Modelling Information Perceiving within Clinical Decision Support using Inverse Reinforcement Learning

Ashish T. S. Ireddy[0000−0003−2964−4371] and Sergey V. Kovalchuk[0000−0001−8828−4615]

ITMO University, Saint Petersburg, Russia
{ireddy,kovalchuk}@itmo.ru

Abstract. Decision support systems in the medical domain is budding field that aims to improve healthcare and overall recovery for patients. While treatment remains specific to individual symptoms, the diagnosis of patients is fairly general. Interpreting the diagnosis and assigning the appropriate care treatment is a crucial part undertaken by medical professionals, however, in critical scenarios, having access to recommendations from a clinical decision support system may prove life-saving. We present a real-world application of inverse reinforcement learning (IRL) to assess the implicit cognitive state of doctors when evaluating decision support data on a patient's risk of acquiring Type 2 Diabetes mellitus (T2DM). We show the underlying process of modelling a Markov Decision Process (MDP) using real-world clinical data and experiment with various policies extracted from sampled trajectories. T The results provide insights into the approach to modelling real-world data into interpretable solutions via IRL.

Keywords: Inverse Reinforcement Learning · Markov Decision processes · Decision Support systems · Clinical decision making · Diabetes mellitus

1 Introduction

The medical sector has forever been one of the most vital industries in the world with a fragile margin of error. From mild cases to severe, diagnosis and selection of appropriate treatments are key to the patient's recovery. While accuracy and precision are key characteristics that healthcare professionals strive to possess, in many cases, they face crucial situations where a diagnosis or opinion falls short causing unforeseen circumstances. Most often a group of doctors consult together and decide on an appropriate course of action but this cycle eventually gets modulated over time (i.e. due to data scarcity, lack of experience etc). Decision support systems (DSS) have taken up the role of providing a layer of confidence and trust for decision-makers to take necessary actions. Despite being moderately used in the medical domain and under an experimental status, the use of Clinical decision support systems (CDSS) has widely increased

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

to incorporate it into general practice. The current state of CDSSs allows professionals to feed data (present and historical) and evaluate the best possible scenarios, optimal solutions, risk conditions and many such parameters that can be fine-tuned to provide the matching recommendations respective to individual patients. However, one of the main features that it lacks is the ability to assess situations as humans do. The underlying impact such as social leverage over fear of treatment can be overruled via CDSS. Understanding the way a human views a situation is complex and varies widely based on each activity [4]. Replicating the cognitive state during an activity via traditional reinforcement learning (RL) methods [14] is tedious and extremely demanding. Hence, learning by observation proves to be a more efficient way of modelling the human mind's cognitive state. Within our work, we provide such a case of modelling the human cognitive state where medical professionals interpret true patient information and prediction data from a diabetes prediction model, to assess if a patient will have Type 2 Diabetes mellitus (T2DM). This data is further evaluated by doctors according to their level of perception, understandability, agreement and usability.

Using Inverse reinforcement learning (IRL), we have modelled and extracted the underlying reward functions based on optimal policies that describe the cognitive state of the doctor and the strategy during the evaluation of patient and prediction model data. We demonstrate the use of Linear IRL using Markov decision processes (MDPs) and provide a basic outlook of apprentice learning using cyclic MDP environments. Our results and investigation provide insight and foundational results to approach IRL and MDPs using expert trajectories from real-world data. Further, we provide early results of complex models for IRL policy extraction using expert trajectories and possible frameworks to interpret reward functions and policies to scenarios.

The remaining part of the paper is structured as follows: section 2 describes the background and related works on IRL and MDPs. Section 3 is the methodology where we elaborate on modelling our MDP and the IRL setup. Section 4 is the results section where the generated simulations are investigated and its features elaborated. Section 5 is the discussion and Section 6 is the conclusion.

