WJEMSD

WASD
WORLD ASSOCIATION FOR SUSTAINABLE DEVELOPMENT

World Journal of
ENTREPRENEURSHIP, MANAGEMENT
AND SUSTAINABLE DEVELOPMENT

ISSN: 2042-5961 (Print) | 2042-597X (Online)

WJEMSD V19 N3/4 2023

OPEN ACCESS

DOI: 10.47556/J.WJEMSD.19.3-4.2023.3

RESEARCH PAPER

Developing an Humanitarian Logistics Framework Using a Reinforcement Learning Technique

Dr Khalid I. AlQumaizi
Assistant Professor of Family Medicine
College of Medicine, AlMaarefa University, Ad Diriyah
Riyadh, Kingdom of Saudi Arabia
Email: kqumaizi@mcst.edu.sa

Prof. Dr Ashit Kumar Dutta
Department of Computer Science and Information Systems
College of Applied Sciences, AlMaarefa University, Ad Diriyah
Riyadh, Kingdom of Saudi Arabia
Email: adotta@mcst.edu.sa

Dr Sultan Alshehri
Department of Pharmaceutical Sciences
College of Pharmacy, AlMaarefa University, Ad Diriyah
Kingdom of Saudi Arabia
Email: sshehri.c@mcst.edu.sa

ABSTRACT

PURPOSE: Humanitarian logistics (HL) refers to co-ordinating relief efforts to ensure disaster victims have timely access to necessary goods. Large-scale catastrophes and catastrophic occurrences sometimes result in a significant resource deficit, making it challenging to allocate limited resources across affected locations to improve emergency logistics' operations. The primary objective of disaster relief is to save lives, alleviate victims' suffering, and protect human dignity in the face of overwhelming odds; however, this is only possible when logistical support is provided for catastrophe victims. The primary objectives of HL are life preservation and post-disaster reduction. Disasters such as earthquakes and tsunamis necessitate the prompt and adequate delivery of emergency relief supplies. There is a lack of an effective optimisation model for allocating resources in HL. This paper therefore presents a reinforcement learning-based framework for optimising the resource allocation processes in HL.

CITATION: AlQumaizi, K.I., Dutta, A.K. and Alshehri, S. (2023): Developing an Humanitarian Logistics Framework Using a Reinforcement Learning Technique. *World Journal of Entrepreneurship, Management and Sustainable Development*, Vol. 19, No. 3/4, pp. 15–32.

RECEIVED: 28 January 2023 / **REVISED:** 11 May 2023 / **ACCEPTED:** 17 May 2023 / **PUBLISHED:** 1 October 2023

DESIGN/METHODOLOGY/APPROACH: The authors employ the State Action Reward State Action (SARSA) algorithm to reduce the complexities in resource allocation. In addition, a transportation plan is developed in order to support victims at remote locations. The suggested algorithm is evaluated against the precise dynamic programming approach and an heuristic algorithm to determine which provides the best quality result.

FINDINGS: The experimental findings reveal that the algorithm outperforms the state-of-the-art methods regarding efficiency and accuracy. In addition, the Q-learning algorithm can deliver solutions closer to optimum or even ideal by improving the training phase.

KEYWORDS: *Humanitarian Logistics; Q-learning; SARSA; Emergency Relief Supplies; Reinforcement Learning*

INTRODUCTION

In the humanitarian sector, a disaster generally implies an urgent need to support victims in the shortest possible time. Planning, executing and regulating the efficient, cost-effective movement and storage of commodities, resources, and information from the point of origin to the place of consumption to reduce the suffering of vulnerable people is known as humanitarian logistics (HL) (Agafonov and Myasnikov, 2021; Yu *et al*., 2021; Yu *et al*., 2018). Most disasters occur unexpectedly, leaving little opportunity for people to prepare for the aftermath. The need for relief supplies, such as food, medication, shelter, water, and other necessities, increases dramatically after devastating events (Yu *et al*., 2019). The availability of humanitarian assistance can be improved by well-executed emergency operations (Khadilkar, 2018). The risks and uncertainties of any disaster make HL operations more difficult.

