Intelligent Medical Decision Making for Sepsis Detection using Reinforcement Learning

Lakshita Singh[1,2], Lakshay Kamra [1], Muskan Agarwal [1], Anjana Gupta[1], H.C. Taneja [1]

[1] *Department of Applied Mathematics, Delhi Technological University, Delhi, India*
[2] *Corresponding Author*

Abstract

When the body's defense against an infection damages its own tissues and causes organ malfunction, it develops sepsis, a catastrophic medical illness. Administering intravenous fluids and antibiotics promptly can increase the patient's chances of survival. In order to determine the best treatment plans for septic patients, this study investigates the application of deep reinforcement learning and continuous state-space models. The method produces clinically comprehensible policies that could assist doctors in intensive care in empowering medical professionals to make informed decisions that ultimately enhance the prospects of patient survival.

Keywords

Dueling Double Deep Q Learning Networks (DDDQN), Sepsis, MIMIC III, Deep-Q.

1. Introduction

Sepsis is a clinical syndrome caused by the invasion of bacteria and/or toxins that triggers a harmful reaction in the body, leading to severe morbidity and mortality [1]. Failure to detect and manage this condition early can result in organ failure, septic shock, and death. To improve patient outcomes, it is crucial to detect sepsis as soon as possible, as each hour of delayed treatment after hypotension increases the risk of dying from septic shock by 7.6%. Recent studies have shown that administering a 3-hour bundle of care for sepsis patients, including a blood culture, broad-spectrum antibiotics, and lactate measurement, can significantly reduce in-hospital mortality [3]. Therefore, timely and aggressive treatment is essential in managing sepsis. Even experienced professionals face difficulties in diagnosing sepsis early and accurately, as its symptoms can be easily confused with those of other medical conditions. However, the electronic health record (EHR) already captures data that could aid in predicting sepsis, despite the challenges that come with the diagnosis of this condition [2]. Hence, early warning scores that rely on data from the EHR hold great promise in detecting early clinical deterioration in real-time. The National Early Warning Score (NEWS) was developed, validated, and implemented by the Royal College of Physicians to detect patients who are acutely decompensating. NEWS employs six physiological variables and compares them to their expected ranges to produce a single composite score. In addition to antibiotics, intravenous fluids and vasopressors are used in severe cases. However, patients' mortality rates vary considerably depending on the fluid and vasopressor therapy methods used, highlighting the importance of making the right choices. In the realm of sepsis management, the absence of tailored real-time decision support has posed significant challenges for healthcare providers despite international efforts to provide general guidelines. In response to this pressing issue, we present a pioneering data-driven approach that leverages advanced deep reinforcement learning (RL) algorithms to optimize sepsis treatment strategies. This study builds upon previous research and seeks to enhance the likelihood of septic patients' survival in the ICU by utilizing continuous- state space models and shaped reward functions to identify the most effective course of action. Our findings represent a crucial contribution to the field of sepsis treatment, as they pave the

IVUS 2023: Information Society and University Studies
EMAIL: lakshitasingh1806@gmail.com (L. Singh); kaylakshay@gmail.com (L. Kamra); muskanagarwal47@gmail.com (M. Agarwal); anjanagupta@dce.ac.in (A.Gupta); hctaneja@dce.ac.in (H.C. Taneja)


way for personalized and real-time decision-making strategies that have the potential to transform patient outcomes and reduce mortality rates. We chose RL over supervised learning because there is a lack of consensus in the medical literature on what constitutes an effective treatment approach. It is worth noting that RL algorithms enable us to derive optimal strategies from training samples that do not correspond to optimal behavior. Our primary emphasis lies in the development of continuous state-space modeling, a sophisticated methodology that utilizes a patient's physiological data from the ICU to represent their current physiological state as a continuous vector at any given instant. We use Deep-Q Learning to determine the appropriate responses. Our study presents remarkable contributions in the realm of patient care, including the generation of treatment plans that have the potential to augment patient outcomes and significantly decrease patient mortality rates. We achieved this by implementing advanced deep reinforcement learning models that incorporate continuous-state spaces and precisely designed reward functions [3].