2 Related works

Learning a task is a process of discovering the outcome after taking a sequence of actions to fulfil an objective. There can be instances when certain actions lead to swift fulfilment of the objective while other approaches might not be conclusive at all. We can view a majority of the day-to-day activities in the real world as a reinforcement problem, where an environment with states, actions and outcomes are defined to reach a certain goal. These activities have already been discovered and documented where the stages ahead involve optimization and improvement of the process itself. A simple example of tuning a guitar shows the effect of over-tuning that leads the strings to snap while under-tuning leads to noise instead of notes. By tuning the guitar to a certain scale, we obtain ideal sound notes and can therefore produce music. However, in open-ended processes, the same

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

cannot be said all the time. Training an autonomous car to mimic the behaviour of a person is one of the most widely used examples to describe the complexity of teaching the penalty, reward and justification for taking specific actions given a situation [11]. Other examples include flying aircraft, learning to play table tennis [9] etc, the direct strategy to fulfil the goal is not explicitly defined yet, and the outline of the problem and its rules are stipulated via observable behaviour of the ideal scenario that is discovered [7]. IRL is aimed at resolving problems where the complete do's and don'ts are not provided by the user and all that remains is a set of expert demonstrations of performing the task from which the ideal behaviour, actions, ethics and rules can be defined by relatively modelling the implicit strategy allowing replication of the behaviour to produce similar results and therefore learning the process itself [2]. Further, from the data standpoint, RL and supervised learning demand explicitly defined pathways and boundary conditions to perform a task on par with humans, IRL stands out by learning over demonstrations and therefore minimizes the initial effort of conditioning the data and preparing the model itself. Depending on certain activities, IRL algorithms can learn existing scenarios and collectively map newer pathways which can prove to be a lucrative incentive in many medical decision making situations where implicit and explicit factors play an interdependent role in decision making. A simple example is the discharge of patient from a hospital due to early recovery impacted by improving weather conditions, traditional approaches may categorize such an occurrence as an outlier or a "by-chance occurrence" without investigating the underlying reasoning. IRL models may be able to derive such relations, thereby, recommending actions often observed in doctors behaviour.

In the community, one of the most common applications of IRL is the gird world and shortest path problem. During our search we found only one such work that provides insight into using IRL to assess human risk-taking characteristics using real-world data [8]. Apprentice modelling [1] is an extension of full-fledged IRL that can take advantage of the full extent of using expert trajectories to find the best reward function and its matching policy relative to the actual process over data. These results can then be interpreted and understood as the plausible cognitive state space or process of taking actions by a human based on the incentive they gain by reaching the final state. Therefore learning the process and its etiquette while also trying to find an undiscovered pathway that may have been overlooked by the human subject [17]. The prospects of using imitation learning are massive yet we believe the problem lies with modelling the data into appropriate state-feature spaces and MDPs. There may be instances where datasets may not directly be modelled into an MDP without additional modifications or modulations of state spaces. Within our work, we have shown such an example of modelling data into an MDP without the need for modulation. Further, there is also the conundrum of using RL to solve an IRL problem being less efficient. An answer to this has been provided by the authors in [14] where they aim to reset the learning trajectory of the IRL model such that follows the start state

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

of the expert's demonstration and avoids exploring all possible combinations in the state space.

3 Modelling perceiving in clinical decision process with CDSS interaction

In this section we define our approach to modelling our decision data for IRL using MDPs as a baseline interpretation. We define notations that will be used throughout the paper here forth.

A set of n (finite) **expert trajectories** $E_T = \{\tau_1; \tau_2; ....\tau_n\}$ constitute to a combination of state action combinations, defined via the following terms:

- $S = \{s_1, s_2, ...s_n\}$ is a finite set of all possible **states** the agent can take in E_T
- $A = \{a_1, a_2, ...a_n\}$ is a set of **actions** an agent can taken in E_T
- $T_{PA}(.)$ are the state **transition probabilities** of moving to state s' from s upon taking action a, i.e. $T(s, a, s')$, extracted from E_T
- $\gamma \in [0, 1)$ is the **discount factor** that dictates the weightage for long-term-short-term reward strategy
- π is the **policy** function that defines the action to be taken in each state i.e. $(\pi : S \rightarrow A)$
- π^* is the **optimal policy** that defines the optimal actions to take in each state such that the generated reward is maximum
- $\tau = \{(s_0, a_1, s_1); (s_1, a_2, s_2); ....(s_{n-1}, a_n, s_n)\}$ is a **trajectory** describing one complete iteration of the agent in the MDP (i.e. until it concludes or reaches an end state)

Traditional RL involves finding the optimal policy of a problem using a defined reward function. IRL is defined as the opposite where the reward function is sought using the perceived optimal policy from a set of expert demonstrations. When taking into account real-world scenarios of IRL i.e. using expert trajectories, we assume that the experts' actions in given trajectories are the optimal behaviour to be followed. With this in mind, given a set of expert trajectories E_T, we assume there are n policies $\{\pi_1, \pi_2, ...\pi_n\}$ that can generate maximal rewards. We use the linear programming approach for IRL as described in [10] to find the maximal reward function for assumed policy π^*. On finding the maximum reward, we iterate through the various policies possible from the data to get a complete overview of reward functions existing within the expert trajectories. Simply put, we run RL inside IRL to find the reward function respective to our assumed policies from the expert trajectories. Figure 1 provides an overview of our setup for the simulation.

