Reinforced Sequential Decision-Making for Sepsis Treatment: The POSNEGDM Framework with Mortality Classifier and Transformer

Dipesh Tamboli, Jiayu Chen, Kiran Pranesh Jotheeswaran, Denny Yu, and Vaneet Aggarwal

***Abstract*—Sepsis, a life-threatening condition triggered by the body's exaggerated response to infection, demands urgent intervention to prevent severe complications. Existing machine learning methods for managing sepsis struggle in offline scenarios, exhibiting suboptimal performance with survival rates below 50%. This paper introduces the POSNEGDM— "Reinforcement Learning with Positive and Negative Demonstrations for Sequential Decision-Making" framework utilizing an innovative transformer-based model and a feedback reinforcer to replicate expert actions while considering individual patient characteristics. A mortality classifier with 96.7% accuracy guides treatment decisions towards positive outcomes. The POSNEGDM framework significantly improves patient survival, saving 97.39% of patients, outperforming established machine learning algorithms (Decision Transformer and Behavioral Cloning) with survival rates of 33.4% and 43.5%, respectively. Additionally, ablation studies underscore the critical role of the transformer-based decision maker and the integration of a mortality classifier in enhancing overall survival rates. In summary, our proposed approach presents a promising avenue for enhancing sepsis treatment outcomes, contributing to improved patient care and reduced healthcare costs.**

***Index Terms*—Machine Learning, Transformer, Sepsis Treatment, Healthcare**

I. INTRODUCTION

Sepsis is a life-threatening medical condition characterized by acute organ dysfunction. According to [1], in a typical year, at least 1.7 million adults in America develop sepsis and at least 350,000 adults who develop sepsis die during their hospitalization or are discharged to hospice. Further, 1 in 3 people who dies in a hospital had sepsis during that hospitalization. Early intervention and appropriate treatment are critical for reducing mortality rates. while the current guidelines for sepsis treatment, such as the Surviving Sepsis Campaign [2], have played a significant role in improving patient outcomes, they also come with their own set of challenges [3]. The primary disadvantage of these standards lies in their one-size-fits-all approach, which may not cater to the unique characteristics and varying responses of individual patients [4]. Furthermore, current methods may not fully take into account the rapidly evolving nature of sepsis, leading to delayed or inadequate treatment modifications [5]. In addressing these challenges, our paper focuses on a learning-based methodology that grounds a continuous treatment strategy in the ongoing condition of the patient. Employing a Reinforcement Learning (RL) approach [6], [7], we leverage its capacity to adapt and optimize treatment plans over time. This affords a dynamic and personalized approach, closely aligned with the evolving needs of the individual.

RL has demonstrated success in addressing medical issues, such as diabetes [8], [9] and mechanical ventilation in Intensive Care Units (ICUs) [10]. RL can effectively model physicians' decision-making processes by leveraging historical patient data and past medical decisions. This allows it to guide real-time treatment decision-making, such as the optimization of drug dosages and the timing of administration [11]. Nevertheless, the application of online RL methods is constrained in practice due to the lack of a simulator providing the subsequent state based on the current patient state and treatment decision. This constraint extends to popular Imitation Learning (IL) techniques like GAIL [12] and AIRL [13], which necessitate a similar simulator. To overcome these challenges, we delve into the realm of Offline RL (ORL) [14]. Unlike traditional RL, which learns from real-time interactions, offline RL leverages pre-collected datasets, making it particularly relevant for sepsis treatment. In our context, the dataset comprises trajectories representing sequences of states, actions, and outcomes observed in past Sepsis cases. This shift to offline RL allows us to learn from both positive and negative trajectories, providing a more comprehensive understanding of effective treatment strategies. The utilization of offline RL is a key aspect of our proposed POSNEGDM framework, enhancing its ability to derive impactful insights from historical patient data.

Specifically, we propose the POSNEGDM framework, a novel approach to assist physicians in identifying effective treatment strategies for sepsis by selectively learning from past expert treatment decisions. POSNEGDM stands for "Reinforcement Learning with Positive and Negative Demonstrations for Sequential Decision-Making." It is an offline RL-based model that leverages both positive and negative trajectories to improve sepsis treatment. The proposed POSNEGDM architecture is built upon a Transformer model [15] and includes a feedback reinforcer for performance improvement. With POSNEGDM, the agent can learn from sequences of decisions leading to positive outcomes while concurrently deviating sequences that result in negative consequences, which is accomplished by promoting treatment decisions that steer towards positive outcomes, such as decreased mortality rates. The key components of the proposed architecture are (i) a mortality classifier, (ii) a Transformer-based model that predicts the next state and action based on the historical patient states and treatments, and (iii) an efficient loss function that includes both next-state and next-action prediction loss as well as a reinforced feedback loss based on the mortality classifier.

The authors are with Purdue University, West Lafayette, IN 47906, USA

As shown in Figure 1, the POSNEGDM framework consists of two main components: a mortality classifier (also works as a feedback reinforcer), and a transformer-based decision maker. The mortality classifier is responsible for predicting the probability of mortality for a given patient based on their clinical data. The transformer-based decision maker takes the patient's clinical data as input and generates a sequence of treatment decisions along with the next states. The feedback reinforcer (trained and frozen Mortality Classifier) takes the next states, evaluates the effectiveness of the treatment decisions, and provides feedback ($L_{survival}$ loss, Eq. 2) to the decision maker to improve future decisions. Once the Decision Maker is ready, it generates actions that lead to the live states and save patients.

This is the first offline RL architecture that aims to reduce mortality by explicitly learning to avoid negative trajectories, making it a promising approach to improve Sepsis treatment. This paper takes a significant step forward by quantifying mortality rates and demonstrating that approaches achieving strong action prediction may not necessarily result in improved mortality outcomes. The key contributions of the work can be summarized as:

1) This paper introduces the POSNEGDM framework, the first offline RL-based model that leverages both positive and negative trajectories to lower mortality rates in sepsis patients. Our innovative design incorporates a transformer-based architecture, a neural network-driven mortality classifier that acts as a feedback reinforcer, and a refined loss function that explicitly guides the transformer to circumvent negative trajectories.
2) Leveraging a meticulously selected set of features, our proposed mortality classifier achieved a remarkable accuracy of 96.7% in predicting mortality among sepsis patients.
3) Our classification-feedback-reinforced transformer architecture is capable of emulating expert decision-making and is shown to save 97.39% patients, a significant improvement compared to the 33.4% (DT) and 43.5% (BC) survival rates. These results underscore the effectiveness of the POSNEGDM framework. In real-world settings, our framework can serve as a valuable tool for physicians, providing sepsis treatment decisions as a reference based on the real-time state of patients.

The article is structured to provide a comprehensive exploration of sepsis treatment and the application of POSNEGDM. The introduction (Section 1) offers an overview of sepsis and delineates the treatment challenges associated with it, introducing the application of machine learning as a potential solution to enhance treatment outcomes. Section 2, the related work, reviews existing literature on machine learning in sepsis treatment, shedding light on the limitations of current approaches. Section 3, the methodology, intricately describes the POSNEGDM framework, elucidating its components, such as the transformer-based model (DUALSIGHT), the mortality classifier (feedback reinforcer). In Section 4, the experimental results showcase the performance of the POSNEGDM framework and provide a comparative analysis with existing machine learning-based algorithms. Section 5, the discussion, delves into the implications of the results and outlines potential avenues for future research in this domain. Finally, Section 6, the conclusion, succinctly summarizes the key findings of the paper and emphasizes the significance of leveraging machine learning for improved sepsis treatment outcomes.

II. RELATED WORKS

This section provides an overview of cutting-edge Machine Learning (ML) algorithms applied to healthcare decision-making, particularly focusing on Offline Reinforcement Learning and Imitation Learning.

A. Offline Reinforcement Learning

The Markov Decision Process (MDP) can be represented as a tuple (S, A, P, R) with states $s \in S$, actions $a \in A$, transition dynamics $P(s'|s,a) \in [0,1]$, and the reward function given by $r = R(s,a)$. A sequence of states, actions, and rewards would comprise a trajectory, which can be represented as $\tau = (s_1, a_1, r_1, \cdots, s_T, a_T, r_T)$. A reinforcement learning (RL) problem aims to learn a policy such that it maximizes the expected return $\mathbb{E}[\sum_{t=1}^{T} r_t]$. Offline RL [14] is proposed to train RL agents using pre-collected data rather than interacting with the environment in real time. This is in contrast to online RL [6], [7], where the agent learns from the interactions with a simulator. Offline RL proves particularly beneficial when real-time data collection via a simulator is challenging or costly. It also presents a more practical solution, given the easier accessibility of trajectories.

In offline RL, an agent is trained on a dataset of previous interactions, often referred to as a replay buffer. This buffer contains tuples composed of the elements: state, action, reward, and next state. Such offline data can typically be gathered from people's routine practice for a given task. Using this information, the agent can learn a policy that correlates states with actions, and a value function that estimates the expected return of a certain state or state-action pair. The derived policy should ideally produce decisions that maximize future return from the current state.

In recent literature on sepsis identification and treatment systems based on offline RL, several approaches have been explored. These approaches aim to model the discrete Markov Decision Process (MDP) [11], [16], [17], [18], and then use online RL approaches. However, the discrete feature space may lead to information loss, which affects the accuracy of online RL. Recent works [19], [20] have modeled MDP in continuous state, while the result relies on the model approximation. In this work, we aim to make decisions without learning the model. Further, these works do not explicitly avoid the negative trajectories. On the contrary, POSNEGDM selects actions from the given set to diverge from the negative trajectories, achieved by maximizing the divergence between the predicted and ground-truth trajectory.

B. Imitation Learning and Behavioral Cloning

Imitation Learning (IL) [21] is a method of training an agent to perform a task by observing and mimicking decision-making sequences from domain experts. IL can be employed

to learn a broad spectrum of tasks, ranging from simpler ones such as line-following to more complex tasks like game-playing.

Behavioral Cloning (BC) represents a specific IL method that trains an agent to mimic expert behavior by learning a policy identical to that of the expert. This employs supervised learning, treating states (x) as input data and actions (y) as target values. For continuous actions, the Mean Squared Error (MSE) loss ($\mathrm{E}((f(s)-a)^2)$) is used, minimizing the difference between predicted ($f(s)$) and actual actions (a). In the case of discrete actions, a cross-entropy loss is employed, measuring the dissimilarity between predicted action probabilities and true action distributions. BC learns to replicate the expert's actions, facilitating the mapping from states to actions through supervised learning. One of the earliest works on BC was proposed by [22] for autopilot applications. Since then, it has been broadly adopted in areas such as autonomous vehicles [23], robotic manipulators [24], and healthcare [25]. In this paper, we employ BC as one of the baseline methods for comparison. However, only decision trajectories leading to positive outcomes are used in the training process, as the agent cannot follow a negative trajectory wherein patients did not survive. Compared with BC, our proposed algorithm demonstrates superior data efficiency by explicitly incorporating negative samples from the dataset.

Negative trajectories can provide crucial insights into which actions should be avoided to prevent repetition of mistakes. Authors of [26] and [27] propose IL models that employ adversarial discriminators to convert negative trajectories into positive ones. However, these techniques necessitate access to simulators, much like the state-of-the-art (SOTA) IL algorithms GAIL [12] and AIRL [13]. In contrast, our proposed method does not depend on a simulator which can be expensive to construct and inaccessible in many application domains. Instead, it recovers a policy from offline data only, facilitated by specially designed objective functions. For instance, when dealing with negative trajectories, we select actions from the given set to diverge from the negative trajectories, achieved by maximizing the divergence between the predicted and ground-truth trajectory.

In our approach, we introduce a novel perspective in Sepsis treatment by explicitly leveraging negative trajectories, a strategy not explored in existing research. While conventional machine learning models, such as Decision Transformer (DT) [28] and Behavioral Cloning (BC) [22], have demonstrated high action prediction accuracy, they lack the explicit consideration of negative outcomes. Our novel framework, named POSNEGDM, explicitly addresses this gap by incorporating a Mortality Classifier that also acts as a Feedback Reinforcer. The Mortality Classifier allows us to assess the potential negative outcome (mortality) of each action, enabling the reinforcement of the model to avoid actions that lead to unfavorable states. This explicit focus on negative trajectories distinguishes our work from previous studies and contributes to the development of a more robust and clinically relevant Sepsis treatment model. The decision to compare our model against DT and BC is motivated by their popularity and effectiveness in predicting actions; however, their omission of explicit consideration for negative trajectories positions our POSNEGDM framework as an innovative and advantageous alternative in improving patient outcomes in Sepsis treatment.

III. PROPOSED APPROACH

In this section, we first delineate the algorithm's components, which include the mortality classifier and the transformer-based dualsight network. Subsequently, we present the complete POSNEGDM framework, with a particular focus on the novel training objectives.

A. Mortality Classifier

We have devised a classifier capable of predicting a patient's mortality likelihood, discerning between probable survival or demise. The classifier takes the current patient state as input and produces a binary result signifying the survival likelihood. The architecture of our mortality classifier comprises five fully-connected layers of size 64, with the training hyperparameters outlined in Table I. A significant challenge encountered during the training process was the presence of a highly imbalanced dataset, with a significantly lower number of deceased patients compared to survivors. To rectify this, we employed Borderline SMOTE [29] to impute minority class data and upsampled the training set.

