Dr. Can See: Towards a Multi-modal Disease Diagnosis Virtual Assistant

Abhisek Tiwari
Indian Institute of Technology, Patna
Patna, India
abhisek_1921cs16@iitp.ac.in

Manisimha Manthena
Manipal Institute of Technology
Manipal, India
manthena.varma@learner.manipal.edu

Sriparna Saha
Indian Institute of Technology, Patna
Patna, India
sriparna@iitp.ac.in

Pushpak Bhattacharyya
Indian Institute of Technology, Bombay
Bombay, India
pb@cse.iitb.ac.in

Minakshi Dhar
All India Institute of Medical Sciences, Rishikesh
Rishikesh, India
minakshi.med@aiimsrishikesh.edu.in

Sarbajeet Tiwari
Midnapore Homoeopathic Medical College and Hospital
Midnapore, India
sarbajeettiwari@gmail.com

ABSTRACT

Artificial Intelligence-based clinical decision support is gaining ever-growing popularity and demand in both the research and industry communities. One such manifestation is automatic disease diagnosis, which aims to assist clinicians in conducting symptom investigations and disease diagnoses. When we consult with doctors, we often report and describe our health conditions with visual aids. Moreover, many people are unacquainted with several symptoms and medical terms, such as mouth ulcer and skin growth. Therefore, visual form of symptom reporting is a necessity. Motivated by the efficacy of visual form of symptom reporting, we propose and build a novel end-to-end Multi-modal Disease Diagnosis Virtual Assistant (MDD-VA) using reinforcement learning technique. In conversation, users' responses are heavily influenced by the ongoing dialogue context, and multi-modal responses appear to be of no difference. We also propose and incorporate a Context-aware Symptom Image Identification module that leverages discourse context in addition to the symptom image for identifying symptoms effectively. Furthermore, we first curate a multi-modal conversational medical dialogue corpus in English that is annotated with intent, symptoms, and visual information. The proposed MDD-VA outperforms multiple uni-modal baselines in both automatic and human evaluation, which firmly establishes the critical role of symptom information provided by visuals [1].

CCS CONCEPTS

• **Computing methodologies** → *Sequential decision making*; **Discourse, dialogue and pragmatics**; • **Applied computing** → *Health care information systems.*

KEYWORDS

Symptom Investigation, Disease Diagnosis, Virtual Assistant, Multimodality, Deep Reinforcement Learning

ACM Reference Format:
Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari. 2022. Dr. Can See: Towards a Multimodal Disease Diagnosis Virtual Assistant. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), Oct. 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557296

1 INTRODUCTION

Disease diagnosis is the initial and most important stage of any disease treatment process. Doctors investigate patients' health conditions and determine diseases by assessing their self-report and other symptoms in the diagnosis phase. Based on the diagnosis and symptom assessment, they decide a treatment procedure. Therefore, the effectiveness of the treatment greatly depends on the accuracy of the diagnosis. As reported by the World Health Organization (WHO), 2013 [15], the world is short of 7.2 million medical workers, which is expected to reach 12.9 million in the upcoming decade. The doctor-per-person ratio continues to be less than one in many countries, as illustrated in a recent report by the WHO, 2019 [14]. These alarming figures firmly suggest that the healthcare system needs to be improved by increasing the number of health workers and utilizing their time more efficiently and critically. With the motivation of efficiently utilizing doctors' time and providing an accessible platform for early diagnosis, automatic disease diagnosis [16] is introduced, which is gaining in demand in both research and industry communities.

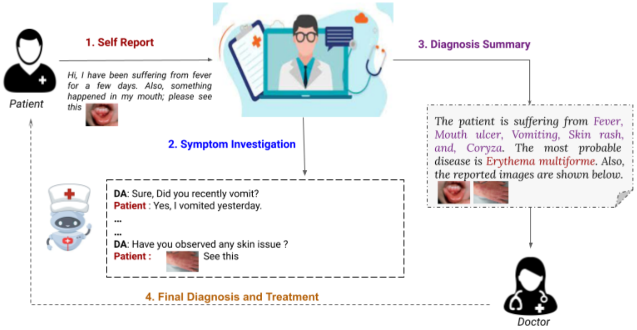
In real life, we often describe our primary complaints and difficulties to doctors with the help of visual aids. The primary motivations for visual communication are - (a) A large population is unfamiliar with medical terms for numerous symptoms, (b) Some symptoms are difficult to describe through text. Furthermore, patients are sometimes confused between two to three closely related symptoms, such as skin dryness and skin rash. Thus, visual reporting seems the most obvious and appropriate solution in such scenarios. However, the existing automatic disease diagnosis systems [8, 10, 16] extract symptoms and signs from users' text messages and fail to utilize patients' complaints, signs, and symptoms described through visual mode. While diagnosis accuracy is certainly the most important aspect of disease diagnosis, making the diagnosis with ease is also critical to end-users stratification with the

[1] The dataset and code are available at https://github.com/NLP-RL/DrCanSee

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557296

diagnosis assistants. Motivated by the inability and the critical limitations, we aim to investigate the role of multi-modal information in disease diagnosis and propose a Multi-modal Disease Diagnosis Virtual Assistant (MDD-VA) that extracts symptoms through both textual and visual modalities through conversation setting and diagnoses a disease accordingly. An illustration has been shown in Figure 1.



Hi, I have been suffering from fever for a few days. Also, something happened in my mouth; please see this

DA: Sure, Did you recently vomit?
Patient : Yes, I vomited yesterday.
...
...
DA: Have you observed any skin issue ?
Patient : See this

The patient is suffering from Fever, Mouth ulcer, Vomiting, Skin rash, and, Coryza. The most probable disease is Erythema multiforme. Also, the reported images are shown below.

Figure 1: An illustration of multi-modal automatic disease diagnosis

It is well said that a man is known by the company he keeps [21]. It has also been observed to be true for words, which later became key for developing different word embedding techniques such as Word2Vec [11]. In dialogue, user's/agent's current response (u_t) largely depends on ongoing dialogue context and the context-dependence is evident across all modalities, including text, visual, and audio. A symptom image shown by a patient at t^{th} dialogue turn is more likely to be relevant to the ongoing dialogue context. Thus, dialogue context can be effectively used to identify symptoms images appearing in the conversation. It can be expressed as follows:

$$I_t = \text{ContextSII}(image_t, C_t) \tag{1}$$

where $ContextSII$, $image_t$, and C_t are context-aware symptom image identifier, symptom image, and dialogue context ($u_{1:t-1}, a_{t-1}$) at tth time step, respectively.