2. BACKGROUND & MOTIVATION

The initial diagnosis of sepsis poses a daunting challenge owing to its inconspicuous presentation, characterized by clinical manifestations resembling those of less severe ailments [5]. While international initiatives try to offer generic recommendations for managing sepsis, doctors at the bedside still lack effective technologies to offer tailored real-time decision support. Developing and validating early warning scores to forecast clinical deterioration and other related outcomes has been the subject of a significant amount of research. For instance, two of the most popular scores used to gauge overall clinical deterioration are the MEWS score and NEWS score. Additionally, the systemic inflammatory response syndrome (SIRS) score (Fig 1) was a component of the initial clinical definition of sepsis, however more recently, other sepsis-specific scores have gained popularity, including SOFA and qSOFA. The Rothman Index, a more complex regression-based method, is also often used to identify general deterioration. In numerous related investigations, multitask Gaussian processes were also used to simulate multivariate physiological time series. Several studies utilized a model that was comparable to ours but placed more emphasis on forecasting vital signs to predict clinical instability.

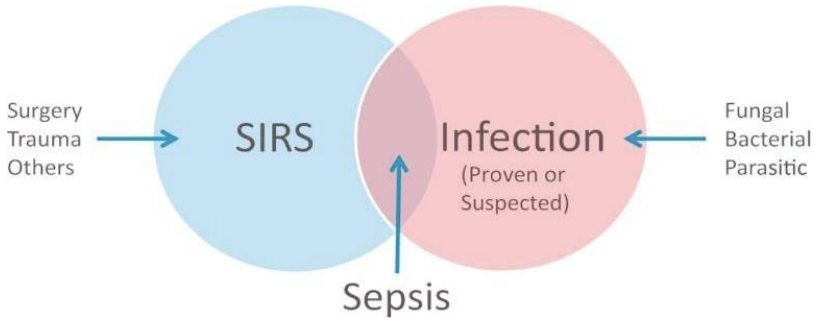

Figure 1: The systematic inflammatory response syndrome [6].

2.1. SERA Algorithm

The SERA algorithm is a risk assessment tool designed to identify patients who may be at risk for sepsis. The algorithm uses both structured and unstructured data from patient consultations to make a prediction. The algorithm is designed to operate on a patient-by-patient basis, with each consultation serving as an analytical unit. It is composed of two interrelated algorithms: the diagnosis algorithm and the early prediction algorithm. When a patient is examined, the diagnosis algorithm determines if the patient is presently suffering from sepsis.

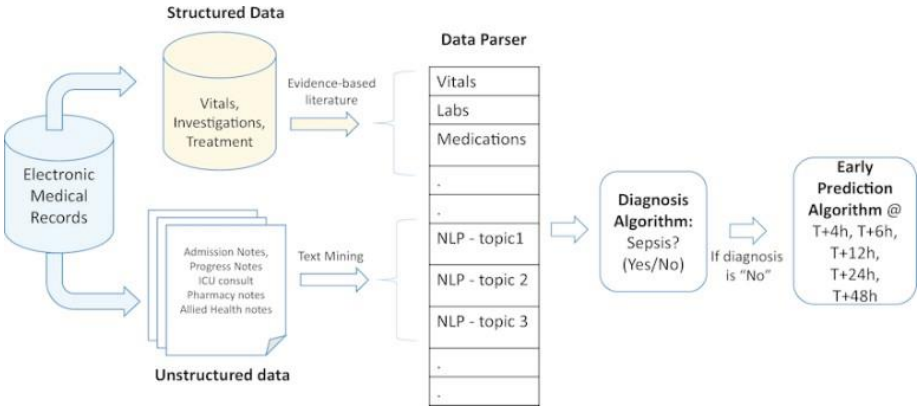

Figure 2: The development process of the SERA algorithm, a tool that utilizes both structured and unstructured data to diagnose and predict sepsis, is depicted in a flow diagram. The algorithm is designed to function in a standard clinical setting where physicians rely on various data sources to analyze and diagnose patients [4].

