Seeing is believing! Towards Knowledge-Infused Multi-modal Medical Dialogue Generation

Abhisek Tiwari[1,*], Shreyangshu Bera[1], Preeti Verma[1], Jaithra Varma Manthena[1], Sriparna Saha[1], Pushpak Bhattacharyya[2], Minakshi Dhar[3] and Sarbajeet Tiwari[4]

[1]Indian Institute of Technology Patna India
[2]Indian Institute of Technology Bombay
[3]All India Institute of Medical Sciences, Rishikesh
[4]Midnapore Homoeopathic Medical College and Hospital
[*]Email: abhisektiwari2014@gmail.com

Abstract

Over the last few years, artificial intelligence-based clinical assistance has gained immense popularity and demand in telemedicine, including automatic disease diagnosis. Patients often describe their signs and symptoms to doctors using visual aids, which provide vital evidence for identifying a medical condition. In addition to learning from our experiences, we learn from well-established theories/ knowledge. With the motivation of leveraging visual cues and medical knowledge, we propose a transformer-based, knowledge-infused multi-modal medical dialogue generation (*KI-MMDG*) framework. In addition, we present a discourse-aware image identifier (DII) that recognizes signs and their severity by leveraging the current conversation context in addition to the image of the signs. We first curate an empathy and severity-aware multi-modal medical dialogue (*ES-MMD*) corpus in English, which is annotated with intent, symptoms, and visual signs with severity information. Experimental results show the superior performance of the proposed *KI-MMDG* model over uni-modal and non-knowledge infused generative models, demonstrating the importance of visual signs and knowledge infusion in symptom investigation and diagnosis. We also observed that the DII model surpasses the existing state-of-the-art model by 7.84%, indicating the crucial significance of dialogue context for identifying a sign image surfaced during conversations. The code and dataset are available at https://github.com/NLP-RL/KI-MMDG.

Keywords: Dialogue System, Clinical Assistance, Automatic Disease Diagnosis, Multi-modality, Knowledge Infusion

1. Introduction

Disease diagnosis is the primary and crucial stage of any medical treatment process. In the diagnosis stage, doctors investigate patients' health conditions and determine diseases by assessing their self-report and other symptoms. In recent few years, dozens of surveys (Lorkowski and Jugowicz, 2020) have indicated an alarming doctor-population ratio. Many countries[1] across the globe have only 0.00001%-0.001% doctors. With the motivation of efficient utilization of doctors' time and providing an accessible platform for early diagnosis, automatic disease diagnosis using artificial intelligence (AI) is gaining huge popularity and demand in both medical research and industry communities (Mintz and Brodie, 2019; Lin et al., 2021).

In real life, we often describe our primary complaints and difficulties to doctors with the help of visual aids (Javaid et al., 2016; Salimi et al., 2018). Some symptoms, namely swelling and skin rashes, are difficult to convey through text. Furthermore, many medical terminologies and symptoms, such as mouth ulcers and skin growth, are unfamiliar to a large percentage of the population. Thus, visual reporting seems to be the most apparent and appropriate solution in such scenarios (Figure 1). However, existing automatic disease diagnosis assistants (Wei et al., 2018; Xia et al., 2020; Chen et al., 2022; Tiwari et al., 2022b) only extract symptoms through users' text messages and fail to utilize patients' signs described through visuals.

A dialogue is a connected sequence of utterances, with each subsequent response generally aligning with the preceding ones. Thus, the dialogue context can effectively boost the assurance of comprehending a message (textual/acoustic/visual) when prior responses (dialogue history) are taken into consideration. Inspired by this, we propose a discourse-aware image identification (DII) module and integrate it with the multi-modal dialogue generation framework. Doctors also utilize diagnosis principles and clinical knowledge bases to gain proficiency in diagnosis tasks (Jiang et al., 2017). Furthermore, they consider sign/symptom severity into account to diagnose a patient accurately. Motivated by these observations, we aim to investigate some fundamental research questions related to multi-modal disease diagnosis.

Research Questions In this paper, we aim to in-

[1]https://data.worldbank.org/indicator/SH.MED.PHYS.ZS

LREC-COLING 2024, pages 14513–14523

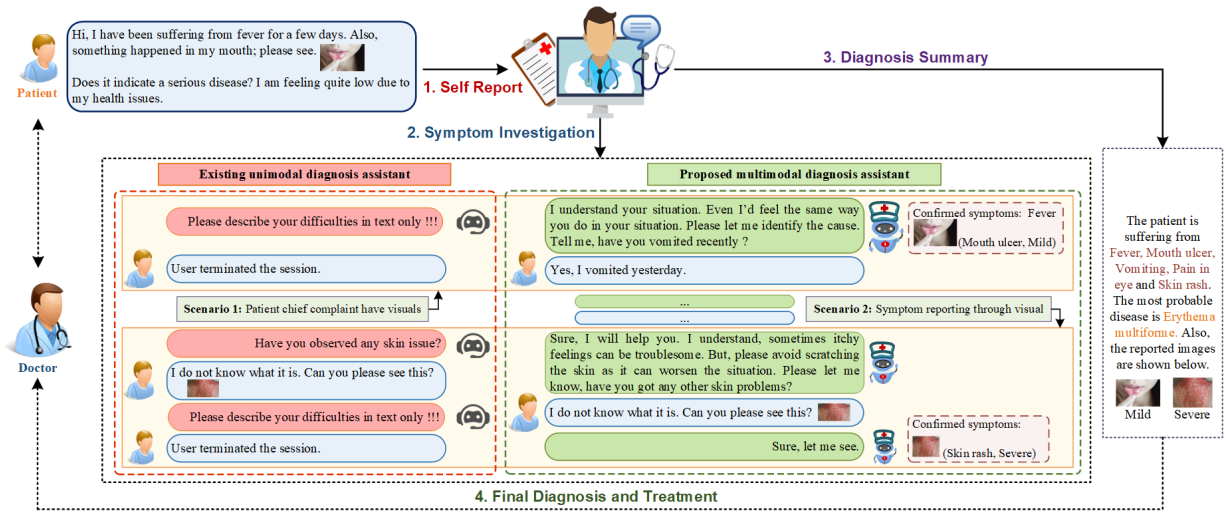

Figure 1: Limitations of existing uni-modal disease diagnosis assistants and the behavior of our proposed multimodal disease diagnosis assistant

