Scalable Online Disease Diagnosis via Multi-Model-Fused Actor-Critic Reinforcement Learning

Weijie He
Department of Computer Science and Technology &
Institute of Artificial Intelligence & BNRist,
Tsinghua University
Beijing, China
hwj19@mails.tsinghua.edu.cn

Ting Chen∗
Department of Computer Science and Technology &
Institute of Artificial Intelligence & BNRist,
Tsinghua University
Beijing, China
tingchen@tsinghua.edu.cn

ABSTRACT
For those seeking healthcare advice online, AI-based dialogue agents capable of interacting with patients to perform automatic disease diagnosis are a viable option. This application necessitates efficient inquiry of relevant disease symptoms in order to make accurate diagnosis recommendations. This can be formulated as a problem of sequential feature (symptom) selection and classification for which reinforcement learning (RL) approaches have been proposed as a natural solution. They perform well when the feature space is small, that is, the number of symptoms and diagnosable disease categories is limited, but they frequently fail in assignments with a large number of features. To address this challenge, we propose a Multi-Model-Fused Actor-Critic (MMF-AC) RL framework that consists of a generative actor network and a diagnostic critic network. The actor incorporates a Variational AutoEncoder (VAE) to model the uncertainty induced by partial observations of features, thereby facilitating in making appropriate inquiries. In the critic network, a supervised diagnosis model for disease predictions is involved to precisely estimate the state-value function. Furthermore, inspired by the medical concept of differential diagnosis, we combine the generative and diagnosis models to create a novel reward shaping mechanism to address the sparse reward problem in large search spaces. We conduct extensive experiments on both synthetic and real-world datasets for empirical evaluations. The results demonstrate that our approach outperforms state-of-the-art methods in terms of diagnostic accuracy and interaction efficiency while also being more effectively scalable to large search spaces. Besides, our method is adaptable to both categorical and continuous features, making it ideal for online applications.

CCS CONCEPTS
• Applied computing → Health informatics; • Information systems → Web applications.

KEYWORDS
online disease diagnosis; self-diagnosis; reinforcement learning
∗Corresponding author.

ACM Reference Format:
Weijie He and Ting Chen. 2022. 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 (KDD ’22), August 14–18, 2022, Washington, DC, USA. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3542672

1 INTRODUCTION
Recent advances in machine learning have facilitated the development of an alternative “AI doctor”, that is, the ability to develop a disease diagnosis model empirically from growing collections of clinical observations. Indeed, automatic disease diagnosis has the potential to benefit both doctors and patients in today’s medical system. Primary care physicians, for example, who are responsible for coordinating patient care among specialists and other levels of care, regularly encounter a variety of illnesses that should be assigned to different departments. In this situation, they may use the automatic disease diagnosis tool to guide what healthcare treatments patients require. Furthermore, the rapid growth of the mobile web has increased customer access to online self-diagnosis. Numerous self-diagnosis tools, such as the well-known symptom checkers from Mayo Clinic [1] and WebMD [3], have been developed to address these needs. Adults are increasingly turning to these online services for advice on their primary concerns prior to deciding whether to see the primary care physicians or which specialists to visit [8]. It is important that online disease diagnosis technologies are not a substitute for medical investigations by a doctor. Physical examinations and diagnostic tests such as X-rays and laboratory tests, which are crucial in the diagnosis of many diseases, could not be done online. However, there is indeed a genuine need for an efficient and accurate online “AI doctor” in healthcare [7].

An “AI doctor” would imitate the diagnosis procedure of human doctors by gathering patients’ symptoms through dialogues until it could confidently make a diagnosis. Naturally, more information regarding possible symptoms could lead to a more accurate diagnosis; for example, most existing questionnaire systems inquire about a considerable number of symptoms from patients [23, 38]. Unfortunately, patients may lose patience during the inquiry process due to the inefficiency of such systems. To overcome this problem, dialogue-based diagnosis agents have been proposed to reduce the number of inquiries while achieving high diagnostic accuracy [39]. Dialogue-based disease diagnosis can be formulated as a cost-sensitive sequential feature selection and classification task in machine learning. Inquiring a symptom to obtain its value is comparable to selecting a feature that comes with a cost. Reinforcement
learning (RL) approaches have demonstrated good performance in sequential feature selection and classification tasks with a small number of features [13, 14, 24, 26, 27, 33, 39, 44]. Figure 1 illustrates how RL works in the process of automatic disease diagnosis. However, RL frequently struggles in dealing with large feature spaces. Without a well-defined exploration strategy, an agent may fail to obtain useful rewards for learning in a large action space where most actions are unrelated to the target.

To address the challenge, we propose a scalable Multi-Model-Fused Actor-Critic RL framework (MMF-AC) that integrates a generative model and a supervised classification model into an actor-critic RL algorithm. The actor-critic algorithm consists of two networks, the actor and the critic, working together to solve the online disease diagnosis problem. At a high level, the actor decides which action should be taken next, and the critic informs the actor how good the action is and how it should adjust. Based on the observed symptoms, the actor of the RL agent must decide whether to continue inquiring (take action) or make a diagnosis (terminate), as well as which subsequent symptom to check if the decision to continue inquiring is made. If the actor decides to terminate the interaction, the supervised classification model is invoked to make a disease diagnosis. Naturally, an unobserved symptom with a high conditional probability given the observed symptoms should be inquired about next since the patient is likely to have it. Hence to speed up convergence during training, we develop a novel actor network by leveraging the Variational AutoEncoder (VAE) [17] that models the conditional probability distributions between observed and unobserved symptoms. In our customized critic network, we make use of the intermediate diagnostic confidence from the supervised classification model to help criticize the actions made by the actor since it would reinforce the agent to inquire about the proper symptoms that contribute to the diagnosis.