In this paper, we aim to investigate the following three questions: ***i.*** *What is the role of visual form of symptom reporting in disease identification? Can a diagnosis assistant diagnose patients more effectively and satisfactorily if it considers patients' visual descriptions in addition to text-based symptoms?* ***ii.*** *Will patients' self-reports via text and images be enough to diagnose them correctly? Does visual reporting of symptoms play a role in patient-user conversations?* ***iii.*** *Can dialogue context help in interpreting an image that surfaced during conversation? Is there a relationship between dialogue context width and the performance of context-aware image identifier?*

To the best of our knowledge, the current work is the first attempt that proposes and builds a multi-modal disease diagnosis system. Also, the dialogue context-aware visual identification is one of the key novelties of our work, which has not been investigated previously in the literature. The key contributions of the work are three-fold, which are as follows:

- We propose and build a Multi-modal Disease Diagnosis Virtual Assistant (MDD-VA) using hierarchical reinforcement learning, which conducts symptom investigations and diagnoses a disease by considering both textual and visual information communicated through conversations.
- Motivated by the role of context in conversation, we propose and integrate a novel Context-aware Symptom Image Identification (ContextSII) module into the MDD-VA framework. Additionally, we have also experimented with a hierarchical method for symptom image identification, where the higher-level model identifies the organ/symptom group, whereas the lower-level model identifies symptoms.
- We curated a conversational medical dialogue corpus in English, where each utterance is annotated with its corresponding intent, symptom, and symptom image information.

2 RELATED WORK

Our work is mainly related to the following two research areas: Automatic disease diagnosis dialogue systems and Symptom/Disease identification using visual aids. We have summarized the relevant works and their limitations in the subsequent paragraphs.

Automatic disease diagnosis dialogue system Tang et al. [16] proposed an ensemble neural network model for symptom checking. The diagnosis model consists of dedicated models for different anatomical parts, which outperformed traditional monolithic systems by a significant margin. In real life, doctors' investigation also depends on patients' personal information, such as age and gender, in addition to patients' reported major difficulties. Motivated by such scenarios, the authors have proposed a context-aware HRL-based dialogue system [8] for symptom investigation followed by disease prediction. The model outperforms the existing non-contextual model [16] and evidences the efficacy of context modeling in the diagnosis process. Wei et al. [23] formulated symptom investigation as a task-oriented dialogue system where the agent extracts symptoms through conversation and diagnoses a disease as per observed symptoms. The obtained results illustrated the significant role of implicit symptoms extracted by the dialogue agent in addition to patient-reported symptoms for accurate diagnosis. Liao et al. [10] proposed an integrated and synchronized two-level policy framework using hierarchical reinforcement learning [4]. The model outperformed the flat policy approach [23] by a significant margin, demonstrating the efficacy of disease group aware symptom investigation.

Symptom/Disease identification using visual aids In [6], the authors have investigated the role of physicians' verbal narratives while examining dermatology images for image-based diagnoses. The results show a significant role of verbal narratives in identifying patients' diseases accurately and efficiently. The work [13] proposed a Multi-modal Attentional Neural Network (MNN) for disease diagnosis from electronic health record (EHR) data. The findings firmly establish the importance of clinical notes, in addition to medical codes, in efficiently identifying disease. In [20], authors have proposed a multi-modal deep learning model for Alzheimer, which leverages features of two different modalities (imaging, genetic and clinical features) for skin lesion identification. The obtained experimental results show that the incorporation of multi-modal information provides complimentary visual features, thereby enhancing the model's discriminative capability.

3 METHODOLOGY

Our proposed automatic disease diagnosis system has two stages: Symptom Investigation (SI) and Disease Diagnosis (DD). We propose an end-to-end diagnosis model that considers end-users utterances in text form and responds to them in text. The detailed architecture of the proposed framework is illustrated in Figure 3. The proposed diagnosis assistant, MDD-VA, works as an assistant to real doctors, which conducts a detailed symptom investigation and extracts relevant symptoms through conversing with a patient depending on chief patient complaints and ongoing dialogue context. MDD-VA also identifies the disease based on the investigated symptoms and forwards a detailed diagnosis report to the doctor. The proposed diagnosis framework consists of three main modules: **i.** Natural language understanding (NLU), **ii.** Symptom Investigation and **iii.** Patient Simulator and Diagnosis Report Generation. The detailed working methodologies of each module have been explained below.

3.1 Natural Language Understanding (NLU)

Natural Language Understanding (NLU) is the first stage of any dialogue system, which takes the user's message (both text and image) as input and identifies user intent and information conveyed by the message. The architecture of the NLU module is illustrated in Figure 4. The NLU component consists of two modules, namely Intent & Symptom module and Symptom Image Identifier module.

Intent and Symptom Module The intent of an utterance signifies user's intention expressed, while the symptom sequence tag identifies symptom information conveyed through the corresponding utterance. The Intent & Symptom module takes user's utterance (text) as input, and the module predicts its intent (self-report, symptom_inform, and visual_inform) and symptom sequence tag. In our proposed framework, we have utilized the joint BERT [3] model to capture the inter-relationships between these two tasks (intent classification and symptom sequence tagging).

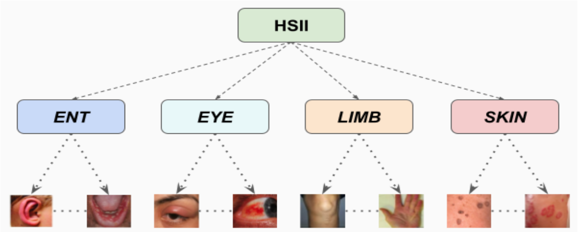


Figure 2: Hierarchical Symptom Image Identifier (HSII) where the higher level model selects a relevant symptom group and the lower level model identifies a symptom from the activated symptom group

Symptom Image Identifier We experimented with Convolution Neural Network (CNN) and some pre-trained image models, such as Inception v3 [24] and DenseNet169 [7]. We also built a hierarchical symptom image classifier (Figure 2) consisting of two levels : Symptom Group Classifier (SGC) and symptom identification. We grouped the multi-modal symptoms into four symptom groups, namely ENT, EYE, LIMB, and SKIN (Table 1). We have utilized the pre-trained DenseNet169 model with two additional convolutional layers to train both level classifiers.

