A Complementarity-Enhanced Mixture of Human-AI Teams for Decision-Making*

Hefei Liang[1], Jiaqi Liu[1] (🖂), Bin Guo[1], and Zhiwen Yu[2,1]

[1] School of Computer Science, Northwestern Polytechnical University, China
[2] Harbin Engineering University, China
`craneflyliang@mail.nwpu.edu.cn`
`{jqliu, zhiwenyu, guob}@nwpu.edu.cn`

Abstract. With the rapid development of deep learning, Artificial Intelligence (AI) has evolved from a mere tool to a collaborator in decision-making, sparking increasing attention to the human-AI cooperation. The Mixture of Experts (MoE) framework, originally proposed to capture domain-specific expertise and now widely adopted in large-scale models, naturally aligns with the requirements of human-AI teams. However, deploying MoE in human-AI cooperation involves two challenges: 1) While machine experts can be continuously optimized during training, human experts remain fixed, significantly reducing the effectiveness of traditional sparse activation strategies; 2) Some existing methods fuse all expert predictions during training phase but select only the highest weighted expert during testing phase, thereby introducing inconsistencies between the two phases. To overcome this, we propose the Complementarity-Enhanced Mixture of Human-AI Teams (CE-MoHAIT) framework. Our approach decomposes the gating network's output into two branches, i.e., a human expert branch and a classifier branch, thereby explicitly modeling the complementarity between human and AI capabilities. Moreover, we introduce a method called Adaptive and Complementary Construction (ACC) that directly optimizes the gating network by constructing weighted labels, enabling the classifier model to compensate for the deficiencies of human experts and ensuring consistent task allocation across training and testing. Experiments on CIFAR-100 and two real-world medical image datasets show that our approach surpasses the existing methods, improving test accuracy by up to 20%, especially with larger teams and weaker experts. Code is available in the repository at `https://github.com/H-F-Liang/CE-MoHAIT`.

Keywords: Human-AI Collaboration in Classification · Human-AI Teams · Human-in-the-Loop.

1 Introduction

With the rapid development of advanced deep learning technologies [2, 3], AI has already achieved or even surpassed human-level performance in many specific domains [1]. However, in some real-world scenarios, especially in high-risk domains

* 🖂 Corresponding author: Jiaqi Liu (e-mail: jqliu@nwpu.edu.cn).

such as healthcare [4, 5], AI still exhibits inherent shortcomings such as limited generalization ability, noise immunity, and interpretability, which underscores the growing need for human-AI collaboration. In human-AI collaboration, how to leverage the complementary capabilities of humans and AI, and assign tasks to the appropriate human or machine based on task characteristics is important [6–8]. Mixture of Experts (MoE) [29] is a framework initially proposed and widely adopted in Large Language Model (LLM), characterized by its ability to perceive knowledge differences across domains or expertise among individuals. This makes MoE potentially well-suited for human-AI collaboration scenarios.

The MoE framework consists of multiple multilayer perceptrons with different knowledge, called machine experts, and a gating network that decides which machine expert to activate according to the task's characteristics. In the training phase, MoE adopts a sparse activation strategy due to the large scale of the model parameters [30], where only a few, or even just one, machine expert is activated during each forward pass. However, When applying MoE to human-AI collaboration scenarios, some machine experts are replaced by human experts and therefore the sparse activation strategy can lead to difficulties in convergence. This is because that the abilities of human experts are fixed, and thus they cannot learn the corresponding knowledge even if activated during training. Moreover, due to the lack of targeted optimization for the gating network, it is challenging for the gating network to leverage the complementary capabilities between humans and machines.

Hemmer et al. [13] proposed a solution that adopts a weighted aggregation of predictions from all experts during the training phase and selects only one expert during the testing phase. However, this weighted aggregation approach introduces inconsistency: during the training phase, multiple predictions are fused by their weights to compute loss, whereas during the testing phase, only the highest-weighted prediction is used, potentially degrading performance and limiting the generalizability of the method. To overcome this limitation, we propose a novel Adaptive and Complementary Construction (ACC) method. It directly optimizes the output of the gating network in MoE by constructing weighted labels, rather than aggregating expert predictions. Specifically, the method directly constructs labels corresponding to the gating network's output weights and allows the machine expert to complement the deficiencies of human experts, significantly enhancing the complementarity between humans and AI models in the MoE framework.

Based on the MoE framework, we propose a Complementarity-Enhanced Mixture of Human-AI Teams (CE-MoHAIT) framework for human-AI collaboration in classification tasks. CE-MoHAIT jointly trains a gating network and one classifier model, i.e., the machine expert, where the gating network determines the weight of each team member, and the task is ultimately assigned to the member with the highest weight. To explicitly consider the complementarity between humans and AI models, we decompose the output weights of the gating network into two branches: a human expert weight branch and a classifier weight branch. The classifier model is designed to complement the deficiencies of human experts. During the training phase, we use the ACC to encourage the gating network to assign tasks that human experts are less competent to the classifier model for learning, thereby maximizing team complementarity. Furthermore, since the task is always assigned to the member with the highest weight during both the training and testing phases, the behavioral consistency between training and testing enhances the performance of the team. Overall, our contributions are as follows:

- We propose a novel framework called Complementarity-Enhanced Mixture of Human-AI Teams (CE-MoHAIT), which explicitly considers the complementarity between humans and AI models by decomposing the output weights of the gating network into a human expert weight branch and a classifier weight branch.
- We introduce a new team loss function that utilizes Adaptive and Complementary Construction (ACC) to construct team weight labels, optimizing two weight branches. This approach enables the classifier to better learn from and complement for the weaknesses of human experts, explicitly maximizing human-AI complementarity and ensuring optimal team performance.
- We conduct comprehensive experiments on the classical CIFAR-100 dataset and two real-world medical image datasets [14, 15], demonstrating the effectiveness and applicability of our method. Our method achieves up to a 20% performance improvement, especially in scenarios where human experts possess lower individual capabilities.