3.1 CDSS data

The dataset used here is from [6] where the authors have performed an experiment to analyse the effect of having decision makers (e.g. doctors, physicians)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

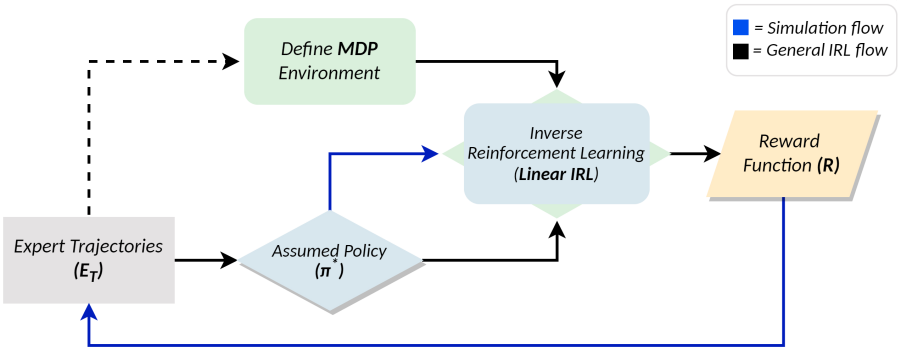

Fig. 1. A schematic overview of our IRL simulation setup where E_T is the real world experts trajectories from which the MDP environment and the policies π are extracted, The MDP tuple is fed to the linear IRL algorithm which assumes optimal policy π^* to be extracted from E_T after which the reward function R is formulated. Our simulation flow involves repeating the process of initialising the MDP respective to each subset of data and all possible policies existing within E_T

supported by information from a prediction model, a FINDRISK measure and case-based explanation to assess the concurrent perceptional state through subjective metrics. Through a survey, physicians were asked to assess the chances of a patient encountering T2DM in the future based on the amount of data presented in three alternative settings:

- ***Setting 1:*** From prediction model only
- ***Setting 2:*** From a FINDRISK scale and prediction model
- ***Setting 3:*** From an explanation, FINDRISK scale and prediction model

The physicians were provided with the patient's basic information (age, height, weight, BMI (body mass index), gender, blood, physical activity level, blood sugar, heredity, arterial pressure) and one of three prediction settings. Physicians were then asked to assess each setting with three subjective perception measures via a Likert scale from 1 (strongly disagree) to 5 (strongly agree), *Understandability* denoting the level of interpretation of the information, *Agreement* corresponding to the acceptance of the model's prediction as per the data and *Usability* reflecting the subsequent usage of the data in further diagnosis. A total of 541 cases of patient assessment data were found to be usable for our experiment.

3.2 Initializing MDP for CDSS data

From the CDSS data we model our MDP as a system of:

- $S : 5$ **states** $= \{End,\ Understandability,\ Agreement,\ Usability,\ Completion\}$

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

– A : 2 **actions** = $\{Continue,\ Terminate\}$
– T_P: Transition probabilities of moving from state s to s' extracted from E_T using the following strategy:

$$T_P(s, a, s') = \frac{\#\ of\ times\ (s\ \rightarrow s'\ )\ occurs\ in\ T_E}{Total\ \#\ of\ occurences\ in\ T_E} \tag{1}$$

Given a trajectory τ of a patient's data evaluated by the physician, the MDP is initialized in the state of understandability and takes an action to either continue or terminate the MDP based on the scored evaluation from the physician respective to each state and metric. The decision to terminate or continue is deterministically relative to the doctors' assessment via a threshold M_T. We have introduced three thresholds for each measure to assess the impact of having strict vs moderate evaluation criteria. Relative to the 1 - 5 Likert scale we have selected thresholds of $M_T = [2, 3, 4]$ to evaluate the changes in reward function per policy over various settings of data. For a complete assessment, the MDP should traverse through the states of understandability, agreement, and usability to finally reach the completion state by taking action *continue* when the given data for each measure is greater than the metric threshold M_T. Likewise, if the condition is not satisfied, by taking action *terminate* the MDP reaches the state of End. Figure 1 describes the MDP and its feature space.