Inspired by the concept of differential diagnosis in healthcare, we create a novel differential reward shaping mechanism. Differential diagnosis is the process of distinguishing a particular disease from others that share similar symptoms [22, 34]. To narrow down the list of possible diseases, physicians typically inquire about specific symptoms, obtain a medical history, conduct a physical examination, or order diagnostic laboratory or imaging tests. This diagnostic process can be considered as inquiring key features in order to confirm or rule out potential diseases. To implement this concept within the context of RL, the RL agent should hypothesize a list of possible diseases based on observed symptoms and then use the reward function to prioritize key symptoms for differential diagnosis. However, estimating the reward function can be very challenging when the number of symptoms is large. Many existing RL algorithms [24, 27, 33] take a greedy approach to constructing reward functions by considering only “positive symptoms”, that is, those that the patient actually experiences, while ignoring all other “negative symptoms”, those that the patient does not have. Because positive symptoms account for a very small proportion of all possible symptoms, the greedy mechanism results in the sparse reward problem, which poses a scalability challenge for RL algorithms. Besides, this greedy mechanism would direct the agent to investigate symptoms that frequently co-occur while ignoring negative symptoms that are critical for differential diagnosis. In comparison, our differential reward shaping mechanism accounts for both the rate of co-occurrence of symptoms and the differential effect by using generative and diagnosis models.

Unlike the majority of previous RL baselines, which have been proven to be effective exclusively on either synthetic or real-world datasets, our approach is effective in both types of datasets. The results demonstrate that our method achieves the state-of-the-art performance on real-world datasets and is more effectively scalable to large search spaces on synthetic datasets, where most RL methods fail. Furthermore, we undertake an evaluation on a real-world diabetes dataset, showing that our approach, unlike conventional baselines, can accommodate continuous feature in addition to categorical symptoms.

Our contributions are highlighted in the following:

- We propose a Multi-Model-Fused Actor-Critic RL framework (MMF-AC) that achieves state-of-the-art performance across various settings. Our approach is suitable for online applications due to its scalability and generalizability;
- Technically, we leverage the characteristics of the VAE and the supervised classification model to enhance the agent with a generative actor and a diagnostic critic network;
- We present a novel differential reward shaping mechanism to address the sparse reward problem in large search spaces.

2 RELATED WORK

2.1 Dialogue-Based Disease Diagnosis

Deep reinforcement learning methods have been extensively utilized to address the problem of automatic disease diagnosis based on dialogue. Tang et al. [39] first formulated the inquiry and diagnosis process as a Markov decision process and then employed reinforcement learning (RL) in a simulated environment. Kao et al. [14] and Peng et al. [33] showed the feasibility of achieving competitive results in medium search spaces. Kachuee et al. [13] suggested a deep Q-learning method with model uncertainty variations as the reward function. In the task of sequential feature acquisition with a specified cost function, Janisch et al. [11, 12] found that RL methods outperformed tree-based methods. Xia et al. [44] developed
an RL agent for inquiry and diagnosis tasks using the Generative
Adversarial Network (GAN) and the policy gradient approach. The
RL agent performed well on two public datasets with four and five
diseases, respectively. However, in most real-world settings, such
tiny number of diseases is impractical. Lin et al. [26] proposed a
RL model combining an inquiry branch implemented through a Q
network with a diagnosis branch involving a generative sampler. It
was tested only on the two small datasets described previously. Liao
et al. [24] developed a hierarchical RL architecture with a master
for worker appointment, several workers for symptom inquiry, and
a distinct worker for disease screening. Liu et al. [27] improved
its convergence employing a pretraining strategy, but they were
limited to manage up to 90 different diseases.

There are also a number of non-RL approaches to inquiry-based
disease diagnosis. Entropy functions have been used to identify
symptoms based on the principle of information gain in a diverse
family of Bayesian inference and tree-based approaches [18, 20, 21,
32, 46, 47]. Wang et al. [40] developed a hierarchical lookup frame-
work based on symptom embedding and disease-specific graph
representation learning. Due to the intractable nature of maximiza-
tion of global information gain or global sensitivity, these methods
frequently rely on greedy or approximation algorithms that result
in low accuracy. Recently, EDDI [28], a framework for active fea-
ture acquisition, was proposed as a method for selecting the next
feature by maximizing a defined information-theoretic reward over
all features. It employs a Partial Variational AutoEncoder (VAE) to
deal with partially observed data when inquiring. The computa-
tional cost of estimating the reward using Monte Carlo sampling,
on the other hand, is too costly to handle a large number of symp-
toms. He et al. [10] adopted EDDI into a framework for online
disease diagnosis. They proposed a bipartite scalable framework
and demonstrated its state-of-the-art performance on a variety of
datasets, as well as its ability to deal with very large feature sets.
Although several novel strategies have been designed to reduce the
computational cost of the reward, the response time is still too slow
to support online applications.

