Towards Trustworthy Automatic Diagnosis Systems by Emulating Doctors' Reasoning with Deep Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

To reduce medical doctor's workload and democratize access to medical care, the automation of the *evidence acquisition* and *diagnosis process* has attracted increasing attention recently. However, most works proposed in the machine learning literature focus solely on improving the prediction accuracy of the patient's pathology. We argue that this objective is insufficient to ensure doctors' acceptability of the system. For doctors to trust the system recommendations, they need to understand how the gathered evidences led to the predicted diseases. In particular, interactions between the system and a patient should emulate doctors' reasoning. To do so, we propose to model the evidence acquisition and automatic diagnosis tasks in a deep reinforcement learning framework by considering three essential aspects of doctors' reasoning, namely using *differential diagnosis* with the *exploration-confirmation* approach while prioritizing *severe pathologies*. We propose metrics for evaluating interaction quality concerning these three aspects. We show that our approach performs better than existing models while maintaining competitive prediction accuracy.

1 Introduction

In recent years, the digital healthcare industry has been growing rapidly, benefiting from the advancements in machine learning (Esteva et al., 2019; Xiao et al., 2018). In particular, telemedicine, i.e., healthcare services provided via digital means, has received much attention (Kichloo et al., 2020). Aiming to reduce human doctors' workload and thereby broaden access to care by automating parts of patients' interaction, telemedicine applications typically involve two significant components, among others: *evidence acquisition* and *automatic diagnosis*. In a typical interaction, a patient first presents his/her chief complaint[1] to the telemedicine system, then the system asks further questions to gather more evidences, and finally, it makes a prediction regarding the underlying disease based on all collected evidences. Importantly, a human doctor would typically review the entire interaction, including the evidence acquisition process as well as the prediction made, before eventually making the final diagnosis and providing next instructions to the patient (e.g., additional exams, prescriptions).

There are many existing works in the machine learning literature that aim to improve the automated evidence acquisition and disease diagnosis steps, using Reinforcement Learning (RL) (Kao et al., 2018; Wei et al., 2018; Yuan and Yu, 2021), Bayesian networks (Guan and Baral, 2021; Liu et al., 2022), or supervised approaches (Luo et al., 2020; Chen et al., 2021). Regardless of the approach, its success is commonly measured by the accuracy of both the collected evidences and the predicted pathology.

[1] A chief complaint is a concise statement in English or other natural language of the symptoms that caused a patient to seek medical care. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7161385/

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

However, a crucial part of patients' interaction is largely overlooked in prior works, namely using the collected evidences and predicted diseases to support human doctors' final decisions. A patient's interaction does not end with models' predictions, and therefore it is important to build models that doctors can *trust*; otherwise, they would need to interact with patients from scratch, defeating the purpose of such models. For doctors to trust its output, the model, in addition to making accurate predictions, needs to behave in a way that mimic human doctors' *reasoning* so that it is understandable by doctors (Druce et al., 2021). The importance of building doctor-understandable machine learning systems for the purpose of translation into medical applications is echoed in prior works in other medical areas (Clough et al., 2019). In particular, when interacting with a patient, senior doctors tend to proceeds under an *organized and logical* diagnostic framework (Mansoor, 2018) to build a *differential diagnosis*, that is a set of plausible underlying pathologies, with a particular attention put on collecting evidences for ruling out *severe pathologies* (Ramanayake and Basnayake, 2018).

In this work, we make a case for explicitly modelling senior doctors' reasoning in designing evidence acquisition and automated diagnosis models using RL. More specifically, we focus on (1) using **differential diagnosis**, rather than the ground truth pathology, as models' training targets, (2) modulating the evidence acquisition process to resemble senior doctors' **exploration-confirmation** approach, and (3) prioritizing **severe pathologies**.

In the remaining sections, we explain the concepts of differential diagnosis, exploration-confirmation, and severe pathologies in the context of medical history taking (§2); we review existing works and their shortcomings (§3); we introduce some notations (§4) and describe our method that improves upon prior works especially in mimicking doctors' reasoning (§5); we provide empirical results demonstrating the benefits of the proposed approach (§6 and 7); finally, we discuss the limitations (§8) before concluding this work (§9). Our main contributions are: (1) we reformulate the task of evidence acquisition and automated diagnosis by introducing doctors' trust as a desideratum, and argue for explicitly designing models towards this goal; (2) we propose a RL agent, CASANDE, that promotes the desired behavior in addition to accurate predictions, by means of predicting differentials and reward shaping; (3) we empirically show that existing strong models on this task, even with modifications, are not sufficient for the proposed goal of mimicking human doctors' reasoning, and that CASANDE improves over existing models while being competitive on conventional metrics.

2 Medical history taking

Medical history is a broad concept encompassing various kinds of information relevant to a patient's concern[2]. It includes, for example, past illnesses, current symptoms, demographics, etc. In taking medical history, a doctor would organize the questions under an *exploration-confirmation* framework. At any point during the interaction, the doctor would have in their mind a *differential diagnosis* of diseases worth considering, which might be modified after further questions. In forming the differential diagnosis, the doctor would treat *severe pathologies* differently by prioritizing questions that could help ruling them out even if they are less likely. We provide more details about these concepts below.

