Details

Datasets For CIFAR-100, we follow the setting of [29] 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 superclasses using asymmetric label noise, where labels can be randomly flipped to other classes within the same super-class. For HAM10000 and Galaxy-zoo, we follow the setting in [78] 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 3 human experts with accuracies 91%, 88% and 99%, and we build 2 setups: 1) using the two pathologists with accuracies 88% and 91%, forming the setup ‘‘Chaoyang2u’’; and 2) using all 3 pathologists, forming the setup ‘‘Chaoyang3u’’. In MiceBone dataset, we use 8 out of 79 annotators who label the whole dataset to represent the experts with accuracies varying from 84% to 86%. In NIH-ChestXray dataset, each image is annotated by 3 experts who label four radiographic findings. Following [28, 54], we focus on the classification of airspace opacity (NIH-AO) because of the balanced prevalence of this finding. The prediction accuracies of the 3 experts in the NIH-AO dataset are approximately 89%, 94%, 80% both in training and

Methods	Type	CIFAR-100	Chaoyang2u	Chaoyang3u	NIH-AO	Micebone	HAM10000	Galaxy-zoo
A-SM	SEHAI-CC	76.41	80.84	82.02	87.33	80.31	87.51	91.28
DCE	SEHAI-CC	76.74	80.73	82.49	88.63	80.34	88.48	91.65
DaF	SEHAI-CC	77.60	81.91	84.07	88.26	81.19	88.43	91.53
MultiL2D	MEHAI-CC	78.84	83.02	87.93	90.14	82.31	89.55	92.38
LECODU	MEHAI-CC	81.17	82.89	85.71	88.19	82.47	90.15	92.67
Ours	MEHAI-CC	88.88	84.79	88.91	91.53	85.27	91.49	93.57

Table 1: Quantitative comparison in terms of the Area Under Accuracy-Coverage Curve (AUACC) [56] of the SOTA SEHAI-CC [8, 14, 54, 85], MEHAI-CC [78, 95] on the human-AI cooperation classification datasets. The best result per benchmark is marked in bold.

testing. Please refer to Appendices B and C in the supplementary material for more details on the datasets, architecture and training parameters.

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. All results are computed from the mean result from 3 runs using the checkpoint obtained at the last training epoch.

Baselines We assess our method in both the 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) [8], Dependent Cross-Entropy (DCE) [85], and defer-and-fusion (DaF) [14] 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 learning to defer to multiple experts (MultiL2D) [78] and learning to complement and to defer to multiple experts (LECODU) [95]. For a fair comparison, all classification for the {SE,ME}HAI-CC methods have the same backbone, and all hyper-parameters are set as previously reported in [54, 78, 95]. To maintain fairness in the accuracy - coverage comparisons, we include the coverage constraint in Eq. (2) into all {SE,ME}HAI-CC methods. In particular, we set the hyper-parameter ϵ , which controls the coverage lower bound, to {0, 0.2, 0.4, 0.6, 0.8} so we can train all methods and plot their corresponding performance using the coverage - accuracy curves.

4.2 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. Additionally, we provide a concise quantitative analysis of Fig. 3 results in Table 1, which shows the *area under the accuracy-coverage curve* (AUACC), where higher AUACC values denote superior accuracy-coverage trade-offs.

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

Image	Human prediction		$f_\theta(\cdot)$	Output of gating model $g_\phi(\cdot)$ [AI, Ex.1, Ex.2, AI + Ex.1, AI + Ex.2] ^T	Final prediction	GT
	m_1	m_2				
	1	0	1	[0.40, 0.00, 0.01, 0.00, 0.59] ^T	0	0
	0	1	1	[0.20, 0.00, 0.01, 0.79 , 0.00] ^T	0	0
	1	1	0	[0.05, 0.00, 0.94 , 0.00, 0.01] ^T	1	1
	1	0	0	[0.02, 0.98 , 0.00, 0.00, 0.00] ^T	1	1
	0	0	1	[0.80 , 0.00, 0.00, 0.00, 0.20] ^T	1	1

Table 2: Examples of the CL2DC classification evaluated on the testing samples of the Galaxy Zoo dataset at 40% coverage.

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 Appendix B of the supplementary material. CL2DC excels by effectively identifying and cooperating with specific experts for their relevant tasks, thereby optimizing decision-making. In the CIFAR-100 dataset, LECODU achieves higher accuracy than SEHAI-CC at low and intermediate coverage levels but it shows lower accuracy than our approach. The performance of MultiL2D is comparable to SEHAI-CC, except at 0% coverage, when MultiL2D is better than SEHAI-CC methods because MultiL2D can successfully identify the best labeller, 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 specifying 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.