2.2 Actor-Critic Reinforcement Learning

Q learning [41] and Deep Q networks [30] are examples of value-
based reinforcement learning methods attempting to compute the
optimal action-value function. While value-based methods are ef-
cient and stable, policy-based methods, such as policy gradient
techniques [25], which conduct gradient ascent optimization on
the performance objective to determine the best policy, are certain
to converge and easy to adapt to various contexts [31]. Taking ad-
vantage of both types of methodologies, the actor-critic algorithm
learns an actor network computing the probabilities of actions and
a critic network computing the state-value function. In particular,
the actor-critic approach optimizes an advantage function rather
than the state-value function. Many variants of actor-critic have
emerged recently. For instance, Mnih et al. [29] developed an asyn-
chronous variant of actor-critic, called asynchronous advantage
actor-critic (A3C). Schulman et al. [35] proposed a practical vari-
ant, Trust Region Policy Optimization (TRPO), by making several
approximations to the theoretically-justified scheme. Proximal Pol-
icy Optimization (PPO) [37] is an approach based on TRPO for
learning a stochastic policy by minimizing a clipped surrogate goal.
Haarnoja et al. [9] also suggested an off-policy maximum entropy
actor-critic method, commonly known as Soft Actor-Critic (SAC),
enabling sample-efficient learning while maintaining stability.

3 METHODOLOGY

In this section, we first present the detailed formulation of the prob-
lem and then introduce the structure of the generative actor and
diagnostic critic, as shown in Figure 2. Next, we elaborate on the
differential reward shaping mechanism. Finally, we give an overview
of the proposed Multi-Model-Fused Actor-Critic algorithm (MMF-
AC), as shown in Figure 3.

3.1 Problem Formulation

In this paper, we aim to learn a dialogue-based agent for automatic
disease diagnosis. The input data consists of a set of explicit symp-
toms that the patient self-reports prior to the doctor’s inquiry and
a set of implicit symptoms that the patient experiences but are
to be acquired through the dialogue. The output is a disease la-
bel diagnosed by the doctor. The objective is to efficiently acquire
the patient’s symptom conditions in order to establish an accurate
diagnosis. This can be described as a Markov Decision Process
(MDP) [5], which can be characterized by a tuple \( M(\mathcal{S}, \mathcal{A}, \mathcal{T}, R, \gamma) \).
\( \mathcal{S} \) (state) is a set of vectors that represents the current state of the
dialogue. \( \mathcal{A} \) (action) specifies the inquiry through a collection of
available actions, including inquiring about a specific symptom or
terminating the dialogue in order to make a diagnosis in our setting.
As a result, the objective is to determine the optimal conversational
policy \( \pi_\theta : \mathcal{S} \rightarrow \mathcal{A} \) that is composed of a series of actions. At each
step \( t \), the policy \( \pi_\theta(a_t|s_t) \) seek to determine the action \( a_t \in \mathcal{A} \)
based on the current state \( s_t \in \mathcal{S} \). The transition function \( \mathcal{T} \) is
the probability distribution over the next state \( s_{t+1} \) given current
state \( s_t \) and action \( a_t \), or \( p(s_{t+1}|s_t, a_t) \). \( R \) is a reward function
that evaluates the gain of the transition \( (s_t, a_t, s_{t+1}) \), and \( \gamma \) is the
discount factor for accumulative rewards at each step. The state-value
function that RL approaches strive to maximize is the expected sum
of discounted rewards given the state \( s_t \) corresponding to policy
\( \pi_\theta \):

\[
V_\pi(s_t) = \mathbb{E}[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots | s_t].
\]
In our setting, when the RL agent terminates the interaction, a supervised classification model is invoked to produce a disease diagnosis. Let $N_D$ denote the number of observed symptoms and $x_O \in \{0, 1\}^{N_O}$ denote the presence (1) or absence (0) of observed symptoms. Similarly, let $N_I$ denote the number of distinct diseases and $y \in \{0, 1\}^{N_D}$ denote a one-hot categorical vector with each dimension indicating whether a specific disease is present (1) or not (0). The diagnosis model is then defined as a mapping $\mathcal{D}$ between the observed symptoms $x_O$ and the disease probability distribution $\hat{y} \in \{0, 1\}^{N_D}$. $\mathcal{D}$ can be any supervised classification model that outputs prediction probabilities. For example, HiePPO [27] applied LightGBM [15] method, while BSODA [10] developed a supervised self-attention model to embed relationships between all features, including diseases and symptoms, and perform diagnosis using the disease embeddings. As BSODA shows great performance on a variety of datasets, we adopt its diagnosis model as $\mathcal{D}$. Please refer to BSODA [10]. Table 1 summarizes some of the notations.