Differential diagnosis Prior work on automatic diagnosis primarily focused on predicting the ground truth pathology, i.e., the disease causing the patient's symptoms(Chen et al., 2021; Zhao et al., 2021; Yuan and Yu, 2021; Guan and Baral, 2021; Wei et al., 2018; Xu et al., 2019; Kao et al., 2018). However, medical doctors consider a set of plausible diseases, known as the *differential diagnosis* or simply *differential*, which is refined throughout the interaction with the patient(Henderson et al., 2012; Guyatt et al., 2002; Rhoads et al., 2017). (Richens et al., 2020) considers the differential in their approach; however it is unclear whether all evidences are provided to the model at the same time or if the model is part of the acquisition process. Our approach considers the differential diagnosis as an essential part of the model. Considering differentials instead of a single pathology has the added benefit of accounting for uncertainty and errors inherent in diagnosis, especially when decisions are made solely based on interactions with patients and without medical exams or tests. For this reason, being able to predict the differentials rather than the ground truth pathology is an important part of gaining doctors' trust in the model.

Exploration-confirmation In addition to considering differentials, inquiring about evidences in a manner similar to the way a doctor would engage with patients is another important factor in

[2]https://www.cancer.gov/publications/dictionaries/cancer-terms/def/medical-history

gaining doctors' trust in automated systems, and this is largely overlooked in prior works as well. Doctors generally proceed according to a diagnostic framework that provides an organized and logical approach for building a differential. More specifically, the interactions conducted by a senior doctor generally consist of two distinct phases: *exploration* followed by *confirmation*. During exploration, the doctor mainly asks questions that keep multiple pathologies under evaluation. In that phase, pathologies can be removed from or added to the differential depending on the patient's answers. During the confirmation phase, the doctor inquires about evidences that strengthen the actual differential (s)he is considering (Mansoor, 2018; Richardson and Wilson, 2015; Rhoads et al., 2017). This two-phase quality of doctors' interactions with patients is a separate dimension of the acquisition process that differs from the objective of predicting differentials, and therefore it should be considered and measured separately from the common paradigm of optimizing prediction accuracy.

Severe pathologies Among all possible diseases, severe pathologies receive more attention than others. Generally, a doctor does not want to miss out on them and, therefore, they would explicitly gather evidences to rule them out from the differential as soon as they become plausible, even if they are less unlikely than other diseases (Rhoads et al., 2017). Emulating this behavior within automated systems would contribute to increasing doctors' trust in these systems.

3 Related work

There have been a variety of prior works on evidence acquisition and automatic diagnosis. For example, Wei et al. (2018) proposes to use DQN (Mnih et al., 2015) to approach the problem. Kao et al. (2018) additionally proposes to create a hierarchy of symptoms and diseases based on anatomical parts and train an agent using Hierarchical Reinforcement Learning (Sutton et al., 1999). Xu et al. (2019), Zhao et al. (2021), and Liu et al. (2022) propose to encode relations among different evidences and evidence-disease pairs to enhance the efficiency of an agent trained using DQN. Peng et al. (2018) uses a policy gradient algorithm (Williams, 1992) and relies on reward shaping functions (Ng et al., 1999; Wiewiora et al., 2003; Devlin and Kudenko, 2012) defined on the state space to favour positive evidence acquisition. The proposed approach also makes use of reward shaping functions, but unlike the work done in (Peng et al., 2018), they are defined on the action space and used to induce bias regarding doctors' reasoning in our agent.

In above mentioned approaches, the space of the evidences to be collected and the one of the diseases to be predicted are merged together. Janisch et al. (2019) considers to deal with those spaces separately and proposes an agent which has two branches correspondingly, i.e. one that decides which evidence to inquire, trained using RL, and the other that predicts the disease, trained using supervised learning. Other approaches enhance the training of the evidence inquiry branch by using information from the classifier branch. For example, Kachuee et al. (2019) relies on Monte-Carlo dropout (Gal and Ghahramani, 2016) to estimate the certainty improvement from the classifier branch output and uses that information as an evidence acquisition reward to train the acquisition branch. Yuan and Yu (2021) proposes an adaptive method to align the tasks performed by the two branches. More specifically, the acquisition branch receives an extra reward of the change in entropy of prediction, and the acquisition process ends when the entropy of the prediction is below a dynamically learned, disease-specific threshold. Our approach also uses the two-branches setting and designs a mechanism to inform the acquisition process when to stop based on information from the classifier branch. Unlike Yuan and Yu (2021), we explicitly introduce a *stop action* whose Q-value is trained to replicate a reward derived from the classification branch.

The evidence acquisition branch can be trained with learning paradigms different from RL. Chen et al. (2021) proposes to model the evidence acquisition process as a sequential generation task and train the agent in a similar manner to BERT's Masked Language Modelling (MLM) objective (Devlin et al., 2019). Similarly, Luo et al. (2020) relies on randomly generated trajectories to train in a supervised way a system to collect evidences from patients. Finally, Guan and Baral (2021) makes use of the Quick Medical Reference belief network (Miller et al., 1986), and applies Bayesian experimental design (Chaloner and Verdinelli, 1995) in the inquiry phase while relying on the Bayes rule to infer the corresponding disease distribution.

