ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240653

HAIformer: Human-AI Collaboration Framework for Disease Diagnosis via Doctor-Enhanced Transformer

Xuehan Zhao[a], Jiaqi Liu[a,*], Yao Zhang[a], Zhiwen Yu[a] and Bin Guo[a]

[a]Northwestern Polytechnical University

Abstract. Online disease diagnosis, gathering the patients' symptoms and making diagnoses through online dialogue, grows rapidly worldwide. Manual-based approach, e.g., Haodaifu, employs real-world doctors, providing high-quality but high-cost medical services. In contrast, machine-based approach, e.g., 01bot, that utilizes machine learning models can make automatic diagnosis but lacks reliable accuracy. While some work has enabled human-AI collaboration in disease diagnosis, their collaboration pattern is simple and needs to be further improved. Therefore, we aim to introduce a doctor-enhanced and low-cost human-AI collaboration pattern. There are two key challenges. 1) How to utilize expert knowledge in doctor feedback to enhance AI's capability? 2) How to design a collaboration workflow to achieve a low-cost doctor workload while ensuring accuracy? To address the above challenges, we propose the Human-AI collaboration framework for disease diagnosis via doctor-enhanced transformer, called HAIformer. Specifically, to enhance AI's capability, we propose a machine module that leverages doctors' medical knowledge through doctor-enhanced attention, using a graph attention-based matrix; to reduce doctor workload, we propose an activation module that uses two units in a cascading manner for human-AI allocation. Experiments on four real-world datasets show that HAIformer can achieve up to 91.2% accuracy with only 18.9% human effort and one-third of dialogue turns. Further real-world clinic study highlights its advantages in practical applications.

1 Introduction

Telemedicine refers to health professionals providing medical services to patients in different locations through the use of information and communication technologies [26]. As the usage on the Web continues to grow, an increasing number of people are participating in online disease diagnosis due to its advantages, including freedom from location constraints. With the COVID-19 pandemic, online disease diagnosis services are experiencing significant growth globally [27, 16]. There are two categories of online disease diagnosis: manual-based and machine-based. Manual-based disease diagnosis, such as Haodaifu [3], American Well [2], and Teladoc Health [4], employs real-world doctors to inquire about symptoms and make the diagnosis, which provides high-quality medical services but brings expensive costs. In contrast, machine-based disease diagnosis, such as Zuoshou Doctor [5] and 01bot [1], employs machine learning models, which have seen significant development in recent years.

Given the rapid advancement of Artificial Intelligence (AI) in medical diagnosis [15, 28, 45], AI agents have shown promising potential in online medical consultation [11, 20]. For the consultation task involving both inquiry and diagnosis, there are currently two types of AI agents: Reinforcement Learning (RL)-based and Transformer-based [31] methods. RL-based models [36, 25, 39, 38, 23, 24, 43, 20, 19, 13, 18] define the dialogue-based diagnostic problem as a Markov decision process [40], interacting with patients to maximize long-term rewards. However, RL-based models often exhibit lower accuracy on large datasets due to sparse reward problems. Transformer-based models [10, 11, 21, 34] redefine the problem as a sequence generation task, achieving state-of-the-art accuracy. However, such models often cause many invalid symptom inquiries, and therefore are not very patient-friendly. More recently, Large Language Models (LLMs) have shown potential benefits for diagnostic medicine. However, their ability to accurately diagnose complex cases is still limited [30]. For instance, the text generated for specific cases exhibits ambiguity.

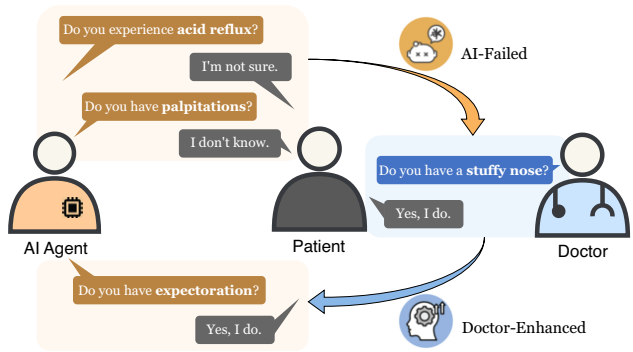


Figure 1: A medical dialogue case based on human-AI collaboration. When AI fails to ask about symptoms, the system activates doctors to conduct inquiries, which enhances AI in turn.

To solve this tough problem, we propose to adopt the Human-Machine Computing (HMC) [41] framework for online medical consultation. The HMC framework has shown its capability to address complex tasks in the medical domain. For example, the human-AI collaboration protocols enhance knee lesion detection performance beyond the capabilities of individual agents [7]; human-LLMs collaboration shows promise in enhancing brain MRI's differential diagnosis [22]; the human-AI collaborative navigation system achieves higher precision and recall in navigating tumour images [17]. Recently, some research has also applied the HMC framework in medical consultation. For example, the phased diagnosis system [12] utilizes doctors to conduct final diagnoses after machine pre-diagnosis. The diagnostic team framework [42] involves both humans and AI

* Corresponding Author. Email: jqliu@nwpu.edu.cn.

in symptom inquiries. However, existing human-AI collaboration frameworks can not explicitly learn from doctor feedback to enhance AI, nor do they define collaborative rules for the entire workflow.

The process of a dialogue-based disease diagnosis based on human-AI collaboration consists of three steps. First, patients input their self-reports to describe their medical condition, which can be summarized as specific symptoms. Second, the system acquires additional symptoms by dialoguing with patients. As shown in Figure 1, in each dialogue turn, when AI continuously fails to output valid symptoms (the symptoms confirmed/denied by the patient), encountering a challenge, expert doctors will participate in symptom inquiry. After the doctors participate, their decisions will enhance AI's performance by providing prior knowledge. Third, when sufficient symptom information is obtained, the inquiry is terminated, and the system diagnoses based on these acquired symptoms.