Logistics integrate emergency planning with response to catastrophes. For this reason, HL is essential for the efficiency and promptness of the response for effective humanitarian programmes (Anuar *et al*., 2019; Kim *et al*., 2019; Klaine *et al*., 2018). In the aftermath phase, humanitarian efforts are made more difficult by disruptions in the supply chain caused by things such as damaged roads and trains, by unknown factors in demand, and weather conditions.

The primary techniques for coping with these difficulties are modelling, optimisation, and simulation (Cao *et al*., 2018). In addition to traditional mathematical and statistical methods, we also employ fuzzy logic, queuing theory, decision theory, and the reference point method (Dubey *et al*., 2021). Academics frequently use modelling and optimisation to deduce answers to questions in emergency HL (Nassar and Yilmaz, 2019). Logistics emergency operations cover a wide range of tasks, from locating facilities to transporting casualties to distributing aid, prepositioning supplies, and even evacuating people from danger areas. Humanitarian and emergency relief efforts are typically carried out by governmental, military, civil society, and other humanitarian organisations (Wu *et al*., 2019). These various emergency operators have the ability to co-ordinate the provision of the most effective help during times of crisis (Wang *et al*., 2021).

Human suffering among survivors is largely caused by a lack of resources (such as food, medicine, and medical aid) or a delay in delivering those resources, often due to their inappropriate allocation (Khan and Javaid, 2020). Extremely high demand will emerge after a catastrophic event

(Jaiswal *et al*., 2020). The decline in survivors' health dramatically affects their need for essential supplies (Agafonov *et al*., 2022; Yan *et al*., 2022; Lu *et al*., 2022; Chamola *et al*., 2020). However, the calamity might wipe away pre-positioned supplies, and any external supplies could be held up in traffic or ruined roadways. Especially at the outset of HL, the practicality of complex supply cannot keep up with rising demand (Yang *et al*., 2020; Lopez *et al*., 2018; Castellanos *et al*., 2018; Mohammed *et al*., 2020; Nadi and Edrissi, 2016). Some survivors may feel unfairly treated, and this could lower their morale and drive them to despair if the few resources are distributed inefficiently. According to Vereshchaka and Dong (2019), adverse emotional reactions can linger for months or even years, significantly reducing survivors' ability to bounce back from traumatic experiences. It is therefore essential to create a delivery strategy that is efficient, effective, and equitable, and that takes into account the suffering of survivors directly when making allocation decisions.

The majority of the presently offered research includes heuristic solution approaches to address challenging optimisation problems (Zou *et al*., 2021; Hachiya *et al*., 2022; Munawar *et al*., 2021). Conventional methods have difficulties overcoming such issues (Lonkar *et al*., 2019; Benkacem *et al*., 2022). However, current heuristic techniques are still insufficiently effective for dealing with situations of this complexity. The study intends to develop a decision-making framework for allocating resources and distributing the materials to the victims using a reinforcement learning (RL) approach. It addresses the multi-objective and non-linear resource allocation issue. In addition, it proposes a transportation plan to reach disaster locations quickly. The contribution of the proposal is as follows:

1. an effective resource allocation technique for providing supporting materials to multiple disaster locations;
2. an intelligent decision-making system to generate a transportation plan to reach the complex disaster location.

The remainder of the paper is organised as follows: the next section presents the related works on resource allocation and transportation plans during disasters. The proposed research methodology is discussed in the following section, the results and discussions follow. The final section concludes the study with its future direction.

RELATED WORKS

The term "humanitarian logistics" refers to the logistics that deal with a crisis management system's pre- and post-disaster logistics, including the acquisition, storage, and distribution of resources, including food, water, medication, and other supplies (Agafonov and Myasnikov, 2021; Yu *et al*., 2021; Yu *et al*., 2018). Recent natural catastrophes have brought much interest from researchers and practitioners to logistics in the context of humanitarian operations. This interest is due to the

demand for agile and competent logistical systems that can handle various disasters, specialised large-scale risk and disruption management, and the consequences of disasters on human lives and the economy (Yu *et al*., 2019). It is predicted that in the next 50 years, the frequency of catastrophes, both natural and artificial, will increase by a factor of five (Khadilkar, 2018). As a result, disaster management systems must prioritise humanitarian logistics as one of the highest priorities.