However, most of these approaches are primarily focused on predicting the single ground truth pathology, as opposed to the differential diagnosis in this work. Also, they have rarely made specific efforts to shape or guide the interaction to resemble how doctors would interact with patients. As a

result, it is debatable how much doctors could trust such systems, even though some may perform well on benchmark datasets, and therefore casting doubt on their applicability in real-world applications.

4 Preliminaries

Let E and D be the number of all evidences (i.e., symptoms and antecedents) and diseases under consideration. Evidences can be *binary* (e.g., *are you coughing?*), *categorical* (e.g., *what is the pain intensity on a scale of 1 to 10*), or *multi-choice* (e.g., *where is your pain located*). The task of automated evidence collection and diagnosis can be viewed as a sequential decision process where each interaction with a patient is formalized using a finite-horizon Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, T)$ with a state space $\mathcal{S} = \mathcal{S}' \cup \{\mathbf{s}_\perp\}$ ($\mathbf{s}_\perp$ is the terminal state), an action space $\mathcal{A}$, a dynamics P, a reward function R, a discount factor γ, and a maximum episode length T. A state $\mathbf{s} \in \mathcal{S}'$ encodes socio-demographic data regarding the patient (e.g., age, sex) as well as the evidences provided by the patient so far. $\mathcal{A}$ is defined as $\{1, \cdots, E, E+1\}$ where the first E elements, referred to as *acquisition actions*, are used to inquire about corresponding evidences, and the last element is the *exit action* that is used to explicitly terminate the interaction. At any point in time, only actions not yet selected in the episode are available for future selection. P is deterministic and updates the current state based on the current action and the patient response if the current turn is less than T, otherwise, the next state is set to $\mathbf{s}_\perp$. Finally, regarding R, each acquisition action is characterized by an inquiry cost r_i, and depending on whether the underlying patient is experiencing the corresponding evidence, it may additionally incur a retrieval reward r_p or a missing penalty r_n. Upon the termination of the interaction, a prediction regarding the patient's disease is made based on the collected evidences.

5 Method

An overview of the proposed approach is depicted in Figure 1. Our model is made of two branches: an evidence acquisition branch in charge of the policy for interacting with patients and a classifier branch in charge of predicting the differential at each turn. Both the classifier and the policy networks rely on a latent representation computed by an encoder from the evidences collected so far from the patient. Importantly, we introduce a *reward shaping module* which leverages the output of the classifier to define auxiliary rewards that help induce the desired properties from doctors' reasoning identified in § 2 in the agent during training. More precisely, the *exploration* and *confirmation* rewards are designed to encourage the two-phase based interactions, whereas the *severity* reward is used to explicitly handle severe pathologies. Finally, the *classification* reward is used to train the agent to recognize when to stop the interaction process. The proposed agent predicts at each turn t the next action $\mathbf{a}_t$ as well as the belief bel_t regarding the patient disease from the current state $\mathbf{s}_t$. We rely on a DQN variant algorithm, Rainbow (Hessel et al., 2018) (without noisy networks Fortunato et al. (2018)), to search for the optimal policy and the objective of the proposed auxiliary rewards is to encourage trajectories similar to those of experienced medical practitioners. In the next sections, we thoroughly describe each of these auxiliary rewards.

5.1 Exploration reward: encouraging the evaluation of multiple pathologies

We want our agent to undertake actions that favour considering several pathologies in the differential diagnosis during the first phase of the interaction. In practice, this means that the agent's belief is constantly changing after inquiring about a new evidence. As more evidences are collected, this fluctuation should gradually reduce until it becomes insignificant. From this intuition, we derived the exploration reward R_{Ex} as

$$R_{\text{Ex}}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \mathbb{1}_{\mathbf{s}_{t+1} \neq \mathbf{s}_\perp} w_{\text{Ex}}(t) \times JSD(bel_t, bel_{t+1}), \tag{1}$$

where JSD is the Jensen-Shannon divergence that computes the dissimilarity between consecutive beliefs, and $w_{\text{Ex}}(t) \in [0, 1]$ is a dynamic weight which aims at controlling the importance of this shaping component as a function of time. Inspired from the shape of the *sigmoid* function, we design $w_{\text{Ex}}(t)$ as a translated version in $[0, T]$ of the "flipped" sigmoid function within the interval $[\bar{x}_{min}, \bar{x}_{max}]$. In other words, using $trans(t) = \frac{t*(\bar{x}_{max} - \bar{x}_{min})}{T} + \bar{x}_{min}$, we have

$$w_{\text{Ex}}(t) = sigmoid(-(trans(t) + \Delta_{\text{Ex}})), \tag{2}$$

where Δ_{Ex} is a hyper-parameter that controls how fast the function saturates at 0 (cf. Appendix A.1).