TABLE I
HYPERPARAMETERS FOR TRAINING THE MORTALITY CLASSIFIER

Parameter	Learning Rate	Weight Decay	Optimizer	Dropout
Value	1e-3	1e-5	Adam	0.2

This classifier is trained to achieve high accuracy in predicting the outcome of a treatment trajectory. Empirically, it has demonstrated an impressive accuracy of 96.7% on the test dataset. It is subsequently utilized to evaluate the quality of treatment decisions derived from various algorithms and to provide feedback for the training of the transformer-based decision maker (as represented in the loss function illustrated in Equation (2)).

B. DUALSIGHT *Decision Maker*

A transformer is particularly well-suited for sequential decision-making tasks, such as Sepsis treatment, due to its ability to capture dependencies and relationships between different states in a sequence effectively. The transformer-based model in the POSNEGDM framework learns from historical data and expert actions by processing input sequences through modality-specific linear embeddings and a positional episodic timestep encoding. The tokens are then introduced into a GPT architecture, which anticipates output in an autoregressive fashion, employing a causal self-attention mask. This self-attention module calculates a weighted sum of the input states, with weights assigned based on the similarity among the states, allowing the model to capture dependencies and relationships between different states in the sequence effectively. The significance of this approach is that it enables the model to learn from both positive and negative trajectories, which is key to improving patient survival rates in sepsis treatment. Additionally, the feedback reinforcer in

the POSNEGDM framework uses the forecasted next states to ascertain the ultimate survival likelihood of patients, which is subsequently utilized as feedback to fortify the decision maker. This unique configuration significantly enhances the decision-maker's performance.

The DUALSIGHT decision maker is trained to produce a sequence of expert-like treatment decisions based on the patient states. Training is done using offline data provided in the MIMIC-III dataset, detailed in Section IV. The actions in the RL setting represent medical interventions, including the administration of intravenous fluids and vasopressor drugs within a 4-hour window. These actions are defined within a 5×5 discrete space, where 0 denotes no drug administration, and the other four non-zero drug dosages are classified into four quartiles. This results in 25 possible discrete action combinations for the two treatments. Regarding the reward function, it assigns $+1$ for positive trajectories and -1 for negative ones, provided only at the final time step of the patient's trajectory. A trajectory is considered positive if the patient survives post-treatment, else it is deemed negative. The agent's ultimate goal is to maximize the cumulative reward throughout the entire treatment period.

We utilize the network structure suggested in the Decision Transformer (DT) [28] for our decision maker. The Transformer model, renowned for its efficacy in sequence processing tasks, has been employed in leading Large Language Models like GPT [30] and BERT [31]. Leveraging a GPT-like architecture, DT has demonstrated exceptional performance, surpassing state-of-the-art offline RL baselines in numerous benchmarks.

As illustrated in Figure 1, our decision maker utilizes a sequence of past states, actions, and returns as input. The return signifies the aggregated future rewards from a given time step t until the episode's termination, encapsulating the expected outcome over the decision enacted. This input sequence undergoes processing through modality-specific linear embeddings, supplemented with a positional episodic timestep encoding to help the model understand the order of events. Subsequently, the tokens are introduced into a GPT architecture, which anticipates output in an autoregressive fashion, employing a causal self-attention mask. This self-attention module calculates a weighted sum of the input states, with weights assigned based on the similarity among the states, which allows the model to capture dependencies and relationships between different states in the sequence effectively.

A crucial distinction between our architecture and that of the Decision Transformer lies in our decision maker's ability to anticipate not only the immediate action but also the ensuing state, thus offering dual insights. The forecasted next states are employed by the mortality classifier to ascertain the ultimate survival likelihood of patients, subsequently utilized as feedback to fortify the decision maker. This unique configuration significantly enhances our decision maker's performance.

C. The Overall Framework: POSNEGDM

The key idea of POSNEGDM architecture is to learn from positive trajectories and move away from negative ones. The POSNEGDM framework offers insightful treatment recommendations, including the administration of intravenous fluids and vasopressor drugs. These suggestions guide clinicians by providing clear actions to consider and potential pitfalls to avoid, thereby bolstering the reliability and safety of their decision-making process. The POSNEGDM architecture comprises two components: (a) an autoregressive decision maker that anticipates the current treatment decision (i.e., action) and subsequent patient states based on historical data, and (b) a mortality classifier that reinforces the decision maker if the final state produced by the preceding networks guides the patient towards a survival state. The Mortality Classifier is independently trained to distinguish between alive and deceased states. In the training of our decision-maker, the Mortality Classifier operates in inference (Feedback Reinforcer) mode rather than training mode to avoid interfering with DUALSIGHT's learning. This enables the DUALSIGHT to produce additional gradients derived from the alive/dead state Feedback Reinforcer's adversarial loss (Eq. 2). These supplementary gradients facilitate improved decision-maker training, enhancing its ability to predict actions that result in alive states and consequently reducing the mortality rate.The objectives for training the Decision Maker with the offline data and mortality classifier are shown as follows.

First, the decision maker is trained to predict the next state by minimizing the Mean Squared Error (MSE) between the predicted state $\hat{s}_t$ and the ground-truth state s_t:

$$L_{state}(\hat{s}_t, s_t) = L_{MSE}(\hat{s}_t, s_t) = \frac{1}{|S|}\sum_{i=1}^{|S|}(\hat{s}_{t,i} - s_{t,i})^2, \quad (1)$$

where $|S|$ is the number of elements in the state vector. As outlined earlier, the forecast accuracy for subsequent states provides extra constraints in training the decision maker, when compared to the Decision Transformer. This design allows the decision maker to extract more relevant and beneficial features from the input, leading to a better equipped model to predict future treatment outcomes. Additionally, the anticipated final state will be fed into the mortality classifier to calculate patient survival rates, which further augments the training feedback.

Next, in order to prevent negative outcomes, the final patient state guided by the sequence of treatments should exhibit a relatively high survival rate. Utilizing the pretrained mortality classifier, we can incorporate another objective that acts as a constraint on survival outcomes, making treatment decisions safer and more reliable. Specifically, we introduce the loss term as follows:

$$\begin{aligned} &L_{survival}(MC, \text{DUALSIGHT}, \tau) \\ &= \mathbb{E}_{\tau \sim P_{data}(\cdot)}[\log(1 - MC(\text{DUALSIGHT}(\tau)))] \end{aligned} \quad (2)$$

Here, τ denotes the decision sequence sampled from an offline dataset with the distribution $P_{data}(\cdot)$, DUALSIGHT(τ) represents the predicted final patient state following a series of treatments provided by the decision maker, MC represents the mortality classifier that gives out the survival rate. The training of the decision maker is geared towards enhancing the patient's survival rate, which in turn minimizes the survival loss $L_{survival}$.

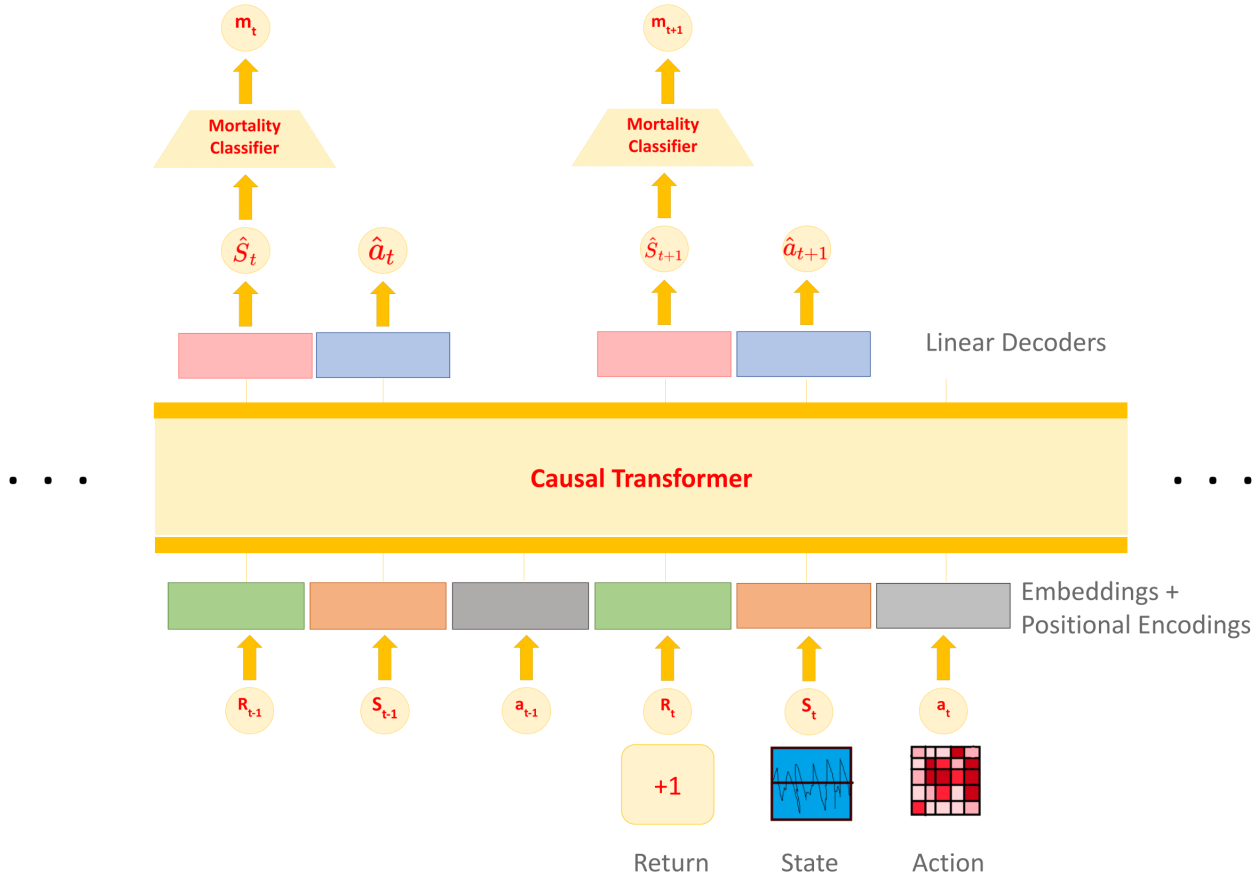


Fig. 1. The DUALSIGHT decision maker takes in states, actions, and returns as input, which are first embedded into linear representations that are specific to each modality. The positional episodic timestep encoding is added to the input to help the model understand the order of events. The tokens are then fed into the GPT architecture, which uses a self-attention mechanism to predict actions and next states. The causal mask ensures that the model can only attend to previous tokens, preserving the causality of the system. The predicted states are subsequently input into the trained Mortality Classifier to assess whether the implemented action guides the patient towards a deceased state. The mortality prediction m_t is employed to influence the DUALSIGHT, compelling it to choose actions aligned with the mortality classifier's prediction of an alive state. This process integrates mortality considerations into the decision-making mechanism, emphasizing the importance of actions that contribute to favorable patient outcomes.

Last, we minimize the action prediction error in the positive trajectories and maximize it in the negative ones. By doing so, we encourage our model to formulate decision rules that not only mimic the expert's actions but also learn to evade actions tied to negative outcomes. This results in a more robust understanding of action-avoidance related to negative trajectories. The loss associated with action prediction is formulated as:

$$L_{action}(\hat{a}_t^+, a_t^+, \hat{a}_t^-, a_t^-) = L_{CE}(\hat{a}_t^+, a_t^+) + \eta L_{CE}(\hat{a}_t^-, a_t^-) \tag{3}$$

In this equation, a^+ corresponds to the actions involved in positive trajectories, whereas a^- represents the actions derived from negative trajectories. L_{CE} represents the cross-entropy loss, and η is a weight factor that assists us in maintaining a balance between adhering to expert trajectories and circumventing negative outcomes.

Integrating the three objectives listed above, we can get an overall loss function as:

$$L_{total} = \alpha L_{action} + \beta L_{state} + \gamma L_{survival} \tag{4}$$

where the parameters α, β, γ control the weight of each term in the overall loss. Rather than merely duplicating treatment decisions (i.e., actions), we propose a three-fold improvement: (1) Empower the agent to predict subsequent states, enriching its understanding of the input information. (2) Explicitly improve the survival rate as assessed by the mortality classifier. (3) Distinctly address action predictions for positive and negative trajectories. As a result, our algorithm can make better use of the treatment sequences, which includes a mixture of positive and negative trajectories, to achieve a significantly higher survival rate.

Algorithm 1 outlines the training procedure for the POS-NEGDM framework, consisting of a Mortality Classifier (MC) and a Decision Maker (DM). The Mortality Classifier is initially trained on the mortality ground truth using cross-entropy loss. The trained Mortality Classifier, denoted as $w_{T_{MC}}$, is then frozen and serves as the Feedback Reinforcer. Subsequently, the Decision Maker is trained using a combination of losses, including survival loss ($\mathcal{L}_{survival}$) adversarially enforced by the Mortality Classifier. The DM's weights, denoted as $w_{T_{DM}}$, are updated iteratively based on the overall loss calculated from the mortality prediction, state prediction, and action prediction. This process aims to enhance the Decision Maker's

ability to predict actions leading to patient survival by leveraging the adversarial feedback from the Mortality Classifier.