There has recently been a surge in HL studies and emergency relief supplies during disasters. Anuar *et al*. (2019) introduced a method for transporting multiple unmanned aerial vehicles (UAVs) to disaster locations; they employed a route optimisation algorithm to identify locations in remote areas. Kim *et al*. (2019) developed an integer linear programming technique for scheduling UAV tasks. Likewise, Klaine *et al*. (2018) used a continuous approximation method for identifying optimal distribution locations and order quantities for disaster relief operations.

Over the past few years, the application of RL has expanded into areas directly connected to HL activities (Cao *et al*., 2018; Dubey *et al*., 2021; Nassar and Yilmaz, 2019). Wu *et al*. (2019) suggested a train route scheduling algorithm for HL management. Using Q-learning, they devised a brand new strategy for solving the optimisation issue of combining several variables. In order to deal effectively with the challenges in the train operation, the authors propose a Q-learning strategy. The authors suggested a Q-learning approach to reduce potential hazards (Wang *et al*., 2021). The RL-based resource allocation technique proposed by Khan and Javaid (2020) may be used in both unicast and broadcast scenarios for V2V communications, making it more flexible and valuable. Jaiswal *et al*. (2020) presented a Q-learning-based approach to train rescheduling. According to the data, Q-learning produces scheduling solutions that are on par with, if not better than, those produced by numerous simple, non-agent-based scheduling techniques (Agafonov *et al*., 2022). To minimise the total priority-weighted delay, the authors proposed a scheduling algorithm based on an RL approach for allocating time for train arrivals and departures.

Using sample-path approaches and stochastic dynamic programming, the authors developed a technique to examine the performance between the urgency of the demand, rescue incentives, and service delays in distributing emergency resources (Yan *et al*., 2022; Lu *et al*., 2022; Chamola *et al*., 2020). The authors built a model for the best allocation of critical supplies to save societal costs, created appropriate heuristic solution techniques, and evaluated the heuristics' efficacy with numerical tests. The authors investigated a resource allocation problem and presented a novel network model to optimise service throughput (Yang *et al*., 2020). The best strategies for allocating resources were determined using a straightforward heuristic method.

A robust model was developed by Lopez (2019) to optimise both efficiency and justice in an emergency resource allocation issue involving many competing afflicted locations and a single relief resource centre. A new approach to allocating scarce resources in an emergency was devised so that officials could more easily determine their preferred allocation strategy. The authors

suggested a multi-objective optimisation model for scheduling emergency resources (Castellanos *et al*., 2018). A multi-objective evolutionary algorithm based on decomposition was offered to address the issue. The two-stage relief chain studied by Mohammed *et al*. (2020) consists of a single location where emergency supplies come at random intervals across time in unpredictable quantities and are distributed to a random sample of catastrophe survivors at a point of distribution (Nadi and Edrissi, 2016). The objective was to identify vehicle dispatching procedures resulting in little unmet demand at the distribution centre. Two graphs were made to show the gap between the objective function value of an heuristic solution and the ideal solution. Using a multi-objective programming model, the authors attempt to maximise the lowest victims' perceived satisfaction while minimising the highest difference in victims' perceived satisfaction across all demand points and phases (Vereshchaka and Dong, 2019). To address this problem, they recommended using a genetic algorithm to map out aid distribution. After a natural catastrophe, an HL framework was developed by Zou *et al*. (2021) in order to provide a model to determine where to set up distribution centres to reach survivors and how to divide up supplies for local distribution. In order to solve the model, it was suggested that an heuristic strategy based on the heuristic concentration integer technique was used.

Currently, UAVs are used for various purposes, including aerial photography, surveying, and pesticide spraying. Improvements in UAV performance and associated technologies have also led to its application in the logistics industry (Hachiya *et al*., 2022). Companies such as Amazon and Walmart are developing new systems that employ UAVs to transport packages to serve their consumers better, while companies such as DHL, Google, and Alibaba have started working on unmanned aerial vehicles. It has also been shown that UAVs may be used to carry supplies during catastrophe situations (Munawar *et al*., 2021). The authors performed several experiments to show how drones may be used to deliver relief supplies in an emergency. Therefore, it is essential to make the most use of UAVs for conveying emergency relief supplies, considering the UAVs' battery life and maximum payload constraints (Lonkar *et al*., 2019). In recent years, a variety of research has been conducted on delivering packages using UAVs. These issues are typically phrased as Unmanned Aerial Vehicle Routing Problems (UAVRPs), a subset of the broader vehicle routing problem family (Benkacem *et al*., 2022).