3.2 Generative Actor

The left panel of Figure 2 displays the structure of the generative actor. Action $a_t$ in our framework is either inquiring about a specific symptom or terminating the dialogue, and it has up to $N_S + 1$ possible values where $N_S$ is the total number of symptoms. Naturally, an unobserved symptom $x$ with a high conditional probability $p(x | x_O)$ should be investigated next since the patient is likely to have this symptom. We model the conditional distributions between observed symptoms $x_O$ and unobserved symptoms $x_I$ in the following. A VAE defines a generative model of the form $p(x, z) = \prod_i p_\Theta(x_i | z)p(z)$ where the data $x$ is generated from latent variables $z$. $p(z)$ is a prior, e.g., spherical Gaussian, and $p_\Theta(x_i | z)$ is presented as a neural network decoder with parameters $\Theta$ specifying a simple likelihood, e.g., Bernoulli. A VAE produces a variational approximation of the posterior $q_\Phi(z|x)$ by using another neural network with parameters $\Phi$ as an encoder. The model can be trained by optimizing a lower bound on the evidence (ELBO):

$$\mathbb{E}_{q_\Phi(z|x)} \log p_\Theta(x_i | z) - \beta \cdot D_{KL}(q_\Phi(z|x) \parallel p(z)), \tag{2}$$

where $\beta$ is the weight to balance the two terms in the expression and $D_{KL}$ is the Kullback-Leibler (KL) divergence between two distributions. Then, $p(x_I | x_O)$ can be obtained by first sampling $\hat{z} \sim q_\Phi(z|x_O)$ from the VAE encoder, followed by $p_\Theta(x_I | \hat{z})$ from the VAE decoder. The most challenging aspect is that, during the inquiry process, observed symptoms are usually incomplete, resulting in up to $O(2^{N_S})$ distinct $q(z|x_O)$. It is impossible to train an independent encoder network for every distribution. To efficiently handle the partial observations, we propose to use the Product-of-Experts (PoE) encoder [43], which takes advantage of the nature of VAE’s latent variables $z$. Specifically, let $e_I$ denote the embedding vector of the $i$-th feature $x_i$ and $e_I = [e_1, e_2]$ denote the concatenated input carrying information of the $i$-th observed feature. Assuming conditional independence among features, the approximated posterior, including a prior Gaussian expert $p(z)$ with mean vector $\mu_0$ and variance vector $\Sigma_0$, is given by

$$q_\Phi(z|x_O) \propto p(z) \prod_i q_\Phi(z|x_i), \tag{3}$$

where $q_\Phi(z|x_i)$ is an inference network denoting the expert associated with the $i$-th observed feature $x_i$. Then we use a Multiple

![Figure 2: Detailed structures of the generative actor and diagnostic critic of MMF-AC.](image)

Layer Perceptron (MLP) as $h_{enc}(\cdot)$ to map the input $e_I$ to a Gaussian distribution in latent space with mean vector $\mu_I$ and variance vector $\Sigma_I$. $h_{enc}(\cdot)$ shares parameters across all features. Because a product of Gaussian experts is itself Gaussian [6], we can quickly compute variance $V = (V_I^{-1} + \sum V_i^{-1})^{-1}$ and mean $\mu = (\mu_I V_I^{-1} + \sum \mu_i V_i^{-1}) V$ for any $q_\Phi(z|x_O)$ in Eqn. 3. The decoder, $p_\Theta(x|z)$, is given by a product of Bernoulli distributions, the probabilities of which are specified by an MLP that accepts $z$ as input.

To develop our generative actor network, we imitate the VAE structure and then use a single layer network $h_{ac}(\cdot)$ with a softmax activation function to map the decoder output to the action space. During training, we pretrain a VAE and then use its parameters to initialize the actor’s encoder and decoder. Only the actor’s decoder will be fine-tuned. In the early stages of training, the actor can use the pretrained parameters to investigate significant symptoms and improve the learning efficiency. The pretrained VAE will also be used in the next sections.

3.3 Diagnostic Critic

The right panel of Figure 2 displays the structure of the diagnostic critic. The critic’s objective is to precisely compute the state-value function $V_s(s_t)$ in Eqn. 1 in order to optimize the policy $\pi_\theta$. At timestep $t$, the state $s_t \in \{1, 0, -1\}^{N_S}$ represents the current symptom information, with 0 indicating a symptom that has not been investigated, 1 indicating a positive symptom or that the patient has experienced the symptom, and -1 indicating a negative symptom or that the patient did not experience the symptom. However, the diagnostic confidence from the diagnosis model is also significant to evaluate the current state, which would reinforce the agent to inquire about the proper symptoms that contribute to the diagnosis. The intermediate diagnostic result and the current timestep $t$ can also help determine when to terminate the interaction and produce a confident final diagnosis. Furthermore, to handle arbitrary
3.4 Differential Reward Shaping

Differential diagnosis is defined as the process of differentiating between the probability of a particular disease versus that of other diseases with similar symptoms that could possibly account for illness in a patient [22]. Based on this concept, we propose a novel reward function \( R(s_t, s_{t+1}) \) to motivate our RL agent not only to examine symptoms with a high co-occurrence, but also to undertake a differential diagnosis process in the manner of real doctors. Inspired by HiePPO [27], we first define \( R(s_t, s_{t+1}) \) as

\[
R(s_t, s_{t+1}) = \begin{cases} 
R_{\text{end}}, & \text{if dialogue ending} \\
R_{\text{diff}}, & \text{otherwise} 
\end{cases}
\]  

where \( R_{\text{end}} \) is the long-term reward for the final diagnosis, and \( R_{\text{diff}} \) is the short-term reward for evaluating the inquiry. When the dialogue ends, we set \( R_{\text{end}} \) to 1 if \( y_{\text{pred}} = y_{\text{true}} \) otherwise to −1. \( y_{\text{pred}} \) is the diagnosis made by \( D \) and \( y_{\text{true}} \) is the true diagnosis.