2 Related work

Traditional deep learning methods mainly focus on optimizing AI systems in isolation, rather than considering human-AI collaboration. As a result, recent research [23–25] has proposed various approaches to coordinate humans and AI models, offering advantages in efficiency, interpretability, and ethical considerations.

Currently, most human-AI collaboration methods fall under the broader concept of Human-in-the-Loop (HITL) [32], which emphasizes continuous human participation and feedback throughout the operation of the system. The core idea of HITL is to integrate expert feedback with data-driven learning strategies to compensate for the limitations of traditional automation in data-scarce or noisy environments, while also providing additional prior knowledge during model optimization. For example, some researchers have embedded human preferences into the reward function [33], achieving significant improvements in policy optimization in deep reinforcement learning. Based on this, many other optimization methods based on human preferences have also been derived [34–37]. In high-risk domains such as finance and healthcare, the complete autonomous decision-making by AI models often lacks transparency and interpretability [39]. Selectively delegating decision-making power to humans can also enhance trust and the interpretability of human-AI collaboration [26–28].

One common approach in this domain is Learning to Defer, which assumes that humans are highly capable but costly decision-makers [9–12]. Its objective is to learn how to defer tasks that the AI model finds uncertain or challenging to human experts. For instance, some work [18] focuses on scenarios with cost constraints and limited expert workload. Other studies emphasize combining the predictions of AI models with those of human experts. For example, Steyvers et al. [19] introduced a Bayesian framework that attempts to integrate human and AI predictions to improve the overall performance of human-AI systems. Many existing Learning to Defer studies focus on single-expert settings, where difficult samples are deferred to one expert. However, this approach can impose a significant workload on the human expert and is impractical in real-world scenarios where a single expert cannot handle all tasks. Consequently, some work has explored deferring challenging samples to multiple experts. For example, Verma et al. [20] built on earlier research by Verma and Nalisnick [21] and refined the softmax surrogate loss introduced by Mozannar and Sontag [22] to propose consistent and calibrated surrogate losses for multi-expert settings. More recently, Zhang et al. [12] combined learning to complement with learning to defer, achieving superior performance in multi-expert settings with noisy labels compared to standalone human experts or AI models.

Another highly relevant yet often overlooked direction in this context is Mixture of Experts (MoE). Initially proposed for modularizing multilayer supervised networks, MoE can be seen as a system composed of multiple independent networks, where each network processes a subset of training data. A gating network is used to determine which model should learn from each data sample [29]. This process closely resembles task allocation in human-AI collaboration. Later, to enable the model to dynamically select a small number of machine experts for computation and reduce computational costs, introduced a sparse gating mechanism-based MoE structure has been proposed [30] , which remains the most mainstream MoE architecture to date. However, dynamically selecting a small number of experts may result in the gating network frequently favoring the more capable experts. To mitigate this issue, another work proposed an auxiliary loss [31] . Inspired by MoE, Hemmer, P., et al. [13] was the first to propose Human-AI Teams, consisting of one classifier and multiple human experts. This approach leverages the classifier to learn tasks that humans struggle with, thereby improving team performance. However, the strategy of using weighted averaging during training while adopting the *argmax* function during testing is not entirely appropriate, as it creates a significant gap between training and testing, negatively impacting generalization.

3 Methodology

In this section, we first present the problem formulation, and then illustrate the proposed Complementarity-Enhanced Mixture of Human-AI Teams (CE-MoHAIT) framework.

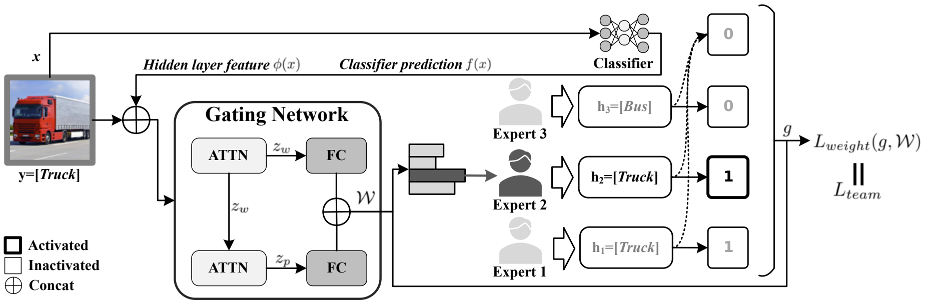


Fig. 1. The framework of the proposed CE-MoHAIT. The left part of the figure illustrates the structure of the gating network and how it generates the team weight $\mathcal{W}$. The right part of the figure demonstrates how we construct the weight label $\mathbf{g}$ to optimize the team weight $\mathcal{W}$.