RESEARCH METHODOLOGY

Distribution is the process through which supplies are sent from storage facilities or hospitals to needy regions (Agafonov and Myasnikov, 2021; Yu *et al*., 2021; Yu *et al*., 2018; Yu *et al*., 2019; Khadilkar, 2018). Smart preparation may achieve maximum aid distribution despite the unpredictability of post-disaster situations. Demand fluctuations, damaged links and facilities, and resource scarcities are all factors that cannot be predicted in the aftermath of a disaster (Anuar *et al*.,

2019; Kim *et al*., 2019; Klaine *et al*., 2018; Cao *et al*., 2018; Dubey *et al*., 2021). In order to save the most lives possible while minimising casualties, it is crucial to distribute aid more effectively so that needs are fulfilled, and gaps are closed. Most of these distribution models are deterministic and have a single purpose. Figure 1 outlines the phases in the proposed research.

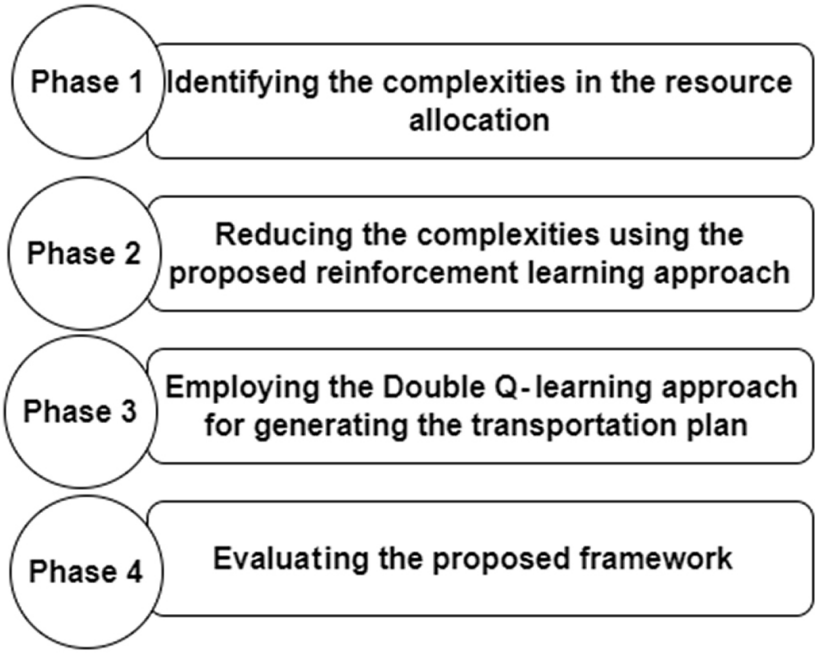

Figure 1: Research Phases
Source: Constructed by authors

In Phase 1, the challenges and limitations of allocating resources to the disaster location from the response centre are identified. Phase 2 discusses the proposed RL approach to overcome the complexities. Phase 3 offers a Double Q-learning approach for generating the transportation plan. Finally, Phase 4 indicates the evaluation metrics for measuring the performance of the proposed method. Therefore, the study adapts the Markov decision process (MDP) model to improve the HL framework's efficiency. The MDP model includes state (S_s), action (A_c), reward (R_w), state transition (S_T), and policy (P). In addition, the policy ($S_s \times A_c \rightarrow R_w$) and density function ($S_T : S_s \times A_c \times S_s \rightarrow [0,1]$), where A_c represents the human elements to distribute the materials, S_T is the

transportation plan to reach the disaster location. R_w is the successful distribution of materials to the victims at the disaster location.

Based on Yu *et al*. (2021), this study intends to reach the location in a critical 72 hour (C) period after a disaster. The decision (D) indicates the response centre decision to allocate resources for multiple disaster locations with relevant transportation mode. Table 1 presents the description of the parameters. In this study, the authors applied State Action Reward State Action (SARSA) (Nassar and Yilmaz, 2019) to optimise the resource allocation processes. SARSA follows the ε-greedy approach and computes the following state by passing the reward to the previous form.