$$\pi(S) \rightarrow A = \begin{cases} Continue, & \text{if } [Understandability\ or\ Agreement\ or\ Usability] > M_T. \\ Terminate, & \text{otherwise.} \end{cases} \tag{2}$$

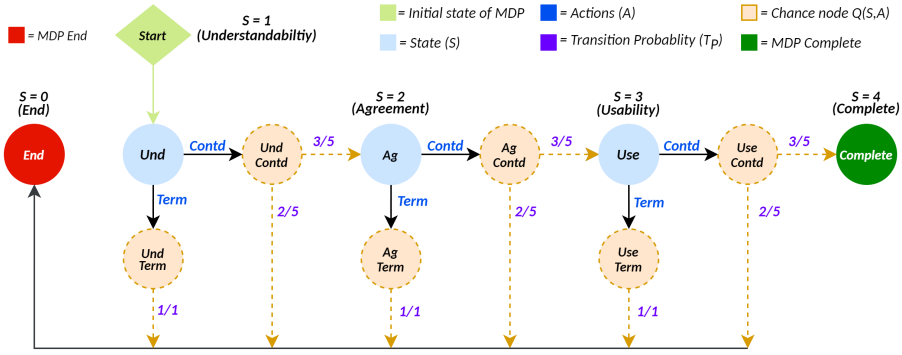

Fig. 2. A visualization of our MDP adapted to CDSS data space. The MDP is initialised in the state of understandability, within each state two actions $Continue\ or\ Terminate$ can be performed, $Q(S, A)$ denotes the chance node and the probability of arriving in state S' on taking action A from which the transition probabilities T_P is extracted

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

4 Case study: T2DM risk prediction perceiving

4.1 Simulating risk Prediciton using MDP

To model the risk perception from CDSS data, we first divided the collected responses of physicians into 4 groups. The three settings as initialized in the previous section and a combined version that holds trajectories of settings 1,2 and 3 together. Next, we set up the MDP as mentioned in the prior section to simulate the flow of the feature space. The transition probabilities T_P were generated using a self-created script 1 where, S and A correspond to the number of states and actions in the MDP, $\tau_n = [Und, Ag, Use]$ is a tuple of metric scores for a set of trajectories, M_T is a metric threshold that is used to generate the transition probability using equation (1).

Algorithm 1: Calculation of transition probabilities

Data: S, A, set of n trajectories $\tau_n = [Und,\ Ag,\ Use]$, $M_T = [2, 3, 4]$
Result: A matrix with size $[S,\ A,\ S]$ with probabilities $\in$ [0, 1]
Feed individual settings data [*setting 1, setting 2, setting 3, combined settings*];
for *n trajectories in* τ_n **do**
  iterate through each patient trajectory;
  if $\tau_n[Und] > M_T$ **then**
    $\#\ of\ A(Continue)_{\tau_n[Und]} += 1$;
  else
    $\#\ of\ A(Terminate)_{\tau_n[Und]} += 1$;
  end
  if $\tau_n[Ag] > M_T$ **then**
    $\#\ of\ A(Continue)_{\tau_n[Ag]} += 1$;
  else
    $\#\ of\ A(Terminate)_{\tau_n[Ag]} += 1$;
  end
  if $\tau_n[Use] > M_T$ **then**
    $\#\ of\ A(Continue)_{\tau_n[Use]} += 1$;
  else
    $\#\ of\ A(Terminate)_{\tau_n[Use]} += 1$;
  end
end
Using equation (1) acquire the transition probability between [0, 1]

We used the linear programming approach to resolve the IRL function as implemented in [3]. We fed the IRL algorithm a tuple (S, A, τ_P, $[\pi]$, R_{max}, γ, $L1$) the number of states S, actions A, transition probabilities T_P, a set of policies $\pi = [\pi_1,\ ....\pi_8]$, a discount factor, maximum reward R_{max}, a discount factor γ and an $L1 \in [0, 1]$ regularization value. The set of policies from our data was defined based on the possible combinations of the MDP reaching the states [*End*, *Und*, *Ag*, *Use*, *Completion*] over the threshold M_T. This results in eight policies as shown in Table1.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

Table 1. All combinations of policy π possible within our MDP space extracted from our CDSS data