On the other hand, the early prediction algorithm ascertains whether sepsis is likely to manifest within the next four hours if the patient does not already have the condition. The algorithm incorporates both structured and unstructured data in its processing. While structured data entails vital signs, investigation results, and treatment details, the unstructured data encompasses clinical notes. Developed to operate in a typical clinical setting, where physicians utilize both types of data to analyze and diagnose patients, the algorithm's construction procedures are illustrated in the elaborate flow diagram presented in Fig 2. Supervised learning, particularly in medical applications, has been hindered by the challenge of frequently missing labels per time point in time series datasets. This issue also affects early diagnosis of sepsis. Prior research has addressed the problem of defining resolved sepsis labels by utilizing ad-hoc approaches. These studies have relied on readily available ad-hoc criteria to predict the onset of sepsis and have used a global time series label, such as an ICD illness number designed for billing purposes, to define resolved sepsis labels.

2.2 Algorithms for the Early Detection of Sepsis

Over the past 10 years, several data-driven approaches for detecting sepsis in the ICU have been proposed. Numerous methods compare only certain clinical scores, including SIRS, NEWS, or MEWS. None of these ratings, meanwhile, are meant to serve as precise, ongoing sepsis risk scores. Doctors now view the SIRS criteria as being non- specific and out of date for the definition of sepsis. A targeted real-time warning score (TREWScore) was presented as an alternative to these scores to predict septic shock, which is a common consequence after sepsis. Notably, even though numerous machine learning techniques have outperformed general-purpose or oversimplified clinical schemes, almost no articles have actually made a direct comparison to other machine learning techniques in the literature. It has been demonstrated that using LSTMs is better than using the InSight model. Modern technology Sepsis prevalence numbers range from 6.6% to 21.4%, and real-world datasets with these prevalence values are typically used to build sepsis detection techniques.

2.3 Reinforcement Learning in Medicine

Reinforcement Learning, an intricate framework for optimizing sequential decision-making, has emerged as a game-changing paradigm. In this sophisticated framework, a Markov Decision Process (MDP), which constitutes a 5-tuple (S, A, r, γ, p), serves as the foundation for its seamless operation. Different applications in the field of medicine have used reinforcement learning. References and surveys offer thorough analyses of applications in critical care and healthcare, respectively. Doctors employed dynamic programming-based approaches to construct the best treatment plans for sepsis using a discrete state representation was crafted by leveraging a 25-dimensional discrete action space and clustering patient physiological readouts. Others have thought about partial observability and continuous state representations. Our suggested decision support system makes decisions, based on a preference score (Fig 3).

2.3.1 Gaussian Process Adapters

It was demonstrated that maximizing a time series end- to-end GP [4] imputation using the gradients of a subsequent classifier outperforms individually improving the classifier and the GP. This technique, also known as GP adapters, is not just for imputed missing data. GP adapters have recently been shown to be a suitable framework for handling the 13 unevenly spaced time series in early sepsis detection. They specifically supported earlier findings that GP adapters outperform traditional GP imputation approaches in time series classification, which call on a separate optimization step unrelated to the classification objective

2.3.2 Markov Decision Process

Typically, mathematical models for sequential decision problems are formulated as Markov decision processes (MDPs), which consist of a tuple M = (S, A, P, r). In this context, S refers to the possible states of the system, A represents the feasible actions that can be taken, P represents the probability distribution for the next state given the current state and action, and r denotes the reward function that assigns a scalar reward to each state-action pair [6] .

