ActiveHAI: Active Collection Based Human-AI Diagnosis with Limited Expert Predictions

Xuehan Zhao[1], **Jiaqi Liu**[1*], **Xin Zhang**[1], **Zhiwen Yu**[2,1] and **Bin Guo**[1]
[1]Northwestern Polytechnical University
[2]Harbin Engineering University
{xuehan.zhao, zhangxin1}@mail.nwpu.edu.cn, {jqliu, zhiwenyu, guob}@nwpu.edu.cn

Abstract

Recent studies indicate that human-AI collaboration performs better than either alone, particularly in medical diagnosis. Beyond collaboration methods that focus on assigning tasks to humans or AI, like deferral, combining human and AI decisions with their confidence scores is emerging as a promising strategy. Due to high cognitive load, doctors often struggle to provide confidence assessments, necessitating explicit human uncertainty evaluation through a limited number of additional expert predictions. There are two challenges. (1) how to actively collect limited yet representative expert predictions? (2) how to accurately evaluate human uncertainty with limited expert predictions? To address the challenges, we propose ActiveHAI, an active human-AI diagnosis method that reduces expert costs through a median-window sampling strategy that actively selects representative samples near the estimated median; and evaluate expert confidence through an evaluator module that integrates sample features and expert predictions, converting them into probability distributions. Experiments on three real-world datasets show that ActiveHAI surpasses doctor and other human-AI methods by 16.3% and 3.6% in accuracy, respectively. Furthermore, ActiveHAI reaches 97.2% relative accuracy, even with just eight expert predictions per class.

1 Introduction

Artificial intelligence (AI) demonstrates remarkable potential in healthcare [Rajpurkar *et al.*, 2022], outperforming cardiology residents in identifying 12-lead electrocardiogram abnormalities [Ribeiro *et al.*, 2020]. However, AI algorithms face challenges like out-of-domain inapplicability, bias, and lack of interpretability [Topol, 2019]. Motivated by this observation, prior work [Yu *et al.*, 2021; Bansal *et al.*, 2021; Gu *et al.*, 2023] has explored human-AI collaboration to leverage their complementary strengths. Beyond the methods that assign inputs to humans or AI, like deferral [Madras *et al.*, 2018; Keswani *et al.*, 2021], combining human and AI decisions is emerging as a promising strategy [Wilder *et al.*, 2021; Steyvers *et al.*, 2022; Zhao *et al.*, 2024]. Additionally, when medical AI is accessible only as a 'locked model' due to privacy constraints, effectively combining doctors' predictions with pretrained AI outputs becomes critical.

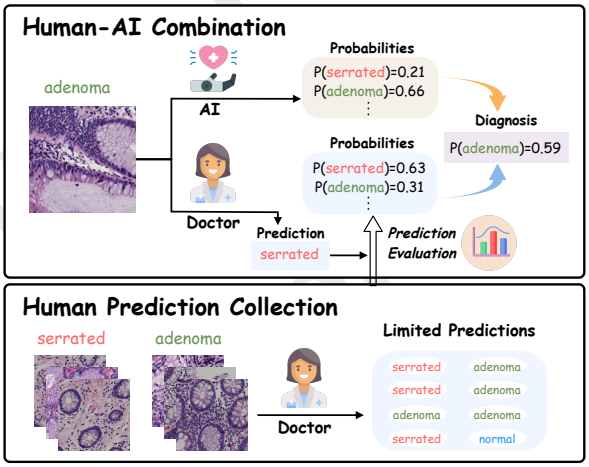

Figure 1: A case of human-AI diagnostic combination based on individual doctors' additional predicted labels. Prediction Evaluation refers to converting expert predictions into probability distributions.

Human-AI decision-making combinations [Steyvers *et al.*, 2022] often require probabilistic distributions from both humans and AI. However, due to cognitive burden, asking doctors to provide confidence scores for all possible labels is impractical. Instead, additional predictions from doctors are needed to infer these scores, but extensive expert predictions incur high costs [Wang *et al.*, 2024; Liu *et al.*, 2024], limiting the feasibility of human-AI diagnostic systems. We aim to make accurate human-AI diagnostics with limited expert predictions. Figure 1 shows that doctor prediction is combined with AI probabilities to generate the final diagnosis.

Existing research [Kerrigan *et al.*, 2021; Gupta *et al.*, 2023; Singh *et al.*, 2023] on combining human class-level predictions with AI probabilistic outputs often leverages confusion matrices, which quantify the relationship between expert predictions and ground truth labels. However, constructing confusion matrices requires substantial expert predictions to accurately estimate probabilities, which is time-consuming and

*Corresponding author

costly, especially for large-scale medical imaging [Willemink *et al.*, 2020; Alzubaidi *et al.*, 2021]. Recent work [Hemmer *et al.*, 2023; Mozannar *et al.*, 2023; Dvijotham *et al.*, 2023; Alves *et al.*, 2024] on human-AI collaborative diagnosis has considered the constraints of limited expert predictions, using a small number of predictions to identify doctors' specific strengths. However, these studies primarily focus on deferring tasks to experts, overlooking the probabilistic distribution of expert predictions.

To improve the accuracy of human-AI diagnosis with limited human prediction labels, there are two challenges. First, given the high cost of inviting experts to predict samples that accurately represent their abilities, *how can we design a strategy to collect a limited yet representative subset of expert predictions actively?* Second, existing methods perform poorly when the number of available expert predictions is limited, *how can we design an efficient method to evaluate human prediction probability distributions with limited predictions?* To address these challenges, we propose two solutions: (i) For the first challenge, we introduce a median-window based active collection strategy. This strategy iteratively selects representative samples near the median of the estimated values and provides expert samples for prediction, thus reducing the cost of expert annotations on medical samples. (ii) For the second challenge, we propose an evaluator module that integrates sample features and expert predictions. This module combines the feature vectors from the pretrained AI with the encoded expert prediction vectors, enabling the evaluation of probability distributions with limited expert predictions.