Symptom group	Symptoms
ENT	Redness in ear, Lip swelling, Mouth ulcer, and Swollen or red tonsils
EYE	Proptosis, Swollen eye, Eye redness, Itchy eyelid, and Eyelid lesion or rash
LIMB	Edema, Foot or toe swelling, Knee swelling, Hand or finger lump or mass, and Neck swelling
SKIN	Cyanosis, Skin growth, Skin dryness, Skin rash, Skin irritation, and Dry or flaky scalp

Table 1: Different multi-modal symptom groups and their symptoms

Context aware Symptom Image Identification Motivated by the role of context in conversation, we propose a dialogue context-aware symptom image identification (ContextSII) model for symptom image identification. The architecture of the ContextSII model is shown in Figure 4. The ContextSII model takes the dialogue context embedding (confirmed symptoms) from the Clinical-BERT [1] model and image features from the DenseNet169 model as input. The concatenated textual and visual representation is passed to a feed-forward neural network, which predicts probability distribution over multi-modal symptom space.

$$I_t = \text{ContextSII}(image_t, C_t) \tag{2}$$

$$C_t =< U_1, A_1, U_2, A_2, ...U_{t-1}, A_{t-1} > \tag{3}$$

where $ContextSII$, $image_t$, and C_t are context-aware symptom image identifier, symptom image, and dialogue context at t^{th} time step, respectively. Here, U_t and A_t are user and agent message at t^{th} turn, respectively.

3.2 Symptom Investigation

Clinics used to have different departments, such as cardiology and pediatrics, to improve investigative efficacy and patient satisfaction. Motivated by the real-world scenario and the promising results obtained by Liao et al. [10], we also formulated a hierarchical structure-based policy for symptom investigation. The higher-level policy (controller) acts as health center receptionist, which activates one of the lower level (Department) policies depending on patients' self-report and other symptoms. The activated department policy conducts department-specific symptom investigations. Finally, the controller policy activates the disease classifier, which projects probable disease as per observed symptoms.

Controller Policy Controller policy is the first level policy, which activates an appropriate department policy (DP$_i$) and disease classifier for symptom inspection and disease projection, respectively. The controller policy selects an action (ac) depending upon current multi-modal dialogue state (S) as follows: $ac = \text{argmax}_i P(A_c^k|S, \pi_c)$ where π_c is the controller policy, A_c^k denotes k^{th} action from the action space. For each action (ac), the agent gets a reward/penalty called extrinsic reward (r$_c$: Reward(S, ac)) as follows:

$$rc_t = \begin{cases} \sum_{i=1}^{n} \gamma_c^i r_{t+i}^d, & \text{if } ac_t = DP_i \\ r_t^d, & \text{if } ac_t = DC \end{cases} \tag{4}$$

where i is the number of turns taken by the activated lower-level policy (DP$_i$). The controller policy π_c has been optimized using

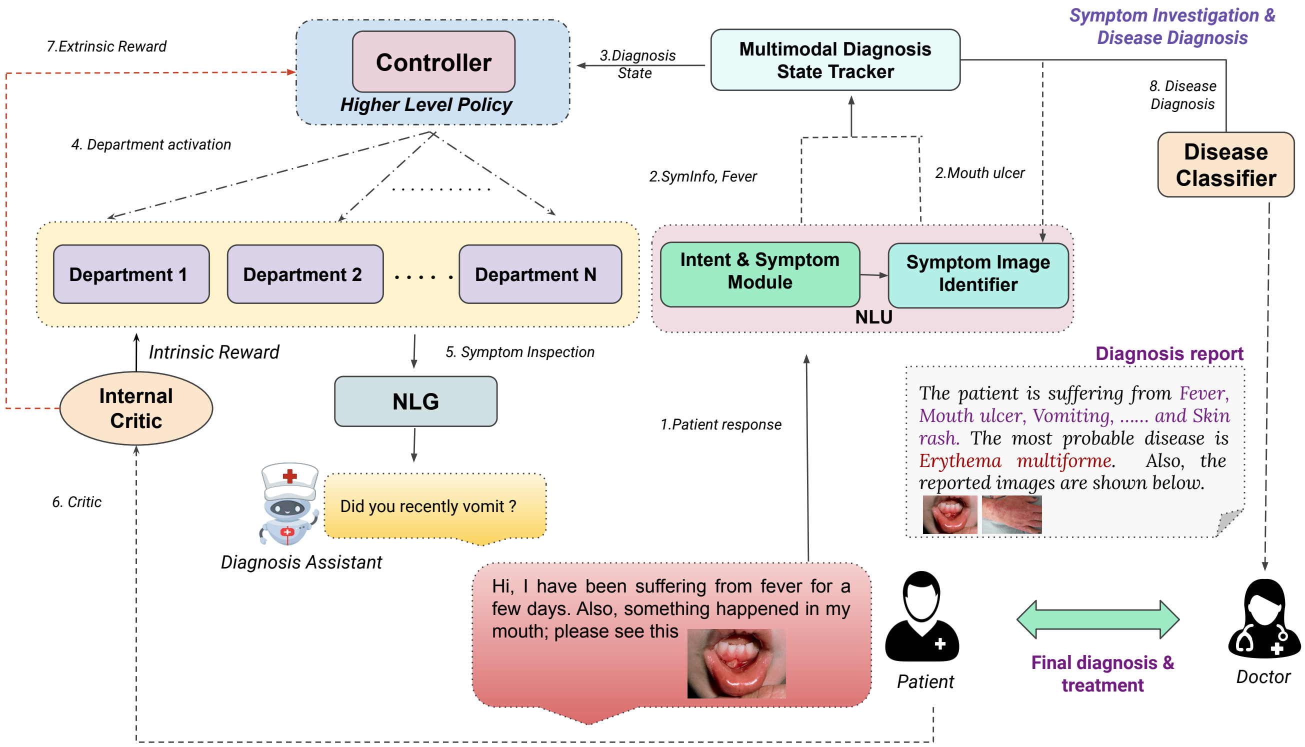


Figure 3: Proposed Multi-modal Disease Diagnosis Virtual Assistant (MDD-VA), which assists doctors by conducting a thorough symptom investigation and provides the symptom report along with the most probable disease as per the investigated symptoms

a value-based deep reinforcement learning technique called Deep Q Network (DQN) [12]. It learns a state-action value function (Q^c (S, ac)), which estimates value for each action (department) for a given dialogue state S (informed symptoms). The policy selects an action with highest Q value i.e., $ac = argmax_i Q^c(S, A_c^i|\pi_c)$. The Q^c function has been calculated and optimized through Bellman equation [2] using temporal difference (TD) loss [17] as follows:

$$Q^c(S_t, ac) = \mathbb{E}[rc_t + \gamma_c \cdot \max_{ac'} Q^c(S_{t+1}, ac')] \tag{5}$$

$$L_t^c = [(rc_t + \gamma_c \cdot \max_{ac' \in A_c} Q^c(S_{t+1}, ac'|\pi_c^{t-1}, \theta^{t-1})) - Q^c(S_t, ac|\pi_c^t, \theta_t)]^2 \tag{6}$$

where L_t^c is the loss at t^{th} time step, which is difference between state-action value calculated through current policy parameter (behavior network: θ_t) and previously freezed policy parameter (target network: θ_{t-1}). Here, ac' denotes next action taken by the agent for the observed next state, S_{t+1}.