Table 2 shows a few examples of the inference of CL2DC at a coverage rate of 40% on test images of Galaxyzoo. Each example includes the test image, the human-provided labels (\mathcal{M}), AI model prediction ($f_\theta(\cdot)$), complementary module prediction $h_\psi(\cdot)$, prediction probability vector by the gating model ($g_\phi(\cdot)$), final prediction of CL2DC, and ground truth (GT) label. Notably, when the AI model or the human experts make individual mistakes, the final prediction tends to be correct, highlighting system robustness. When the AI model is correct, $g_\phi^{\text{AI}}(\cdot)$ tends to have a high probability suggesting that the AI

model can be trusted. When the L2D options are selected, that usually happens with very high probability for one of the options in $g_{\phi}^{\text{L2D}_j}(\cdot)|_{j=1}^M$ and quite low value for $g_{\phi}^{\text{AI}}(\cdot)$, suggesting a complete lack of trust in the AI model. On the other hand, when one of the L2C options are selected, notice that both $g_{\phi}^{\text{AI}}(\cdot)$ and one of the options in $g_{\phi}^{\text{L2C}_j}(\cdot)|_{j=1}^M$ show high values, indicating that the AI model can be partially trusted.

4.3 Ablation studies

In Fig. 4a, we study the hyper-parameter λ from Eq. (5) on the accuracy-coverage performance of CL2DC in CIFAR-100. The graphs illustrate a clear trend: when $\lambda = 1$, the accuracy is distinctively low at lower coverage values because of the relatively high weight for the constraint defined in Eq. (4), but when we decrease λ , there is an improvement in accuracy for almost all coverage values.

Next, we investigate the effect of altering the number and quality of experts in the experimental setting. Focusing on the “Chaoyang2u” setup, the outcomes, displayed in Fig. 4b and Table 1, show that our solution outperforms other HAI-CC methods more distinctly when compared with the original results in Fig. 3d that use all three pathologists for the “Chaoyang3u” setup.

We then study the influence L2D and L2C in our CL2DC method on Galaxyzoo and Micebone datasets in Fig. 4d and Fig. 4e. In Fig. 4d, CL2DC w/o L2D outperforms CL2DC w/o L2C at large coverage values, which means 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 AI model prediction. CL2DC outperforms CL2DC w/o L2C and CL2DC w/o L2D at all coverage values, showing the advantage of integrating both L2D and L2C into HAI-CC. In Fig. 4e, 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 L2D and L2C.

We also study the scalability of CL2DC when increasing the number of experts on CIFAR-100. We generate other seven synthetic experts, who have similar accuracy rates as described in Appendix B, i.e., each expert performs correctly on random 6/7 super-classes, while making 50% mistakes via instance-dependent noise on the remaining super-classes. Fig. 4c shows the accuracy-coverage curves with our methods, where the number of available experts increases from 2 to 6. In Fig. 4f, we evaluate the AUACC as the number of available experts increases. A significant improvement is observed when the number of experts increases from 2 to 3. This is because, in our setting, the first three expert high-accuracy sets of super-classes cover all super-classes, whereas the first two sets cover only about two-thirds of the super-classes. Consequently, it is unsurprising that the AUACC tends to converge as the number of experts continues to increase, indicating that additional redundant experts do not contribute to integrating more effective information or improving predictions.

