Knowledge-Based Systems 242 (2022) 108292

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning

Abhisek Tiwari *, Sriparna Saha, Pushpak Bhattacharyya

Department of Computer Science and Engineering, Indian Institute of Technology, Patna, India

ARTICLE INFO

Article history:
Received 31 July 2021
Received in revised form 21 January 2022
Accepted 22 January 2022
Available online 31 January 2022

Keywords:
Automatic disease diagnosis
Symptom investigation
Virtual diagnosis assistant
Bayesian learning
Deep reinforcement learning

ABSTRACT

Disease diagnosis is an essential and critical step in any disease treatment process. Automatic disease diagnosis has gained immense popularity in recent years, owing to its efficacy, easy accessability and reliablity. The major challenges for the diagnosis agent are inevitably large action space (symptoms) and varieties of diseases, which demand either rich domain knowledge or an intelligent learning framework. We propose a novel knowledge-infused context-driven (KI-CD) hierarchical reinforcement learning (HRL) based diagnosis dialogue system, which leverages a bayesian learning-inspired symptom investigation module called potential candidate module (PCM) for aiding context-aware, knowledge grounded symptom investigation. The PCM module serves as a context and knowledge guiding companion for lower-level policies, leveraging current context and disease-symptom knowledge to identify candidate diseases and potential symptoms, and reinforcing the agent for conducting an intelligent and context guided symptom investigation with the information enriched state and an additional critic known as learner critic. The knowledge-guided symptom investigation extracts an adequate set of symptoms for disease identification, whereas the context-aware symptom investigation aspect substantially improves topic (symptom) transition and enhances user experiences. Furthermore, we also propose and incorporate a hierarchical disease classifier (HDC) with the model for alleviating symptom state sparsity issues, which has led to a significant improvement in disease classification accuracy. The proposed framework outperforms the current state-of-the-art method on the multiple benchmarked datasets and, in all evaluation metrics other than dialogue length (diagnosis success rate, average match rate, symptom identification rate, and disease classification accuracy by 7.1 %, 0.23 %, 19.67 % and 8.04 %, respectively), which firmly establishes the efficacy of the proposed bayesian learning-inspired context-driven symptom investigation and disease diagnosis methodology [1].

1. Introduction

Disease diagnosis is a primary and critical step in any disease treatment process. Since last decade, electronic health records (EHRs) [1] based systems have emerged as promising fields for automatic diagnosis [2,3]. However, it requires huge labor and effort to develop an accurate and robust EHR. Also, the diagnosis capability of EHRs used to be disease-specific, which necessitates a dedicated EHR system for each disease. The report by WHO, 2019 [4] discloses that there are many countries where doctor per 1000 people is less than one. The figure strongly suggests that the healthcare system needs to be improved by expanding the number of health workers and better utilizing their time. There are mainly two sub-stages in disease diagnosis systems: i. Symptom investigation, and ii. Disease identification. Doctors spend a substantial amount of time on determining disease/condition. To alleviate such extensive efforts and to make better use of doctors' time, researchers have introduced a new paradigm for conducting symptom investigations and diagnosing disease automatically. The paradigm attempts to imitate real-world scenarios, where doctors either diagnose a disease or refer to a clinical test based on patients' difficulties and symptoms. The initial objectives of automatic diagnosis systems were early diagnosis and assistance to real doctors in symptom investigation rather than replacing them, which helps in better utilization of doctors' time and hence a more economical system. In recent years, these automatic diagnosis systems assist doctors effectively and evolving as standalone diagnosis systems because of their reliable performance, scalability, and cost-effective aspect.

In real life, doctors diagnose a substantial number of diseases by conducting an in-depth symptom investigation only. Nevertheless, they require further evidence, such as lab reports in some cases, before reaching to a conclusive disease. However, the symptom investigation seems essential also for suggesting

* Corresponding author.
E-mail addresses: abhisek_1921cs16@iitp.ac.in (A. Tiwari), sriparna@iitp.ac.in (S. Saha), pushpakbh@gmail.com (P. Bhattacharyya).

[1] The code is available at https://github.com/NLP-RL/KI-CD

https://doi.org/10.1016/j.knosys.2022.108292
0950-7051/

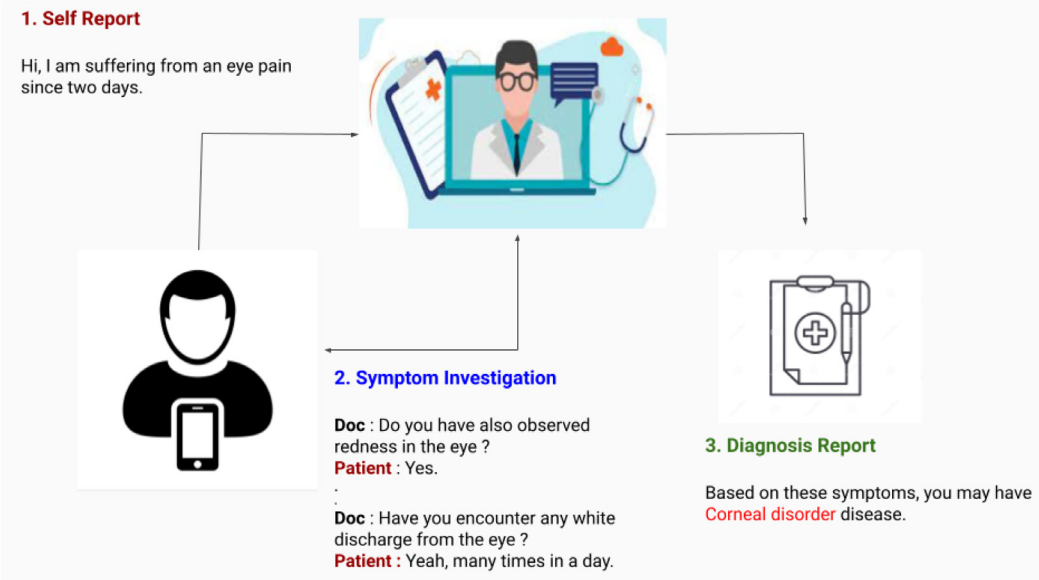


Fig. 1. Illustration of Automatic Disease Diagnosis System using an example, where the patient reports about his/her suffering symptom (eye pain), the doctor investigates some more relevant symptoms for disease grounding and then provides the diagnosis report (Corneal disorder).

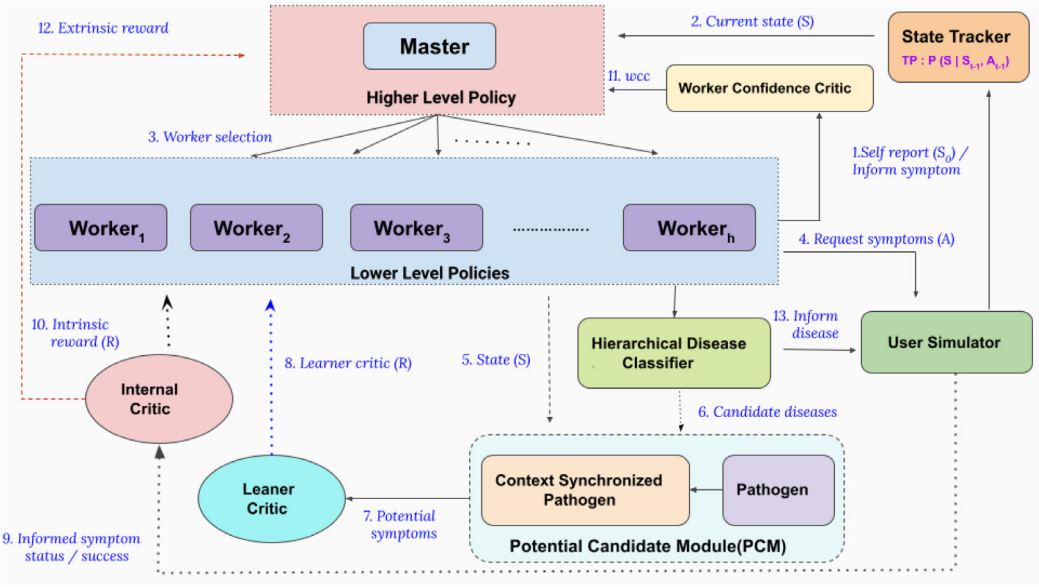


Fig. 2. Proposed hierarchical reinforcement learning framework with learner critic controlled by potential candidate module (PCM) and hierarchical disease classifier (HDC) : The higher level policy (Master) selects one of the worker/group (like medical department such as pediatrics) policies depending upon patient's self-report & dialogue context, and the selected lower-level policy conducts department/group-specific symptom investigation.

lab reports. Motivated by this real-world scenario, Wei et al. [5] formulated disease diagnosis as a task-oriented dialogue system (interactive system), where the dialogue agent converses with a patient for extracting symptom information and diagnoses a disease depending upon the extracted symptoms. An illustration of a typical automatic diagnosis dialogue system has been shown in Fig. 1. In automatic disease diagnosis dialogue systems, patients initiate dialogue by informing his/her chief complaint (self-report), the system inspects a series of symptoms depending upon the reported symptoms & ongoing context and diagnoses a disease depending upon the status (presence/absence) of these extracted symptoms including patient self-reported symptoms.

In a typical task-oriented dialogue system [6,7], the agent serves users' task based on specified task information throughout the conversation. But, it is hard for the diagnosis task to diagnose a patient based on only his/her informed symptoms, i.e., self-report. Most people are not aware enough of the disease/medical domain, so their self-reports may not contain adequate symptoms or information for diagnosing the correct disease. It emphasizes the role of doctors who have expertise in collecting adequate and essential information (symptoms) required for diagnosing patients correctly. Thus, a typical slot-filling dialogue agent will not be sufficient for the task; it requires an advanced dialogue system. In real life, when a patient visits a clinic, he/she is allocated to a certain department, such as a paediatrician or a physician, based on his or her self-report. Then, the assigned doctor consults with the patient for diagnosis. A similar kind of hierarchical structure has been formulated for dialogue policy

learning through Hierarchical Reinforcement Learning [8], which leads to better diagnosis performance [9]. However, the obtained result shows that the agent diagnoses incorrect diseases almost half of the time because of the large number of diseases and overlapping symptoms across these diseases. Disease diagnosis is a crucial and sensitive stage of any disease treatment, so an incorrect diagnosis may adversely impact someone's health.

Motivation and Research Question : Disease diagnosis is the first stage in any medical treatment process, where doctors spend a significant amount of time identifying disease/condition. A detailed symptom investigation report prepared by a junior doctor /an artificial intelligence-based agent can significantly reduce the burden of conducting detailed and long symptom investigations. Thus, they can focus more on disease identification and the treatment process. Diagnosis process deals with exponential state space, i.e., for D diseases and n symptoms, a diagnosis policy needs to learn to map symptom space of size 2^n-1 to disease space (D). Thus, a trivial mapping of symptoms set to diseases seems intractable, which emphasizes the need for knowledge-aware, experience-guided disease diagnosis policy. In the real world, when a doctor diagnoses a patient, he or she concocts a list of possible diagnoses based on the patient's symptoms. The doctor then investigates any possible signs of the hypothesized condition, allowing him to diagnose patients with greater confidence and fewer conversation turns. It might be irritating for patients if the VA/virtual doctor investigates some irrelevant symptoms, i.e., the VA needs to understand potential symptoms based upon current context. The context/current symptom status can surely guide the VA to extract more relevant and appropriate symptoms, leading to a more accurate diagnosis.