3.1 Problem Formulation

A human-AI team for a k-class classification task consists of one classifier and m human experts. The classifier outputs a prediction denoted by $f(x) \in \mathbb{R}^k$, where $f(x)$ represents the predicted probability distribution over the k possible classes. Each human expert provides a prediction in the form of a k-dimensional one-hot vector $\mathbf{h} \in H$, indicating their chosen class. Given a training sample $x_i \in \mathcal{X}$ with ground truth label $y_i \in \mathcal{Y}$ and related human prediction $\mathbf{h}_i \in H$, the training dataset is presented as $D = \{(x_i, y_i, \mathbf{h}_i)\}_{i=1}^n \sim P$, where $n = |D|$ is the total number of samples and P is an unknown data distribution.

The final prediction is selected from these $m+1$ candidates through a gating network $g : \mathbb{R}^k \times \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}^{m+1}$, Its input consists of three components: the prediction of classifier $f(x) \in \mathbb{R}^k$, the d-dimensional feature vector extracted by the classifier $\phi(x) \in \mathbb{R}^d$, and the original data sample $x \in \mathcal{X}$, formally expressed as $g(f(x), \phi(x), x)$. The network outputs a $(1+m)$-dimensional vector $\mathcal{W} = [p, \mathbf{w}]$, where $p \in \mathbb{R}$ denotes the confidence in the prediction of classifier, and $\mathbf{w} = [w_1, \ldots, w_m] \in \mathbb{R}^m$ are the scores assigned to the human experts.

The final team prediction $\hat{y}_{\text{team}}$ is determined by selecting the team member with the highest weight. Our goal is to minimize the team loss defined as

$$L_{\text{team}}(f, g, x, y, \mathbf{h}) = \mathbb{E}_{(x,y,\mathbf{h})\sim P}\left[l\left(y, \hat{y}_{\text{team}}\right)\right]. \tag{1}$$

To minimize team loss, it is essential to consider the differences in capabilities among human-AI teams members, especially the complementary characteristics between humans and AI.

3.2 Implementation

The overall framework of the proposed framework is shown in Figure 1. The system can be considered as a collaborative decision-making team consisting of one classifier and m human experts. First, the sample x is input to the classifier, which outputs the extracted hidden layer features $\phi(x)$ and the predicted result $f(x)$. These outputs, along with the sample x, serve as inputs to the gating network. The function of the gating network is to coordinate task allocation. It outputs a weight vector for the team members $\mathcal{W}$, and based on $\mathcal{W}$, we select the most suitable member of the team to make the final decision.

In human-AI collaboration methods, expert predictions cannot directly compute loss for backpropagation. These methods typically obtained a weighted vector through fusion and calculated the loss between the weighted vector and the true labels for backpropagation. However, this approach often relies on predictions from multiple members, which can lead to dependency on certain members. Furthermore, during testing, if only the prediction from the member with the highest weight is selected, this results in an inconsistency between the training and testing phases. To address this issue, we propose an approach called Adaptive and Complementary Construction (ACC), which directly optimizes the member weights $\mathcal{W}$ and explicitly considers the complementarity between humans and AI, thereby improving the team performance.

Gating Network As shown in Figure 1, the gating network takes the feature vector as input and outputs the team member weights $\mathcal{W}$. Its primary function is to perceive the differences and complementarities between human and machines. To achieve this, we employ two attention layers to extract features from human experts and the classifier separately. We begin by extracting the topic features related to expert capability

$$z_w = \text{AttentionLayer}(x, f(x), \phi(x)), \quad z_w \in \mathbb{R}^{d_h}, \tag{2}$$

where d_h denotes the dimension of the hidden layer. Meanwhile, to explicitly leverage human-AI complementarity, we need to obtain a topic feature associated with human-AI complementarity

$$z_p = \text{AttentionLayer}(x, g(x), M(x), z_w), \quad z_p \in \mathbb{R}^{d_h}. \tag{3}$$

The topic feature z_p is used to represent the capability differences between the classifier and the human experts. In the attention layer, multi-dimensional features are mapped to Q, K, and V

$$Q = W_Q \cdot z, \quad K = W_K \cdot z, \quad V = W_V \cdot z, \tag{4}$$

where $Q, K, V \in \mathbb{R}^{h \times d_k}$, h is the number of attention heads, d_h is the hidden layer dimension, and $d_k = d_h/h$. Attention weights are then computed and used to obtain the weighted output

$$\mathbf{Z} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{5}$$

Finally, after passing through the linear layer, p and $\mathbf{w}$ are concatenated to obtain the output:

$$\begin{aligned} \mathbf{w} = \text{MLP}(z_w), \quad p = \text{MLP}(z_p), \\ \mathcal{W} = \text{Concat}(p, \mathbf{w}). \end{aligned} \tag{6}$$