To achieve doctor-enhanced and low-cost human-AI collaborative consultations, there are two challenges. First, previous work mainly focuses on when to invoke humans for assistance, but lacks a study on how to utilize the expert information contained in human feedback. So, *How to better utilize doctor feedback in order to improve diagnosis accuracy?* Second, due to the high cost of doctor resources, *how to design a human-AI collaboration workflow that can effectively reduce doctor workload?* To address the above challenges, we propose two modules. (i) For the first challenge, we design a doctor-enhanced machine module that enhances the attention scores of doctor-related symptoms through a doctor matrix, thereby improving the module's performance and further promoting successful diagnosis. Additionally, we construct a unique symptom-graph for each dataset and then use a graph attention network to build an adaptive doctor matrix. (ii) For the second challenge, we design an activation module that utilizes two units in a cascaded manner to select either doctors or AI for further symptom inquiries, optimizing human-AI collaboration. On the other hand, the activation module alters the input distribution of machine diagnosis, enabling our framework to utilize only a few crucial symptoms for high-accuracy diagnosis.

In this paper, we propose a human-AI collaboration framework for disease diagnosis via doctor-enhanced Transformer. Before each dialogue turn, the proposed activation module decides whether to terminate the inquiry and give the final diagnosis based on the acquired symptoms. If the symptoms are insufficient, the activation module conducts a low-cost human-AI allocation with two units in a cascaded manner, calling the machine module or the doctor to inquire. Additionally, the machine's capability is enhanced by the expert information in doctor feedback through the proposed graph attention-based doctor matrix. Our contributions are as follows:

- **Human-AI Collaboration**: We propose a novel human-AI collaboration framework (HAIformer) for disease diagnosis, which is, to the best of our knowledge, the first attempt to make the diagnosis with high accuracy but limited inquiry turns.
- **Enhancement Design**: We propose a doctor-enhanced attention, which leverages doctors' prior knowledge in consultations to enhance AI's capability in symptom inquiry.
- **Allocation Design**: We propose a low-cost human-AI collaboration workflow, where doctors handle complex cases unsolvable by machines, thus reducing the generation of invalid symptoms.
- **Experiment Study**: The experiments on four real-world datasets show that HAIformer outperforms state-of-the-art models by 5.6% in diagnosis accuracy with one-third of inquiry turns. Additionally, we conduct real clinical experiments, in which HAIformer outperforms the AI and the manual systems. For reproducibility, we release the code and data in https://github.com/mercyzi/HAIformer.git.

2 Related Work

This paper is related to two research areas, namely, dialogue-based disease diagnosis and human-AI collaboration in medical diagnosis.

2.1 Dialogue-Based Disease Diagnosis

The medical dialogue system is divided into end-to-end and pipeline-based approaches. For end-to-end systems, during the COVID-19 pandemic, Zhou et al. [44] developed a medical dialogue system and successfully achieved generating doctor-like responses. Weng et al. [37] proposed the HoT method to guide LLMs in medical conversation question-answering, achieving accuracy levels close to those of real doctors. However, LLMs face many ethical issues in healthcare [29], such as bias and privacy concerns. For pipeline-based systems, Wei et al. [36] constructed a disease diagnosis dialogue system, in which they utilized DQN to gather patient symptoms and employed a fully connected network for disease classification. He et al. [19] further proposed a multi-model-fused RL framework to address the sparse reward problem in large search spaces. Chen et al. [10] pointed out the inefficiency of RL in decision-making and proposed for the first time a symptom sequence generation method based on Transformer. Wang et al. [35] proposed a multi-expert consultation framework using open-source LLMs, becoming the current state-of-the-art AI model. However, the additional medical knowledge and ethical issues introduced by LLMs still require further investigation.

2.2 Human-AI Collaboration in Medical Diagnosis

The researchs [6, 33] show that human-AI collaboration has broader applicability in high-risk medical decision-making. For example, Fogliato et al. [14] studied the design details of human-AI collaboration in radiology diagnosis and found that radiologists' diagnoses are more consistent with AI suggestions when AI inferences are displayed immediately. Calisto et al. [8, 9] constructed a real-world clinical doctor-AI workflow for breast cancer image classification. The research findings indicate an improvement in clinical doctors' accuracy, and the clinical field is accepting human-AI systems for breast cancer diagnosis. Online disease diagnosis has also recently started using human-AI collaboration. For example, Chen et al. [12] proposed a human-machine collaborative diagnosis system that ensures diagnosis accuracy. In the inquiry phase, AI employs a DQN-based method to inquire about symptoms; in the diagnosis phase, doctors examine AI's preliminary diagnosis and provide the final diagnosis. Zhao et al. [42] proposed a human-AI diagnostic team framework where doctors and AI collaborate in symptom inquiries through RL-based allocation strategies. This framework ensures high diagnostic accuracy while minimizing the workload for doctors. However, most frameworks involve only humans and AI collaborating on individual problems, such as classification problems, where AI provides advice for humans in the decision-making process. In this situation, AI does not fully utilize the additional knowledge from human feedback to enhance itself. In addition, it lacks a human-AI collaboration rule that runs through the entire complex workflow.

3 System Overview

In this section, we formulate the problem in human-AI collaboration and give a brief introduction to the proposed framework.

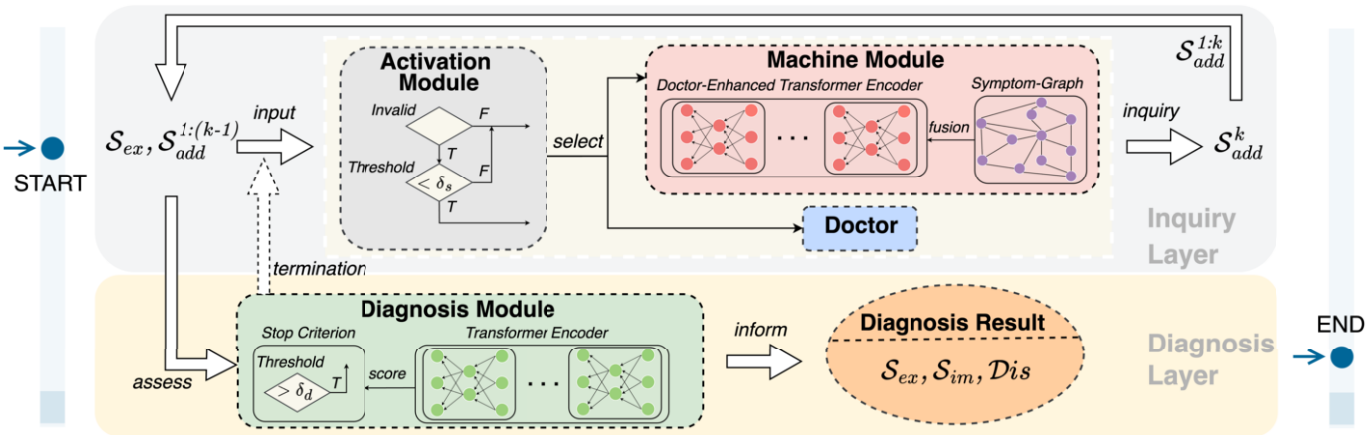


Figure 2: Human-AI Collaboration Framework for Online Disease Diagnosis via Doctor-Enhanced Transformer. The patient initiates an online consultation (left), then interacts with the inquiry layer and the diagnosis layer (central body), and finally, the system provides the diagnosis result (right). The dashed arrow (termination) indicates that if the known symptom information is sufficient, the dialogue is terminated.