In this work, we mainly investigate the following two research questions. ***i.*** *In real life, doctors generally go in reverse order, i.e., they hypothesize a set of candidate diseases based on the patient's self-report and informed symptoms, and they prioritize potential symptoms of candidate diseases. The interleaving of symptom exploration and exploitation in this bayesian learning-inspired diagnosis approach results in adequate and appropriate symptom inspection. This work investigates the role of such a bayesian learning-inspired diagnosis process with interleaving symptom exploration–exploitation and proposes a bayesian learning-based knowledge-infused dialogue agent for automated disease diagnosis setting (Fig. 2).* ***ii.*** *In healthcare systems, there are various medical departments, and symptom investigation, diagnosis, and treatment take place in the appropriate departments. However, the existing automated diagnosis systems have incorporated only department-aware symptom investigation, which improves diagnosis accuracy significantly because of efficient symptom exploration. The proposed work also incorporates a hierarchical disease classifier (department and disease) and studies its impact in an automated disease diagnosis system.*

The main contributions of the paper are as follows :

- We propose a novel knowledge-infused, context-driven task-oriented virtual agent for autonomous disease diagnosis that employs a bayesian learning-inspired diagnosis strategy to enhance context-aware, knowledge-based symptom research, resulting in relevant symptom investigation and accurate diagnosis.
- We have also developed and incorporated a Hierarchical Disease Classifier (HDC) in the proposed framework. The work is the first attempt, which builds and incorporates a hierarchical disease classification method in automatic disease diagnosis setting. Results show that the HDC elevates disease classification performance by a significant margin of 8.04% because of more focused symptom/disease representation and more effective and grounded symptom investigation.
- The proposed agent, KI-CD, outperforms the state of the art technique in all the metrics other than dialogue length, i.e., diagnosis success rate, AMR, SIR and disease classification accuracy by 7.1%, 0.23%, 19.67% and 8.04%, respectively.

The rest of the paper is organized as follows: Section 2 highlights related works and the weaknesses/limitations of existing automatic disease diagnosis systems. A list of notations and their meaning has been reported in Table 1. Section 3 illustrates the problem formulation. The proposed context-aware, knowledge-guided diagnosis method is described in Section 4. Section 5 provides dataset statistics and details. Experimental setup and the obtained results are reported in Section 6. A case study has been mentioned in Section 7. The final section, conclusion (Section 8), summarizes the research work and outlines a few future directions in the line of the proposed research work.

2. Background and related work

Since the last decade, automatic disease diagnosis has become one of the most trending research topics in the Artificial Intelligence (AI) community because of medical data availability, the advancement of machine learning /deep learning techniques, and most importantly, promising & reliable performance of employed frameworks. The existing diagnosis approaches could broadly be categorized into two methods: 1. Disease Diagnosis, 2. Symptom Investigation and Disease Diagnosis. The former learns to diagnose an appropriate disease for a given patient's difficulties /symptoms, which is either collected by real doctors or EHRs. The latter first investigates patients' symptoms and then diagnoses disease depending upon symptom complications. Our work belongs to the latter approach, where the proposed agent first collects adequate symptoms depending upon patients' self-report and investigates symptoms, and then diagnoses an appropriate disease diagnosis based on collected symptoms. Although the second paradigm is in its beginning phase of research and deployment, it has considerably demonstrated its needs through the effectiveness and reliable performances of recently proposed frameworks.

Reinforcement learning and Knowledge infused learning : Reinforcement Learning (RL) [10] is a very well-known and acceptable solution for any optimization task such as dialogue policy learning. The central component of any dialogue system is dialogue policy, which can effectively be framed as a Markov Decision Process(MDP) [11] and optimized through an RL algorithm such as Deep Q Network(DQN) [12]. RL suffers from the curse of dimensionality in case of a complex problem where the environment state grows exponentially, which is a primary motivation of Hierarchical Reinforcement Learning (HRL) [13]. Option [14] is one of the HRL frameworks, which divides a complex problem into smaller ones by providing an abstraction over action space. In recent few years, HRL has evolved as one of the most promising matches for dialogue policy optimization for complex tasks [15–17]. In the deep learning technique, models learn tasks from underlying humongous raw data. In addition to the amount and quality of data, the data representation technique plays an important part in the learning process [18]. In [19,20], the authors have proposed a novel concept known as knowledge infused learning (K-IL), which incorporates external knowledge in addition to raw input data in order to learn the underlying task more rapidly and effectively. In our case, we utilized a disease-symptom knowledge graph as external knowledge and infused it into our reinforcement learning agent's state space representation.

Automatic disease diagnosis system : In [21], the authors have categorized diseases into a number of groups based on

anatomy and then trained an individual policy for each group. The proposed methodology utilizes a rule-based system for activating group policies and then investigates symptoms as per the selected policy. In case of large number of disease/groups with huge overlapping symptoms across groups, such rule based system will require huge manual effort and time. Kao et al. [22] have proposed a context symptom checker which shows that context (patient's personal information) such as patient's gender and age in addition to symptom status can enhance diagnosis performance.

Task-oriented diagnosis dialogue system : In [5], authors have proposed a unified flat dialogue policy for disease diagnosis, which shows that the additional symptoms (in addition to self-report) extracted through conversation significantly improve diagnosis performance. Xu et al. [23] extended the previous work [5] and proposed a Knowledge Routed Dialogue System (KR-DS) that incorporates external rich medical knowledge graph, in policy learning (Flat DQN) for a diagnosis system with only five diseases. Creating and incorporating such a rich medical knowledge graph (disease symptom relation, relations among different symptoms, and knowledge routed graph for policy decision) are very challenging and expensive tasks. Also, it becomes intractable when the number of diseases and symptoms are huge. The work [9] proposed a disease diagnostic system having large number of diseases using HRL based dialogue policy. The reported results suggest that as the number of diseases grows, integrated hierarchical policies are more successful than flat policy.

Limitations : Diagnosis success hugely depends on the appropriateness and distinctiveness of extracted symptoms from symptom investigation, and thus an intelligent and robust symptom investigation is the key to diagnosing patients correctly. Most of the existing works [5,9,22] employed a completely data-driven approach, where the system's efficacies are bounded by the scale and scope of the underlying dataset. These systems rely entirely on previous experiences for conducting symptom investigation depending upon patient chief complaint and dialogue context. However, real doctors' behaviors also depend on their knowledge and learning in addition to experiences. There is a recent work [23], which incorporated the disease-symptom knowledge through a rich medical knowledge graph in diagnosis learning process. The proposed method has shown a significant improvement on the DX dataset with five diseases and 41 symptoms. However, creating and incorporating such a rich medical knowledge graph (disease symptom relation, relations among different symptoms, and knowledge routed graph for policy decision) for a system with a sufficiently large number of diseases and symptoms is very challenging and expensive.

3. Problem formulation

In this section, we formulate the problem using MDP and describe our proposed HRL framework with two levels, which are being utilized for policy learning. The automatic disease diagnosis task can be formulated as a Markov Decision Process (MDP) [11]. The MDP consists of 4 attributes (S, A, TP, R): S is state space, A is action space, TP is state transition probability, and R is reward model. The state (s, s $\in S$) is composed of *virtual agent's current intent(request symptom /inform disease), current requested symptom, user informed symptoms, all confirmed symptoms, dialogue turn no., PCM state (potential disease and potential symptoms), and, reward (including learner critic)*. The action space (A) comprises of all symptoms, and each action corresponds to a symptom examination. Here, TP is state transition function that defines determines next state (s') for a given current state and action taken by the agent, i.e., $S' = TP(S, a)$. In the tuple, the last term is R which indicates reward model that implicitly supervises reinforcement learning agent for learning an underlying task. The agent gets a reward if it inspects a relevant symptom as per dialogue context or diagnoses patient with the correct disease. The agent also gets a penalty if it re-inspects an already inspected symptom. We have introduced an additional reward called learner critic, which reinforces the agent to inspect context-aware, knowledge grounded symptoms (potential symptoms). The goal of the agent is to learn an optimal policy (π) which maps state to most appropriate action, i.e.,

$$a^* = argmax_i \pi^*(a_i|s_i) \tag{1}$$

where i ranges over action space, s_i is the current dialogue state and a^* is the most appropriate action for the given state s_i and an optimal policy(π^*). A well known and one of the most established methods for optimizing a policy is reinforcement learning [24] where an agent learns through a trial and error approach. The agent's objective is to find an optimal dialogue policy that maximizes cumulative reward (R) over an episode.

$$R = \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot r_t \tag{2}$$

where N is the number of dialogue sessions in an episode, T is the maximum turn of ith session, $\gamma \in [0,1]$ is discounted factor and r_t is immediate reward which agent receives at tth time step. The discount factor (γ) determines the relative value of present and future rewards, i.e., it decides the horizon of successive rewards provided in response to action (a) on state s. If $\gamma = 0$, the agent will care only about immediate reward and $\gamma = 1$, the agent will consider the immediate reward and all consecutive rewards equally important. For most reinforcement learning problems, the discount factor is usually between 0.9 and 0.99. Thus, an agent learns to select an action, which maximizes cumulative discounted reward and leads the agent to achieve the long-term

Table 1
Symbol Table: Notation and their meaning.

Symbol	Meaning
s/s_t	Current dialogue state/dialogue state at tth time step
A	Action space
s'	Next dialogue state
R	Total reward
r_t	Reward at tth time step
Q (s, a)	Cumulative reward for action a in state s
a'	Next action
r_t^e	Extrinsic reward received by the master agent at tth time step
a_t	Agent's action at tth time step
θ_m, θ_w	Master's policy parameter, Worker's policy parameter
r_t^w	reward received by the worker policy at tth time step
π	Dialogue policy
γ_m	Master policy discount factor (0.9)
γ^w	Worker policy discount factor (0.9)
S	set of entire symptoms
S_t	set of extracted symptoms till tth dialogue turn
P (D $\|S_t$)	Probability distribution of diseases for the given symptom set S_t at tth time step
I(s_t)	probability of the most probable disease d for given symptom set s_t
HRL	Hierarchical reinforcement learning
PCM	Potential candidate module
KI-CD	Knowledge infused context driven
HDC	Hierarchical disease classifier
WCC	Worker confidence critic
RL	Reinforcement learning
DQN	Deep q network
DDQN	Double deep q network
AMR	Average match rate
SIR	Symptom investigation rate
UDP	Unified dialogue policy

goal. The discounted cumulative reward ($\sum_{t=0}^{T} \gamma^t \cdot r_t$) can be approximated through the following Bellman equation [25] :

$$Q^*(s, a) = \mathbb{E}[r + \gamma \cdot max_{a' \in A} Q^*(s', a')] \quad (3)$$

The Q function ($Q(s, a)$) is the state–action value function, which learns to approximate agent action (a)'s appropriateness and efficacy for a given state/observation, s. Here, $Q^*(s, a) = max_{\pi^*} Q^{\pi}(s, a)$, is an optimal state–action ($Q(s, a)$) value which represents maximum possible cumulative discounted reward for action a for a given state s. s is the current state, a is the current action which leads to new state s'. A policy (π) will be an optimal policy (π^*) if and only if it selects the most optimal action for every state and action, i.e., $Q^{\pi}(s, a) = Q^*(s, a)$ for $\forall s \in SS, \forall a \in A$. The symbol SS indicates state space. For an instance, consider a dialogue state (s) = {cold} and the two actions a, aa = inspect(cough), inspect(shoulder pain). Initially (t=0), $Q(s, a) = 0, Q(s, aa) = 0$. The agent selects an action a on the given dialogue state s, it gets an immediate reward of r (let us say 0.8 as the action (cough) is quite relevant to state (cold)) and leads to a new state s'. The $Q(s, a)$ value will be updated as follows : $Q(s, a) = r(0.8) + \gamma \cdot Q(s', a') = 0.8 + \gamma \cdot 0 = 0.8$. But if the agent would have taken the action aa, the agent gets a penalty of r' (let us say -0.5 as the inspected symptom, shoulder pain is not much relevant to cold) and the state action will be updated as follows : $Q(s, aa) = r'(-0.5) + \gamma \cdot Q(s', a') = -0.5 + \gamma \cdot 0 = -0.5$. The $Q(s, a)$ values keep on updating with training experiences until they converge and the converged Q state value function is utilized for estimating the most appropriate action for a given dialogue state in testing.

Policy optimization : One of the major challenges for automatic disease diagnostic systems is to deal with a vast symptom and disease space (action space). A single/unified policy fails to learn an optimal behavior for conducting symptom investigation and disease identification in a diagnostic system with large no. of symptoms and diseases [9]. The main reason behind the failure is the underlying huge action space (all symptoms + all diseases + additional actions such as greeting and close), which requires to be handled by a single policy. Thus, we have utilized a hierarchical structure of two layers for policy learning. The first layer policy (master) selects an appropriate lower policy, and the selected lower-level policy (worker) conducts symptom investigation using the bayesian inspired diagnosis method, controlled and reinforced through the PCM and learner critic, respectively. Both the master and worker policies are optimized using Deep Q Networks (DQNs) [12]. We selected the DQN algorithm for policy optimization primarily because of its proven efficacy for dialogue optimization task [5,6,26]. Also, the DQN algorithm is more sample efficient and converges much faster compared to policy-based algorithms such as policy gradient [27]. We have also experimented with two more algorithms, namely, Double deep Q network (Double DQN) [27] and Dueling deep Q network (Dueling DQN) [28].

