IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 June 2022, accepted 17 July 2022, date of publication 25 July 2022, date of current version 29 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193494

RESEARCH ARTICLE

Learning and Assessing Optimal Dynamic Treatment Regimes Through Cooperative Imitation Learning

SYED IHTESHAM HUSSAIN SHAH[1,2], ANTONIO CORONATO[1,3], MUDDASAR NAEEM[1], AND GIUSEPPE DE PIETRO[1], (Member, IEEE)

[1]Institute for High Performance Computing and Networking (ICAR), CNR, 80131 Napoli, Italy
[2]Department of Engineering, Università Degli Studi Di Napoli Parthenope, 80143 Napoli, Italy
[3]Department of Computer Science, Università Telematica Giustino Fortunato, 82100 Benevento, Italy

Corresponding author: Muddasar Naeem (muddasar.naeem@icar.cnr.it)

ABSTRACT Dynamic Treatment Regimes (DTRs) are sets of sequential decision rules that can be adapted over time to treat patients with a specific pathology. DTR consists of alternative treatment paths and any of these treatments can be adapted depending on the patient's characteristics. Reinforcement Learning (RL) and Imitation Learning (IL) approaches have been deployed for obtaining optimal treatment for a patient but, these approaches rely only on positive trajectories (i.e., treatments that concluded with positive responses of the patient). In contrast, negative trajectories (i.e., samples of non-responding treatments) are discarded, although these have valuable information content. We propose a Cooperative Imitation Learning (CIL) method that exploits information from both negative and positive trajectories to learn the optimal DTR. The proposed method reduces the chance of selecting any treatment which results in a negative outcome (negative response of the patient) during the medical examination. To validate our approach, we have considered a well-known DTR which is defined for the treatment of patients with alcohol addiction. Results show that our approach outperforms those that rely only on positive trajectories.

INDEX TERMS Inverse reinforcement learning, imitation learning, dynamic treatment regime, reinforcement learning, cooperative imitation learning, Markov decision process.

I. INTRODUCTION

Dynamic Treatment Regimes (DTRs) are sequences of decision rules that implement adaptive treatment strategies. A DTR allows clinicians to personalize the treatment of a patient depending on his/her step-by-step response to the treatment [1], [2]. Nowadays, DTRs is considered part of precision medicine because it determines the choice of treatment for the patient based on his/her condition. In general, DTRs can be designed and assessed by exploiting the Sequential Multiple Assignment Randomized Trial (SMART) method [3], [4] which includes a sequence of observations and treatments.

The associate editor coordinating the review of this manuscript and approving it for publication was Kostas Kolomvatsos.

Figure 1 shows the DTR of patients with alcohol addiction. According to the structure of the DTR and the SMART method, each participant to the experiment is initially associated with one of the available treatments. Treatment is modified based on the patient's response. In the Figure 1, a circle with the letter *R* indicates a *Randomization* stage where patients are given other available treatments.

In the proposed scenario, three basic treatments are contemplated: psychological therapy (PSY), medication therapy (MED) and telephone monitoring (TM). In addition, a combination is also possible, such as Psychological and Medication Therapy (PSY + MED).

Responders are those participants who have reported less than two drinking days over the last two months. In contrast, patients who have reported more drinking days are classified as *Non-Responders*. *Non-Responders* are re-randomized for

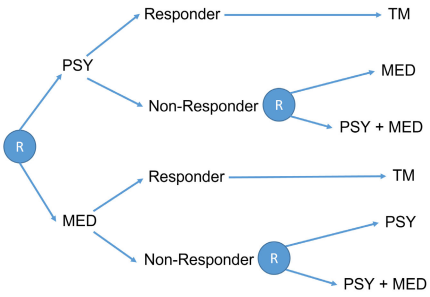

FIGURE 1. A Dynamic Treatment Regime.

the successive stages of the treatment. For example, if a patient has not responded properly to a Psychological treatment (PSY), the next type of treatment will be chosen randomly among the Medical or the combination of Medical and Psychological one. Instead, *Responders* to the first-stage treatment will successively be checked only via Telephone Monitoring (TM).

Application of Machine Learning (ML) techniques in biomedical [5], minimizing medication errors during home treatment [6], [7] vision enhancing scheme for low vision impairments [8], risk management [9], communication [10], and healthcare [11], [12] domains has been increased extensively in recent years.

The structure of a DTR, however, fits naturally into the Reinforcement Learning (RL) [13] problem. In a RL problem, we aim at learning the optimal policy; i.e., the policy that maximizes the cumulative reward. A similar goal is pursued by the clinicians who search for the optimal DTR; i.e., the one that maximizes the response of the patients to the treatment. This can be modeled as a reward function and, thus, the problem of searching the optimal DTR as an RL one.

However, in many RL problems the defined reward function can be very scattered (specially in a large state space) and it can be extremely complicated to identify which actions allow the best end result [14]. Patients with similar conditions may respond differently to the given treatments. Some patients respond while others may not. Reward function, in that case, can be designed under experts (clinicians) guidance to find better treatments for all patients [15].

On the other hand, Inverse Reinforcement Learning (IRL) [16], [17] obtains the reward function from the demonstrations(expert trajectories) given by an expert [18]. It is also the foundation of Learning from Demonstration (LD) [19], which aims at reproducing the demonstrated behaviour.

We define positive trajectories as those episodes that report good results after the therapeutic process and negative trajectories as those episodes that fail to reduce the addiction to alcohol. Usually, these IRL and Imitation Learning (IL) methods consider the positive trajectories only to learn the optimal policy, but it can be intuitively figured out that the information carried by negative trajectories (which has been largely ignored) can help to speed up or improve the learning process.

To address these limitation, we proposed a Cooperative Imitation Learning (CIL) model to learn the optimal policies, where the negative trajectories are also taken into account. The proposed algorithm aims to learn policies π^{θ} that replicate positive trajectories and avoid negative trajectories.

This article is organized as follows. Related work is described in Section II. In particular, a comparison between different methods of IL in DTRs is discussed. Section III presents technical background to Markov Decision Process (MDP), RL, and IRL. Section IV describes the proposed approach for learning the optimal DTR from positive and negative trajectories. Section V mainly concerns with the theoretical analysis of the proposed approach. The discussion about experimental model, data description and results is reported in Section VI. We concludes the paper in Section VII.

II. RELATED WORK

DTR, during to time-varying treatment process where the treatment is frequently tailored to a person's dynamic state, oversimplify personalized medicine [20], [21]. DTRs are also referred as adaptive treatment strategies [22].

A framework to estimate properly defined ''optimal'' DTRs using a time-varying instrumental variable is developed in [23]. Authors have derived a Bellman equation under partial identification and utilized it to define a generic class of estimands known as instrumental variable optimal DTRs. However, the analysis in this paper depends on the assumption of bounded concentration coefficients. It is important to evaluate whether this assumption can be relaxed if additional structural assumptions are imposed. A new Subgoal conditioned hierarchical Imitation Learning framework is proposed in [24]. Authors sequentially set a subgoal for each sub-task using high-level policy and without prior knowledge. Moreover, a self-supervised learning technique is used to learn an effective representation for each subgoal to get rid of prior knowledge.