Algorithm 1 POSNEGDM

MC: Mortality Classifier, DM: Decision Maker, R_t: Reward at timestep $t \in \{-1, 1\}$, a_t: Action at timestep t, S_t: State at timestep t, T_{MC}: Total iterations for MC, T_{DM}: Total iterations for MC, FR: Feedback Reinforcer with frozen MC weights, M: Mortality GroundTruth, m: Predicted Mortality, CEL: Cross-Entropy Loss, ∇g_w: gradient.

Training Mortality Classifier
Initialize w_{MC} randomly
for $t = 1$ **to** T_{MC} **do**
    $S, M \leftarrow$ sample batch from $data_{train}$
    $m \leftarrow MC(S, w_{t_{MC}})$
    $\mathcal{L} \leftarrow CEL(m, M)$
    Calculate $\nabla g_{w_{t_{MC}}}$ from $\mathcal{L}$ and update $w_{t_{MC}}$
end for
$w_{T_{MC}}$ will be trained the Mortality Classifier
Frozen $\overline{w_{T_{MC}}}$ will act as the Feedback Reinforcer
Training POSNEGDM
Initialize w_{DM} randomly
for $t = 1$ **to** T_{DM} **do**
    $d_{batch} \leftarrow$ sample from $data_{train}$
    $S_t, a_t \leftarrow d_{batch}$
    $S_{t+1}, a_{t+1} \leftarrow DM(S_t, a_t, w_{t_{DM}})$
    $m_{t+1} \leftarrow MC(S_{t+1}, \overline{w_{T_{MC}}})$
    $\mathcal{L}_{survival} \leftarrow CEL(m_{t+1}, 1)$ (Forcing m_{t+1} to be 1 to adversarially train DM to take actions resulting in alive state)
    Calculate $\mathcal{L}_{state}$ and $\mathcal{L}_{action}$
    Calculate $\nabla g_{w_{t_{DM}}}$ from $\mathcal{L}_{total}$ and update $w_{t_{DM}}$
end for
$w_{T_{DM}}$ will be the trained Decision Maker

IV. SEPSIS DATA DESCRIPTION

The MIMIC-III database [32] is a comprehensive collection of de-identified clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, accessible to researchers globally under a data use agreement. It spans over a decade and contains data for 53,423 adult hospital admissions and 7870 neonate admissions, covering 38,597 distinct adult patients with a median age of 65.8 years and in-hospital mortality of 11.5%. The database offers detailed information about patients' demographics, vital signs, laboratory test results, treatments, and outcomes, as well as free-text notes recorded by healthcare providers. The median length of an ICU stay is 2.1 days and a hospital stay is 6.9 days, with an average of 4,579 charted observations and 380 laboratory measurements available per hospital admission. The MIMIC-III database is widely used for research in critical care and offers a unique opportunity for researchers to study sepsis and other critical illnesses using a large, diverse, and rich dataset.

Sepsis MIMIC-III data is a subset of the MIMIC-III database, specifically focused on the patients who were diagnosed with sepsis during their ICU stay. It can be used to train models for sepsis detection and treatment, as well as for researching sepsis and its associated outcomes. The Sepsis MIMIC-III dataset comprises 19,614 treatment trajectories specific to sepsis patients. In order to design our experiment, we randomly allocate 30% of this data for testing purposes, while the remaining 70% is utilized for training. The following table presents a brief summary of this dataset. Here, positive trajectories correspond to patient survival instances, while negative ones denote cases of patient fatality. The mortality rates for both the training and testing sets approximate 9.5%.

TABLE II
OVERVIEW OF THE SEPSIS DATASET

Type	Positive	Negative	Total
Train	12411	1319	13730
Test	5321	563	5884
Total	17732	1882	19614

The Sepsis MIMIC-III dataset typically consists of several types of patient data, including demographic information, vital signs, laboratory test results, treatments, and outcomes. The input and output features that we are considering is listed as follows.

The data preprocessing phase comprised selecting pertinent features, discretizing them, and transforming them into a 4-hour window. Key physiological parameters, encompassing demographics, laboratory values, vital signs, and intake/output events, were systematically gathered for each patient. To facilitate analysis, the data were aggregated into 4-hour windows, with the mean or sum recorded—depending on the nature of the data—when multiple data points were available within a given window. Notably, we have used the same data preprocessing steps as detailed in [16].

Input features:

- Demographic information: patient age, sex, weight, height.
- Vital signs: heart rate, blood pressure, temperature, oxygen saturation.
- Laboratory test results: white blood cell count, lactate level, blood glucose, creatinine.
- Treatments: medications administered, fluid boluses, mechanical ventilation.
- Length of stay: the number of days the patient spent in the ICU.
- Organ dysfunction: whether or not the patient experienced organ dysfunction during their ICU stay.

Output features:

- Diagnosis: whether or not the patient was diagnosed with sepsis.
- Mortality: whether or not the patient died during their ICU stay.

V. EXPERIMENTAL RESULTS

In Table III, we present the hyperparameters used for POSNEGDM, with the exception of those pertaining to the mortality classifier, which are outlined in Table I. The DUALSIGHT model consists of three layers and utilizes a single attention head with an embedding dimension of 128. During

TABLE III
HYPERPARAMETERS FOR POSNEGDM

Hyperparameter	Value
Number of layers	3
Number of attention heads	1
Embedding dimension	128
Batch size	64
Context Length	3
Nonlinearity	Relu
Dropout rate	0.1
Learning rate	$1e-4$
Weight decay	$1e-4$
Warm-up steps	10000
α: Weight of action prediction loss	1
β: Weight of state prediction loss	0.1
γ: Weight of survival loss	1

training, we use a batch size of 64 and a context length of 3. The model employs the ReLU activation function as the non-linearity and includes a dropout layer with a rate of 0.1 to prevent overfitting. To avoid overfitting further, we set the learning rate to 1e-4 and apply a weight decay of 1e-4. The training process incorporates a warm-up phase of 10,000 steps, where the learning rate gradually increases from 0 to the specified value. The parameters α, β, γ control the weight of each term in the overall loss (i.e., Equation (4)). The hyperparameters, including the learning rate, number of epochs, and layers, underwent fine-tuning through an iterative process within a standard k-fold cross-validation setup. This procedure involved adjusting these settings and evaluating the model's performance on a validation dataset to identify configurations leading to the highest validation accuracy.

Assessing off-policy models presents a significant challenge as it is difficult to estimate the impact of executing a learned policy on patient mortality without a simulator. To address this challenge, we utilize a pretrained mortality classifier known for its high accuracy in predicting survival states. Our primary evaluation metric is the mortality rate, which indicates the proportion of patients experiencing negative outcomes as a result of the model's recommended treatments. It is important to note that the mortality rate considers the patient's death in any state along the trajectory.

In terms of baselines, we compare our POSNEGDM model to two established approaches: the state-of-the-art offline reinforcement learning algorithm, Decision Transformer (DT) [28], and the widely used imitation learning algorithm, Behavioral Cloning (BC) [23]. It is worth noting that other imitation learning algorithms like GAIL [12] and AIRL [13] require a simulator, making them unsuitable as baselines in our specific scenario. Our code is publicly available at https://github.com/Dipeshtamboli/PosNegDM-Reinforced-Sequential-Decision-Making-for-Sepsis-Treatment.

A. Achieving Low Mortality with POSNEGDM

Table IV showcases the performance of our proposed algorithm, POSNEGDM ($\alpha = 1, \beta = 0.1, \gamma = 1$) benchmarked against the Decision Transformer [28] and a Behavioral Cloning model [24]. The evaluation was conducted using test data from the Sepsis MIMIC-III dataset. The table includes several metrics: (i) action prediction accuracy for surviving patients in the test data (positive data), (ii) mortality rate among patients who were alive (positive data), (iii) mortality rate among patients who did not survive (negative data), and (iv) overall mortality rate. In order to obtain this metrics, we consider the last 10 time-steps of each trajectory. To assess action prediction on positive test trajectories, we compare the actions taken at each time-step in our approach with those in the expert data. Regarding mortality assessment, we input states from a test trajectory and observe whether the next predicted state indicates that the patient remains alive or not. The evaluation employs a trained mortality classifier in inference mode to make this determination. It is essential to clarify that the mortality classifier is trained independently on the training set and does not confer an advantage to the POSNEGDM framework. The same mortality classifier is used consistently to assess the performances of Decision Transformer (DT) and Behavioral Cloning (BC). This process is repeated for all states in the trajectory. If at any point, the patient is predicted not to be alive, it is classified as a negative mortality case. In contrast, if the patient is predicted to be alive throughout the trajectory, it is considered a positive mortality case. This mortality is called Step-by-step method, we will also consider a complete trajectory method in the next subsection.

The results demonstrate a high action prediction accuracy of 94.6% for the proposed algorithm, which is comparable to the baselines. However, it is important to note that the proposed algorithm significantly outperforms the baselines in terms of mortality rate. Simply achieving or mimicking actions is insufficient to prevent mortality, as any incorrect action within the trajectory can result in a patient not surviving.

The proposed algorithm, with its careful treatment to avoid mortality, achieves a remarkably low mortality rate of 2.61%. In contrast, the Behavioral Cloning model had a mortality rate of 56.5%, and the Decision Transformer model had a mortality rate of 66.6%. It's significant to note that the figure 3.6% suggests that with the application of our model's medical decisions, 96.4% of patients, who otherwise would not have survived under the physicians' treatments, ultimately survived. This represents a substantial improvement. By mandating the prediction of state outputs and incorporating an adversarial mortality classifier, the model is empowered to enhance its mortality rate predictions. The feedback reinforcer, represented by the mortality classifier, plays a crucial role in offering additional guidance to the model, facilitating learning from errors. In totality, the amalgamation of these distinctive features and mechanisms is the driving force behind the superior performance of the POSNEGDM framework when compared to existing machine learning-based algorithms. These findings indicate that our proposed method is better suited for avoiding negative outcomes, such as mortality, which is a critical factor in clinical decision-making.

Therefore, the POSNEGDM model holds the potential to significantly enhance the accuracy and reliability of clinical decision-making, ultimately improving patient outcomes in sepsis treatment.

B. Ablation studies

In this section, we conduct ablation studies to evaluate the impact of the three objective terms in the overall loss function

TABLE IV
PERFORMANCE COMPARISON OF ALGORITHMS

Algorithm	Action Prediction Accuracy on Positive Test Data %	Mortality %		
		Positive Data	Negative Data	Total
POSNEGDM	94.6	**2.5**	**3.6**	**2.61**
Decision Transformer	94.3	68.2	51.5	66.6
Behavioral Cloning	**95.1**	57.5	46.7	56.5

TABLE V
IMPORTANCE OF THE ACTION PREDICTION LOSS (L_{action}). HERE, $\beta = 0.1$ AND $\gamma = 1$.

Loss importance (α)	Action Prediction Accuracy on Positive Test Data %	Mortality %	
		Step by Step	Complete Traj.
0.0	4.6	6.12	0.36
0.1	91.6	5.61	0.65
0.3	94	6.07	0.63
0.5	93.6	3.33	0.18
0.8	92.8	3.06	0.40
1.0	**94.6**	**2.61**	**0.18**

TABLE VI
IMPORTANCE OF SUBSEQUENT STATE PREDICTION LOSS (L_{state}). HERE, $\alpha = 1$ AND $\gamma = 1$.

Loss importance (β)	Action Prediction Accuracy on Positive Test Data %	Mortality %	
		Step by Step	Complete Traj.
0.0	94.7	N\A	N\A
0.1	94.6	**2.61**	**0.18**
0.3	92.9	3.57	0.36
0.5	93.5	6.33	0.40
0.8	92.9	7.54	0.20
1.0	92.8	10.02	0.67

(Equation (4)) on system performance, as delineated in Tables V to VII.

Here, we compare action prediction accuracy on the positive data and mortality. Action prediction accuracy measures how well our model is able to mimic an expert's actions. Mortality gives us an idea of how well the model is able to drive patient states to alive states by taking proper actions. Given that we only have information from the trajectories available and no direct access to the model, we have calculated mortality in two different ways. In the Step-by-step method, we input states from a test trajectory and check if the next predicted state is an alive state or not. We do it for all the states in the trajectory. In the Complete-trajectory state, we start with the initial test state and allow the model to generate the next ten actions and corresponding states. If any of the states results in a dead state, we consider that the patient has died in that trajectory. These mortality results are for both (positive and negative) test trajectories (i.e., total mortality).

TABLE VII
IMPORTANCE OF THE SURVIVAL LOSS ($L_{survival}$). HERE, $\alpha = 1$, $\beta = 0.1$.