vestigate the following three research questions: (i) *Does a diagnosis assistant diagnose patients more accurately and satisfactorily if it considers visual signs and their severity in addition to symptoms conveyed through text?* (ii) *Can dialogue context help in identifying a sign image and its severity, which appears during the conversation?* (iii) *What impact might global knowledge, such as knowledge of symptom-disease associations, have on the diagnosis ability of diagnosis assistants? Does the mechanism of knowledge infusion influence its efficacy?*

A major bottleneck for the development of AI-driven healthcare technology is the lack of structured medical datasets. Despite the huge significance of the visual form of sign and symptom reporting, we could not find a single visual sign-aided disease diagnosis dialogue corpus in English. Motivated by the limitation and to investigate the research questions, we first attempt to curate an Empathy and severity-aware Multi-Modal Medical Dialogue (*ES-MMD*) corpus. The dataset bridges the following three gaps: (a) Textual-visual symptom investigation and disease diagnosis, (b) Utterance semantic understanding (intent, symptom, and symptom severity), and (c) Empathetic dialogue responses.

The key contributions of the paper are as follows:

- We curate an Empathy and severity-aware Multi-Modal Medical Dialogue (*ES-MMD*) corpus in English, where each utterance is annotated with its corresponding intent, sign, symptom, and severity level.
- We propose a transformer-based Knowledge-Infused, Multi-modal Medical Dialogue Generation (*KI-MMDG*) framework, which leverages a discourse-aware selective filtering strategy for knowledge distillation and a natural language understanding (NLU) module for semantic understanding of textual-visual utterances. The *KI-MMDG* model involves three key novel components: (a) Utterance semantic incorporated dialogue generation, (b) Discourse aware knowledge selection, and (c) Discourse-aware image identification.
- Our proposed *KI-MMDG* exhibits a substantial performance improvement over several non-knowledge infused uni-modal medical dialogue generation models across a variety of evaluation metrics, including human evaluation. Additionally, the DII model surpasses existing pre-trained image models in both symptom identification (by 7.84%) and severity recognition (by 2.63%), achieving state-of-the-art performance for image identification in dialogue settings.

2. Related Work

Automatic disease diagnosis dialogue system Wei et al. (Wei et al., 2018) formulated symptom investigation as a task-oriented dialogue system where the agent extracts symptoms through conversation and diagnoses a disease as per observed symptoms. Zhong et al. (Zhong et al., 2022) have proposed an integrated and synchronized two-level policy framework using hierarchical reinforcement learning (Dietterich, 2000). The model outperformed the flat policy approach (Wei et al., 2018) by a significant margin, demonstrating the efficacy of disease department-aware symptom investigation. Xu et al. (2023) (Xu et al., 2023) demonstrated the effectiveness of utilizing an information graph, incorporating clinicians' dialogue acts and patients' medical entities from their utter-

ances within the dialogue context, to generate a satisfactory medical response.

Knowledge aware Dialogue Generation In (Ham et al., 2020), the authors have emphasized the implication of utilizing fine-grained information (database and dialogue act) of dialogue history in order to generate an adequate and knowledge-consistent responses. However, the model has a high annotation requirement, and it must have annotated data even during inference time. The work (Liu et al., 2021) showed that how we design our input (prompt) to generative models has a significant impact on the model's performance. Sun et al., (Sun et al., 2021) proposed a knowledge-infused dialogue language understanding and language generation model that leverages additional relevant knowledge for providing relevant context in order to categorize the text and generate follow-up responses. In this work (Zhao et al., 2022), the authors introduced MedPIR, leveraging knowledge-aware dialogue graphs and recall-enhanced generators, to overcome the challenge of capturing essential information from extensive medical dialogues.

Multi-modal dialogue system The work (O'Hare and Smeaton, 2009) is the first attempt to explore the effectiveness of context (such as background, clothing, and style) in accurately identifying a person's image. Ge et al. (Ge et al., 2017) proposed a novel deep convolutional neural network with a saliency feature descriptor for capturing discriminative features of two different modalities for skin lesion identification. In (Venugopalan et al., 2021), authors have proposed a multi-modal deep learning model for Alzheimer disease prediction, which utilizes three different modalities, namely imaging, genetic, and clinical test data, for analyzing patients' conditions. In (Tiwari et al., 2022a), a multi-modal diagnosis assistant was presented, featuring a dialogue policy trained through hierarchical reinforcement learning with a template-based natural language generator.

3. Dataset

We first extensively investigated the existing benchmark medical diagnosis dialogue corpora, and the summary is presented in Table 1. Motivated by the unavailability of visual and severity-aided diagnosis conversational dataset and the efficacy of an end-to-end multi-modal dialogue system, we make the very first attempt to develop a multi-modal conversational disease diagnosis corpus named Empathetic and Severity-aware Multi-modal Medical Dialogue (*ES-MMD*) in English.