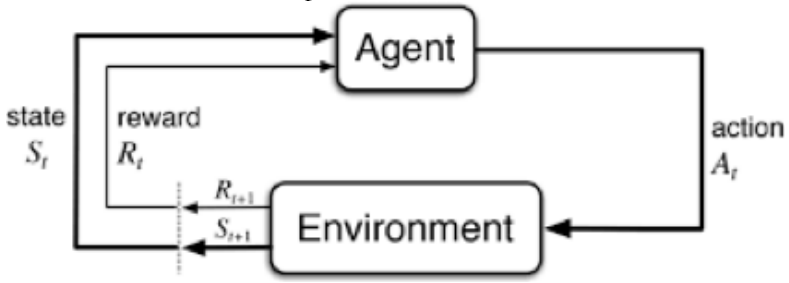

Figure 3: The proposed approach of deep reinforcement learning boasts a sophisticated network architecture with numerous features and an aesthetically pleasing design [8].

Above Fig 3 shows the use of Markov Decision Process (MDP) to model time-varying state spaces in reinforcement learning. Amidst the realm of artificial intelligence, the agent observes the state of the environment at each timestep denoted by s(t), and executes an action a(t) followed by earning a reward r(t), leading to a transition to a new state s(t+1). The ultimate goal of the agent lies in the maximization of the projected discounted future reward, popularly known as the "return", by choosing the most suitable activities. Previous studies have applied reinforcement learning in healthcare contexts, including treating septic patients using models with discretized state and action-spaces. In this study, we used value-iteration procedures to discover an ideal policy, determined by contrasting the Q-values obtained under it with those of a doctor's 14 policy. We improved upon this by utilizing continuous state-space models, deep reinforcement learning, and a clinically oriented reward function. We also evaluated how the learned policies serve patients of varying severity levels.

We have used the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database [7] to conduct this research. It is a freely available dataset which provides us with comprehensive data. This data is generally taken every four hours and the data is recorded, when several data points are present. It yields a feature vector of dimensions 48x1 at a given time ‘t’ and the state at this time is called St. The data provided by MIMIC-III is focused on patients with sepsis-3 symptoms. Since, we must apply multiple queries on the data set, so to filter out relevant data, it is first pre-processed using postgreSQL [9] and relations and tables are built [reference], further the processed data is analyzed to get a MIMIC TABLE, which is directly used as the data source for out RL Algorithm.

3. DATASET

We have used the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database [7] to conduct this research. It is a freely available dataset which provides us with comprehensive data. This data is generally taken every four hours and the data is recorded, when several data points are present. It yields a feature vector of dimensions 48x1 at a given time ‘t’ and the state at this time is

called St. The data provided by MIMIC-III is focused on patients with sepsis-3 symptoms. Since, we must apply multiple queries on the data set, so to filter out relevant data, it is first pre-processed using postgreSQL [9] and relations and tables are built [reference], further the processed data is analyzed to get a MIMIC TABLE, which is directly used as the data source for out RL Algorithm.

4. HARDWARE/SOFTWARE REQUIREMENTS

Our project on Data Analytics was implemented on a Windows operating system with Jupyter Notebook. The experiments have partly been conducted with Ryzen 7 CPU, 16GB RAM, and Nvidia RTX 3050 GPU with 4GB Memory. These requirements were enough to run python and any desired ML algorithm.
The project majorly uses Python. Some important libraries are Matplotlib, NumPy and Pandas and is carried in the Jupyter notebook.