Department Policy The departmental/lower level policies (DP$_i$: π^i) are responsible for symptom examination corresponding to their departments. These departmental policies learn to select an appropriate action (symptom) depending upon the current multi-modal dialogue state. It selects an action (a$_i$) as: $a_i = \text{argmax}_j Q^i(A_{ij}|S, \pi^i)$ where Q^i is state-action value function of i^{th} department policy (π^i) and A_{ij} is j^{th} action of i^{th} departmental policy. The multi-modal state, S, consists of patient self-report, users' provided image information, inspected /informed symptoms' status (both textual and visual), dialogue turn, agent's previous actions, and reward. The size of the action space of each policy is N$_i$ + 1, where N$_i$ is the number of symptoms in i^{th} department. The additional action is to return the control to the controller policy. For each action (a$_d^t$), the activated departmental policy gets a reward/penalty (intrinsic reward: r$_e$) from environment (user) based on its appropriateness to current diagnosis state as follows:

$$r_e = \begin{cases} +s, & \text{if } success \\ +A(SII(I_t), a_d^t), & \text{if } I_t \in U_t \\ +ts, & \text{if } U(a_d^t) = True \\ -p, & \text{if } repetition \\ 0 & Otherwise \end{cases} \tag{7}$$

where s, ts, and p are reward/penalty for successful diagnosis, inspecting a relevant symptom from which the patient is suffering, and repetitive behavior, respectively. The diagnosis assistant gets a positive (+ts) reward if it predicts the correct symptom. The agent also gets penalized if it inspects an already examined symptom. When the user's message contains a symptom image (I_t), the agent gets a reward proportional to the association between the diagnosis assistant's requested symptom (a_d^t) and predicted symptom tag ($A(SII(I_t), a_d^t)$). The lower association score between the agent's requested symptom and the user-displayed image indicates less relevance to the requested symptom, and thus the agent receives less reward. The association score between two symptoms (S$_i$, S$_j$) is computed by utilizing the frequency of their co-occurrence as

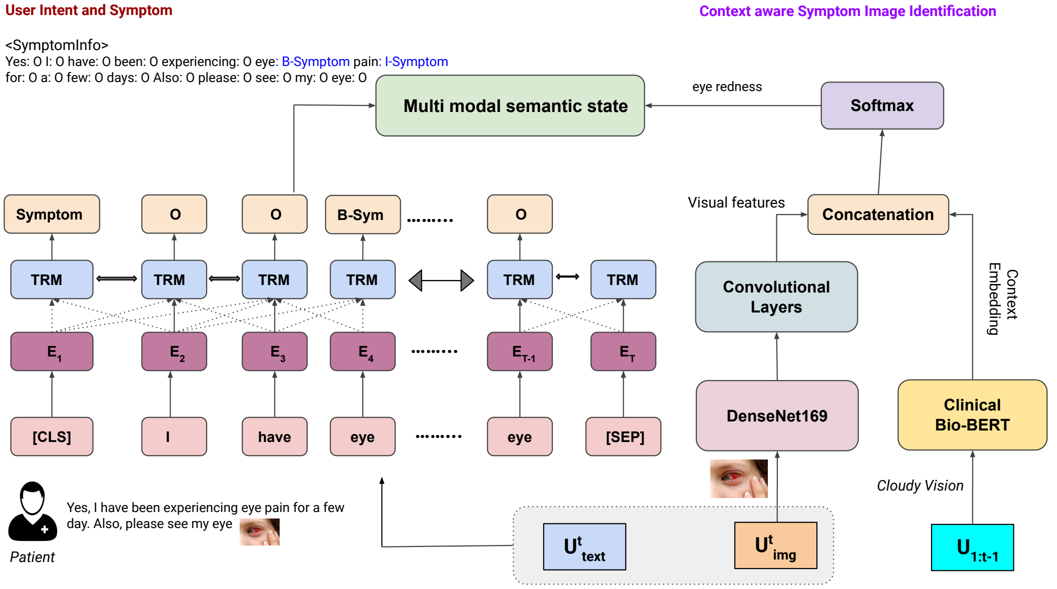

User Intent and Symptom
<SymptomInfo>
Yes: O I: O have: O been: O experiencing: O eye: B-Symptom pain: I-Symptom
for: O a: O few: O days: O Also: O please: O see: O my: O eye: O
Context aware Symptom Image Identification
Multi modal semantic state
eye redness
Softmax
Visual features
Concatenation
Symptom
O
O
B-Sym
O
TRM
TRM
TRM
TRM
TRM
TRM
Convolutional Layers
Context Embedding
E1
E2
E3
E4
ET-1
ET
[CLS]
I
have
eye
eye
[SEP]
DenseNet169
Clinical Bio-BERT
Cloudy Vision
Yes, I have been experiencing eye pain for a few day. Also, please see my eye
Patient
Ut text
Ut img
U1:t-1

Figure 4: Proposed architecture of natural language understanding module (NLU), which compromised of two main modules, namely Intent & Symptom module and context-aware symptom image identification (ContextSII)

follows:

$$A(S_i, S_j) = \frac{n(S_i, S_j)}{\sum_k n(S_i, S_k)} \tag{8}$$

where $n(S_i, S_j)$ is the number of cases where S_i and S_j have co-occurred. The term k ranges in the entire symptom space (Sy). These department policies (π_d^i) have also been optimized using dedicated DQN policy networks using Equations 5 and 6.

3.3 Patient Simulator and Diagnosis Report Generation

We have developed a pseudo environment/user simulator similar to the popular task-oriented user simulators [9, 10]. The user simulator initializes each diagnosis session with a diagnosis case from training samples. At the first turn of conversation, the patient simulator informs self-report (all explicit symptoms) to the diagnosis agent either through text or visual aid and asks to identify the disease/condition. Then, the simulator responds to the diagnosis assistant's request through text or visual aid as per the sampled diagnosis case during the conversation.

Disease classification is the final stage, which diagnoses a disease depending upon the extracted symptoms (including patient self-report). The disease classifier is a two-layered deep neural network, which takes a one-hot encoding representation of symptom status as input and predicts the probability distribution over all diseases. We utilize a template-based module for the agent's response generation, which takes the agent's requested symptoms as input and converts them into user-readable form. It creates a diagnosis report that includes all of the observed symptoms as well as the identified ailment and presents it to end-users.