Two RL-based techniques (Direct Augmented V-Learning and Safe Augmented V-Learning) are proposed in [25]. The performance of the proposed methods has been evaluated using clinical data and synthetic data. However, research work is required to examine the interpretability of the policies that authors have obtained using the proposed two methods to make sure of effective use in practice. Optimal DTRs estimation is done in [26] employing information extraction from the available unstructured clinical text. Authors combine information extraction and optimal DTR estimation to derive patient characteristics and then utilized tree-based RL for estimating multistage optimal DTRs. However, the accuracy of DTR estimation mostly depends on the quality of unstructured clinical notes. The advantage of information extraction may be limited in case of less additional informative content available in the clinical free-text.

Two methods using causal forests and causal trees and are based on a data-driven estimation of heterogeneous treatment

effects are presented in [27]. These methods learn non-linear relationships and control for time-varying confounding. Both models have been evaluated on synthetic data and then utilized real-world data from intensive care units. It would be useful to implement this technique in the operational decision-making of healthcare providers to evaluate the feasibility for patients in the field.

A Privacy-Preserving RL method for DTRs using health data is introduced in [28]. The authors first present computation protocols based on Cheon's approximate homomorphic encryption technique for implementing comparison, exponentiation, maximum, and division and then develop a homomorphic reciprocal of square root protocol that only requires one approximate computation. In the end, an asynchronous advantage actor-critic RL algorithm is designed. Similarly, an outcome-weighed learning method for the decision function in DTRS in a boosting scheme is presented in [29] in line with prediction in supervised learning.

Behavior Cloning (BC) and RL [30], [31] are two techniques exploited to learn DTRs where the first method can be used to recover the clinician's policies provided that the Electronic Health Record (EHR) is optimal and plentiful. BC [32] learns the policy through supervised learning by direct mapping of states to the actions. However, without considerable improvement during training, BC introduces a compounding error [33] over the trajectories length. On the other hand, RL and Deep Reinforcement Learning (DRL) methods are based on maximizing the aggregated reward [14], [30] by directly learning a policy.

To find the best fitting rewards function, the researcher adopted the IRL approaches [34], [35]. IRL learns the reward function by using expert demonstrations. The Maximum likelihood IRL (MLIRL) utilizes an estimate of the gradient of the likelihood function. It states that the likelihood of the data-set may be expressed by the product of the likelihood of the state-action pairs [36]. Researchers have discussed a technique of Maximum Margin Planning (MMP) [37] to learn a reward function that improves the expert policy than alternative policies. Gaussian Process (GP) [38] is adopted in continuous state space to recover both rewards information and uncertainty. Deep GP model [39], mounds many hidden GP layers and it can learn the complex reward patterns even with limited demonstrations. However, unlike the IRL approaches, the preference of the IL [40] is to learn policy by minimizing the Jensen-Shannon divergence between learned policy and expert policy directly [41]. IL approaches to model the intention and preference of the demonstrator.

These existing IL and Adversarial Imitation Learning (AIL) techniques mostly use positive trajectories for learning purposes. The information in the negative trajectories is being ignored which we believe, is important in learning the preference of an agent and avoiding making mistakes.

TABLE 1. Description of related notations.

Notation	*Description*
$s_t \in S$	The set of state which consists of the patient's demographics
$a_t \in A$	What medication is chosen for a patient at time step t.
$T(\acute{s}\|s,a)$	Probability of ending up in state $\acute{s}$ by taking action a in state s. $T \in [0,1]$.
$R(\acute{s}\|s,a)$	Reward due to action a in state s, where $R \in \mathbb{R}$
$0 \leq \gamma \leq 1$	Discount factor $\in [0,1]$. A value of zero gives more weight to immediate rewards and a value close to one gives more weight to long term rewards.
$\pi = \{\pi^+, \pi^-, \pi^\theta\}$	The behavior policy, which consists of positive policies π^+, negative policies π^- and learned policies π^θ
$\tau^+, \tau^-, \tau^\theta$	Trajectories, which are the sequences of pairs $< state_j, action_j >$, being $\tau^i = (s_1^i, a_1^i, s_2^i, a_2^i, \ldots)$. Positive trajectories τ^+, negative trajectories τ^- and learned trajectories τ^θ are associated to π^+, π^- and π^θ respectively.
$\rho_\pi : S \times A \rightarrow \mathbb{R}$	The distribution of state-action pairs when policy π interact with the environment.
D_a	The Adversarial discriminator.
D_c	The Cooperative discriminator.

To address these limitations associated with existing methods, we proposed a CIL model to learn the optimal policies. These learned policies should be similar to positive policies. To achieve this goal, two discriminators (adversarial discriminator AD and cooperative discriminator CD) have been used. AD plays a role in minimizing the difference between positive and learned policies. On the other hand, the CD distinguishes between positive and learned policies from negative policies (detail is presented in section-IV).

III. TECHNICAL BACKGROUND

A. REINFORCEMENT LEARNING (RL)

RL is the domain of ML where the learning process is guided by interactions with the environment, without any prior knowledge, to achieve a goal. The RL algorithms generally satisfy the Markov Decision Process (MDP) based on Markov properties. It does not take into account previous information when taking action in the current state.

MDP is a tuple (S, A, T, R, γ) [42] as described in table 1.

The objective of RL approaches is to learn the behavior of the surrounding environment by repetitive interactions. The **Agent** is the component that interacts with the environment. The agent selects action a_t in state s_t at time t. After, the environment updates its status s_{t+1} and returns a reward r_{t+1}, which may be positive or negative. The Agent aims at learning the optimal **policy** by using a try&error approach. It tries to maximize the cumulative reward, which is provided by the **Value Function** ($V^\pi(s)$). Such a function gives the expected reward for the policy π given the current state $s_t = s$ as in equations 1 and 2:

$$V^\pi(s) = E_\pi\{R_t|s_t = s\} = \sum_{a\in A(s)} \pi(s,a) \sum_{s_{t+1}\in S} T^a_{ss_{t+1}}\{r(s,a) + \gamma V^\pi(s_{t+1})\} \quad (1)$$

being

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots . \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad (2)$$

The optimal policy is denoted as π^*. The state value function (V) for the optimal policy π^* using the Bellman equation is given as in equation 3:

$$V^*(s) = \max_a \sum_{s_{t+1} \in S} T^a_{ss_{t+1}} (R^a_{ss_{t+1}} + \gamma V^*(s_{t+1})) \quad (3)$$

The reward function defined for the very large state space RL problem is usually sparse and imprecise. Under such conditions, IRL has been successfully adopted to automatically estimate the reward function from a set of given trajectories or expert demonstrations.

B. INVERSE REINFORCEMENT LEARNING (IRL)

IRL is about learning the expert knowledge by observing its decisions during the decision-making process [43]. This set of techniques aims at finding a reward function that explains the expert behavior (derived from a series of demonstrations).

Expert demonstrations, called trajectories, are supposed to be part of the optimal policies [44]. A policy may be optimal for several distinct reward functions. Therefore, the objective is to obtain a reward function that best reflects the optimal policy pi^* derived from expert demonstrations.