To evaluate the inquiry quality, we define a function \( \text{Diff}(s_t, x) \) to estimate the differential effect on the current state \( s_t \) if the patient has a certain symptom \( x \):

\[
\text{Diff}(s_t, x, s'_t) = |D_{KL}[y \mid \mathcal{D}(s_t)] - D_{KL}[y \mid \mathcal{D}(s'_t)]|,
\]

where \( s'_t \) assumes that the patient has symptom \( x \). The two KL divergences measure the difference between the true diagnosis result and the disease distribution predicted by the diagnosis model \( \mathcal{D} \). Hence the absolute value of the difference between the two terms indicates the significance of \( x \) given \( y \) and \( \mathcal{D} \). Then, depending on whether or not \( x \) is one of the implicit symptoms of the patient, we design two slightly different short-term rewards. If the patient truly experienced symptom \( x \), \( R_{\text{diff}} \) is simply \( \text{Diff}(s_t, x, s_{t+1}) \):

\[
R_{\text{diff}} = \text{Diff}(s_t, x, s_{t+1})
\]

since \( s'_t \) is equal to \( s_{t+1} \). Otherwise, for negative symptom \( x \), we have

\[
R_{\text{diff}} = p(x|\tilde{x}_O) \cdot \text{Diff}(s_t, x, s'_t) - \alpha,
\]

where the conditional probability \( p(x|\tilde{x}_O) \in (0, 1) \) reduces the reward for unrelated symptoms based on observed data, which can be computed by the pretrained VAE as described in Sec. 3.2. \( \alpha \) is a constant penalty factor for lengthy interactions.

Our reward shaping function is more instructive than other approaches’ greedy reward schemes [24, 27, 33], which only consider positive symptoms. For example, HiePPO [27] simply returns a constant reward when a positive symptom is inquired about or when the rank of the true label’s probability in the diagnosis result is moved up after an inquiry. The greedy scheme causes the sparse reward problem and offers a scalability barrier because positive symptoms account for a tiny proportion of all possible symptoms.

3.5 Multi-Model-Fused Actor-Critic Reinforcement Learning

We incorporated a generative actor network (Sec. 3.2), a diagnostic critic network (Sec. 3.3), a pretrained VAE, a supervised diagnosis model, and a differential reward shaping function to create a Multi-Model-Fused Actor-Critic reinforcement learning framework (MMF-AC), as shown in Figure 3. The structures of the actor and the critic networks are depicted in Figure 2. The solid lines in Figure 3 represent the MMF-AC workflow within a single round of inquiry and diagnosis. When the actor decides to inquire about a specific symptom from the patient, the diagnosis model will produce an intermediate prediction based on the patient’s response. The differential reward shaping function defined in Sec. 3.4 then uses the prediction to compute the reward and update the agent. If the actor chooses to terminate the dialogue, the final diagnosis would be the result of the last intermediate prediction. The dotted lines in Figure 3 illustrate the effects of the generative and diagnosis models in MMF-AC. Apart from producing a diagnosis prediction, the diagnosis model feeds the diagnostic confidence to the critic. The pretrained VAE initializes the actor’s parameters, and the conditional probability computed by the pretrained VAE is utilized for reward shaping. Additionally, the critic makes use of the pretrained VAE’s ability to deal with arbitrary partial observations.

In the training procedure, we begin by training the supervised diagnosis model that receives fully observed symptoms \( x \) as input. We drop a random fraction of the complete features \( x \) to obtain \( \tilde{x} \), imitating arbitrary partial observations during the inquiry process. Then we pretrain a VAE model by the ELBO in Eqn. 2, which outputs \( \hat{p}_O(\tilde{x} | z) \) to recover the input \( x \) as described in Sec. 3.2. After that, we initialize the actor’s encoder and decoder with the pretrained VAE’s parameters and only the decoder is fine-tuned in the follow-up training. Next, we train the actor and critic with Proximal Policy Optimization (PPO) [37]. The critic network is trained using a mean squared error loss function: 

\[
L_{\text{MSE}} = \sum (V_{\text{pred}} - V_\pi(s_t))^2,
\]

where \( V_{\text{pred}} \) is the output of the critic and \( V_\pi(s_t) \) is calculated by Eqn. 1.
According to the action-critic algorithm [29] and the Generalized Advantage Estimator (GAE) [36], let $\delta_t^\pi = r_t + \gamma V(s_{t+1}) - V(s_t)$ and the advantage function optimized at timestep $t$ is

$$A_t = \sum_{l=0}^{T} (\gamma \lambda)^l \delta_{t+l}^\pi,$$

(8)

where $T$ is the final step, $\gamma$ is the discount factor, and $\lambda$ is a hyperparameter of GAE. According to the clipped PPO objective function [37], the goal of actor training is to maximize the objective function:

$$L(\theta) = E_\pi [\min(o_t(\theta)A_t, clip(o_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t) + \eta \cdot E(\theta)],$$

(9)

where $o_t(\theta) = \frac{\pi(\theta_s)}{\pi_0(\theta_s)}$ is the probability ratio and $\theta$ is the parameters of the actor network. $E(\theta) = -\pi(\theta_s)ln(\pi_0(a_t|s_t))$ is the entropy of the policy $\pi_0$ to encourage more exploration and $\eta$ is a small weight. To avoid inquiring repeatedly, we set the probabilities of symptoms mentioned in the actor’s output to zero. A maximum number of inquiries $N_T$ is also set to avoid lengthy interactions. Finally, when conducting the final diagnosis, we use the VAE to impute unobserved symptoms before feeding them to the diagnosis model to provide a more accurate diagnosis.