Loss importance (γ)	Action Prediction Accuracy on Positive Test Data %	Mortality %	
		Step by Step	Complete Traj.
0.0	**94.7**	60.41	10.29
0.1	94.3	5.21	0.49
0.3	93.9	3.95	0.45
0.5	93.4	**2.32**	0.36
0.8	93.9	2.41	**0.18**
1.0	94.6	2.61	**0.18**

Table V showcases how the action prediction loss impacts the POSNEGDM system's performance. Excluding L_{action} (i.e., setting $\alpha = 0$) results in a significant drop in action prediction accuracy on expert data, while increasing α enhances overall performance, as indicated by total mortality.

Table VI reveals the effect of the subsequent state prediction loss on the POSNEGDM system's performance. Without L_{state} (i.e., setting $\beta = 0$), subsequent state prediction and consequent mortality estimation from the classifier are unattainable. In addition, we can see that the total mortality increases as β increases.

In Table VII, we evaluate the impact of the survival loss on the performance of the POSNEGDM system. We observe a significant increase in mortality when $L_{survival}$ is omitted, thereby underscoring its significance. Furthermore, we find that the performance with $\gamma = 0.8$ is not only comparable to that of $\gamma = 1.0$, but it actually results in better mortality results than the parameter setting we employed for Table IV.

Overall, the results highlight the variability of mortality rates across different hyperparameters. The parameters α, β, and γ provide control over POSNEGDM's training objectives, allowing us to balance considerations such as prioritizing negative trajectory avoidance or achieving high action prediction accuracy. The impact of varying these parameters on the model's behavior and performance is demonstrated in Tables V, VI, and VII for actions, states, and adversarial training, respectively. Importantly, it should be noted that higher action prediction accuracy does not necessarily correlate with lower mortality rates. It is essential to explicitly study and consider mortality, which had not been explored in the context of decision making prior to our work, to the best of our knowledge.

VI. CONCLUSION

Sepsis is a severe medical condition where rapid, precise interventions can significantly decrease mortality rates. Existing guidelines, however, fall short of providing real-time, individualized treatment decisions. Reinforcement Learning (RL) has shown promise in the medical field by modeling physicians' decision-making processes but relies on a simulator. To circumvent these challenges, we proposed the POSNEGDM framework, an Offline RL-based solution. This innovative method assists physicians in devising effective treatment strategies for sepsis patients by learning from experts' decisions (e.g., MIMIC-III dataset). The architecture of POSNEGDM includes a Transformer model and a feedback reinforcer, enabling the agent to learn from successful decision sequences while veering away from those leading to negative

outcomes. Our model could replicate expert decision-making with 94.6% accuracy and showed an improvement in survival rates, saving 97.39% of patients compared to survival rates of 33.4% (DT) and 43.5% (BC). The POSNEGDM framework hence emerges as a promising tool in assisting physicians in making informed, timely treatment decisions for sepsis patients, underscoring its potential in significantly improving patient outcomes.

Our work shows promise in enhancing sepsis treatment, with future considerations for deploying the model in clinical practice. Ethical and legal aspects, such as patient privacy and informed consent, require careful attention. The model may need periodic adaptation to address real-world clinical complexities. We also acknowledge limitations, including dependence on training data quality, generalization challenges, and interpretability issues, as directions for future work.

REFERENCES

[1] C. for Disease Control and Prevention, "What is sepsis?" 2023, accessed: 2023-07-13. [Online]. Available: https://www.cdc.gov/sepsis/what-is-sepsis.html

[2] L. Evans, A. Rhodes, W. Alhazzani, M. Antonelli, C. M. Coopersmith, C. French, F. R. Machado, L. Mcintyre, M. Ostermann, H. C. Prescott *et al.*, "Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021," *Intensive care medicine*, vol. 47, no. 11, pp. 1181–1247, 2021.

[3] C. Fleischmann-Struzek, L. Mellhammar, N. Rose, A. Cassini, K. E. Rudd, P. Schlattmann, B. Allegranzi, and K. Reinhart, "Incidence and mortality of hospital- and ICU-treated sepsis: results from an updated and expanded systematic review and meta-analysis," *Intensive Care Med.*, vol. 46, no. 8, pp. 1552–1562, Aug. 2020.

[4] C. Rhee and E. J. Septimus, "Heterogeneity in clinical presentations of sepsis: Challenges and implications for "one-size-fits-all" time-to-antibiotic measures," *Critical Care Medicine*, vol. 50, no. 5, pp. 886–889, 2022.

[5] S. Ranjit and N. Kissoon, "Challenges and solutions in translating sepsis guidelines into practice in resource-limited settings," *Translational Pediatrics*, vol. 10, no. 10, p. 2646, 2021.

[6] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, 2018, pp. 1861–1870.

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[8] M. A. Raheb, V. R. Niazmand, N. Eqra, and R. Vatankhah, "Subcutaneous insulin administration by deep reinforcement learning for blood glucose level control of type-2 diabetic patients," *Computers in Biology and Medicine*, vol. 148, p. 105860, 2022.

[9] M. Voets, K. Møllersen, and L. A. Bongo, "Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *PloS one*, vol. 14, no. 6, p. e0217541, 2019.

[10] H. Li and X. Xie, "Deep learning-based prediction of mechanical ventilation reintubation in intensive care units," in *City, Society, and Digital Transformation: Proceedings of the 2022 INFORMS International Conference on Service Science*, 2022, pp. 15–22.

[11] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature medicine*, vol. 24, no. 11, pp. 1716–1720, 2018.

[12] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[13] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adverserial inverse reinforcement learning," in *International Conference on Learning Representations*, 2017.

[14] R. F. Prudencio, M. R. Maximo, and E. L. Colombini, "A survey on offline reinforcement learning: Taxonomy, review, and open problems," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi, "Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach," in *Machine Learning for Healthcare Conference*, 2017, pp. 147–163.

[17] P. Kaushik, S. Kummetha, P. Moodley, and R. S. Bapi, "A conservative q-learning approach for handling distribution shift in sepsis treatment strategies," *arXiv preprint arXiv:2203.13884*, 2022.

[18] X. Peng, Y. Ding, D. Wihl, O. Gottesman, M. Komorowski, H. L. Liwei, A. Ross, A. Faisal, and F. Doshi-Velez, "Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning," in *AMIA Annual Symposium Proceedings*, vol. 2018, 2018, p. 887.

[19] A. Raghu, M. Komorowski, and S. Singh, "Model-based reinforcement learning for sepsis treatment," *arXiv preprint arXiv:1811.09602*, 2018.

[20] Z. Wang, H. Zhao, P. Ren, Y. Zhou, and M. Sheng, "Learning optimal treatment strategies for sepsis using offline reinforcement learning in continuous space," in *International Conference on Health Information Science*, 2022, pp. 113–124.

[21] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys*, vol. 50, no. 2, pp. 1–35, 2017.

[22] C. Sammut, S. Hurst, D. Kedzier, and D. Michie, "Learning to fly," in *Proceedings of the 9th International Workshop on Machine Learning*, 1992, pp. 385–393.

[23] T. V. Samak, C. V. Samak, and S. Kandhasamy, "Robust behavioral cloning for autonomous vehicles using end-to-end imitation learning," *SAE International Journal of Connected and Automated Vehicles*, vol. 4, no. 12-04-03-0023, 2021.

[24] J. Choi, H. Kim, Y. Son, C.-W. Park, and J. H. Park, "Robotic behavioral cloning through task building," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2020, pp. 1279–1281.

[25] I. Bica, D. Jarrett, and M. van der Schaar, "Invariant causal imitation learning for generalizable policies," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3952–3964, 2021.

[26] 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 *Proceedings of The Web Conference 2020*, 2020, pp. 1785–1795.

[27] S. I. H. Shah, A. Coronato, M. Naeem, and G. De Pietro, "Learning and assessing optimal dynamic treatment regimes through cooperative imitation learning," *IEEE Access*, vol. 10, pp. 78 148–78 158, 2022.

[28] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.

[29] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*, 2005, pp. 878–887.

[30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[31] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[32] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

APPENDIX A
ADDITIONAL EXPERIMENTS

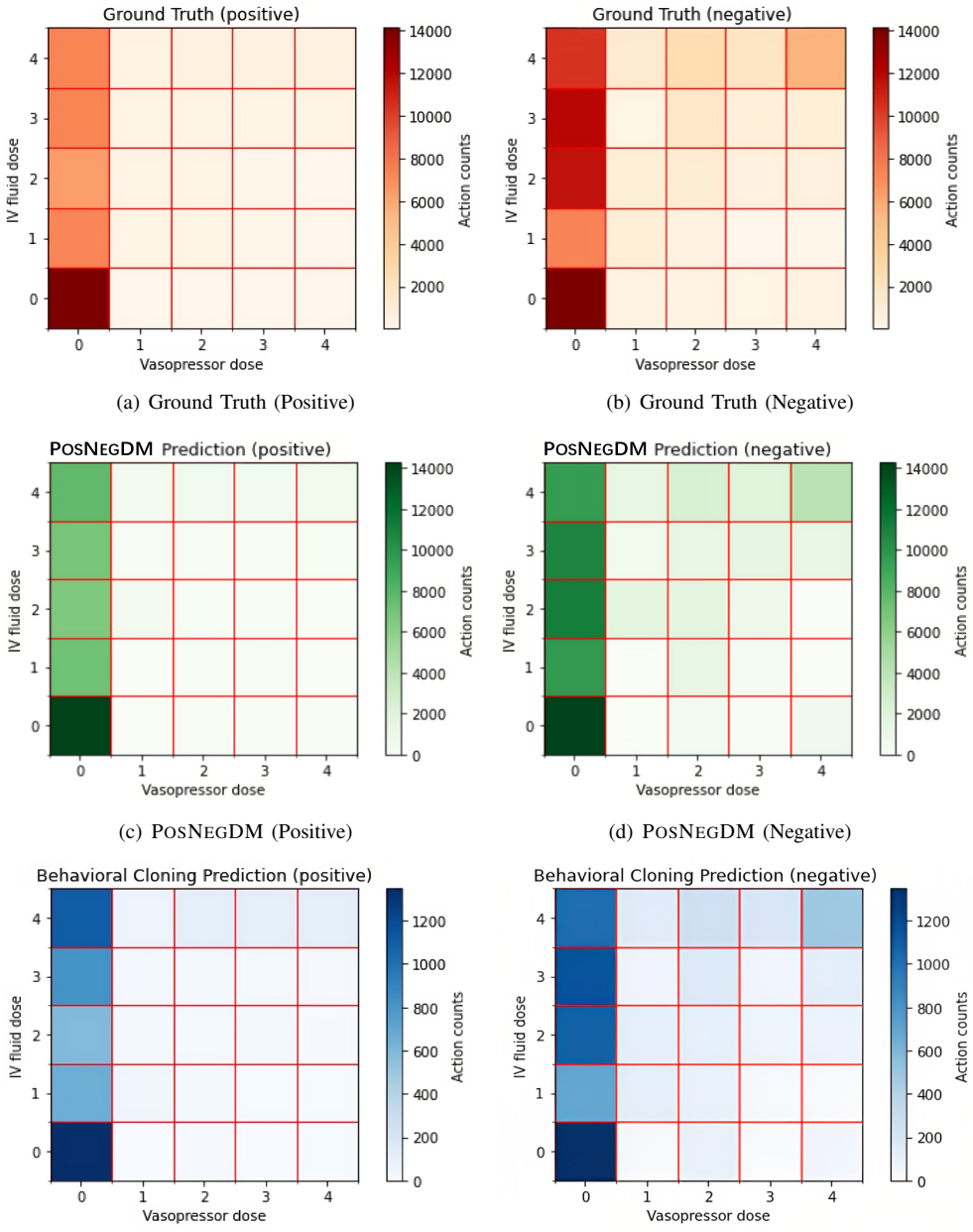
This section includes the sensitivity experiments and additional 2D histograms that visualize the aggregated actions recommended by the physician (ground truth), POSNEGDM, and Behavioral Cloning (BC). These histograms provide a qualitative analysis of the actions recommended by each method and highlight the differences in decision-making strategies.

In Table VIII, we present the results of a sensitivity experiment conducted to assess the stability of our model across different random seeds. The experiment involves five runs with varied random seeds, maintaining fixed hyperparameters ($\alpha = 1$, $\beta = 0.1$, $\gamma = 1$). The table reports the action prediction accuracy on positive test data and mortality rates using both step-by-step and complete trajectory methods. The mean (μ) and standard deviation (σ) across the five runs are provided, demonstrating consistent performance with an average action prediction accuracy of 94.36%, a mortality rate of 2.29% (step-by-step), and 0.11% (complete trajectory). These results underscore the robustness of our model, indicating minimal variability in outcomes across different random seeds.

TABLE VIII
SENSITIVITY EXPERIMENT TO CHECK STABILITY OF RESULTS WITH 5 RUNS WITH DIFFERENT RANDOM SEEDS FOR $\alpha = 1$ HERE, $\beta = 0.1$ AND $\gamma = 1$.