Data Collection We found a single diagnostic dataset in English named Synthetic dataset (SD) (Zhong et al., 2022), which includes patients' symptoms (self-report and implicit symptoms) and their corresponding diseases for each sample. We analyzed the dataset thoroughly with the help of two clinicians. We identified 23 symptoms from the SD dataset that are either hard to specify through text or are not commonly known. We selected only 17 signs/symptoms out of the list for multi-modal conversation creation, primarily due to the lack of sufficient images of other signs/symptoms. We then collected the symptom images from open-source platforms and filtered out inappropriate and blurry images with the help of clinicians.

Dataset	Language	Conversation	Intent	Symptom	Multimodality	Severity
RD (Wei et al., 2018)	Chinese	×	×	×	×	×
DX (Xu et al., 2019)	Chinese	✓	×	✓	✓	×
M^2 - MedDialogue (Yan et al., 2021)	Chinese	✓	×	✓	×	×
MedDialog-EN (Zeng et al., 2020)	English	×	×	×	×	×
MedDG (Liu et al., 2020)	Chinese	✓	×	✓	×	×
SD (Zhong et al., 2022)	English	×	×	×	×	×
ES-MMD (ours)	English	✓	✓	✓	✓	✓

Table 1: Statistics of existing medical datasets for disease diagnosis task

***ES-MMD*: Dialogue Creation and Annotation** We selected 100 random diagnosis cases from the SD dataset. The SD dataset comprises samples in a database format, including case ID, self-report, implicit symptoms, and final diagnosis. The two clinical authors tagged the collected visual sign images into one of the three severity categories: mild, moderate, or severe. With the help of these clinicians, we curated a conversation-based sample dataset corresponding to the 100 diagnosis cases and annotated them with their corresponding intent, symptoms, and image information. With the help of the curated sample conversational dataset and the detailed guidelines provided by the clinicians, three biology graduates created conversations for a subset of SD dataset samples (1742 cases) and annotated them with intent and symptoms. Note that all samples are unique, and sample information for each one is taken from the SD dataset. In order to measure the annotation agreement among the annotators, we calculated the Fleiss kappa (Fleiss et al., 2013), which was found to be 0.76, indicating a significant uniform annotation. The detailed statistics of the *ES-MMD* dataset and a conversation from the curated corpus are reported in Table 2 and Figure 2. Further statistics have been provided in the Appendix section.

Role of Intent and Symptom Annotation In order to make an end-to-end medical diagnosis dialogue system capable of communicating with users in natural language form, we developed a dyadic multimodal dialogue corpus and tagged each utterance of the conversations with intent & symptom information.

Significance of Multi-modality During a conversation with a doctor, we often use visuals when

Attribute	Value
# of Dialogues	1742
# of Utterances	12466
Avg. dialogue length	7.16
# of intents	3
# of diseases	90
# of symptoms	266
# of signs	17
# of symptom images	1805

Table 2: Statistics of *ES-MMD* Dataset

Figure 2: A conversation from the curated *ES-MMD* corpus

we are unsure of symptom names, or it is difficult to describe some symptoms/signs precisely through text. For example, many people are unaware that the images shown in Figure 3 (column 2) are instances of mouth ulcers. Thus, we curated a multi-modal disease diagnosis dialogue corpora that includes both textual and visual symptom reporting.

Importance of Symptom Severity Information In real life, doctors do not diagnose a disease only based on the presence of some symptoms and signs. They also consider their severity information in disease diagnosis, which helps them in narrowing possible disease space and identify patient disease effectively. Some images for a few symptoms with their severity levels are shown in Figure 3.

Role of Empathy During the consultation, patients' comfort and users' satisfaction are also crucial. This helps build trust between patient and doctor and increases patient compliance. For example, in Figure 2, the patient shares his insecurity about his skin condition. The agent replies empathetically, ensuring the patient is not alone in these situations and that others have also suffered from such situations and felt the same way.

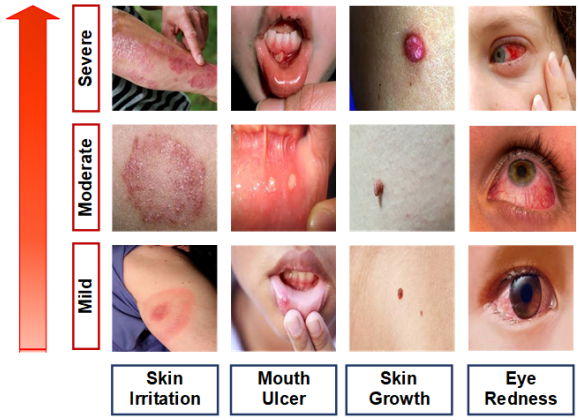

Figure 3: Some visual samples of signs

4. Methodology

The proposed end-to-end knowledge-infused multimodal medical dialogue generation system, *KI-MMDG*, is shown in Figure 5. There are three key stages: Natural Language Understanding (NLU), Knowledge Infusion, and Response Generation. The working methodologies for each module are explained and illustrated in the following sections.

4.1. Natural Language Understanding

We introduce a natural language understanding (NLU) module (Figure 4) coupled with the generation model to extract medical entity information from user utterances (Yadav et al., 2018). The NLU component consists of two modules, namely Intent & Symptom module (left side) and the Discourse aware image identifier (right side). Intent & Symptom module takes the user's utterance (text: U^t_{text}) as input, and the module predicts its intent and symptom sequence tag. In our proposed framework, we have utilized the joint BERT (Chen et al., 2019) model, which jointly optimizes both intent identification and symptom sequence labeling tasks.