In this paper, we propose ActiveHAI, an active collection based human-AI diagnostic combination method that enhances diagnostic accuracy with limited expert predictions. First, we use the proposed median-window sampling method to actively select samples for expert predictions, iteratively generating fully labeled data. Second, we train the evaluator module using data that includes expert predictions and ground truth labels, enabling improved evaluation by jointly encoding sample features and expert predictions. Finally, we combine the human prediction probability distributions generated by the evaluator module with the pretrained AI's probabilistic outputs to compute joint probabilities, ultimately producing the final human-AI combined prediction.

Our main contributions are as follows:

- **Active Collection**: We propose the median-window active collection algorithm to actively select expert predictions for human-AI diagnosis, enabling efficient evaluation of human prediction probability distributions.
- **Human Prediction Evaluation**: We propose an evaluator module that enhances the ability to transform expert predictions into probability distributions by leveraging pretrained feature layers and prediction embeddings.
- **Experiment Study**: Experiments on three real-world datasets show that the proposed method outperforms individual human and other human-AI collaboration methods by 16.3% and 3.6% in diagnosis accuracy, respectively. For reproducibility, we release the code and data in https://github.com/mercyzi/ActiveHAI.git.

2 Related Work

2.1 Medical Diagnosis with Limited Labels

Researchers have explored various techniques to improve predictions in data-scarce scenarios. For instance, [Hemmer *et al.*, 2023] utilized semi-supervised learning to generate artificial labels with limited annotations, enabling deferred learning. [Chae and Kim, 2023] applied transfer learning in medical image analysis to enhance accuracy with small medical datasets. [Kotia *et al.*, 2021] focused on few-shot learning to learn effective feature representations from a limited number of labeled samples in medical imaging. However, these approaches often rely on large amounts of unlabeled data and domain similarity, which may not always be feasible in real-world medical applications.

In contrast, active learning methods [Liu *et al.*, 2020; Budd *et al.*, 2021] aim to efficiently identify the most informative subsets of unlabeled samples, thus enabling more effective training of AI models and reducing the expensive annotation burden typically associated with medical image data. For example, [Tang *et al.*, 2023] employed active learning to tackle label scarcity, achieving more accurate identification of gastrointestinal diseases. [Zhang *et al.*, 2024a] proposed an interactive image annotation framework that improves prostate MRI image segmentation accuracy with fewer interactive annotations. However, few studies have explored how to actively select limited expert predictions in order to evaluate human expert capabilities effectively.

2.2 Human-AI Collaborative Diagnosis

In human-AI collaborative classification [Fragiadakis *et al.*, 2024], humans and AI systems collaborate and complement each other to tackle more complex tasks. It is primarily divided into Learning to Defer (L2D) and Learning to Combine (L2C) [Zhang *et al.*, 2024b]. L2D [Madras *et al.*, 2018; Hemmer *et al.*, 2022; Dvijotham *et al.*, 2023] refers to assigning the final decision to either the expert or the AI. For example, [Mozannar and Sontag, 2020] proposed a framework based on classifiers and rejectors that defers the decision on chest X-rays to experts. [Verma and Nalisnick, 2022] introduced a calibrated L2D system that defers diagnosing skin lesions to experts. [Mao *et al.*, 2024] extended the L2D model to multiple medical experts.

In this paper, we focus on L2C, which combines AI and human predictions for the final diagnosis, ensuring higher accuracy by leveraging the complementary strengths of both systems and capturing information that machines may miss [Groh *et al.*, 2022]. For example, [Kerrigan *et al.*, 2021] developed an algorithm that combines human predictions with machine model probabilities, improving accuracy and reliability. [Steyvers *et al.*, 2022] proposed a Bayesian framework that uses varying confidence scores, demonstrating that a hybrid human-AI combination outperforms either prediction type individually. [Gupta *et al.*, 2023] combined calibrated model probabilities with expert predictions, improving the alignment of model outputs with expert predictions. However, these studies do not address the limited availability of expert-predicted labels in the medical field, which may lead to suboptimal accuracy in human-AI diagnosis.

3 Problem Formulation

In this section, we consider human-AI decision combination in medical diagnosis tasks. For a given medical instance $x \in \mathcal{X}$, we assume that we have access to the probability vector $p_h(x) \in P_h$ predicted by a human expert and the probability vector $p_a(x) \in P_a$ predicted by a pretrained AI model. The goal is to accurately predict the true label $y \in Y = \{1, ..., K\}$ with K denoting the number of classes. Based on Bayesian theory, assuming that the predictions $p_h(x)$ and $p_a(x)$ are conditionally independent, we compute their joint probability using the product rule. Similar to [Kerrigan *et al.*, 2021], we represents the prediction probability of class j as:

$$p(y = j|p_h(x), p_a(x)) = \frac{p_h(x)_j \cdot p_a(x)_j}{\sum_{k=1}^{K} (p_h(x)_k \cdot p_a(x)_k)}, \quad (1)$$

where $p_h(x)_j$ and $p_a(x)_j$ represent the prediction probabilities of class j by humans and AI, respectively.

However, it is infeasible in practice to directly obtain $p_h(x)$, and it is costly for doctors to evaluate confidence levels when making predictions. Our premise for human-AI collaboration is to train an evaluator module $E : \mathcal{X} \times H \rightarrow P_h$ that models the probability distribution of individual doctor predictions $h \in H = \{1, \ldots, K\}$. Rather than directly converting h into a one-hot encoding, we use historical expert prediction data to fit a more fine-grained probability distribution. The capabilities of individual doctors vary across different disease feature spaces, and ground-truth labels in medical diagnosis datasets typically require consensus among multiple experts. For example, three expert pathologists independently annotate the slides in a colon slide dataset [Zhu *et al.*, 2021], and the consistent results are used as the true labels. Therefore, doctors collaborating with AI models need to provide additional self-predictions to train the evaluator module for distribution fitting.