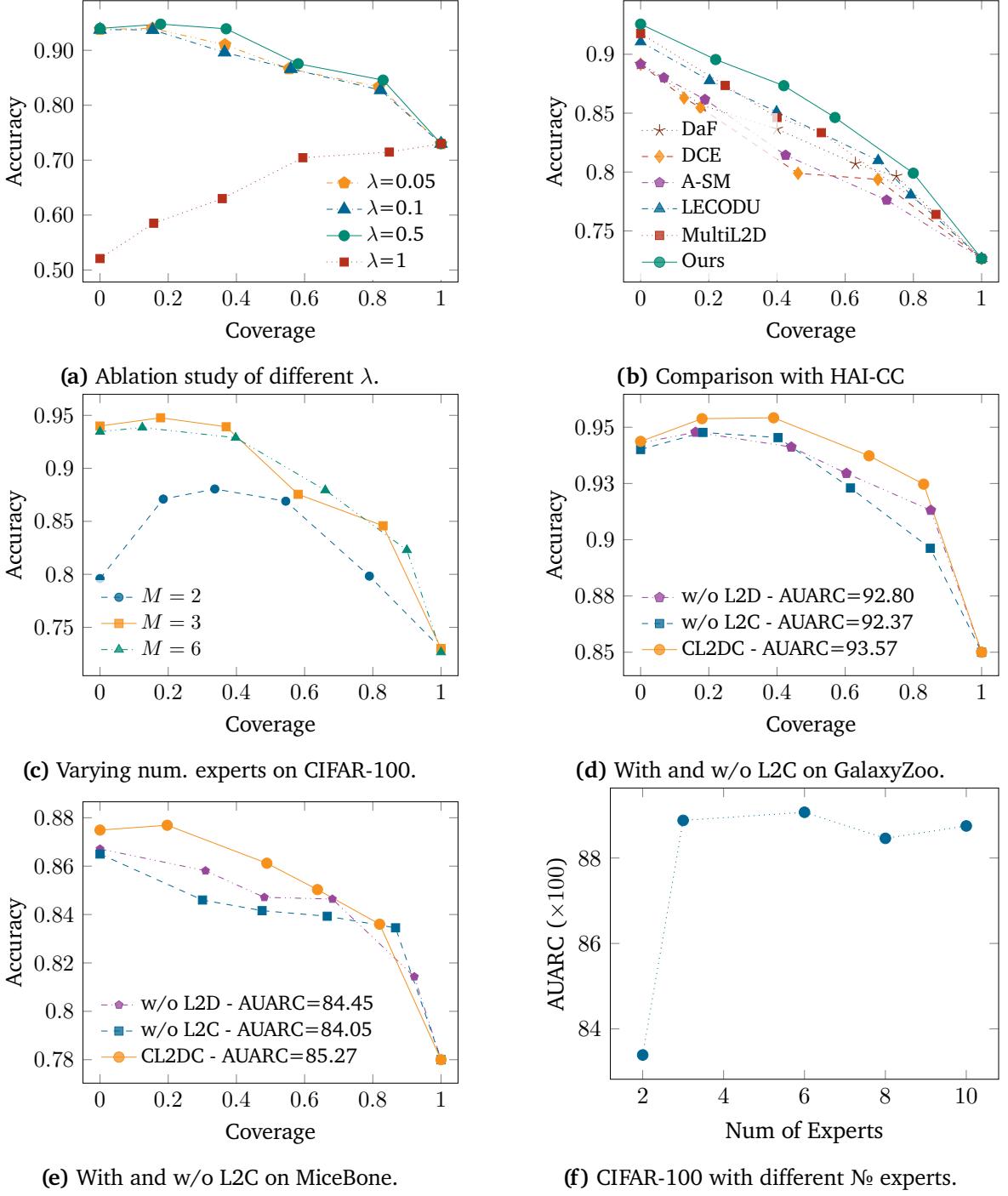


Figure 4: Accuracy-coverage curves in various evaluations: (a) different penalty coefficient λ on CIFAR-100 dataset, (b) comparison between our method and competing HAI-CC methods on Chaoyang dataset with 2 experts, (c) ablation study when varying the number of experts on CIFAR-100 dataset, (d) and (e) ablation study with and without L2C and L2C on Galaxyzoo and Micebone, respectively, and (f) evaluation of AUARC when varying the number of experts on CIFAR-100 dataset.

5 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, which consequently 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.

The proposed CL2DC has a limitation with the selection of multiple human experts. In a real-world decision-making system, multiple specific experts may be engaged in the decision process (e.g., clinical diagnosis). In future work, we will develop a new hybrid intelligent system to select a sequence of specific human experts to make decision collaboratively. Another potential issue of CL2DC is the fact that we need roughly balanced training sets to avoid overfitting to majority classes. We plan to address this problem by leveraging learning methods that are robust to imbalanced distributions.

References

- [1] Jean V Alves, Diogo Leitão, Sérgio Jesus, Marco OP Sampaio, Javier Liébana, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. Cost-sensitive learning to defer to multiple experts with workload constraints. *Transactions on Machine Learning Research*, 2024. [2](#), [4](#)
- [2] Eric Arazo, Diego Ortego, Paul Albert, Noel O’Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pages 312–321, 2019. [5](#)
- [3] Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-AI teams. In *International Joint Conference on Artificial Intelligence*, 2022. [4](#)
- [4] Steven P Bamford, Robert C Nichol, Ivan K Baldry, Kate Land, Chris J Lintott, Kevin Schawinski, Anže Slosar, Alexander S Szalay, Daniel Thomas, Mehri Torki, Dan Andreescu, Edward M. Edmondson, Christopher J. Miller, Phil Murray, M. Jordan Raddick, and Jan Vandenberg. Galaxy Zoo: The dependence of morphology and colour on environment. *Monthly Notices of the Royal Astronomical Society*, 393(4):1324–1352, 2009. [3](#), [7](#), [22](#)
- [5] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate AI the best teammate? Optimizing AI for teamwork. In *AAAI Conference on Artificial Intelligence*, pages 11405–11414, 2021. [4](#)
- [6] Björn Barz and Joachim Denzler. Do we train on test data? Purging CIFAR of near-duplicates. *Journal of Imaging*, 6(6):41, 2020. [3](#), [7](#), [21](#), [22](#)
- [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023. [1](#)