Random Seed	Action Prediction Accuracy on Positive Test Data %	Mortality %	
		Step by Step	Complete Traj.
1	94.6	2.61	0.18
2	93.6	1.89	0.18
3	94.1	2.61	0.0
4	94.9	2.04	0.18
5	94.6	2.29	0.0
$\boldsymbol{\mu \pm \sigma}$	**94.36 $\pm$ 0.51**	**2.29$\pm$0.33**	**0.11$\pm$0.1**

Figure 2 presents 2D histograms that visualize the aggregated actions recommended by the physician (Ground Truth), POSNEGDM, and Behavioral Cloning (BC). These histograms are analogous to those used for qualitative analysis in [16]. The horizontal and vertical bins correspond to the prescribed dosages of Vasopressor and IV fluid, respectively, based on the given policy. An action value of 0 signifies that no drugs are administered to the patient, while an increasing action value corresponds to a higher dosage. These non-zero drug dosages are represented in quartiles. Each grid cell denotes a specific action, with the color reflecting its frequency of occurrence. Notably, the POSNEGDM model appears to mirror the ground truth more accurately for positive cases than the BC model does, a distinction especially noticeable in the action tuple (0,4) for the positive data. The results from Decision Transformer is not included, since it deviates from the ground truth to an even greater extent than Behavioral Cloning.



(a) Ground Truth (Positive)

(b) Ground Truth (Negative)

(c) PosNegDM (Positive)

(d) PosNegDM (Negative)

(e) BC (Positive)

(f) BC (Negative)

Fig. 2. The three rows in the visualization represent the policies as provided by physicians, PosNegDM, and Behavioral Cloning (BC) respectively, each applied to both positive and negative test data. The axis labels correspond to the discretized action space, where '0' signifies no drug administration, and '4' indicates the maximum dosage of a particular drug. Each grid cell represents a specific action, with its color indicating the frequency of its occurrence.

Author Response to JBHI-02530-2023.R1

We would like to express our gratitude to the Editor-in-Chief and the Associate Editor for handling our revised manuscript. Further, we would like to thank the anonymous reviewers for their thorough and insightful comments. In the following, we provide a response to the different comments by the reviewers. We note that since the comments include the previous response document, the previous response document with mentioned edits is also attached.

APPENDIX A
ASSOCIATE EDITOR

Comment 1: Incorporate Changes in Response Letters: In the response to Comment 1 in the Introduction section, the authors mentioned adding references but did not include the modified text. It would be beneficial to explicitly include the updated sentences in the response for clarity.
Response: Thanks for the comment. We have fixed this in the previous response letter. Specifically, the response now says the following:

Thanks a lot for the comment. We have added the references to these lines.

In addressing these challenges, our paper focuses on a learning-based methodology that grounds a continuous treatment strategy in the ongoing condition of the patient. Employing a Reinforcement Learning (RL) approach [6], [7], we leverage its capacity to adapt and optimize treatment plans over time. This affords a dynamic and personalized approach, closely aligned with the evolving needs of the individual.

To overcome these challenges, we delve into the realm of Offline RL (ORL) [14]. Unlike traditional RL, which learns from real-time interactions, offline RL leverages pre-collected datasets, making it particularly relevant for sepsis treatment.

Some added part in the Related Works: In recent literature on sepsis identification and treatment systems based on offline RL, several approaches have been explored. These approaches aim to model the discrete Markov Decision Process (MDP) [11], [16], [17], [18], and then use online RL approaches. However, the discrete feature space may lead to information loss, which affects the accuracy of online RL. Recent works [19], [20] have modeled MDP in continuous state, while the result relies on the model approximation. In this work, we aim to make decisions without learning the model. Further, these works do not explicitly avoid the negative trajectories. On the contrary, POSNEGDM *selects actions from the given set to diverge from the negative trajectories, achieved by maximizing the divergence between the predicted and ground-truth trajectory.*

Comment 2: Detailed Response to Suggestions: The authors have not provided specific responses to the suggestions in Section H. It would be advisable to address each suggestion explicitly, indicating whether the suggestions will be implemented or explaining the reasons for any decisions not to implement them.
Response 2: Thanks a lot for this. We have now revised the Appendix H in the previous responses (Author Response to JBHI-02530-2023 v1). Three of the comments are addressed in the future work part, and the one comment is addressed in the experimental results. Please see the Section H in the previous response letter for detailed response for this.

Comment 3: Consistent Formatting: In Section E. Sepsis data description, there is a repetition of ”Comment 1” and ”Response 1.” This should be corrected for clarity.
Response:

We would like to kindly note that the numbering of comments and responses was based on the section numbers provided here. As Section E of the response letter contained only one comment, it was labeled as comment 1 again. We understand that the response letter will not be included in the published paper, and therefore, we feel that adjusting the format of the response letter may not be necessary. However, we have endeavored to maintain consistency in the formatting of the response letter throughout. Thank you for your understanding.

APPENDIX B
REVIEWER #1

Comment 1: Highlight the importance of real-world validation and integration with healthcare systems for practical feasibility.
Response 1: Real-world validation of the POSNEGDM framework involves comparing its output with actual patient outcomes and incorporating physician feedback to assess accuracy. This process ensures alignment with healthcare professionals' expertise. Given the increasing integration of digital health solutions, real-world validation is feasible and valuable for assessing practical effectiveness.

Once validated, seamless integration with existing healthcare systems becomes achievable. The framework's design, which considers current and past states to predict the next action, facilitates integration with tabular databases commonly used in healthcare settings. This streamlined integration enhances the framework's practical feasibility and adoption in clinical practice, ultimately improving sepsis treatment and patient care.

We have added the following in the conclusion section to incorporate this:

Our work shows promise in enhancing sepsis treatment, with future considerations for deploying the model in clinical practice. Ethical and legal aspects, such as patient privacy and informed consent, require careful attention.

Comment 2: It is advisable for the authors to include a discussion on the limitations of their proposed approach and experimental setup. This can provide a more comprehensive understanding of the study's scope, potential constraints, and areas for future improvement.
Response 2: Our study highlights limitations in the POSNEGDM framework and its experimental setup, including reliance on training data quality and limited interpretability. Additionally, our focus on offline settings with limited data may overlook real-time sepsis treatment complexity. To address these, future research will enhance framework robustness and applicability, improving training data quality, interpretability, and exploring real-time datasets.

We have added the following in the conclusion section to incorporate this: We also acknowledge limitations, including dependence on training data quality, generalization challenges, and interpretability issues, as directions for future work.

Comment 3: Provide a detailed discussion on planned future work, exploring advancements and extensions of the POSNEGDM framework.
Response 3: The POSNEGDM framework has shown promise in improving sepsis treatment outcomes, but future work is needed to enhance its robustness for real-world clinical use. Focus areas include improving interpretability, assessing scalability and generalizability, integrating with existing healthcare systems, and addressing limitations in the mortality classifier and transformer-based model. These efforts aim to contribute to better patient care and outcomes in sepsis management.

We have added the following in the conclusion section to incorporate this: Our work shows promise in enhancing sepsis treatment, with future considerations for deploying the model in clinical practice. Ethical and legal aspects, such as patient privacy and informed consent, require careful attention. The model may need periodic adaptation to address real-world clinical complexities. We also acknowledge limitations, including dependence on training data quality, generalization challenges, and interpretability issues, as directions for future work.

APPENDIX C
REVIEWER #2

We believe that the comments in AE comments are that of Reviewer #2. Thus, we have responded to the comments in the response to AE.

Author Response to JBHI-02530-2023 v1

We would like to express our gratitude to the Editor-in-Chief and the Associate Editor for handling our manuscript. Further, we would like to thank the anonymous reviewers for their thorough and insightful comments. In the following, we present a point-to-point response to each individual comment from the reviewers and elaborate on how the manuscript has been revised accordingly. Corresponding changes in the paper are marked in blue color.

APPENDIX A
EDITOR-IN-CHIEF

Recommendation:

Manuscript requires a Major Revision before it can be accepted for publication in J-BHI.

Response: Thanks for giving us the opportunity to revise the paper. We hope that the revision addresses the raised comments.

APPENDIX B
ASSOCIATE EDITOR

Comment:

First of all, thanks for your trust on getting published your results on J-BHI. All the reviewers have agreed that there is a lot of potential on your research but there are some flaws that impede the acceptance in its current stage. Therefore, decision is reject and resubmit, based on the different comments from the reviewers. If you can address such concerns (ie. stress more clearly the novelty of your approach and make public the code) and provide the requested information, we would be happy to review again the paper. Please use and answer the different recommendations and comments raised in the reviews as that will improve the quality of your paper, being key for final acceptance/rejection.

Response: Thank you for your detailed feedback. We have meticulously addressed these comments from the reviewers in the following sections.

APPENDIX C
REVIEWER #1

Comment:

The authors have introduced POSNEGDM for sepsis treatment. The paper is well-written and the experiments are well-designed. This method is built from the transformer model and a feedback reinforcer. However, the novelty is not enough. I expect the authors to describe more details of the causal transformer in figure 1. The details of the algorithm are needed.

Response:

We would like to thank the reviewer for going over the paper. Regarding the novelty, we would like to mention the key points of novelty in the paper, expanded also in the revision:

- Integration of Mortality Classifier that reinforces treatment decisions towards positive outcomes, thereby improving patient survival rates significantly.
- Transformer-Based Decision Maker predicting next action and state, allowing the model to anticipate both the immediate action and the ensuing state, providing dual insights and enhancing the decision maker's performance.
- The POSNEGDM framework is the first offline RL-based model that leverages both positive and negative reinforcement to reduce mortality by explicitly learning to avoid negative trajectories.

We have added the following in the Introduction section in the revision for the details of casual transformer:

As shown in Figure 1, the POSNEGDM framework consists of two main components: a mortality classifier (also works as a feedback reinforcer), and a transformer-based decision maker. The mortality classifier is responsible for predicting the probability of mortality for a given patient based on their clinical data. The transformer-based decision maker takes the patient's clinical data as input and generates a sequence of treatment decisions along with the next states. The feedback reinforcer (trained and frozen Mortality Classifier) takes the next states, evaluates the effectiveness of the treatment decisions, and provides feedback ($L_{survival}$ loss, Eq. 2) to the decision maker to improve future decisions. Once the Decision Maker is ready, it generates actions that lead to the live states and save patients.

The Overall Framework: POSNEGDM (SubsectionIII-C) has been revised to reflect Algorithm 1 and the details as follows: Algorithm 1 outlines the training procedure for the POSNEGDM framework, consisting of a Mortality Classifier (MC) and a Decision Maker (DM). The Mortality Classifier is initially trained on the mortality ground truth using cross-entropy loss. The trained Mortality Classifier, denoted as $w_{T_{MC}}$, is then frozen and serves as the Feedback Reinforcer. Subsequently, the Decision Maker is trained using a combination of losses, including survival loss ($\mathcal{L}_{survival}$) adversarially enforced by the Mortality Classifier. The DM's weights, denoted as $w_{T_{DM}}$, are updated iteratively based on the overall loss calculated from the mortality prediction, state prediction, and action prediction. This process aims to enhance the Decision Maker's ability to predict actions leading to patient survival by leveraging the adversarial feedback from the Mortality Classifier.

We note that Algorithm 1 table is also added in this revision.