3.1 Problem Formulation

The user goal for each sample consists of three parts: explicit symptoms S_{ex}, implicit symptoms S_{im}, and a disease label Dis. The online diagnosis based on human-AI collaboration typically involves two stages: the inquiry stage and the diagnosis stage. In the inquiry stage, the Human doctor and AI ask the patient for additional symptoms $S_{add} \subset S_{im}$ based on patient's self-reported symptoms, i.e., S_{ex}, aiming to gather more implicit symptoms and thereby facilitate disease diagnosis. Formally, the additional symptoms is denoted as $S_{add} = S^H_{add} \cup S^{AI}_{add}$, where $S^H_{add} = \{s^H_1, \ldots, s^H_M\}$ denotes the M symptoms inquired by the Human doctor, and $S^{AI}_{add} = \{s^{AI}_1, \ldots, s^{AI}_N\}$ are inquired by AI. We formulate the problem in the inquiry stage as a sequence generation task, with S_{im} serving as the target generated sequence. It is worth noting that, due to the involvement of doctors, the process of AI generating symptoms is partially autoregressive. Our system aims to maximize the objective function:

$$\prod_{S_{add} \subset S_{im}} P\left(S_{im} - S_{add} \mid S_{ex}, S^H_{add}, S^{AI}_{add}\right), \tag{1}$$

with a limited consumption of human effort, i.e., the number $|S^H_{add}|$ of implicit symptoms obtained by Human doctor.

In the diagnosis stage, AI conducts disease diagnosis based on the known symptoms (explicit symptoms and additional symptoms) of the patient. Essentially, disease diagnosis is a classification problem in machine learning, with the objective of maximizing the likelihood:

$$P\left(Dis \mid S_{ex}, S_{add}\right), \tag{2}$$

We note that there is a relationship between inquiry and diagnosis: inquiry serves as the foundation for diagnosis, while diagnosis is the ultimate goal of inquiry. Therefore, our framework aims to generate effective symptom sequences with limited human effort and only a few dialogue turns, and further accurately predicts diseases.

3.2 Human-AI Collaboration Framework for Online Disease Diagnosis

The framework of HAIformer is shown in Figure 2, which consists of an inquiry layer for interacting with patients and a diagnosis layer for disease classification of patients. The specific roles of these two layers in our framework are described as follows.

A) **Inquiry Layer.** The inquiry layer has a loop structure that takes explicit symptoms S_{ex} and additional symptoms $S^{1:(k-1)}_{add}$ in k-th turn as input, where $S^{1:(k-1)}_{add} = \cup_{i=1}^{k-1} S^i_{add}$ represents the symptoms queried from 1-th turn to k-1-th turn. Firstly, the symptom input goes through the activation module, which is responsible for selecting either the machine module or the doctor to inquire about symptoms from the patient. Then, the selected machine/doctor outputs an additional symptom S^k_{add} to be asked in the k-th turn, while updating the next input $S^{1:k}_{add}$ after the patient's answer. For the doctor, in experiments conducted on real datasets, we simulate expert doctors based on ground truth. In real-world experiments, we invite real doctors to evaluate our framework.

1) **Machine Module.** It utilizes the doctor-enhanced Transformer to generate symptom sequences and incorporates the Symptom-Graph. It will be detailed in Section 4.1.1 & 4.1.2.

2) **Activation Module.** It uses two units (Invalid Unit and Threshold Unit) in a serial form to determine which one to activate. It will be detailed in Section 4.1.3.

B) **Diagnosis Layer.** The diagnosis module in the diagnosis layer also plays a role in the loop structure of multi-turn dialogues: it assesses the existing symptoms before each turn of medical dialogue. If there is sufficient symptom information, the dialogue is terminated directly; otherwise, the loop continues. At the end of the conversation, the diagnosis module informs the diagnosis result, including known symptoms and the predicted disease.

1) **Diagnosis Module.** It utilizes the Transformer to predict diseases and employs a stop criterion to determine whether to terminate the dialogue. It will be detailed in Section 4.2.

4 Methodology

In this section, we describe the implementation methodology of the inquiry layer and diagnosis layer in HAIformer.

4.1 Implementation of Inquiry Layer

In this subsection, we first illustrate the implementation of the machine module with doctor-enhanced Transformer. Then, we introduce the doctor matrix with graph learning, which is used to enhance doctor-related symptoms. Finally, we describe the activation module.

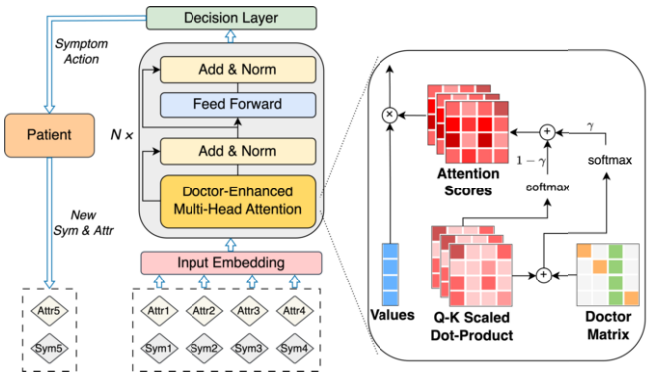


Figure 3: The implementation details of the machine module.