- [8] Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. In *Advances in Neural Information Processing Systems*, 2024. [3](#), [4](#), [8](#), [9](#)
- [9] Zhi Cao, Enhong Chen, Ye Huang, Shuanghong Shen, and Zhenya Huang. Learning from crowds with annotation reliability. In *International Conference on Research and Development in Information Retrieval*, pages 2103–2107, 2023. [5](#)
- [10] G. Carneiro. *Machine Learning with Noisy Labels: Definitions, Theory, Techniques and Solutions*. Elsevier Science, 2024. [4](#)
- [11] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pages 1507–1517. PMLR, 2021. [4](#)
- [12] Mohammad-Amin Charusaie and Samira Samadi. A unifying post-processing framework for multi-objective learn-to-defer problems. *Advances in Neural Information Processing Systems*, 2024. [2](#)
- [13] Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, pages 2972–3005. PMLR, 2022. [4](#)
- [14] Mohammad-Amin Charusaie, Amirmehdi Jafari Fesharaki, and Samira Samadi. Defer-and-fusion: Optimal predictors that incorporate human decisions. In *5th Workshop on practical ML for limited/low resource settings*, 2024. [2](#), [3](#), [4](#), [8](#), [9](#)
- [15] Jian Chen, Ruiyi Zhang, Tong Yu, Rohan Sharma, Zhiqiang Xu, Tong Sun, and Changyou Chen. Label-retrieval-augmented diffusion models for learning from noisy labels. In *Advances in Neural Information Processing Systems*, 2024. [5](#)
- [16] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *AAAI Conference on Artificial Intelligence*, 2021. [5](#)
- [17] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016. [4](#)
- [18] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021. [1](#), [2](#), [4](#)
- [19] Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *The 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018. [1](#)
- [20] Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations*, 2024. [1](#)
- [21] Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. Instance-dependent noisy label learning via graphical modelling. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2288–2298, 2023. [5](#)

- [22] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence*, 2017. 4
- [23] Hui Wen Goh, Ulyana Tkachenko, and Jonas Mueller. CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. In *NeurIPS Human in the Loop Learning Workshop*, 2022. 5
- [24] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Conference on Fairness, Accountability, and Transparency*, pages 90–99, 2019. 1
- [25] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*, 2018. 5
- [26] Mark D Halling-Brown, Lucy M Warren, Dominic Ward, Emma Lewis, Alistair Mackenzie, Matthew G Wallis, Louise S Wilkinson, Rosalind M Given-Wilson, Rita McAvinchey, and Kenneth C Young. OPTIMAM mammography image database: a large-scale resource of mammography images and clinical data. *Radiology: Artificial Intelligence*, 3(1):e200103, 2020. 1
- [27] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018. 4
- [28] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming effective human-AI teams: Building machine learning models that complement the capabilities of multiple experts. In *International Joint Conference on Artificial Intelligence*, pages 2478–2484, 2022. 2, 4, 8, 23
- [29] Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. Learning to defer with limited expert predictions. In *AAAI Conference on Artificial Intelligence*, pages 6002–6011, 2023. 4, 8
- [30] Marek Herde, Lukas Lührs, Denis Huseljic, and Bernhard Sick. Annot-Mix: Learning with noisy class labels from multiple annotators via a mixup extension. In *European Conference on Artificial Intelligence*, 2024. 5
- [31] Lee Jaehwan, Yoo Donggeun, and Kim Hyo-Eun. Photometric transformer networks and label adjustment for breast density prediction. In *International Conference on Computer Vision Workshops*, 2019. 4
- [32] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021. 5
- [33] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2018. 4
- [34] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. In *Advances in Neural Information Processing Systems*, pages 4421–4434, 2021. 4