4. Proposed framework

In an automatic diagnosis dialogue system, patients and doctors converse about patients' difficulties, complications, and other symptoms to identify the disease they are suffering. In a typical case, a patient first informs his/her major difficulties (explicit symptoms), doctors conduct a detailed symptom investigation after enquiring about a set of symptoms, the patient responds to each of them whether he is suffering from such symptom or not or even not sure. Finally, the doctor informs the most probable disease as per the investigated symptoms. Thus, there are mainly two sub-tasks: 1. Symptom investigation, and 2. Disease identification. The proposed system utilizes a potential candidate module (PCM) and leaner critic incorporated hierarchical dialogue for context-driven knowledge-guided symptom investigation. Once the agent collects adequate symptom information, it activates a hierarchical disease classifier for disease identification. The detailed architecture of the proposed model has been illustrated in Fig. 2.

Dialogue policy leaning. As mentioned, the agent's primary challenge is inevitably large action space; to alleviate this problem, we have extended the RL problem formulation to a hierarchical structure with two-level policies. The proposed architecture using options framework [29] has been shown in Fig. 2. It is composed of two-level policies: the higher one named as the master policy that decides which worker (department) should be assigned, and the lower one called worker policy, which takes primitive actions such as symptom request. Diseases are divided into h groups (D_1, D_2... D_h) as per the International Classification of Diseases (ICD-10-CM). Each worker policy (W_i) corresponds to a set of diseases (D_i) having a set of symptoms, $S_i \subset S$. The framework has five main components: Master, Worker, Hierarchical Disease Classifier, Internal Critic and Potential Candidate module & Learner Critic.

Master. The master policy is responsible for activating an appropriate worker (W_i) or disease classifier depending upon the current dialogue state. The action space of master policy is composed of worker policies (W_1, W_2 .... W_h) and disease classifier (D). At each time stamp t, it takes dialogue state (s_t) as input and selects either any lower policy (W_i) or D which leads to a new state (s_{t+1}). The master gets a reward r_t as follows :

$$r_t^m = \begin{cases} \sum_{t_l=1}^{n} \gamma_m^i r_{t+t_l}^e, & \text{if } a_t^m = W_i \\ r_t^e, & \text{if } a_t^m = D \end{cases} \quad (4)$$

where n is the number of primitive actions taken by worker W_i, γ_m is the discounted factor of the master policy and r_t^e is the extrinsic reward of the master at turn t. In response to the taken master action (a_t^m: W_i), the triggered worker agent/policy takes a series of actions (a_i), i.e., the worker agent inspects a series of disease group-specific symptoms and each time it gets an intrinsic reward (r) depending upon the executed action's appropriateness to the current state. The agent selects an action depending upon the observed state and policy (π), i.e., $a = \pi(s)$. In our case, the policy function is optimized using a deep neural network, where the policy parameter (θ) is analogous to the neural network's weight (WW^*). It can be expressed as follows : $a = \pi_\theta(s|WW^*)$ where s is dialogue state, pi is the policy function with parameter $\theta = WW^*$. The master action gets the reward for action, a_t^m, which is discounted sum of rewards accomplished by the triggered worker for its taken action (a_i). Thus, it is known as an extrinsic reward. The agent's goal is to maximize total extrinsic reward over an episode. Thus, the temporal difference error (TD) [30] is backpropagated to the top-level policy network as follows:

$$L_\theta^m = [(r^m + \gamma_m max_{a'} Q_t^m(s', a', \theta_m^-)) - Q_{t-1}^m(s, a, \theta_m)]^2 \quad (5)$$

where L_θ^m is the loss at tth time step, which is difference between state value $Q(s, a)$ calculated through current policy parameter (θ_t) and previous policy parameter (θ_{t-1}). Here, θ_m is the current policy network parameter, θ_m^- is the frozen parameter during the last training iteration.

Worker. The worker policy (W_i) gets invoked by the master policy, which takes updated state (s_t) as input and selects the most appropriate primitive action (a_t). The action space of a worker policy (W_i) consists of its group symptom request and disease classification action. The worker ((W_i) gets an immediate intrinsic reward (r_t) for action a_t. The agent gets the reward/penalty (r_t) depending upon its current action's (a_t) appropriateness to

the current state (s_t). It is called intrinsic reward because the worker receives the reward as a direct result of its current action (a_t). The worker policies are also optimized using a dedicated deep q network corresponding to each disease group (disease department).

Internal critic. During symptom investigation, the worker agents (W) take action (inspecting a probable symptom) depending on current dialogue state, which includes all the status of all already inspected symptoms. For each action, the worker agent gets a reward/penalty from the underlying environment (user) depending upon the action's relevance with current dialogue state. The reward/penalty is provided to the worker as a critic, and the agent learns to optimize its behavior by maximizing the cumulative reward. Also, the cumulative discounted worker's reward (extrinsic reward: r_t^m) provided to the master agent guides for selecting an appropriate worker (W_i) for a given context/state. In our case, the worker gets intrinsic reward as Eq. (6) where match$_i$ signifies that the worker requests a symptom that the patient is suffering from. N_{SubMax} is the upper bound of the number of turns allowed in a single disease group (worker).

$$r_t^w = \begin{cases} +w_1, & \text{if } match_i = 1 \\ -w_2, & \text{if repetition, or turn=}N_{SubMax} \\ +SR, & \text{if success} \\ 0, & \text{Otherwise} \end{cases} \tag{6}$$

Potential Candidate Module (PCM) and Learner Critic. In the disease diagnosis task, doctor/autonomous agent predicts a disease (d) based on investigated symptom status SS = $\{S^i|$ i = 1,2,…,n$\}$, i.e., d = $argmax_i P(D_i|SS)$. It is challenging to investigate an adequate and minimal set of symptoms for proper diagnosis because of huge no. of associated symptoms. So, in real life, doctors generally go in reverse order, i.e., they hypothesize a set of candidate diseases ($C = C_1, C_2 ... C_m$) based on the patient's self-report and informed symptoms(S_t), and then they inspect potential symptoms of candidate disease. It can be expressed as follows:

$$P(D|S_t) = \frac{P(S_t|C_i)P(C_i)}{P(S_t)} \tag{7}$$

where C_i = argmax$_i$ P(C_i | S_t) and S_t is patient's symptom status at tth turn. In PCM, Pathogen module takes candidate disease (C_t) as input and outputs x topmost contributing symptoms (PCS$_t$). In order to extract the most contributing symptoms of disease, this module utilizes Disease-Symptom knowledge where each disease corresponds to its respective symptoms with some confidence values. The pathogen triggers the next layer only if the topmost disease potential score is more than a threshold value (C_{th}) to avoid exploitation with very limited symptom information. The next layer, Context Synchronized Pathogen, filters out non-appropriate symptoms as per the context by discarding symptoms that are present in either self-report or dialogue context (dialogue state). The knowledge about candidate disease and potential symptoms is infused in dialogue state, which can be expressed as follows :

$$C_t = \{C_i, P(C_i)|S_t, i = 1, 2..m\} \tag{8}$$

$$PCS_t^i = PCM(C_t^i) \tag{9}$$

$$s_t = StateUpdate_{i=1}^{i=m}(PCS_t^i, C_t^i) \tag{10}$$

where C_t represents most probable m candidate diseases along with their probabilities for the given symptom set (S_t) at time t, PCS_t^i signifies k most contributing symptoms of candidate disease C_i at tth turn. Then, the current state (s_t) is infused with PCS_t. This infused knowledge feeds the agent's experience (learning) to the dialogue state, leading to an intelligent investigation, i.e., the agent inspects mostly relevant symptoms that the patient might have. It guides the agent through a critic called Learner Critic, which is defined as follows :

$$r_t^l = \begin{cases} w_3 \cdot P(C_t), & \text{if } ars_t \in PCS_t \text{ and } is_t = 1 \\ -w_4 \cdot P(C_t), & \text{if } ars_t \notin PCS_t \text{ and } is_t = 0 \\ 0, & \text{Otherwise} \end{cases} \tag{11}$$

where $P(C_t) = \frac{\sum_{k=1}^{k=m} P(C_t^k)}{m}$, ars_t denotes agent requested symptom at turn t, is_t = 1 signifies that the patient suffers from the symptoms which the agent has requested at tth turn. Here $P(C_t^k)$ denotes probability of kth candidate disease C_k at tth time step.

Worker Confidence Critic. Sparse reward is one of the well-known problems of reinforcement learning, which either leads to non-convergence or convergence at local optima. To avoid this issue and motivate the master agent for triggering an appropriate worker, an additional immediate reward is provided to the master agent at each time step t, which is as follows :

$$wcc_t = w_5 \cdot max(0, \quad [I(s_t) - I(s_{t-1})]) \tag{12}$$

Here, $I(s_t) = P(d_{s_t} \mid s_t)$ represents immediate reward, which is the probability of disease d (disease with maximum probability) for the given non terminal dialogue state (s_t) and w_5 is a reward parameter.

User Simulator. A user simulator is an essential and crucial component for training an RL agent as it needs real-time interaction. The user simulator initializes each dialogue session with a user goal randomly selected from training samples. It informs self-report (all explicit symptoms) to the agent at first turn and requests disease. Then, during the conversation, the simulator responds to each agent's request for symptoms as per the sampled goal. If the number of turns reaches the maximum limit or the agent informs a disease, it ends the conversation.

Hierarchical Disease Classifier (HDC). In order to alleviate the sparsity issue, we build a hierarchical disease classier, which consists of a worker/group classifier followed by a disease classifier. Once a worker (W$_i$) selects disease classification as action, this module gets activated with input as patient's symptoms and returns the probability distribution over all diseases belonging to the worker (W$_i$). In the lower level of the hierarchical disease classifier, each deep learning-based disease classifier is a three-layer /one hidden layer neural network with the structure as follows: Input layer (size - number of symptoms in a disease group), Hidden layer (256 neurons) and Output layer (size: number of diseases in a disease group, i.e., 10). The architecture is shown in Fig. 3.

5. Dataset

The proposed agent is trained and tested on the publicly available large medical dataset, Synthetic dataset (SD) [9] which is constructed based on SymCat[2] database, and the diseases are grouped into different disease groups (G1, G4, G5, G6, G7, G12, G13, G14, G19) as per International Classification of Disease (ICD-10-CM)[3]. To the best of our knowledge, it is the only publicly available English dataset to train a conversational agent for disease diagnosis. The data statistics are shown in Table 2. The number of unique symptoms of different ontological disease groups are shown in Fig. 4(a). Fig. 4(b) depicts symptom space across

[2] www.symcat.com

[3] https://www.cdc.gov/nchs/icd/

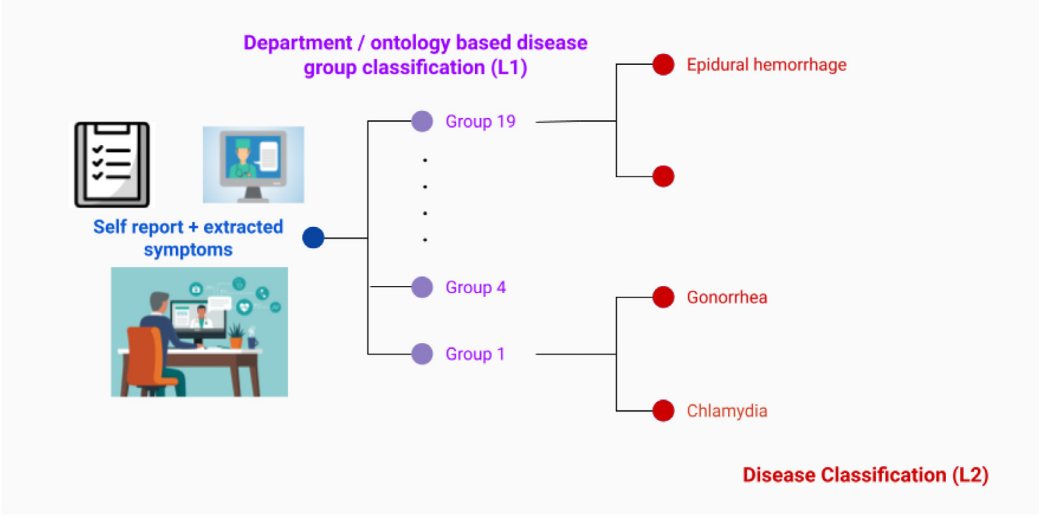


Fig. 3. Architecture of Hierarchical Disease Classifier (HDC), where the first level classifier selects disease group and the second level classifiers diagnose the patient's suffering disease.