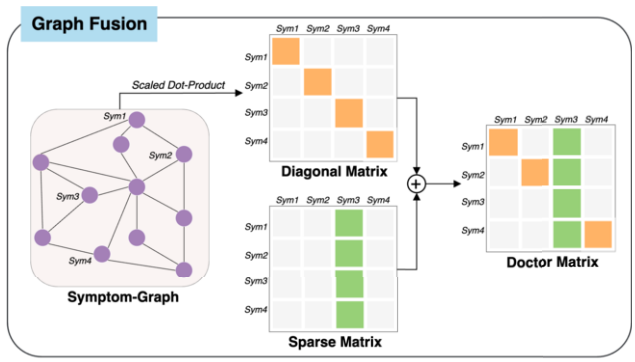


Figure 4: The process of generating a doctor matrix, where "Sym3" is the symptom inquired by the doctor. The doctor matrix is filled with green or orange elements from sparse and diagonal matrices.

4.1.1 *Machine Module with Doctor-Enhanced Transformer*

Figure 3 shows the overview of our proposed machine module. The backbone is a stack of Transformer blocks with N layers. The known symptoms and their attributes are encoded using an embedding layer to create contextual representations. Subsequently, these representations are continuously updated through multiple layers of Transformer blocks, and finally, a symptom action is generated as output via the decision layer. The symptom action taken by the machine module will fill the "symptom" slot, generating templated natural language for dialogue with the patient, for example, "Do you have [runny nose]?" where "runny nose" is the symptom. For each symptom inquiry, the patient will respond True or False as attributes of positive/negative symptoms, and if the symptom does not belong to the implicit symptoms S_{im}, the response will be UNK. After the patient's response, the new additional symptom and its attribute will be added to the input for the next turn.

The input embedding $\boldsymbol{X}_0$ of known symptoms is the sum of these symptom embeddings and attribute embeddings:

$$\boldsymbol{X}_0 = Embedding\left(S_{ex}, S_{add}\right) + Embedding\left(A_{ex}, A_{add}\right), \quad (3)$$

where A_{ex} and A_{add} represent the attributes of explicit symptoms and additional symptoms, respectively. There are three symptom attributes: True, False, and UNK, reflecting the relationship between the symptom and the patient. The embedding layer maps the symptom and the attribute input sequences to vectors with the same dimension, thereby obtaining the initial contextual representation $\boldsymbol{X}_0 \in \mathbb{R}^{l\times d}$, where l denotes the sequence length, and d represents the dimensionality of token embedding vectors. We then use doctor-enhanced Transformer layers to generate contextual representations:

$$\boldsymbol{X}_n = Transformer_{doctor-enhanced}\left(\boldsymbol{X}_{n-1}\right), \quad (4)$$

where $n \in [1, N]$ indicates that the variables X are at the n-th layer. Each Transformer layer applies doctor-enhanced multi-head self-attention operation, followed by a fully connected feed-forward network. Multi-head attention enables the model to jointly attend to information across different dimensions, thereby enhancing its generalization capability:

$$Multihead\left(Q, K, V\right) = Concat\left(head_1, \ldots head_h\right)\boldsymbol{W}^O, \quad (5)$$

where h represents the number of heads, $\boldsymbol{W}^O \in \mathbb{R}^{d\times d}$ is the parameter matrix to be learned.

Specifically, the self-attention operation first applies linear transformations W_i^q, W_i^k, and $W_i^v \in \mathbb{R}^{d\times d_h}$ on the contextual representation to obtain a triplet consisting of the query Q_i, the key K_i, and the value $V_i \in \mathbb{R}^{l\times d_h}$ for each head, where d_h is the dimensionality of each head. Then, the global attention score S_i^{glb} and the doctor-enhanced attention score S_i^{doc} are calculated as follows:

$$\boldsymbol{S}_i^{glb} = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_h}}\right), \quad (6)$$

$$\boldsymbol{S}_i^{doc} = softmax\left(\frac{Q_i {K_i}^T}{\sqrt{d_h}} + M^{doc}\right), \quad (7)$$

where $M^{doc} \in \mathbb{R}^{l\times l}$ is an attention matrix that emphasizes the symptoms asked by doctors, which will be introduced later. Finally, we integrate global and doctor-enhanced attentions to calculate each head:

$$head_i = \left(\gamma S_i^{doc} + (1-\gamma)\, S_i^{glb}\right) V_i, \quad (8)$$

where i denotes the i-th head of the multi-head operation and $\gamma \in [0, 1]$ is a hyper-parameter used to adjust the weight of the doctor-enhanced attention score S_i^{doc}. A larger γ value indicates a greater emphasis on symptoms asked by the doctor.

After obtaining a unique token vector from the N-th layer, we further utilize the decision layer to output symptom actions by the linear transformation $W^A \in \mathbb{R}^{d\times n_s}$, where n_s represents the number of symptom categories. The machine module's training objective is multi-label (i.e., $S_{im} - S_{add}$), so we can use RL terminology to describe the learning process, similar to Chen et al. [11]. When the machine module outputs a symptom, if the symptom is valid (True/False) and the pre-diagnosis is correct, the reward is $R_{pos} = 5.0$; otherwise, the reward is $R_{neg} = -0.2$. Therefore, we train the machine module by minimizing the loss function $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} R_i P_\theta(s_i^{AI}), \quad (9)$$

where N represents the total number of symptoms queried by the machine module, θ denotes the trainable model parameters, $P_\theta(s_i^{AI})$ signifies the conditional probability of the machine module with parameters θ taking the i-th symptom action s_i^{AI}, and R_i represents the reward returned by the i-th symptom action. Before training the machine module, we initialize the model parameters θ through self-supervised learning with maximum likelihood objectives.

4.1.2 *Doctor Matrix with Graph Learning*

As illustrated in Figure 4, we learn a doctor matrix M^{doc} to incorporate doctors' prior knowledge into the Transformer, enabling the

machine module to generate symptoms more relevant to doctors. The motivation is to enhance the attention scores for doctor-related symptoms, allowing attention heads to focus on more valuable features.