We define the binary data containing medical instances and the ground-truth labels as $D^u = \{(x_i, y_i)\}_{i=1}^{K \cdot u}$, and the ternary data that additionally includes human expert predictions as $D^l = \{(x_i, y_i, h_i)\}_{i=1}^{K \cdot l}$. Here, l represents the number of samples with human predictions for each class, $u = n - l$ denotes the remaining number of samples per class, and n is the total number of samples in each class. Acquiring individual expert predictions incurs additional costs and human effort, particularly in the medical domain. Therefore, we consider achieving efficient human-AI diagnostic collaboration at a lower cost by evaluating the predictive probability distribution under a limited number of expert predictions l. The goal is to leverage the distribution derived from limited expert predictions l to approximate the distribution of full expert predictions n, thereby enabling a more accurate human-AI diagnostic combination.

In this study, the network parameters of the pretrained AI model $M : \mathcal{X} \rightarrow P_a$ are fixed, and we redefine the human-AI combination probability of class j as:

$$p(y = j|x, h) = \frac{E_\theta(x, h)_j \cdot M(x)_j}{\sum_{k=1}^{K} (E_\theta(x, h)_k \cdot M(x)_k)}, \quad (2)$$

where $E_\theta(x, h)$ and $M(x)$ represent the outputs of the proposed evaluator module and the AI model, respectively, and θ denotes the trainable parameters of the evaluator module.

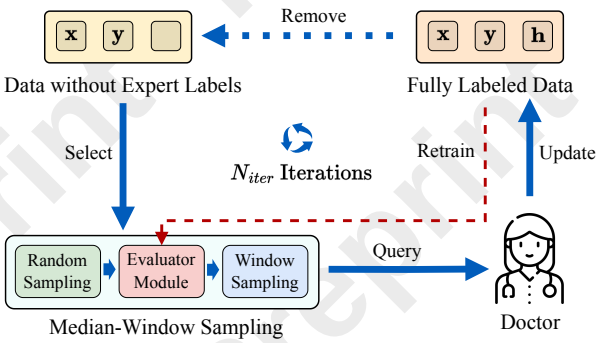

Figure 2: Iteratively collect human predictions via median-window active collection. The red dashed line indicates that each iteration will train the evaluator module with new, fully labeled data.

4 Approach

In this section, we describe the proposed method for fitting the probability distribution using limited expert predictions. The approach consists of two main parts: (1) collecting representative limited expert predictions to construct the fully labeled data D^l; (2) training the proposed evaluator module on D^l to fit the probability distribution of the expert predictions.

4.1 Active Collection of Human Predictions

The first component of our method is the active selection of samples for annotation by medical experts. By selecting the most representative data samples, we effectively train the evaluator module to reduce the cost of expert predictions.

In active learning, the algorithm selects the most informative data samples to train AI models and efficiently reduce labeling costs. Unlike traditional active learning approaches: (1) They involve human annotation for samples with unknown true labels, whereas, in our task, the true labels of all samples are known. (2) Their goal is to improve the performance of the AI model itself. In contrast, our goal is to enhance the performance of the evaluator to estimate the human predictive probability distribution better. To this end, we propose a novel sample collection strategy, Median-Window Active Collection (MWAC), to collect predictions from medical experts.

As shown in Figure 2, we iteratively collect human predictions. Through median-window sampling, a subset of samples from D^u is selected for expert annotation, generating a new fully labeled dataset D^l. Furthermore, the updated D^l from each iteration is used as training data for the evaluator module, which is retrained and utilized in subsequent median-window sampling iterations. Specifically, the median-window sampling consists of three primary steps: (1) *Random Sampling*: Since disease diagnosis is a multi-class problem, we aim to uniformly evaluate the predictive probability distribution of experts across different classes. To this end, we uniformly and randomly sample N samples per class from D^u. (2) *Estimation Generation*: For the N samples in each class, in order to select the most representative samples, we use the evaluator module to calculate the estimates e_i:

$$e_i = \left| E(x_i, \tilde{h}_i)_{\tilde{h}_i} - 0.5 \right|, \quad (3)$$

Algorithm 1: Median-Window Active Collection

Input: D^u: data without expert labels.
N_{iter}: number of active learning iterations.
l: number of human predictions for each class.
$QueryExpert$: ask expert to label samples.
$TrainEvaluator$: train the evaluator module.
Output: D^l: fully labeled data after active learning.

Initialization:
for $k \leftarrow 1$ **to** K **do**
  Randomly choose l_1 samples for class k from D^u.
  $L_1 \leftarrow QueryExpert(l_1)$.
  Add L_1 to D^l, and remove samples from D^u.
$E \leftarrow TrainEvaluator(D^l)$.
for $n \leftarrow 2$ **to** N_{iter} **do**
  for $k \leftarrow 1$ **to** K **do**
    Randomly choose N samples for k from D^u.
    $\{(x_i, e_i)\}_{i=1}^{N} \leftarrow$ `Estimate(`N, E`)`.
    Sort $[(x_i, e_i)]_{i=1}^{N}$ by e_i in descending order.
    Construct median-window $[W_s, W_s + W_l]$.
    Randomly choose l_n samples from window.
    $L_n \leftarrow QueryExpert(l_n)$.
    Add L_n to D^l, and remove samples from D^u.
  $E \leftarrow TrainEvaluator(D^l)$.
return D^l.
Function `Estimate(`N, E`)`**:**
  $S \leftarrow \emptyset$.
  for *each instance* x_i, y_i *from* N **do**
    $\tilde{h}_i \leftarrow y_i$.
    $e_i \leftarrow \left| E(x_i, \tilde{h}_i)_{\tilde{h}_i} - 0.5 \right|$.
    $S \leftarrow S \cup \{(x_i, e_i)\}$.
  return S.

where, $\tilde{h}_i = y_i$ assumes that expert predictions align with the ground truth labels, and $E(x_i, \tilde{h}_i)_{\tilde{h}_i}$ denotes the probability output of the evaluator module for the pseudo expert prediction $\tilde{h}_i$ category. (3) *Window Sampling*: We rank the N samples in descending order based on their estimated values and construct a window with a starting index W_s and a length W_l. This window corresponds to the interval $[W_s, W_s + W_l]$ within the ordered sample list $[0, N]$. The MWAC algorithm is described in detail in Algorithm 1.

