scientific reports

Check for updates

OPEN

Towards knowledge-infused automated disease diagnosis assistant

Mohit Tomar[1,2], Abhisek Tiwari[1,2] & Sriparna Saha[1]✉

With the advancement of internet communication and telemedicine, people are increasingly turning to the web for various healthcare activities. With an ever-increasing number of diseases and symptoms, diagnosing patients becomes challenging. In this work, we build a diagnosis assistant to assist doctors, which identifies diseases based on patient–doctor interaction. During diagnosis, doctors utilize both symptomatology knowledge and diagnostic experience to identify diseases accurately and efficiently. Inspired by this, we investigate the role of medical knowledge in disease diagnosis through doctor–patient interaction. We propose a two-channel, knowledge-infused, discourse-aware disease diagnosis model (*KI-DDI*), where the first channel encodes patient–doctor communication using a transformer-based encoder, while the other creates an embedding of symptom-disease using a graph attention network (GAT). In the next stage, the conversation and knowledge graph embeddings are infused together and fed to a deep neural network for disease identification. Furthermore, we first develop an empathetic conversational medical corpus comprising conversations between patients and doctors, annotated with intent and symptoms information. The proposed model demonstrates a significant improvement over the existing state-of-the-art models, establishing the crucial roles of (a) a doctor's effort for additional symptom extraction (in addition to patient self-report) and (b) infusing medical knowledge in identifying diseases effectively. Many times, patients also show their medical conditions, which acts as crucial evidence in diagnosis. Therefore, integrating visual sensory information would represent an effective avenue for enhancing the capabilities of diagnostic assistants.

The development of the Internet was primarily aimed at providing global access to information. In the last few years, the Internet has become one of the most popular and reliable platforms for accessing healthcare-related information. A survey by Cohen et al.[1] found that more than 65% of US adults use the Internet for performing several healthcare-related activities. Over the past 5 years, numerous surveys have highlighted an alarming population-to-doctor ratio in different countries, emphasizing the urgent need for improvements in healthcare systems. According to the report of the World Health Organisation (WHO), 2013[2], there is a shortage of 7.2 million health workers globally which can reach 12.9 million in the upcoming decade. With the motivation of assisting doctors and utilizing their time more efficiently, there has been a significant rise in the popularity of artificial intelligence-based virtual assistants and tools for various medical activities, including automatic disease diagnosis. The objective of Automatic Disease Diagnosis (ADD)[3–6] is to support doctors by performing an initial examination of symptoms. It also diagnoses disease from the conversation between the patient and the doctor. First, the user reports their problems and symptoms (called explicit symptoms) in their self-report, and then the agent inquires about additional symptoms (called implicit symptoms) to diagnose the disease. Hence, an automatic disease diagnosis system can be summarised as a system where an agent inquires about symptoms step by step and then can diagnose disease based on implicit and explicit symptoms. Hence, in a healthcare setting that incorporates this system, when a patient visits a doctor, the doctor is provided with comprehensive information about the patient and his/her situation. Some automatic disease diagnoses systems, such as Mayo Clinic, Babylon Healthcare, and GMAN[7], are already deployed, which are being extensively used by both hospitals and end-users.

In online communication with doctors, patients first inform their chief complaints, known as self-reports to doctors. Based on the chief complaint, a doctor is assigned, who conducts a detailed symptom investigation and extracts relevant symptoms through chat. An example is shown in Figure 1. Over the last 5 years, significant efforts have been made by both the dialogue and healthcare communities to develop an artificial

[1]Department of Computer Science and Engineering, Indian Institute of Technology, Patna 801103, India. [2]These authors contributed equally: Mohit Tomar and Abhisek Tiwari. ✉email: sriparna@iitp.ac.in

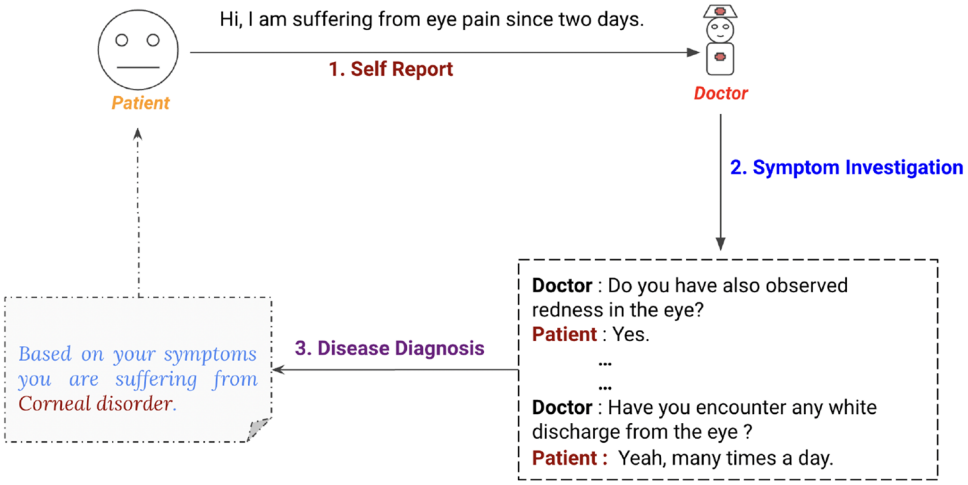

Figure 1. An illustration of online symptom investigation and disease diagnosis.

intelligence-based diagnosis assistant that can act as a third eye for disease diagnosis[8,9]. In the study[3], they introduced a task-oriented dialogue system, which collects patient self-reported information and extracts further signs and symptoms during conversational interactions. In[6], the authors have illustrated the impact of different reward functions utilized to provide feedback to a reinforcement learning-based diagnosis system on diagnosis efficacy. Following the work, the work[5] incorporated a medical department-driven disease diagnosis system, which illustrated superior performance in terms of both quantitative and qualitative metrics. However, most of the diagnosis assistants[5,10] are based on some data-driven approaches, which learn solely from existing and underlying medical corpora. Thus, given the scarcity of publicly available medical corpora, they are likely to result in a model with local knowledge concentrated in the underlying corpus. In real life, doctors also learn from knowledge bases and well-established principles in addition to diagnosis experience.

It is common for us to communicate only our most prominent and urgent health concerns to doctors during consultations. However, doctors do not rely solely on our reported symptoms to make diagnoses and prescribe treatments. Instead, they conduct a thorough investigation to arrive at a conclusive diagnosis. This is necessary because we tend to report only the most common and noticeable symptoms, overlooking other potential clues. They collect additional evidence to better understand medical issues and treat them properly. Motivated by the above two observations, we aim to build knowledge-infused, discourse-aware disease identification (KI-DDI) that incorporates an external knowledge graph and uses an attention mechanism that emphasizes the importance of self-report in the whole conversation. We also create a symptom disease knowledge graph (S–S–D) where symptoms and diseases act as nodes, and an edge between them is treated as a co-occurrence of both of them. We then determine edge weights using the symptom frequency–inverse disease frequency (sf–idf) method, inspired by the term frequency–inverse document frequency (tf–idf) method[11]. The edge weight between a symptom and a disease determines the co-occurrence of the symptom and disease. In KI-DDI, we pass the whole dialog to the language model to extract embedding. We then extract symptoms from the dialog and retrieve the sub-graph from the knowledge graph relevant to the dialog. We then form a joint graph by connecting the dialog node and subgraph. Finally, we obtain graph embedding by considering the mean pool, and then an attention mechanism is used to calculate the weighted sum of the dialog node and self-report, where graph embedding is infused with dialogue embedding to perform disease classification.

In the last few years, there have been tremendous efforts by both research and industrial communities to automatize many medical operations to assist physicians[12]. Nevertheless, the exploration and outcome of these efforts are limited primarily due to the lack of an adequate amount of medical data[13]. For example, there is not a single conversational disease diagnosis dialogue corpus in English. Motivated by the limitation, we curate an empathetic medical dialogue dataset called Empathical Dialogue Dataset. We annotate each utterance of a conversation with its corresponding intent and symptom information. There are two types of intent tags: Symptom and Affirmative. Symptom intent indicates the presence of a symptom, and Affirmative intent indicates that the patient agrees with the doctor, but the symptom is not present in the patient's utterance. The role of empathy in this dataset is that it helps patients feel trusted and cared for by the doctor. The dialogue corpus bridges the following gaps in the medical diagnosis research community: (a) End-to-End communications directly with end-users in the English language, (b) Medical utterance understanding modules could be pre-trained using the curated corpus for symptom extraction, and (c) Context coherent response generation.