To learn this matrix M^{doc}, we first construct a unique symptom-graph for each dataset and pretrain it by neighbors masking. Then, we query the node feature $g \in \mathbb{R}^d$ of the doctor symptom s^H from the symptom-graph, perform scaled dot-product with the context representation X, and then fill the resulting diagonal vector $V^{dia} \in \mathbb{R}^l$ into a diagonal matrix $M^{dia} \in \mathbb{R}^{l \times l}$:

$$\boldsymbol{V}^{dia} = softmax\left(\frac{g\boldsymbol{W}^g X^T}{\sqrt{d}}\right), \tag{10}$$

$$\boldsymbol{M}_{ij}^{dia} = \begin{cases} V_i^{dia}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}, \tag{11}$$

where $W^g \in \mathbb{R}^{d \times d}$ is the trainable parameter matrix. Additionally, since the self-attention operation requires normalization of row vectors through the softmax function, to enhance the weight of doctor symptoms, we also construct a sparse matrix $M^{spa} \in \mathbb{R}^{l \times l}$:

$$\boldsymbol{M}_{ij}^{spa} = \begin{cases} 1, & \text{if } s_j \in s^H \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where s_j represents the j-th symptom in the symptom input sequence. By combining M_{ij}^{dia} and M_{ij}^{spa}, we obtain the doctor matrix M^{doc} to enhance symptoms related to doctors:

$$\boldsymbol{M}_{ij}^{doc} = \begin{cases} M_{ij}^{spa}, & \text{if } s_j \in s^H \\ M_{ij}^{spa} + M_{ij}^{dia}, & \text{otherwise} \end{cases}, \tag{13}$$

For each directed symptom-graph corresponding to the dataset, each node corresponds to a symptom in the dataset, and edges represent the relational information between symptoms. We utilize GAT [32] to aggregate the relationship information among symptoms, extracting the structural feature g_i for each node i:

$$\boldsymbol{g}_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \boldsymbol{W}_1 h_j\right), \tag{14}$$

$$\boldsymbol{\alpha}_{ij} = \frac{\exp\left(LeakyReLU\left(\boldsymbol{W}_2[h_i; h_j]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp(LeakyReLU(\boldsymbol{W}_2[h_i; h_k]))}, \tag{15}$$

where W_1 and W_2 are the trainable parameters. h_j is the representation of symptom j, $\mathcal{N}_i$ represents the neighbors of symptom i, and σ is the sigmoid function. α_{ij} is the attention coefficient computed by the attention mechanism applying the LeakyReLU activation.

4.1.3 Activation Module

In online medical consultations, doctors possess superior capabilities to machines regarding symptom accuracy. To better utilize limited medical resources and minimize ineffective queries by AI, we design the following human-AI collaboration workflow: when the machine module struggles to generate relevant symptoms effectively, the activation module prompts doctors to query the symptoms. Specifically, based on the machine module's output, the activation module decides whether to involve doctors in the next turn using two units: 1) Invalid Unit: During each turn, if the machine module consecutively generates $N_{IS} = 2$ invalid symptoms, it transitions to the next unit for assessment; otherwise, the machine module continues its inquiry. 2) Threshold Unit: At the beginning of a new turn, if the confidence level of newly generated symptoms by the machine module falls below the threshold δ_s, it prompts the activation module to engage doctors; otherwise, the machine module continues its inquiry.

Algorithm 1 Training process mechanism of HAIformer

```
Construct a symptom-graph for each dataset
Pretrain symptom-graph by neighbors masking
Pretrain machine module θ_m by self-supervised learning
Pretrain diagnosis module θ_d using the original sample
for i = 1, 2, ..., N_1 do
    Machine module takes symptom actions s^a
    Environment generates reward feedback R
    Update θ_m by minimizing L(θ) = −Σ_{i=1}^{N} R_i P_θ(s_i^{AI})
    for n = 1, 2, ..., N_2 do
        Get additional symptoms via inquires
        Fine-tuning θ_d by optimizing cross-entropy loss
    end for
    θ_d ← Pretrained diagnosis module θ_d
end for
```

Furthermore, we stipulate that each time a doctor is activated, they only handle one symptom inquiry task. To prevent long intervals between multiple calls to doctors, we set a limit that doctors will not be called after the turn λ_d. Additionally, the inquiry layer also has a maximum dialogue turn limit λ_{max}; if the dialogue turn exceeds λ_{max}, the dialogue ends and the diagnosis result is informed.

4.2 Implementation of Diagnosis Layer

The backbone of the diagnosis module is also composed of stacked encoder layers of the Transformer. After obtaining the sequence representation from the N-th layer, we aggregate it into a single representation using an average pooling operation.

$$AvgPooling(x) = \frac{1}{l}\sum_{i=1}^{l} x_i \tag{16}$$

where l represents the length of the symptom sequence, and x_i denotes the i-th symptom vector in the symptom sequence. Then, the single representation is mapped to the output space required for the disease classification task using a fully connected layer $W^d \in \mathbb{R}^{d \times n_d}$, where n_d represents the number of disease categories. The training loss of the diagnosis module is calculated based on the cross-entropy loss between the predicted probabilities and the true labels.

In a multi-turn dialogue system, we need to balance the number of dialogue turns and the accuracy of disease diagnosis. Specifically, the diagnosis module makes a preliminary diagnosis based on the known symptoms. If the probability of disease prediction exceeds the threshold δ_d, the diagnosis module terminates the dialogue and informs the diagnosis result to the patient; otherwise, the dialogue continues. In summary, the training process of our proposed framework is presented in Algorithm 1.

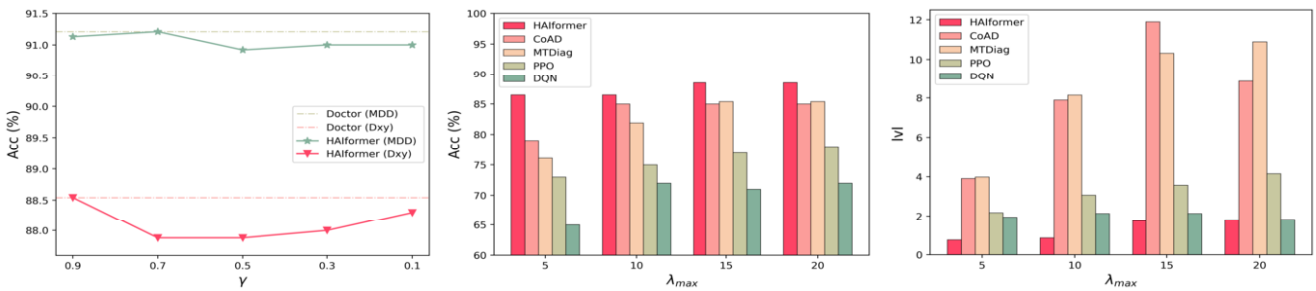
5 Experiments

5.1 Settings

Datasets. To validate the applicability of our proposed framework in real-world scenarios, we conduct experiments on four publicly available real datasets instead of using the synthetic dataset [43]. The four datasets, MZ-4 [36], Dxy [39], MZ-10 [11], and MDD [21], all contain explicit symptoms, implicit symptoms, and a disease label, covering 4, 5, 10, and 12 disease types, respectively. The MZ-4, Dxy, and MDD datasets are medical diagnosis dialogue datasets, while the MZ-10 dataset is a corpus with multi-level fine-grained annotations.