Since the estimated value e is defined as the absolute difference between the probability output by the evaluator for the true label and 0.5, the closer this probability is to 1 or 0, the larger the estimated value e. This reflects an extreme estimation of human predictions. Conversely, when the probability is closer to 0.5, the estimated value becomes smaller, representing a more ambiguous estimation of human predictions. We argue that the median of the estimated values better represents the collective characteristics of the sample set, enabling a more reasonable estimation of human predictions, which is validated in the experimental section. Therefore, we select the sampling window's starting index W_s based on the estimated values' median. Finally, we perform sampling within this window to obtain a limited number of expert predictions.

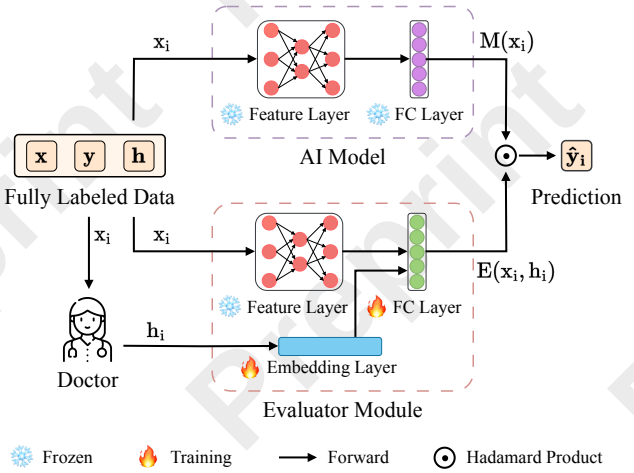

Figure 3: Human-AI decision combination based on proposed evaluator module. The human-AI collaborative decision-making process is achieved by calculating the joint distribution of the human prediction probability distribution and the AI prediction probability distribution, where the human prediction probability distribution is estimated using the evaluator module.

4.2 Design of Evaluator Module

As shown in Figure 3, for a medical instance with features x, expert doctors and the AI model independently provide predicted labels and probabilities. The evaluator module processes the expert's predicted label to generate a probability distribution, and the final human-AI collaborative prediction $\hat{y}$ is computed using Equation (2). The proposed evaluator module is designed to assess the probability distribution of doctors' decisions, represented as a probability vector over multiple categories.

Existing methods for combining human predictions and AI probabilities usually rely on confusion matrices, which quantify the alignment between expert predictions and ground truth labels. However, their reliability declines with limited annotations. To address this, we propose an evaluator module integrating expert predictions with sample features, enabling robust evaluation in scenarios with fewer annotations. Specifically, in the evaluator module, we first extract feature vectors from instance x using a feature layer:

$$f_x = \text{FeatureLayer}(x; \theta_x), \tag{4}$$

where θ_x is the feature layer parameter from the pretrained AI model. Then, we pass the human prediction h through the embedding layer to obtain the embedding vector:

$$f_h = \text{EmbeddingLayer}(h; \theta_h), \tag{5}$$

where θ_h is the embedding layer parameter. Finally, the feature vector and embedding vector are added together and then passed through a linear layer to obtain the final human prediction probability distribution:

$$E(x, h) = \text{Softmax}(\text{LinearLayer}(f_x + f_h; \theta_l)), \tag{6}$$

where θ_l is the linear layer parameter. We can train the evaluator module by calculating the cross entropy loss between human-AI combination prediction $\hat{y}$ and true labels y:

$$\mathcal{L} = -\sum_{i=1}^{K \cdot l} y_i \log(p_\theta(y = \hat{y}_i | x_i, h_i)), \tag{7}$$

where θ represents the set of θ_h and θ_l, excluding θ_x, as the feature layer network parameters are fixed.

5 Experiments

In this section, we conduct extensive experiments on real-world medical diagnosis datasets to evaluate our method.

5.1 Experimental Settings

Datasets. We extensively evaluate the proposed method on three datasets: MZ-10 [Chen *et al.*, 2023], DR-5 [Ju *et al.*, 2022], and Chaoyang-3 [Zhu *et al.*, 2021]. MZ-10 is a medical consultation dataset, while Chaoyang-5 and DR-3 are medical imaging datasets.

MZ-10: A large-scale corpus for evaluating medical consultation systems. This dataset covers 331 symptoms and 10 pediatric diseases. As the dataset lacks doctor-predicted labels, we artificially generate three sets of expert prediction labels (D_1, D_2, and D_3) to represent doctors with varying levels of expertise. For each set of artificial expert labels, we sample probabilities P from a uniform distribution with mean a, where a denotes overall accuracy and P represents accuracy for each class. Given the machine model's accuracy of 0.69, we set a to 0.65, 0.7, and 0.75 to simulate doctors with expertise slightly below, equal to, and above the AI model.

DR-5: A diabetic retinopathy dataset with five grading labels. It includes prediction labels from multiple ophthalmologists, and we select the ophthalmologist who provides the most significant number of expert predictions. Due to the limited number of samples for proliferative diabetic retinopathy, we randomly sample 200 samples from each category to ensure sufficient representation and unbiased evaluation.

Chaoyang-3: A colon slide dataset from Chaoyang Hospital, with prediction labels from the expert pathologist. Our study does not consider the diagnosis of adenocarcinoma, as the prediction accuracy for adenocarcinoma samples by experts approaches 100%, which limits the complementarity of human-AI collaboration in this category. We randomly sample 600 slides from each of the other three categories (normal, serrated, and adenoma).

Baselines. We compare our proposed ActiveHAI with the three following baselines:

CM [Kerrigan *et al.*, 2021; Gupta *et al.*, 2023; Singh *et al.*, 2023]: A human-AI decision combination method that evaluates the probability distribution of human predictions using a confusion matrix, followed by Bayesian fusion of human and AI probabilities.

PCM [Hemmer *et al.*, 2023]: A method that generates pseudo-human predictions to extend the confusion matrix, enabling a similar human-AI decision combination as CM.