IRL can be modeled to an optimization problem and consequently Linear Programming (LP) can be employed for the solution [45]. LP method assumes that the expert policy always produce a higher expected value $\mathbb{E}[V^*(\acute{s})]$ than the expected value of any other estimated policy $\mathbb{E}[V^{\pi}(\acute{s})]$ [46]. It is defined as in equation 4.

$$\forall_{s_0 \,\epsilon\, S}, \forall_{a \,\epsilon\, A}: \quad \mathbb{E}_{\acute{s} \backsim T(\acute{s}|s,\pi(s))}\{V^*(\acute{s})\} \geq \mathbb{E}_{\acute{s} \backsim T(\acute{s}|s,a)}\{V^{\pi}(\acute{s})\} \quad (4)$$

LP algorithm aims to maximize the minimum difference between expert values and the estimated values as in equation 5.

$$\max \sum_{s \in S_0} \min_{a \in A|\pi(s)} \{\mathbb{E}_{\acute{s} \backsim T(\acute{s}|s,\pi(s))}\{V^*(\acute{s})\} - \{\mathbb{E}_{\acute{s} \backsim T(\acute{s}|s,a)}\{V^{\pi}(\acute{s})\}\} \quad (5)$$

LP is useful if all parameters of RL environment are known. In many practical applications we do not have complete knowledge of the environment (i.e., transition probabilities are not known) then the max-margin IRL [47] technique can be utilized. Max-margin IRL method assumes that the reward function can be represented as a linear function of known basis Φ_i [48]:

$$R = \sum_i w_i \phi_i. \quad (6)$$

where weight vector $w : (||w||_1 \leq 1)$ minimizes the Euclidean distance $(||\mu(\hat{\pi}) - \mu_E||_2)$ between the expert feature expectation μ_E and the estimated feature expectation $\mu\hat{\pi}$ [48].

$$\max_{w:||w||_2 \leq 1} \min_{j \in \{0..(i-1)\}} (||w^T \mu(\hat{\pi}) - w^T \mu_E||_2) \quad (7)$$

$$\leq ||w^T||_2 \, ||\mu(\hat{\pi}) - \mu_E||_2 \leq \epsilon \quad (8)$$

The value of w is considered optimal when the Euclidean distance $(||\mu(\hat{\pi}) - \mu_E||_2)$ becomes smaller than a predefined threshold value ϵ. The Max-Margin IRL method examines almost all policies to find the best one. In large state space, it is computationally complex to investigate all actions and states.

IV. SYSTEM MODEL

The flowchart of proposed CIL methodology is shown in Figure 2. The patient model emulates the responses of a real patient to treatment. We are given a set of positive and negative trajectories. Trajectories τ are the sequence of state-action pairs such as $\tau = (s_0, a_0, s_1, a_1 \ldots..)$. States represent the response of the patients (i.e., responder, non-responder, etc.) and actions represent the recommended treatments (i.e., MED, PSY, TM, $PSY + MED$, etc.). Positive trajectories consist of the samples that result in a successful outcome (e.g., successful medical treatments). On the other hand, negative trajectories τ^- refer to failures or bad outcomes (e.g, unsuccessful medical treatments). The goal of CIL is to identify the positive trajectories and stay away from the negative trajectories. To achieve this objective two discriminators, the Adversarial Discriminator D_a (AD) and the Cooperative Discriminator D_c (CD), have been used. AD plays a role in minimizing the difference between positive policies and learned policies. On the other hand, the CD distinguishes positive and learned policies from negative policies. Both the discriminators are used to update the Q-table which eventually updates the learned behavior policy π^{θ}.

Positive trajectories τ^+ and negative trajectories τ^-, some time called demonstrations, are generated from the positive behaviour policies π^+ and negative behavior policies π^- respectively, as given below in equations 9 and 10.

$$\begin{aligned} \tau^+ &= [(s_1^{1+}, a_1^{1+}, s_2^{1+}, a_2^{1+}, .., s_d^{1+}, a_d^{1+}), \\ &\qquad (s_1^{2+}, a_1^{2+}, s_2^{2+}, a_2^{2+}, .., s_d^{2+}, a_d^{2+}), \ \ldots] \\ &= [\tau^{1+}, \tau^{2+}, \ldots] \end{aligned} \quad (9)$$

$$\begin{aligned} \tau^- &= [(s_1^{1-}, a_1^{1-}, s_2^{1-}, a_2^{1-}, .., s_d^{1-}, a_d^{1-}), \\ &\qquad (s_1^{2-}, a_1^{2-}, s_2^{2-}, a_2^{2-}, .., s_d^{2-}, a_d^{2-}), \ \ldots] \\ &= [\tau^{1-}, \tau^{2-}, \ldots] \end{aligned} \quad (10)$$

Occupancy measure $\rho_{\pi}: S \times A \rightarrow \mathbb{R}$ for a policy $\pi \in \Pi$ is defined as:

$$\rho_{\pi}(s, a) = \pi(a \mid s) \sum_{t=0}^{T} \gamma P(s_t = s \mid \pi) \quad (11)$$

It represents the distribution of state-action pairs by following a policy π with discount factor γ. State $s_t \in S$ represents the patient's condition, while action a_t represents the recommended medication. Practically, we compare the difference between positive behavior policies π^+ and learned

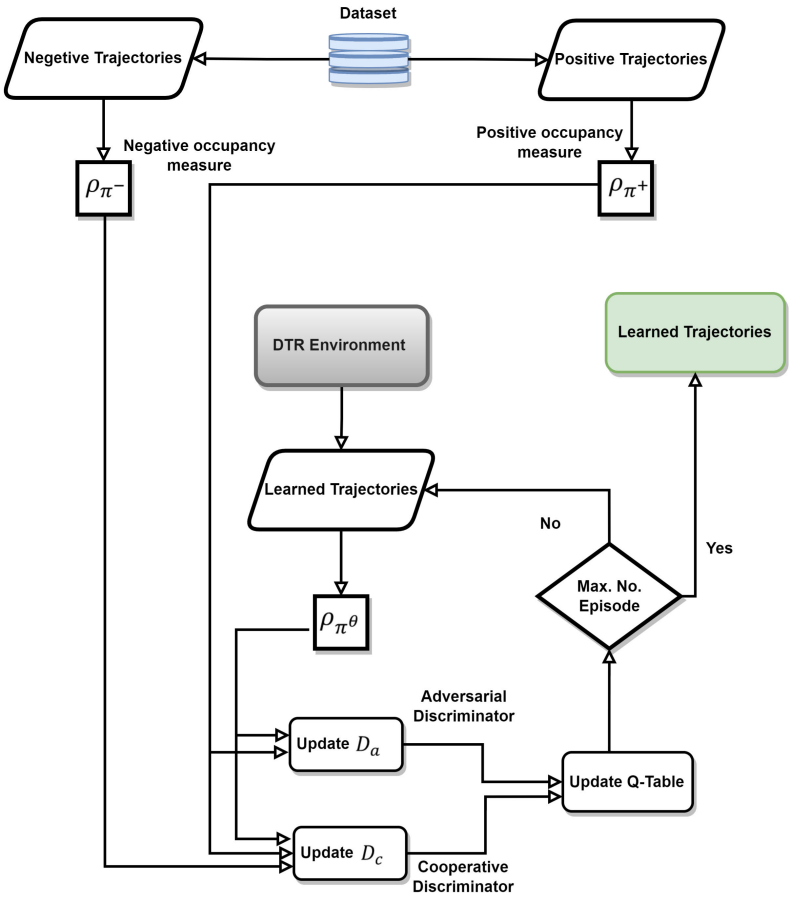

FIGURE 2. Flow chart of proposed model for DTR.