Table 1: Experimental results on four real-world datasets.

Model	MZ-4 dataset			Dxy dataset			MZ-10 dataset			MDD dataset		
	Acc	Turn	HE	Acc	Turn	HE	Acc	Turn	HE	Acc	Turn	HE
PPO [24]	73.2	6.3	–	74.6	3.3	–	–	–	–	–	–	–
Diaformer [10]	74.2	15.3	–	82.9	13.1	–	59.5	14.5	–	86.0	18.9	–
DxFormer [11]	73.2	20.0	–	83.7	20.0	–	66.6	20.0	–	86.2	20.0	–
CoAD [34]	75.0	13.4	–	85.0	10.5	–	62.0	19.9	–	84.9	18.3	–
MTDiag [21]	75.9	17.9	–	85.4	12.5	–	–	–	–	89.1	13.8	–
Doctor	78.5	3.2	3.19	88.5	1.8	1.76	70.3	5.4	5.35	91.2	2.6	2.64
Doctor-SC	78.5	0.8	0.80	88.5	0.7	0.68	70.3	3.1	3.09	91.2	0.5	0.53
HAIformer	**78.5**	1.4	**0.14**	**88.5**	2.3	**0.23**	**70.3**	6.6	**0.86**	**91.2**	0.8	**0.10**
w/o δ_s	78.5	1.4	0.19	88.5	2.3	0.24	70.3	6.6	0.94	91.1	0.8	0.11
w/o δ_d	78.5	9.3	1.37	88.5	9.3	0.79	70.3	12.0	1.84	91.2	8.2	0.70
w/o M^{doc}	77.9	1.4	0.14	87.9	2.3	0.23	69.8	6.6	0.86	90.8	0.8	0.10



(a) Effect of Weight γ on Acc

(b) Effect of Max Turn λ_{max} on Acc (left) and IvI (right)

Figure 5: Impact of two hyperparameters.

Baselines. We compare our framework with several baseline models, including the models involving only machines and the models involving only doctors. **DQN** [36], **PPO** [24], **Diaformer** [10], **DxFormer** [11], **CoAD** [34], and **MTDiag** [21] are state-of-the-art models involving only machines. **Doctor** and **Doctor-SC** are two models involving only doctors. The former ends the conversation when all implicit symptoms have been asked, while the latter ends the conversation based on the stop criterion.

Metrics. Our experiments mainly utilize three metrics: Diagnosis Accuracy (Acc), Dialogue Turn (Turn), and Human Effort (HE). **Acc** is the accuracy of disease diagnosis. **Turn** is the average number of dialogue turns. **HE** is specifically used to assess human resource cost, which equals the number of doctor inquiries. Additionally, to assess the system's user-friendliness for patients, we use the Invalid Inquiry (**IvI**) metric, which equals the number of invalid inquiries.

5.2 *Overall Performance*

In Table 1, we report the performance of HAIformer and baseline models on four real-world datasets. Our framework achieves higher accuracy on four datasets than the best model involving only machines by 3.4%, 3.6%, 5.6%, and 2.4%, respectively, indicating that HAIformer can provide higher-quality healthcare services. Additionally, our framework reduces dialogue turns by 92.2%, 81.6%, 67.0%, and 94.2%, significantly reducing the number of interactions with patients and avoiding unnecessary inquiries. Our framework is the same accurate with models involving only doctors, and the reduction in human effort is 82.5%, 66.2%, 72.2%, and 81.1%, demonstrating that HAIformer optimizes human-machine collaboration rather than overusing doctor inquiries. It is worth noting that Doctor-SC's dialogue turns are fewer than HAIformer's because the doctors' performance is superior to the machines.

5.3 *Ablation Study*

To validate the effectiveness of each component in HAIformer, we conduct some ablation experiments to analyze the impact of δ_s, δ_d, and M^{doc}. The role of the confidence threshold δ_s in the activation module is to optimize human-machine resource allocation. In Table 1, we observe that δ_s reduced human effort by 26.3%, 4.2%, 8.5%, and 9.1% on the MZ-4, Dxy, MZ-10, and MDD datasets, respectively. The stop criterion in the diagnosis module is used to end the dialogue, and δ_d can timely detect whether the symptom information is sufficient, significantly reducing dialogue turns and human effort. The doctor matrix M^{doc} is used to enhance the symptom attention scores of doctors, improving disease diagnosis accuracy because doctors with medical knowledge can compensate for AI's instability.

5.4 *Impact of two hyperparameters*

Effect of Weight γ. While computing attention heads, we use the weight γ to control the proportion of doctor-enhanced attention scores. As shown in Figure 5(a), we explore the impact of different γ values (ranging from 0.9 to 0.1) on the accuracy of our framework on the Dxy and MDD datasets. We observe that on the MDD dataset, Acc first increases and then decreases, reaching its maximum at γ = 0.7. In contrast, on the Dxy dataset, Acc first decreases and then increases, reaching its maximum at γ = 0.9. Additionally, adopting an appropriate weight γ can enable HAIformer to achieve the same accuracy as Doctor, indicating that proper emphasis on doctor-related symptoms can enhance the performance of the machine module.

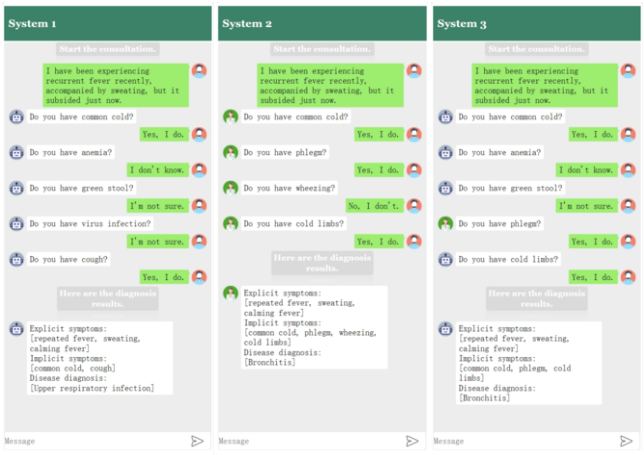
Effect of Max Turn λ_{max}. As shown in Figure 5(b), we conduct experiments on the Dxy dataset with different max turns (i.e., 5/10/15/20). HAIformer's Acc at different λ_{max} outperform Transformer-based models (CoAD, MTDiag) and RL-based models

Table 2: Real-world experimental results under the conditions of four testing systems.