Collab [Zhang *et al.*, 2024b]: A human-AI decision combination method that directly combines human predictions and AI probabilities through a collaboration module.

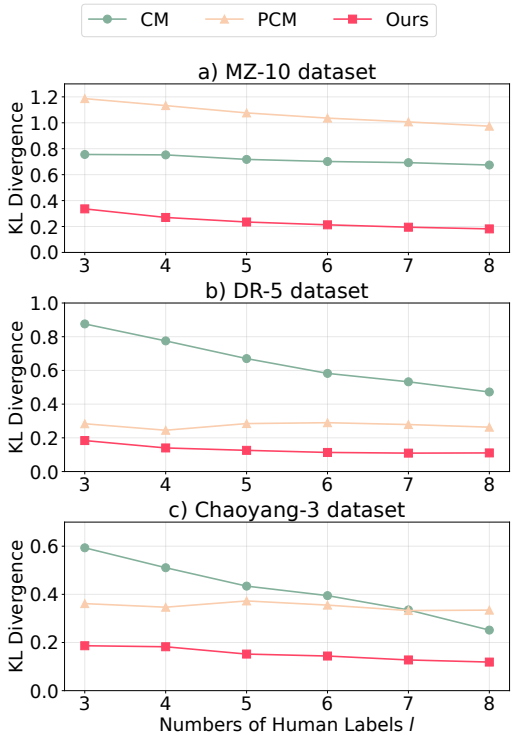

Figure 4: KL divergence under l human expert predictions for our method and the confusion matrix method on three datasets. For MZ-10, we present results on D_1.

Metrics. We use two widely used metrics to evaluate the proposed method's performance: accuracy and KL divergence. Accuracy measures the overall performance of the human-AI combination, while KL divergence quantifies the difference between the human probability distribution under limited predictions and the distribution under all predictions, thus evaluating the quality of the probability distribution fitting.

Implementation Details. We implement ActiveHAI using PyTorch on a single NVIDIA 3090 GPU. For the feature layer, we employ a two-layer Transformer [Vaswani *et al.*, 2017] encoder for MZ-10, and a pretrained EfficientNet-B1 [Tan and Le, 2019] model for DR-5 and Chaoyang-3. The embedding layer dimension is set to 512. The evaluator module is trained for 100 epochs using the Adam optimizer with a learning rate of 3×10^{-4}. For MZ-10, we sample the probability P ten times for each set of artificial expert predictions, repeating each sampling experiment five times. For DR-5 and Chaoyang-3, we perform five-fold cross-validation, repeating each fold ten times.

We consider the following numbers of expert predictions: $l \in \{3, 4, 5, 6, 7, 8, \text{All}\}$. Initially, the number of expert predictions is set to 2, and it increases by 1 in each iteration. The random sampling size N is set to 100, and the median-window length W_l is set to 5. For D_1, D_2, and D_3 in MZ-10, the window starting points W_s are set to 65, 50, and 50, respectively. For DR-5, W_s is set to 55, and for Chaoyang-3, W_s is set to 50.

	l	3	4	5	6	7	8	All
D_1	AI	68.81 (±0.50)	68.81 (±0.50)	68.81 (±0.50)	68.81 (±0.50)	68.81 (±0.50)	68.81 (±0.50)	68.81 (±0.50)
	Human	65.79 (±0.92)	65.79 (±0.92)	65.79 (±0.92)	65.79 (±0.92)	65.79 (±0.92)	65.79 (±0.92)	65.79 (±0.92)
	CM	76.27 (±2.17)	76.09 (±1.84)	76.42 (±1.30)	76.45 (±1.42)	76.60 (±1.38)	76.85 (±1.27)	**84.57** (±1.18)
	PCM	72.36 (±2.00)	73.27 (±1.82)	74.08 (±1.71)	74.59 (±1.60)	75.03 (±1.53)	75.37 (±1.60)	**84.57** (±1.18)
	Collab	72.88 (±2.35)	74.83 (±2.30)	76.55 (±1.95)	77.59 (±1.60)	78.37 (±1.68)	78.98 (±1.45)	82.41 (±1.15)
	ActiveHAI (Ours)	**79.84** (±1.65)	**80.56** (±1.73)	**81.15** (±1.74)	**81.42** (±1.60)	**81.70** (±1.52)	**81.88** (±1.44)	83.92 (±1.24)
	w/o MWAC	78.24 (±2.35)	78.98 (±2.10)	79.75 (±1.93)	80.07 (±1.71)	80.51 (±1.65)	80.86 (±1.52)	83.92 (±1.24)
	w/o EM	77.40 (±1.40)	77.07 (±1.32)	76.70 (±1.19)	76.25 (±1.13)	75.94 (±1.25)	75.70 (±1.29)	84.57 (±1.18)
D_2	Human	70.67 (±1.00)	70.67 (±1.00)	70.67 (±1.00)	70.67 (±1.00)	70.67 (±1.00)	70.67 (±1.00)	70.67 (±1.00)
	CM	79.52 (±2.83)	79.67 (±2.34)	79.39 (±2.14)	79.40 (±1.80)	79.28 (±1.67)	79.55 (±1.48)	**86.29** (±1.29)
	PCM	75.85 (±2.05)	76.50 (±1.79)	76.91 (±1.88)	77.47 (±1.66)	77.88 (±1.61)	78.15 (±1.48)	**86.29** (±1.29)
	Collab	75.92 (±3.23)	77.90 (±2.43)	78.84 (±2.21)	79.79 (±1.93)	80.31 (±1.85)	80.75 (±1.77)	84.14 (±1.08)
	ActiveHAI (Ours)	**81.64** (±2.04)	**82.73** (±1.71)	**83.30** (±1.64)	**83.55** (±1.50)	**83.77** (±1.33)	**83.84** (±1.41)	85.55 (±1.20)
	w/o MWAC	80.66 (±2.20)	81.40 (±1.88)	81.68 (±2.01)	82.04 (±2.09)	82.25 (±1.94)	82.56 (±1.85)	85.55 (±1.20)
	w/o EM	80.55 (±2.03)	80.21 (±1.74)	79.78 (±1.70)	79.36 (±1.74)	79.06 (±1.57)	78.72 (±1.50)	86.29 (±1.29)
D_3	Human	75.70 (±1.11)	75.70 (±1.11)	75.70 (±1.11)	75.70 (±1.11)	75.70 (±1.11)	75.70 (±1.11)	75.70 (±1.11)
	CM	82.85 (±2.89)	82.44 (±2.76)	82.60 (±1.86)	82.59 (±1.68)	82.59 (±1.69)	82.77 (±1.65)	**88.14** (±1.42)
	PCM	79.51 (±1.50)	80.00 (±1.60)	80.44 (±1.59)	80.75 (±1.65)	81.06 (±1.64)	81.20 (±1.56)	**88.14** (±1.42)
	Collab	79.17 (±2.73)	80.35 (±2.52)	81.56 (±2.08)	82.30 (±2.06)	82.73 (±1.85)	83.13 (±1.85)	86.12 (±1.49)
	ActiveHAI (Ours)	**83.90** (±2.56)	**84.92** (±2.27)	**85.27** (±2.06)	**85.70** (±1.90)	**85.83** (±1.82)	**85.99** (±1.85)	87.78 (±1.65)
	w/o MWAC	82.90 (±3.40)	83.50 (±2.97)	83.87 (±2.82)	84.22 (±2.62)	84.64 (±2.28)	84.92 (±2.14)	87.78 (±1.65)
	w/o EM	83.51 (±2.16)	83.36 (±1.67)	82.97 (±1.62)	82.58 (±1.63)	82.18 (±1.55)	81.94 (±1.55)	88.14 (±1.42)