4 EXPERIMENTS

4.1 Datasets

Obviously, due to patient privacy and security concerns, it is difficult to obtain large-size real-world medical data from the public domain for training and evaluation. Nevertheless, computational approaches should be evaluated for their generalizability and ability to handle large datasets. Hence synthetic datasets containing thousands of features are used to evaluate a model’s ability to handle large feature spaces. MMF-AC shows excellent performance on both synthetic and real-world datasets. All datasets are publicly accessible and do not contain any personally identifiable information.

4.1.1 Real-World Medical Dialogue Datasets. The MuZhi medical dialogue dataset [42] was collected from the dialogue data from the pediatric department of a Chinese online healthcare website (https://muzhi.baidu.com/). It includes 66 symptoms associated with 4 diseases: children’s bronchitis, children’s functional dyspepsia, infantile diarrhea infection, and upper respiratory infection.

The Dxy medical dialogue dataset [45] contains data from another popular Chinese online healthcare website (https://dxy.com/). It includes 41 symptoms for 5 diseases: allergic rhinitis, upper respiratory infection, pneumonia, children hand-foot-mouth disease, and pediatric diarrhea. All patients in the dataset are infants. These two datasets have been structured in the form of disease-symptoms pairs and split into a training set and a testing set.

4.1.2 Real-World Diabetes Dataset. Kachuee et al. [13] compiled a diabetes dataset from the NANational Health and Nutrition Examination Survey (NAHNES) data [2]. It consists of 45 features, such as demographic information, lab results (total cholesterol, triglyceride, etc.), examination data (weight, height, etc.), and questionnaire answers (smoking, alcohol, etc.). In contrast to the above symptom checking datasets, some of these features are continuous, and each feature is associated with a cost suggested by an expert in medical studies. The fasting glucose values were used to define three classes based on standard threshold values: normal, pre-diabetes, and diabetes. The dataset is randomly split into three subsets, including 15% for testing, 15% for validation, and the rest 70% for training.

4.1.3 Synthetic SymCAT Dataset. SymCAT is a symptom-disease database [4]. There is information about the symptoms associated with each disease, as well as the marginal probabilities associated with these symptoms. Using SymCAT, we construct a synthetic dataset in the manner described by Peng et al. [33]. We randomly select a disease and its associated symptoms from among all disorders, and then conduct a Bernoulli trial on each extracted symptom based on the associated probability to create a symptom set. For example, if we take a sample of the disease "abscess of nose," we would observe the following symptoms: "cough" and "fever," which occur in 73% and 62% of cases, respectively. Then, based on these probabilities, we generate one data instance by sampling Bernoulli random variables. In this way, we created three diagnostic tasks with 200, 300, and 400 diseases, respectively, and for each task, we sampled $10^6$, $10^7$, and $10^4$ records, respectively for training, validation, and testing.

4.1.4 Synthetic HPO Dataset. The Human Phenotype Ontology (HPO) [19] provides a standardized vocabulary for phenotypic abnormalities associated with human diseases. The HPO's initial domain of application was rare disorders. There are 123,724 HPO terms associated with rare diseases and 132,620 with common diseases. Following He et al. [10], we chose diseases and related symptoms from the public rare disease dictionary, along with their marginal probabilities, and used the aforementioned process for SymCAT to create synthetic datasets. Two diagnostic tasks containing 500 and 1,000 diseases were created for evaluation.

4.2 Experimental Protocol

MMF-AC is compared with the following baselines:

- GAMP [44] uses Generative Adversarial Network (GAN) to implement an RL agent for dialogue-based diagnosis, which has only been proven to be effective on MuZhi and Dxy.
- FlatPPO [27] is a simplified version of HiePPO [27], a customized hierarchical RL approach, which has been applied to MuZhi and Dxy. HiePPO can only take up to 90 diseases, which is too small to be included as our baseline for synthetic datasets.
- REFUEL [33] is a deep RL method that has been evaluated on medium sized synthetic datasets.
- OL [13] is based on deep Q-learning and designed for cost-sensitive feature selection in datasets such as the diabetes dataset.
- BSODA [10] is a non-RL framework, which has shown state-of-the-art performance on a variety of datasets and the ability to deal with a large number of features.

Patients’ self-reports, if available, are used at the beginning of the inquiring process. MuZhi and Dxy have already had self-reports. For the synthetic datasets, a positive symptom is chosen uniformly at random as a patient self-report. The diabetes dataset is evaluated at random as a patient self-report. The diabetes dataset is evaluated for its generalizability and ability to handle large datasets. Nevertheless, computational approaches should be evaluated for their generalizability and ability to handle large datasets. Hence synthetic datasets containing thousands of features are used to evaluate a model’s ability to handle large feature spaces. MMF-AC shows excellent performance on both synthetic and real-world datasets. All datasets are publicly accessible and do not contain any personally identifiable information.
Table 2: A case study for actual inquired symptoms on the Dxy dataset.