Discourse-aware Image Identification (DII) Motivated by the importance of context in conversation, we propose a discourse-aware image identification (DII) model for sign and severity identification from an image. The architecture of the DII model is shown in Figure 4 (right side). It takes the dialogue context (confirmed symptoms and signs) embedding from the Clinical-BERT (Alsentzer et al., 2019) and image features from the VGG19 model (Simonyan and Zisserman, 2014) as input. The concatenated textual and visual representation is passed through two feed-forward neural networks - one identifies sign/symptom (I_t), and the other recognizes its severity (Sev_t). Mathematically, it can be expressed as follows:

$$I_t, Sev_t = \text{DII}(image_t, C_t) \quad (1)$$

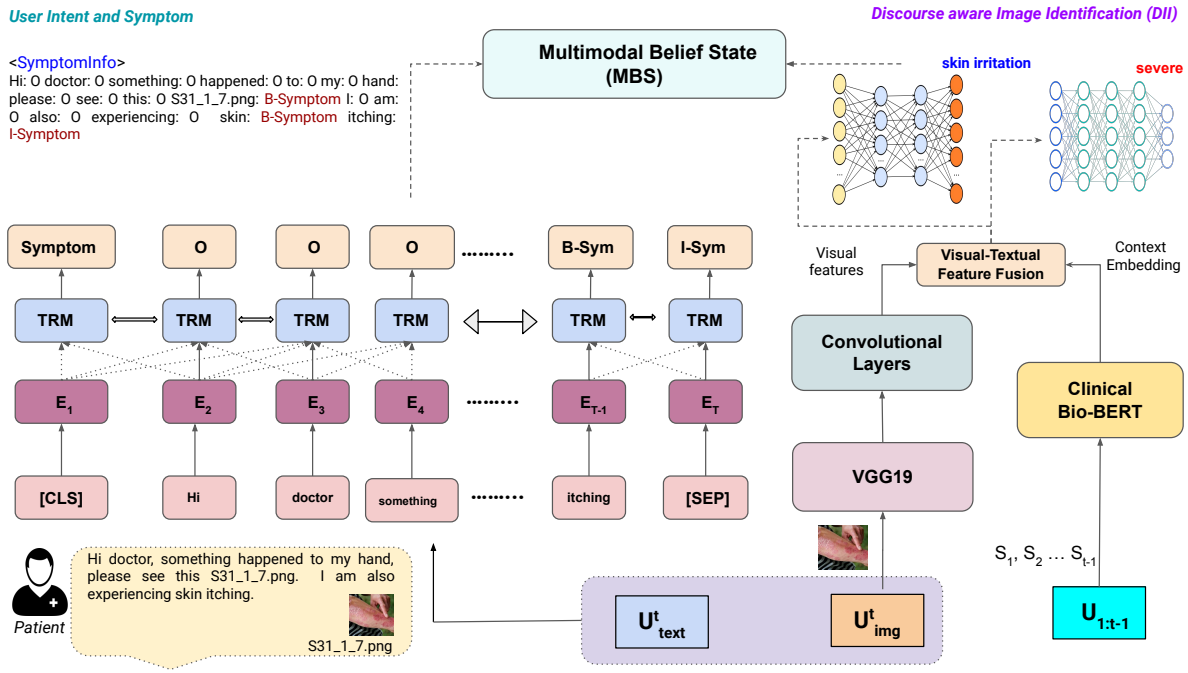

Figure 4: Proposed natural language understanding model which identifies medical entities, visuals, and severity of symptoms

$$C_t =< s_1, s_2.....s_{t-1} > \quad (2)$$

where DII, $image_t$, and C_t are discourse-aware image identifier, sign image, and dialogue context at tth time step, respectively. Here, s_i denotes the symptom conveyed in i^{th} user utterance.

4.2. Knowledge Infusion

Clinical knowledge aids physicians in narrowing the scope of an investigation given a context. Inspired by the observation, we incorporate a symptom-symptom/symptom-disease (S-S-D) knowledge graph in the medical dialogue generation task. We created the knowledge graph (S-S-D) from the ES-MDD dataset, where symptoms and diseases are nodes. An edge between two nodes indicates their co-occurrence. The edges are weighted through the symptom frequency-inverse disease frequency (sf-idf) method (Ramos et al., 2003). The edge weights between symptom-disease e(s, d) and symptom-symptom e(s_i, s_j) are computed as follows:

$$e(s,d) = sf(s,d) * idf(s,d) \quad (3)$$

$$sf(s,d) = \frac{n_{sd}}{\sum_k n_{kd}} \quad (4)$$

$$idf(s) = \log \frac{|D|}{|d : s \in disease_j|} \quad (5)$$

$$e(s_i, s_j) = \frac{n(s_i, s_j)}{\sum_k n(s_i, s_k)} \quad (6)$$

where n_{sd} is the number of cases where symptom (s) has occurred with the disease, d. k ranges in symptom space, and $|D|$ signifies the total number of diseases.

Discourse-aware Selective Filtering (DSF) The knowledge graph is substantially large; thus, infusing the entire relationship does not seem feasible or effective for response generation. We propose and employ a novel discourse-aware selective filtering (DSF) knowledge distillation method for extracting relevant relations depending on the conversation context. The DSF function is defined in Algorithm 1. It chooses a subset of the knowledge graph (KG_{t+1}) that contains the nodes (symptoms) and their relationships discussed in previous patient and doctor utterances (C_t).