Table 1: Diagnosis accuracy under different numbers of l human expert predictions for the synthetic experts D_1, D_2, and D_3 on MZ-10.

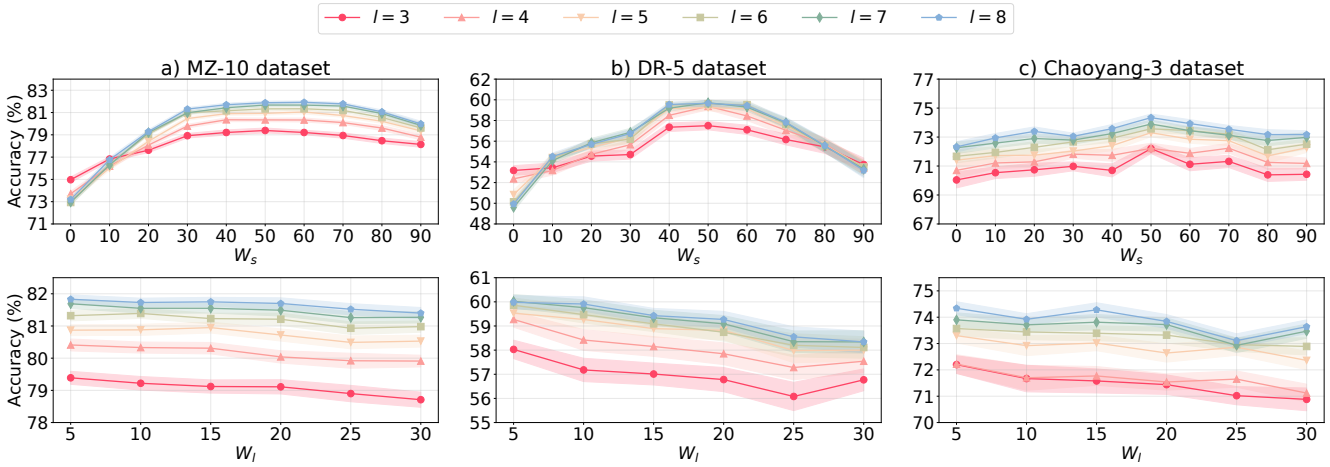

Figure 5: Effect of the median-window parameters W_s and W_l in diagnosis accuracy on three datasets. For MZ-10, we present results on D_1.

5.2 Overall Performance

We present the accuracy performance of the human-AI diagnosis combination models on the MZ-10, DR-5, and Chaoyang-3 datasets in Tables 1, 2, and 3, respectively. Overall, when the number $l \leq 8$ of human expert predictions, the proposed ActiveHAI method outperforms all baseline methods in diagnosis accuracy. Specifically, the results show that when $l \leq 8$, our method achieves an average accuracy improvement of 17.8% (on MZ-10), 14.8% (on DR-5), and 13.2% (on Chaoyang-3) compared to human predictions. Moreover, our method also surpasses existing human-AI combination models by 4.6% (on MZ-10), 1.5% (on DR-5), and 2.8% (on Chaoyang-3) under the same conditions. These results highlight the significant advantage of our method under limited human predictions. Furthermore, when $l =$ All, our method achieves accuracy comparable to other approaches, demonstrating its ability to maintain performance under complete human predictions.

Figure 4 shows the performance of human prediction probability distribution fitting for varying numbers of available human expert predictions on the MZ-10, DR-5, and Chaoyang-3 datasets. Since Collab directly combines human predictions with AI probabilities, these results highlight the quality differences between our method, CM, and PCM in converting human predictions into probability distributions. On the MZ-10 dataset, our ActiveHAI method reduces KL divergence by an average of 66.9% and 77.9% compared to CM and PCM, respectively, under limited expert predictions. On the DR-5 dataset, ActiveHAI reduces KL divergence by an average of 79.8% and 52.3% compared to CM and PCM,