policies π^θ by their generated trajectories. D_a: $S \times A \rightarrow (0, 1)$ is known as adversarial discriminator. It estimates the probabilities that an action on a state belongs to the positive policies π^+, rather than learned policies π^θ. It can be updated according to:

$$\max_{D_a}(\mathbb{E}_{\rho\pi^+}[\log(D_a(s,a))] + \mathbb{E}_{\rho\pi^\theta}[\log(1 - D_a(s,a))]) \quad (12)$$

Optimizing D_a and π^θ are opposite goals. D_a minimizes the transition probabilities of state-action pairs that are generated by π^θ. On the other hand, π^θ aims at maximizing the transition probabilities of D_a by making a mistake. The Cooperative Discriminator D_c: $S \times A \rightarrow (0, 1)$, instead, differentiates the positive policies π^+ and the learned policies π^θ from the negative ones π^-. The objective function can be defined as:

$$\max_{D_c}(\mathbb{E}_{\rho\pi^+,\rho\pi^\theta}[\log(D_c(s,a))] + \mathbb{E}_{\rho\pi^-}[\log(1 - D_c(s,a))]) \quad (13)$$

This function is naturally cooperative because the objectives of D_c and π^θ are the same. Both are trying to maximize the probability that learned trajectories be similar to the positive trajectories. Adversarial and cooperative discriminators are used to refine the learning policy π^θ. Objective functions of D_a can be represented as a minimization problem of Jensen–Shannon(JS) divergence $D_{JS}(\rho_{\pi^\theta} \parallel \rho_{\pi^+})$ between ρ_{π^θ}, ρ_{π^+}. Instead, objective functions of D_c can be represented as a maximization $D_{JS}(\rho_{\pi^\theta} + \rho_{\pi^+} \parallel \rho_{\pi^-})$, of Jensen–Shannon(JS) divergence between ρ_{π^θ}, ρ_{π^+} and ρ_{π^-} [49]. Given a fixed policy π^θ, optimal discriminators D_a^* and D_c^* can be defined as:

$$D_a^*(s,a) = \frac{\rho_{\pi^+}(s,a)}{\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a)} \quad (14)$$

$$D_c^*(s,a) = \frac{\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a)}{\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a) + \rho_{\pi^-}(s,a)} \quad (15)$$

The value of an action (e.g., medical treatment) in a particular state (e.g., patient's condition) is represented by the Q-value as:

$$Q(s,a) = \hat{\mathbb{E}}_{\tau_i^\theta}[\omega_\alpha \log(D_a(s,a)) + \omega_\beta \log(D_c(s,a))] \quad (16)$$

where $\omega_\alpha \in [0, 1]$ and $\omega_\beta \in [0, 1]$ are the balancing factors that control the importance of the Adversarial and Cooperative discriminators. The aim of updating learned policy π^θ is to learn the positive trajectories τ^+, while staying

away from negative ones τ^-. The Q-value plays an important role in finding this objective because the higher the Q-value in a state, the more likely this action will be chosen by the agent.

$$\pi^\theta = \arg\max_a Q(s,a) \quad \forall s \in S \tag{17}$$

Under this setting, the learned policy is updated where D_a push π^θ close to π^+, while D_c separates π^θ from π^-.

V. THEORETICAL ANALYSIS

In this section we provide the theoretical analysis of convergence of learned policy. Given enough time and capacity, we would like the learned policy π^θ to converge to the positive distribution ρ_{π^+}. The learned policy π^θ represents the probability distribution ρ_{π^θ} of state-action pairs. This analysis is mainly based on the theorems presented in [49].

The objective function for proposed algorithm can be defined as:

$$\begin{aligned} &J_{\pi_\theta, D_a, D_c} \\ &= \mathbb{E}_{\rho\pi^\theta}[\log(1 - D_a(s,a))] + \mathbb{E}_{\rho\pi^+}[\log(D_a(s,a))] \\ &\quad - \mathbb{E}_{\rho\pi^+, \rho\pi^\theta}[\log(D_c(s,a))] - \mathbb{E}_{\rho\pi^-}[\log(1 - D_c(s,a))] \end{aligned} \tag{18}$$

For simplicity, balancing factors ω_α, ω_β (defined in equation 16) are set to 1. The objective function can be written as follows:

$$\begin{aligned} J_{\pi_\theta, D_a, D_c} = [\int_{s,a} &[\rho_{\pi^+} \log(D_a(s,a)) + \rho_{\pi^\theta} \log(1 - D_a(s,a)) \\ &- (\rho_{\pi^+} + \rho_{\pi^\theta}) \log(D_c(s,a)) \\ &- \rho_{\pi^-} \log(1 - D_c(s,a))] dsda] \end{aligned} \tag{19}$$

To find the optimal values of adversarial and cooperative discriminators (D_a^*, D_c^*), we maximize the function inside integral w.r.t D_a and D_c. Setting the derivative w.r.t D_a and D_c to 0 yields the following results:

$$\frac{\rho_{\pi^+}(s,a)}{D_a} - (\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a)) = 0 \tag{20}$$

$$\begin{aligned} &\frac{\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a)}{D_c} - (\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a) \\ &\quad + \rho_{\pi^-}(s,a)) = 0 \end{aligned} \tag{21}$$

Thus it can be verified from above equation that:

$$D_a^*(s,a) = \frac{\rho_{\pi^+}(s,a)}{\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a)} \tag{22}$$

$$D_c^*(s,a) = \frac{\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a)}{\rho_{\pi^+}(s,a) + \rho_{\pi^\theta}(s,a) + \rho_{\pi^-}(s,a)} \tag{23}$$

We have obtained the maximum solution as the second derivation: $-\frac{\rho_{\pi^+}(s,a)}{D_a^2}$ and $-\frac{\rho_{\pi^+}(s,a)+\rho_{\pi^\theta}(s,a)}{D_c^2}$ are non-positive.

Occupancy measure $\rho_\pi(s,a)$ indicates the distribution of state-action pairs when the agent interacts with the environment under the policy π. By inserting the values of D_a^* and D_c^* in equation 18, the proposed algorithm minimizes the following imitation learning algorithm,

$$\min_{\pi^\theta} D_{JS}(\rho_{\pi^+} \parallel \rho_{\pi^\theta}) - D_{JS}(\rho_{\pi^\theta} + \rho_{\pi^+} \parallel \rho_{\pi^-}) \tag{24}$$

It represents the minimization of JS divergence between probability distributions which encourages the learned trajectories to replicate the positive trajectories and stay away from the negative ones.