Team Loss To leverage MoE for instance allocation in a human-AI collaboration team, the loss computation of MoE needs to be modified. In MoE, a sparse activation strategy is typically used, where only the selected networks are trained. However, in a human-AI collaboration team, the predictions provided by selected human experts are non-differentiable, indicating that the weights output by the gating network are difficult to optimize. Therefore, in [13], a weighted fusion strategy was adopted during training to obtain the final team prediction, ensuring that gradients propagate back to the gating network. Following this work, we have

$$\hat{y}_{team} = f(\mathbf{x}_i)w_1 + \sum_{i=2}^{1+m} w_i h_{i-1}. \tag{7}$$

However, this leads to a dependency of the prediction results of the team on multiple members. In real-world scenarios, due to considerations of collaboration efficiency and cost, we only select the best member to make the prediction. Thus, using the above strategy for training may affect the robustness and generalization of team predictions. In contrast, our method mitigates the inconsistency between training and testing behaviors and explicitly accounts for human-AI complementarity during the training phase by constructing weight labels.

First, in both training and testing phases, we select only the best-performing member. For each sample, we take the member with the highest weight, obtaining its predicted logits vector t_j where $j = \arg\max \mathcal{W}$, and compute the cross-entropy loss with the ground truth label y_i as follows

$$L_{\mathrm{CE}} = -\sum_{i=1}^{N}\sum_{j=1}^{k} y_j^{(i)} \log\left(\frac{\exp(t_j^{(i)})}{\sum_{j=1}^{k}\exp(t_j^{(i)})}\right). \tag{8}$$

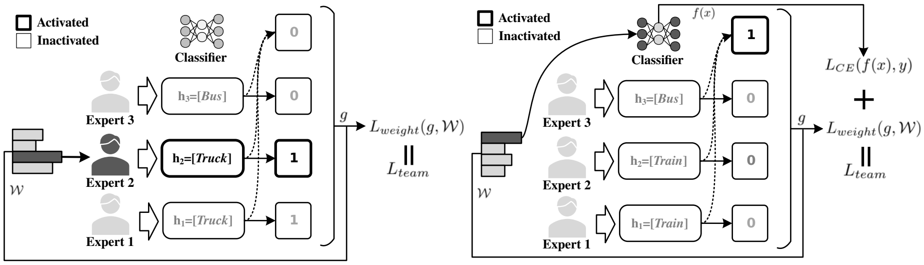
But it should be noted, that when the member represented by j is a human expert, the predicted label does not have a gradient; hence, the loss cannot be backpropagated. This loss works only when j corresponds to a classifier, meaning it is used exclusively for optimizing the classifier.

Next, we consider the team weight vector $\mathcal{W}$ output by the gating network. It consists of two branches, i.e., the human expert weighting branch $\mathbf{w}$ and the classifier weighting branch p, as in Equation (6). This addresses the issue that the predictions provided by human experts are *non-differentiable*, so we need a method to construct target labels for $\mathcal{W}$ during the training phase and this method is the Adaptive and Complementary Construction (ACC) that we mentioned earlier.

In the ACC method, for the human expert weight branch, during the training phase, we know the predictions of all experts, and therefore we can determine whether the prediction of human expert for the current sample is correct, denoted as

$$g_{ij} = 1\{h_{i,j} = y_i\}. \tag{9}$$

Then, for the classifier weight branch, we have illustrated two cases in Figure 2 for an intuitive explanation. To enhance the complementarity between the



*The case where the expert **can solve** the task*

*The case where the expert **cannot solve** the task*

Fig. 2. An intuitive explanation of constructing weight labels. We illustrate two different cases: the left part of the figure presents the **case where experts can solve the task**, while the right part of the figure shows the **case where experts cannot solve the task**.

classifier and human experts, it is essential to ensure that when all human experts make incorrect predictions, the classifier should be selected and learn the current sample, denoted as

$$g_{i1} = 1\left\{\sum_{j=1}^{1+m} \delta_i(j) = 0\right\}, \tag{10}$$

where $\delta_i(j) = g_{ij} = 1\{h_{i,j} = y_i\}$. Thus, the target label vector constructed for the gating network using the ACC method is

$$\mathbf{g}_i = (g_{i1}, \ldots, g_{i(1+m)}) \in \{0,1\}^{1+m}. \tag{11}$$

Then, we use binary cross-entropy to calculate the weight loss for the gating

$$L_{\text{weight}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{1+m}\Bigg(g_{ij}\log\sigma(\mathcal{W}_{ij}) + (1-g_{ij})\log(1-\sigma(\mathcal{W}_{ij}))\Bigg), \tag{12}$$

where $\sigma(z) = \frac{1}{1+\exp(-z)}$.

Finally, the overall team loss in Equation (1) can be expressed as the sum of Equation (8) and Equation (12):

$$L_{\text{team}} = L_{\text{CE}} + L_{\text{weight}}. \tag{13}$$

4 Experiments

In this section, we evaluate our proposed approach on three datasets. First, we simulate the performance of our approach under different team member abilities and team sizes using the classic CIFAR-100 dataset, explaining how our approach optimizes team performance. Next, we further validate our approach on two real-world medical image datasets, NIH and Chaoyang. Particularly, the NIH dataset includes annotations from 22 radiologists, providing comprehensive human expert labeling.

4.1 Experimental Setup

Datasets We evaluated our approach on three datasets. The first is the CIFAR-100 dataset, used for simulation experiments. The other two are real-world medical image datasets, NIH and Chaoyang, containing annotations from multiple human experts.