Research questions and hypotheses

In this paper, we investigate the following three research questions: **(i)** Are self-reports from patients sufficient for an accurate diagnosis? *We hypothesize that the patient's self-report (first utterance by the patient) alone is insufficient for disease diagnosis. We performed empirical studies that showed that self-report is insufficient in diagnosing disease as the model achieves poor diagnosis accuracy. Thus, it indicates the need for further symptom investigation*. **(ii)** How does the medical knowledge graph influence the disease diagnosis model's performance?

We hypothesize that the external medical knowledge graph aids in disease diagnosis. We showed through empirical studies that our model KI-DDI incorporating external knowledge outperformed the baseline models without external knowledge. This shows external knowledge provides valuable insights in diagnosing disease. **(iii)** Does the mechanism of knowledge infusion impact the efficacy of disease diagnosis? *We hypothesize that incorporating knowledge using graph structure is an efficient mechanism of infusing them. We showed through empirical studies that when external knowledge is infused in a graph structure, it performs better than when it is infused in a linear structure. Thus, adding external knowledge in a graph structure helps diagnose disease.*

Key contributions

The key contributions of the work are threefold, which are enumerated as follows:

- We propose a two-channeled knowledge-infused, discourse-aware disease identification (KI-DDI) model that leverages external medical knowledge encoded through a context-aware filtered knowledge graph for identifying diseases accurately and efficiently from patient–doctor communications.
- We first curate a conversational medical dialogue corpus named Empathical dialogue dataset in English, where each utterance is annotated with its corresponding intent and slot information.
- The proposed KI-DDI model achieves a significant improvement over an existing state-of-the-art model and establishes a new benchmark for the conversation-driven diagnosis problem.

Background

Related works

The research primarily pertains to the following areas: electronic health records, automatic disease diagnosis, graph neural network, knowledge infusion, and Dynamic Uncertain Causality Graph. In the subsequent paragraphs, we provide summaries of the pertinent works in these domains.

Electronic health records (EHR)

During the early 2000s, systems based on Electronic Health Records (EHR) were introduced with the goal of aiding patients, driven by the motivation to provide virtual assistance to individuals in rural areas[14] In BEHRT[15], the authors developed a transformer-based model for mining electronic health records (EHR). It uses patients' EHR data to perform multi-label classification for given all possible diseases. It is also capable of personalized recommendations, and it can incorporate concepts such as diagnosis, medication, and measurements. In[16] authors proposed a reinforcement learning algorithm based on EHR to optimize the sequential processing of diseases. It considers both physiological variables and major disease factors during EHR modeling to improve the interpretability of the model. It utilizes Deep Q Learning (DQN)[17] algorithm to explore the optimal insulin dosage for the patients. In[18], authors handled the problem of Generalized Anxiety Disorder (GAD) and Major Depressive Disorder (MAD) using an ensemble of machine learning pipelines (Support Vector Machine, XG Boost, K Nearest Neighbor, Random Forest, Logistic Regression, Neural Network). It also utilized SHAP values to highlight which features had the major impact on the prediction for each disease. In Med-BERT[19], authors adapt BERT[20] in the EHR setting. As an input, it receives three types of embeddings: the diagnosis code, the order of code within each visit, and the position of each visit. It achieved remarkable performance when fine-tuned on EHR. In Med7[21], authors introduced a named-entity recognition model that is trained on EHR. The goal of the model is to recognize seven categories such as drug names, route of administration, frequency, dosage, strength, form, and duration.

Automatic disease diagnosis

The utilization of an Electronic Health Record (EHR) system necessitates the coordination and synchronization of multiple devices[22]. To streamline the process, researchers have introduced a novel technique for automatically diagnosing non-fatal or sensitive diseases. In this approach, an interactive system conducts symptom investigation and provides disease diagnoses[23]. Wei et al.[3] devised a task-oriented dialogue procedure for symptom investigation. In this process, the agent gathers symptoms through conversation and subsequently diagnoses a disease based on the observed symptoms. In[10], the authors presented a context-aware symptom checker, which also models patients' personal information, such as gender and age, in disease diagnosis. The experimental results of the contextual model confirm the vital role of patients' personal information in executing an appropriate and efficient diagnosis. Liao et al. (Liao 2020 Task) have proposed an integrated and synchronized two-level policy framework using hierarchical reinforcement learning[24]. The model demonstrated superior performance compared to the flat policy approach[3] by a considerable margin. In[25], authors considered disease diagnosis as a generation process. It uses a symptom attention framework for the generation of symptoms and diagnosis. It uses an orderless training mechanism.

Graph neural network

Graph Convolutional Network (GCN)[26] uses graph data and updates the node embedding depending upon the neighboring nodes. Graph Attention Network (GAT)[27] uses an attention mechanism to get the embedding of nodes depending on which neighboring node is relevant. In Graph Transformer[28], authors proposed the adaptation of the transformer network to graphs. It uses an attention mechanism that depends on a neighboring connection for each node. In[29], authors present a systematic approach to building a scalable graph transformer. Its time complexity is linear in the number of nodes and edges. In[30], authors used a modified Markov decision kernel to derive Simple Spectral Graph Convolution. Upon using this method, it trades off between low and high

pass filter bands, which capture global and local context for each node. In[31], authors found that using reversible connections with deep networks allows the effective training of an overparameterized graph neural network. In[32], authors showed the limitation of using static attention in a graph attention network (GAT). They further developed the dynamic attention mechanism, which attends dynamically to neighboring nodes depending on the query node, to overcome the limitation.

Knowledge graph and knowledge infusion
Numerous studies have been conducted to integrate external knowledge into the language model. ERNIE[33] adds external knowledge by infusion of token embedding and entity embedding. It has an information fusion layer that mixes token embedding and its corresponding entity embedding. In[34], it retrieves the subgraph based on the entities. Further, it forms the joint graph of language embedding and subgraph and applies Graph Neural Network (GNN) for knowledge infusion. GreaseLM[35] focuses on the deep fusion of embeddings from the language model and the graph neural network using a modality interaction unit over multiple layers. In[36] authors utilized self-supervised learning methods such as masked language modeling and knowledge graph link prediction for learning joint representation of text and knowledge graph. In[37] authors studied the capabilities of the multimodal BERT model in storing the grammatical and linguistic knowledge that is learned with the help of objects in images. In[38], authors developed the method of knowledge prompting in which they first extracted knowledge from a language model, and later they used that knowledge in question-answering tasks.

Dynamic uncertain causality graph (DUCG)
has been used for the purpose of the clinical diagnosis. DUCG[39] has been utilized to diagnose vertigo by incorporating symptoms, signs, medical histories, etiology, and pathogenesis. Also, Cubic DUCG[40] has been used for fault diagnosis for complex systems by representing dynamic casualties in the system fault spreading process in a compact manner and conducting accurate reasoning. Also, in the context of diagnosing and treating Hepatitis B., DUCG[41] based diagnosis and Treatment Unification Model is utilized. It uses Reverse logic gates to enhance the accuracy of treatment planning.

Problem formulation

The proposed model aims to identify the disease of the patient based on patient–doctor interaction. Thus, the input to an autonomous system will be dialogue, and the output will be disease. A dialogue can be regarded as sequences of patient and doctor utterances, i.e., $D =< (P_1, D_1)(P_2, D_2) \cdots\cdots (P_n, D_n) >$ where (P_i, D_i) denote ith utterance of patient and doctor, respectively, and n signifies the total number of turns in the dialogue. The disease identification through patient–doctor dialogue can be expressed as follows:

$$d = argmax_j P(Dis_j | \{(P_1, D_1)(P_2, D_2) \cdots\cdots (P_n, D_n)\}, \theta) \quad (1)$$

where Dis is the set of diseases, the term, θ denotes the diagnosis model's parameter.

Dataset

We begin by investigating the benchmark medical diagnosis dialogue datasets, and the findings are presented in Table 1. We could not find a single dyadic conversational diagnosis dataset in English, which motivated us to curate a new medical dialogue corpus. Doctors usually engage with and respond empathetically to their patients, which increases patient compliance and further helps in building trust between patient and doctor. We developed an Empathetic Medical (Empathical) Dialogue dataset with the help of the benchmarked SD[42] dataset and clinical guidelines provided by medical experts.