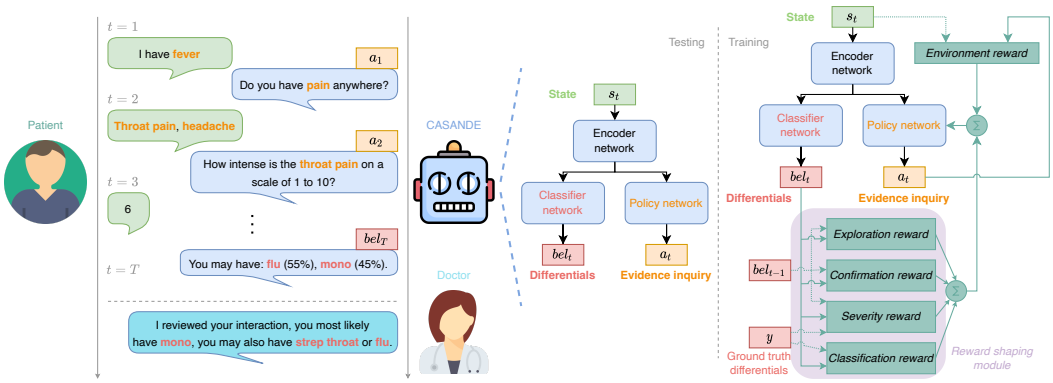

Figure 1: **Left**: An synthetic example of a patient's interaction with our proposed model. Importantly, human doctors' role of final decision maker is largely overlooked in prior works. The example is for illustration purpose only and is not meant as medical advice. **Middle**: The agent consists of two parts: an evidence acquisition module and a classifier. **Right**: The reward shaping module leverages the output of the classifier to define auxiliary rewards that help induce desired features of medical doctor's reasoning in the agent.

5.2 Confirmation reward: Strengthening the agent's belief in the pathology

In the second phase, the agent should inquire about evidences that help strengthen its belief regarding the differential diagnosis. This means that the more the agent collects evidences, the closer its belief should be to the ground truth distribution y. The importance of this shaping component should gradually increase as the agent moves towards the end of the interaction. From this intuition and inspired from potential-based reward shaping Ng et al. (1999), we derived the confirmation reward R_{Co} as

$$R_{\text{Co}}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = -\mathbb{1}_{\mathbf{s}_{t+1} \neq \mathbf{s}_\perp} w_{\text{Co}}(t) \times (\gamma\, CE(bel_{t+1}, y) - CE(bel_t, y)), \tag{3}$$

where CE is the cross-entropy function, and $w_{\text{Co}}(t) \in [0, 1]$ controls the importance of this shaping component as a function of time. Like $w_{\text{Ex}}(t)$, $w_{\text{Co}}(t)$ is a translated version of the sigmoid function with the difference that it increases over time. Thus, we have:

$$w_{\text{Co}}(t) = sigmoid((trans(t) + \Delta_{\text{Co}})), \tag{4}$$

where Δ_{Co} is a hyper-parameter controlling how fast the function saturates at 1 (cf. Appendix A.1).

5.3 Severity reward: evidence gathering for ruling out severe pathologies

We want our agent to collect evidences that help ruling out severe pathologies as soon as they become plausible. One proxy to achieve this for the agent is to behave in such a way to monotonically increase the number $SevOut$ of severe pathologies that are not in both the ground truth differential and its predictions through time. As such, we define the severity reward R_{Sev} as

$$R_{\text{Sev}}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \mathbb{1}_{\mathbf{s}_{t+1} \neq \mathbf{s}_\perp \& SevOut_{t+1} \neq SevOut_t} (\gamma SevOut_{t+1} - SevOut_t). \tag{5}$$

Here, a pathology is not part of a predicted distribution if its probability is below a given threshold.

5.4 Classification reward

This reward is designed to provide the agent feedback regarding how good is its final predicted differential with respect to the ground truth differential y when the interaction process is over. Let $SevIn_t$ be the number of severe pathologies that are part of both the ground truth differential y and the predicted belief bel_t and let Sev_y be the number of severe pathologies in y. We define $\mathcal{V}(\mathbf{s}_t, y) = -CE(bel_t, y) + w_{\text{si}} \frac{SevIn_t}{Sev_y}$ as a measure of the quality of the belief prediction from state $\mathbf{s}_t$ where CE is the cross-entropy function, and the second term, controlled by the hyper-parameter w_{si}, is meant to deal with the missing of severe pathologies. We thus define the classification reward R_{Cl} as

$$R_{\text{Cl}}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \mathbb{1}_{\mathbf{s}_{t+1} = \mathbf{s}_\perp} \mathcal{V}(\mathbf{s}_t, y). \tag{6}$$

5.5 Training Method

The above mentioned auxiliary rewards are combined together according to the Equation 7 where the α weights are hyper-parameters. The resulting function F is then combined with the environment reward leading to a new reward function $R' = R + F$ that is used to train the policy network.

$$F = \alpha_{\text{Ex}} R_{\text{Ex}} + \alpha_{\text{Co}} R_{\text{Co}} + \alpha_{\text{Sev}} R_{\text{Sev}} + \alpha_{\text{Cl}} R_{\text{Cl}}. \tag{7}$$

For simplicity, this section relies on DQN to present the loss for training each part of our agent, with the extension to Rainbow easy to do. Let θ and ϕ be the agent and target network parameters.