l	3	4	5	6	7	8	All
AI	48.90 (±0.86)	48.90 (±0.86)	48.90 (±0.86)	48.90 (±0.86)	48.90 (±0.86)	48.90 (±0.86)	48.90 (±0.86)
Human	51.80 (±1.08)	51.80 (±1.08)	51.80 (±1.08)	51.80 (±1.08)	51.80 (±1.08)	51.80 (±1.08)	51.80 (±1.08)
CM	53.81 (±4.69)	55.41 (±4.21)	56.70 (±3.54)	57.76 (±3.24)	58.44 (±3.21)	58.42 (±3.10)	62.60 (±1.74)
PCM	57.72 (±4.39)	58.27 (±3.43)	58.70 (±5.40)	58.52 (±4.47)	59.14 (±3.45)	59.13 (±3.32)	62.60 (±1.74)
Collab	51.01 (±4.85)	53.67 (±4.80)	55.62 (±4.30)	56.93 (±2.97)	57.88 (±2.90)	57.92 (±3.08)	**63.72** (±2.32)
ActiveHAI (Ours)	**58.03** (±3.82)	**59.27** (±3.19)	**59.53** (±3.13)	**59.86** (±2.92)	**60.02** (±2.86)	**59.98** (±2.79)	62.39 (±2.31)
w/o MWAC	56.13 (±4.53)	57.62 (±3.87)	58.35 (±3.96)	59.18 (±3.38)	59.63 (±2.99)	59.77 (±3.09)	62.39 (±2.31)
w/o EM	55.63 (±3.91)	56.91 (±3.89)	56.87 (±3.95)	56.99 (±3.65)	57.38 (±3.48)	59.50 (±3.13)	62.60 (±1.74)

Table 2: Diagnosis accuracy under different numbers of l human expert predictions on DR-5.

l	3	4	5	6	7	8	All
AI	68.28 (±1.37)	68.28 (±1.37)	68.28 (±1.37)	68.28 (±1.37)	68.28 (±1.37)	68.28 (±1.37)	68.28 (±1.37)
Human	64.72 (±0.70)	64.72 (±0.70)	64.72 (±0.70)	64.72 (±0.70)	64.72 (±0.70)	64.72 (±0.70)	64.72 (±0.70)
CM	69.02 (±6.10)	70.05 (±5.59)	71.27 (±3.73)	72.03 (±3.36)	72.32 (±3.20)	73.03 (±2.97)	76.00 (±1.51)
PCM	69.96 (±4.50)	70.29 (±3.94)	70.26 (±4.30)	71.00 (±4.26)	71.27 (±4.28)	71.28 (±4.29)	76.00 (±1.51)
Collab	66.39 (±5.51)	68.42 (±4.98)	69.12 (±4.53)	70.03 (±4.28)	71.27 (±3.87)	70.76 (±3.65)	76.41 (±1.29)
ActiveHAI (Ours)	**72.20** (±3.36)	**72.22** (±3.55)	**73.30** (±2.78)	**73.57** (±2.72)	**73.90** (±2.65)	**74.34** (±2.47)	**77.06** (±1.64)
w/o MWAC	71.30 (±3.54)	71.86 (±3.70)	72.63 (±2.96)	73.26 (±2.78)	73.48 (±2.90)	73.96 (±3.10)	77.06 (±1.64)
w/o EM	69.45 (±4.85)	70.82 (±4.02)	72.05 (±3.11)	72.31 (±2.65)	72.73 (±2.65)	73.27 (±2.11)	76.00 (±1.51)

Table 3: Diagnosis accuracy under different numbers of l human expert predictions on Chaoyang-3.

respectively. On the Chaoyang-3 dataset, ActiveHAI reduces KL divergence by an average of 62.7% and 56.8% compared to CM and PCM, respectively. These results indicate that our method significantly improves the effectiveness of fitting distributions under limited expert predictions.

5.3 Ablation Study

We conduct a series of ablation studies to validate the effectiveness of each component of our method. First, we introduce the MWAC algorithm to select limited samples for human prediction actively. Second, we propose the evaluator module to transform human predictions into probability distributions. Tables 1, 2, and 3 demonstrate that both the MWAC algorithm and the evaluator module improve the accuracy of the human-AI diagnostic combination when $l \leq 8$. When the active collection strategy is replaced by random collection for obtaining human predictions, the accuracy decreases, particularly by 1.8% when $l = 3$. This result validates the effectiveness of the MWAC algorithm. Similarly, without the evaluator module and relying on the existing confusion matrix method, the accuracy drops significantly, highlighting the importance of the evaluator module.

5.4 Effect of Parameters W_s and W_l in Median-Window

We use median-window sampling to implement the MWAC algorithm. W_s represents the starting position of the window, which defines the candidate sample interval. If W_s is close to 0 or N, the window primarily selects samples with extreme or ambiguous estimates. Conversely, if W_s is close to $N/2$, the window primarily selects samples near the median of the estimated values. W_l defines the size of the window. To assess the effect of the median-window on the accuracy of ActiveHAI across the three datasets, we set $N = 100$ and adjust $W_s \in \{0, 10, 20, 30, 40, 50, 60, 70, 80, 90\}$ and $W_l \in \{5, 10, 15, 20, 25, 30\}$. The results are shown in Figure 4. The results indicate that as W_s increases, diagnostic accuracy initially improves, reaching a peak in the range of 40 to 70, and then decreases. In contrast, diagnostic accuracy declines as W_l increases. This aligns with our intuition that, under limited human predictions, selecting samples near the median of estimated values for human prediction evaluation improves distribution fitting and enhances the accuracy of human-AI collaborative diagnosis. Moreover, for different numbers of predictions l, the impact of W_s and W_l on accuracy remains consistent. This consistency suggests that the proposed median-window sampling method applies to various scenarios involving limited expert predictions and demonstrates a degree of generalization capability.

6 Conclusion

In this paper, we propose ActiveHAI, a human-AI diagnostic combination method designed to improve diagnostic accuracy with limited expert predictions. Through median-window active collection, we efficiently select human predictions, and our evaluator module enhances the evaluation of human predictions by integrating expert predictions with sample features. Experimental results on three real-world datasets demonstrate that ActiveHAI outperforms individual doctor performance and other human-AI collaboration methods. This method provides a promising solution for scenarios with high expert prediction costs and limited human resources, enhancing the potential of human-AI diagnosis.