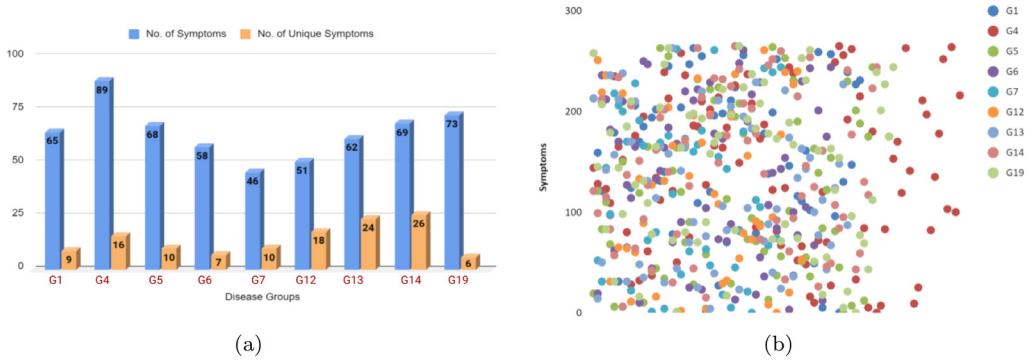


Fig. 4. a. Distribution of total number of symptoms and unique symptoms across the disease groups, b. Symptom space across different disease groups.

different disease groups, demonstrating that the symptom sets of different groups are hardly isolated or well separated. Fig. 5 shows the number of unique symptom distributions across different diseases. It can be observed that there are very few unique symptoms across diseases /groups, i.e., huge overlapping among multiple groups and different diseases, which significantly intensifies the complexity of the diagnosis problem.

Fig. 5 illustrates the distribution of number of unique symptoms and total number of symptoms across different disease groups (G1, G4). There are nine disease groups, and each group has ten diseases. The left plot contains counts of total symptoms (blue) and unique symptoms (red) corresponding to the diseases mentioned. The right plots show the number of total symptoms and unique symptoms of different diseases of group 1 (G1) across the symptoms of diseases belonging to group 1 only. For instance, there are 12 symptoms for Chlamydia, and it does not contain a single symptom that is unique across all diseases (left plot, first data point). The disease (Chlamydia) contains four unique symptoms across its corresponding diseases as there are only symptoms of the group's diseases(only ten diseases). The distributions for other groups (G5, G6, G7, G12, G13, G14, and G19) are reported in the Appendix section.

6. Experiments and results

We trained, evaluated, and compared our proposed framework with the existing state-of-the-art method and other baselines on the existing benchmark dataset.

Table 2
SD dataset statistics.

Entries	Value
Total no. of patient samples	30,000
Total no. of diseases	90
Total no. of disease groups	9
Total no. of symptoms	266
Avg. length of self reports	1
Avg. length of implicit symptoms	2.6

6.1. Implementation details

The agent is trained and tested with 80% (24,000) and 20% (6,000) of total data, respectively. For sufficient exploration and to avoid local optimum, the proposed model (both master and department policies) utilizes ϵ — greedy technique in training, i.e., it explores the action space with ϵ probability, and with remaining $1 - \epsilon$ probability, it selects action through the policy networks by providing current state as observation. Both master and worker policies are four-layered neural networks (NN) with an input layer, two hidden layers of 512 neurons, and an output layer. The hierarchical disease classifier consists of two layers: disease group classifier and worker-wise disease classifier. There is one hidden layered neural network in the first level, with 256 neurons having an input layer with a size equivalent to symptom space (no. of symptoms: 266) and an output layer with nine neurons (disease group). The lower layer has nine department-specific disease classifiers, which are also one hidden layered

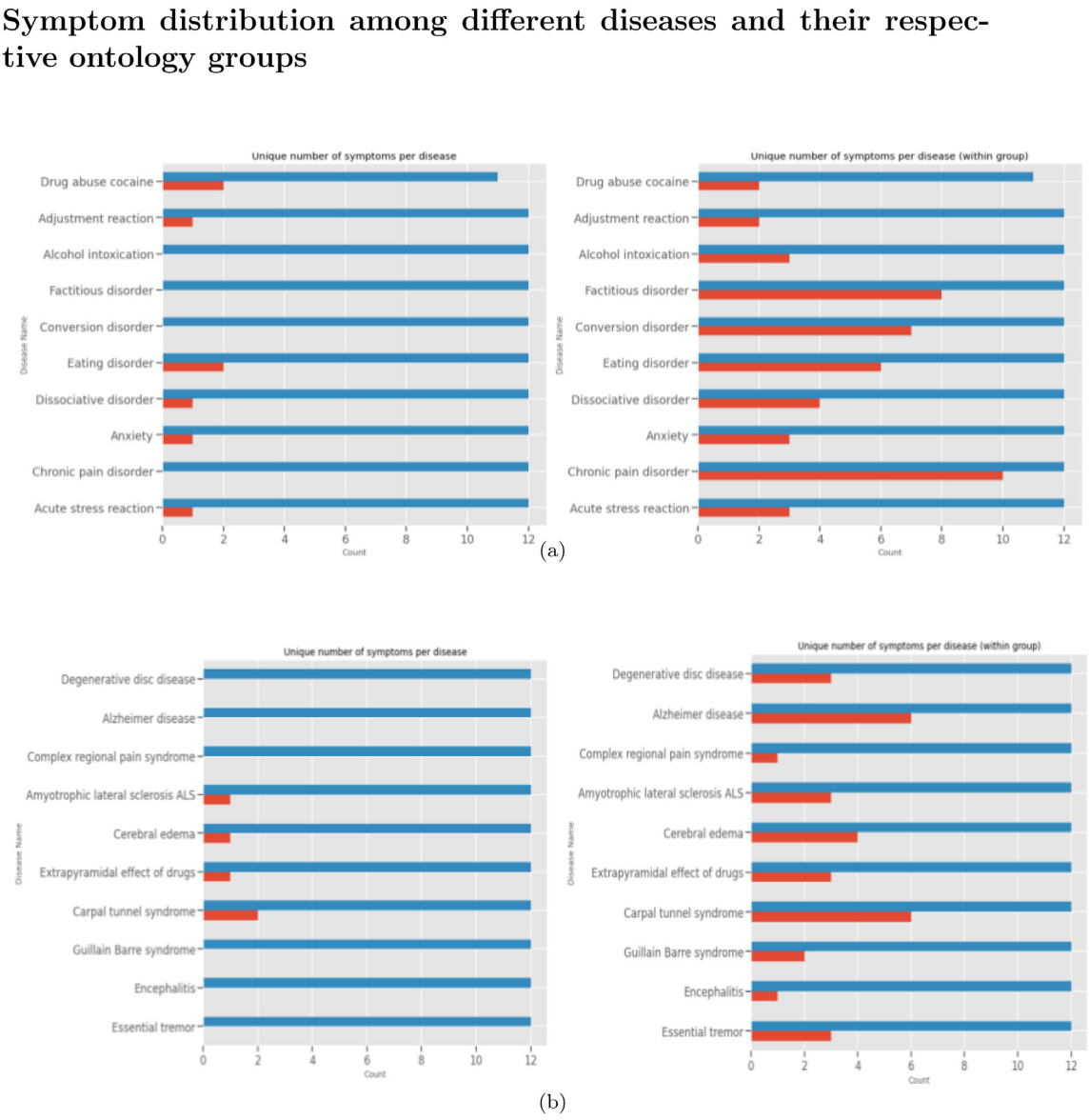


Fig. 5. The left plots of each figure show unique symptom count (red color) distribution across all diseases, whereas the right plots illustrate unique symptom count (red color) across diseases of their respective group. a. Disease Group 1, b. Diseasse Group 4.

deep neural network having an input layer with a size equivalent to the number of symptoms of the corresponding group and an output layer with 10 nodes (no. of diseases in each disease group). The other hyperparameters are as follows : learning rate — 0.0004, batch size — 100, optimizer — Adam, loss function — CrossEntropy. All the hyperparameters and parameters are chosen through rigorous empirical analysis, and the final values are reported in Table 3. We have also mentioned the pseudo-code below.

6.2. Performance

We have utilized the most popular automatic diagnosis evaluation metrics (diagnosis success rate, dialogue length, and disease classifier accuracy) [5,6,23] for evaluating our proposed agent's performance, comparing with state of the art methods and other baselines. We have also reported the learning curves of different agents during training in Fig. 6. In addition to these metrics, we have also evaluated the model in terms of average matching rate (AMR) and symptom identification rate (SIR) for measuring the effectiveness of symptom investigation. AMR can be defined as the proportion of agent's symptom request that patient truly suffers, whereas SIR is the ratio of the number of extracted implicit symptoms to the total number of implicit symptoms. The metrics can be defined as follows:

$$\textbf{Success rate} = \frac{\sum_{i=1}^{i=EL} DS_i}{EL} \tag{13}$$

$$\textbf{Avg reward} = \frac{\sum_{i=1}^{i=EL} \sum_{j=1}^{j=t} r_{ij}}{EL} \tag{14}$$

$$\textbf{Average match rate (AMR)} = \frac{\sum_{i=1}^{i=EL} m_i/r_i}{EL} x100,\ m_i \tag{15}$$

$$\textbf{Symptom identification rate (SIR)} = \frac{\sum_{i=1}^{i=EL} m_i/t_i}{EL} x100 \tag{16}$$

where EL (Episode length) denotes no. of simulated dialogues in an episode, $DS_i = 1$ if the ith dialogue ends successfully, i.e., the agent informs correct disease, otherwise 0. The term, r_{ij} represents reward received by agent in jth turn of ith dialogue session of an episode. Here t represents no. of dialogue turns

Table 3
Hyper-parameter and parameter values utilized in the proposed and the baseline (HRL) methodology.