- Python: It is designed to be easy to read and write, with a clean syntax and an emphasis on readability and simplicity. Its popularity stems from its ease of use, powerful standard library, and large number of third-party modules and packages. Python's community development model and open-source license have also contributed to its widespread adoption and continued growth.
- NumPy: A distinguished Python package that caters to the realm of numerical analysis and scientific computing. It endows an unparalleled N-dimensional array object and a vast array of mathematical operations that can be performed effortlessly on these arrays. The versatility of NumPy makes it a quintessential tool for researchers, scientists, and analysts across various fields such as physics, engineering, economics, machine learning, and more.
- Matplotlib: A plotting library for Python that allows users to create high-quality, publishable graphs and visualizations. It provides a range of visualization tools, from simple line charts to 3D charts.
- Pandas: A library for data manipulation and analysis. It provides data structures for efficiently storing and querying large datasets, as well as powerful tools for data cleaning, aggregation, and visualization. Pandas is widely used in data science, finance, and other fields dealing with large amounts of data.
- Scikit-learn: A formidable machine learning library for Python, presents an array of powerful tools beyond just model selection and evaluation. Its rich repertoire boasts of an exquisite set of methods for classification, regression, clustering, and dimensionality reduction, all crafted to elevate the art of machine learning to the next level.
- PostgreSQL: PostgreSQL is a powerful open-source relational database management system. It is widely used in web applications, data science, and other fields that require robust and scalable data storage.
- TQDM: TQDM is a library for creating progress bars in Python. It is often used in long-running processes, such as data processing or model training, to provide users with feedback on the progress of an operation.

5. RESEARCH METHODOLOGY

5.1. Action Space

We will work with a discrete action space for this research. The action space [10] defined is a 5x5 matrix which will cover maximum vasopressor dose and Intravenous fluids dose over a period of four hours. The action space is defined such that it covers all the non-zero dosages of VP and IV fluids, measured per/dosage and converted into an integer value by the concatenation of dosage, drug and the time stamp. All the zero dosage entries will be represented by 0 bin value. Medical data and records are very uncertain when it comes to finding the appropriate tuples for the action space, that’s why we are going forward with the standard i.e. total IV fluid dosage and max Vasopressor dosage as our key tuples for the action space.

5.2. Reward Function

We need a successfully working reward function [10] to map each state-action pair with a numeric value which will intrinsically define the value of that state, which will finally help our model to reach conclusions and not just predictions. We basically measure the lactate levels defining the cell hypoxia and SOFA score which gives a numeric value to measure organ failure in sepsis patients to define the overall health of a sepsis-3 patient. These two measures are the key features of our reward function where increase in SOFA score and lactate levels will result in a negative reward. For a terminal patient, at the time of ending state, the state is rewarded positive if he survives, otherwise negative.

5.3. Model-Used

The model used in this research is Dueling Double Deep Q Learning Networks (DDDQN) [11], which is based on a variant of DQN can be seen in Fig 4.

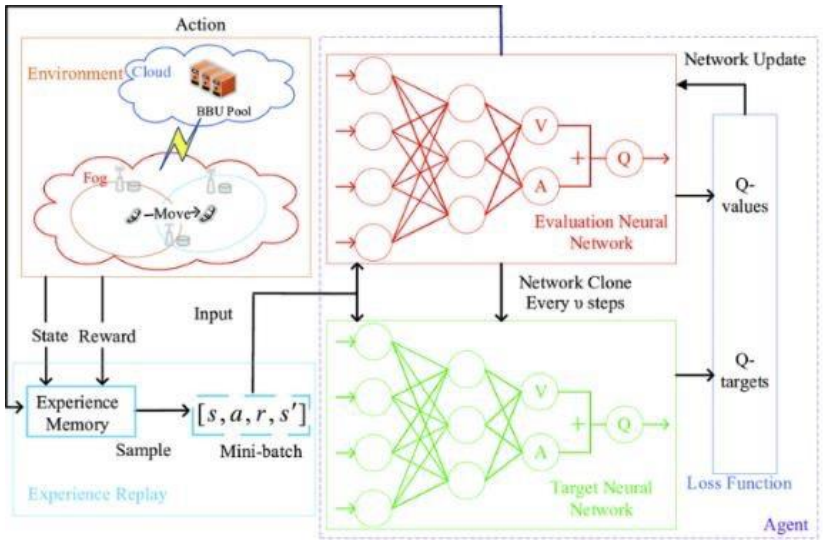

Figure 4. An illustration of dueling deep-Q-network.