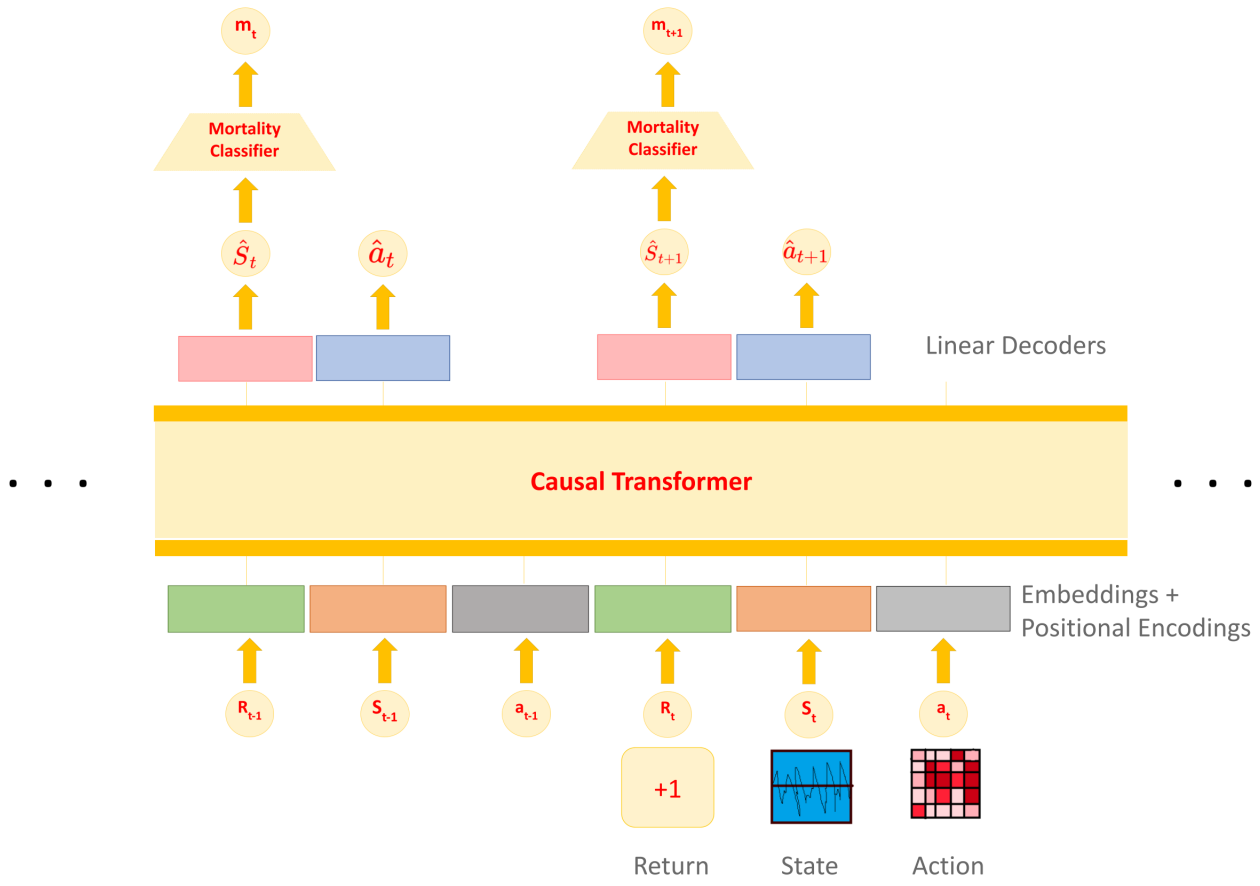


Fig. 3. The DUALSIGHT decision maker takes in states, actions, and returns as input, which are first embedded into linear representations that are specific to each modality. The positional episodic timestep encoding is added to the input to help the model understand the order of events. The tokens are then fed into the GPT architecture, which uses a self-attention mechanism to predict actions and next states. The causal mask ensures that the model can only attend to previous tokens, preserving the causality of the system. The predicted states are subsequently input into the trained Mortality Classifier to assess whether the implemented action guides the patient towards a deceased state. The mortality prediction m_t is employed to influence the DUALSIGHT, compelling it to choose actions aligned with the mortality classifier's prediction of an alive state. This process integrates mortality considerations into the decision-making mechanism, emphasizing the importance of actions that contribute to favorable patient outcomes.

APPENDIX D
REVIEWER #2

Summary of review:
Thank you for your compelling article. The relevance of the topic and the innovative approach presented make this work a potential fit for IEEE JBHI. However, I have some concerns regarding the methodology employed, and I believe that addressing these concerns transparently and adequately is crucial before a further decision can be made. Given that the work utilizes a publicly available dataset, I would personally consider it an acceptance requirement for the authors to provide the full code necessary to replicate the results presented.

Response: We express our gratitude to the reviewer for their review of our paper. We have uploaded our code here: https://github.com/Dipeshtamboli/PosNegDM-Reinforced-Sequential-Decision-Making-for-Sepsis-Treatment. We have added the link in the paper.

Comment 1:
The mortality classifier appears to be a central element of your work, but its operational details are not entirely clear. My primary concerns are:

(a) On which part of the MIMIC-III dataset was the classifier trained (training set, test set, or both)?
(b) How were the respective hyperparameters selected?
(c) Additionally, the reported accuracy of 96.7% seems almost unnaturally high.
  (i) Could you contextualize this figure by comparing it with similar works?
  (ii) Also, how was the accuracy calculated, especially given the imbalanced nature of the dataset? This question extends to the entire article.

Response 1:

(a) First, the training of the Mortality Classifier and the Decision Maker is completely disjoint. Once we train the Mortality Classifier, we freeze the parameters of it and only use it in the *eval* mode. We trained the Mortality Classifier using a 70-30% split between the training and test sets (Table II). Subsequently, we evaluated the classifier's performance on the test set, and our reported accuracy is based on this independent test set assessment. These details are also in the Algorithm details Algorithm 1 outlines the training procedure for the POSNEGDM framework, consisting of a Mortality Classifier (MC) and a Decision Maker (DM). The Mortality Classifier is initially trained on the mortality ground truth using cross-entropy loss. The trained Mortality Classifier, denoted as $w_{T_{MC}}$, is then frozen and serves as the Feedback Reinforcer. Subsequently, the Decision Maker is trained using a combination of losses, including survival loss ($\mathcal{L}_{survival}$) adversarially enforced by the Mortality Classifier. The DM's weights, denoted as $w_{T_{DM}}$, are updated iteratively based on the overall loss calculated from the mortality prediction, state prediction, and action prediction. This process aims to enhance the Decision Maker's ability to predict actions leading to patient survival by leveraging the adversarial feedback from the Mortality Classifier.

(b) The hyperparameters, including the learning rate, number of epochs, and layers, underwent fine-tuning through an iterative process within a standard k-fold cross-validation setup. This procedure involved adjusting these settings and evaluating the model's performance on a validation dataset to identify configurations leading to the highest validation accuracy. We have added in Section V that The hyperparameters, including the learning rate, number of epochs, and layers, underwent fine-tuning through an iterative process within a standard k-fold cross-validation setup. This procedure involved adjusting these settings and evaluating the model's performance on a validation dataset to identify configurations leading to the highest validation accuracy.

(c) (i) As we are the first one proposing the mortality classifier and its use case, we cannot compare it with any previous work.

(ii) Mortality Classifier predicts if the given input state is a deceased state or not (1 or 0). We are using batches with a balanced number of samples from both classes to adjust for the unbalanced nature of the dataset. We are calculating that accuracy similar to the binary classification task by checking if the prediction matches with the ground truth. Given that we have a trained mortality classifier, we are calculating the survival rate by calculating predicted trajectories that do not contain any deceased states at any point.

Comment 2:

In the evaluation (Subsection A. Achieving Low Mortality with POSNEGDM): The mortality rates reported, how exactly were they calculated? Were they derived using the mortality classifier? If so, this raises another concern as it's expected that an approach incorporating this model in its loss function would perform notably better than models that do not (the baselines). This aspect is particularly crucial since improved mortality rates appear to be the main advantage of the framework presented. This needs to be clarified and addressed further. There's also a concern that the mortality classifier may leak test data if used improperly.

Response 2: Mortality rates are determined during model testing by assessing whether the predicted next state is categorized as alive or deceased. The evaluation employs a trained mortality classifier in inference mode to make this determination. It is essential to clarify that the mortality classifier is trained independently on the training set and does not confer an advantage to the POSNEGDM framework. The same mortality classifier is used consistently to assess the performances of Decision Transformer (DT) and Behavioral Cloning (BC). Integrating the mortality classifier's feedback in the objective makes it possible for our algorithm to explicitly avoid negative trajectories, which is our key novelty and has been shown to significantly lower the mortality. Both the mortality classifier and decision maker use the same training dataset as our baselines. The test dataset is disjoint with the training one, so there is no test data leakage when using the mortality classifier.

We have modified subsection V-A to have the aforementioned changes.

Comment 3:

How were the hyperparameters of the entire framework selected? The article suggests that these were evaluated on the test set, implying that the provided performance metrics may represent an upper bound and possibly indicate a degree of overfitting. Please clarify this aspect.

Response 3:

We have revised the evaluation results and the Ablation Studies V-B to reflect the following:

The hyperparameters, including the learning rate, number of epochs, and layers, underwent fine-tuning through an iterative process within a standard k-fold cross-validation setup. This procedure involved adjusting these settings and evaluating the model's performance on a validation dataset to identify configurations leading to the highest validation accuracy.

The parameters α, β, and γ provide control over our model's training objectives, allowing us to balance considerations such as prioritizing negative trajectory avoidance or achieving high action prediction accuracy. The impact of varying these parameters on the model's behavior and performance is demonstrated in Tables V, VI, and VII for actions, states, and adversarial training, respectively.

In Table VIII, we present the results of a sensitivity experiment conducted to assess the stability of our model across different random seeds. The experiment involves five runs with varied random seeds, maintaining fixed hyperparameters ($\alpha = 1$, $\beta = 0.1$, $\gamma = 1$). The table reports the action prediction accuracy on positive test data and mortality rates using both step-by-step

and complete trajectory methods. The mean (μ) and standard deviation (σ) across the five runs are provided, demonstrating consistent performance with an average action prediction accuracy of 94.36%, a mortality rate of 2.29% (step-by-step), and 0.11% (complete trajectory). These results underscore the robustness of our model, indicating minimal variability in outcomes across different random seeds.

We also note that higher action prediction accuracy does not necessarily correlate with lower mortality rates, and that it is essential to explicitly study and consider mortality in decision making.

We have added Table VIII to Subsection A. Moreover, we have also modified Section V to include how we have chosed the hyperparameters.

Comment 4:
It would be beneficial to see multiple runs, along with respective statistics, to evaluate the stability of the results presented. Moreover, if possible, repeating the experiments with multiple train/test splits would add immense value to the robustness of the findings.

Response 4: In Table VIII, we present the results of a sensitivity experiment conducted to assess the stability of our model across different random seeds. The experiment involves five runs with varied random seeds, maintaining fixed hyperparameters ($\alpha = 1$, $\beta = 0.1$, $\gamma = 1$). The table reports the action prediction accuracy on positive test data and mortality rates using both step-by-step and complete trajectory methods. The mean (μ) and standard deviation (σ) across the five runs are provided, demonstrating consistent performance with an average action prediction accuracy of 94.36%, a mortality rate of 2.29% (step-by-step), and 0.11% (complete trajectory). These results underscore the robustness of our model, indicating minimal variability in outcomes across different random seeds.

We have added Table VIII in Section A.

Comment 5: Minor Points:

1) A more extensive discussion of the results and a comparison with similar research would have been appreciated.
2) Lines 46-49: Please provide a reference.
3) EQ 2: Couldn't the loss become undefined if MC(.) evaluates to 1?

Response 5:

1) In our approach, we introduce a novel perspective in Sepsis treatment by explicitly leveraging negative trajectories, a strategy not explored in existing research. While conventional machine learning models, such as Decision Transformer (DT) and Behavioral Cloning (BC), have demonstrated high action prediction accuracy, they lack the explicit consideration of negative outcomes. Our novel framework, named POSNEGDM, explicitly addresses this gap by incorporating a Mortality Classifier that also acts as a Feedback Reinforcer. The Mortality Classifier allows us to assess the potential negative outcome (mortality) of each action, enabling the reinforcement of the model to avoid actions that lead to unfavorable states. This explicit focus on negative trajectories distinguishes our work from previous studies and contributes to the development of a more robust and clinically relevant Sepsis treatment model. The decision to compare our model against DT and BC is motivated by their popularity and effectiveness in predicting actions; however, their omission of explicit consideration for negative trajectories positions our POSNEGDM framework as an innovative and advantageous alternative in improving patient outcomes in Sepsis treatment.
   We have revised the Related Work (Section II) to reflect the above changes.
2) Added [3].
3) This equation is the same as the discriminator part of Eq. 1 from the GAN paper [33]. Also, when MC(.) is 1, meaning DUALSIGHT is outputting a state that is not alive, resulting in a huge loss value (PyTorch takes care of the numerical instability) that we need to optimize. If DUALSIGHT is trained, it will produce alive states that will make the loss value zero. The survival loss, denoted as $L_{survival}(MC, \text{DUALSIGHT}, \tau)$, is formulated as follows:

$$L_{survival}(MC, \text{DUALSIGHT}, \tau) = \mathbb{E}_{\tau \sim P_{data}(\cdot)}[\log(1 - MC(\text{DUALSIGHT}(\tau)) + \epsilon)],$$

   where $1 >> \epsilon > 0$ is introduced to address numerical instability, a practice akin to [33]. This regularization term is not explicitly mentioned in the paper for brevity and clarity.
   [33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

APPENDIX E
REVIEWER #3

Summary of review:
The authors have used reinforcement learning-based classifier for sepsis identification and treatment systems. The reviewer appreciates the effort of the author but the following suggestions may be included for the further improvement of the manuscript.

Response:
We extend our sincere appreciation to the reviewer for their valuable insights and comments on our manuscript. We have addressed/included the suggestions in the main paper.

Comment 1:
A review of recent literature regarding sepsis identification and treatment systems based on offline RL could be discussed.
Response 1: In recent literature on Sepsis identification and treatment systems based on RL, several approaches have been explored. [11] use k-means clustering to discretize the feature representation and model the transition distribution within this discrete space of the Markov Decision Process (MDP), employing Q-value iteration for deriving an effective treatment policy. [16] and [18] adopt variations of Deep Q-learning, with [16] utilizing raw physiological features and [18] employing a recurrent autoencoder to capture temporal information. [18] further introduces a Mixture of Experts approach, combining the Deep Q-learning policy with a kernel-based policy. In work by [19], continuous state-space model-based RL is considered, incorporating an environment model to capture transition dynamics and utilizing the Proximal Policy Optimization algorithm ([7]) for learning treatment policies. The feature representation in this study involves the concatenation of raw physiological features over multiple timesteps, integrating temporal information.

Despite the advancements in these approaches, there are notable limitations. The discretization of feature representations in the work by [11] may lead to information loss, and the effectiveness of Q-value iteration depends on the accuracy of the discretization. Deep Q-learning variants face challenges in preprocessing techniques, with potential limitations in capturing nuanced temporal patterns. The continuous state-space model-based RL, as employed by [19], may have scalability issues and requires careful modeling of transition dynamics. These limitations highlight the need for further research to address challenges and enhance the robustness of sepsis treatment models based on offline RL.