Hyper-parameter Parameter	Value (Proposed methodology)	Value (Baseline, HRL)
No. of training episodes (N)	5000	5000
No. of dialogue sessions in an episode	100	100
Learning rate (lr)	0.0005	0.0005
Epsilon (ϵ)	0.1	0.1
Turn limit	26	26
Batch size	100	100
Department's turn limit (N_{SubMax})	5	5
Master's discount factor (γ_m)	0.95	0.90
Worker's discount factor (γ^w)	0.95	0.90
No. of potential symptoms of candidate disease(x)	5	NA
Reward for success (SR)	+3	+3
No. of potential disease (m)	1	NA
Potential disease probability threshold (C_{th})	0.5	NA
Reward parameters {w_1, w_2, w_3, w_4, w_5}	{1, −1, 0.95, −0.95, 0.95}	{1, −1, NA, NA, NA}
Optimizer	Adam	Adam
Loss function	cross entropy	cross entropy

in *j*th dialogue session. m_i indicates the total number of agent's requested symptoms, which belongs to the patient's suffering symptoms. Here, r_i signifies the total number of symptoms requested by agent during *i*th conversation. The term t_i denotes the total number of true implicit symptoms of the patient *i*th conversation. The different baselines and state of the art methods which are used to compare our proposed approach are as follows :

- **SVM ex**: It is a support vector machine (SVM) [31] based model which takes a patient's self-report (only explicit symptoms) as input and predicts one of the diseases. The model has been trained in a supervised fashion, where input is a one-hot encoding of explicit symptoms (patient self-report) and output is a disease tag.
- **SVM ex-im**: It is also an SVM model, but it considers both explicit as well as implicit symptoms and classifies diseases. The model considers patients' symptoms (both explicit and implicit) as inputs, and it predicts probability distribution over diseases. This is the same as the previous model, with one difference that it also takes implicit symptoms in addition to patient self-report as input. As it considers patients' entire symptoms, its performance should be treated as upper bound for any reinforcement learning-based dialogue agent.
- **Unified Dialogue Policy (UDP) /Flat DQN**: It is a unified dialogue policy optimized through a single DQN with an action space that includes both disease and its symptoms. The methodology is similar to the state-of-the-art method, DS-MD [5] which utilizes flat DQN for diagnosis on a much smaller dataset with only four diseases. There is only one policy network, which takes context (symptoms) as observation and takes appropriate action from action space, A. The action space (A) includes both symptoms and diseases, i.e., $A = symptom \cup disease$. Thus, the agent either inspects a symptom or diagnoses a disease depending upon extracted symptoms (through conversation) and chief patient complaint(self-report). The policy network is a four-layered neural network with architecture as follows: Input layer (size - number of symptoms), Two hidden layers of size 512 neurons, and Output layer (size - number of symptoms + number of diseases).
- **HRL**: HRL [9] is the state-of-the-art method for the automatic disease diagnosis task, which has been trained and validated on the largest English benchmarked SD dataset. The proposed methodology employs an HRL framework that consists of two-level policies and a disease classifier. The higher-level policy is responsible for choosing an appropriate disease group, while the lower-level policies are responsible for conducting disease department aware symptom investigation. The higher-level policy triggers one of the lower-level policies depending upon observation state (symptoms), and the lower-level policy conducts symptom investigation followed by disease prediction by the disease classifier. Thus, there are *N* (number of disease groups)+1 dialogue policies, each consisting of four layers of neural networks having two hidden layers of size 512. The disease classifier is also a deep learning-based model with one hidden layer of 256 neurons.
- **KI-CD with only PCM** : Our proposed model (KI-CD) with only PCM component, i.e., KI-CD without hierarchical disease classifier.
- **KI-CD with only HDC** : Our proposed model (KI-CD) with only hierarchical disease classifier, i.e., KI-CD without PCM.

Fig. 6(a) illustrates how success rates of the different agents are increasing over training episodes, which signifies the agent behavior is getting optimized. We also observed that the KI-CD_HDC is doing significantly better in training because of improved disease classification accuracy; however, it does not perform that way with unseen cases and fails to generalize and identify new cases. Fig. 6(b) shows dialogue length of different baselines over training episodes. Our proposed agent's dialogue length is comparatively higher than the HRL agent as it attempts to investigate significantly more symptoms for ensuring right diagnosis. Fig. 6(c), 6(d) shows match rate and symptom identification rate over training episodes, respectively. Our proposed agents' SIRs (#correctly guessed symptom / # implicit symptom) are much higher than the HRL agent, which depicts the agent's intelligence /learning of requesting symptom in proportion to length of implicit symptom through the infused knowledge of potential candidate disease and its symptoms. Fig. 6(e), 6(f) illustrates reward and diagnosis success of our best performing (KI-CD_PCM) policy.

The complete results have been reported in Table 4, which determine the effectiveness of the proposed framework. The reported values (Table 4) are average of 5 runs. We have also experimented with two more deep reinforcement learning algorithms, namely Double deep Q network (DDQN) and Dueling deep Q network (Dueling DQN). The obtained results are reported in Tables 5 and 6. The summaries of the findings are as follows:

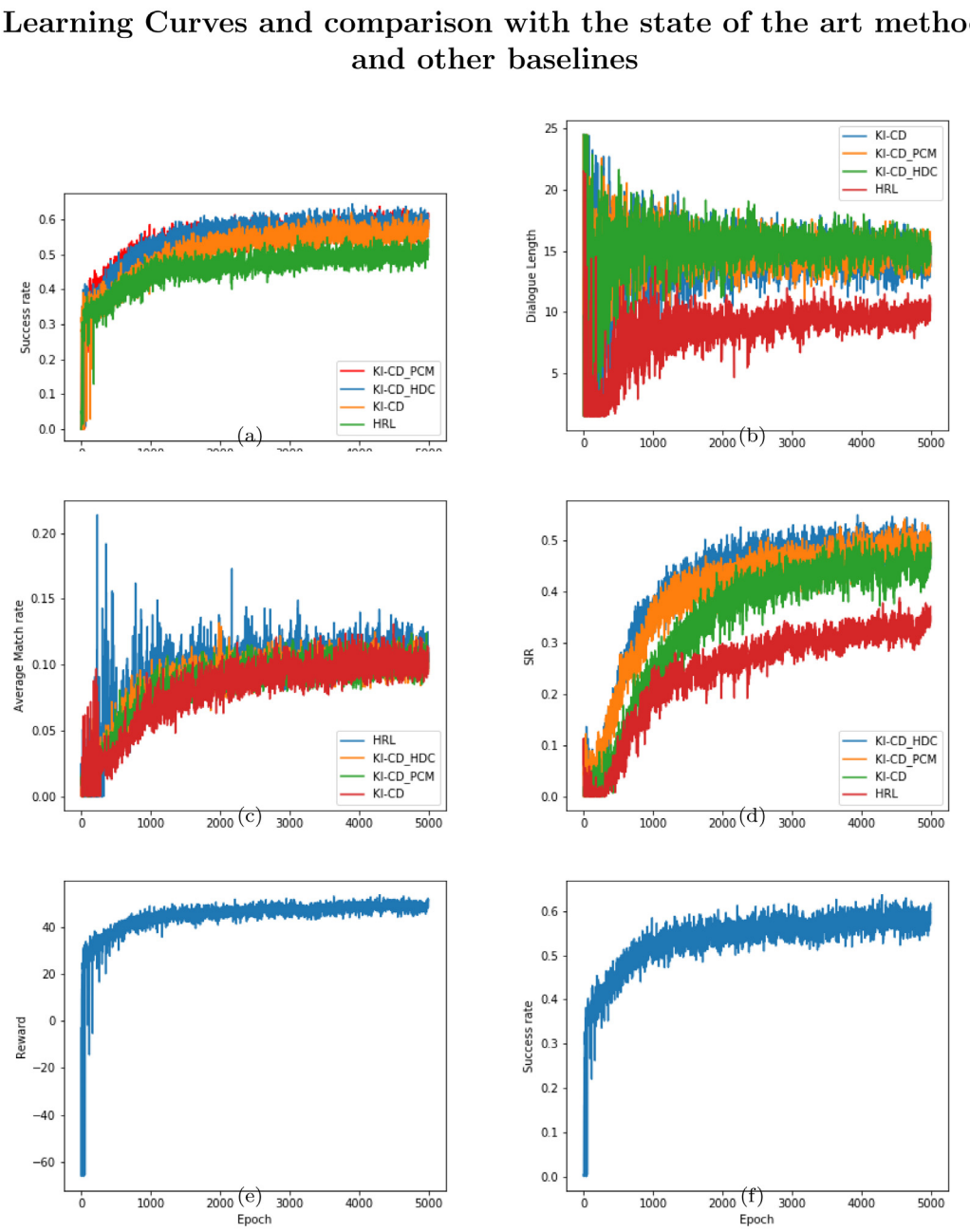


Fig. 6. Learning Curves and comparison with the state of the art method and other baselines a. Success rate over training episodes, b. Dialogue Length over training episodes, c. Average match rate (AMR) over training episodes, d. Symptom identification rate over training episodes, e. Reward curve of KI-CD_PCM agent over training episodes, f. Success rate of KI-CD_PCM agent over training episodes.

Table 4
Performance of our proposed agents, state-of-the-art model (HRL) and other baselines with DQN as policy optimization algorithm. The reported values are average of five iterations, and ± indicates the obtained standard errors.

Agent	Success rate	Dialogue length	AMR(%)	SIR(%)	Disease classifier accuracy (%)
SVM ex	0.322	NA	NA	NA	NA
UDP (Liu et al. 2018)	0.342 ± 0.0056	5.34 ± 0.0153	02.41 ± 0.0011	01.26 ± 0.0005	NA
HRL(Liao et al. 2020)	0.504 ± 0.018	12.95 ± 0.704	10.49 ± 0.002	29.56 ± 0.0016	0.4980 ± 0.0022
KI-CD_HDC (KI-CD with HDC only)	0.566 ± 0.0074	14.93 ± 0.0885	11.01 ± 0.0014	50.38 ± 0.0036	0.5974 ± 0.0034
KI-CD_PCM (KI-CD with PCM only)	0.575 ± 0.0083	15.20 ± 0.1040	10.72 ± 0.0011	49.23 ± 0.0034	0.5784 ± 0.0047
KI-CD (both PCM & HDC)	0.569 ± 0.0058	16.11 ± 0.0850	09.95 ± 0.0012	48.33 ± 0.0031	0.5975 ± 0.0029
SVM ex-im	0.731	NA	NA	NA	NA

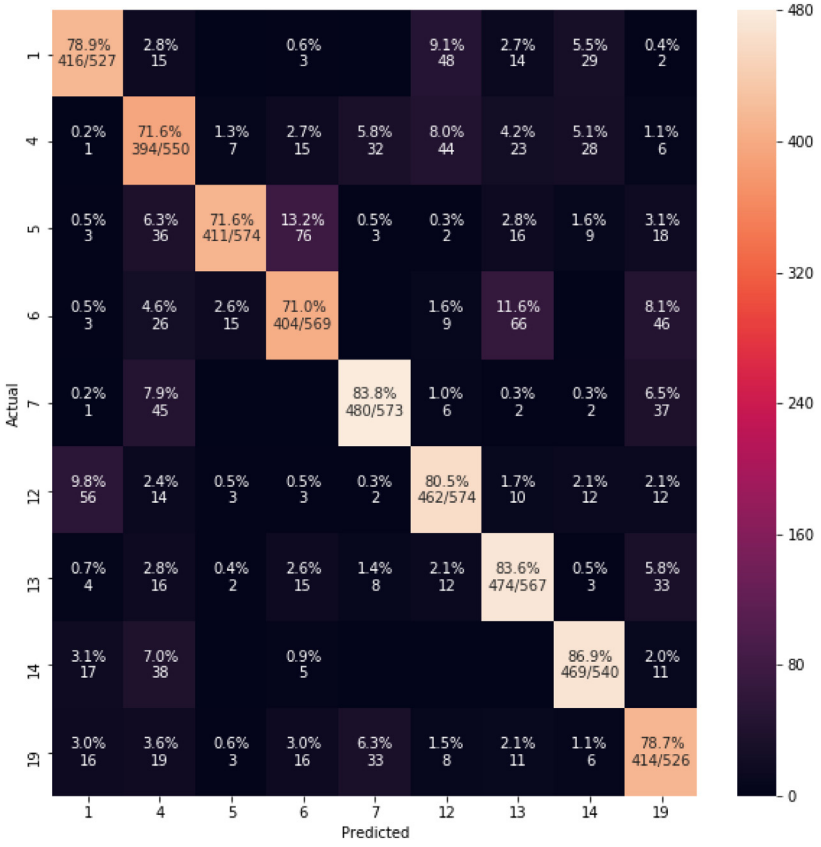


Fig. 7. Confusion matrix across different disease groups.

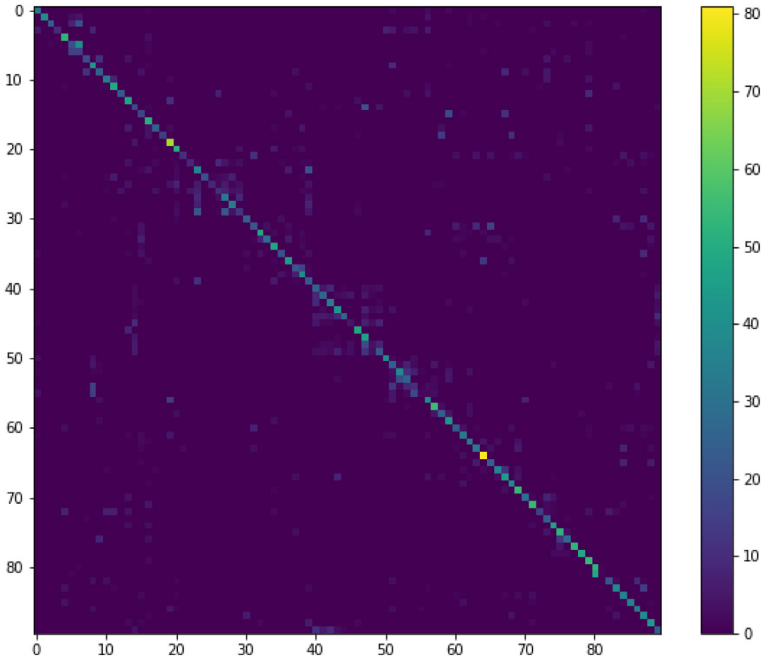


Fig. 8. Confusion matrix across different diseases.