Model	Acc	Turn	HE
CoAD	77.5	10.0	–
HAIformer	82.5	2.0	0.4
RealDoc	91.3	3.9	3.9
HAI-Doc	91.3	4.2	2.6

Table 3: Summary of doctors' questionnaire. For Q6, 7=Very strongly prefer HAIformer over RealDoc

ID	Question	HAIformer	RealDoc
Q1	Is the interface of this system easy to operate?	5.0(sd=1.0)	4.5(sd=0.5)
Q2	How do you feel about the mental pressure when handling tasks?	2.0(sd=1.0)	3.5(sd=1.5)
Q3	Can the system reduce doctor's workload?	5.0(sd=1.0)	N/A
Q4	Can the system help you screen patients' symptoms?	6.0(sd=0.0)	N/A
Q5	Would you like to use the system in the future?	6.5(sd=0.5)	N/A
Q6	Overall Preference	6.5(sd=0.5)	



(a) CoAD (AI) (b) RealDoc (Human) (c) HAIformer

Figure 6: Interface for online consultation system. The experiments on the system 4 (HAI-Doc) are conducted on the HAIformer system.

(PPO, DQN), especially at λ_{max} of 5. This indicates that HAIformer can maintain high accuracy even with fewer max turns. Overall, as λ_{max} increases, each model's Invalid Inquiry (IvI) also increases. We observe that HAIformer's IvI is significantly better than that of Transformer-based models and RL-based models. Further analysis reveals that HAIformer can achieve high diagnosis accuracy at the cost of fewer invalid inquiries, which is patient-friendly.

5.5 Clinical Evaluation

We recruit two experienced clinical doctors and forty volunteer patients who have experienced bronchitis, dyspepsia, diarrhea, or upper respiratory infection. As shown in the figure 6, we design four testing systems to support clinical evaluation:

- **CoAD**: the system involving only machines;
- **RealDoc**: the system involving only real-world doctors;
- **HAIformer**: the system using our human-AI framework;
- **HAI-Doc**: the system involving real-world doctors conducting final consultations after the diagnosis made by HAIformer;

Each doctor participates in two studies: the first doctor sequentially uses RealDoc and HAI-Doc to provide consultation services to patients, while the second doctor sequentially uses HAI-Doc and RealDoc. The average duration of each study is approximately 65 minutes. Patients also use CoAD for additional testing. During the evaluation process, we record three metrics (Acc, Turn, HE) for four systems (HAIformer results based on HAI-Doc before the final doctor consultation). Finally, doctors provide feedback by completing a questionnaire of six seven-scaled Likert questions. In order to avoid bias, we refer to CoAD, RealDoc, and HAIformer as "System 1", "System 2", and "System 3".

Table 2 presents the experimental results of four systems. Compared to CoAD, HAIformer only requires 20% of Turn and HE of 0.4, leading to a 6.5% improvement in Acc. This demonstrates that HAIformer is more suitable for clinical settings compared to models involving only AI. When compared to RealDoc, HAI-Doc only uses 67% of HE. This implies that our framework has the potential to be integrated into the actual process of doctor consultations, as it saves human resources without compromising accuracy.

As shown in Table 3, compared to RealDoc, HAIformer significantly reduces the mental pressure on doctors during consultations while also being more accessible to operation. Additionally, doctors indicate that HAIformer has partially alleviated their workload and helped them screen symptoms. Finally, doctors express a higher likelihood of using HAIformer in the future. Overall, doctors prefer our proposed human-AI collaboration system over manual systems.

6 Conclusion

In this paper, we collaborate AI with expert doctors to inquire about symptoms, combining doctors' prior knowledge to enhance AI for online disease diagnosis tasks. We introduce a human-AI collaboration framework called HAIformer, which utilizes doctor feedback to improve diagnosis accuracy with a limited doctor workload. After conducting experiments on four public real-world datasets, our results show that our framework has higher accuracy than AI models and significantly reduces the number of dialogue turns required during diagnosis. More importantly, our human subject experiments show that, doctors prefer our framework compared to manual methods and think that HAIformer reduces their mental pressure, highlighting its advantages in practical applications.

6.1 Ethical Statement

All information related to patient privacy in this dataset has been meticulously eliminated. Furthermore, the datasets have undergone a comprehensive manual review to confirm that they contain no identifiable or offensive pieces of information. In our real-world experiments, volunteer patients were informed of the simulated environment, and both patient and doctor identities were anonymized during the online system interactions. The real-world deployment of AI in medical diagnosis indeed raises ethical concerns. Although our framework achieves promising results, the AI errors caused by inadequate data may bring potential harm to patients when directly applying the method as a diagnostic system.

Acknowledgements

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, 61960206008), and the National Science Fund for Distinguished Young Scholars (No.62025205).