Case	*Policy*	*Understandability*	*Agreement*	*Usability*
1	[0, 0, 0, 0, 0]	Terminate	Terminate	Terminate
2	[0, 0, 0, 1, 0]	Terminate	Terminate	Continue
3	[0, 0, 1, 1, 0]	Terminate	Continue	Continue
4	[0, 1, 1, 1, 0]	Continue	Continue	Continue
5	[0, 1, 1, 0, 0]	Continue	Continue	Terminate
6	[0, 1, 0, 0, 0]	Continue	Terminate	Terminate
7	[0, 0, 1, 0, 0]	Terminate	Continue	Terminate
8	[0, 1, 0, 1, 0]	Continue	Terminate	Continue

The resultant of the IRL algorithm is a reward function of specific weights for each state respective to the policy. We chose our γ value to be 0.9,a long-term strategy, relative to predicting the patients' assessment in the future and not in the present state. When a smaller γ value was selected we observed that reward was only awarded in a policy where all three states had a *continue* action and the state of *Complete* was reached, for other policies the reward was 0. This can be attributed to having a short-term strategy of γ. Similarly, our $L1$ regularization value was also set to 0.9 as smaller values produced no plausible reward as output.

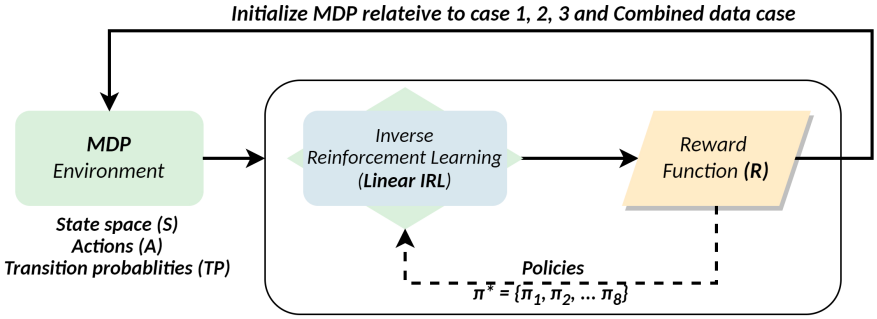

Fig. 3. Overview of the simulation process to extract reward function R for all possible policies in a given data setting. The MDP is generated for each data setting (including the transition probabilities T_P) after which the IRL algorithm is fed with MDP tuple and a set of assumed optimal policies $[\pi^* = \pi^1, ...\pi^8]$ over which the maximum reward is extracted. The same process is carried out for other cases of data.

4.2 Inferring reward functions for trajectories

We initialized the MDP for settings 1, 2, 3 and a combined version. The transition probabilities T_P reflecting the physicians' assessment for each setting varied

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

as the level of data fed was different. This was also observed in the metric ratings provided by physicians when filtered by threshold M_T. It was observed that *setting 1* consisting only of a model prediction data had the highest metric ratings of all cases. *Setting 2* adds one more level of information i.e. FINDRISK score, notably the agreement metric for *setting 2* was lower than the latter metrics, this may be attributed to a steep FINDRISK score. *Setting 3* consists of the most information among all cases where the model prediction, FINDRISK score and an explanation are provided to the physician making it more interpretable. Collectively, *setting 3* has a higher amount of data fed to the physician with specificity in details this provides a better score across all metrics over *setting 2* yet not better than *setting 1*. We assume that due to the broad level of details in *setting 1*, the chances of assigning a higher score is more likely, however having greater details as in *setting 3* may have caused minor disagreements with the results which therefore have been reflected in the scores. The behaviour of the metric scores across levels of data can be seen in Figure 4.

From our simulations, we acquired 96 reward functions for 8 policy combinations over 4 settings of data and 3 metric thresholds. We describe the behaviour and attributes of the reward function as per the metric threshold M_T. When $M_T = 2$, across all 4 settings, we saw the same reward function produced. For policies where two or more Continue actions were performed, the reward was awarded to the concurrent state reached. However, for all other policies and states, the reward acquired was zero. Table 2 describes the reward function acquired for the policies under the threshold.