Table 1: Summary of Parameters

Parameters	Description
$S_{.T.}$	State transition
P	Policy
S_s	State
A_c	Action
R_w	Reward
α	Learning rate
Γ	Discount factor
C	Critical 72 hours
A_{c_t}	Action at time t
S_{s_t}	State at time t
∈	Episodes

Source: Constructed by authors

Eqn.1 represents the space and actions at the time (t).

$$Q(L_t, A_{c_t}) \leftarrow Q(L_t, A_{c_t}) + \propto \left[R_{w_{t+1}} + \gamma Q(L_{t+1}, A_{c_{t+1}}) - Q(L_t, A_{c_{t+1}}) \right] \quad (1)$$

Where L_t is space, A_{c_t} is action, R_w is a reward at time t, and γ is a constant.
Eqn. 2 reflects the optimal decision at time t.

$$D_t = \{D_1, D_2, \ldots\ldots, D_N\}, D_t \leq C \quad (2)$$

Figure 2 highlights the RL-based HL framework for effective resource allocation. Figure 3 outlines the state transitions and decisions of the RL approach.

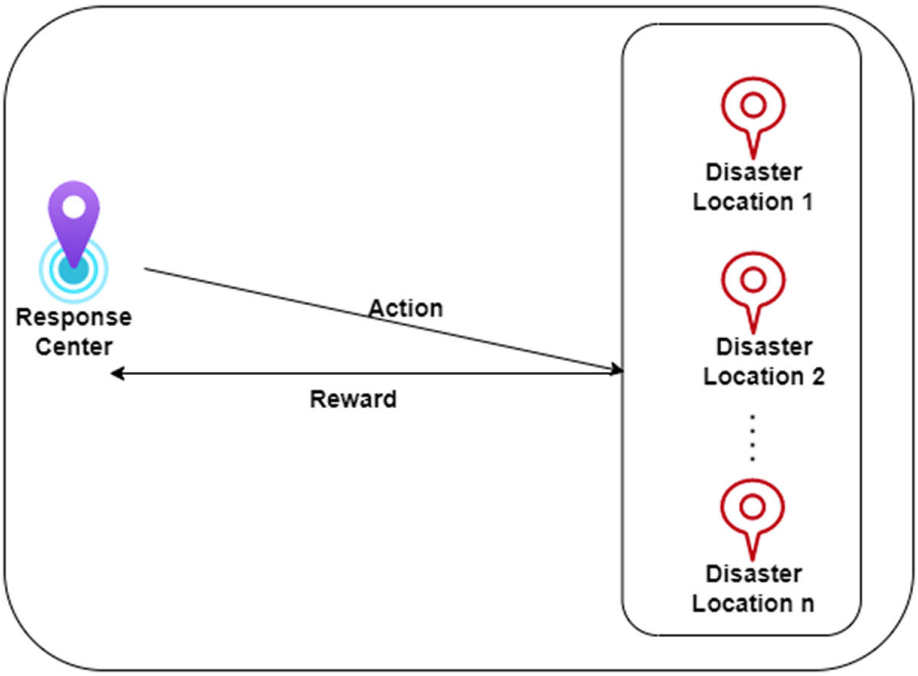

Figure 2: Reinforcement Learning Based HL Framework
Source: Constructed by authors

SARSA follows an ε-greedy approach for deciding the actions according to the policy. Eqns. 3, 4, and 5 outline the policy for distributing the materials over given actions, the intention being to identify a policy P^* to optimise the S_s and A_c pair $Q^*(S_{s_t}, A_{c_t})$.

$$P^*(A_c|S_s) = P\left(A_{C_t} = (A_c|S_{s_t}) = S\right) \tag{3}$$

$$P^* = argMax_{A_{c_t \to A}} Q^*(S_{s_t}, A_{c_t}) \tag{4}$$

$$Q^*(S_{s_t}, A_{c_t}) = \overset{Max}{A_{c_t}} Q^P(S_s, A_c) \tag{5}$$

Eqn. 6 shows the optimum value of $Q^*(S_{s_t}, A_{c_t})$ with learning rate (L).