- CIFAR-100: This dataset consists of 60,000 32×32 color images across 100 classes, with 600 images per class. It is split into 50,000 training images and 10,000 testing images;
- NIH: This dataset, collected by the National Institutes of Health (NIH) Clinical Center, includes chest X-ray images with annotations from 22 radiologists. It contains 4,374 chest X-ray images with up to three possible symptoms per image, making it a multi-label classification dataset;
- Chaoyang: This dataset consists of 6,160 colon slide patches, each with a resolution of 512×512. Each patch includes three noisy labels provided by pathologists, and each image belongs to one of four categories.

Since the CIFAR-100 dataset does not contain annotations from multiple human experts, we simulated experts with different levels of expertise to generate corresponding expert labels. For a team of m human experts, we assume that each expert can perfectly classify a subset of categories. Specifically, we sample the capability value of each expert as $c_i \sim \mathcal{N}(c_{\text{mean}}, c_{\text{std}})$, where $i \in \{1, ..., m\}$, c_i denotes the number of categories that expert i specializes in, c_{mean} represents the average number of categories an expert is proficient in, and c_{std} is the standard deviation of expertise distribution. In our experiments, we set different values of m, c_{mean}, and c_{std} to simulate the performance of teams with varying sizes and levels of expertise.

The NIH and Chaoyang datasets contain human expert annotations, and thus we directly use them in our experiments. The Chaoyang dataset features comprehensive expert annotations, allowing us to form an expert team with two experts and train on the full dataset. However, in the NIH dataset, not every sample is annotated by all experts. To ensure reliability and consistency, we selected pairs of experts and retained samples with the most overlapping annotations.

Baselines We compared the performance of our approach, CE-MoHAIT, with six baseline methods.

- One Classifier: Single classifier model;
- Classifier Team: A team consisting of m classifier models;
- Random Expert: A team consisting of m experts, where each instance is randomly assigned to one expert for prediction;
- Expert Team: A team consisting of m experts, where the expert with the highest weight selected by a gating network makes the prediction;
- JSF: A team consisting of one classifier and m experts, with separate loss calculations for the classifier and the expert team [38];

Table 1. Team accuracies of our approach and the baselines on the CIFAR-100 dataset. To evaluate the impact of different expert ability settings on method performance, we sample the number of categories each expert can classify from a normal distribution $\mathcal{N}(c_{mean}, c_{std})$, where c_{mean} represents the average number of categories an expert can classify, and c_{std} represents the variance of expert abilities.

Method	CIFAR-100		
	$\mathcal{N}(25,5)$	$\mathcal{N}(50,5)$	$\mathcal{N}(75,5)$
One Classifier	78.50(±0.20)	78.50(±0.20)	78.50(±0.20)
Classifier Team	77.06(±0.49)	77.06(±0.49)	77.06(±0.49)
Random Expert	27.13(±1.38)	51.80(±1.01)	76.81(±0.79)
Expert Team	40.02(±1.83)	65.94(±1.65)	87.80(±0.88)
JSF	63.39(±0.93)	54.25(±1.33)	78.75(±1.37)
HAIT	57.95(±1.60)	69.52(±2.12)	90.38(±0.71)
CE-MoHAIT(Ours)	**80.31(±0.46)**	**85.08(±0.67)**	**94.62(±0.46)**

- HAIT: A team consisting of one classifier and m experts, with overall system loss computed through weighted fusion of all member predictions [13];

Training Details For all the experiments, we adopted ResNet-18 [16] pre-trained on ImageNet-1K [17] as the feature extraction network. The Adam optimizer and a cosine annealing scheduler were used, with the cosine annealing period set to one-fifth of the total training epochs. On CIFAR-100, we used 40,000 images for training, 10,000 for validation, and 10,000 for testing. The initial learning rate was set to 2×10^{-4}, the batch size was 512, and the model was trained for 50 epochs. On the NIH and Chaoyang datasets, due to their smaller scales, we employed 10-fold cross-validation with an initial learning rate of 2×10^{-4}, a batch size of 64, and 20 epochs per fold. All the reported experimental results were obtained by repeating training five times with fixed but different random seeds.

4.2 Experimental Results and Analysis

In Table 1, we present the results on the CIFAR-100 dataset. we fixed the number of experts in the team to 2 and reported the team performance when the expert capability c_{mean} was set to 25, 50, and 75. More detailed experimental results on the CIFAR-100 dataset are illustrated in Figure 3. The three subfigures illustrate controlled experiments that investigate the effects of team size and individual human expert capability. Specifically, we vary the team size from 2 to 12 with a step of 2. The average human expert capability c_{mean} is set to 25, 50, and 75, with a figixed standard deviation c_{std} of 5. The results demonstrate that our approach consistently outperforms the current state-of-the-art Human-AI Teams (HAIT) across different team sizes and human expert capabilities. Notably, when the average human expert capability is low ($c_{\text{mean}} = 25$), our approach achieves

Table 2. Team accuracies of our approach and the baselines including standard errors on the NIH and Chaoyang datasets. In the NIH dataset, each radiologist participating in the annotation process has a unique ID. We select two experts to form an expert team, with four different pairs used for the experiment.