The proposed algorithm exploits the same concept as described in Adversarial Cooperative Imitation Learning (ACIL). In the ACIL approach, parameters of D_a and D_s are updated by taking the expectations of gradient values over the trajectories (τ^+, τ^- and τ^θ) as given bellow:

$$\hat{\mathbb{E}}_{\tau^+}[\nabla \log(D_a(s,a))] + \hat{\mathbb{E}}_{\tau^\theta}[\nabla \log(1 - D_a(s,a))] \tag{25}$$

$$\hat{\mathbb{E}}_{\tau^+, \tau^\theta}[\nabla \log(D_c(s,a))] + \hat{\mathbb{E}}_{\tau^-}[\nabla \log(1 - D_c(s,a))] \tag{26}$$

In contrast, in our proposed algorithm, we update the parameters of D_a and D_c by finding the immediate optimal values of occupancy measures ρ_π as given in algorithm-1. Updating D_a and D_c in such a way makes our approach computationally and mathematically simpler than the ACIL approach.

Another difference that makes our approach simpler and faster, during the process of learning optimal dynamic treatment regimes, is the method of updating learned trajectories. In the ACIL method, learned policies are updated through the Trust Region Policy Optimization (TRPO) [50] approach. TRPO tries to update the learned trajectories iteratively by solving a trust region optimization problem. On the other hand, in the proposed algorithm we used the current maximum Q-value to update the learned policies as mentioned in equation 17 which makes the proposed algorithm comparatively simple.

VI. EXPERIMENT

In this section, we present the results of experiments that have been conducted to evaluate the proposed model.

Firstly, we describe the dataset and test. Next, we present results and validation.

A. MODEL SETUP AND DATASET DESCRIPTION

The DTR under examination (Figure 1) is a two stage decision process. We have completed and mapped such a process into the MDP Model as shown in Figure 3.

Each patient is assigned to one of two possible initial treatments: psychology (*PSY*) or medicine (*MED*) randomly during the trial. As already described in section I, participants are classified as responders (*Res*) or non-responders (*NR*) in response to the first treatment depending on whether participants do (or do not) have had more than two heavy-drinking days over the last two months.

We have built a model of the patient to emulate his/her behavior, as a responder or non-responder, that is described by probability distributions and depends on some patient characteristics such as gender, age, cultural level, and

Algorithm 1 CIL Algorithm for Optimal DTR

Given: Data set, discount factor γ, termination criteria,
Initialize: Input Parameter, number of iteration T, Positive trajectories τ^+, Negative trajectories τ^- generated by behavior policies π^+ and π^- respectively, length of trajectories N, number of trajectories M.
Estimate occupancy measures ρ_{π^+}, ρ_{π^-} where $\rho_\pi(s,a) = \pi(a|s)\sum_M \gamma P(s_t = s|\pi)$
Assign random values to Q_Table, $Q(s,a) = rand()\ \forall$ states and actions
for $(i = 1;\ i \leq No.\ of\ iteration;\ i^{++})$ **do**
  Generate τ^θ by $\pi^\theta = \arg\max_a Q(s,a)\ \forall s \in S$
  Estimate $\rho_{\pi^\theta}(s,a) = \pi(a|s)\sum_M \gamma P(s_t = s|\pi)$
  Adversarial Discriminator $D_a(s,a) = \frac{\rho_{\pi^+}}{\rho_{\pi^+} + \rho_{\pi^\theta}}$
  Cooperative Discriminator $D_c(s,a) = \frac{\rho_{\pi^+} + \rho_{\pi^\theta}}{\rho_{\pi^+} + \rho_{\pi^\theta} + \rho_{\pi^-}}$
  Update $Q(s,a) = \mathbb{E}_{\tau^\theta}[\omega_\alpha\ \log(D_a(s,a)) + \omega_\beta\ \log(D_c(s,a))]$ with balancing terms $\omega_\alpha, \omega_\beta \in [0,1]$
end for
Return: τ^θ and Q-Table.

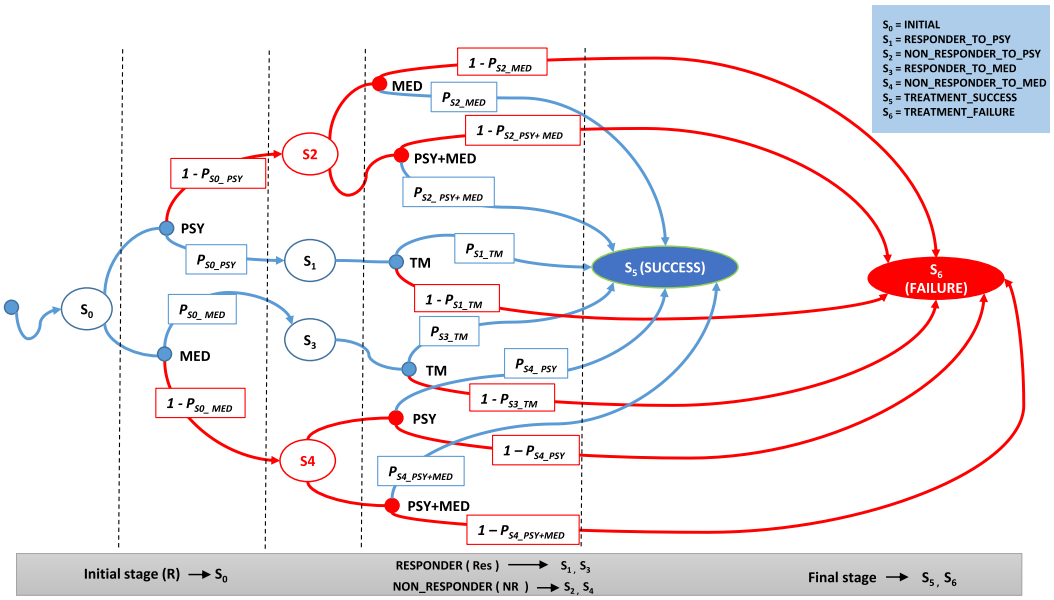

FIGURE 3. A Markov Decision Process mapping the DTR.

neighborhood of residence. Any observable trajectory of the dataset is encoded in sequences of data such as: $(O_1, A_1, O_2, A_2, O_3)$. Where O_1, O_2, and O_3 are the pre-treatment information, intermediate outcomes and final outcomes respectively. $A1$ and $A2$ are the randomized treatments. In the addiction management study, for example, O_1 may include comorbidity, gender, age, level of addition, etc. While O_2 may contains the participant's binary response (Res, NR), side effects, and adherence to the initial treatment. Similarly, O_3 could be the number of non-heavy-drinking days over an under-observed period. On the other hand, $A1$ can be PSY or MED and $A2$ can be one of the TM, PSY, MED and $PSY + MED$ treatments.

The model, shown in figure 3, consists of seven states $\{S_0, \ldots, S_6\}$ and four types of treatments $\{'MED', 'PSY', 'TM', 'PSY + MED'\}$ that a clinician or an artificial agent can select. Probability distributions can be represented as $P_{(S_X \| ACTION)}$, where P is the probability that the patient is 'Responder', S_X is the current state and $ACTION$ is the selected treatment.