| Doctor       | Source       | Inquired Symptoms                     | Diagnosis       |
|--------------|--------------|---------------------------------------|-----------------|
| BSODA        | Convulsion,  | Rubby Nose, Rubbing Nose, Coughing Up Phlegm | Allergic Rhinitis|
| MMF-AC       | Loose Stools, Fever, Coughing Up Phlegm, Vomiting | Pneumonia       |
| Doctor       | Vomiting,     | Stuffy Nose                           | Pneumonia       |

Table 3: Performance of MMF-AC and previous works on two medical dialogue datasets.

| Method   | MuZhi       | Dxy        |   |
|----------|-------------|------------|---|
|          | Accuracy (%)| #Inquiries|   |
| REFUEL   | 71.8        | 6.20       |   |
| GAMP     | 73          | -          | 76.9| 2.68|
| FlatPPO  | 73.2        | 6.32       | 74.6| 3.31|
| BSODA    | 73.1        | 8.80       | 80.2| 7.49|
| MMF-AC   | 73.9        | 4.47       | 82.7| 6.08|

Table 4: Performance of OL [13], BSODA [10] and MMF-AC on the diabetes dataset.

| Method   | Accuracy (%) | Costs |
|----------|--------------|-------|
| OL       | 81.66        | 63.99 |
| BSODA    | 86.11        | 23.49 |
| MMF-AC   | 87.18        | 20.31 |

for the synthetic HPO 1,000-disease task. The discount factor $\gamma$ in Eqn. 8 is set to 0.95 for the synthetic SymCAT dataset and 0.9 otherwise. $\lambda$ in Eqn. 8 is set to 0.9 and $\epsilon$ in Eqn. 9 is set to 0.1. In Eqn. 9, $\eta$ is set to 0.001 for the synthetic HPO dataset and to 0.01 for all other datasets. We train the networks using the Adam optimizer [16]. The learning rates of both the actor and critic are $5 \times 10^{-5}$ and the batch size is 512. $\eta$ and the learning rates would linearly decay to one-tenth with iterations. The upper limit of dialogue rounds $N_T$ is set to 20. The evaluation metrics are diagnostic accuracy and the average number of rounds in dialogues. Our objective is to achieve high diagnostic accuracy while limiting the rounds of dialogues.

4.3 Results

4.3.1 Real-World Medical Dialogue Datasets. Table 3 compares the results for MMF-AC and previous works on the MuZhi and Dxy datasets. Since GAMP [44] did not disclose their source code or the average number of inquiries on MuZhi, we only present its self-reported accuracy on MuZhi. The results show that our method MMF-AC outperformed the other techniques on MuZhi, achieving the highest accuracy and requiring the fewest inquiry rounds. Our method is also the most accurate on Dxy, although it generates more queries than REFUEL [33], GAMP [44], and FlatPPO [27]. Overall, MMF-AC remains the best method because its accuracy is at least 5.8 points higher than the other methods. Additionally, the number of inquiries generated by MMF-AC is less than that by the previous SOTA method, BSODA [10].

Using the Dxy dataset, we conducted a case study to investigate the actual inquired symptoms of BSODA, MMF-AC and the real doctor. The results are summarized in Table 2 with the inquired implicit symptoms highlighted in bold. The upper case is a 40-day infant who had Pneumonia but was easily misdiagnosed as having Allergic Rhinitis due to the presence of both nasal symptoms and infantile pneumonia signs, such as spitting bubbles and difficult breathing. It is worth noting that infant may exhibit no symptoms of pneumonia infection. Alternatively, they may vomit, have a fever and cough, or have diarrhea. These are not typical symptoms of Allergic Rhinitis. As a result, MMF-AC inquired about Loose Stools, Fever and Vomiting in order to differentiate the two illnesses. While BSODA tended to inquire more about nasal symptoms based on the initial "Clear Nasal Discharge", resulting in an erroneous misdiagnosis. It demonstrates that, using differential reward shaping, MMF-AC makes inquiries based on the differential effect rather than depending on co-occurrence of symptoms such as BSODA.

Another case involves a 46-day old infant who is suffering from Pneumonia. The key implicit symptom is Difficult Breathing. Although both BSODA and MMF-AC inquired about Difficult Breathing, we observe that MMF-AC required fewer inquiries before terminating the interaction and providing a confident diagnosis, which demonstrates its efficiency.

4.3.2 Real-World Diabetes Dataset. Apart from symptom checking tasks, Table 4 shows an evaluation using the real-world diabetes dataset that has hybrid types of features, such as laboratory test results and questionnaire answers. The total costs for all 45 features are 185. It should be noted that symptom checking baselines such as REFUEL [33] and FlatPPO [27] cannot be directly applied to these continuous features since they only construct reward functions for implicit symptoms instead of continuous features as in the diabetes dataset. These baselines are incapable of differentiating the significance of various continuous features in a patient. For MMF-AC, however, these feature types have no effect on the differential function in Eqn. 5, so it could handle these features only with the reward in Eqn. 6. We observe that both BSODA [10] and MMF-AC outperform OL [13] in terms of accuracy while having a lower
Table 5: Performance of REFUEL [33], BSODA [10] and MMF-AC on synthetic datasets.