Empathical dataset creation and annotation

During our investigation into the benchmarked conversation dataset, we found the SD dataset[42], which has a database of 30K diagnosis cases covering over 90 diseases and 266 symptoms. We considered the SD dataset as a reference for creating the new conversational dataset because of its variety and credibility. We sampled 100 random diagnosis examples from the SD dataset. With the help of two clinicians, we formed a conversation-based sample dataset corresponding to the 100 diagnosis cases and annotated it with its intent and symptom information. Then we employed three medical students for the creation and annotation of dialogues based on the SD

Dataset	Language	Conversation	Intent	Symptom
RD[3]	Chinese	×	×	×
DX[43]	Chinese	✓	×	✓
M^2-MedDialogue[44]	Chinese	✓	×	✓
MedDialog-EN[45]	English	×	×	×
MedDG[46]	Chinese	✓	×	✓
SD[42]	English	×	×	×
Empathical (ours)	English	✓	✓	✓

Table 1. Comparison of the existing medical datasets for diagnosing disease.

dataset samples. The students created a large dialogue corpus of 1367 diagnosis conversations by following the sample dataset and the detailed guidelines with the curated sample dataset. In order to measure the annotation agreement among the annotators, we calculated the Fleiss kappa[47], which was 0.76, indicating a strong agreement among annotators. The dataset statistics are reported in Table 2.

Clinical and ethical guidelines

As the medical field is highly sensitive and specialized, clinical validity holds paramount importance. We have strictly followed the guidelines established for legal, ethical, and regulatory standards in medical research during the dataset curation process. The key guidelines provided to the annotators are as follows: (i) An annotator should not add or remove any entity in a conversation corresponding to the reported diagnosis sample in the benchmark SD dataset. (ii) No individual's personal information, which might disclose their identity, should be present in any statement within a dialogue or the entire dialogue itself. (iii) Any personal or sensitive information shared in the conversations should be properly de-identified to protect the privacy of individuals. (iv) The use of profanity or offensive language is strictly prohibited in conversations. (v) It is important to use the correct medical terminology in the transcript to ensure that the information is understood correctly by healthcare professionals. (vi) If the intent of a counseling talk is unclear, mark it, and it will be examined and confirmed by a medical professional. Furthermore, the created corpus by the annotators is thoroughly checked and verified by the clinicians. We have also obtained approval from our institute's ethical committee, IIT Patna to employ the dataset and carry out the research (IITP/EC/2022-23/07).

Role of dyadic conversation and intent/symptom annotation

Natural language understanding (NLU) is the first stage of a conversation system which aims to recognize users' intentions (intent) and key information from their utterances. In order to make a disease diagnosis system that can be used for communicating directly with humans in language, NLU is necessary. Thus, we first curate the dyadic corpus and train the NLU module with the corpus. Here we have two kinds of intents (a) *Symptom*, which means the presence of a symptom and (b) *Affirmative* which means the patient is agreeing with a doctor, but there is no mention of the symptom in the patient's utterance.

Purpose of intent and symptom information

Identifying intent and slot are two key tasks in NLU, which are vital for communicating with humans effectively. So, for building the NLU module, we have tagged intent and slot (here symptom) information for every utterance by the user (Figure 2).

Role of empathy

Patients' comfort and user satisfaction are of the utmost importance during doctor–patient consultations. This helps build trust between patient and doctor and increases patient compliance. Moreover, patients' recovery rate gets better when they connect with a doctor on common grounds, which boosts their mental well-being.

Methodology

We proposed a two-stage discourse-aware disease diagnosis framework; the two stages are (a) symptom investigation encoding and (b) external relevant knowledge infusion. The proposed architecture is illustrated in Figure 3. The rationale behind the model is that for obtaining dialog and self-report embedding, we pass through the transformer encoder, i.e., SapBERT, and to diagnose the disease properly, we take help from external knowledge. To represent this external knowledge, we identify which diseases are more commonly linked with symptoms in the conversation and form a knowledge subgraph between symptoms and diseases. Then, to identify which symptom and its associated disease is more important in diagnosing disease, we form a joint graph between dialog embedding, symptoms, and diseases and apply Graph Attention Network. Finally, we use joint graph embedding to attend to dialog and self-report embedding to determine which is more critical for diagnosing, and then we diagnose the patient's disease. External knowledge helps aid clinicians by providing relevant information on which diseases are more closely linked with a particular symptom and providing knowledge expansion.

The model is comprised of three parts: **(i)** Symptom Investigation Encoding: Dialog and Self-Report Encoder, which generates the embedding for complete dialog between doctor and patient and also embeds the self-report given by the patient. Self-report signifies patients' chief complaints/major difficulties expressed by themselves.

Attribute	Value
No. of dialogues	1367
No. of utterances	8962
Utterance tags	Intent and symptom
Avg. dialogue length	6.56
intent tags	Symptom, affirmative
No. of diseases	90
No. of symptoms	228

Table 2. Statistics of Empathical Dataset.

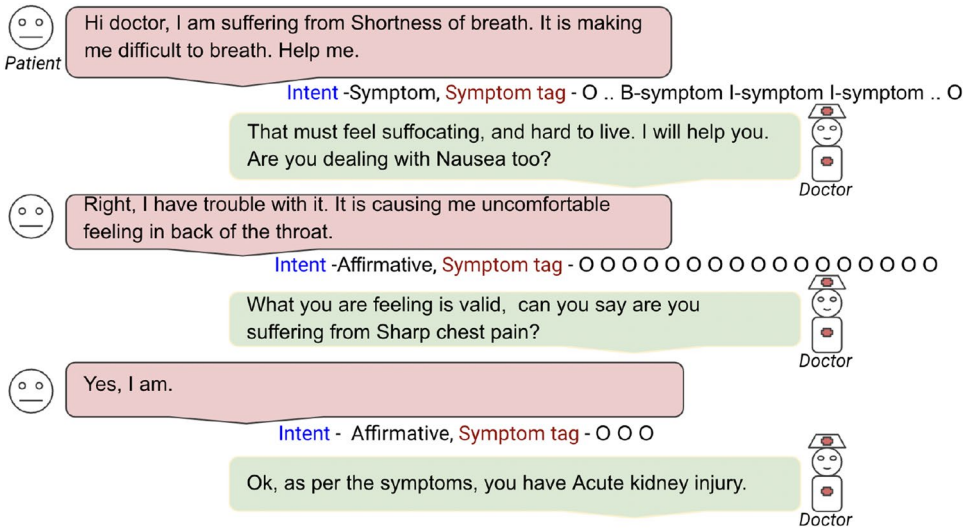

Figure 2. A dialogue sample from the curated Empathical dataset. Conversation between patient and doctor having symptom and intent tagged.