- [35] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165, 2021. [2](#), [4](#)
- [36] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018. [5](#)
- [37] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. [7](#), [21](#), [22](#)
- [38] Diogo Leitão, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. Human-AI collaboration in decision-making: Beyond learning to defer. In *ICML Workshop on Human-Machine Collaboration and Teaming*, 2022. [4](#)
- [39] Junnan Li, Richard Socher, and Steven CH Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. [4](#)
- [40] Minghao Liu, Jiaheng Wei, Yang Liu, and James Davis. Do humans and machines have the same eyes? Human-machine perceptual differences on image classification. *arXiv preprint arXiv:2304.08733*, 2023. [4](#)
- [41] Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Mitigating underfitting in learning to defer with consistent losses. In *International Conference on Artificial Intelligence and Statistics*, 2024. [4](#)
- [42] Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, pages 21475–21496. PMLR, 2023. [5](#)
- [43] Thodoris Lykouris and Wentao Weng. Learning to defer in content moderation: The human-ai interplay. *arXiv preprint arXiv:2402.12237*, 2024. [1](#)
- [44] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, 2018. [4](#)
- [45] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, Alexander Ding, Greg S. Corrado, Daniel Tse, and Shravya Shetty. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020. [3](#), [7](#), [23](#)
- [46] Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, 2023. [4](#)
- [47] Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *Advances in neural information processing systems*, 2024. [2](#)
- [48] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Principled approaches for learning to defer with multiple experts. In *International Symposium on Artificial Intelligence and Mathematics*, 2024. [4](#)
- [49] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable h -consistent and Bayes-consistent loss functions for learning to defer. In *Advances in neural information processing systems*, 2024. [2](#)
- [50] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Regression with multi-expert deferral. In *International Conference on Machine Learning*, 2024. [2](#), [4](#)

- [51] Zahra Mirikhrajji, Kumar Abhishek, Saeed Izadi, and Ghassan Hamarneh. D-LEMA: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1837–1846, 2021. 5
- [52] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020. 2, 4
- [53] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. Teaching humans when to defer to a classifier via exemplars. In *AAAI Conference on Artificial Intelligence*, pages 5323–5331, 2022.
- [54] Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? Exact algorithms for learning to defer to humans. In *International Conference on Artificial Intelligence and Statistics*, pages 10520–10545. PMLR, 2023. 2, 4, 8, 9, 23
- [55] Hussein Mozannar, Jimin Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Effective human-ai teams via learned natural language rules and onboarding. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [56] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, pages 65–81. PMLR, 2009. 9
- [57] Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, pages 29292–29304, 2022. 2, 4
- [58] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999. 7
- [59] Nastaran Okati, Abir De, and Manuel Rodriguez. Differentiable learning under triage. In *Advances in Neural Information Processing Systems*, pages 9140–9151, 2021. 2, 4
- [60] Diego Ortego, Eric Arazo, Paul Albert, Noel E O’Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6606–6615, 2021. 5
- [61] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019. 23
- [62] Clara Punzi, Roberto Pellungrini, Mattia Setzu, Fosca Giannotti, and Dino Pedreschi. AI, meet human: Learning paradigms for hybrid decision making systems. *arXiv preprint arXiv:2402.06287*, 2024. 1
- [63] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. In *Machine Learning for Health Symposium*, 2018. 2, 4

- [64] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *International Conference on Machine Learning*, pages 889–896, 2009. 5
- [65] Mengye Ren, Wenyuan Zeng, Binh Yang, and R. Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, 2018. 5
- [66] Lars Schmarje, Claudius Zelenka, Ulf Geisen, Claus-C Glüer, and Reinhard Koch. 2D and 3D segmentation of uncertain local collagen fiber orientations in shg microscopy. In *German Conference on Pattern Recognition*, pages 374–386. Springer, 2019. 3, 7, 23
- [67] Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, and Reinhard Koch. Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215–33232, 2022.
- [68] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, Claudius Zelenka, Rainer Kiko, Jenny Stracke, Nina Volkmann, and Reinhard Koch. A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. In *European Conference on Computer Vision*, pages 363–380. Springer, 2022. 3, 7, 23
- [69] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 4
- [70] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–AI complementarity. *National Academy of Sciences*, 119(11):e2111547119, 2022. 1, 2, 4
- [71] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pages 32633–32653. PMLR, 2023. 4
- [72] Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In *International Conference on Artificial Intelligence and Statistics*, 2024. 2, 4
- [73] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019. 5
- [74] Shahroz Tariq, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. A2C: A modular multi-stage collaborative decision framework for human-AI teams. *arXiv preprint arXiv:2401.14432*, 2024. 2, 4
- [75] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 3, 7, 22
- [76] Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pages 22184–22202. PMLR, 2022. 2, 4