Method	Chaoyang	NIH			
		ID=(357,121)	ID=(249,124)	ID=(357,117)	ID=(249,296)
One Classifier	78.23(±0.28)	84.70(±0.13)	84.59(±0.12)	83.13(±0.21)	83.63(±0.08)
Classifier Team	76.88(±0.25)	85.09(±0.17)	84.83(±0.23)	83.44(±0.16)	83.69(±0.27)
Random Expert	83.88(±0.23)	88.04(±0.48)	88.30(±0.49)	89.15(±0.37)	84.58(±0.62)
Expert Team	90.57(±0.12)	90.98(±0.12)	88.54(±0.50)	95.34(±0.00)	91.34(±0.05)
JSF	86.30(±0.01)	90.90(±0.19)	88.26(±0.25)	94.94(±0.69)	90.52(±1.08)
HAIT	90.39(±0.05)	91.01(±0.16)	88.76(±0.50)	95.34(±0.00)	91.36(±0.06)
CE-MoHAIT(Ours)	**91.05(±0.29)**	**91.50(±0.75)**	**88.79(±0.41)**	**95.41(±0.13)**	**91.60(±0.43)**

an improvement of 15% to 18% over HAIT [13], depending on the team size. These results highlight the robustness of our approach across varying team sizes and expert capabilities. Additionally, we observe that the gating network retains significant potential for improvement in expert assignment when human experts have lower individual capabilities.

In Table 2, we further present the validation results on two real-world medical image datasets, NIH and Chaoyang. We focus on one class in the NIH dataset called *airspace opacity*, which accounts for 49.5% of the data. It is a common pulmonary manifestation indicating pneumonia or other fluid-related pathologies. We selected experts with a larger amount of data, forming four groups of paired experts, and annotated the ID of each expert. For the Chaoyang dataset, we construct teams using two out of three available experts, as the remaining expert serves as the ground truth standard of dataset. Compared to the synthetic dataset results, the performance gains on real-world datasets are relatively smaller. This observation aligns with the results shown in Figure 3, where the performance gap diminishes as individual expert capabilities increase. However, in real-world scenarios, considerations such as expert workload and cost often lead to situations where experts specialize in a narrow range of tasks, resulting in lower average expert capabilities. This further underscores the advantage of our approach in practical applications.

In Table 3, we analyze the influence of two key factors on our method using the CIFAR-100 dataset: 1) the removal of the weight loss, defined in Equation (12) and computed from the weight labels constructed by the ACC method; and 2) the investigation of the effect of the ratio between the two weight branches—the classifier weight branch and the human expert weight branch—on team performance. When Equation (12) is removed, the method no longer explicitly accounts for the complementarity between humans and AI, resulting in a significant decline in team performance. Moreover, even with an increased number of human experts, tasks cannot be appropriately allocated to the suitable members, sometimes leading to even worse performance. Additionally, under

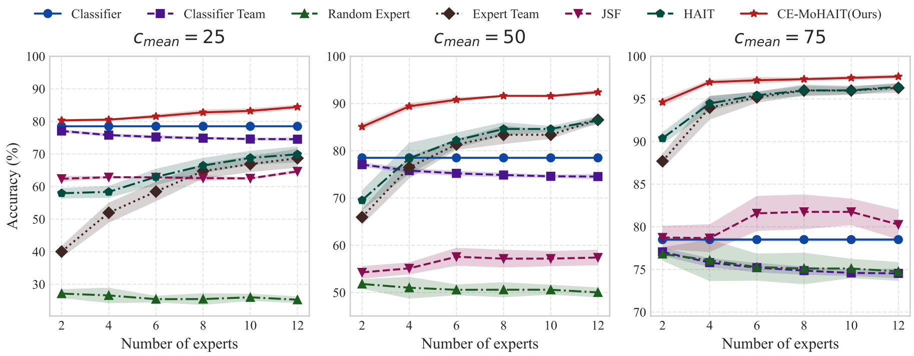


Fig. 3. Team test accuracy of our approach (CE-MoHAIT) in different expert ability(c_{mean}) with increasing team size on the CIFAR-100. Shaded regions display standard errors.

Table 3. The experimental results on CIFAR-100. We conducted experiments under the settings where the number of human experts (m) is 2 and 12, and the human expert capability (c_{mean}) is 25, 50, and 75, respectively, and analyzed the influences of removing the weight loss and varying the ratio between the classifier and human expert weight branches on team performance.