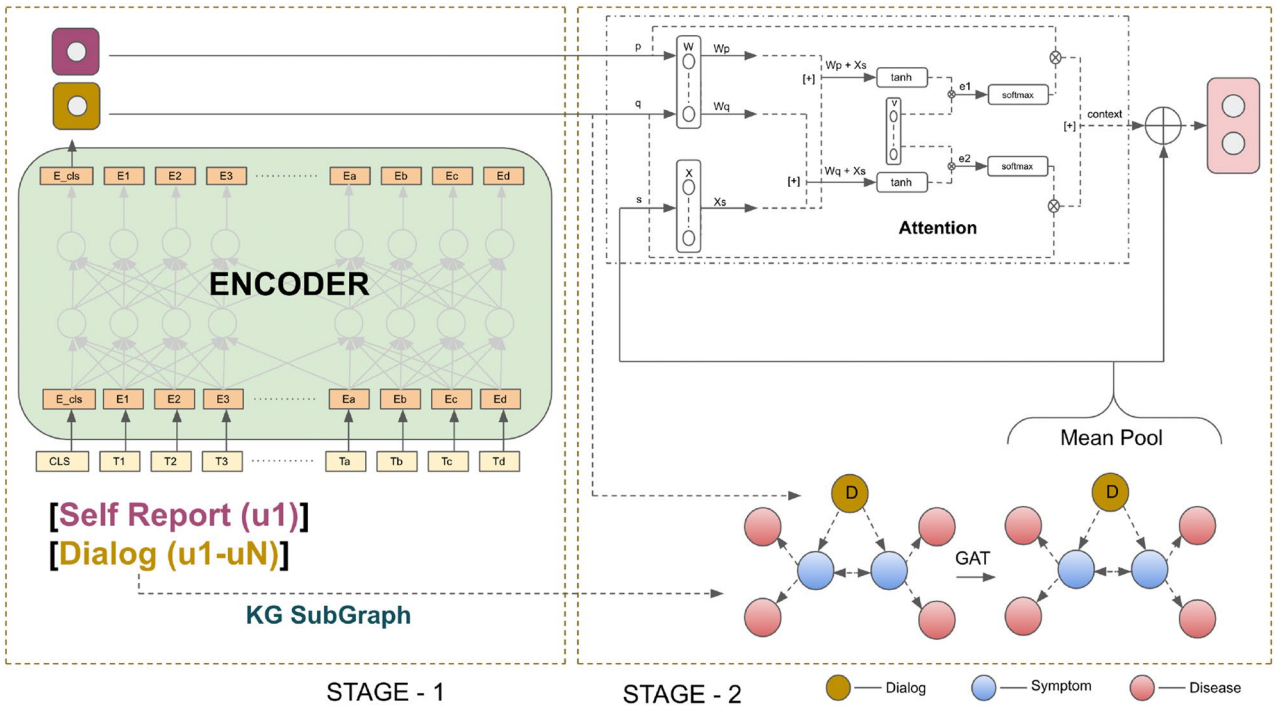

Figure 3. KI-DDI: Self Report and Dialog are passed through the language model to obtain their embedding. The blue nodes are symptoms and Red nodes are diseases linked to symptoms. A joint Graph is formed by connecting the dialog node to all symptom nodes.

(ii) Knowledge Infusion: Knowledge Graph Extraction for extracting relevant sub-graphs from the knowledge graph to emphasize information relevant to the context. **(iii)** Disease Diagnosis Network. We have discussed and demonstrated the working principle of each module in the following sections.

Symptom investigation encoding

Symptom investigation is the foundational and essential component of disease diagnosis. Patients first report their chief complaints; doctors conduct a thorough investigation and diagnose accordingly. Thus, encoding the investigation report efficiently is critical to the autonomous disease diagnosis model. Usually, doctors diagnose a disease based on the set of symptoms experienced by patients; however, they prioritize a few symptoms, particularly the patient's self-reported symptoms, in diagnosis. Thus, we segregate self-reported symptoms from

the other extracted symptoms and infuse a weighted vector of this information into the diagnosis prediction model. To encode patient self-report and dialog, we have utilized SapBERT[48] to capture the semantic meaning of patient–doctor utterances (Figure 3). We have utilized sr_{strat}, sr_{end} for denoting the self-reports start and end, respectively. Two more special tokens, *pat*, and *doc* have been used to signify the starting position of patient and doctor utterances, respectively. We use SapBERT to get contextualized encoded representations (S and C) from the vectors (Equations 2 and 3).

$$S = LM(|sr_{start}|SR|sr_{end}|) \tag{2}$$

$$C = LM(|pat|Us_{1:t}|doc|Do_{1:t}) \tag{3}$$

where Us_i and Do_i denote ith user utterance and doctor utterance, and LM is the notion for the utilized language model (SapBERT).

Knowledge infusion

Clinical knowledge helps clinicians narrow the investigation space and use the information gathered efficiently during the diagnostic process. Thus, we aim to infuse the knowledge structure in the disease diagnosis framework.

Knowledge graph construction

Here, we first created the knowledge graph (S-S-D) from the Empathical dataset, where symptoms (S) and diseases (D) 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[11] which involves applying the technique term frequency–inverse document frequency (TF–IDF) in symptom disease settings. Here, symptom frequency is equivalent to term frequency, and inverse disease frequency is equivalent to inverse document frequency. The edge weights between symptom–disease e(s, d, D) and symptom–symptom e(s_i, s_j) are computed as follows:

$$e(s, d, D) = sf(s, d) * idf(s, D) \tag{4}$$

$$sf(s, d) = \frac{n_{sd}}{\sum_k n_{kd}} \tag{5}$$

Here, n_{sd} is the number of cases where symptom (s) has occurred with the disease, d. k ranges in symptom space. The term $sf(s, d)$ represents the raw count of the co-occurrence of a symptom s with disease d divided by the co-occurrence of every symptom with disease d.

$$idf(s, D) = \log \frac{|D|}{|d : s \in disease_j|} \tag{6}$$

Here $|D|$ signifies the total number of diseases. The term $idf(s, D)$ represents the logarithmic fraction of diseases containing the symptom s obtained by dividing the total number of diseases by the number of diseases having symptom s and then taking the logarithm of that quotient.

$$e(s_i, s_j) = \frac{n(s_i, s_j)}{\sum_k n(s_i, s_k)} \tag{7}$$

The term $e(s_i, s_j)$ represents the number of times symptoms s_i and s_j occur together, divided by the co-occurrence of symptom s_i with all other symptoms. The intuition behind the symptom frequency–inverse disease frequency (sf–idf) is that the weight of symptom disease depends on the factor that if a symptom occurs with a particular disease and it also co-occurs with a large number of diseases, its inverse disease frequency will be close to zero (the denominator in "idf" will be closer to numerator) so the weight of that symptom and disease will be lower (since weight is product of symptom frequency(sf) and inverse disease frequency (idf)), indicating that the symptom is loosely associated with the disease. If a symptom occurs with a particular disease and it also co-occurs with a very small number of diseases, then inverse disease frequency will be large (the denominator in "idf" will be much smaller than the numerator) so the weight of that symptom disease will be much higher, indicating that symptom is closely associated with that particular disease.

Knowledge distillation

While knowledge is crucial, focusing on relevant knowledge is more significant while solving a task. Thus, infusing the entire medical knowledge with the proposed disease diagnosis setup would be ineffective and may even deteriorate the performance. Thus, the proposed model extracts a subset of the knowledge graph depending on context (patients' symptoms) dynamically. It first extracts medical entities (signs and symptoms) from the conversation using joint BERT[49] language model and filters the knowledge graph considering the top K associated diseases of the symptoms present in the conversation. We experimented with various K values (1, 2, 3).

Initialization: KG_distill = {}
Input: Current Knowledge Graph - KG_distill, Complete Knowledge Graph - KG_original, Dialog - dialog, Disease list - disease_list, Symptoms extraction from text - SymptomExtractor.
Output: Filtered Knowledge Graph (KG_distill)

```
symptom_list = []
for symptom in SymptomExtractor(dialog) do
    symptom_list.append(symptom)
end for
for i in len(symptom_list) do
    for j in len(symptom_list) do
        if i != j then
            s_i = symptom_list[i], s_j = symptom_list[j]
            symp_symp_weight_1 = KG_original[s_i][s_j]
            symp_symp_weight_2 = KG_original[s_j][s_i]
            KG_distill.append({s_i, s_j, symp_symp_weight_1})
            KG_distill.append({s_j, s_i, symp_symp_weight_2})
        end if
    end for
end for
for i in len(symptom_list) do
    s_i = symptom_list[i]
    for j in KG_original[s_i] do
        disease_count = 0
        if j in disease_list then
            d_j = j
            symp_dis_weight = KG_original[s_i][d_j]
            KG_distill.append({s_i, d_j, symp_dis_weight})
            disease_count += 1
            if disease_count == 3 then
                break
            end if
        end if
    end for
end for
return KG_distill
```

Algorithm 1. Discourse-aware selective filtering (DSF)

Graph attention network and knowledge infusion
We always prefer to analyze structured data, which helps us summarize effectively and take action path accordingly. Similar behavior has been observed for autonomous models, and thus, graph-based models are gaining huge popularity for developing models with a considerable amount of data[50]. Motivated by efficacy, we build a graph attention (GAT) network over the relevant knowledge graph (S–D) and infuse it with context for disease diagnosis. In GAT, the vertex *i* of l-th layer can be described by the following equation

$$h_i^{(l)} = LeakyReLU\left(\sum_{j\in N_i} \alpha_{ij} W_h h_j^{(l-1)}\right) \tag{8}$$

where N_i is the first hop neighbour of vertex i, $W_h \in \mathbb{R}^{d_1' \times d_1}$ is trainable parameter. The attention weight α_{ij} for vertex *i* is calculated as follows:

$$\alpha_{ij} = \frac{exp(LeakyReLU(a^{\mathsf{T}}[W_h h_i || W_h h_j])}{\sum_{k\in N_i} exp(LeakyReLU(a^{\mathsf{T}}[W_h h_i || W_h h_k])} \tag{9}$$

where $a \in \mathbb{R}^{2d_1'}$ is a trainable parameter. Here || means concatenation. Finally, we obtain graph embedding by taking the mean pool of embedding of each vertex in JointGraph.

$$s = MeanPool(h_v^{(L)} | v \in JointGraph) \in \mathbb{R}^{d_1} \tag{10}$$

where *L* denotes the last layer of GAT. *MeanPool* is the average of node features across node dimensions. *JointGraph* means graph obtained after adding dialog node to knowledge subgraph. The obtained global mean pool from the graph is passed to the attention layer.