- [77] Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. On the calibration of learning to defer to multiple experts. In *ICML Workshop on Human-Machine Collaboration and Teaming*, 2022. 2, 4
- [78] Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 11415–11434. PMLR, 2023. 2, 3, 4, 8, 9, 22
- [79] Haobo Wang, Ruixuan Xiao, Yiwen Dong, Lei Feng, and Junbo Zhao. ProMix: combating label noise via maximizing clean sample utility. In *International Joint Conference on Artificial Intelligence*, 2023. 5, 23
- [80] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale chestX-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017. 3, 7, 23
- [81] Hongxin Wei, RENCHUNZI Xie, Lei Feng, Bo Han, and Bo An. Deep learning from multiple noisy annotators as a union. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 5
- [82] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2021. 3
- [83] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification. 1
- [84] Tong Wei, Hao-Tian Li, Chun-Shu Li, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. Vision-language models are strong noisy label detectors. In *Advances in Neural Information Processing Systems*, 2024. 5
- [85] Zixi Wei, Yuzhou Cao, and Lei Feng. Exploiting human-AI dependence for learning to defer. In *International Conference on Machine Learning*, 2024. 3, 4, 8, 9
- [86] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *International Joint Conference on Artificial Intelligence*, 2021. 2, 4
- [87] Junde Wu, Huihui Fang, Zhaowei Wang, Dalu Yang, Yehui Yang, Fangxin Shang, Wenshuo Zhou, and Yanwu Xu. Learning self-calibrated optic disc and cup segmentation from multi-rater annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 614–624. Springer, 2022. 5
- [88] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *International Conference on Learning Representations*, 2022. 21
- [89] Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. Faster meta update strategy for noise-robust deep learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 5
- [90] Bodi Yuan, Jianyu Chen, Weidong Zhang, Hung-Shuo Tai, and Sara McMains. Iterative cross learning on noisy labels. In *Winter Conference on Applications of Computer Vision*, pages 757–765, 2018. 4

- [91] Zizhao Zhang and Tomas Pfister. Learning fast sample re-weighting without reward data. In *International Conference on Computer Vision*, 2021. 5
- [92] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, pages 8778–8788, 2018. 4
- [93] Zizhao Zhang, Han Zhang, SercanÖ. Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9291–9300, 2020. 5
- [94] Zheng Zhang, Kevin Wells, and Gustavo Carneiro. Learning to complement with multiple humans (LECOMH): Integrating multi-rater and noisy-label learning into human-AI collaboration. *arXiv preprint arXiv:2311.13172*, 2023. 2, 4
- [95] Zheng Zhang, Wenjie Ai, Kevin Wells, David Rosewarne, Thanh-Toan Do, and Gustavo Carneiro. Learning to complement and to defer to multiple users. In *European Conference on Computer Vision*, pages 144–162. Springer, 2025. 2, 3, 4, 5, 8, 9
- [96] Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE Transactions on Medical Imaging*, 41(4):881–894, 2021. 3, 7, 22

A Training and testing algorithms

We provide the detailed algorithms used in the training and testing of CL2DC in Algorithms 1 and 2, respectively.