Method	m=2			m=12		
	$\mathcal{N}(25, 5)$	$\mathcal{N}(50, 5)$	$\mathcal{N}(75, 5)$	$\mathcal{N}(25, 5)$	$\mathcal{N}(50, 5)$	$\mathcal{N}(75, 5)$
CE-MoHAIT$_{9:1}$	80.00(±0.51)	83.76(±0.39)	93.80(±0.33)	81.95(±1.16)	92.18(±0.38)	97.66(±0.18)
CE-MoHAIT$_{7:3}$	80.29(±0.24)	84.81(±0.77)	94.55(±0.45)	83.92(±0.65)	92.35(±0.42)	97.49(±0.15)
CE-MoHAIT$_{3:7}$	80.41(±0.42)	85.04(±0.62)	94.59(±0.45)	84.26(±0.73)	92.34(±0.56)	**97.74(±0.17)**
CE-MoHAIT$_{1:9}$	**80.64(±0.25)**	84.91(±0.72)	**94.62(±0.42)**	84.15(±1.00)	**92.47(±0.42)**	97.62(±0.17)
CE-MoHAIT$_{w/o\ WL}$	79.65(±0.31)	76.79(±1.35)	91.07(±1.34)	58.72(±5.08)	74.08(±6.71)	87.99(±1.99)
CE-MoHAIT	80.31(±0.46)	**85.08(±0.67)**	**94.62(±0.46)**	**84.61(±0.69)**	91.87(±0.75)	97.30(±0.31)

team sizes of 2 and 12, we sequentially adjusted the ratio between the classifier weight branch and the human expert weight branch to 1:9, 3:7, 7:3, and 9:1. The results indicate that when the ratio is 1:9, that is, when the classifier weight constitutes a lower proportion—the team performance tends to be better. Intuitively, this is likely because, under the MoE framework, the gating network's task assignment is highly random during the early stages of training; if the classifier weight branch's proportion is too high, it may converge too rapidly and mistakenly learn tasks that are better suited for human experts, thereby undermining the intended human-AI complementarity.

Finally, to demonstrate the stability of our method, we further plotted the training curves in Figure 4. We selected training processes with team sizes of 2 and 12 and expert capabilities of 25, 50, and 75, respectively. The experimental results show that our method exhibits a significant advantage in both final performance and convergence speed, confirming its effectiveness and stability.

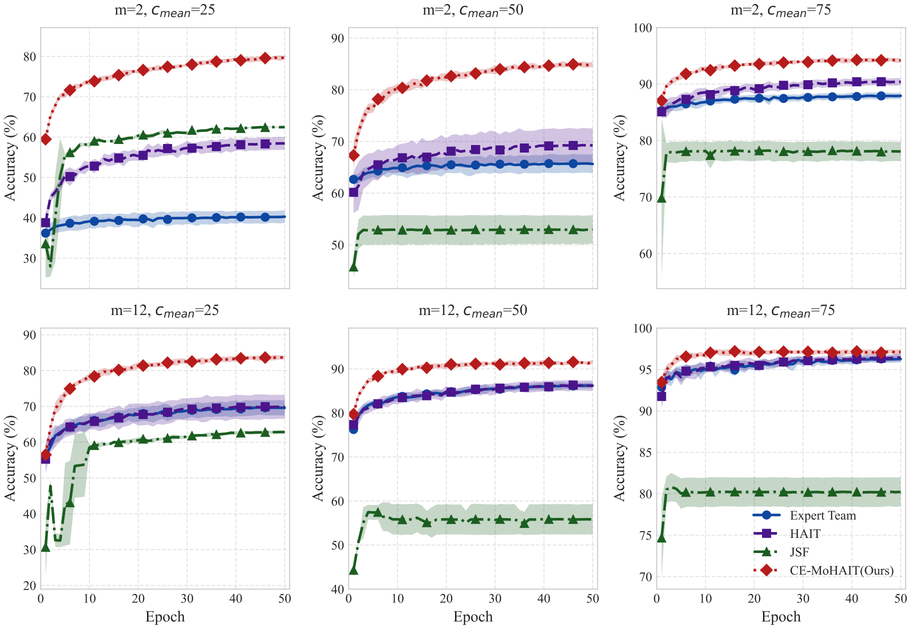


Fig. 4. Team test accuracy during training under setting where the number of human experts (m) is set to 2 and 12, and the expert ability (c_{mean}) is set to 25, 50, and 75, respectively. Shaded regions display standard errors.

5 Conclusion

In this paper, we propose a framework called Complementarity-Enhanced Mixture of Human-AI Teams (CE-MoHAIT). Based on the MoE framework, CE-MoHAIT is applied to human-AI collaboration scenarios to construct a human-AI team consisting of one classifier model and m human experts. By splitting the output weights of the gating network into two branches—the classifier weight branch and the human expert weight branch—we enhance the complementarity between humans and AI within the team. Moreover, we adopt an Adaptive and Complementary Construction (ACC) method to specifically construct weight labels that directly optimize the gating network's output weights, thereby yielding a novel team loss function. Experimental results demonstrate that our method leads to a significant improvement in team performance.

However, the experimental results also reveal a potential shortcoming of CE-MoHAIT. As the team size increases, the overall performance improvement remains limited, indicating that a more fine-grained approach is still required to perceive the differences in capabilities among team members, enhance their complementarity, and thus better allocate tasks. In future work, we plan to introduce more effective methods to address this issue.

Acknowledgments. This work was supported in part by the National Key R&D Program of China (No.2021ZD0113305), the National Natural Science Foundation of China (No.62372381), and the National Science Fund for Distinguished Young Scholars (No.62025205).