Algorithm 1 Discourse-aware Selective Filtering (DSF)

Initialization: KG = {(s_i, s_j, a_{ij})} where s_i, s_j are nodes, a_{ij} is the edge weight.
Input: Current Knowledge Graph (KG_t), PSR: Patient Self Report, and Current Discourse (C_t)
Output: Filtered Knowledge Graph (KG_{t+1})

C_t = {(s_0), (s_1), (DII(i_2), s_2)...(s_t)}
KG_{t+1} = KG_t
Potential_diseases (PD) = $\Pi_{i=1}^{i=3}$ i^{th} most_associated_disease (PSR)
for d in PD **do**
    triplet = (PSR, d, a_{PSR-d}) $\Rightarrow$ a_{PSR-d}: edge (PSR, d) weight (Equation 3)
    KG_{t+1} = append(KG_{t+1}, triplet)
end for
for s in C_t[-1] **do**
    ps = $\Pi_{j=1}^{j=3}$ j^{th}-most_associated_symptom(s)
    for k in ps **do**
        triplet = (s, k, a_{s-k})
        KG_{t+1} = append(KG_{t+1}, triplet)
    end for
end for
return KG_{t+1}

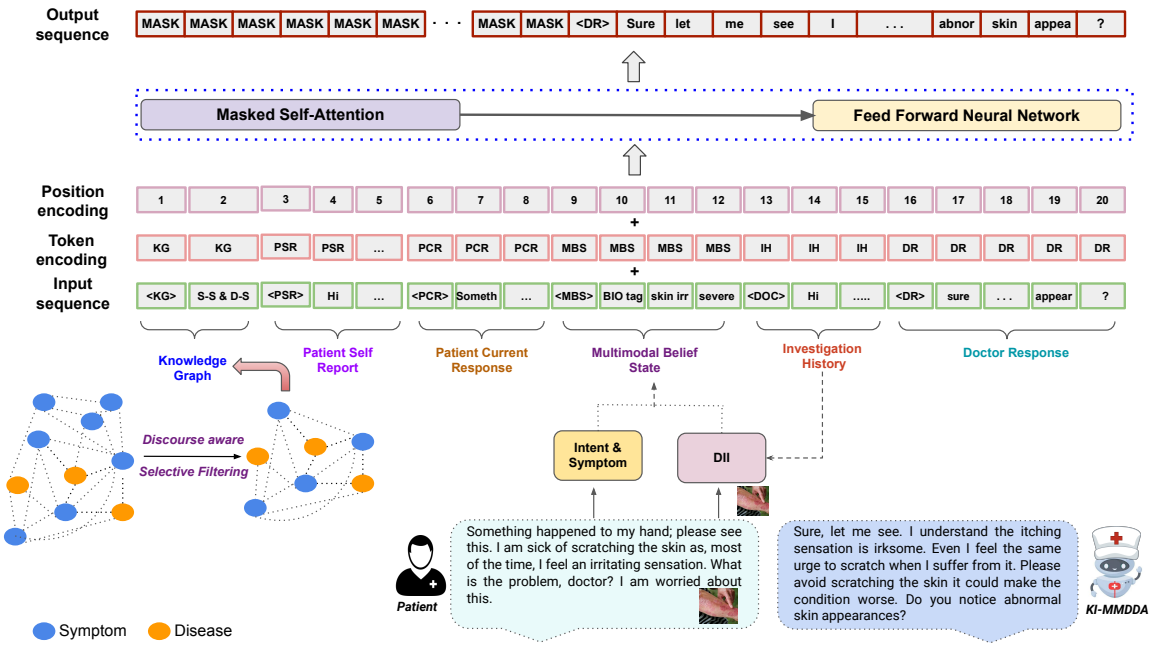

Figure 5: Proposed architecture of the knowledge-infused multi-modal medical dialogue generation (*KI-MMDG*) framework

4.3. Response Generation

The proposed NLU module extracts semantic information from user utterances and encodes the information, namely intent, symptoms, and visual signs with severity, in the multi-modal belief state (MBS). We fine-tuned the pre-trained DialoGPT (Zhang et al., 2020) and conditioned the response on discourse-filtered S-S-D knowledge graph and Multimodal Belief State (MBS). MBS includes both textual and visual information, namely symptoms, signs, and its severity pertaining to current utterance. The model considers patient utterance, dialogue context, and distilled knowledge triplets segregated with special tokens as input. In the second stage, the model takes token encoding, representing different segments of the input sequence. In the final layer of input encoding, positional encoding is added to preserve the order of input word sequences. The sequences are fed into the masked self-attention layer for attending to the importance of different segments. The attended token is fed to feed-forward networks, which autoregressively generate tokens for doctor's response. We utilize categorical cross-entropy (CE) loss for measuring the dissimilarity between the generated response and its respective gold response.

4.4. Experimental Setup

We have utilized the PyTorch framework for implementing the proposed models. The proposed *KI-MMDG* model was trained for 10 epochs on an RTX 2080 Ti GPU, which took around 2 hours. The base of the proposed *KI-MMDG* model is a GPT2. The generation models have been trained and evaluated with 80% and 20% samples of the curated dialogue dataset, respectively. The hyperparameter values for dialogue generation are as follows: batch size (4), learning rate, α (6.25e-5), and optimizer (Adam). The hyperparameter values for discourse-aware image identification (DII) model are as follows: train-test split (80%, 20%), batch size (32), leaning rate (1e-3), Optimizer (Adam), no. of convolution layers (3).

5. Result

We employed the most popular automatic evaluation metrics, namely BLEU, ROUGE, and METEOR to evaluate the generation quality of the proposed model. All the reported values in the following tables are statistically significant as we have performed Welch's t-test (Welch, 1947) at a 5% significance level. Table 3 shows the performance of the intent & symptom module for user utterance identification and symptom labeling tasks. Based on the experiments, we report the following answers (with evidence) and observations to our investigated research questions (RQ).