Policy network loss function: The Q-value of the *exit* action does not depend on any following state. Also, the reward received when explicitly exiting is the classification reward. Therefore, to synchronize the agent branches, the *exit* action Q-value needs to be reconciled with the classification reward. This should be done at each interaction time step during the training, given that the agent needs to have a good estimate of the expected Q-value of the *exit* action to select this action. Thus,

$$Loss_Q(B) = \frac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}, y) \in B} \left[[Q_t - Q_\theta(\mathbf{s}_t, \mathbf{a}_t)]^2 + \mathbb{1}_{\mathbf{s}_{t+1} \neq \mathbf{s}_\perp} \left[Q_\theta(\mathbf{s}_t, \mathbf{a}_{exit}) - \mathcal{V}(\mathbf{s}_t, y) \right]^2 \right] \tag{8}$$

where B is a batch from the replay buffer, $Q_\theta(\mathbf{s}_t, \mathbf{a}_{exit})$ is the predicted Q-value for the pair $(\mathbf{s}_t, \mathbf{a}_{exit})$, $Q_t = r'_t + \mathbb{1}_{\mathbf{s}_{t+1} \neq \mathbf{s}_\perp} \gamma \max_{\mathbf{a} \in \mathcal{A}} Q_\phi(\mathbf{s}_{t+1}, \mathbf{a})$ is the target Q-value for the pair $(\mathbf{s}_t, \mathbf{a}_t)$ with $r'_t = r_t + F(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$.

Classifier network loss function: The classifier is updated using the following loss at the end of an interaction:

$$Loss_C(B) = \frac{1}{2} \mathop{\mathbb{E}}_{(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}, y) \in B} \left[\mathbb{1}_{\mathbf{s}_{t+1} = \mathbf{s}_\perp} CE(bel_t, y) \right]. \tag{9}$$

Training process: Using the loss functions defined in Equations 8 and 9, the two branches are alternatively updated until the stopping criteria are satisfied. Both updates share the same batch of data B which is sampled from the replay buffer. The training process is described in Appendix A.2.

6 Experiments

Datasets The datasets used in prior works such as DX (Wei et al., 2018), Muzhi (Xu et al., 2019), SymCAT (Peng et al., 2018), HPO (Guan and Baral, 2021), and MedlinePlus (Yuan and Yu, 2021) do not provide differential diagnosis information, and therefore they are ineligible for validating the approach proposed in this work. Besides, as supported by Yuan and Yu (2021), the patients simulated using SymCAT are not sufficiently realistic for testing automatic diagnosis systems. We instead use the DDXPlus dataset (Fansi Tchango et al., 2022) for that purpose. In addition to providing differential diagnosis information, the DDXPlus dataset, unlike the above mentioned datasets, is not restricted to binary evidences, but also has categorical and multi-choice evidences. Furthermore, the DDXPlus dataset contains 49 pathologies and 223 evidences. Finally, the dataset is split into 3 subsets: a training subset with more than 10^6 synthetic patients, a validation and a test subsets containing roughly 1.4×10^5 synthetic patients each.

Baselines We consider 4 baselines, AARLC (Yuan and Yu, 2021), Diaformer (Chen et al., 2021), BED (Guan and Baral, 2021), and ASD which is an extension on Luo et al. (2020). AARLC demonstrates SOTA results on the SymCAT dataset while Diaformer shows competitive results on the Muzhi and DX datasets. BED, an approach that does not require training, has impressive results on the HPO dataset. All these methods are designed for only binary evidences, and thus they are adapted to deal with the DDXPlus dataset. Also, because they were designed to predict the ground truth pathology, we also modify these models and train them to predict the differentials. Additional details are provided in Appendix B.

Experimental settings Each patient in the DDXPlus dataset is characterized by a chief complaint, which is presented to the models at the beginning of the interaction, a set of additional evidences that the models need to discover through inquiry, a differential diagnosis, and a ground truth pathology.

Table 1: Evaluation performance on the test set. Values indicate the average of 3 runs, and numbers in brackets indicate 95% confidence intervals. BED is deterministic and related results are from 1 run. ↓ indicates lower is better and ↑ indicates higher is better.

			Values expressed in percentage (%) ↑				
	Method	**IL ↓**	**GTPA@1**	**DDF1**	**DSF1**	**PER**	**TVD**
Trained With Differential	**CASANDE**	19.71 (0.46)	69.7 (3.54)	**94.24 (0.55)**	**73.88 (0.34)**	**98.39 (0.86)**	**6.2 (0.87)**
	Diaformer	18.41 (0.07)	73.62 (0.60)	83.3 (2.10)	69.32 (0.73)	92.92 (0.30)	1.58 (0.01)
	AARLC	**14.36 (0.56)**	62.06 (1.21)	75.7 (4.38)	67.49 (2.09)	49.9 (7.91)	2.32 (0.20)
	ASD	17.86 (0.88)	67.71 (1.19)	83.69 (1.57)	65.06 (2.30)	88.18 (1.12)	4.36 (0.38)
Trained Without Differential	**CASANDE**	19.53 (0.87)	98.8 (0.54)	30.84 (0.32)	10.62 (0.22)	**98.07 (1.91)**	**6.97 (1.26)**
	Diaformer	18.45 (0.33)	91.81 (1.95)	30.38 (14.72)	**19.90 (31.07)**	92.61 (1.12)	0.60 (0.23)
	AARLC	6.73 (1.35)	99.21 (0.78)	31.28 (0.38)	10.96 (0.26)	32.78 (13.92)	3.65 (0.55)
	ASD	17.99 (3.57)	97.15 (1.70)	31.31 (0.29)	10.81 (0.29)	88.45 (5.78)	3.84 (0.46)
	BED	**5.47**	99.47	31.01	10.78	18.62	3.91