4 DATASET

We first extensively investigated existing benchmark medical diagnosis dialogue corpora, and the summary is presented in Table 2. Despite the huge importance of visual form of symptom reporting, we neither find a single multi-modal dialogue corpus nor a diagnosis assistant that allows patients (end-users) to report their difficulties through visual mode. Motivated by the efficacy of an end-to-end multi-modal dialogue system, we make an attempt to develop a multi-modal conversational disease diagnosis corpus named Visual Medical Disease Diagnosis (Vis-MDD) by leveraging the SD dataset with the guidance of two clinicians.

Data Collection We found only one diagnosis data, SD [10] in English, where each sample contains patients' symptoms (self-report and implicit symptoms) and their corresponding disease. We first analyzed the dataset thoroughly with the help of two clinicians. We identified 31 symptoms from the SD dataset that are either hard to specify through text or are not commonly known. We selected only 17 symptoms out of these identified symptoms, such as *Mouth ulcer, Skin growth* for multi-modal conversation creation primarily due to the lack of sufficient images of other symptoms. We then collected the symptom images from open source platforms, such as Google, and filtered out inappropriate pictures with the help of the clinicians.

Dialogue Creation and Annotation We selected 100 random diagnosis cases from the SD dataset, each having patient self-reported symptoms, implicit symptoms requested by the doctor, and diagnosed disease. We curated a sample conversational data corresponding to the 100 diagnosis cases and annotated them with their corresponding intent, symptom, and image information with the help of clinicians. Then, we employed three biology graduates to create and annotate 1500 dialogues based on detailed guidelines and a curated sample dataset provided by the clinicians. In order to measure

Dataset	Language	Size	Conversation	Intent	Symptom	Multi-modality	E2E
RD [23]	Chinese	710 S, 67 Sym, 4 D	×	×	×	×	×
DX [25]	Chinese	510 S, 41 Sym, 5 D	✓	×	✓	✓	×
M^2 - MedDialogue [26]	Chinese	1557 S, 2468 E, 276 D	✓	×	✓	×	×
MedDialog-EN [27]	English	0.26 million S, 96 D	×	×	×	×	×
SD [10]	English	30,000 S, 266 Sym, 90 D	×	×	×	×	×
Vis-MDD (ours)	English	1500 S, 266 Sym, 90 D	✓	✓	✓	✓	✓

Table 2: Statistics of the existing medical dataset for disease diagnosis task. Here, S, Sym, D, and E refer to sample, symptom, disease, and entity, respectively

the annotation agreement among the annotators, we calculated the kappa coefficient (k), which was found to be 0.78, indicating a significant uniform annotation. The detailed statistics of the SD dataset, some symptom image samples, and a conversation from the curated Vis-MDD corpus have been reported in Table 3, Figure 5, and Figure 6, respectively.

Attribute	Value
# of Dialogues	1500
# of Utterances	6936
Utterance tags	Intent, Symptom, Image info
Average dialogue length	4.62
# of intents	3
# of diseases	90
# of symptoms	266
# of visual symptoms	17
# of symptom images	1805
# of multi-modal utterances	725

Table 3: Statistics of Vis-MDD Dataset

Role of Multi-modality During a conversation with a doctor, we strive to convey our symptoms accurately and efficiently. Therefore, we often use visuals when we are unsure of symptom names or it is difficult to convey some symptoms precisely through text. Such few symptoms are reported in Figure 5. For example, many people are unaware that figures (Figure 5, column 2) are instances of mouth ulcer. Thus, we curated a multi-modal disease diagnosis dialogue corpus that includes both textual and visual symptom reporting.

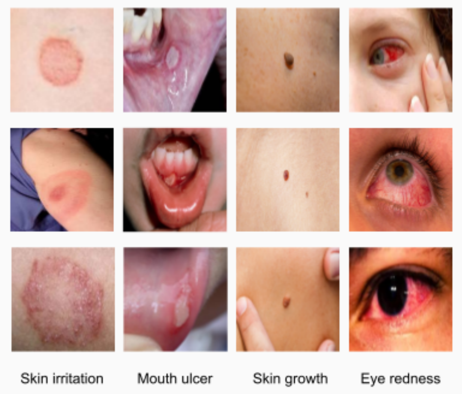
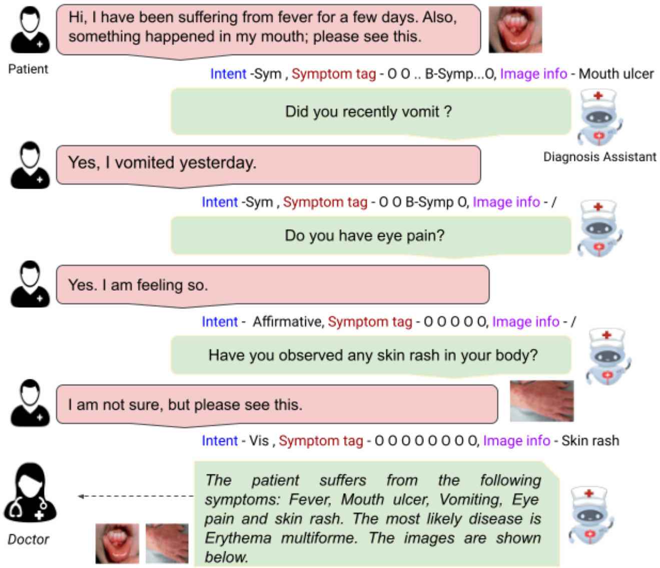


Figure 5: Few image samples for some multi-modal symptoms

Role of Dyadic Conversation Natural language understanding is the first stage of a dialogue system that recognizes users' intention and information from their messages. Both of the modules require conversational data for training. All the existing medical diagnosis datasets [10, 23] in the English language are in database form where each entry corresponds to patients' suffering symptoms (explicit and implicit) and corresponding diseases. To the best of our knowledge, the curated dataset, Vis-MDD, is the first medical diagnosis dialogue corpus in English. Thus, in order to make an end-to-end medical diagnosis dialogue system, we developed a dyadic dialogue corpus having doctor-patient conversations.



Hi, I have been suffering from fever for a few days. Also, something happened in my mouth; please see this.

Intent -Sym , Symptom tag - O O .. B-Symp...O, Image info - Mouth ulcer

Did you recently vomit ?

Yes, I vomited yesterday.

Intent -Sym , Symptom tag - O O B-Symp O, Image info - /

Do you have eye pain?

Yes. I am feeling so.

Intent - Affirmative, Symptom tag - O O O O O, Image info - /

Have you observed any skin rash in your body?

I am not sure, but please see this.

Intent - Vis , Symptom tag - O O O O O O O O, Image info - Skin rash

The patient suffers from the following symptoms: Fever, Mouth ulcer, Vomiting, Eye pain and skin rash. The most likely disease is Erythema multiforme. The images are shown below.