Algorithm Proposed Knowledge Infused, Context Driven (KI-CD) Automatic Disease Diagnosis Algorithm

Initialize : State Tracker, User environment (User), Experience replay memory for master policy (M), Experience replay memory for worker policy (M_{w1}, M_{w2}, ... M_{w9}), Disease classifier replay memory (DCR), Master policy(π^m), Worker policies (π^{w1},π^{w2} ....π^{w9}) - Two policy networks : i. Target network (θ), with the state-action value function $\widehat{Q}$(S,a) , ii. Behavioral network ($\widehat{\theta}$) with the state-action value function, Q(S,a), Hierarchical disease classifier (HDC)

repeat

  Reset environment, State , candidate diseases (C), potential symptoms (PS), learner critic (lc) = 0, intrinsic reward = 0, extrinsic reward = 0, worker confidence critic (wcc) = 0, patient data (DS), done = false

  repeat

    State = Update(state tracker, patient self reported symptoms)

    master action (a^m)=argmax$_j$ π^m (A_j | State) ▷ A = W_1, W_2, .. W_9, DC (disease classifier)

    activated worker policy (π^w) = π^{wp}, p = a^m if $a^m \neq$ DC, else activate disease classifier (DC)

    worker action (a^w) = argmax $_k$ π^w (A_k | State)

    candidate disease (C) = HDC (State)

    potential symptoms (PS) = PCM (State, C)

    lc = Learner Critic (PS, a^w)

    user action (u) = User(a^w) ▷ u ∈ {true, false, not sure}

    intrinsic reward(ir) = Internal Critic(u, done)

    r = ir + lc

    wcc = Worker Confidence Critic (State, a^w)

    mr = r + wcc ▷ mr denotes master's reward

    extrinsic reward (er) = $\begin{cases} \sum_{t'=1}^{n} \gamma_m^t mr_{t+t'}, & \text{if } a^m = W_i \\ mr, & \text{if } a^m = DC \end{cases}$

    done = $\begin{cases} 1, & \text{if } a^m = \text{DC} \\ 0, & \text{Otherwise} \end{cases}$

    Next state (S') = Update(state, a^w, u)

    true disease = disease(DS)

    Append the experience tuple (State,a^w, S', r, done) to the worker's experience replay memory(M_{wp})

    State = S'

  until done

  D = HDC(state) ▷ disease prediction

  Append the experience tuple (State, a^m, S', er, done) to the master's experience replay memory(M), Append (State, true disease) to the disease classifier replay memory (DCR)

  if(TrainingEpisode % 100)

  train disease classifier networks by back propagating the categorical cross entropy loss between predicted and true disease probability distribution

    E = Sample random mini-batch of experiences from M ▷ Master policy, workers' polices (W_1, W_2 .... W_9)

    target$_i$ = $\begin{cases} r, & \text{if done = 1} \\ r + \gamma \cdot max_{a'} \widehat{Q}(S', a'), & \text{Otherwise} \end{cases}$

  $\theta_{k+1} = \theta_k - \alpha \cdot [(Q(state, a) - target)^2]$ ▷ a : agent action, r : reward

  $\widehat{\theta} = \theta$

until convergence ▷ Number of training episodes

Table 5
Performance of different agents with Double DQN as policy optimization algorithm. The mentioned values are the average of five random consecutive experiment iterations.

Agent	Success rate	Dialogue length	AMR(%)	SIR(%)	Disease classifier accuracy (%)
UDP (Liu et al. 2018)	0.337	5.43	2.82	1.32	NA
HRL(Liao et al. 2020)	0.478	8.57	13.80	29.24	48.4
KI-CD_HDC (KI-CD with HDC only)	0.524	13.16	11.10	38.23	56.2
KI-CD_PCM (KI-CD with PCM only)	0.528	13.91	10.14	37.87	53.1
KI-CD (both PCM & HDC)	0.523	14.71	11.02	40.03	54.8

Table 6
Performance of different agents with Dueling DQN as policy optimization algorithm. The mentioned values are the average of five random consecutive experiment iterations.

Agent	Success rate	Dialogue length	AMR(%)	SIR(%)	Disease classifier accuracy (%)
UDP (Liu et al. 2018)	0.357	5.42	2.48	1.16	NA
HRL(Liao et al. 2020)	0.426	7.02	13.52	22.26	43.01
KI-CD_HDC (KI-CD with HDC only)	0.488	8.10	15.18	33.44	52.29
KI-CD_PCM (KI-CD with PCM only)	0.494	11.5	12.64	35.88	48.80
KI-CD (both PCM & HDC)	0.507	10.60	13.70	39.04	53.57

Table 7
Performance of our proposed agents and HRL agent on MDD dataset. The reported values are average of five iterations.

Agent	Success rate	Dialogue length	AMR(%)	SIR(%)	Disease classifier accuracy (%)
HRL(Liao et al. 2020) with DQN	0.771	5.21	19.50	29.10	77.14
KI-CD_PCM with DQN	0.808	6.16	14.66	32.10	81.79
HRL(Liao et al. 2020) with DDQN	0.753	9.61	18.80	24.20	75.34
KI-CD_PCM with DDQN	0.778	6.37	15.30	31.80	77.85
HRL(Liao at al., 2020) with Dueling DQN	0.751	4.90	23.70	21.66	75.31
KI-CD_PCM with Dueling DQN	0.757	4.68	16.40	27.79	75.68

- The proposed agent, KI-CD_PCM, outperforms the state-of-the-art method (HRL) in all metrics other than dialogue turn by a significant margin. It improves diagnosis success rate, AMR, SIR and disease classification accuracy by 7.1%, 0.23%, 19.67% and 8.04%, respectively. Furthermore, the proposed model has also outperformed all the baselines and state-of-the-art methods (HRL) with the other two algorithms (DDQN and Dueling DQN).
- Diagnosis success rate is the most crucial factor for any automatic disease diagnosis system, which directly influences the useability of the system. Our proposed agents are taking a few more turns compared to the state-of-the-art agent (HRL) in order to ensure sufficient investigation and correct diagnosis.
- The incorporated additional knowledge companion (PCM) guides the agent in conducting an intelligent and grounded symptom investigation, which leads to an improved diagnosis success rate, AMR, and SIR. Also, it elevates the disease classifier performance because it provides more grounded and relevant symptoms as input to the disease classifier for disease prediction.
- The utilization of the hierarchical disease classifier alleviates input state (symptom status) sparsity issue by a large extent and utilizes a more focused symptom/disease representation, leading to significantly improved classification accuracy (HRL - 0.4980, KI-CD$_{HDC}$ - 0.5974).
- The KI-CD (both PCM and HDC) agent does not perform as expected, i.e., better than both KI-CD_HDC and KI-CD_PCM. In a detailed analysis, we observe that the restriction of current disease group for candidate disease and potential symptom recommendation degrades the quality and relevance of potential symptoms. This restriction occurs because the KI-CD agent utilizes the hierarchical classifier also for candidate disease prediction and hence potential symptom recommendation.
- The qualitative analysis shows that the proposed RL agent acquires some human-like intelligent decision-making behavior, such as inspection/projection depending upon the environment state.

There is only one dataset (SD) with disease department annotation available in English. However, we experimented with the Medical dialogue dataset (MDD), having 12 diseases and 118 symptoms. As the dataset does not contain disease department tag, we considered 12 disease groups (D1, D2 ... D12), each having one disease. There are 1912 and 239 training and testing samples, respectively. The obtained results have been reported below. Our proposed model diagnoses patients more accurately compared to the state-of-the-art method (HRL) with all three algorithms (DQN, Double DQN, and Dueling DQN). As there is only one disease in each group, a hierarchical disease classifier would not be appropriate; hence, we experimented with our proposed model with the PCM only (KI-CD_PCM). The results have been reported in Table 7. The proposed method outperforms baselines in all three algorithms, which firmly establishes the efficacy of the proposed model.

6.3. Analysis

In order to analyze the performance of the proposed framework in different disease groups, we have also evaluated the different worker's performance in terms of disease diagnosis accuracy. The obtained performances are reported in Table 8. Although there are many factors such as no. of unique symptoms (Fig. 4(b)), avg length of implicit symptom, no. of disease with zero or very less unique symptom across other diseases of groups (Fig. 5) that influences simultaneously to the workers' performance, we observe some relations and probable reasons of such performances. The workers (G4, G14) having significantly less avg no. of implicit symptoms (< 2) are performing significantly better than others. The workers (G7, G5) with many diseases having zero or significantly less no. of unique symptoms (Fig. 5) perform very poorly. We have selected the final values of different parameters (Table 3) empirically. The performance of the proposed model with different parameter settings has been reported in Table 9.

Error Analysis and Weakness of the proposed system We also plotted confusion matrix of the HDC disease classifier across groups and diseases. Fig. 7 shows group classification (Top level) performance across groups while Fig. 8 illustrates disease classification performance. An extensive error analysis highlights the following weaknesses of the proposed system.

- There is a huge number of cases where the disease classifier has predicted wrong diseases, but the predicted diseases belong to the same group of actual diseases (Fig. 7 is denser than Fig. 8). The classifier fails to distinguish between two very close diseases because of many common symptoms.
- In some cases, the learner critic misleads the agent if the disease classifier provides inappropriate /wrong candidate disease with high confidence, leading to the investigation of potential symptoms of the wrongly predicted candidate disease.
- As the KI-CD agent utilizes a hierarchical disease classifier (HDC), the negative symptom information may not contribute as effectively as a diagnostic system with a unified disease classifier because disease space is limited by group/worker in HDC.
- While analyzing the failed dialogues, we observe that the role of associations between symptoms is also a key factor in the disease diagnosis task. Sometimes, the proposed system lacks to capture these associations and diagnoses a patient with a wrong disease but very similar to the true disease despite initial correct candidate prediction. An example of such a scenario is illustrated in Table 11.

7. Case study

In order to evaluate and compare our proposed model qualitatively, we have reported a common dialogue session (patient data - Table 10.1) generated by the state of the art method, HRL

Table 8
Performance of baseline's disease classifier and Hierarchical disease classifier at different hierarchy levels.

Classifier	Avg. no of implicit symptoms	Accuracy(%)
Disease Classification(Liao at al., 2020)	2.6	49.80
First Layer: Group Classifier	NA	79.02
Worker1 (G1)	3.23	72.42
Worker2 (G4)	1.71	94.10
Worker3 (G5)	2.67	64.77
Worker4 (G6)	2.83	82.59
Worker5 (G7)	2.78	58.47
Worker6 (G12)	2.04	66.09
Worker7 (G13)	2.48	82.87
Worker8 (G14)	1.58	86.61
Worker9 (G19)	2.91	72.60
Second Layer: Disease Classifier (avg(L_{acc}))		75.61
KI-CD Disease Classifier($Group_{acc}$ · avg(L_{acc}))		59.74

Table 9
Performance of the proposed KI-CD_PCM model with different parameter setting.