Table 2. Reward functions of all policies π under a threshold $M_T = 2$ for a all four settings of data

Policy	***End***	***Understandability***	***Agreement***	***Usability***	***Complete***
[0, 0, 1, 1, 0]	0	0	0	0	**10**
[0, 1, 1, 1, 0]	0	0	0	**10**	**10**
[0, 1, 1, 0, 0]	0	0	0	**10**	0
[0, 1, 0, 1, 0]	0	0	0	0	**10**

With a threshold of $M_T = 3$, across all settings, we observed the reward function to follow the behaviour as in the previous threshold scenario indicated in Table 2. However, for settings 2 and 3 we observed an exception in the reward produced for reaching *Understandability* state in the policies where there are 3 continued actions performed. The observed reward function is shown in Table 3

Having $M_T = 4$, unlike in previous threshold scenarios, we observe the reward function to emphasize penalties to reaching a state over rewards. In all settings, 1, 2, 3 and combined where the policy has the action *continue* occurring in states of Agreement and Usability, a negative reward was observed when reaching the End state and a small negative reward when reaching the state of agreement with usability action being *continue*. Table 4 shows the reward function observed.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

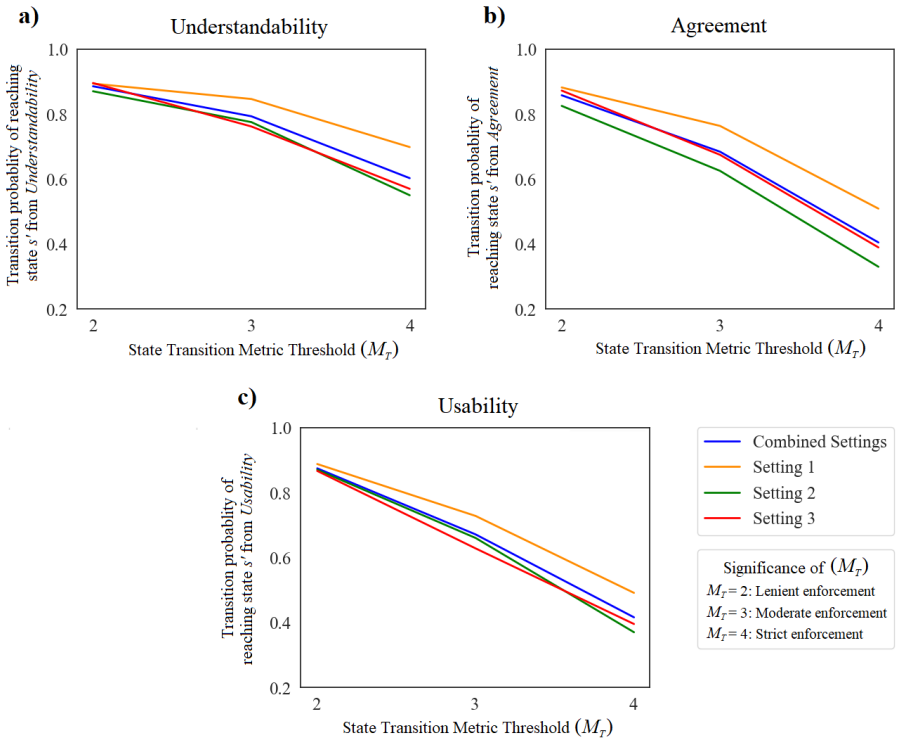

Fig. 4. Observed transition probabilities of trajectories of settings 1, 2, 3 and combination when traversing between the states with metric threshold set to lenient, moderate and strict enforcement of the implicit policy; (a) Transition probability of *Understandability* over thresholds M_T; (b) Transition probability of *Agreement* over thresholds M_T (c) Transition probability of *Usability* over thresholds M_T

Table 3. Reward functions of all policies π under a threshold $M_T = 3$ for a all four settings of data

Policy	***End***	***Understandability***	***Agreement***	***Usability***	***Complete***
[0, 0, 1, 1, 0]	0	0	0	0	**10**
[0, 1, 1, 1, 0]	0	**-10**	0	**10**	**10**
[0, 1, 1, 0, 0]	0	0	0	**10**	0
[0, 1, 0, 1, 0]	0	0	0	0	**10**

Overall, we observe that with a low threshold $M_T = 2$, the reward function is lenient and allows for a broader range of policies to have higher rewards this is due to many trajectories easily overcoming the threshold and therefore having a higher transition rate. With the threshold $M_T = 3$, the reward function can be termed as locally balanced by adding a penalty term to the understandability state when having a complete policy case. This may be attributed to the inter-

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

Table 4. Reward functions of all policies π under a threshold $M_T = 4$ for a all four settings of data

Policy	*End*	*Understandability*	*Agreement*	*Usability*	*Complete*
[0, 0, 0, 0, 0]	**10**	0	0	0	0
[0, 0, 1, 1, 0]	**-10**	0	**-2.1**	0	0
[0, 1, 1, 1, 0]	**-10**	0	0	0	0

pretation of the data i.e. to have data with high usability and agreement, it must be well understandable. When threshold $M_T = 4$, only a few trajectories can qualify to have a high rate of transition, therefore creating a baseline of negative reward for reaching the End state.