DDDQN minimizes the error between target and output. We use neural approximation of Q * (s,a) which yields optimal value function. If we include θ in this function, we can easily calculate the output of the networks, i.e. Q(s, a; θ).

The desired output given by the model is $Q_{target} = r + \gamma * max_{a`}(s`, a`, \theta)$ where we have sets of the form < s, a, r, s`>. To minimize the expected loss between Q_{target} and Q_{output} we introduce a stochastic batch gradient descent in your model. Moreover, since the target values are highly volatile, addition of an extra network, which is dynamically upgraded, helps to improve the overall yield. The basic Deep Q Networks are not very efficient due to the problem of overestimation, which is very persistent in these networks, which generally leads to incorrect predictions and large error ranges. This problem of overestimation is due to the presence of Max of Q value for the next state in the Q learning update equation. This is solved through a better variant of DQN, i.e. double deep Q networks (DDQN), where we calculate the Q value by a feed-forward pass on the main network rather than using the main network directly for calculation of Q values. Now, to solve another problem, i.e. when we find optimal treatments, we have to ignore the influence of the previous state if it has a positive reward and correct action is to be taken at the present time stamp. For this we turn to Dueling deep Q network (DDDQN), where the values of 19 tuples action and state given by Q(s,a) are divided in two parts namely, estimation of advantage of a stage representing the quality of chosen action and estimation of flow of value representing the quality of chosen state. This yields a fully formed Dueling Double-Deep Q Network in Fig 5 having 2 hidden layers of size 128, combining the above ideas [11]. The training of the model based on this methodology gives us the optimal state of a patient as, $\pi * (s) = \arg max_a Q(s,a)$.

.

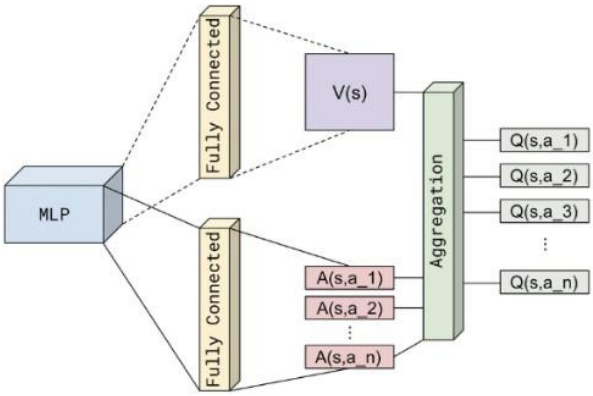

Figure 5: Architecture of DDDQN.

5.4. RL Algorithm

Start with the initialization of two Q networks, D3QN-A network defined as $Q^A(s,a;\theta^A)$ and D3QN-B network defined as $Q^B(s,a;\theta^B)$. Here all the parameters in the networks are defined by θ^A and θ^B. Secondly, the aproach involves initializing the Experience Reply with (s_t,s_{t+1},a_t,r_t). s_t has two components, the first being a comprehensive feature comprising a basic feature and salience map, and the second being a historical experience vector that preserves previously used action indexes. Our approach initializes a 20*13d vector to represent historical experience, with a maximum exploring step of 20 and 13 action numbers.