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

Table 3: Performance of the joint intent & symptom module

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	BLEU	ROUGE 1	ROUGE L	METEOR
DLGNet (Oluwatobi and Mueller, 2020)	21.84	9.66	4.21	2.10	9.45	26.86	25.05	21.98
DLGNet with VSI-KG	**25.48 (3.64↑)**	**12.26 (2.60↑)**	**6.86 (2.65↑)**	**3.51 (1.41↑)**	**12.02 (2.57↑)**	**29.45(2.59↑)**	**28.82(2.57↑)**	**25.79(3.77↑)**
BART (Lewis et al., 2020)	23.19	12.34	7.32	4.37	11.80	27.77	27.37	29.66
BART with VSI-KG	**25.69(2.50↑)**	**15.07(2.73↑)**	**9.41(2.09↑)**	**5.62(1.25↑)**	**13.95(2.15↑)**	**29.93(2.23↑)**	**29.58(2.21↑)**	**32.41(2.75↑)**
DialoGPT (Zhang et al., 2020)	26.59	16.16	10.52	6.92	15.05	30.63	30.22	34.07
KI-MMDG	**28.53(1.94↑)**	**18.41(2.25↑)**	**12.28(1.76↑)**	**8.34(1.42↑)**	**16.89(1.84↑)**	**32.69(2.06↑)**	**32.25(2.03↑)**	**36.52(2.45↑)**

Table 4: Performance of different baselines and proposed models incorporated with the proposed visual sign and knowledge (VS-KG) guided disease diagnosis component

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	BLEU	ROUGE 1	ROUGE L	METEOR
DLGNet with only KG	23.54	8.65	4.85	0.99	9.51	29.18	26.01	22.78
DLGNet with only VS	23.59	10.65	5.81	2.44	10.62	28.31	25.45	26.69
DLGNet with only VSI	24.25	10.77	5.91	3.40	11.08	29.37	27.72	25.42
BART with only KG	25.03	14.20	9.30	5.91	13.61	29.04	28.60	31.11
BART with only VS	25.37	14.85	9.39	5.63	13.81	29.58	29.12	32.35
BART with only VSI	25.69	14.98	9.65	5.93	14.06	29.92	29.43	31.88
DialoGPT with complete KG(w/o DSF)	1.58	0.73	0.48	0.28	0.77	1.99	1.95	2.20
DialoGPT with only KG	27.72	17.27	11.39	7.70	16.02	31.74	31.21	35.42
DialoGPT with only VS	27.53	17.08	11.47	7.82	15.98	31.54	31.13	35.29
DialoGPT with only VSI	27.11	16.94	11.45	7.74	15.81	31.23	30.91	35.07

Table 5: Ablation study– performances of the proposed *KI-MMDG* model with different components. Here, KG, VS, VSI, and DSF refer to the knowledge graph, visual symptom, visual & severity information, and discourse-aware filtering

RQ 1. Does a diagnosis assistant diagnose patients more accurately and satisfactorily if it considers visual signs/symptoms in addition to textual/verbal symptoms?The performances attained by different baselines and the proposed *KI-MMDG* model are reported in Table 4. The obtained improvements (across all evaluation metrics) over uni-modal baselines firmly demonstrate the importance of utilizing visual sign information in the disease diagnosis process. Furthermore, the proposed model also performs superior in human evaluation (Table 9), exhibiting the co-relation between user satisfaction and the flexibility of providing signs/symptoms through visuals.

RQ 2. Can dialogue context help in identifying a sign image and its severity, which appears during the conversation? We first experimented with the existing pre-trained vision models for symptom image identification and severity recognition, and the obtained results are reported in Table 6. The performances obtained by the proposed discourse-aware image identifier (DII) are reported in Table 6 (for image identification) and Table 7 (for severity recognition). The results show that the inclusion of discourse context helped these models identify both signs and severity more efficiently. The result demonstrates that neither a very long history nor merely the immediate context performs ideally but rather a context of some earlier utterances.

RQ 3(a). What impact might additional information, such as knowledge of symptom-disease associations, have on diagnosis assistants' diagnosis ability? The knowledge-infused models surpassed models w/o additional knowledge for every case in all evaluation metrics (Table 4). The model proposed includes three different modalities of input, apart from dialogue utterances, which include knowledge graph (KG), visual sign (VS), and visual sign's severity information (VSI). In order to understand the efficacy of each component, we have performed a detailed ablation study, and the obtained results are provided in Table 5. To comprehend the individual impact of each modality, we disabled the other modalities during experimentation. For instance, when using only KG with DLGNet, the model takes into account only the KG and dialogue utterances without considering visual information or severity for doctor response generation. The results indicate that knowledge infusion has led to improvements across all evaluation metrics in all three settings. We observed that the infusion of

Model	Accuracy (%)	F1-Score
CNN (Li et al., 2014)	40.99	0.4247
Inception v3 (Xia et al., 2017)	66.14	0.6475
Inception v3 + Conv Layers	72.29	0.7163
DenseNet121 (Huang et al., 2017)	68.17	0.6712
DenseNet121 + Conv Layers	75.58	0.7412
DenseNet169 (Serte et al., 2022)	72.27	0.7157
DenseNet169 + Conv Layers	78.51	0.7734
VGG19 + Conv Layers (Gupta et al., 2022)	81.11	0.7924
DII with CW=1	82.82	0.8271
DII with CW=2	85.27	0.8505
DII with CW=3	**88.95** (**7.84** ↑)	**0.8703** (**0.0779** ↑)
DII with CW=4	81.59	0.8171