Parameter	Success rate	Dialogue length	AMR(%)	SIR(%)	Disease classifier accuracy (%)
Maximum turn limit-26	0.5514	14.23	10.52	43.51	55.22
Maximum turn limit-30	0.5554	13.68	11.22	43.11	56.60
γ_m, γ^w-0.95, 0.95	0.5304	15.69	9.96	41.87	53.80
reward for success (SR) - +4	0.5478	14.07	10.01	41.81	55.20
reward for success (SR) - +2	0.5671	14.35	10.96	47.35	57.78
penalty for each turn - −0.1	0.5432	16.60	9.07	41.92	54.56
potential disease prediction probability (C_{th}) - 0.4	0.5478	16.32	7.96	38.04	54.90
potential disease prediction probability (C_{th}) - 0.6	0.5488	14.52	9.10	38.54	54.96
learning rate (α) - 0.0003	0.5266	16.53	9.11	39.12	52.46
learning rate (α) - 0.0007	0.5586	16.28	9.21	47.18	56.54
epsilon (ϵ) - 0.03	0.5542	15.39	9.78	46.93	56.06
epsilon (ϵ) - 0.15	0.5516	17.67	8.66	43.35	55 .84

Table 10
The sampled diagnosis tasks from the SD dataset, which contains 1 explicit symptom, 3 & 1 implicit symptoms and Endometrial hyperplasia & Anxiety as true disease.

Sample	Disease	Self report	Implicit symptoms
1	Endometrial hyperplasia	Vaginal bleeding after menopause: True	Heavy menstrual flow: True, Unpredictable menstruation: True, Pelvic pain: True
2	Anxiety	Anxiety and nervousness: True	Depressive or psychotic symptoms: True

Table 11
A sample (Patient data — Table 10.2) where the patient informs his/her difficulty (self report — Anxiety and nervousness) followed by symptom investigation and wrong disease prediction.

Candidate disease	Agent	Patient
Anxiety	Do you have headache ?	No, I do not have this symptom
Anxiety	Do you have Depressive or psychotic symptoms ?	Yes, I feel Depressive or psychotic
Dissociative disorder	you may have Dissociative disorder (probability 0.89)	Over, close conversation

Table 12
HRL (Liao at al., 2020) agent's performance on the selected samples (Table 10.1): It informs **incorrect disease (Endometrial cancer)** after symptom investigation as below.

Turn	Agent	Patient
Self-report	/	Hi, I had vaginal bleeding after menopause. I am also feeling anxiety and nervousness. Can you please tell, what could be the disease ?
1	Ok, Do you have also sharp abdominal pain ?	No, I am not feeling abdominal pain.
2	Do you suffer from painful menstruation ?	No, I do not feel painful menstruation.
3	Have you encountered unpredictable menstruation ?	Yes, I have unpredictable menstruation.
4	Are you feeling involuntary urination ?	No, I have not encounter such problem.
5	Have you seen pus in sputum ?	No, I have not seen pus in sputum.
6	Have you felt hot flashes ?	No, I have not felt so.
7	Do you had Long menstrual period ?	No, I had not encountered long menstrual period.
8	Ok, Do you have pelvic pain ?	Yes, I have pelvic pain.
9	Ok, you may have Endometrial cancer as per the symptoms.	Thank you.

Table 13
KI-CD_PCM's performance on the selected sample (Table 10.1): It informs **correct disease (Endometrial hyperplasia)** after symptom investigation as follows.

Turn	Candidate disease	Potential Symptoms	Agent	Patient
Self-report	/	/	/	Hi, I had vaginal bleeding after menopause. I am also feeling anxiety and nervousness. Can you please tell, what could be the disease ?
1	Endometrial cancer	Vaginal bleeding after menopause, Intermenstrual bleeding, Pelvic pain, Hot flashes, Long menstrual periods	Sure, are you suffering from painful urination ?	No, I do not feel so.
2	Endometrial cancer	-Do-	Have you felt heavy menstrual flow ?	yeah, true, I felt heavy menstrual flow.
3	Endometrial cancer	-Do-	Ok, have you encountered intermenstrual bleeding ?	No, I do not have this problem.
4	Endometrial cancer	-Do-	Have you experienced hot flash ?	No, I have not felt hot flash.
5	Endometrial hyperplasia,	Vaginal bleeding after menopause, Heavy menstrual flow, Unpredictable menstruation, Involuntary urination, Pelvic pain	Fine, Have you noticed vaginal bleeding after menopause ?	No, I have not experienced anything like this.
6	Endometrial hyperplasia	-Do-	Ok, Do you have pelvic pain ?	Yes, I have pelvic pain.
7	Endometrial hyperplasia	-Do-	Have you experienced frequent vaginal discharge ?	No, I do not have this problem.
8	Endometrial hyperplasia	-Do-	Fine, have you encountered unpredictable menstruation ?	True, I had unpredictable menstruation.
9	Endometrial hyperplasia	/	Ok, You are most likely to have Endometrial hyperplasia disease.	Thank you.

and our proposed agent in Table 12 and 13, respectively. Disease diagnosis is a critical step in the disease treatment process. An incorrect diagnosis may affect someone's health. Thus, diagnosis accuracy is the topmost and essential requirement for any automatic diagnosis system. In the diagnosis session (Table 12), the HRL agent requests some symptoms and informs a susceptible disease (Endometrial cancer) incorrectly. The performance of our proposed agent has been shown in Table 13.

The proposed agent, KI-CD_PCM investigates comparatively more relevant and grounded symptoms led through the dialogue context and infused knowledge (Candidate disease and Potential score) as per the context provided by the PCM module. Initially, the candidate disease was the same as the inaccurately informed disease by the HRL model due to a lack of context/investigated symptom, but as the agent gathers more information driven by the dialogue context and PCM module (potential disease and potential symptom), it gets updated. The agent investigates more symptoms, particularly aligned with the potential symptom set, and diagnoses the patient with correct disease (Endometrial hyperplasia) with high confidence. We also found that, even if the candidate disease is not the actual disease, the indicated prospective symptoms are very related to patients' implicit symptoms, resulting in better symptom investigation and patient satisfaction with the system.

8. Conclusion and future work

The work proposes a knowledge-infused context-driven hierarchical reinforcement learning (HRL) framework for automatic disease diagnosis. It incorporates a PCM as a knowledge companion, which assists the proposed agent in intelligent and effective symptom investigation. We have also developed a hierarchical disease classifier that alleviates symptom state sparsity issue and enhances classification capability. Although the inclusion of a hierarchical disease classifier (HDC) with the vanilla HRL framework outperforms the baselines and the state-of-the-art model, combining it with the potential candidate module (PCM) degrades the performance, owing to the restriction of candidate disease and potential symptom recommendations to only current disease group. The experimental results (both quantitative and qualitative) demonstrate that the proposed agent, KI-CD_PCM outperforms state of the art and baselines by a significant margin in all metrics other than dialogue turn (diagnosis success rate, AMR, SIR, and disease classification accuracy by 7.1%, 0.23%, 19.67% and 8.04%, respectively) on the publicly available dataset. Also, the rigorous analysis and case study establish the efficacy of the proposed framework qualitatively.

In future, we would like to extend the work in these three directions. We would like to incorporate a rich dialogue state representation capable of capturing associations amongst symptoms and their candidate disease, leading to more effective and intelligent behavior. We would also like to develop a more robust and effective disease classifier through advanced feature engineering techniques. Often, patients face difficulty in informing medical symptoms by text only; multi-modal input may enhance both patients' experience and the agent's effectiveness.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research. Abhisek Tiwari graciously acknowledges the Prime Minister Research Fellowship (PMRF) supported by the Government of India for conducting this research. We would also like to express our gratitude to Dr. Minakshi Dhar, AIMS Rishikesh, for her insightful comments and suggestions.

Appendix

See Fig. 9, Fig. 10, Fig. 11, Fig. 12

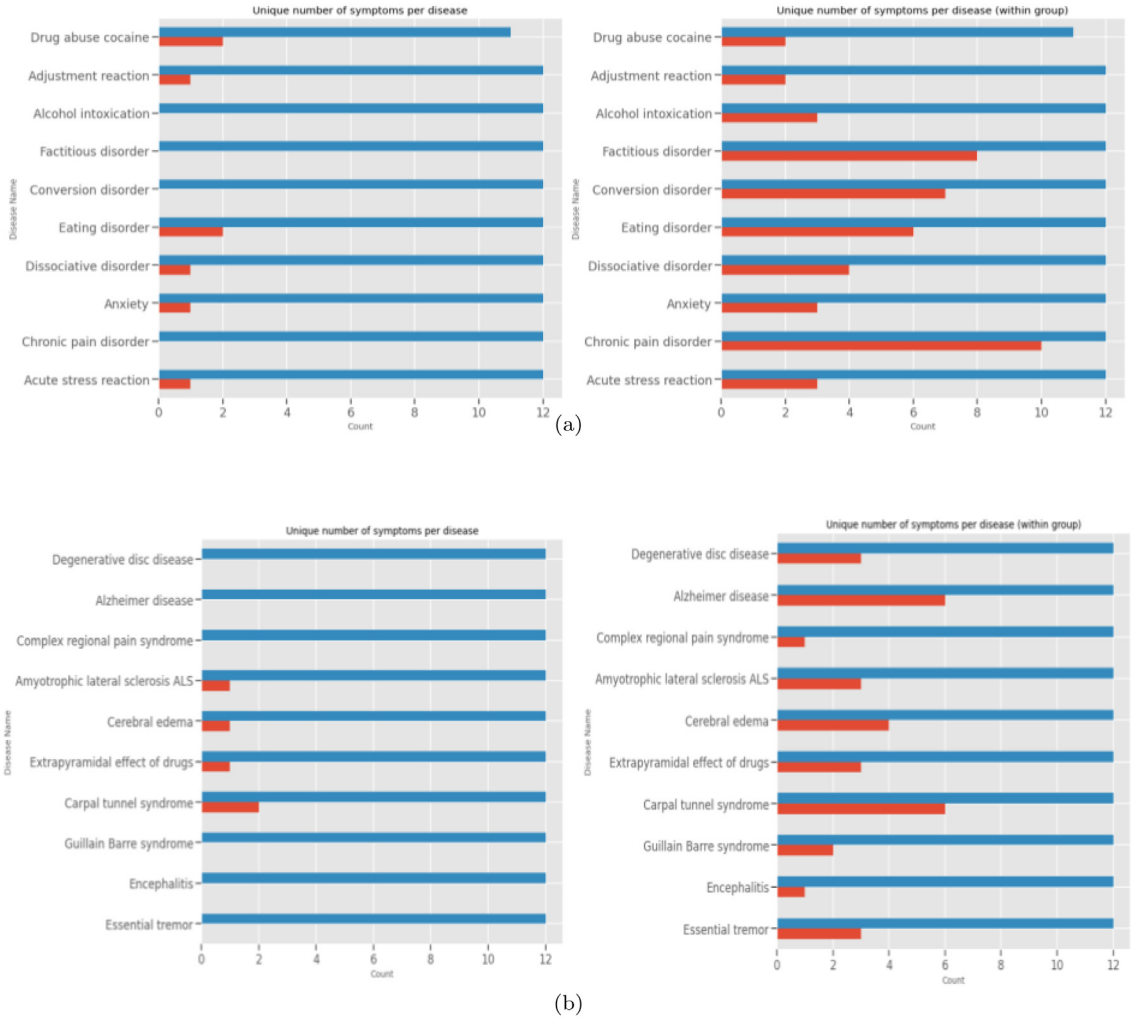
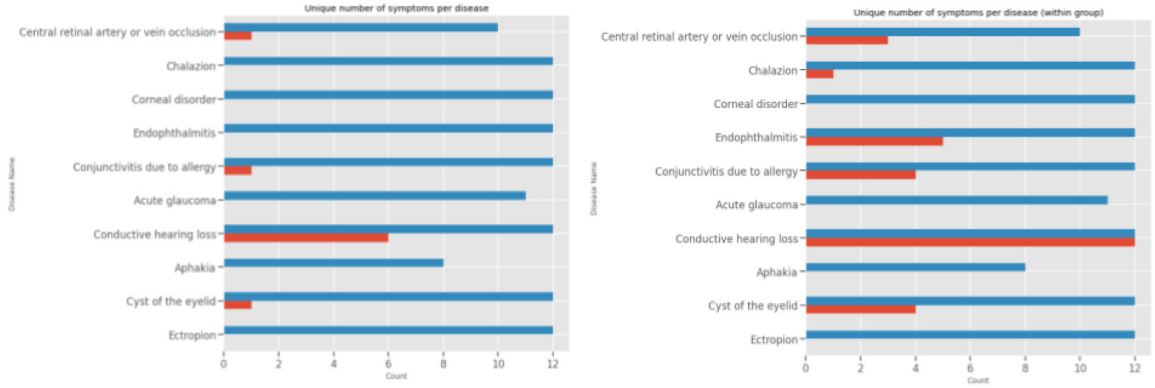
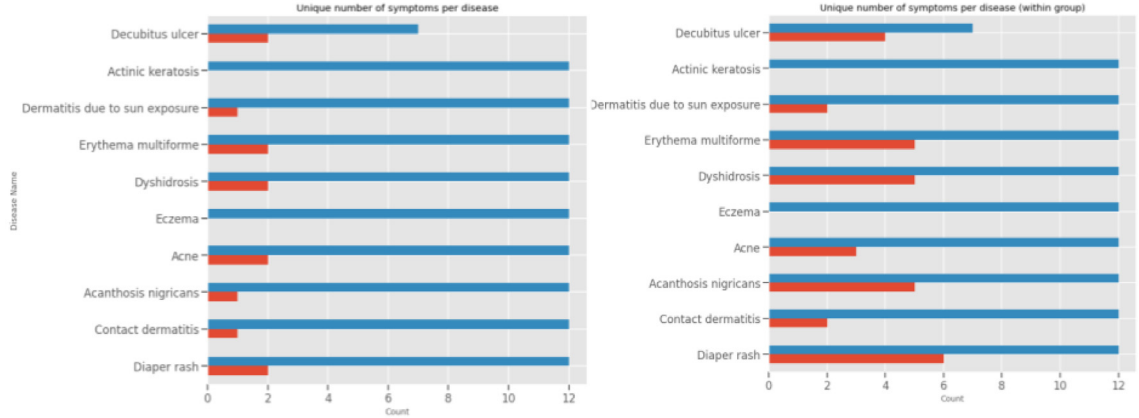