Our proposed POSNEGDM framework addresses these limitations by introducing a Transformer-based model, a mortality classifier, and a feedback reinforcer, outperforming existing ML-based algorithms. Notably, our model explicitly considers negative trajectories, a novel approach not explored in previous works, enhancing its ability to make informed, patient-specific treatment decisions.

We added in the related work: In recent literature on Sepsis identification and treatment systems based on RL, several approaches have been explored. [11] discretized the feature representation and model the transition distribution within this discrete space of the Markov Decision Process (MDP), employing Q-value iteration for deriving an effective treatment policy. [16], [17], [18] adopt variations of Deep Q-learning. In work by [19], continuous state-space model-based RL is considered, incorporating an environment model to capture transition dynamics and utilizing the Proximal Policy Optimization algorithm for learning treatment policies. While these works focus on online approaches, we focus on offline RL.

For relating other approaches for Sepsis, we added:

In recent literature on Sepsis identification and treatment systems based on offline RL, several approaches have been explored. These approaches aim to model the discrete Markov Decision Process (MDP) [11], [16], [17], [18], and then use online RL approaches. However, the discrete feature space may lead to information loss, which affects the accuracy of online RL. Recent works [19], [20] have modeled MDP in continuous state, while the result relies on the model approximation. In this work, we aim to make decisions without learning the model. Further, these works do not explicitly avoid the negative trajectories. On the contrary, POSNEGDM selects actions from the given set to diverge from the negative trajectories, achieved by maximizing the divergence between the predicted and ground-truth trajectory.

Comment 2:
Results related to the mortality classifier in Section III-A might confuse the reader about which test data is being used.
Response 2: The training of the Mortality Classifier and the Decision Maker is entirely disjoint in our approach. Once we train the Mortality Classifier, we freeze its parameters and only use it in the evaluation mode. We trained the Mortality Classifier using a 70-30% split between the training and test sets (Table II). Subsequently, we evaluated the classifier's performance on the test set, and our reported accuracy is based on this independent test set assessment. It is essential to note that the Mortality Classifier is trained independently of POSNEGDM, and its training does not directly affect the training of POSNEGDM. However, we ensured that the Mortality Classifier is trained on the same training data used for POSNEGDM, specifically the training set of the dataset (Table II). Additionally, the final mortality is calculated based on the predicted state, as predicted by DUALSIGHT, and is different from the test set used to validate the Mortality Classifier's performance.

To further explain, we added Algorithm 1 Table and description: Algorithm 1 outlines the training procedure for the POSNEGDM framework, consisting of a Mortality Classifier (MC) and a Decision Maker (DM). The Mortality Classifier is initially trained on the mortality ground truth using cross-entropy loss. The trained Mortality Classifier, denoted as $w_{T_{MC}}$, is then frozen and serves as the Feedback Reinforcer.

Comment 3: Results regarding cross-validation techniques could be added to prove the robustness of the proposed method against data-biasing.
Response 3: Table VIII displays the outcomes of a sensitivity experiment designed to evaluate the robustness of our model under various random seeds. The experiment comprises five runs with distinct random seeds while maintaining constant hyperparameters ($\alpha = 1$, $\beta = 0.1$, $\gamma = 1$). The table provides details on action prediction accuracy for positive test data and mortality rates, employing both step-by-step and complete trajectory methods. The mean (μ) and standard deviation (σ) across the five runs showcase consistent performance, with an average action prediction accuracy of 94.36%, mortality rates of 2.29% (step-by-step) and 0.11% (complete trajectory). These findings underscore the stability and reliability of our model, revealing

minimal variability in outcomes across different random seeds.
We have added Table VIII in Appendix A.
Comment 4: A comparison of the proposed method with recent offline RL methods for sepsis treatment, especially with the other approaches on the MIMIC-III dataset, could be added.
Response 4: As pioneers in leveraging negative trajectories for both predicting actions at specific states and reinforcing our model to guide actions towards patient survival, direct comparison with previous methodologies is not feasible. Our unique approach of utilizing negative trajectories aims to avoid deceased states, setting our method apart. Nevertheless, we have conducted comparative evaluations with two state-of-the-art methods focused on action prediction accuracy. While these benchmarks excel in predicting actions, their models lack the capability to systematically train toward transitioning patient states to a state of vitality.

Comment 5: References should be proofread.
Response 5: We have taken a pass on the references.

APPENDIX F
REVIEWER #4

We express our gratitude to the reviewer for their thoughtful feedback and constructive comments. We have carefully addressed each comment section by section.

A. Abstract

Comment 1:
The abstract contains some jargon (comprehensive ablation studies) and abbreviations such as 'ML' (Machine Learning), 'DT' (Decision Transformer), 'BC' (Behavioral Cloning), etc. that may not be immediately understandable to a broad audience. It is advisable to avoid abbreviations in the abstract and clarify terms for better accessibility.
Response 1: Revised Abstract —
Sepsis, a life-threatening condition triggered by the body's exaggerated response to infection, demands urgent intervention to prevent severe complications. Existing machine learning (ML) methods for managing sepsis struggle in offline scenarios, exhibiting suboptimal performance with survival rates below 50%. This paper introduces the POSNEGDM framework, utilizing an innovative transformer-based model and a feedback reinforcer to replicate expert actions while considering individual patient characteristics. A mortality classifier with 96.7% accuracy guides treatment decisions towards positive outcomes. The POSNEGDM framework significantly improves patient survival, saving 97.39% of patients, outperforming established ML algorithms (Decision Transformer and Behavioral Cloning) with survival rates of 33.4% and 43.5%, respectively. Additionally, ablation studies underscore the critical role of the transformer-based decision maker and the integration of a mortality classifier in enhancing overall survival rates. In summary, our proposed approach presents a promising avenue for enhancing sepsis treatment outcomes, contributing to improved patient care and reduced healthcare costs.
Comment 2:
While the abstract mentions that "this paper introduces the POSNEGDM framework," it would be beneficial to briefly describe what POSNEGDM stands for and its components in a few words to give readers a clearer picture right from the abstract.
Response 2: POSNEGDM stands for "Reinforcement Learning with Positive and Negative Demonstrations for Sequential Decision-Making." It is an offline RL-based model that leverages both positive and negative trajectories to improve sepsis treatment. The key components of the proposed architecture are a mortality classifier, a Transformer-based model that predicts the next state and action based on the historical patient states and treatments, and an efficient loss function that includes both next-state and next-action prediction loss as well as a reinforced feedback loss based on the mortality classifier. We have added this explanation in the Introduction (Section I) and Abstract.

B. Introduction

Comment 1:
The introduction does a good job of highlighting the critical issue of sepsis and the need for improved treatment strategies. However, the problem statement could be more explicitly defined, especially the challenges of one-size-fits-all approaches, delayed or inadequate treatment modifications, and limited prediction capabilities. To support these problem statements, consider adding references to relevant studies, reports, or statistics that highlight the issues you mention.
Response 1: Thanks a lot for the comment. We have added the references to these lines.

In addressing these challenges, our paper focuses on a learning-based methodology that grounds a continuous treatment strategy in the ongoing condition of the patient. Employing a Reinforcement Learning (RL) approach [6], [7], we leverage its capacity to adapt and optimize treatment plans over time. This affords a dynamic and personalized approach, closely aligned with the evolving needs of the individual.

To overcome these challenges, we delve into the realm of Offline RL (ORL) [14]. Unlike traditional RL, which learns from real-time interactions, offline RL leverages pre-collected datasets, making it particularly relevant for sepsis treatment.

Some added part in the Related Works: In recent literature on sepsis identification and treatment systems based on offline RL, several approaches have been explored. These approaches aim to model the discrete Markov Decision Process (MDP) [11], [16], [17], [18], and then use online RL approaches. However, the discrete feature space may lead to information loss, which affects the accuracy of online RL. Recent works [19], [20] have modeled MDP in continuous state, while the result relies on the model approximation. In this work, we aim to make decisions without learning the model. Further, these works do not explicitly avoid the negative trajectories. On the contrary, POSNEGDM selects actions from the given set to diverge from the negative trajectories, achieved by maximizing the divergence between the predicted and ground-truth trajectory.

Comment 2:

It is mentioned in the introduction that "the introduction of machine learning into sepsis treatment can provide an additional layer of predictive, personalized, and data-driven decision-making support, potentially improving patient outcomes," it would be helpful to briefly elaborate with supporting reference on how machine learning can address the identified challenges in sepsis treatment.

Response 2: We have removed the sentences to make them specific for reinforcement learning. In addressing these challenges, our paper focuses on a learning-based methodology that grounds a continuous treatment strategy in the ongoing condition of the patient. Employing a Reinforcement Learning (RL) approach [6], [7], we leverage its capacity to adapt and optimize treatment plans over time. This affords a dynamic and personalized approach, closely aligned with the evolving needs of the individual.

Comment 3:

The transition from discussing machine learning to introducing Offline Reinforcement Learning (RL) could be smoother. A brief explanation of what offline RL is and why it is particularly relevant to sepsis treatment would aid reader comprehension.

Response 3:

Thanks a lot for the comment. We have first made the transition from machine learning to reinforcement learning, as mentioned in the last comment.

We add the transition to offline reinforcement learning as follows:

To overcome these challenges, we delve into the realm of Offline RL. Unlike traditional RL, which learns from real-time interactions, offline RL leverages pre-collected datasets, making it particularly relevant for sepsis treatment. In our context, the dataset comprises trajectories representing sequences of states, actions, and outcomes observed in past sepsis cases. This shift to offline RL allows us to learn from both positive and negative trajectories, providing a more comprehensive understanding of effective treatment strategies. The utilization of offline RL is a key aspect of our proposed POSNEGDM framework, enhancing its ability to derive impactful insights from historical patient data.

We have modified the Introduction Section to reflect this.

Comment 4:

Please mention the organization of the article and what each section will cover. This can provide readers with a roadmap of what to expect.

Response 4:

The article is organized as follows:

The introduction (Section 1) offers an insightful overview of sepsis and delineates the treatment challenges associated with it, introducing the application of machine learning as a potential solution to enhance treatment outcomes. Section 2, the related work, reviews existing literature on machine learning in sepsis treatment, shedding light on the limitations of current approaches. Section 3, the methodology, intricately describes the POSNEGDM framework, elucidating its components, such as the transformer-based model (DUALSIGHT), the mortality classifier (feedback reinforcer). In Section 4, the experimental results showcase the performance of the POSNEGDM framework and provide a comparative analysis with existing machine learning-based algorithms. Section 5, the discussion, delves into the implications of the results and outlines potential avenues for future research in this domain. Finally, Section 6, the conclusion, succinctly summarizes the key findings of the paper and emphasizes the significance of leveraging machine learning for improved sepsis treatment outcomes.

We have added the above response in the Introduction section.

C. Related Works

Comment 1:

When discussing prior applications of offline RL to sepsis treatment, it would be insightful to provide a brief comparative analysis of these approaches. For instance, highlighting the shortcomings of prior methods and how the proposed approach addresses those gaps can be helpful.

Response 1: Thanks for the comment. We have revised the paragraph as In recent literature on Sepsis identification and treatment systems based on offline RL, several approaches have been explored. These approaches aim to model the discrete Markov Decision Process (MDP) [11], [16], [17], [18], and then use online RL approaches. However, the discrete feature space may lead to information loss, which affects the accuracy of online RL. Recent works [19], [20] have modeled MDP in

continuous state, while the result relies on the model approximation. In this work, we aim to make decisions without learning the model. Further, these works do not explicitly avoid the negative trajectories. On the contrary, POSNEGDM selects actions from the given set to diverge from the negative trajectories, achieved by maximizing the divergence between the predicted and ground-truth trajectory.

Comment 2:

The section described Imitation Learning (IL) methods, including Behavioral Cloning (BC), effectively. However, it would be useful to briefly explain how BC works, especially for readers who might not be familiar with the concept.

Response 2: Behavioral Cloning (BC) employs supervised learning, treating states (x) as input data and actions (y) as target values. For continuous actions, the Mean Squared Error (MSE) loss ($\mathrm{E}((f(s) - a)^2)$) is used, minimizing the difference between predicted ($f(s)$) and actual actions (a). In the case of discrete actions, a cross-entropy loss is employed, measuring the dissimilarity between predicted action probabilities and true action distributions. BC learns to replicate the expert's actions, facilitating the mapping from states to actions through supervised learning.

We have added the above in the Related Work section.

Comment 3:

It would be beneficial to provide a bridge between the related works and the proposed POSNEGDM framework. Specifically, it would be helpful to explain how the proposed approach differs from or improves upon prior methods, especially regarding the utilization of negative trajectories.

Response 3: The last line now indicates that: Further, these works do not explicitly avoid the negative trajectories. On the contrary, POSNEGDM selects actions from the given set to diverge from the negative trajectories, achieved by maximizing the divergence between the predicted and ground-truth trajectory.

D. Proposed Approach

Comment 1:

The section clarifies the model's architecture well. However, briefly mentioning the choice of the Transformer architecture and elaborating on how it learns from historical data and expert actions, and its significance, would enhance clarity.