From the reward functions, we can interpret a broad perspective of the cognitive state during the evaluation of the patient data by physicians. With the increase in the level of data (i.e. the addition of FINDRISK and explanation to the model prediction) we observe an improved and inclusive reward function that alots rewards for close and full completion while penalties for reaching an End state. The increase in metric threshold M_T is observed to have a diminishing reward incentive with penalties being more prevalent. this may be attributed to the amount of CDSS trajectories that have high ratings provided by the physician and can be further assumed that cases with high scores are more likely to pass through all stages of the MDP.

5 Discussion

In this section we address issues and open topics that we came across during our work.

Perceptional state of AI and Human for collaboration: Modelling the cognitive state of a decision maker (human) can be considered as achieving partial progress towards active human-artificial intelligence (AI) collaboration. The rest of this process lies in crafting the environment, validation metrics and the AI model itself. One of the ways this can be achieved with great accuracy is by trying to interpret the perceptional state of the human mind and the AI model from either perspective [5]. On a general basis, the human decision maker views the AI model (DSS) as an accessory tool providing suggestions, rather than considering it as a capable collaborator in the decision-making process. This can be attributed to the nature of AI models being tediously sensitive or providing recommendations and solutions that are broadly accurate but not specific enough. There is also the factor of trust in AI as it is relatively new to fully formal usage in the real world [15]. Decision-makers often see their recommendations with skepticism and prefer to re-validate solutions to ensure there are no errors, this implicitly brings about another layer of work and scrutiny which decision-makers have to take on, thereby reducing the trust and dependence on AI. Parallelly, from the AI's perspective, there is a bigger dimension of unanswered questions that are not explicitly demanded yet should be investigated

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

to aid the AI model's evolution to provide more aligning solutions to the human decision-maker [12] e.g. if a physician decision maker selects to discharge a patient 2 days earlier than scheduled yet the AI models suggestion was the latter, the reasoning behind such a decision is left explicitly unanswered to the AI model but implicitly answered to the physician. This may cause a loss of understanding from the AI's perspective considering the action taken. Hence, we believe that by using metrics to evaluate decisions in the form of subjective and descriptive measures, while also training the AI model to perceive the possible cognitive state of a human in a situation, the recommendations and alignment of AI models to that of the human decision-maker will greatly improve, therefore positively impacting trust and reliability in AI models.

Modelling the data a greater issue than IRL: During our experiment, we encountered numerous instances where modelling the real-world CDSS data into an MDP was a tedious and gruelling task. Initially, we assumed the possibility of using all patient characteristics, metrics evaluations and settings together in a single MDP, yet the development of such an MDP was unsuccessful as it led to multi-dimensional state spaces with non-comprehensible solutions. As IRL aims to learn the complete policy and behaviour of experts within the trajectories, we can model the MDP features space with nested dimensions or with branched recurrences thereby reducing the computation complexity by a significant level. E.g. within our CDSS data, we tried to model a 125-state space MDP where each state corresponds to an iteratively expanding variable with 5 actions that correspond to the metric evaluation from the physician, yet the layer of patient data here was not included in the state space as it wasn't rational in its solution. In one of our approaches to model a multi-dimensional state space with patient data included, we created a theoretically working state-action feature space MDP, however, when trying to feed the CDSS data into the designed architecture the complexity of data extraction and processing was high. Further, when trying to feed the data to the IRL algorithm, the dimensionality of input tuples was unable to be resolved without splitting and transforming the data into simpler formats (i.e. a four-dimensional state space had to be transformed into a single dimension vector to fit into the tuple format and it may cause it to lose its accurate weightage).

6 Conclusion and future work

The results of our work provide a first-stage implementation of IRL to assess the cognitive state of the human mind using real-world data from a clinical decision support system and its evaluation performed by physicians. We uncover the underlying policies and reward functions using linear programming IRL that explains the cognitive state of the physician during the analysis of patient data and decision support data to predict the chances of the patient acquiring T2DM in the future. We demonstrate our construction of MDP, our approach to performing IRL using RL and investigate the reward functions over a set of policies

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

under various levels of information provided to the physician and the effect of having data thresholds that modulate the reward function.