Patient's behavior and their responses to the treatments are dependent on some pre-treatment information as reported in Table 2: Recommended treatments at fist stage (e.g PSY and MED) are dependent on the following pre-treatment features:

- Location: $\{downtown(d), hill - station(o)\}$. It represents the neighborhood of residence.
- Gender= $\{Male(M), Female(F)\}$. It represents the the sex of the patient.
- Cultural-Level= $\{high(h), medium(m), low(l)\}$. It represents and socio-cultural level of patient.

Whereas, the treatment selected at the second stage depends on the kind of response of the patient (i.e Responder, non-responder) to the previous treatment.

TABLE 2. Features of (SMART) design schematic for DTR.

Features	*Stages*
Pre-treatment Features	*First Stage*
Location, Gender, Cultural Level	PSY , MED
First Response Features	*Second Stage*
Responder, Non-responder	TM, MED, PSY, MED+PSY

TABLE 3. Recovered trajectories through CIL.

Trajectories	Pre-treatment Features {M-O-L}				
	State	*Action*	*State*	*Action*	*Final − State*
Positive Trajectories	S_0	MED	S_3	TM	$Success$
	S_0	MED	S_4	$PSY + MED$	$Success$
	S_0	MED	S_3	TM	$Success$
	S_0	MED	S_3	TM	$Success$
	S_0	MED	S_4	$PSY + MED$	$Success$
Negative Trajectories	S_0	PSY	S_1	TM	$Failure$
	S_0	PSY	S_2	$PSY + MED$	$Failure$
	S_0	PSY	S_1	TM	$Failure$
	S_0	MED	S_3	TM	$Failure$
	S_0	MED	S_4	$PSY + MED$	$Failure$
Learned Trajectories	S_0	MED	S_3	TM	$Success$
	S_0	MED	S_3	TM	$Success$
	S_0	MED	S_3	TM	$Success$
	S_0	MED	S_4	$PSY + MED$	$Success$
	S_0	MED	S_3	TM	$Success$
	S_0	MED	S_3	TM	$Success$

In order to validate the proposed approach, we have generated a dataset by adopting the SMART method [3], [51], [52]. For different pre-treatment parameters, we have generated 500 positive trajectories (τ^+) that ends as a successful treatment (i.e the patient is reponder) and 500 negative trajectories (τ^-) which represents the failure of the treatments (i.e the patient is non-reponder). Both trajectories are used in the proposed algorithm to guide the RL agent to learn the optimal trajectory. An example of positive trajectories with trajectory length $l = 3$ is $\tau^+ = [S_0, MED, S_2, TM, Success]$.

B. RESULT

This subsection intends to present the outcomes of experimental setup that was used during experiments. In addition, we have also compared the result of our scheme to other well known methods such as: Optimism in the Face of Uncertainty (OFU-DTR) [53], MDP based approach [54], and Q-learning based method [55].

A patient may have different pre-treatment parameters and accordingly we may have varying dynamic treatment regimes. Similarly, there are positive and negative trajectories against each combination of pre-treatment parameters in the dataset. We have conducted experiments and validated the proposed methodology on them. The goal of the proposed algorithm is to mimic the positive trajectories and avoid the negative ones.

An example of recovered trajectories is shown in Table 3. *States* represent the patient responses while *Actions* represent the recommended treatments. We assumed that $M - O - L$ (i.e, {Gender = Male (M); Residence = Hill-station (O); and Cultural level = Low (L)}) are the pre-treatment features.

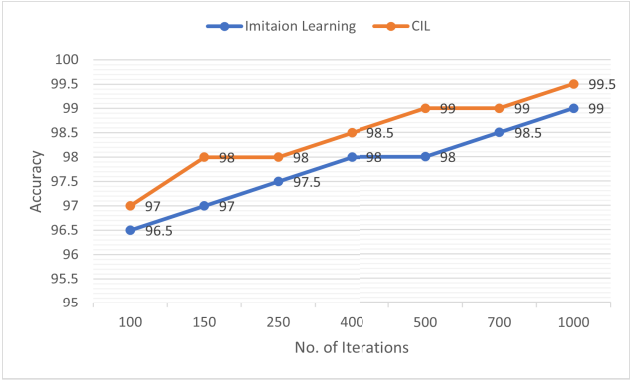

FIGURE 4. Accuracies comparison.

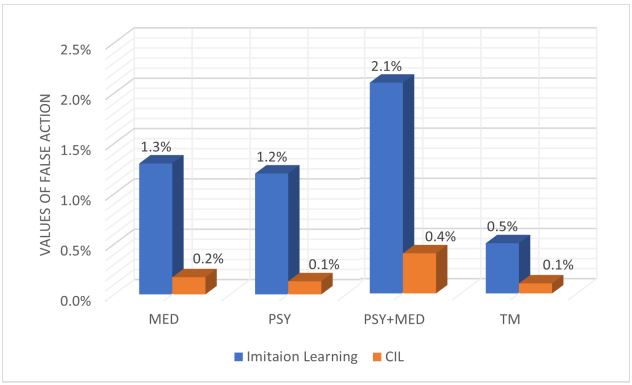

FIGURE 5. Imitation learning in uncertain environment.

Under this consideration, a small batch of positive and negative trajectories is collected as shown in Table 3. The final state in case of positive trajectories is "success", while it is "failure" in case of negative trajectories. Learned trajectories in Table 3 represent the recommended treatments at each state by executing the CIL algorithm. It can be seen that learned trajectories have the same sequence of state-action pairs as positive trajectories and are different from negative trajectories.

We have used accuracy as a performance metric to indicate the ratio of correct selection of a treatment in a given state. Initially, we experimented only with positive trajectories and later we have performed experiments by both using positive and negative trajectories. The result of obtained for accuracy parameter against both schemes is shown in Figure 4. It is evident from Figure 4 that the accuracy of the proposed algorithm is better.

Furthermore, to mitigate the problem of overfitting i.e. the repetition of the same outcome, we introduced 15% of uncertainty to the environment. This is to say that if *PSY* treatment is recommended at state s_0 (refer to Figure 3), then the probability that the next state will be s_1 or s_2 is 0.85 while a probability of 0.15 that next state will be s_3 or s_4.

Results of both Imitation Learning (with only positive trajectories) and CIL in an uncertain environment are shown in the Figure 5. Patients can be divided into different groups

TABLE 4. Comparison in terms of accuracy.

S/No	Approaches	No. of episode	Accuracy
1	OFU-DTR [53]	200	93 %
2	PSRL [54]	200	92 %
3	Q-learning [55]	200	90 %
4	Proposed Algorithm	200	98 %

based on the pre-treatment features. Some treatments, that are recommended by observing patient's responses, are considered positive treatments for a group of patients and these may be considered negative treatments to the other groups of patients. The bars in figure-5 represents a comparison of how many times the negative treatment, which may lead to the negative outcome, is proposed by each of the algorithms. In other words, graph present the errors and mistakes the RL agent made each time while proposing a treatment.

On 1000 trials, Imitation Learning algorithm (where only positive trajectories are used as input) proposed 13 times 'MED' treatments that were negative treatments for a chosen group of patients. On the other hand, in CIL case, this error is repeated only twice. Similarly, the number of times negative treatments were recommended by both the algorithm are shown in the graph. Among all, 'PSY+MED' is the negative treatment that is proposed most of the time in each case. For all groups of patients, CIL algorithm made fewer mistakes in recommending negative treatments as compared to the Imitation Learning algorithm.