Initialize $Q^A(s,a;\theta^A)$ and $Q^B(s,a;\theta^B)$ with parameters θ^A and θ^B
Initialize Experience Reply pool
for $episode$ = 1,$episode < N_episodes$ do
reset the environment and process initial state s_0
  for t = 1,$t < Max_steps$ do
    select a cropping action a_t according to ε-greedy
    perform action a_t to get $s(t+1)$ and the instant reward
    put (s_t, s_{t+1}, a_t, r_t) into Experience Reply pool
    randomly choose a set of (s_i, s_{i+1}, a_i, r_i)
    randomly choose the update network of A or B
    if $episode \% update_step = 0$ && update A:
      $a^* = \arg\max_a Q^A(s_{i+1}, a, \theta_t^A)$
      $y_t^A = \begin{cases} r_{i+1}, s_{i+1} ends \\ r_{i+1} + \gamma Q^B(s_{i+1}, a^*; \theta_t^B), else \end{cases}$
      $loss = (y_t^A - Q^A(s_t, a_t; \theta_t^A))^2$, update θ^A with gradient descent
    else if $episode \% update_step = 0$ && update Bčž
      $a^* = \arg\max_a Q^B(s_{i+1}, a, \theta_t^B)$
      $y_t^B = \begin{cases} r_{i+1}, s_{i+1} ends \\ r_{i+1} + \gamma Q^A(s_{i+1}, a^*; \theta_t^A), else \end{cases}$
      $loss = (y_t^B - Q^B(s_t, a_t; \theta_t^B))^2$, update θ^B with gradient descent
    end if
    $s_t \rightarrow s_{t+1}$
  end for
end for

Figure 6: RL Algorithm

The art of state representation lies in its ability to incorporate valuable historical experiences and emulate the human decision-making process, thus facilitating informed decision-making in the present. The feature extraction part generates s0 as the initial state. Subsequently, states are sent to the Agent, which uses the ε-greedy algorithm to select the current cropping action, followed by execution of the chosen cropping action and obtaining the cropped image. This process is repeated, and each one-step cropping operation (s_t,s_{t+1},a_t,r_t) is recorded in the experience reply pool. The maximum number of cropping steps is 20, and in the training process, N_episodes is set to 160,000, with a group of records randomly selected for learning each time. See Fig 6 for the algorithm steps.

6. RESULTS

On a held-out test set which was accurately on 50 epochs, the y-axis of the graph depicts mortality rates, which fluctuate based on the variance between recommended dosages dictated by the optimal policy and those administered by healthcare providers, which serves as the return of action. This difference was computed and correlated with whether the patient lived or passed away in the hospital for

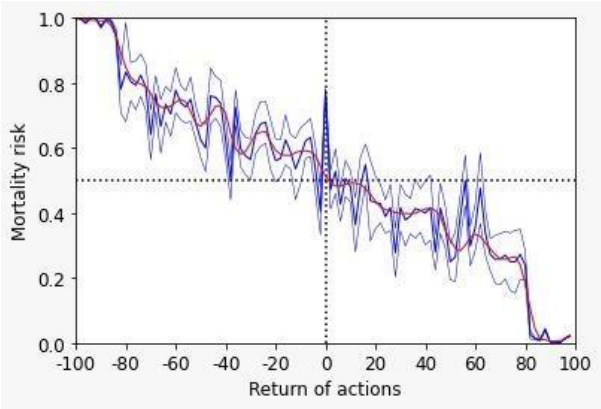

Figure 7: Plot between the return of the clinician's policy and patients' mortality.

each timestep as shown in Fig 7, enabling the computation of observed mortality. In Fig 8, With a 95% confidence level, this bound would always be higher than the clinicians' guideline if enough models were produced. The statistical safety of the novel artificial intelligence (AI) policy in question is a topic of current discourse in theory which is maximized by this model selection method.