Figure 6: A conversation from the curated Vis-MDD dialogue corpus

Role of Intent and Symptom Annotation Intent and slot identifications are two primary tasks for NLU, which are essential for developing an end-to-end dialogue system capable of communicating with users in natural language form. Thus in order to create the NLU module, we tagged intent and slot (here symptom) information corresponding to each user utterance of a conversation (Figure 6).

Ethical Consideration We have strictly followed the guidelines established for legal, ethical, and regulatory standards in medical research during the Vis-MDD curation process. With this in mind, we have not added or removed any entity in a conversation corresponding to the reported dialogues in the SD dataset. Also, the curated dataset does not reveal users' identities. Moreover, the annotation guidelines are provided by two clinical authors, and the dataset is thoroughly checked and corrected by them. Therefore, we ensure that the Vis-MDD dataset and each stage of its creation procedure do not violate any ethical principles.

5 EXPERIMENTAL SETUP

We have utilized the PyTorch framework for implementing the proposed end-to-end disease diagnosis assistant. The joint intent and slot module have been trained and evaluated on the conversational dialogue data with 80% and 20% of total utterances, respectively. The image classifier models have been trained and tested with 90% and 10% of total symptom images, respectively. The proposed diagnosis model has been trained and evaluated on 80% (24,000) and 20% (6,000) patients' samples of the dataset, respectively. The final selected hyperparameter values are as follows: s (66), ts (+44), p (-44), N (maximum dialogue length limit: 22), Batch size (32), learning rate (0.001 for image classification, 0.0001 for dialogue policy learning), Optimizer (Adam), Gamma (0.95 for controller, γ_c and 0.9 for department policies, γ_d).

6 RESULTS AND DISCUSSION

We have utilized the most popular automatic diagnosis evaluation metrics (diagnosis success rate, dialogue length, match rate, match rate2, and disease classifier accuracy) [9, 10, 23] for evaluating the proposed agent's performance. We have experimented with three reinforcement learning algorithms, namely DQN [12], Double DQN [19], and Dueling DQN [22]. Match rate, a metric, is the ratio of the number of true symptoms (extracted through conversation) to the total number of agent's symptom requests. Match rate 2 signifies the ratio of the count of extracted symptoms (with true status) to the total number of symptoms in the patient's implicit symptom set.

To determine the efficacy of the proposed method, we have compared our model with the following baselines and state-of-the-art models. **i. SVM-ex**: SVM-ex is a support vector machine (SVM) model [5], which considers only patients' self-reports (explicit symptoms) for identifying their diseases. **ii. Unified Dialogue Policy (UDP) / Flat policy**: An unified policy [23] that conducts both symptom investigation and diagnosis. It either requests (inspects) a symptom from the entire action space or selects a disease as per observation state. **iii. HRL**: HRL [10] is one of the state-of-the-art models for the automatic disease diagnosis task, which has been trained and validated on the largest English benchmarked dataset. The proposed methodology employs an HRL framework that consists of two-level policies and a disease classifier, **iv. KI-CD**: The KI-CD model [18] utilizes a knowledge-infused context-driven diagnosis method for symptom investigation. The diagnosis assistant identifies potential diseases based on context and prioritizes the symptoms of these diseases for further examination. **v. MDD-VA_MSR**: MDD-VA_MSR is our proposed model, which considers visual symptoms reported in only patients' self-reports along with text-based symptoms extracted from conversations.

Task	Accuracy(%)	F1-Score
Intent classification	95.49	0.9388
Symptom labeling	92.04	0.9131

Table 4: Performance of the joint intent and symptom module

Model	Accuracy(%)	F1-Score
CNN	40.99	0.4247
Inception v3	66.14	0.6475
Inception v3 + Conv Layers	72.29	0.7163
DenseNet121	68.17	0.6712
DenseNet121 + Conv Layers	75.58	0.7412
DenseNet169	72.27	0.7157
DenseNet169 + Conv Layers	**78.51**	**0.7734**

Table 5: Performance of different pre-trained image classifiers

Model	Accuracy(%)	F1-Score
HSC: First layer (L1): SGC	93.24	0.9301
ENT (C1)	98.47	0.9729
EYE (C2)	80.31	0.7998
LIMB (C3)	92.22	0.9187
SKIN (C4)	71.66	0.7121
Second layer (L2): avg(C1, C2, C3, C4))	85.66	0.8508
HSC: Overall (L1 * L2)	**79.86**	**0.7913**

Table 6: Performance of two layered hierarchical symptom image identification module

Model	Accuracy	F1-Score
ContextSII (context window = 1)	83.43	0.8287
ContextSII (context window = 2)	84.66	0.8445
ContextSII (context window = 3)	**88.34**	**0.8824**
ContextSII (context window = 4)	85.89	0.8521
ContextSII (context window = 5)	82.28	0.8195

Table 7: Performance of proposed context-aware image identification module

The results obtained by the joint intent and symptom identifier are reported in Table 4. Table 5 shows the performances of different pre-trained imagenet models for identifying symptom images reported by patients. In Table 6, we have described the performance obtained by the hierarchical symptom image identifier (HSII). Table 7 illustrates the performance of the context-aware symptom image identifier, which outperformed both pre-trained image identifiers and hierarchical symptom classifier (HSC) models by a large margin of 9.84% and 8.45%, respectively. The huge improvement achieved by the context-aware image identification model firmly establishes the effectiveness of leveraging dialogue context in identifying visual symptoms that surface during conversations. Furthermore, we examined how context width influences the identifier's accuracy. The performances of the proposed disease diagnosis assistant (MDD-VA) and different baselines are reported in Table 8 and Table 9. All the reported result values are statistically significant at 5% significant level ($p < 0.05$).