Ethical Statement

We train a model in medical diagnosis to assess expert prediction probability distributions, aiming to improve the accuracy of human-AI diagnostic combinations. A broader contribution of this work is the proposed active collection strategy that

selects limited representative samples to evaluate the probability distributions of doctors' decisions efficiently. The approach could theoretically be used to estimate the capabilities of individual expert doctors in disease diagnosis. However, real-world deployments must undergo training on larger-scale real-world data for ethical considerations. AI models learned from insufficient and incomplete medical datasets may pose considerable prediction risks. It is important to note that the public datasets used in our experiments have had all patient privacy-related information meticulously removed.

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

[Alves *et al.*, 2024] 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. *arXiv preprint arXiv:2403.06906*, 2024.

[Alzubaidi *et al.*, 2021] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J Humaidi, Omran Al-Shamma, Mohammed A Fadhel, Jinglan Zhang, Jesus Santamaría, and Ye Duan. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13(7):1590, 2021.

[Bansal *et al.*, 2021] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11405–11414, 2021.

[Budd *et al.*, 2021] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical image analysis*, 71:102062, 2021.

[Chae and Kim, 2023] Jinyeong Chae and Jihie Kim. An investigation of transfer learning approaches to overcome limited labeled data in medical image analysis. *Applied Sciences*, 13(15):8671, 2023.

[Chen *et al.*, 2023] Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817, 2023.

[Dvijotham *et al.*, 2023] Krishnamurthy Dvijotham, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Robert Stanforth, Nick Pawlowski, Patricia Strachan, Zahra Ahmed, Shekoofeh Azizi, Yoram Bachrach, et al. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29(7):1814–1820, 2023.

[Fragiadakis *et al.*, 2024] George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*, 2024.

[Groh *et al.*, 2022] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, 2022.

[Gu *et al.*, 2023] Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang'Anthony' Chen. Augmenting pathologists with navipath: design and evaluation of a human-ai collaborative navigation system. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[Gupta *et al.*, 2023] Sumeet Gupta, Shweta Jain, Shashi Shekhar Jha, Pao-Ann Hsiung, and Ming-Hung Wang. Take expert advice judiciously: Combining groupwise calibrated model probabilities with expert predictions. In *ECAI 2023*, pages 956–963. IOS Press, 2023.

[Hemmer *et al.*, 2022] 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 *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, page 2478, 2022.

[Hemmer *et al.*, 2023] Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. Learning to defer with limited expert predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6002–6011, 2023.

[Ju *et al.*, 2022] Lie Ju, Xin Wang, Lin Wang, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, Tongliang Liu, and Zongyuan Ge. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE transactions on medical imaging*, 41(6):1533–1546, 2022.

[Kerrigan *et al.*, 2021] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34:4421–4434, 2021.

[Keswani *et al.*, 2021] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165, 2021.

[Kotia *et al.*, 2021] Jai Kotia, Adit Kotwal, Rishika Bharti, and Ramchandra Mangrulkar. Few shot learning for medical imaging. *Machine learning algorithms for industrial applications*, pages 107–132, 2021.

[Liu *et al.*, 2020] Jingya Liu, Liangliang Cao, and Yingli Tian. Deep active learning for effective pulmonary nodule

detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 609–618. Springer, 2020.

[Liu *et al.*, 2024] Jiaqi Liu, Fengming Zhang, Xin Zhang, Zhiwen Yu, Liang Wang, Yao Zhang, and Bin Guo. hmcodetrans: Human-machine interactive code translation. *IEEE Transactions on Software Engineering*, 2024.

[Madras *et al.*, 2018] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

[Mao *et al.*, 2024] Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. *Advances in neural information processing systems*, 36, 2024.

[Mozannar and Sontag, 2020] 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.

[Mozannar *et al.*, 2023] 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.

[Rajpurkar *et al.*, 2022] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

[Ribeiro *et al.*, 2020] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020.

[Singh *et al.*, 2023] Sagalpreet Singh, Shweta Jain, and Shashi Shekhar Jha. On subset selection of multiple humans to improve human-ai team accuracy. In *Proceedings of the 2023 international conference on autonomous agents and multiagent systems*, pages 317–325, 2023.

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

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[Tang *et al.*, 2023] Suigu Tang, Xiaoyuan Yu, Chak Fong Cheang, Yanyan Liang, Penghui Zhao, Hon Ho Yu, and I Cheong Choi. Transformer-based multi-task learning for classification and segmentation of gastrointestinal tract endoscopic images. *Computers in Biology and Medicine*, 157:106723, 2023.

[Topol, 2019] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[Verma and Nalisnick, 2022] 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.

[Wang *et al.*, 2024] Hui Wang, Zhiwen Yu, Yao Zhang, Yanfei Wang, Fan Yang, Liang Wang, Jiaqi Liu, and Bin Guo. hmos: An extensible platform for task-oriented human–machine computing. *IEEE Transactions on Human-Machine Systems*, 2024.

[Wilder *et al.*, 2021] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1526–1533, 2021.

[Willemink *et al.*, 2020] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

[Yu *et al.*, 2021] Zhiwen Yu, Qingyang Li, Fan Yang, and Bin Guo. Human–machine computing. *CCF Transactions on Pervasive Computing and Interaction*, 3:1–12, 2021.

[Zhang *et al.*, 2024a] Yu Zhang, Jing Chen, Xiangxun Ma, Gang Wang, Uzair Aslam Bhatti, and Mengxing Huang. Interactive medical image annotation using improved attention u-net with compound geodesic distance. *Expert systems with applications*, 237:121282, 2024.

[Zhang *et al.*, 2024b] 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, 2024.

[Zhao *et al.*, 2024] Xuehan Zhao, Jiaqi Liu, Yao Zhang, Zhiwen Yu, and Bin Guo. Haiformer: Human-ai collaboration framework for disease diagnosis via doctor-enhanced transformer. In *ECAI 2024*, pages 1495–1502. IOS Press, 2024.

[Zhu *et al.*, 2021] 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.