Table 6: Performance of different models for visual signs identification

Model	Accuracy (%)	F1-Score
VGG19+ Conv Layers (Gupta et al., 2022)	50.65	0.5086
DII with CW=1	49.34	0.4907
DII with CW=2	52.63	0.5233
DII with CW=3	**53.28** (**2.63** ↑)	**0.5117** (**0.0091** ↑)
DII with CW=4	50.01	0.4771

Table 7: DII with varying dialogue context window for severity recognition

Context	Gold	DialoGPT	DialoGPT+VSI	DialoGPT+KG	*KI-MMDG*
Hello doctor, I am suffering from Lower body pain. And because of it, I cannot sleep properly	I understand; it must be hard. Do you also have back pain?	Please be patient; it can be treated. Do you feel headache?	Please let me investigate. Do you feel headache ?	I understand that the pain would be hard. Do you also feel back pain?	I can understand how hard it is for you to sleep with this pain. Let me check, do you have back pain?
I am having weight gain. I feel like I look ugly because of it. ... Do you feel fatigue? yes	You don't have to think like that. You will be fine. You seem to have Cushing syndrome	Weight gain is a common problem these days. You will be fine if you take care of it. Do you have aches all over?	Weight gain is a common problem these days and can be completely curable as well. Are you suffering from Acne or pimples?	I can understand your concern. Weight gain can be due to various reasons. Let me ask you a few questions to get a better idea of its cause. Do you have skin lesion?	You are not alone. Weight gain is a common problem these days and can be completely curable as well. Based on your symptoms, you have Cushing syndrome.
Hello doctor, can you please check my eye? I am not sure what is wrong with it [moderate Swollen eye].	Is there any neck swelling?	And do you have any skin swelling? Please show.	Do you have any skin rash?	Do you have any skin rash? Please show me	Do you have neck swelling? Please show your neck.

Table 8: Some generated responses by different models for some common test cases

the symptoms and diseases relation significantly narrows the exploration space for the agent, resulting in context-based symptom inspection and disease diagnosis.

RQ 3(b). Does the mechanism of knowledge infusion impact on the adequacy of dialogue response generation While knowledge is crucial, focusing on relevant knowledge is more significant while solving a task. A similar behavior has been observed in the knowledge infusion experiment; the model that utilizes the proposed DSF knowledge distillation method (Table 5: DialoGPT with only KG) performs significantly superior to the model that utilizes the entire knowledge graph (Table 5: DialoGPT with complete KG w/o DSF). The poor performance was observed because the major portion of the input (KG) was always the same. As a result, the model is unable to distinguish between different dialogue contexts.

Human Evaluation We also conducted a human evaluation of 100 randomly selected test samples. For this evaluation, two professionals with expertise in the medical field and an independent researcher, distinct from the authors, were engaged to assess the generated samples by various models, all while maintaining the confidentiality of the model names. The evaluation encompasses five key metrics: easiness of symptom investigation (E), investigation relevance (IR), empathy (Emp), diagnosis time (T), and relevance of predicted disease (RPD). All the samples are rated on a scale from 1 (extremely poor) to 5 (excellent). The obtained scores by different models are reported in Table 9.

Model	E	IR	Emp	T	RPD	Avg.
DialoGPT	2.46	2.14	2.51	2.38	1.98	2.29
DialoGPT with only KG	2.64	3.42	2.79	2.45	2.80	2.82
DialoGPT with only VSI	3.02	2.67	2.66	**2.73**	2.21	2.66
KI-MMDG	**3.18**	**3.66**	**2.92**	2.62	**2.44**	**2.96**

Table 9: Human evaluation scores for different generation models

6. Analysis

The comprehensive analyses of the performances of different models and case studies lead to the following key observations: (i) We observed an interesting finding that the knowledge infusion has not only improved symptom inspection relevancy but also significantly influenced the model's ability to generate appropriate and empathetic responses because of the in-depth understanding of user concerns and potential diseases (Table 9 and Table 8). (ii) Dialogue context significantly matters in sign image identification surfaced during dialogue (Table 6, Figure 6, and Figure 7). The DII model gets the context (skin-related issues - abnormal appearing skin and itching of skin) and successfully identifies the image, whereas the non-context aware image identifier gets it confused with neck swelling. The context (skin lession) helped in identifying the severity level of the skin dryness image (Figure 7). The skin lesion information may indicate that the severity level is mild because skin lesion usually occurs with mild skin dryness. (iii) The proposed model suffers from sequence agnostic symptom investigation, i.e., when a fever request appears at a specific turn of a dialogue and the agent inquires about cold symptom. The agent gets penalized even when the patient suffers from cold, and the information is present at an upcoming turn.

Figure 6: Efficacy of incorporating context for identifying an image

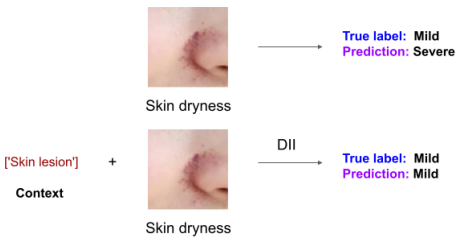

Figure 7: Importance of context for identifying severity of a visual sign

We also observed some weaknesses with the proposed model, which are as follows: (i) In medical dialogue generation, medical entities, such as symptoms and diseases, are more important than other words. However, the utilized traditional categorical cross-entropy loss function does not capture the behavior. (ii) Sequence agnostic symptom investigation: When a fever request is present in the corpus at a specified turn in a dialogue, and the agent requests information about the symptom, cold. Then, the agent gets penalized for this case even when the inspection of cold is also present in the dialogue at an upcoming turn. (iii) In real life, doctors take into account a patient's gender and age to perform more precise and effective diagnoses; we have not incorporated such demographic data in our work due to the limited scope of the primary database used for benchmarking the new dialogue dataset. In the future, we aim to investigate the effectiveness of such personal information, in addition to clinical signs and symptoms, for generating an adequate ad relevant medical dialogue response.