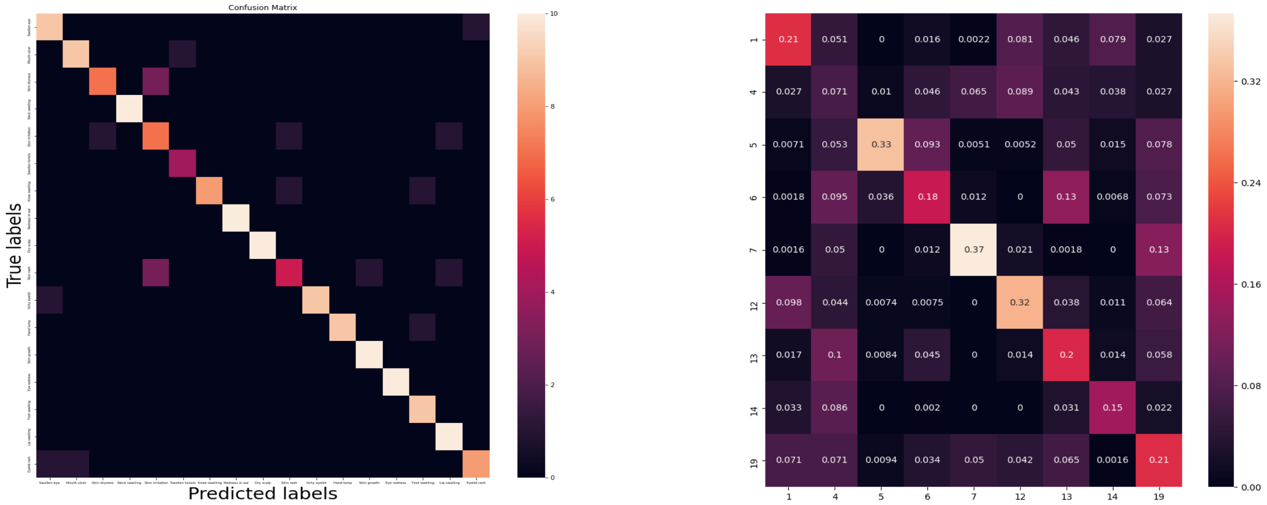
Findings With the obtained results and observations, the raised research questions can be answered as follows: **i.** *What is the role of visual form of symptom reporting in disease identification?* The improvement (both quantitative and qualitative - Tables 8, 9 and Figure 8) obtained by the proposed MDD-VA firmly establishes the crucial role of visual form of symptom reporting. Furthermore, the MDD-VA performed the best in human evaluation, demonstrating the efficacy and necessity of the multi-modal diagnosis assistant. **ii.** *Will patients' self-reports via text and images be enough to diagnose them correctly? Does visual reporting of symptoms play a role in patient-user conversations?* No, patient self-report (text and images) is not sufficient for identifying the associated disease accurately.

Model	Policy Optimization	Diagnosis success rate	Dialogue length	AMR	AMR2	Disease classifier accuracy (%)
SVM-ex		0.322	/	/	/	32.20
HRL [10]		0.393	7.89	9.32	17.70	39.98
MDD-VA-MSR	Dueling DQN	0.410	7.38	11.82	19.96	41.35
MDD-VA		**0.414**	**7.12**	**14.52**	**23.76**	**42.11**
HRL [10]		0.424	8.77	9.45	17.70	0.4202
MDD-VA_MSR	Double DQN	0.428	8.14	12.32	22.54	42.22
MDD-VA		**0.442**	**8.07**	**15.64**	**27.26**	**43.64**
HRL [10]		0.439	11.37	7.22	24.37	0.4237
MDD-VA_MSR	DQN	0.464	10.96	8.80	27.70	46.41
MDD-VA		**0.487**	**10.53**	**10.66**	**32.16**	**48.27**
KI - CD [18] w/o multi-modality		52.69	16.09	8.48	40.80	51.78
KI - CD with MSR	DQN	53.42	15.90	8.97	44.24	53.76
KI - CD with multi-modality		**54.04**	**15.76**	**10.48**	**47.14**	**54.38**

Table 8: Performance of the proposed MDD-VA and other baselines with Hierarchical symptom image identifier (HSII)

Model	Policy Optimization	Diagnosis success rate	Dialogue length	AMR	AMR2	Disease classifier accuracy (%)
HRL [10]		0.393	7.89	9.32	17.70	39.98
MDD-VA_MSR	Dueling DQN	0.409	7.40	11.86	20.36	41.43
MDD-VA		**0.415**	**7.15**	**13.70**	**22.58**	**42.42**
HRL [10]		0.424	8.77	9.45	17.70	42.02
MDD-VA_MSR	Double DQN	0.439	8.23	12.58	22.85	44.32
MDD-VA		**0.449**	**8.16**	**14.44**	**26.50**	**45.89**
UDP [23]		0.342	5.34	2.41	1.26	NA
HRL	DQN	0.439	11.37	7.22	24.37	0.4237
MDD-VA-MSR		0.476	10.80	9.10	28.20	47.66
MDD-VA		**0.497**	**10.66**	**10.12**	**31.54**	**48.90**
KI - CD [18] w/o multi-modality		0.514	16.05	8.64	43.47	51.89
KI - CD [18] with MSR	DQN	0.544	15.78	9.20	44.40	54.90
KI - CD [18] with multi-modality		**0.547**	**15.59**	**10.50**	**47.72**	**55.68**

Table 9: Performance of the proposed MDD-VA and other baselines with Context-aware symptom image identifier



(a)

(b)

Figure 7: a. Confusion matrix of Context-SII (context window = 3), b. Confusion matrix for failed diagnoses - The diagonal elements show the percentage of times that the agent diagnoses an incorrect disease despite predicting right disease group

The diagnosis assistant (MDD-VA) outperformed MDD-VA_MSR by a large margin across all policies and symptom image identifier, which clearly demonstrates the importance of symptoms extracted by the assistant during conversations. **iii.** *Can dialogue context help* *in interpreting an image that surfaced during conversation?* The context aware symptom image identifier (ContextSII) outperforms both pre-trained imagenet models (Table 5) and hierarchical symptom image identifier (Table 6) significantly. Thus, the answer is yes; the obtained evidence (Table 7) firmly strengthens the hypothesis that dialogue context helps in identifying an image that surfaced during the dialogue. **iv.** *Is there any co-relation between dialogue context width and context-aware image identifier's performance?* Yes, the results obtained by the ContextSII (Table 7) show a positive correlation between dialogue context window size and its performance. The performance increases as context size increases (context size = 1, 2, 3), but up to a certain point. The model performance diminishes for too wide dialogue contexts, primarily because of the variety of symptoms and numerous possible conditions associated with them. **v** In Dueling DQN, Double DQN, and DQN, the dialogue length increases as the agent's performance (Tables 8, 9) improves, which shows a positive correlation between diagnosis success rate and diagnosis time (dialogue length).

Human Evaluation To rule out the possibility of under informative assessment done by automatic metrics, we conducted the human evaluation of 100 randomly selected test samples. In this assessment, medical domain experts, including three researchers, out of which two are clinicians, have been employed to evaluate the generated samples. The samples are assessed based on *easiness of symptom investigation/diagnosis, investigation relevance, coherence, diagnosis time, and relevance of predicted disease.* The average scores obtained for different diagnosis assistants are portrayed in Figure 8. The key observations are as follows: **i**. The UDP model usually diagnoses disease based on only the chief complaints reported by the patient rather than investigating more symptoms. **ii.** When both textual and visual symptoms/signs are provided, the MDD-VA efficiently identifies patients' diseases. This is primarily due to the high relevance of images to diseases. **iii.** HRL / KI-CD agents (without multi-modality) completely flatter and inspect random symptoms when patients' self-report is comprised of only visual complaints.