Attention layer
In some cases, patient-reported data is crucial to understanding disease, while in other cases, symptoms extracted by physicians are crucial. We use additive attention[51] to compute attention. Here, the output of the GAT acts as a query, and self-report encoding and encoded dialog act as values. We take the weighted average of self-report encoding and dialog encoding and concatenate that with GAT output to finally pass it through the linear layer to perform disease classification. It can be expressed as follows:

$$e_i = v^{\mathsf{T}} tanh(W_1 h_i + W_2 s) \tag{11}$$

Here query $s \in \mathbb{R}^{d_1}$ represents GAT output, $i \in \{1, 2\}$ where values $h_1 \in \mathbb{R}^{d_2}$ means self report encoding and $h_2 \in \mathbb{R}^{d_2}$ means dialog encoding. Here $W_1 \in \mathbb{R}^{d_3 \times d_2}$, $W_2 \in \mathbb{R}^{d_3 \times d_1}$ and $v \in \mathbb{R}^{d_3}$ are learnable parameters. Here $e_i \in \mathbb{R}$, $e = [e_1; e_2] \in \mathbb{R}^2$. The attention value α and final context is determined as follows:

$$\alpha = softmax(e) \in \mathbb{R}^2 \tag{12}$$

$$context = \sum_{i=1}^{2} \alpha_i h_i \in \mathbb{R}^{d_2} \tag{13}$$

Finally, the attended context is passed to the disease diagnosis network for disease prediction.

Disease diagnosis network

We hypothesized that only patient self-report is not enough for disease diagnosis; we also need to consider doctor–patient interaction and additional medical knowledge for diagnosing patients effectively. Thus, our prediction network leverages all three components. We utilize self-report, doctor–patient interaction, and medical knowledge (joint graph) and pass the concatenation of attended vector (patient self-report, patient–doctor utterances, and knowledge graph) from the previous stage and joint graph embedding to a fully connected feed-forward network.

$$h_f = \sigma(W[s; context] + b) \tag{14}$$

Here σ is the softmax activation function. $W \in \mathbb{R}^{n \times (d_1 + d_2)}$ and $b \in \mathbb{R}^n$. n is the number of diseases.

$$\hat{y} = argmax_i P(D_i | h_f) \tag{15}$$

where i ranges over the set of diseases. We have utilized categorical cross entropy for calculating loss, which can be expressed as below.

$$L = -\sum_{i=1}^{m} \sum_{j=1}^{n} y_j^{(i)} \log(\hat{y}_j^{(i)}) \in \mathbb{R} \tag{16}$$

where m is the number of training examples, n is the number of diseases. Here, $y_j^{(i)}$ is the ground truth label, and $\hat{y}_j^{(i)}$ is the predicted disease distribution label for ith dialogue.

Experimental setup

We have used the curated Empathical dataset for training and evaluating the proposed model. We divided the dataset as follows: 70% training, 10% validation, and 20% testing. We have utilized the PyTorch framework for implementing the proposed discourse-aware disease diagnosis model. We use SapBERT[48] for encoding the dialog. In Table 3, we have listed the final values of hyperparameters. These values have been chosen through empirical experimentation using the validation dataset. The dataset we use is in English and created based on a benchmarked medical database SD Dataset[42]. The proposed model has been trained, validated, and tested with the dataset. The model works for English; however, it can be adapted to another language with minimal change, such as multi-lingual tokenizer incorporation. We use the BERT tokenizer, capable of processing slang words based on its pre-trained vocabulary, which includes a mix of formal and informal language from diverse sources. If a slang word is present in the vocabulary, BERT tokenizes it like any other word; otherwise, it may employ subword tokenization for out-of-vocabulary terms. Furthermore, the model's ability to handle slang depends on

Hyperparameters	Selected values
Max sequence length	512
Batch size	16
GAT layers	2
GAT hidden dim	384
GAT attention heads	3
GAT dropout	0.5
Attention layer hidden dim 1	768
Attention layer hidden dim 2	384
Attention layer projection dim	64
Optimizer	Adam
Loss function	CrossEntropyLoss
Learning rate	1e−3
Epochs	25

Table 3. Different hyperparameters and their values.

exposure to such terms during pre-training. While it excels with common slang, it may struggle with more niche or emerging expressions. Furthermore, the dataset and code are available at https://github.com/NLP-RL/KI-DDI.

Ethical consideration

While creating the dataset, we followed guidelines aligned with medical research's legal, ethical, and regulatory standards. We utilized a benchmarked dataset named SD (Zhong et al.[42]) to construct a conversational corpus. It's important to note that the dataset includes samples with user consent. With this in mind, we have not added or removed any entity in a conversation corresponding to the reported dialogues in the benchmark SD dataset. Also, the curated dataset does not disclose users' personal information. Hence, we ensure that the Empathical dataset and each step of its formation do not violate ethical and clinical principles. We have also obtained approval from our institute's ethical committee, IIT Patna, to employ the dataset and carry out the research (IITP/EC/2022-23/07). Please note that the research does not involve any human beings or living entities.

Informed consent and privacy

We utilized a benchmarked dataset named SD (Zhong et al.[42]) to construct a conversational corpus. It's important to note that the dataset includes samples with user consent. They do not include any personal patient information, such as names, ages, or genders. Instead, they solely contain information about symptoms discussed during conversations with doctors and the identified diseases by the doctors.

Societal ramifications

Over the last 5 years, numerous surveys and reports have consistently highlighted an imbalanced doctor-to-population ratio. These findings strongly advocate for addressing the concerning statistics by augmenting the healthcare workforce and optimizing their time more effectively. With the objective of aiding doctors and streamlining early diagnosis, the suggested automated disease diagnosis assistant plays a pivotal role in assisting healthcare professionals in precisely identifying illnesses. The research delves into the impact of knowledge infusion on disease identification through doctor–patient conversations. Rigorous experiments and human analyses across diverse algorithms underscore the substantial influence of knowledge infusion in deducing diseases.

Reproducibility

We have used the curated Empathical dataset for training and evaluating the proposed model. We divided the dataset as follows: 70% training, 10% validation, and 20% testing. We have utilized the PyTorch framework for implementing the proposed discourse-aware disease diagnosis model. We use SapBERT (Liu et al.[38]) for encoding the dialog. In Table below, we have listed the final values of hyperparameters. These values have been chosen through empirical experimentation using the validation dataset. The dataset we use is in English and created based on a benchmarked medical database SD Dataset. The proposed model has been trained, validated, and tested with the dataset. The model works for English; however, it can be adapted to another language with minimal change, such as multi-lingual tokenizer incorporation. We have provided details of our experimental setup, including hyperparameter values and evaluation metrics, and made our code available (https://github.com/NLP-RL/KI-DDI).

Accession codes

We have made a GitHub repository that contains the curated conversational dataset and the experimental setup (code). The dataset and code are available at https://github.com/NLP-RL/KI-DDI.

Result and discussion

In order to comprehend the efficacy and limitations of the proposed model, we compared it with the following baselines and state-of-the-art models. The baselines and state-of-the-art models are as follows:

- *BioLinkBert*[52]—It is the pretraining method that uses links between different documents to train BERT. BioLinkBert is pretrained on PubMed articles with citation links on two self-supervised tasks masked language modeling and document relation prediction.
- *KrissBert*[53]—It trains PubMedBERT using entity list to generate self-supervised mention examples of biomedical entities and further it uses contrastive learning for training.
- *KI-CD*[54]—It has a potential candidate (PCM) module which is based on Bayesian learning for symptom investigation. Also, it uses Hierarchical Reinforcement Learning for diagnosing disease.
- *Coder*[55]—It uses contrastive learning along with Unified Medical Language System (UMLS) knowledge graph to produce the embedding for medical terms.
- *SapBert*[48]—It trains the language model in a way that uses hard positive and hard negative samples to align synonymous biomedical entities. It uses a UMLS knowledge graph.