7. Conclusion

In this work, we proposed a knowledge-infused, multi-modal medical dialogue generation (*KI-MMDG*) framework that leverages a discourse-aware selective filtering technique for knowledge distillation and a natural language understanding module for semantic understanding of textual-visual utterances. Furthermore, we also proposed a discourse-aware image identification (DII) model that exploits dialogue context to identify an image and its severity effectively. The proposed *KI-MMDG* model outperforms several transformer-based dialogue generation models in both automatic and human evaluations by a significant margin. The obtained improvements and detailed ablation study firmly establish the efficacy of (a) visual signs, (b) discourse-aware selective filtering (DSF) for knowledge infusion, and (c) discourse information for identifying an image surface during the conversation. In the future, we would like to develop a new loss function for dialogue settings, which semantically evaluates the significance of generated utterances in relation to context in addition to n-gram matching at the utterance level.

Ethical Consideration

The medical field is highly sensitive and specialized, and thus clinical validity holds paramount importance. We have strictly followed the guidelines established for legal, ethical, and regulatory standards in medical research during the *ES-MMD* curation process. With this in mind, we have not added or removed any entity in a conversation corresponding to the reported diagnosis sample in the SD dataset. We ensure that there are no copyrighted images in the curated dataset. Also, the curated dataset does not reveal users' identities. The annotation guidelines are provided by two clinical authors, and the dataset is thoroughly checked and corrected by them. Furthermore, we have also obtained approval from our institute's healthcare committee and ethical review board (ERB) to employ the dataset and carry out the research.

Acknowledgement

Abhisek Tiwari acknowledges the generous support provided by the Prime Minister Research Fellowship (PMRF) grant, which facilitated the research. The Prime Minister's Research Fellows (PMRF) scheme aims at enhancing the quality of research in various higher educational institutions across the country. Dr. Sriparna Saha expresses gratitude for receiving the Young Faculty Research Fellowship (YFRF) Award. This recognition is supported by the Visvesvaraya Ph.D. Scheme for Electronics and IT under the Ministry of Electronics and Information Technology (MeitY), Government of India, and administered by Digital India Corporation (formerly Media Lab Asia), has significantly contributed to the advancement of the research.

8. References

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*.

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. 2022. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4432–4440.

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

Thomas G Dietterich. 2000. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Zongyuan Ge, Sergey Demyanov, Rajib Chakravorty, Adrian Bowling, and Rahil Garnavi. 2017.

Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 250–258. Springer.

Jaya Gupta, Sunil Pathak, and Gireesh Kumar. 2022. Bare skin image classification using convolution neural netowrk. *International Journal of Emerging Technology and Advanced Engineering*, 12(01).

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 583–592.

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*, pages 4700–4708.

Fatimah Zara Javaid, Jonathan Brenton, Li Guo, and Maria F Cordeiro. 2016. Visual and ocular manifestations of alzheimer's disease and their use as biomarkers for diagnosis and progression. *Frontiers in neurology*, 7:55.

Yicheng Jiang, Bensheng Qiu, Chunsheng Xu, and Chuanfu Li. 2017. The research of clinical decision support system based on three-layer knowledge base model. *Journal of healthcare engineering*, 2017.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. 2014. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 13362–13370. AAAI Press.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *CoRR*, abs/2010.07497.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Jacek Lorkowski and Agnieszka Jugowicz. 2020. Shortage of physicians: a critical review. *Medical Research and Innovation*, pages 57–62.

Yoav Mintz and Ronit Brodie. 2019. Introduction to artificial intelligence in medicine. *Minimally Invasive Therapy & Allied Technologies*, 28(2):73–81.

Neil O'Hare and Alan F Smeaton. 2009. Context-aware person identification in personal photo collections. *IEEE Transactions on Multimedia*, 11(2):220–228.

Olabiyi Oluwatobi and Erik Mueller. 2020. Dlgnet: A transformer-based model for dialogue response generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 54–62.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Shirin Salimi, Muireann Irish, David Foxe, John R Hodges, Olivier Piguet, and James R Burrell. 2018. Can visuospatial measures improve the diagnosis of alzheimer's disease? *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:66–74.

Sertan Serte, Ali Serener, and Fadi Al-Turjman. 2022. Deep learning in medical imaging: A brief review. *Transactions on Emerging Telecommunications Technologies*, 33(10):e4080.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari. 2022a. Dr. can see: towards a multi-modal disease diagnosis virtual assistant. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1935–1944.

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

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):1–13.

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)*, pages 201–207.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

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

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1062–1069.

Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li. 2023. Medical dialogue generation via dual flow modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6771–6784.

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*, volume 33, pages 7346–7353.

Shweta Yadav, Asif Ekbal, and Sriparna Saha. 2018. Feature selection for entity extraction from multiple biomedical corpora: A pso-based approach. *Soft Computing*, 22(20):6881–6904.

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*.

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)*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Yu Zhao, Yunxin Li, Yuxiang Wu, Baotian Hu, Qingcai Chen, Xiaolong Wang, Yuxin Ding, and Min Zhang. 2022. Medical dialogue response generation with pivotal information recalling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4763–4771.

Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*.