Moreover, we present a comparison in terms of accuracy in Table-4 to demonstrate the effectiveness of the proposed approach to the existing work. In [53] authors introduced an algorithm of Optimism in the Face of Uncertainty (OFU-DTR). The learning process is comprised of optimistic planning, policy execution, and model updating. At 200 number of episode, OFU-DTR achieved 93% accuracy. While Posterior Sampling for Reinforcement Learning (PSRL) [54] updates a prior distribution over MDP and selects one sample from this posterior. It achieved 92% accuracy. In [55] authors have employed Q-learning to determine the optimal strategy. In an uncertain DTR environment, it achieved 90 % accuracy. On the other hand, the proposed algorithm achieved 98 % accuracy over 200 number of the episode which is the highest among all.

It can be seen that the use of negative trajectories has not only shown superior performance but also can prevent the agent from making false or negative actions (treatments) and positive trajectories guide the agent to take good actions. This is why the use of CIL is fruitful in an uncertain environment.

VII. CONCLUSION

W have presented CIL method to learn the optimal dynamic treatment regime by exploiting information from both trajectories (positive and negative). The adversarial discriminator is responsible for minimization of discrepancies between positive trajectories (e.g. survived patients) and learned trajectories while the cooperative discriminator is used to separate the learned trajectories from negative trajectories (e.g. deceased patients). Adversarial discriminator and cooperative discriminator play an important role in updating Q-table and hence improving learned trajectories.

The proposed algorithm performs better in a case where we do not have the complete knowledge of the environment (e.g reward function, transition probabilities). Experiments have demonstrated that the proposed algorithm provides better dynamic treatment regimes for people with alcohol addiction.

LIST OF ACRONYMS

AI	Artificial Intelligence
MDP	Markov Decision Process
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
DTR	Dynamic Treatment Regime
IRL	Inverse Reinforcement Learning
IL	Imitation Learning
LP	Linear Programming
AIL	Adversarial Imitation Learning
CIL	Cooperative Imitation Learning
AD	Adversarial Discriminator D_a
CD	Cooperative Discriminator D_c
LD	Learning from Demonstration
ML	Machine Learning
BC	Behavior Cloning
MLIRL	Maximum likelihood IRL
GP	Gaussian Process
MMP	Maximum Margin Planning
SMART	Sequential Multiple Assignment Randomized Trial

REFERENCES

[1] B. Chakraborty and S. A. Murphy, "Dynamic treatment regimes," *Annu. Rev. Statist. Appl.*, vol. 1, pp. 447–464, Jan. 2014.
[2] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," *Artif. Intell. Med.*, vol. 109, Sep. 2020, Art. no. 101964.
[3] S. A. Murphy, "An experimental design for the development of adaptive treatment strategies," *Statist. Med.*, vol. 24, no. 10, pp. 1455–1481, May 2005.
[4] M. Bakhouya, R. Campbell, A. Coronato, G. D. Pietro, and A. Ranganathan, "Introduction to special section on formal methods in pervasive computing," *ACM Trans. Auton. Adapt. Syst.*, vol. 7, no. 1, pp. 1–9, 2012.
[5] S. S. Deepika and T. V. Geetha, "Pattern-based bootstrapping framework for biomedical relation extraction," *Eng. Appl. Artif. Intell.*, vol. 99, Mar. 2021, Art. no. 104130.
[6] M. Ciampi, A. Coronato, M. Naeem, and S. Silvestri, "An intelligent environment for preventing medication errors in home treatment," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116434.
[7] M. Naeem and A. Coronato, "An AI-empowered home-infrastructure to minimize medication errors," *J. Sensor Actuator Netw.*, vol. 11, no. 1, p. 13, Feb. 2022.
[8] C. Lodato and P. Ribino, "A novel vision-enhancing technology for low-vision impairments," *J. Med. Syst.*, vol. 42, no. 12, pp. 1–13, Dec. 2018.
[9] G. Paragliola and M. Naeem, "Risk management for nuclear medical department using reinforcement learning algorithms," *J. Reliable Intell. Environ.*, vol. 5, no. 2, pp. 105–113, Jul. 2019.
[10] M. Naeem, G. De Pietro, and A. Coronato, "Application of reinforcement learning and deep learning in multiple-input and multiple-output (MIMO) systems," *Sensors*, vol. 22, no. 1, p. 309, Dec. 2021.
[11] M. Cinque, A. Coronato, and A. Testa, "A failure modes and effects analysis of mobile health monitoring systems," in *Proc. Innov. Adv. Comput., Inf., Syst. Sci., Eng.* Cham, Switzerland: Springer, 2013, pp. 569–582.