Fig. 9. The left plots of each figure show unique symptom count distribution across all diseases, whereas the right plots illustrate unique symptom count across diseases of their respective group. a. Disease group 5, b. Disease group 6.

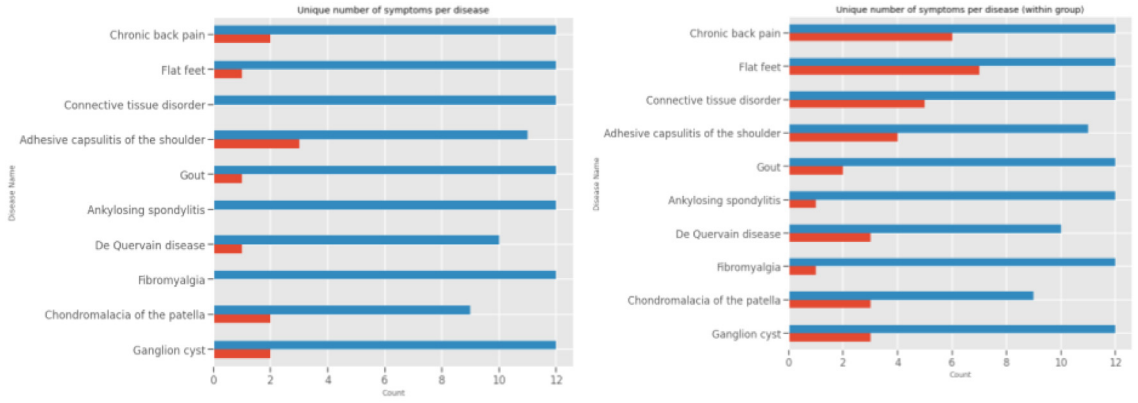


(a)

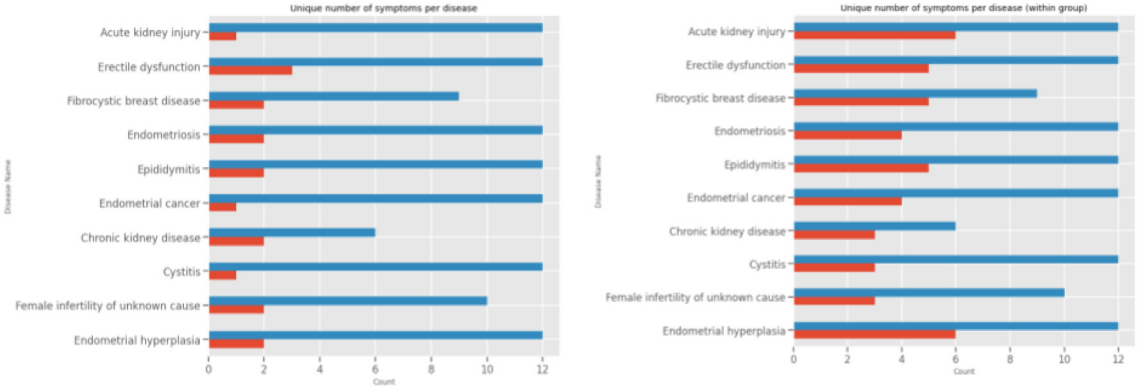


(b)

Fig. 10. The left plots of each figure show unique symptom count distribution across all diseases, whereas the right plots illustrate unique symptom count across diseases of their respective group. a. Disease group 7, b. Disease group 12.

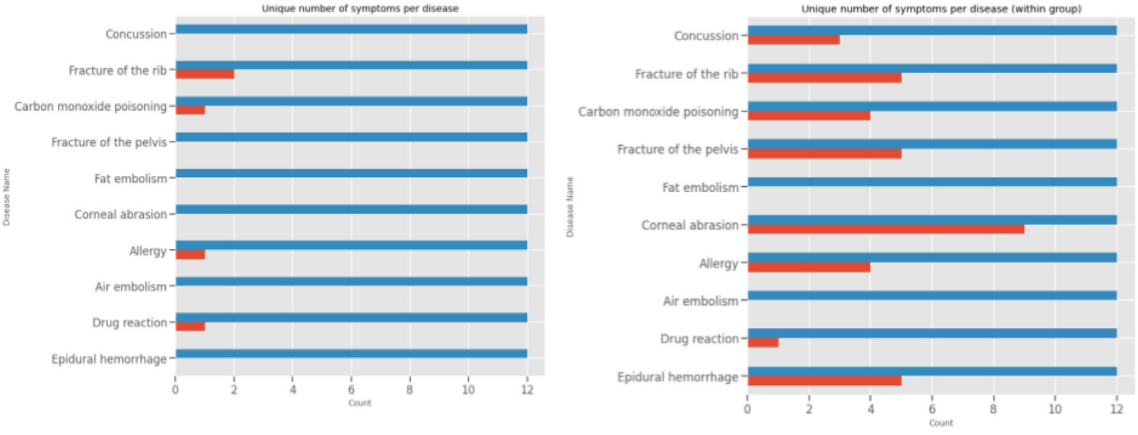


(a)



(b)

Fig. 11. The left plots of each figure show unique symptom count distribution across all diseases, whereas the right plots illustrate unique symptom count across diseases of their respective group. a. Disease group 13, b. Disease group 14.



(a)

Fig. 12. The left plots of each figure show unique symptom count distribution across all diseases, whereas the right plots illustrate unique symptom count across diseases of their respective group. a. Disease group 19.

References

[1] M.R. Cowie, J.I. Blomster, L.H. Curtis, S. Duclaux, I. Ford, F. Fritz, S. Goldman, S. Janmohamed, J. Kreuzer, M. Leenay, et al., Electronic health records to facilitate clinical research, Clin. Res. Cardiol. 106 (2017) 1–9.

[2] R.J. Byrd, S.R. Steinhubl, J. Sun, S. Ebadollahi, W.F. Stewart, Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records, Int. J. Med. Inform. 83 (2014) 983–992.

[3] S.R. Jonnalagadda, A.K. Adupa, R.P. Garg, J. Corona-Cox, S.J. Shah, Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of hfpef patients for clinical trials, J. Cardiovasc. Transl. Res. 10 (2017) 313–321.

[4] N. Ramakrishnan, B.K.T. Vijayaraghavan, R. Venkataraman, Breaking barriers to reach farther: A call for urgent action on tele-icu services, Indian J. Crit. Care Med.: Peer-Reviewed, Official Publ. Indian Soc. Crit. Care Med. 24 (2020) 393.

[5] Z. Wei, Q. Liu, B. Peng, H. Tou, T. Chen, X.-J. Huang, K.-F. Wong, X. Dai, Task-oriented dialogue system for automatic diagnosis, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 2 : Short Papers, 2018, pp. 201–207.

[6] X. Li, Y.-N. Chen, L. Li, J. Gao, A. Celikyilmaz, End-to-end task-completion neural dialogue systems, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing, 1: Long Papers, 2017, pp. 733–743.

[7] B. Liu, G. Tur, D. Hakkani-Tur, P. Shah, L. Heck, Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 2018, pp. 2060–2069.

[8] T.G. Dietterich, Hierarchical reinforcement learning with the maxq value function decomposition, J. Artificial Intelligence Res. 13 (2000) 227–303.

[9] K. Liao, Q. Liu, Z. Wei, B. Peng, Q. Chen, W. Sun, X. Huang, Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning, 2020, arXiv preprint arXiv:2004.14254.

[10] A.G. Barto, R.S. Sutton, Reinforcement learning, Handb. Brain Theory Neural Netw. (1995) 804–809.

[11] E. Levin, R. Pieraccini, W. Eckert, Using markov decision process for learning dialogue strategies, in: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), Vol. 1, 1998, pp. 201–204.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533.

[13] A.G. Barto, S. Mahadevan, Recent advances in hierarchical reinforcement learning, Discrete Event Dyn. Syst. 13 (2003) 41–77.

[14] R. Parr, S. Russell, Reinforcement learning with hierarchies of machines, in: Advances in Neural Information Processing Systems, 1998, pp. 1043–1049.

[15] P. Budzianowski, S. Ultes, P.-H. Su, T.-H. Mrkšić, I. Casanueva, L.M.R. Barahona, M. Gasic, Sub-domain modelling for dialogue management with hierarchical reinforcement learning, in: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, 2017, pp. 86–92.

[16] B. Peng, X. Li, L. Li, J. Gao, A. Celikyilmaz, S. Lee, K.-F. Wong, Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2231–2240.

[17] J. Liu, F. Pan, L. Luo, Gochat: Goal-oriented chatbots with hierarchical reinforcement learning, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1793–1796.

[18] Xue-Wen Chen, Xiaotong Lin, Big data deep learning: challenges and perspectives, IEEE access 2 (2014) 514–525.

[19] Amit Sheth, Manas Gaur, Ugur Kursuncu, Ruwan Wickramarachchi, Shades of knowledge-infused learning for enhancing deep learning, IEEE Internet Computing 23 (6) (2019) 54–63.

[20] Manas Gaur, Ugur Kursuncu, Amit Sheth, Ruwan Wickramarachchi, Shweta Yadav, Knowledge-infused deep learning, in: Proceedings of the 31st ACM Conference on Hypertext and Social Media, 2020, pp. 309–310.

[21] K.-F. Tang, H.-C. Kao, C.-N. Chou, E.Y. Chang, Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning, in: NIPS Workshop on Deep Reinforcement Learning, 2016.

[22] H.-C. Kao, K.-F. Tang, E. Chang, Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[23] L. Xu, Q. Zhou, K. Gong, X. Liang, J. Tang, L. Lin, End-to-end knowledge-routed relational dialogue system for automatic diagnosis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7346–7353.

[24] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.

[25] L. Baird, Residual algorithms: Reinforcement learning with function approximation, in: Machine Learning Proceedings, Vol. 1995, 1995, pp. 30–37.

[26] H. Cuayahuitl, S. Yu, A. Williamson, J. Carse, et al., Deep reinforcement learning for multi-domain dialogue systems, CoRR (2016).

[27] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double q-learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, 2016.

[28] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas, Dueling network architectures for deep reinforcement learning, in: International Conference on Machine Learning, 2016, pp. 1995–2003.

[29] R.S. Sutton, D. Precup, S. Singh, Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning, Artificial Intelligence 112 (1999) 181–211.

[30] G. Tesauro, Temporal difference learning and td-gammon, Commun. ACM 38 (1995) 58–68.

[31] V. Franc, V. Hlavác, Multi-class support vector machine, in: Object Recognition Supported By User Interaction for Service Robots, Vol. 2, 2002, pp. 236–239.