Algorithm 1 The training procedure of the proposed method

```

1: procedure CL2DC TRAINING( $\mathcal{D}, \varepsilon, K, \lambda$ )
2:    $\triangleright \mathcal{D} = \{(\mathbf{x}_i, \mathcal{M}_i)\}_{i=1}^N$ : training dataset consists of samples and annotations of  $M$   $\triangleleft$ 
   human experts
3:    $\triangleright \varepsilon$ : the lower bound of the targeted coverage value  $\triangleleft$ 
4:    $\triangleright K$ : the number of iterations  $\triangleleft$ 
5:    $\triangleright \lambda$ : a hyper-parameter used in the penalty function  $\triangleleft$ 
6:   initialise the parameters of AI classifier, gating and complement models:  $\theta, \phi$  and  $\psi$ 
7:   initialise penalty hyper-parameter:  $\beta_0 > 0$ 
8:   obtain consensus labels:  $(\hat{\mathbf{y}}_i)_{i=1}^N \leftarrow \text{GET CONSENSUS LABELS}(\mathcal{D}, \theta)$ 
9:   for  $k = 1 : K$  do
10:    initialise  $p_{\text{AI}} \leftarrow 0$   $\triangleright$  to accumulate “coverage”
11:    initialise  $L \leftarrow 0$   $\triangleright$  to accumulate the training loss
12:    for each sample indexed by  $i$  in  $\mathcal{D}$  do
13:      calculate the probability of selection:  $g_\phi(\mathbf{x}_i)$ 
14:      calculate the losses  $\ell(\mathbf{x}_i, \hat{\mathbf{y}}_i, \mathcal{M}_i, \theta, \psi)$  defined in Eq. (3)
15:       $L \leftarrow L + g_\phi(\mathbf{x}_i)^\top \ell(\mathbf{x}_i, \hat{\mathbf{y}}_i, \mathcal{M}_i, \theta, \psi)$   $\triangleright$  accumulate the loss on sample  $i$ -th
16:       $p_{\text{AI}} \leftarrow p_{\text{AI}} + g_\phi^{(\text{AI})}(\mathbf{x}_i)$   $\triangleright$  accumulate coverage
17:    calculate the penalty of the constraint:  $c(\phi, \varepsilon) \leftarrow \max(0, \varepsilon - p_{\text{AI}}/N)^2$   $\triangleright$  defined in Eq. (4)
18:     $\beta_k \leftarrow \lambda(\beta_{k-1} + k)$   $\triangleright$  hyper-parameter associated with the penalty function
19:     $L \leftarrow L/N + \beta_k c(\phi, \varepsilon)$   $\triangleright$  training loss
20:    update:  $\theta, \phi, \psi \leftarrow \text{SGD}(L)$ 
21:   return  $\theta, \phi, \psi$ 

22: procedure GET CONSENSUS LABELS( $\mathcal{D}, \theta$ )
23:    $\triangleright \mathcal{D} = \{(\mathbf{x}_i, \mathcal{M}_i = \{m_i^{(j)}\}_{j=1}^M)\}_{i=1}^N$ : the multi-rater noisy label dataset  $\triangleleft$ 
24:    $\triangleright \theta$ : the parameter of a classifier  $\triangleleft$ 
25:   get the majority vote label:  $\bar{m}_i \leftarrow \text{MAJORITY VOTE}(\{m_i^{(j)}\}_{j=1}^M), i = 1 : |\mathcal{D}|$ 
26:   train classifier on majority vote:  $\theta \leftarrow \text{TRAIN CLASSIFIER}(\{(\mathbf{x}_i, \bar{m}_i)\}_{i=1}^{|\mathcal{D}|}, \theta)$ 
27:   obtain consensus labels:  $(\hat{\mathbf{y}}_i)_{i=1}^{|\mathcal{D}|} \leftarrow \text{CROWDLAB}(\{(f_\theta(\mathbf{x}_i), \mathcal{M}_i)\}_{i=1}^{|\mathcal{D}|})$   $\triangleright$  defined in Eq. (1)
28:   return  $(\hat{\mathbf{y}}_i)_{i=1}^{|\mathcal{D}|}$ 

```

B Datasets

CIFAR-100 [6, 37] has 50k training images and 10k testing images, with each image belonging to one of 100 classes categorised into 20 super-classes. We follow the instance-dependent label noise [88] to generate synthetic labels representing a synthetic expert.

Algorithm 2 The proposed testing procedure

```

1: procedure CL2DC TESTING( $(\mathbf{x}_{N+1}, (m_{N+1}^{(j)})_{j=1}^M, \theta, \phi, \psi, \varepsilon)$ )
2:    $\triangleright \mathbf{x}_{N+1}$ : a testing sample
3:    $\triangleright m_{N+1}^{(j)}$ : the annotations of the testing samples made by expert indexed by  $j$ 
4:    $\triangleright \theta$ : parameter of AI classifier
5:    $\triangleright \phi$ : parameter of gating model
6:    $\triangleright \psi$ : parameter of complementary module
7:    $\mathbf{p} \leftarrow g(\mathbf{x}_{N+1}; \phi)$   $\triangleright$  probability of selection
8:    $\mu \leftarrow \text{argmax}_\mu \mathbf{p}_\mu$ 
9:   if ( $\mu = 1$ ) then  $\triangleright$  AI predict alone
10:    |  $\tilde{y}_{N+1} \leftarrow f_\theta(\mathbf{x}_{N+1})$ 
11:   else if  $2 \leq \mu \leq M + 1$  then  $\triangleright$  learning to defer option
12:    |  $\tilde{y}_{N+1} \leftarrow m_{N+1}^{(\mu-1)}$ 
13:   else  $\triangleright$  learning to complement option
14:    |  $\tilde{y}_{N+1} \leftarrow h_\psi(f_\theta(\mathbf{x}_{N+1}), m_{N+1}^{(\mu-M-1)})$ 
15:   return  $\tilde{y}_{N+1}$ 

```

In particular, each expert performs 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, where labels can be randomly flipped to other classes within the same super-class. In the experiments, we evaluate ours and competing methods using three synthetic experts. In addition, because about 10% of testing images in CIFAR-100 [37] are duplicated or almost identical to the ones in the training set, in our training and testing, we use ciFAIR-100 [6], which replaces those duplicated images by different images belonging to the same class.