We allow interactions to have a maximum of $T = 30$ turns. We tune the hyper-parameters for each model, including the baselines, separately on the validation set and use the resulting optimal set of parameters to report the performances. For more details, see Appendix A.4.

Evaluation metrics We report on the interaction length (IL), the recall of the inquired evidences (PER), the ground truth pathology accuracy (GTPA@1), and the F1 score of the predicted differential diagnosis (DDF1). Additionally, we compute the F1 score regarding the prediction of the severe pathology (DSF1). Finally, we report on the total variation distance (TVD), a proxy metric used to quantify the portion of the space of the probability distribution considered during an interaction. See Appendix D for the definition of these metrics.

7 Results

Disease prediction and evidence acquisition Table 1 depicts the performances obtained at the end of the interaction process by each model. On recovering the ground truth pathology, CASANDE performs on par with the baselines. However, when trained with the differential, CASANDE is significantly better on DDF1 than the baselines, outperforming the second-best model by more than 10% absolute. Meanwhile, they are on par on that metric when the differential is not used at training time. Likewise, CASANDE outperforms the baselines in recovering the severe pathologies (DSF1) with an absolute margin of more than 4% when trained with the differential, while it behaves similarly as the baselines (except Diaformer) otherwise. The observed difference between the resulting performance for each training setting suggests that relying only on the ground truth pathology at training time is not enough to accurately predict the differential diagnosis at inference time. Furthermore, with the highest TVD scores, CASANDE shows its ability to explore wide regions of the probability space throughout the interactions compared to the baselines without sacrificing the end performances.

CASANDE also achieves the best performance on the PER metric demonstrating its ability to collect relevant evidences. We also observe that the PER score usually improves, for a given model, when the differential is used during training. This suggests that differentials are beneficial for collecting evidences. Finally, CASANDE has the longest interaction length . This is unsurprising, however, as CASANDE is trained to emulate doctors' reasoning. In particular, the exploration phase may lead to the model asking more questions.

Trajectory quality evaluation We plot on Figure 2 the *confirmation score* versus the *exploration score* throughout the trajectories for the different models. The exploration score captures how distant are two consecutive agent's predictions while the confirmation score captures how close is the agent prediction w.r.t. the target distribution (see Appendix D.2 for the definitions). Ideally, an agent would start with a high (resp. low) exploration (resp. confirmation) score (upper-left corner), gradually decrease (resp. increase) it, until the score reaches a low (resp. high) value at the end of the interaction (lower-right corner). As it can be observed, this trend is highly followed by CASANDE which exhibits the highest exploration (resp. confirmation) score at the beginning (resp. end) of the interaction while, unlike other models, it is consistently moving towards the lower-right corner of the chart.

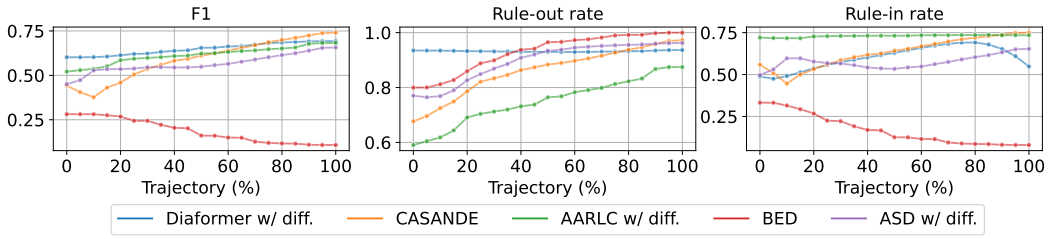

Figure 3: Average F1 score, rule-in rate and rule-out rate on severe pathologies throughout trajectories. For each trajectory, their values are calculated at each turn and values from turns that are distant from each other of 5% of the interaction length are selected (i.e., 21 positions in total). Then, the average value for each such position is computed throughout the dataset. The results are on the test set.

Handling of severe pathologies We are interested in analysing the pace at which the severe pathologies are ruled out from or ruled in within the predicted differentials throughout the trajectories. We show in Figure 3 the curves corresponding to the average values of (1) the F1 scores, i.e., DSF1, (2) the *rule-out* rate, and the *rule-in* rate of severe pathologies throughout the trajectories (see Appendix D.3 for the definitions). As observed, CASANDE improves significantly on the F1 score throughout a trajectory, and eventually performs the best among all models. This shows that the capability of CASANDE for predicting severe pathologies is heavily grounded on the evidence it gradually collects. This evidence-based behavior is further confirmed by rule-out and rule-in rate as CASANDE is the only method that significantly improves on both rates throughout the trajectories.