References

[1] 01bot. https://01bot.baidu.com/.
[2] Amwell. https://patients.amwell.com/.
[3] Haodf. https://www.haodf.com/.
[4] Teladochealth. https://www.teladochealth.com/.
[5] Zuoshouyisheng. https://open.zuoshouyisheng.com/.
[6] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.
[7] F. Cabitza, A. Campagner, and L. M. Sconfienza. Studying human-ai collaboration protocols: the case of the kasparov's law in radiological double reading. *Health information science and systems*, 9:1–20, 2021.
[8] F. M. Calisto, C. Santiago, N. Nunes, and J. C. Nascimento. Introduction of human-centric ai assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies*, 150:102607, 2021.
[9] F. M. Calisto, C. Santiago, N. Nunes, and J. C. Nascimento. Breastscreening-ai: Evaluating medical intelligent agents for human-ai interactions. *Artificial Intelligence in Medicine*, 127:102285, 2022.
[10] J. Chen, D. Li, Q. Chen, W. Zhou, and X. Liu. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4432–4440, 2022.
[11] W. Chen, C. Zhong, J. Peng, and Z. Wei. Dxformer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations. *Bioinformatics*, 39(1):btac744, 2023.
[12] Y. Chen, J. Liu, Z. Yu, H. Wang, L. Wang, and B. Guo. Hm-mds: A human-machine collaboration based online medical diagnosis system. In *2022 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2384–2389, 2022.
[13] A. Fansi Tchango, R. Goel, J. Martel, Z. Wen, G. Marceau Caron, and J. Ghosn. Towards trustworthy automatic diagnosis systems by emulating doctors' reasoning with deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24502–24515, 2022.
[14] R. Fogliato, S. Chappidi, M. Lungren, P. Fisher, D. Wilson, M. Fitzke, M. Parkinson, E. Horvitz, K. Inkpen, and B. Nushi. Who goes first? influences of human-ai workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1362–1374, 2022.
[15] H. Fujita. Ai-based computer-aided diagnosis (ai-cad): the latest review to read first. *Radiological physics and technology*, 13(1):6–19, 2020.
[16] S. Goyal, S. Chauhan, and P. Gupta. Users' response toward online doctor consultation platforms: Sor approach. *Management decision*, 60 (7):1990–2018, 2022.
[17] H. Gu, C. Yang, M. Haeri, J. Wang, S. Tang, W. Yan, S. He, C. K. Williams, S. Magaki, and X. 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.
[18] S. Guo, K. Liu, P. Wang, W. Dai, Y. Du, Y. Zhou, and W. Cui. Rdkg: A reinforcement learning framework for disease diagnosis on knowledge graph. In *2023 IEEE International Conference on Data Mining*, pages 1049–1054, 2023.
[19] W. He and T. Chen. Scalable online disease diagnosis via multi-model-fused actor-critic reinforcement learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4695–4703, 2022.
[20] W. He, X. Mao, C. Ma, Y. Huang, J. M. Hernàndez-Lobato, and T. Chen. Bsoda: a bipartite scalable framework for online disease diagnosis. In *Proceedings of the ACM Web Conference 2022*, pages 2511–2521, 2022.
[21] Z. Hou, Y. Cen, Z. Liu, D. Wu, B. Wang, X. Li, L. Hong, and J. Tang. Mtdiag: an effective multi-task framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14241–14248, 2023.
[22] S. H. Kim, S. Schramm, C. Berberich, E. Rosenkranz, L. Schmitzer, K. Serguen, C. Klenk, N. Lenhart, C. Zimmer, B. Wiestler, et al. Human-ai collaboration in large language model-assisted brain mri differential diagnosis: A usability study. *medRxiv*, pages 2024–02, 2024.
[23] J. Lin, K. Wang, Z. Chen, X. Liang, and L. Lin. Towards causality-aware inferring: A sequential discriminative approach for medical automatic diagnosis. *arXiv preprint arXiv:2003.06534*, 2020.
[24] Z. Liu, Y. Li, X. Sun, F. Wang, G. Hu, and G. Xie. Dialogue based disease screening through domain customized reinforcement learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1120–1128, 2021.
[25] Y.-S. Peng, K.-F. Tang, H.-T. Lin, and E. Chang. Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. *Advances in neural information processing systems*, 31, 2018.
[26] V. L. Raposo. Telemedicine: The legal framework (or the lack of it) in europe. *GMS health technology assessment*, 12, 2016.
[27] E. Richardson, D. Aissat, G. A. Williams, N. Fahy, et al. Keeping what works: remote consultations during the covid-19 pandemic. *Eurohealth*, 26(2):73–76, 2020.
[28] Ł. Struski, D. Rymarczyk, A. Lewicki, R. Sabiniewicz, J. Tabor, and B. Zieliński. Promil: Probabilistic multiple instance learning for medical imaging. In *ECAI 2023*, pages 2210–2217. IOS Press, 2023.
[29] K. TSIMA. The reproducibility issues that haunt health-care ai. *Nature*, 613, 2023.
[30] E. Ullah, A. Parwani, M. M. Baig, and R. Singh. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic Pathology*, 19(1):1–9, 2024.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
[32] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
[33] D. Wang, L. Wang, Z. Zhang, D. Wang, H. Zhu, Y. Gao, X. Fan, and F. Tian. "brilliant ai doctor" in rural clinics: challenges in ai-powered clinical decision support system deployment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–18, 2021.
[34] H. Wang, W. C. Kwan, K.-F. Wong, and Y. Zheng. Coad: Automatic diagnosis through symptom and disease collaborative generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 6348–6361, 2023.
[35] H. Wang, S. Zhao, Z. Qiang, N. Xi, B. Qin, and T. Liu. Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107*, 2024.
[36] Z. Wei, Q. Liu, B. Peng, H. Tou, T. Chen, X.-J. Huang, K.-F. Wong, and X. Dai. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 201–207, 2018.
[37] Y. Weng, B. Li, F. Xia, M. Zhu, B. Sun, S. He, K. Liu, and J. Zhao. Large language models need holistically thought in medical conversational qa. *arXiv preprint arXiv:2305.05410*, 2023.
[38] Y. Xia, J. Zhou, Z. Shi, C. Lu, and H. Huang. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1062–1069, 2020.
[39] L. Xu, Q. Zhou, K. Gong, X. Liang, J. Tang, and L. Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7346–7353, 2019.
[40] S. Young, M. Gašić, B. Thomson, and J. D. Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.
[41] Z. Yu, Q. Li, F. Yang, and B. Guo. Human–machine computing. *CCF Transactions on Pervasive Computing and Interaction*, 3:1–12, 2021.
[42] X. Zhao, J. Liu, Z. Yu, and B. Guo. Hadt: Human-ai diagnostic team via hierarchical reinforcement learning. In *Proceedings of the 2024 SIAM International Conference on Data Mining*, pages 860–868, 2024.
[43] C. Zhong, K. Liao, W. Chen, Q. Liu, B. Peng, X. Huang, J. Peng, and Z. Wei. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*, pages 3995–4001, 2022.
[44] M. Zhou, Z. Li, B. Tan, G. Zeng, W. Yang, X. He, Z. Ju, S. Chakravorty, S. Chen, X. Yang, et al. On the generation of medical dialogs for covid-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.
[45] L. Zhu, L. L. Chan, T. K. Ng, M. Zhang, and B. C. Ooi. Deep co-training for cross-modality medical image segmentation. In *ECAI 2023*, pages 3140–3147. IOS Press, 2023.