In the future, we are planning to extend the study into a deeper investigation of physicians' professional behaviour and decision-making. The ultimate goal of the study is to improve human-AI interaction in the CDSS context with better interaction and collaboration between decision-makers and AI agents. One of the ways to achieve this goal is to attain a better understanding of human intentions and decisions via IRL, technology acceptance models, etc. An important future direction is the incorporation of model and human correctness, agreement, and final decision performance metrics into the model. Also, we aim to include more patient data and create a denser MDP feature space to improve the specificity of the reward function obtained along with using diverse IRL algorithms such as deep learning and Bayesian approach [16,13]. We also plan to include newer real-world datasets for various diseases and clinical decisions that can provide more unexplored dimensions of data therefore providing a dynamic representation of the cognitive state during decision-making scenarios.

Acknowledgement. This research is financially supported by The Russian Science Foundation, Agreement #24-11-00272.

References

1. Abbeel, P., Ng, A.Y.: Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the twenty-first international conference on Machine learning. p. 1 (2004)
2. Adams, S., Cody, T., Beling, P.A.: A survey of inverse reinforcement learning. Artificial Intelligence Review **55**(6), 4307–4346 (2022)
3. Alger, M.: Inverse reinforcement learning (2017). `https://doi.org/10.5281/zenodo.555999`, `https://doi.org/10.5281/zenodo.555999`
4. Damacharla, P., Javaid, A.Y., Gallimore, J.J., Devabhaktuni, V.K.: Common metrics to benchmark human-machine teams (hmt): A review. IEEE Access **6**, 38637–38655 (2018)
5. Howes, A., Jokinen, J.P., Oulasvirta, A.: Towards machines that understand people. AI Magazine **44**(3), 312–327 (2023)
6. Kovalchuk, S.V., Kopanitsa, G.D., Derevitskii, I.V., Matveev, G.A., Savitskaya, D.A.: Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. Journal of Biomedical Informatics **127**, 104013 (2022)
7. Lee, K., Rucker, M., Scherer, W.T., Beling, P.A., Gerber, M.S., Kang, H.: Agent-based model construction using inverse reinforcement learning. In: 2017 Winter Simulation Conference (WSC). pp. 1264–1275. IEEE (2017)
8. Liu, Q., Wu, H., Liu, A.: Modeling and interpreting real-world human risk decision making with inverse reinforcement learning. arXiv preprint arXiv:1906.05803 (2019)
9. Muelling, K., Boularias, A., Mohler, B., Schölkopf, B., Peters, J.: Learning strategies in table tennis using inverse reinforcement learning. Biological cybernetics **108**, 603–619 (2014)
10. Ng, A.Y., Russell, S., et al.: Algorithms for inverse reinforcement learning. In: Icml. vol. 1, p. 2 (2000)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20

11. Phan-Minh, T., Howington, F., Chu, T.S., Tomov, M.S., Beaudoin, R.E., Lee, S.U., Li, N., Dicle, C., Findler, S., Suarez-Ruiz, F., Yang, B., Omari, S., Wolff, E.M.: Driveirl: Drive in real life with inverse reinforcement learning. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 1544–1550 (2023). https://doi.org/10.1109/ICRA48891.2023.10160449
12. Pinski, M., Benlian, A.: Ai literacy-towards measuring human competency in artificial intelligence (2023)
13. Ramachandran, D., Amir, E.: Bayesian inverse reinforcement learning. In: IJCAI. vol. 7, pp. 2586–2591 (2007)
14. Swamy, G., Wu, D., Choudhury, S., Bagnell, D., Wu, S.: Inverse reinforcement learning without reinforcement learning. In: International Conference on Machine Learning. pp. 33299–33318. PMLR (2023)
15. Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., Seaborn, K.: Trust in human-ai interaction: Scoping out models, measures, and methods. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–7 (2022)
16. Wulfmeier, M., Ondruska, P., Posner, I.: Deep inverse reinforcement learning. CoRR, abs/1507.04888 (2015)
17. Ziebart, B.D., Maas, A.L., Bagnell, J.A., Dey, A.K., et al.: Maximum entropy inverse reinforcement learning. In: Aaai. vol. 8, pp. 1433–1438. Chicago, IL, USA (2008)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63772-8_20