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision, pp. 1026–1034. IEEE Computer Society, USA (2015). https://doi.org/10.1109/ICCV.2015.123
2. Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010. Curran Associates Inc., USA (2017)
4. Summers, R.: Nih chest x-ray dataset of 14 common thorax disease categories. NIH Clinical Center: Bethesda, MD, USA (2019)
5. Bilic, P., Christ, P., Li, H.B., et al.: The liver tumor segmentation benchmark (lits). Medical image analysis **84**, 102680 (2023)
6. Strouse, D.J., McKee, K.R., Botvinick, M., et al.: Collaborating with humans without human data. In: Proceedings of the 35th International Conference on Neural Information Processing Systems (2021)
7. Hemmer, P., Schemmer, M., Kühl, N., et al.: Complementarity in human-AI collaboration: Concept, sources, and evidence. arXiv preprint arXiv:2404.00029 (2024)
8. Zhao, Xuehan, et al. "HAIformer: Human-AI Collaboration Framework for Disease Diagnosis via Doctor-Enhanced Transformer." ECAI 2024. IOS Press, 2024. 1495-1502.
9. Mozannar, H., Sontag, D.: Consistent estimators for learning to defer to an expert. In: International conference on machine learning, pp. 7076–7087. PMLR (2020)
10. Raghu, M., Blumer, K., Corrado, G., et al.: The algorithmic automation problem: Prediction, triage, and human effort. arXiv preprint arXiv:1903.12220 (2019)
11. Hemmer, P., Thede, L., Vössing, M., et al.: Learning to defer with limited expert predictions. In: Proceedings of the AAAI Conference on Artificial Intelligence **37**(5), 6002–6011 (2023)
12. Zhang, Z., Ai, W., Wells, K., et al.: Learning to Complement and to Defer to Multiple Users. In: European Conference on Computer Vision, pp. 144–162. Springer (2024)
13. Hemmer, P., Schellhammer, S., Vössing, M., et al.: Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. In: International Joint Conference on Artificial Intelligence (2022)
14. Majkowska, A., Mittal, S., Steiner, D.F., et al.: Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology **294**(2), 421–431 (2020)
15. Zhu, C., Chen, W., Peng, T., et al.: Hard sample aware noise robust learning for histopathology image classification. IEEE transactions on medical imaging **41**(4), 881–894 (2021)

16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
17. Deng, J., Dong, W., Socher, R., et al.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255 (2009)
18. Alves, J.V., Leitão, D., Jesus, S., et al.: Cost-sensitive learning to defer to multiple experts with workload constraints. arXiv preprint arXiv:2403.06906 (2024)
19. Steyvers, M., Tejeda, H., Kerrigan, G., Smyth, P.: Bayesian modeling of human–AI complementarity. Proceedings of the National Academy of Sciences **119**(11), e2111547119 (2022)
20. Verma, R., Barrejón, D., Nalisnick, E.: Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In: International Conference on Artificial Intelligence and Statistics, pp. 11415–11434 (2023)
21. Verma, R., Nalisnick, E.: Calibrated learning to defer with one-vs-all classifiers. In: International Conference on Machine Learning, pp. 22184–22202 (2022)
22. Mozannar, H., Sontag, D.: Consistent estimators for learning to defer to an expert. In: International conference on machine learning, pp. 7076–7087 (2020)
23. Agarwal, N., Moehring, A., Rajpurkar, P., Salz, T.: Combining human expertise with artificial intelligence: Experimental evidence from radiology. National Bureau of Economic Research (2023)
24. Pradier, M.F., Zazo, J., Parbhoo, S., et al.: Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible. AMIA Summits on Translational Science Proceedings **2021**, 525 (2021)
25. Wilder, B., Horvitz, E., Kamar, E.: Learning to Complement Humans. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, pp. 1526–1533 (2020). https://doi.org/10.24963/ijcai.2020/212
26. Chiou, E.K., Lee, J.D.: Trusting automation: Designing for responsivity and resilience. Human factors **65**(1), 137–165 (2023)
27. Shin, D.: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International journal of human-computer studies **146**, 102551 (2021)
28. Lu, Z., Yin, M.: Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2021)
29. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
30. Shazeer, N., Mirhoseini, A., Maziarz, K., et al.: Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In: International Conference on Learning Representations (2017)
31. Lepikhin, D., Lee, H., Xu, Y., et al.: GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In: International Conference on Learning Representations (2021)
32. Munro, R.: Human-in-the-Loop Machine Learning. Manning Publications (2021)
33. Knox, W.B., Stone, P.: Tamer: Training an agent manually via evaluative reinforcement. In: 2008 7th IEEE international conference on development and learning, pp. 292–297 (2008)
34. Rafailov, R., Sharma, A., Mitchell, E., et al.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36**, 53728–53741 (2023)

35. Azar, M.G., Guo, Z.D., Piot, B., et al.: A general theoretical paradigm to understand learning from human preferences. In: International Conference on Artificial Intelligence and Statistics, pp. 4447–4455 (2024)
36. Ethayarajh, K., Xu, W., Muennighoff, N., et al.: KTO: Model Alignment as Prospect Theoretic Optimization. In: Proceedings of the 41st International Conference on Machine Learning (2024)
37. Xu, H., Sharaf, A., Chen, Y., et al.: Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In: Forty-first International Conference on Machine Learning (2024)
38. Keswani, V., Lease, M., Kenthapadi, K.: Towards unbiased and accurate deferral to multiple experts. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 154–165 (2021)
39. Kroll, J.A.: Why AI is Just Automation. Brookings Institution (2021)