Evaluation metrics

We utilize the most popular classification evaluation metrics, namely accuracy, F1-Score, and Jaccard similarity for evaluating the performance of different diagnosis models. The different metrics are defined as follows:

- *Accuracy* It is defined as the number of correct predictions divided by the total number of predictions. It is represented as follows:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{17}$$

For a binary classification task, it can also be expressed as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{18}$$

Where TP—True Positive, TN—True Negative, FP—False Positive, FN—False Negative.

True Positive (TP)—Number of examples predicted to be positive by the machine learning model and its label is actually positive.

True Negative (TN)—Number of examples predicted to be negative by the machine learning model and its label is actually negative.

False Positive (FP)—Number of examples predicted to be positive by the machine learning model and its label is actually negative.

False Negative (FN)—Number of examples predicted to be negative by the machine learning model and its label is actually positive.

- *F1 score* is the harmonic mean of the precision and recall.

$$F1 = \frac{2 * (precision * recall)}{(precision + recall)} \tag{19}$$

Precision—It indicates the proportion of positive predictions that are actually correct. It is given as the ratio of True positive divided by the sum of True Positive and False Positive.

$$Precision = \frac{(TP)}{(TP + FP)} \tag{20}$$

Recall—It indicates the proportion of actual positives that are identified correctly. It is given as the ratio of True Positive divided by the sum of True Positive and False Negative.

$$Recall = \frac{(TP)}{(TP + FN)} \tag{21}$$

- *Jaccard Similarity* is defined as the size of the intersection divided by the size of the union of two label sets. It compares predicted labels to the ground truth labels for a sample. For ground truth label set "a" and predicted label set "b", it is given as:

$$Jaccard(a, b) = \frac{|a \cap b|}{|a \cup b|} \tag{22}$$

All the reported values in the following tables are statistically significant, which are validated using the statistical *t*-test[56] at a significant level of 5%. The obtained performance by the joint BERT natural language understanding model for intent and symptom identification is provided in Table 4. With the conducted experiments and performance comparison with the state-of-art/baseline models, the raised research questions (RQs) can be answered as follows.

RQ1: Are self-reports from patients sufficient for accurate diagnosis?

Table 5 shows the efficacy of models that utilize only patient self-report for diagnosing a disease. The model that considers both self-reports and symptoms extracted by clinicians is way superior in terms of diagnostic accuracy. It firmly establishes the importance of the detailed symptom investigation conducted by clinicians. It is primarily due to the inadequacy of self-reports to accurately recognize patient diseases. It's also obvious because most of the time, we report symptoms that used to be common across several diseases, such as cold, cough, and fever. It shows the doctor needs to further investigate symptoms in addition to patient self-reports. Hence, the answer is no; we need further symptom investigation (in addition to patient self-report) to diagnose accurately.

RQ2: How does the medical knowledge graph influence the disease diagnosis model's performance?

The performance obtained by the state-of-the-art model and our proposed knowledge-infused disease diagnosis models are reported in Table 6. It shows that KI-DDI improved the performance of disease diagnosis by a

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	Jaccard
SRE + Linear	23.80	0.183	0.122
Knowledge	26.49	0.197	0.135
SRE + Knowledge_1	24.78	0.198	0.136
SRE + Knowledge_2	24.90	0.201	0.140
SRE + Knowledge_3	24.53	0.190	0.131

Table 5. Performance of model using Self Report with Linear and Knowledge. Here Knowledge_2 means every symptom (blue node see Figure 3) has at most two diseases (red node) connected to it.

Model	Accuracy	F1-score	Jaccard
BioLinkBERT[52]	47.25	0.4067	0.3129
KrissBERT[53]	57.14	0.4977	0.4091
KI-CD[54]	57.84	–	–
Coder[55]	59.70	0.5612	0.4630
SapBERT[48]	61.53	0.5801	0.4834
KI-DDI	**64.10**	**0.6035**	**0.5099**

Table 6. Performance of the proposed KI-DDI model. Significant values are in bold.

margin of 2.57% compared to the SapBERT[48] model. Hence, we conclude that the knowledge graph is helpful in improving disease diagnosis accuracy. We also show the Top 1, 3, and 5 disease coverage accuracies for different models in Figure 4.

RQ3: Does the mechanism of knowledge infusion impact the efficacy of disease diagnosis?

In addition to the knowledge being essential for autonomous models, its representation also matters. As humans, we always prefer to have information presented in a structured manner. With this motivation, we investigated the performance of different models having knowledge incorporated with different approaches. The obtained findings are reported in Table 7. It demonstrates that the model that infuses medical knowledge using a graph structure outperformed the model that employs a linear structure by a significantly large margin of 4.76%. Hence,

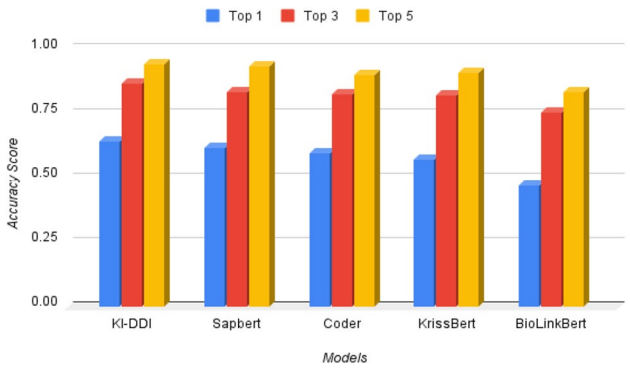

Figure 4. Disease diagnosis accuracy of different models.

Model	Accuracy (%)	F1-score	Jaccard
SRE + Linear	23.80	0.183	0.122
SRE + Knowledge	**24.90**	**0.201**	**0.140**
DE + Linear	58.97	0.5306	0.4331
DE + Knowledge	**63.73** (**4.76** ↑)	**0.5752**	**0.4796**

Table 7. Performance comparison of adding knowledge in self report and dialog. Significant values are in bold.

the answer is yes, knowledge representation matters and the graph-based symptom–disease infusion is more effective in disease diagnosis than infusing it as a linear vector.

Ablation study

We have also conducted an ablation study to comprehend the importance of different components and concepts linked to the proposed model. The obtained findings are summarized in Table 8. It leads to the following evidence and observations: **i.** We see that concatenating dialog encoding with knowledge graph embedding improves the performance. **ii.** We also observed that the behavior of constantly increasing knowledge width does not lead to superior performance, mainly because extraneous and large information sizes are included. **iii.** We see that using an attention mechanism between self-report encoding and dialog encoding leads to improvement.

Analysis

The comprehensive analysis of the performances of different models led to the following major observations: **(i)** We analyze the performance of different models on common test cases and one such case study is shown in Figure 5. Our model correctly diagnosed the disease, while the other models misclassified the disease. This can be attributed to the knowledge infusion mechanism that the model is able to attend to symptoms that are more important for diagnosing the disease. **(ii)** In order to exploit the structure of medical departments in healthcare systems, we also experimented with a hierarchical-based disease classifier. The first layer classifier triggers an appropriate medical department, and the activated disease classifier identifies the disease. The obtained results are reported in Table 9. **(iii)** In the case of hierarchical classification, we observed that the model identifies disease groups/medical departments quite adequately, but it gets confused among the diseases of the same medical group. **(iv)** We report the impact of variation of layers of GAT on the model's performance in Figure 7. We find

Model	Accuracy	F1-score	Jaccard
Knowledge_1	58.60	0.5353	0.4474
Knowledge_2	**60.31**	**0.5479**	**0.4610**
Knowledge_3	58.48	0.5337	0.4446
DE + Knowledge_1	63.36	**0.5909**	**0.4987**
DE + Knowledge_2	62.39	0.5884	0.4903
DE + Knowledge_3	**63.73**	0.5752	0.4796
KI-DDI_1	**64.10**	**0.6035**	**0.5099**
KI-DDI_2	63.61	0.5969	0.5073
KI-DDI_3	63.24	0.5911	0.5007

Table 8. Result of the ablation study, which shows the efficacy of different components of the proposed model. Significant values are in bold.