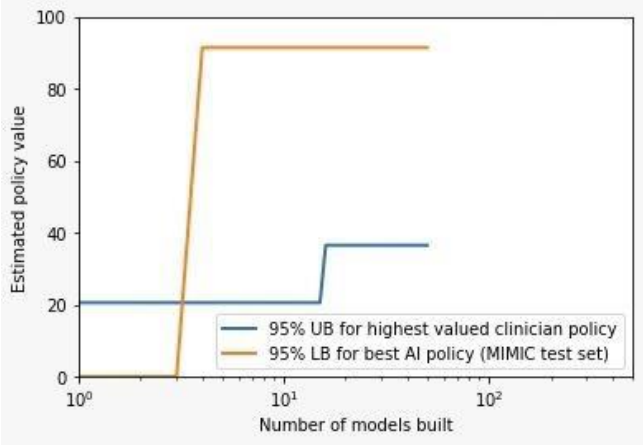

Figure 8: During the generation of 500 models of MIMIC III, the 95% lower bound of the optimal AI policy is compared with the 95% upper bound of the most esteemed clinician policy established to ascertain the range of values within which they operate.

7. CONCLUSION

Employing deep learning in this research, the problem of treating sepsis patients is being addressed in a practical manner. The study investigated fully continuous state-space/discrete action space models to discover the most efficient treatment options, learning an estimate for the best action-value function, using Dueling Double-Deep Q networks, Q *(s, a). It was discovered that the resulting continuous state space model generated interpretable regulations that might enhance sepsis treatment. The taught policies will be put through a patient evaluation and contrasted with other investigative algorithms in future study. The results of this study may significantly influence medical practice for sepsis

identification. The use of a model like described in the papers can anticipate the onset of sepsis that could lessen the vexing issue of alarm fatigue that plagues the existing clinical scoring systems, improve patient outcomes, and lessen the burden on the healthcare system because sepsis is a condition that is poorly understood and challenging for practitioners to diagnose. Although the focus of this work was on early sepsis identification, it would be simple to adapt the techniques to other clinical events of relevance, such as cardiac arrests, code blue occurrences, ICU admissions, and cardiogenic shock. This will enable practitioner to employ the techniques in a real-world clinical context, and the model's usefulness can be objectively demonstrated by gathering information on the reliability of the warnings it raises and how it is applied on the actual wards.

8. REFERENCES

[1] Futoma, J. (2018). "Gaussian process-based models for clinical time series in healthcare" (Doctoral dissertation, Duke University).

[2] Raghu, Aniruddh, et al. "Deep reinforcement learning for sepsis treatment." arXiv preprint arXiv:1711.09602 (2017).

[3] Tardini, Elisa et al. "Optimal Treatment Selection in Sequential Systemic and Locoregional Therapy of Oropharyngeal Squamous Carcinomas: Deep Q-Learning With a Patient-Physician Digital Twin Dyad." Journal of medical Internet research vol. 24,4 e29455. 20 Apr. 2022, doi:10.2196/29455.

[4] Goh, Kim & Wang, Le & Yeow, Adrian & Poh, Hermione & Li, Ke & Yeow, Joannas & Tan, Gamaliel. (2021). "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature Communications", doi:10.1038/s41467-021- 20910-4.

[5] Jonsson A. "Deep Reinforcement Learning in Medicine", Kidney Dis 2019;5:18-22. doi: 10.1159/000492670.

[6] Littman, Michael L. "A tutorial on partially observable Markov decision processes." Journal of Mathematical Psychology 53.3 (2009): 119-125.

[7] Johnson, A., Pollard, T., & Mark, R. (2019). MIMIC-III Clinical Database Demo(v14) . PhysioNet. https://doi.org/10.13026/C2HM2Q.

[8] Cao, LiChun & ZhiMin,. (2019), "An Overview of Deep Reinforcement Learning", CACRE2019: Proceedings of the 2019 4th International Conference on Automation, Control and Robotics Engineering. 1-9. 10.1145/3351917.3351989.

[9] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

[10] Joseph M. Carew, "Tech Target blog," Tech Target Enterprise AI. [Online]. URL: techtarget.com/definition/reinforcement-learning.

[11] Mary Mammen, Priyanka & Kumar, Hareesh. (2019), "Explainable AI: Deep Reinforcement Learning Agents for Residential Demand Side Cost Savings in Smart Grids."