HAM10000 [75] has about 10k training and 1,500 testing dermatoscopic images categorised into seven types of human skin lesions. These seven categories can be grouped further into two super-classes: *benign* and *malignant*. We follow the setting presented in Multi-L2D [78] to simulate two experts based on these two super-classes, each following an asymmetric label noise, similarly to CIFAR-100. In particular, the accuracy of the two experts is around 90%, where the first expert makes 5% and 15% of labelling mistakes on super-classes *malignant* and *benign*, respectively, while the second expert only makes 15% and 5% of labelling mistakes on the super-classes *malignant* and *benign*, respectively.

Galaxy Zoo [4] consists of 60k images of galaxies and the corresponding answers from hundreds of thousands of volunteers to classify their shapes. We follow the setup in Multi-L2D [78] that uses the response to the first question "*Is the object a smooth galaxy or a galaxy with features/disk?*" as the ground truth labels of a binary classification to simulate two synthetic experts. In particular, the first expert makes 5% and 15% of labelling mistakes on *smooth galaxy* and *galaxy with features/disk*, respectively, while the second expert makes 15% and 5% of labelling mistakes on *galaxy with features/disk* and *smooth galaxy*, respectively.

Chaoyang [96] comprises 6,160 colon slide patches categorised into four classes: *normal*, *serrated*, *adenocarcinoma*, and *adenoma*, where each patch has *three noisy labels annotated by three pathologists*. In the original Chaoyang dataset setup, the training set has patches with multi-rater noisy labels, while the testing set only contains patches that all experts agree on a single label. We have restructured the dataset to ensure that 1) *both*

training and testing sets contain multiple noisy labels; 2) experts have similar performance in training and testing sets; and 3) patches from the same slide do not appear in both the training and testing sets. This setting results in a partition of 4,533 patches for training and 1,627 patches for testing, and the accuracy of the three experts are 91%, 88%, 99%, assuming that the majority vote forms the ground truth annotation.

MiceBone [66–68] has 7,240 second-harmonic generation microscopy images, with each image being annotated by one to five professional annotators, where the annotation consists of one of three possible classes: *similar collagen fiber orientation, dissimilar collagen fiber orientation, and not of interest due to noise or background*. Only 8 out of 79 annotators label the whole dataset. We, therefore, use these 8 annotators to represent the experts in our experiment. Using the majority vote as the ground truth, the accuracy of those 8 experts are from 84% to 86%. As the dataset is divided into 5 folds, we use the first 4 folds as the training set, and the remaining fold as the test set.

NIH-ChestXray [45, 80] contains an average of 3 manual labels per image for four radiographic findings on 4,374 chest X-ray images [45] from the ChestXray8 dataset [80]. We focus on the classification of airspace opacity (NIH-AO) because only this finding’s prevalence is close to 50%, without heavy class-imbalance problem. Following [28, 54], a total of 2,412 images is for training and 1,962 images are for testing. The prediction accuracy of the 3 experts in the NIH-AO dataset is approximately 89%, 94%, 80% both in training and testing.

C Implementation Details

C.1 Architecture

All methods are implemented in PyTorch [61] and run on Nvidia RTX A6000. For experiments performed on CIFAR-100 dataset, we employ ProMix [79] to train the AI model formed by two PreAct-ResNet-18 as the LNL AI models. For Chaoyang, we use a ResNet-34 for the AI model, and for other datasets, we train the AI model with a ResNet-18 using a regular CE loss minimisation with a ground truth label formed by the majority-voting of experts. The gating model uses the same trained backbones as the ones used for the AI model. The complementary module is represented by a two-layer multi-layer perceptron (MLP), where each hidden layer has 512 nodes activated by Rectified Linear Units (ReLU). On CIFAR-100 the AI model achieves 73.42% accuracy on the testing set. The AI models on Chaoyang, NIH-AO, Micebone, HAM10000, and Galaxy-zoo datasets achieve 72.65%, 85.37%, 78.12%, 78.06%, and 85.24%, respectively.

C.2 Training

For each dataset, the proposed human-AI system is trained for 200 epochs using SGD with a momentum of 0.9 and a weight decay of 5×10^{-4} . The batch size used is 256 for all datasets. The initial learning rate is set at 0.01 and decayed through a cosine annealing. For training the whole HAI-CC method, the ground truth labels are set as the consensus labels obtained via CROWDLAB. For testing, the ground truth label is either available from the dataset (e.g., CIFAR-100, HAM10000, Galaxy-zoo) or from majority voting (e.g., MiceBone, Chaoyang, NIH-ChestXray).