Interestingly, when focusing on the rule-out curve, it is noticeable that CASANDE performs the second-worst at the beginning, trailing the best performing model by more than 20%, but it improves to the second-best at the end, with the gap smaller than 3%. This suggests that at the beginning CASANDE is more lenient on including severe diseases in its differentials, resulting in more false positives than others. However, this is a desired behavior of CASANDE, as during the exploration phase it is expected to consider unlikely but severe diseases before gathering enough evidences to rule them out. In other words, at the beginning of a trajectory where few evidences are gathered, a model should err on the side of caution by keeping unlikely but severe diseases into consideration.

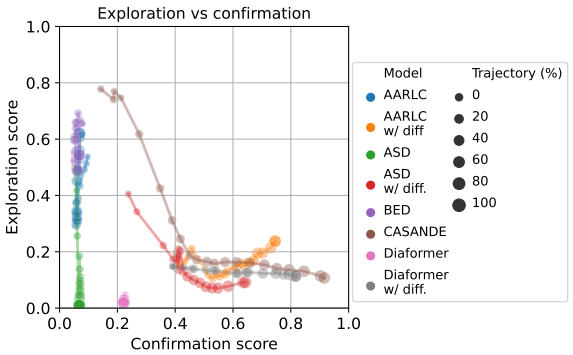

Figure 2: Confirmation score vs exploration score throughout trajectories on average for each model on test set. For each trajectory, the scores are computed for each turn. We select the score values at turns distant from each other of 5% of the trajectory length (i.e., 21 positions in total). Finally, we average the scores obtained for each such position throughout the dataset.

Ablation studies We perform an ablation analysis to illustrate the contribution of each auxiliary reward component of CASANDE to the overall performance. Table 2 shows the results when some reward components are deactivated. We can observe that having both the confirmation and classification rewards do help the agent achieving strong performances in predicting the differential diagnosis. Also, as expected, the exploration score contributes in increasing the interaction length. Finally, we observe a slight decrease of the DSF1 metric when the severity reward is not used. Additional studies are provided in Appendix E.

8 Limitations and potential negative social impact

Limitations In this work, we set out to include doctors' trust as part of the desiderata in building evidence acquisition and automated diagnosis systems. In doing so, we focus our efforts on three essential aspects of doctors' reasoning, namely differential diagnosis, exploration-confirmation, and

Table 2: Impact of the auxiliary reward components on the differential prediction and evidence collection, as measured on the validation set. Values are in % except for IL.

R_{Ex}	R_{Co}	R_{Sev}	R_{Cl}	**IL**	**DDF1**	**DSF1**	**PER**
×	×	×	×	**17.67**	93.41	73.58	98.41
×	×	×	✓	17.76	93.02	73.21	97.91
✓	✓	✓	×	18.92	92.95	**74.05**	97.95
×	✓	✓	✓	18.03	94.4	73.74	97.96
✓	×	✓	✓	18.28	93.91	73.98	98.36
✓	✓	✓	✓	19.96	**94.44**	73.99	**98.6**

severe pathologies. We choose to focus on them as advised by our collaborating physicians. However, to most accurately evaluate doctors' trust, the proposed method has to be deployed so as to allow doctors different from the ones we collaborate with to judge. Additionally, it is reasonable to assume that doctors in other locations and/or specialties may have different ways of engaging with patients. Therefore, while we provide evidence in medical literature (§2) to show the identified desiderata are widely applicable, we recognize that there may be cases where this work does not apply. In this work, we conduct experiments on synthetic patients, mainly due to the lack of real-patient datasets that contain differential diagnosis. We recognize that synthetic patients can be different from real patients in various and important ways, and that therefore results reported on synthetic patients may not extend to real patients.

Potential negative social impact As we have made clear from the beginning, evidence acquisition and automated diagnosis systems are not substitutes for human doctors, but rather they are supportive tools for doctors, who should make the final decisions. However, the predictions of such systems potentially might be provided to patients as the final medical advice, without human doctors' intervention. In such cases, the instructions that patients receive may be misleading or erroneous, and thus can do more harm than good to patients' health.

9 Conclusion

We reflected on the task formulation of evidence acquisition and automatic diagnosis that are essential for telemedicine services, and introduced doctors' trust as an additional desideratum. Concretely, we argued emulating doctors' reasoning is critical for gaining doctors' trust, and we identified three essential features of doctors' reasoning that models can mimic. We proposed a novel RL agent with these features built in. We showed empirically existing models are insufficient for imitating doctors' reasoning. We then demonstrated the importance of the explicit modelling of differential diagnosis, and the efficacy of our model on emulating doctors, while being competitive on conventional metrics. This work is a first step towards reshaping the research in automatic diagnosis systems, and there is abundant potential for future work to explore. For instance, incorporating doctors' reasoning into supervised approaches is a direction worth investigating.

References

Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273 – 304, 1995.

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. Diaformer: Automatic diagnosis via symptoms sequence generation. *arXiv preprint arXiv:2112.10433*, 2021.