[12] M. Naeem, G. Paragliola, and A. Coronato, "A reinforcement learning and deep learning based intelligent system for the support of impaired patients in home treatment," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114285.
[13] M. Naeem, S. T. H. Rizvi, and A. Coronato, "A gentle introduction to reinforcement learning and its application in different fields," *IEEE Access*, vol. 8, pp. 209320–209344, 2020.
[14] A. Raghu, M. Komorowski, L. Anthony Celi, P. Szolovits, and M. Ghassemi, "Continuous state-space models for optimal sepsis treatment—A deep reinforcement learning approach," 2017, *arXiv:1705.08422*.
[15] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi, "Deep reinforcement learning for sepsis treatment," 2017, *arXiv:1711.09602*.
[16] S. Russell, "Learning agents for uncertain environments," in *Proc. 11th Annu. Conf. Comput. Learn.*, 1998, pp. 101–103.
[17] B. Michini and J. P. How, *Bayesian Nonparametric Inverse Reinforcement Learning* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7524. Springer, 2012, pp. 148–163. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-33486-3_10
[18] S. Zhifei and E. M. Joo, "A survey of inverse reinforcement learning techniques," *Int. J. Intell. Comput. Cybern.*, vol. 5, no. 3, pp. 293–311, 2012.
[19] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, May 2009.
[20] S. A. Murphy, "Optimal dynamic treatment regimes," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 65, no. 2, pp. 331–355, May 2003.
[21] J. M. Robins, "Optimal structural nested models for optimal sequential decisions," in *Proc. 2nd Seattle Symp. Biostatistics*. Cham, Switzerland: Springer, 2004, pp. 189–326.
[22] A. Oetting, J. Levy, R. Weiss, and S. Murphy, "Statistical methodology for a smart design in the development of adaptive treatment strategies," *Causality Psychopathol., Finding Determinants Disorders Cures*, vol. 8, pp. 179–205, Jan. 2011.
[23] S. Chen and B. Zhang, "Estimating and improving dynamic treatment regimes with a time-varying instrumental variable," 2021, *arXiv:2104.07822*.
[24] L. Wang, R. Tang, X. He, and X. He, "Hierarchical imitation learning via subgoal representation learning for dynamic treatment recommendation," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, Feb. 2022, pp. 1081–1089.
[25] S. Saghafian, "Ambiguous dynamic treatment regimes: A reinforcement learning approach," 2021, *arXiv:2112.04571*.
[26] N. Zhou, R. D. Brook, I. D. Dinov, and L. Wang, "Optimal dynamic treatment regime estimation using information extraction from unstructured clinical text," *Biometrical J.*, vol. 64, no. 4, pp. 805–817, Apr. 2022.
[27] T. Blümlein, J. Persson, and S. Feuerriegel, "Learning optimal dynamic treatment regimes using causal tree methods in medicine," 2022, *arXiv:2204.07124*.
[28] X. Sun, Z. Sun, T. Wang, J. Feng, J. Wei, and G. Hu, "A privacy-preserving reinforcement learning approach for dynamic treatment regimes on health data," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–16, Nov. 2021.
[29] S. Eguchi and O. Komori, "Outcome weighted learning in dynamic treatment regimes," in *Minimum Divergence Methods in Statistical Machine Learning*. Cham, Switzerland: Springer, 2022, pp. 197–216.
[30] L. Wang, W. Zhang, X. He, and H. Zha, "Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2447–2456.
[31] Y. Zhang, R. Chen, J. Tang, W. F. Stewart, and J. Sun, "LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1315–1324.
[32] D. A. Pomerleau, "Efficient training of artificial neural networks for autonomous navigation," *Neural Comput.*, vol. 3, no. 1, pp. 88–97, May 1991.
[33] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 627–635.
[34] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 1.
[35] S. Zhifei and E. Meng Joo, "A survey of inverse reinforcement learning techniques," *Int. J. Intell. Comput. Cybern.*, vol. 5, no. 3, pp. 293–311, Aug. 2012.
[36] H. Ratia, L. Montesano, and R. Martinez-Cantin, "On the performance of maximum likelihood inverse reinforcement learning," 2012, *arXiv:1202.1558*.
[37] J. A. Bagnell, N. Ratliff, and M. Zinkevich, "Maximum margin planning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 729–736.
[38] Z.-J. Jin, H. Qian, and M.-L. Zhu, "Gaussian processes in inverse reinforcement learning," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 1, Jul. 2010, pp. 225–230.
[39] M. Jin, A. Damianou, P. Abbeel, and C. Spanos, "Inverse reinforcement learning via deep Gaussian process," 2015, *arXiv:1512.08065*.
[40] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4565–4573.
[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[42] R. S. Sutton, A. G. Barto, and H. Klopf, *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press, 2016.
[43] S. I. H. Shah and G. De Pietro, "An overview of inverse reinforcement learning techniques," in *Proc. 17th Int. Conf. Intell. Environ.*, vol. 29, 2021, p. 202.
[44] S. I. H. Shah and A. Coronato, "Learning tasks in intelligent environments via inverse reinforcement learning," in *Proc. 17th Int. Conf. Intell. Environ. (IE)*, Jun. 2021, pp. 1–4.
[45] A. Y. Ng, "Algorithms for inverse reinforcement learning," in *Proc. ICML*, vol. 1, 2000, p. 2.
[46] S. Sharifzadeh, I. Chiotellis, R. Triebel, and D. Cremers, "Learning to drive using inverse reinforcement learning and deep Q-networks," 2016, *arXiv:1612.03653*.
[47] D. Choi, T.-H. An, K. Ahn, and J. Choi, "Future trajectory prediction via RNN and maximum margin inverse reinforcement learning," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 125–130.
[48] S. I. H. Shah and A. Coronato, "Inverse reinforcement learning through max-margin algorithm," in *Proc. 17th Int. Conf. Intell. Environ.*, vol. 29, 2021, p. 190.
[49] L. Wang, W. Yu, X. He, W. Cheng, M. R. Ren, W. Wang, B. Zong, H. Chen, and H. Zha, "Adversarial cooperative imitation learning for dynamic treatment regimes," in *Proc. Web Conf.*, Apr. 2020, pp. 1785–1795.
[50] Q. Shen, Y. Li, H. Jiang, Z. Wang, and T. Zhao, "Deep reinforcement learning with robust and smooth policy," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8707–8718.
[51] P. W. Lavori and R. Dawson, "A design for testing clinical strategies: Biased adaptive within-subject randomization," *J. Roy. Stat. Soc. A, Statist. Soc.*, vol. 163, no. 1, pp. 29–38, Jan. 2000.
[52] P. W. Lavori and R. Dawson, "Dynamic treatment regimes: Practical design considerations," *Clin. Trials*, vol. 1, no. 1, pp. 9–20, Feb. 2004.
[53] J. Zhang, "Designing optimal dynamic treatment regimes: A causal reinforcement learning approach," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11012–11022.
[54] I. Osband, D. Russo, and B. Van Roy, "(More) efficient reinforcement learning via posterior sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/6a5889bb0190d0211a991f47bb19a777-Abstract.html
[55] E. E. M. Moodie, B. Chakraborty, and M. S. Kramer, "Q-learning for estimating optimal dynamic treatment rules from observational data," *Can. J. Statist.*, vol. 40, no. 4, pp. 629–645, Dec. 2012.

SYED IHTESHAM HUSSAIN SHAH received M.S. degree in electrical engineering (specialization in signal processing) from International Islamic University, Pakistan. He is currently pursuing the Ph.D. degree with University of Naples Parthenope, Italy. He is currently working with the Institute for High Performance Computing and Networking (ICAR), National Research Council of Italy. He is also a Visiting Researcher at the Eindhoven University of Technology, Netherlands. His research interests include developing machine learning techniques (reinforcement learning, inverse-reinforcement learning, and imitation learning) and their application in the biomedical and interdisciplinary domains.

ANTONIO CORONATO is currently the Head of the Research Group on Self-Learning and Self-Adaptive Systems, Institute for High Performance Computing and Networking, National Research Council of Italy. He is also the Co-Founder and the Co-Editor-in-Chief of the *Journal of Reliable Intelligent Environments*. He has participated to the organization of many international workshops and conferences, including several editions of Intelligent Environments. He has also edited Special Issues of *ACM Transactions on Autonomous Systems*, *ACM Transactions on Embedded Systems*, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS and other journals.

MUDDASAR NAEEM received the bachelor's degree in electronics from Quaid-i-Azam University Islamabad, Pakistan, in 2010, the master's degree in telecommunication and networks from Iqra University Islamabad, Pakistan, in 2015, and the Ph.D. degree in information and communication technology and engineering from the University of Naples Parthenope, Naples, Italy, in 2021. He was worked as a Visiting Lecturer at the Department of Electronics, Institute of Information Technology, Quaid-i-Azam University Islamabad. Since 2018, he has been serving as a Researcher at the Institute for High Performance Computing and Networking, National Researcher Council of Italy, Naples. His current research interests include application of AI methods, in particular reinforcement learning in healthcare, and communication.

GIUSEPPE DE PIETRO (Member, IEEE) is currently the Director of the Institute of High Performance Computing and Networking (ICAR), CNR. He is also an Adjunct Professor with the College of Science and Technology, Temple University. He has been actively involved in many European and national projects, with industrial cooperations too. He has authored over 150 scientific papers published in international journals and conferences and he is involved in many program committees and journal editorial boards. His current research interests include pervasive and mobile computing, clinical decision support systems, and software architectures for e-health. He is an International Member of KES.

Open Access funding provided by 'Consiglio Nazionale delle Ricerche' within the CRUI CARE Agreement