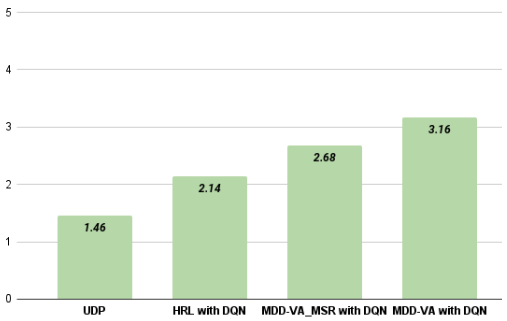


Figure 8: Human scores obtained by different diagnosis assistants

7 ANALYSIS

The detailed analyses of the performances of different models lead to the following key observations: **i.** We observed that the diagnosis assistant inspects less relevant symptoms and predicts an inaccurate disease if it misidentifies a symptom image. **ii.** The findings and case studies clearly demonstrate the vital role of dialogue context in identifying an image surface during a conversation. Furthermore, we found the model failed primarily when the dialogue context contains a variety of symptoms/symptoms that rarely occur with the symptom/sign shown in the image. One such case study has been reported in Figure 9. **iii.** In some cases, the joint intent and slot module fail to identify the complete symptom name for symptoms with multiple words, such as Lower abdominal pain. Figure 7a shows the confusion matrix of the context SII module (confusion: Skin dryness & Skin irritation, skin rash & Skin irritation). We have also reported the confusion matrix of the proposed MDD-VA diagnosis assistant for failure to diagnose in Figure 7b. The figure indicates that the agent fails, primarily because it gets confused among diseases within its corresponding disease group with a high rate of common symptoms.

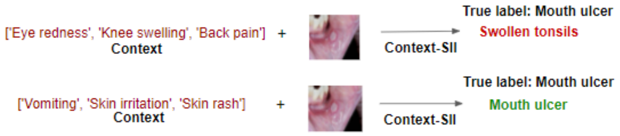


Figure 9: Performance of the context-aware symptom image identifier for same symptom image with different contexts

8 CONCLUSION

In this work, we investigated the role of multi-modal information provided by visual clues in symptom investigation and disease diagnosis processes. We proposed a novel multi-modal disease diagnosis virtual assistant (MDD-VA) to assist doctors in symptom investigation and diagnosis process, which extracts symptoms through both text and visual aids and utilizes them to identify disease accurately. We also introduced and incorporated a novel dialogue context-aware symptom identification model into the proposed framework, which leverages diagnosis/dialogue context information for identifying symptoms from images shown by the users accurately. Furthermore, we have also curated a multi-modal medical conversational dialogue corpus named Vis-MDD, where each utterance is annotated with its corresponding intent, symptom, and image information. The obtained extensive results, including human evaluation and comparisons with different uni-modal baselines and state-of-the-art models, firmly establish the significance of visual information in the disease diagnosis process. In addition to the presence of a symptom, its intensity and time-stamp/duration also have significant impacts in the disease diagnosis process. In future, we would like to build a virtual diagnosis assistant that can also extract symptoms' intensity and duration information and utilize them in the diagnosis process.

9 ACKNOWLEDGEMENT

Abhisek Tiwari graciously acknowledges the Prime Minister Research Fellowship (PMRF) Award supported by the Government of India for conducting this research. 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.

REFERENCES

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323* (2019).

[2] Leemon Baird. 1995. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*. Elsevier, 30–37.

[3] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909* (2019).

[4] Thomas G Dietterich. 2000. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of artificial intelligence research* 13 (2000), 227–303.

[5] Vojtech Franc and Václav Hlavác. 2002. Multi-class support vector machine. In *Object recognition supported by user interaction for service robots*, Vol. 2. IEEE, 236–239.

[6] Xuan Guo, Rui Li, Qi Yu, and Anne R Haake. 2017. Modeling Physicians' Utterances to Explore Diagnostic Decision-making.. In *IJCAI*. 3700–3706.

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[8] Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[9] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-End Task-Completion Neural Dialogue Systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 733–743.

[10] Kangenbei Liao, Qianlong Liu, Zhongyu Wei, Baolin Peng, Qin Chen, Weijian Sun, and Xuanjing Huang. 2020. Task-oriented Dialogue System for Automatic Disease Diagnosis via Hierarchical Reinforcement Learning. *arXiv preprint arXiv:2004.14254* (2020).

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[12] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[13] Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. 2019. Mnn: multimodal attentional neural networks for diagnosis prediction. *Extraction* 1 (2019), A1.

[14] Nagarajan Ramakrishnan, Bharath Kumar Tirupakuzhi Vijayaraghavan, and Ramesh Venkataraman. 2020. Breaking Barriers to Reach Farther: A Call for Urgent Action on Tele-ICU Services. *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine* 24, 6 (2020), 393.

[15] Kristine Rasmussen, José Marcano Belisario, Petra A Wark, Joseph Antonio Molina, Stewart Lee Loong, Ziva Cotic, Nikos Papachristou, Eva Riboli-Sasco, Lorainne Tudor Car, Eve Marie Musulanov, et al. 2014. Offline eLearning for undergraduates in health professions: a systematic review of the impact on knowledge, skills, attitudes and satisfaction. *Journal of global health* 4, 1 (2014).

[16] Kai-Fu Tang, Hao-Cheng Kao, Chun-Nan Chou, and Edward Y Chang. 2016. Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In *NIPS Workshop on Deep Reinforcement Learning*.

[17] Gerald Tesauro. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38, 3 (1995), 58–68.

[18] Abhisek Tiwari, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning. *Knowledge-Based Systems* 242 (2022), 108292.

[19] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[20] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific reports* 11, 1 (2021), 1–13.

[21] Joseph B Walther, Brandon Van Der Heide, Sang-Yeon Kim, David Westerman, and Stephanie Tom Tong. 2008. The role of friends' appearance and behavior on evaluations of individuals on Facebook: Are we known by the company we keep? *Human communication research* 34, 1 (2008), 28–49.

[22] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. 2016. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1995–2003.

[23] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 201–207.

[24] Xiaoling Xia, Cui Xu, and Bing Nan. 2017. Inception-v3 for flower classification. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 783–787.

[25] Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7346–7353.

[26] Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhumin Chen, Zhaochun Ren, and Huasheng Liang. 2021. Mˆ 2-MedDialog: A Dataset and Benchmarks for Multi-domain Multi-service Medical Dialogues. *arXiv preprint arXiv:2109.00430* (2021).

[27] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. MedDialog: Large-scale Medical Dialogue Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.