James R Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink, Andrew P King, and Julia A Schnabel. Global and local interpretability for cardiac mri classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 656–664. Springer, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Sam Michael Devlin and Daniel Kudenko. Dynamic potential-based reward shaping. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 433–440. IFAAMAS, 2012.

Jeff Druce, James Niehaus, Vanessa Moody, David Jensen, and Michael L Littman. Brittle ai, causal confusion, and bad mental models: Challenges and successes in the xai program. *arXiv preprint arXiv:2106.05506*, 2021.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Arsene Fansi Tchango, Zhi Wen, Rishab Goel, and Joumana Ghosn. DDXPlus: A new Dataset for Medical Automatic Diagnosis. *preprint*, 2022. URL `https://github.com/bruzwen/ddxplus/blob/main/Dialogue01_Dataset.pdf`.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, et al. Noisy networks for exploration. In *International Conference on Learning Representations*, 2018.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Hong Guan and Chitta Baral. A bayesian approach for medical inquiry and disease inference in automated differential diagnosis. *arXiv preprint arXiv:2110.08393*, 2021.

Gordon Guyatt, Drummond Rennie, Maureen Meade, Deborah Cook, et al. *Users' guides to the medical literature: a manual for evidence-based clinical practice*, volume 706. AMA press Chicago, 2002.

Mark Henderson, Lawrence M Tierney, and Gerald W Smetana. *The patient history: Evidence-based approach*. McGraw Hill Professional, 2012.

Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3959–3966, 2019.

Mohammad Kachuee, Orpaz Goldstein, Kimmo Kärkkäinen, Sajad Darabi, and Majid Sarrafzadeh. Opportunistic learning: Budgeted cost-sensitive learning from data streams. In *International Conference on Learning Representations*, 2019.

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

Asim Kichloo, Michael Albosta, Kirk Dettloff, Farah Wani, Zain El-Amir, Jagmeet Singh, Michael Aljadah, Raja Chandra Chakinala, Ashok Kumar Kanugula, Shantanu Solanki, et al. Telemedicine, the current covid-19 pandemic and the future: a narrative review and perspectives moving forward in the usa. *Family medicine and community health*, 8(3), 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Wenge Liu, Yi Cheng, Hao Wang, Jianheng Tangi, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. "my nose is running.""are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. *arXiv preprint arXiv:2204.13953*, 2022.

Hongyin Luo, Shang-Wen Li, and James R. Glass. Knowledge grounded conversational symptom detection with graph memory networks. In *CLINICALNLP*, 2020.

Andre Mansoor. *Frameworks for internal medicine*. Lippincott Williams & Wilkins, 2018.

R Miller, FE Masarie, and JD Myers. Quick medical reference (qmr) for diagnostic assistance. *M.D. Computing : Computers in Medical Practice*, 3:34–48, 1986.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, volume 99, pages 278–287, 1999.

Yu-Shao Peng, Kai-Fu Tang, Hsuan-Tien Lin, and Edward Chang. REFUEL: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. *Advances in Neural Information Processing Systems*, 31:7322–7331, 2018.

R.P.J.C. Ramanayake and Basnayake Mudiyanselage Duminda Bandara Basnayake. Evaluation of red flags minimizes missing serious diseases in primary care. *Journal of Family Medicine and Primary Care*, 7:315 – 318, 2018.

Jacqueline Rhoads, Julie C Penick, et al. *Formulating a Differential Diagnosis for the Advanced Practice Provider*. Springer Publishing Company, 2017.

W. Scott Richardson and Mark C. Wilson. *The Process of Diagnosis*. McGraw-Hill Education, New York, NY, 2015. URL `jamaevidence.mhmedical.com/content.aspx?aid=1183877015`.

Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):1–9, 2020.

Adam Stooke and Pieter Abbeel. rlpyt: A research code base for deep reinforcement learning in pytorch. *CoRR*, abs/1909.01500, 2019. URL `http://arxiv.org/abs/1909.01500`.

Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2033. URL `https://aclanthology.org/P18-2033`.

Eric Wiewiora, Garrison W Cottrell, and Charles Elkan. Principled methods for advising reinforcement learning agents. In *Proceedings of the 20th International Conference on Machine Learning*, pages 792–799, 2003.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, may 1992.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33017346. URL `https://doi.org/10.1609/aaai.v33i01.33017346`.

Hongyi Yuan and Sheng Yu. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *arXiv preprint arXiv:2112.00733*, 2021.

Xinyan Zhao, Liangwei Chen, and Huanhuan Chen. A weighted heterogeneous graph-based dialog system. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Sections 5, 6, 7.
   (b) Did you describe the limitations of your work? [Yes] See Section 8.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 8.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See Section 8.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] the code is provided in the supplementary and it contains a readme with instructions needed to reproduce the main experimental results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Data splits is described in Section 6. Other training details are described in Appendix C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 7
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6.
   (b) Did you mention the license of the assets? [Yes] The license for the code is attached with the code in the supplementary.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include code in the supplementary.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The data used in this project is publicly available under the CC-BY licence.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The data is synthetic and does not contain personally identifiable information (See Section 6).

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]