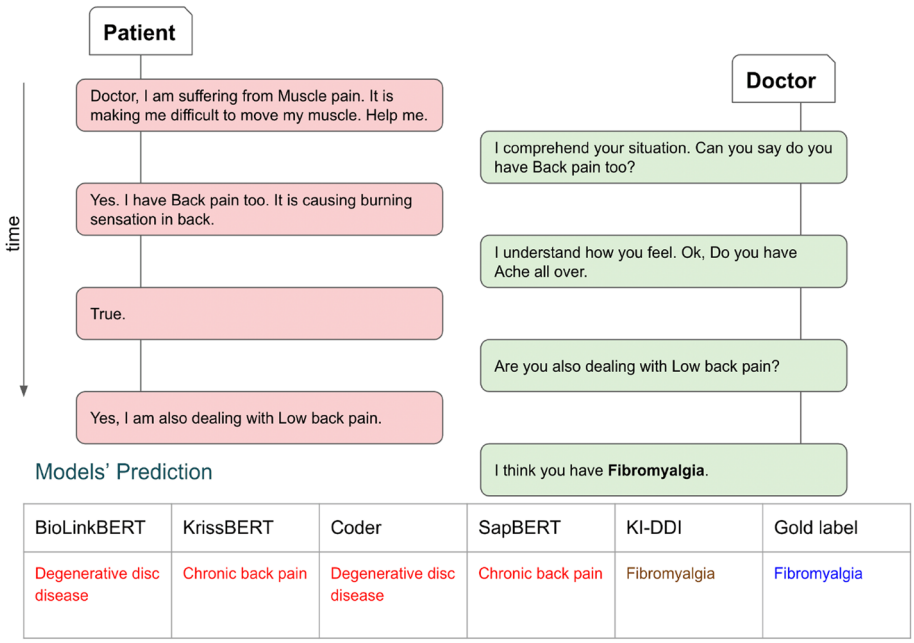

Figure 5. Performance of KI-DDI and other models on a common test case.

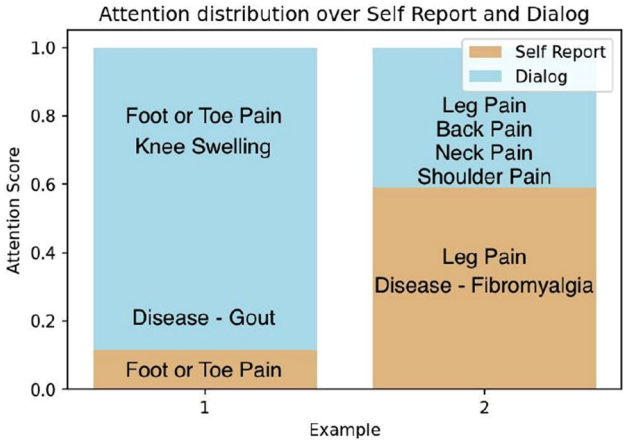

Figure 6. Distribution of attention scores for two test examples.

Model	Group Acc.	Accuracy	F1-score	Jaccard
KI-DDI_1	79.24	58.97	0.555	0.457
KI-DDI_2	**81.19**	**62.27**	**0.583**	**0.487**
KI-DDI_3	80.70	60.07	0.565	0.467

Table 9. Hierarchical classification. Group Acc—Group Classification Accuracy, Acc—Disease classification within that group. Significant values are in bold.

that upon increasing layers up to two model's performance increases then it starts decreasing. **(v)** Sometimes patient self-report is vital to disease, whereas other times symptoms extracted by doctors are critical (Figure 6). We must thus take into account both and make a diagnosis that is appropriate to the situation rather than relying solely on one.

Our model also has some data biases like the majority of deep learning models; however, it is minimal while evaluating its impact. The model is trained on textual data, and some diseases have few examples, so our model is biased toward identifying diseases with many training examples. Also, many symptoms are expressed visually, and our model doesn't integrate multi-modal input. Our model is trained on a single language corpus, i.e., English; its effectiveness is reduced in code-mixed scenarios. Our model has low diagnostic accuracy (64.10%). Therefore, it can give inaccurate diagnoses and shouldn't be used in real-world medical settings. But our model (KI-DDI) performs relatively better in the Top3 (86.8%) and Top5 (94.01%) accuracy in disease diagnosis.

Table 10 shows that Bio Link Bert and KI-CD models take the highest train and inference time. BioLinkBert takes longer because of the bigger model size (having a total parameter count of 333 Million), and KI-CD takes longer train time because its architecture consists of 10 hierarchical models to train and has a longer inference time because it performs symptom investigation and disease diagnosis. In contrast, the remaining model only performs disease diagnosis. Models SapBert, KrissBert, and Coder take approximately the same train and inference time because of nearly the same parameter count (around 109 Million and 7 Thousand trainable parameters). KI-DDI takes longer because it is larger than SapBERT (as it involves SapBERT and Graph Attention Network), having (110 Million total parameters and 622 Thousand trainable parameters) but less time than BioLinkBERT because of its smaller model size.

Model	Training time (for 1 epoch) (s)	Inference time (s)
KI-DDI	22.26	5.00
BioLinkBert	30.34	8.27
SapBert	9.55	2.57
KrissBert	7.91	2.52
Coder	9.35	2.52
KI-CD	30.14	9.44*

Table 10. Train and Inference time comparison of various models. Here, * means the model performs symptom investigation along with disease diagnosis.

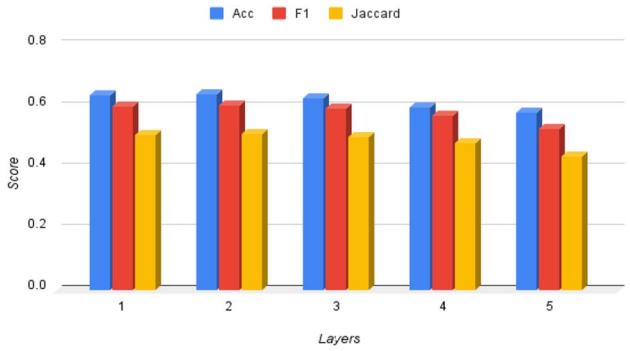

Figure 7. Performance of KI-DDI model upon varying layers.

Scalability Our model KI-DDI utilizes a joint graph by incorporating a knowledge graph subgraph. It requires diseases associated with the symptoms. We experimented with each symptom associated with one to three diseases, and our model is scalable with more than three diseases associated with a symptom.

Reliability Our model achieves an accuracy of 64.10% for disease diagnosis, which is low for diagnosing diseases in real-world settings. Hence, our model is not suitable for practical applications. But our model performs well in the Top3 (86.8%) and Top5 (94.01%) diagnosis accuracy. This shows that our model is getting confused to diagnose diseases linked common symptoms but works well in case of diagnosing diseases in the Top3 and Top5 settings.

Robustness We have tested the robustness of our model concerning the number of layers and diseases linked with the symptoms and provided the results in Figure 7 and Table 8.

Limitations

While the proposed KI-DDI has demonstrated superior performance compared to baseline models, certain limitations have been observed. The key limitations are as follows: (i) The model has been trained and evaluated solely on a single-language corpus, specifically English. It exhibits reduced effectiveness when encountering code-mixed sentences. Therefore, an important avenue for future work is the incorporation of multilingual capabilities. (ii) The model's performance across different diseases is influenced by the frequency of disease samples in the training data. Consequently, it may not perform well when there are very few samples available for certain diseases. Therefore, it is essential to integrate few-shot learning capabilities to address this limitation. (iii) Many symptoms are often conveyed through visual cues, but the model currently operates exclusively with text-based data. In future developments, we aim to integrate a multi-sensory input processing module into the diagnostic assistant.

Conclusion

In this paper, we investigate the importance of knowledge infusion and doctor-driven symptom research in identifying patients' illnesses through dialogue. We presented a two-channel knowledge-infused, discourse-aware disease identification (KI-DDI) model that leverages external medical knowledge encoded through a context-aware filtered knowledge graph for identifying diseases accurately. We first developed a conversational disease diagnosis dataset in English, which is comprised of patient–doctor communication and annotated with semantic information (intent and symptom). The proposed model outperformed baselines and the existing state-of-the-art model significantly across all evaluation metrics. With the rigorous set of experiments conducted, the work evidences the paramount importance of (a) medical knowledge infusion, (b) doctor's collected symptoms (in addition to the patient's self-reported symptom), and (c) structured approaches for the knowledge representation. We note that the model's performance with respect to a specific disease is directly correlated with the quantity of samples available for that disease in the dataset. To mitigate this effect and ensure effective performance even for diseases with limited samples, the inclusion of a few-shot learning module could be considered. 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. Thus, we would like to extend the work by investigating the role of multi-modality in symptom investigation and diagnosis and building a multimodal diagnosis dialogue system.