Response 1: We have revised the DUALSIGHT Decision Maker subsection (III-B) as follows:

A transformer is particularly well-suited for sequential decision-making tasks, such as Sepsis treatment, due to its ability to capture dependencies and relationships between different states in a sequence effectively. The transformer-based model in the POSNEGDM framework learns from historical data and expert actions by processing input sequences through modality-specific linear embeddings and a positional episodic timestep encoding. The tokens are then introduced into a GPT architecture, which anticipates output in an autoregressive fashion, employing a causal self-attention mask. This self-attention module calculates a weighted sum of the input states, with weights assigned based on the similarity among the states, allowing the model to capture dependencies and relationships between different states in the sequence effectively. The significance of this approach is that it enables the model to learn from both positive and negative trajectories, which is key to improving patient survival rates in sepsis treatment. Additionally, the feedback reinforcer in the POSNEGDM framework uses the forecasted next states to ascertain the ultimate survival likelihood of patients, which is subsequently utilized as feedback to fortify the decision maker. This unique configuration significantly enhances the decision-maker's performance.

Comment 2:

The section mentions the decision maker's unique ability to anticipate both immediate actions and ensuing states. Please explain more clearly why this dual insight is essential for improving sepsis treatment and how it contributes to the framework's overall effectiveness.

Response 2: The decision maker's unique ability to anticipate both immediate actions and ensuing states is essential for improving sepsis treatment because it provides dual insights into the patient's condition. By anticipating the immediate action, the decision maker can take steps to address the patient's current state. However, by also anticipating the ensuing state, the decision maker can take steps to prevent the patient's condition from deteriorating further. This is particularly important in sepsis treatment, where the condition can progress rapidly and lead to organ failure and death if not treated promptly.

For instance, focusing solely on action learning, we observe that Decision Transformer (DT) and Behavioral Cloning (BC) exhibit action prediction accuracies comparable to our algorithm. However, their incapacity to discern actions conducive to avoiding unfavorable patient outcomes distinguishes our approach.

Comment 3:

The overall explanation of the POSNEGDM framework is clear, but it should emphasize how the mortality classifier and the decision maker work together to enhance patient survival rates, ensuring the reader understands their synergy.

Response 3: We have revised the Overall Framework: POSNEGDM (Subsection III-C) as follows:

The Mortality Classifier is independently trained to distinguish between alive and deceased states. In the training of our decision-maker, the Mortality Classifier operates in inference (Feedback Reinforcer) mode rather than training mode to avoid interfering with DUALSIGHT's learning. This enables the DUALSIGHT to produce additional gradients derived from the alive/dead state Feedback Reinforcer's adversarial loss (Eq. 2). These supplementary gradients facilitate improved decision-maker training, enhancing its ability to predict actions that result in alive states and consequently reducing the mortality rate.

E. Sepsis data description

Comment 1:
It is mentioned that the data is preprocessed following steps detailed in another reference ([6]). It might be beneficial to briefly summarize what these preprocessing steps entail or why they are necessary to understand the data preparation better.

Response 1: We have revised the Sepsis Data Description (Section IV) to reflect following:
The data preprocessing phase comprised selecting pertinent features, discretizing them, and transforming them into a 4-hour window. Key physiological parameters, encompassing demographics, laboratory values, vital signs, and intake/output events, were systematically gathered for each patient. To facilitate analysis, the data were aggregated into 4-hour windows, with the mean or sum recorded—depending on the nature of the data—when multiple data points were available within a given window.

F. Experimental Results

Comment 1:
It is mentioned that hyperparameters were deliberately chosen, but is not elaborated on the reasoning behind these choices. Explaining why specific hyperparameter values were selected can provide more insight into the model's design and its impact on performance.
Response 1: We have revised the Ablation Studies V-B to reflect the following:
The hyperparameters, including the learning rate, number of epochs, and layers, underwent fine-tuning through an iterative process within a standard k-fold cross-validation setup. This procedure involved adjusting these settings and evaluating the model's performance on a validation dataset to identify configurations leading to the highest validation accuracy.
The parameters α, β, and γ provide control over our model's training objectives, allowing us to balance considerations such as prioritizing negative trajectory avoidance or achieving high action prediction accuracy. The impact of varying these parameters on the model's behavior and performance is demonstrated in Tables V, VI, and VII for actions, states, and adversarial training, respectively.

In Table VIII, we present the results of a sensitivity experiment conducted to assess the stability of our model across different random seeds. The experiment involves five runs with varied random seeds, maintaining fixed hyperparameters ($\alpha = 1$, $\beta = 0.1$, $\gamma = 1$). The table reports the action prediction accuracy on positive test data and mortality rates using both step-by-step and complete trajectory methods. The mean (μ) and standard deviation (σ) across the five runs are provided, demonstrating consistent performance with an average action prediction accuracy of 94.36%, a mortality rate of 2.29% (step-by-step), and 0.11% (complete trajectory). These results underscore the robustness of our model, indicating minimal variability in outcomes across different random seeds.

We also note that higher action prediction accuracy does not necessarily correlate with lower mortality rates, and that it is essential to explicitly study and consider mortality in decision making.

We have added Table VIII to Subsection A. Moreover, we have also modified Section V to include how we have chosed the hyperparameters.

Comment 2:
While the results indicated the superiority of POSNEGDM in terms of mortality rates, it would be beneficial to include a discussion on why it outperforms the baselines, perhaps by highlighting the unique features or mechanisms that contribute to its success.
Response 2: We have revised Achieving Low Mortality with POSNEGDM (Subsection V-A) to add the following:
By mandating the prediction of state outputs and incorporating an adversarial mortality classifier, the model is empowered to enhance its mortality rate predictions. The feedback reinforcer, represented by the mortality classifier, plays a crucial role in offering additional guidance to the model, facilitating learning from errors. In totality, the amalgamation of these distinctive features and mechanisms is the driving force behind the superior performance of the POSNEGDM framework when compared to existing machine learning-based algorithms.

Comment 3:
Authors mentioned additional experiments but did not provide any specific details. Expanding on what these experiments entail and their results would enhance the comprehensiveness of the findings.
Response 3: The "Additional Experiments" section of the paper presents additional analyses and visualizations to evaluate the POSNEGDM framework's performance further. Specifically, the section includes the sensitivity experiments and additional 2D histograms that visualize the aggregated actions recommended by the physician (ground truth), POSNEGDM, and Behavioral Cloning (BC). These histograms provide a qualitative analysis of the actions recommended by each method and highlight the differences in decision-making strategies. The x-axis represents the different discrete actions selected at a given timestep, where 0 represents no drug given, and increasing actions refer to higher drug dosages, where quartiles represent drug dosages. The y-axis represents the frequency of each action taken by the physician or the model. The chart shows the aggregation of actions taken over all timesteps for those cohorts.

We have added the above response in the Additional Experiments (Subsection A).

G. Overall

1) The paper does not compare the framework's performance with other state-of-the-art sepsis treatment approaches, making it difficult to evaluate its effectiveness.
   Response: We propose a novel method where we are considering negative trajectories to utilize as well as predicting the next action in discrete space along with the mortality. Previous works do not calculate the mortality or the action prediction in the same fashion.
2) The study focuses on offline settings with limited data, which may not fully capture the complexity and variability of real-time sepsis treatment scenarios.
   Response: While we acknowledge the potential limitation of offline settings with limited data, it's important to note that the use of the MIMIC-III dataset in sepsis studies is a common practice in the existing literature. The MIMIC-III database is a comprehensive collection of de-identified clinical data from the Beth Israel Deaconess Medical Center, spanning over a decade and including information on a substantial number of adult hospital admissions. The dataset is widely accessible to researchers globally under a data use agreement, providing valuable insights into real-world clinical scenarios.
3) The paper does not address the potential challenges or limitations of implementing the POSNEGDM framework in real-world clinical settings, such as integration with existing healthcare systems or the need for additional resources.
   Response: Our work has shown potential in this, and thus, we are considering adding it to our future work; however, we need to consider the following points as well.
   - Ethical and legal considerations that need to be taken into account when deploying the model in clinical practice, such as issues related to patient privacy and informed consent.
   - The model may not be able to capture all of the complexities and nuances of real-world clinical practice, and it may need to be adapted or updated over time as new data and insights become available

   We added:Our work shows promise in enhancing sepsis treatment, with future considerations for deploying the model in clinical practice. Ethical and legal aspects, such as patient privacy and informed consent, require careful attention. The model may need periodic adaptation to address real-world clinical complexities. We also acknowledge limitations, including dependence on training data quality, generalization challenges, and interpretability issues, as directions for future work.
4) The study does not provide a detailed analysis of the interpretability of the framework's decision-making process, which may be important for gaining trust and acceptance from healthcare professionals. Moreover, Lack of interpretability can be a significant limitation when it comes to deploying AI systems in clinical settings.
   Response: The paper notes that the framework provides insightful treatment recommendations, including the administration of intravenous fluids and vasopressor drugs, which is a significant improvement over existing ML-based algorithms. Additionally, the framework utilizes a feedback reinforcer to improve patient survival rates, which suggests that the framework's decisions are based on a combination of patient data and the learning model. However, we have added this limitation in the conclusions section. Our work shows promise in enhancing sepsis treatment, with future considerations for deploying the model in clinical practice. Ethical and legal aspects, such as patient privacy and informed consent, require careful attention. The model may need periodic adaptation to address real-world clinical complexities. We also acknowledge limitations, including dependence on training data quality, generalization challenges, and interpretability issues, as directions for future work.
5) The limitations of the mortality classifier and transformer-based model used in the framework are not explicitly discussed, leaving room for further investigation and improvement.
   Response: The mortality classifier and transformer-based model in the framework exhibit limitations, such as their dependence on the quality of training data, potential challenges in generalizing to unseen patient populations, interpretability issues, adaptability to evolving clinical practices, and ethical and legal considerations. These constraints highlight areas for further investigation and improvement in future research, aiming to enhance the models' robustness and applicability in real-world clinical settings. We added: Our work shows promise in enhancing sepsis treatment, with future considerations for deploying the model in clinical practice. Ethical and legal aspects, such as patient privacy and informed consent, require careful attention. The model may need periodic adaptation to address real-world clinical complexities. We also acknowledge limitations, including dependence on training data quality, generalization challenges, and interpretability issues, as directions for future work.

H. Suggestions

1) To evaluate the effectiveness and feasibility of the POSNEGDM framework in improving sepsis treatment outcomes, it is necessary to explore and evaluate it in real-world clinical settings and to ensure the model's safety and effectiveness in practice. And to facilitate seamless implementation and adoption in clinical practice, the POSNEGDM framework should be integrated and tested with existing healthcare systems.
   Response: Our model is designed with deployment capabilities, and we acknowledge the necessity of future studies to validate its performance and safety in real-world scenarios. We are optimistic about the potential for practical application

and look forward to future research initiatives that can explore the clinical utility of the POSNEGDM framework. Specifically, we have added in the revised future work: Our work shows promise in enhancing sepsis treatment, with future considerations for deploying the model in clinical practice.

2) To gain trust and acceptance from healthcare professionals, it is important to investigate the interpretability of the framework's decision-making process.
   Response: Recognizing the crucial aspect of gaining trust and acceptance from healthcare professionals through interpretability, we acknowledge that our current work lacks a dedicated module to explain the decision-making process explicitly. However, we plan to address this in our future work by incorporating an interpretable component that can elucidate the actions taken by the framework. We appreciate the significance of this aspect and are committed to enhancing the transparency and interpretability of the POSNEGDM framework to align with the needs of healthcare professionals. Specifically, we have added in the future work: We also acknowledge limitations, including dependence on training data quality, generalization challenges, and interpretability issues, as directions for future work.
3) To determine the performance and superiority of the POSNEGDM framework, it should be compared with other state-of-the-art sepsis treatment approaches.
   Response: While we understand the importance of benchmarking the POSNEGDM framework against other sepsis treatment approaches, the unique methodology we employ, particularly in calculating mortality, makes direct comparisons challenging. The incorporation of negative trajectories sets our approach apart, and a direct comparison is not possible. We have chosen Decision Transformer (DT) and Behavioral Cloning (BC) for comparison, as they excel in imitating actions and predicting the next state with high accuracy (95.1% for BC and 94.3% for DT). However, it is noteworthy that despite their impressive accuracy, these models fall short of effectively reducing mortality rates.
   We have added the following in the revision:
   In terms of baselines, we compare our POSNEGDM model to two established approaches: the state-of-the-art offline reinforcement learning algorithm, Decision Transformer (DT) [28], and the widely used imitation learning algorithm, Behavioral Cloning (BC) [23]. It is worth noting that other imitation learning algorithms like GAIL [12] and AIRL [13] require a simulator, making them unsuitable as baselines in our specific scenario.
4) To assess the scalability and generalizability of the framework across different patient populations and healthcare settings, studies should be conducted.
   Response: Conducting studies to assess scalability and generalizability across different datasets and patients thoroughly is on our agenda as we continue to advance the framework's applicability. However, we are limited by the available data for the study. We have added in the future work: We also acknowledge limitations, including dependence on training data quality, generalization challenges, and interpretability issues, as directions for future work.