| Task             | Method          | Top1 | Top3 | Top5 | #Inquiries |
|------------------|-----------------|------|------|------|------------|
| SymCAT (200 Diseases, 328 Symptoms) | REFUEL         | 53.76 | 73.12 | 79.53 | 8.24       |
|                  | BSODA          | 55.65 | 80.71 | 89.32 | 12.02      |
|                  | MMF-AC         | 59.00 | 84.66 | 92.45 | 12.54      |
| SymCAT (300 diseases, 349 Symptoms) | REFUEL         | 47.65 | 66.22 | 71.79 | 8.39       |
|                  | BSODA          | 48.23 | 73.82 | 84.21 | 13.10      |
|                  | MMF-AC         | 51.91 | 78.24 | 87.55 | 13.80      |
| SymCAT (400 diseases, 355 Symptoms) | REFUEL         | 43.01 | 59.65 | 68.89 | 8.92       |
|                  | BSODA          | 44.63 | 69.22 | 79.54 | 14.42      |
|                  | MMF-AC         | 45.80 | 72.14 | 82.43 | 14.73      |
| HPO (500 diseases, 1901 Symptoms) | REFUEL         | 64.33 | 73.14 | 75.34 | 8.09       |
|                  | BSODA          | 76.23 | 84.17 | 86.99 | 5.34       |
|                  | MMF-AC         | 88.39 | 93.96 | 94.96 | 4.77       |
| HPO (1000 diseases, 3599 Symptoms) | REFUEL         | 40.08 | 62.67 | 67.42 | 14.19      |
|                  | BSODA          | 67.03 | 75.94 | 79.27 | 10.62      |
|                  | MMF-AC         | 63.04 | 77.99 | 81.43 | 10.61      |

4.4 Ablation Study

We performed an ablation study to investigate the effects of the VAE and the differential reward shaping mechanism using the synthetic HPO 500-disease task. In Figure 4, "MMF-AC w/o RS" denotes that we omit the differential reward shaping function and instead set $R_{diff}$ simply to 0.25 when inquiring about an implicit symptom. "MMF-AC w/o RS and VAE" denotes that we omit the effects of both the reward shaping and the VAE. The generator is then a simple MLP, and $x_t$ encoded by the VAE in the critic is replaced by the observed state $s_t$. The number in the parenthesis denotes the average number of implicit symptoms queried. We observe that, with an increasing number of epochs, MMF-AC without partial observations handled by the VAE converges slowly in the initial stage and eventually to a horrible policy, highlighting the critical role of the generative model in large search spaces. Without differential reward shaping, MMF-AC is able to inquire more implicit symptoms based on symptom co-occurrence, but performs poorly due to the omission of the differential effect of symptoms. It indicates that the differential effect of symptoms is more important than their co-occurrence in making an accurate diagnosis.

5 CONCLUSION

We offer MMF-AC, a novel Multi-Model-Fused Actor-Critic reinforcement learning approach for online disease diagnosis that is scalable to large feature spaces. It incorporates a generative model and a supervised classification model into an actor-critic RL algorithm. We also develop a novel reward shaping mechanism inspired by differential diagnosis. The experimental results on both real-world and synthetic datasets demonstrate the effectiveness of our approach, which outperforms prior RL and non-RL methods in their respective settings [10, 13, 27, 33, 44]. The scalability to large feature spaces and generalizability for handling hybrid feature types make our approach suitable for online applications.

with only a slightly lower performance in the Top 1 category of the 1,000-disease task, MMF-AC remains the most competitive and efficient, and it is more suitable for online applications.
ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (2021YF1201303 and 2019YFB140404), the National Natural Science Foundation of China (grants 61872218, 61972218, and 61972219), the Beijing Municipal Science and Technology Commission (grant Z191100000419002), and Beijing National Research Center for Information Science and Technology (BNRist). We are grateful to the many people who have contributed to the development of our software, especially the developers of Theano, Tensorflow, and PyTorch.

REFERENCES

[1] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).
[2] Sebastian Cross, Ahmed Mourad, Guido Zuccon, and Bevan Koopman. 2021. AHEAD Research Inc. 2017. SymCAT: Symptom-based, computer assisted triage. http://www.symcat.com.
[3] Richard Bellman. 1957. A Markovian decision process. Journal of mathematics and mechanics 6, 5 (1957), 679–684.
[4] Susannah Fox and Maeve Duggan. 2013. Health Online 2013. Pew Research Center.
[5] Yanshuai Cao and David J Fleet. 2014. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. arXiv preprint arXiv:1410.7827 (2014).
[6] Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. In Thirty-Second AAAI Conference on Artificial Intelligence.
[7] A. Boycott, et al. 2017. The human phenotype ontology in 2017. Methods of Information in Medicine, Vol. 56, No. 1, pp. 27–37.
[8] Hajin Shim, Sung J. Hong, and Eunho Yang. 2018. Joint active feature acquisition and classification with variable-size set encoding. In Advances in neural information processing systems 30 (2017), 3735–3744.
[9] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In International conference on machine learning. PMLR, 1889–1897.
[10] Lin Xu, Yefeng Zheng. 2021. Online Disease Diagnosis with Inductive Heterogeneous Graph Convolutional Networks. In Proceedings of the Web Conference 2021. 3349–3358.
[11] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. Machine learning 8, 3-4 (1992), 279–292.
[12] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2017. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438 (2015).
[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017).
[14] Hajin Shim, Sung J. Hong, and Eunho Yang. 2018. Joint active feature acquisition and classification with variable-size set encoding. In Advances in neural information processing systems 30 (2017), 3735–3744.