Received: 11 June 2023; Accepted: 27 January 2024
Published online: 11 June 2024

References

1. Cohen, R. A. & Adams, P. F. Use of the internet for health information: United states, 2009. In *NCHS Data Brief* 1–8 (2011).
2. George, P. P. *et al.* Online elearning for undergraduates in health professions: A systematic review of the impact on knowledge, skills, attitudes and satisfaction. *J. Glob. Health***4** (2014).
3. Wei, Z. *et al.* 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 (2018).

4. Teixeira, M. S., Maran, V. & Dragoni, M. The interplay of a conversational ontology and ai planning for health dialogue management. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* 611–619 (2021).
5. Liao, K. *et al.* Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning. arXiv:2004.14254 (2020).
6. Peng, Y.-S., Tang, K.-F., Lin, H.-T. & Chang, E. Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. *Adv. Neural. Inf. Process. Syst.* **31**, 7322–7331 (2018).
7. Yuan, Q., Chen, J., Lu, C. & Huang, H. The graph-based mutual attentive network for automatic diagnosis. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence* 3393–3399 (2021).
8. Yu, C., Liu, J., Nemati, S. & Yin, G. Reinforcement learning in healthcare: A survey. *ACM Comput. Surv. (CSUR)* **55**, 1–36 (2021).
9. Kumar, Y., Koul, A., Singla, R. & Ijaz, M. F. Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *J. Ambient Intell. Human. Comput.* 1–28 (2022).
10. Kao, H.-C., Tang, K.-F. & Chang, E. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 32 (2018).
11. Ramos, J. *et al.* Using tf–idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning* vol. 242, 29–48 (Citeseer, 2003).
12. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc. J.* **6**, 94 (2019).
13. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: Review, opportunities and challenges. *Brief. Bioinform.* **19**, 1236–1246 (2018).
14. Ventres, W. *et al.* Physicians, patients, and the electronic health record: An ethnographic analysis. *Ann. Fam. Med.* **4**, 124–131 (2006).
15. Li, Y. *et al.* Behrt: Transformer for electronic health records. *Sci. Rep.* **10**, 1–12 (2020).
16. Li, T., Wang, Z., Lu, W., Zhang, Q. & Li, D. Electronic health records based reinforcement learning for treatment optimizing. *Inf. Syst.* **104**, 101878 (2022).
17. Mnih, V. *et al.* Playing Atari with deep reinforcement learning. arXiv:1312.5602 (2013).
18. Nemesure, M. D., Heinz, M. V., Huang, R. & Jacobson, N. C. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci. Rep.* **11**, 1–9 (2021).
19. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, 86 (2021).
20. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018).
21. Kormilitzin, A., Vaci, N., Liu, Q. & Nevado-Holgado, A. Med7: A transferable clinical natural language processing model for electronic health records. *Artif. Intell. Med.* **118**, 102086 (2021).
22. Menachemi, N. & Collum, T. H. Benefits and drawbacks of electronic health record systems. *Risk Manage. Healthc. Policy* **4**, 47 (2011).
23. Tang, K.-F., Kao, H.-C., Chou, C.-N. & Chang, E. Y. Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In *NIPS Workshop on Deep Reinforcement Learning* (2016).
24. Dietterich, T. G. Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.* **13**, 227–303 (2000).
25. Chen, J., Li, D., Chen, Q., Zhou, W. & Liu, X. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 36, 4432–4440 (2022).
26. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907 (2016).
27. Veličković, P. *et al.* Graph attention networks. arXiv:1710.10903 (2017).
28. Dwivedi, V. P. & Bresson, X. A generalization of transformer networks to graphs. arXiv:2012.09699 (2020).
29. Rampášek, L. *et al.* Recipe for a general, powerful, scalable graph transformer. arXiv:2205.12454 (2022).
30. Zhu, H. & Koniusz, P. Simple spectral graph convolution. In *International Conference on Learning Representations* (2021).
31. Li, G., Müller, M., Ghanem, B. & Koltun, V. Training graph neural networks with 1000 layers. In *International Conference on Machine Learning* 6437–6449 (PMLR, 2021).
32. Brody, S., Alon, U. & Yahav, E. How attentive are graph attention networks? arXiv:2105.14491 (2021).
33. Zhang, Z. *et al.* Ernie: Enhanced language representation with informative entities. arXiv:1905.07129 (2019).
34. Yasunaga, M., Ren, H., Bosselut, A., Liang, P. & Leskovec, J. QA-GNN: Reasoning with language models and knowledge graphs for question answering. arXiv:2104.06378 (2021).
35. Zhang, X. *et al.* Greaselm: Graph reasoning enhanced language models for question answering. arXiv:2201.08860 (2022).
36. Yasunaga, M. *et al.* Deep bidirectional language-knowledge graph pretraining. arXiv:2210.09338 (2022).
37. Milewski, V., de Lhoneux, M. & Moens, M.-F. Finding structural knowledge in multimodal-bert. arXiv:2203.09306 (2022).
38. Liu, J. *et al.* Generated knowledge prompting for commonsense reasoning. arXiv:2110.08387 (2021).
39. Dong, C., Wang, Y., Zhang, Q. & Wang, N. The methodology of dynamic uncertain causality graph for intelligent diagnosis of vertigo. *Comput. Methods Programs Biomed.* **113**, 162–174 (2014).
40. Dong, C. & Zhang, Q. The cubic dynamic uncertain causality graph: A methodology for temporal process modeling and diagnostic logic inference. *IEEE Trans. Neural Netw. Learn. Syst. RD* **31**, 4239–4253. https://doi.org/10.1109/TNNLS.2019.2953177 (2020).
41. Deng, N. & Zhang, Q. The application of dynamic uncertain causality graph based diagnosis and treatment unification model in the intelligent diagnosis and treatment of hepatitis B. *Symmetry* **13**, 1185 (2021).
42. Zhong, C. *et al.* Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics* (2022).
43. Xu, L. *et al.* End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 33, 7346–7353 (2019).
44. Yan, G. *et al.* M^2-meddialog: A dataset and benchmarks for multi-domain multi-service medical dialogues. arXiv:2109.00430 (2021).
45. Zeng, G. *et al.* Meddialog: Large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020).
46. Liu, W. *et al.* Meddg: A large-scale medical consultation dataset for building medical dialogue system. *CoRR*arXiv:2010.07497 *(2020)*.
47. Fleiss, J. L., Levin, B. & Paik, M. C. *Statistical Methods for Rates and Proportions* (John Wiley & Sons, 2013).
48. Liu, F., Shareghi, E., Meng, Z., Basaldella, M. & Collier, N. Self-alignment pretraining for biomedical entity representations. arXiv:2010.11784 (2020).
49. Chen, Q., Zhuo, Z. & Wang, W. Bert for joint intent classification and slot filling. arXiv:1902.10909 (2019).
50. Zhang, Z., Cui, P. & Zhu, W. *IEEE Trans. Knowl. Data Eng.* (Deep learning on graphs A survey, 2020).
51. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473 (2014).
52. Yasunaga, M., Leskovec, J. & Liang, P. Linkbert: Pretraining language models with document links. arXiv:2203.15827 (2022).
53. Zhang, S. *et al.* Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP* **2022**, 868–880 (2022).
54. Tiwari, A., Saha, S. & Bhattacharyya, P. A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning. *Knowl. Based Syst.* **242**, 108292 (2022).

55. Yuan, Z. *et al.* Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *J. Biomed. Inform.* **126**, 103983 (2022).
56. Welch, B. L. The generalization of student's' problem when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).

Author contributions

Mohit Tomar (M.T.): Conceptualization, Data analysis, Experimentation, Validation, Analysis, Investigation, Visualization, and Writing; Abhisek Tiwari (A.T.): Conceptualization, Data analysis, Experimentation, Validation, Analysis, Investigation, Visualization, and Writing; Sriparna Saha (S.S.): Conceptualization, Data analysis, Validation, Analysis, Visualization, and Writing. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-53042-y.

Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.