$$Q^*(S_{s_t}, A_{c_t}) \leftarrow Q(S_{s_t}, A_{c_t}) + L\left[R_{w_t} + \lambda\left(\sum_{A_{c_t} \in A_c} P(A_c|S_s) + Q(S_{s_{t+1}}, A_{c_{t+1}}) - Q(S_{s_t}, A_{c_t})\right)\right] \tag{6}$$

Eqn. 7 shows the process of the weight factor (W_f) computation using the gradient descent approach.

$$\Delta W_f = P\left[R_{w_t} + \lambda Q^P(S_s, A_c, W_f) - Q^P(S_s, A_c)\right] \nabla W_f Q^P(S_s, A_c, W_f) \tag{7}$$

Response centre selects an action with probability (P_b) during time slot (t), Eqn. 8 shows the action (A_{c_t}).

$$\begin{aligned} T_r &= \underset{a_{c_t}, S_s \in A_{c_t}}{Max} Q^P(S_s, A_c) \\ &= 1 - \epsilon,\ 0 < \epsilon < 1 \end{aligned} \tag{8}$$

The authors apply the Agafonov and Myasnikov (2021) approach for generating the transportation plan (TP). Eqns. 9 and 10 show the Double Q-learning approach. The outcome of the analysis generates a binary value: "0" indicates a UAV and "1" represents the manned vehicle.

$$Q^A_{t+1}(S_{s_t}, A_{c_t}) = (1-\epsilon) Q^A_t(S_{s_t}, A_{c_t}) + \epsilon \left(R_{w_t} + \gamma Q^B_t \left(S_{s_{t+1}}, argmax_{\in} Q^A_t \left(S_{s_{t+1}}, \epsilon \right) \right) \right) \tag{9}$$

$$Q^B_{t+1}(S_{s_t}, A_{c_t}) = (1-\epsilon) Q^B_t(S_{s_t}, A_{c_t}) + \epsilon \left(R_{w_t} + \gamma Q^A_t \left(S_{s_{t+1}}, argmax_{\in} Q^B_t \left(S_{s_{t+1}}, \epsilon \right) \right) \right) \tag{10}$$

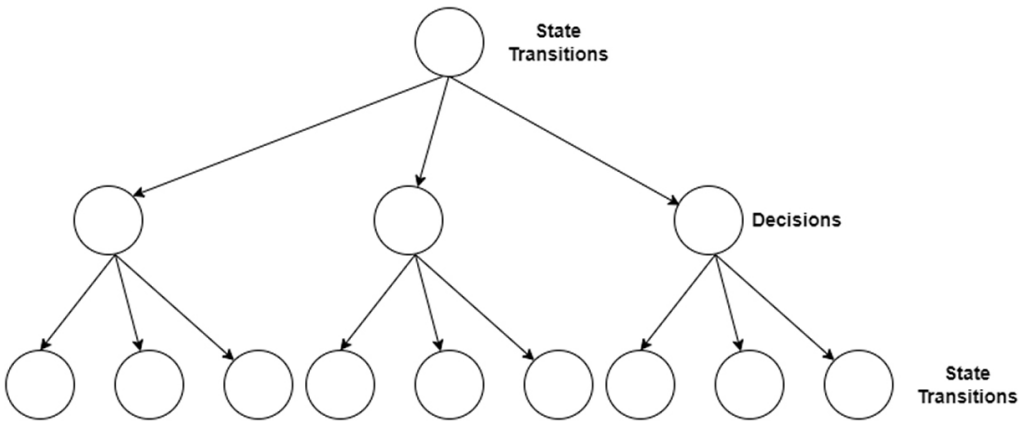

Figure 3: State transitions and Decisions

Source: Constructed by authors

The learning rate (α) shows the efficiency of an agent in gaining new information. The higher value of α indicates the improved efficiency of the model. The agent achieves value between 0 and 1 during

the experiment. The nearest value to 1 indicates the highest reward to the agent. In this study, α is set to 0.6 in order to identify the average efficiency. The discount factor (Γ) is used to identify the future rewards. The range of Γ is between 0 and 1. The higher value of Γ indicates that the model approaches towards the future rewards. In this study, Γ is set to 0.8 to balance future rewards. Figure 4 highlights the SARSA algorithm for generating objective resource allocation and transportation plan values.

Algorithm: SARSA based HL framework

1. Initialise ε, λ, P, W
2. Compute distance (d) between resource centre and disaster location, resource (R)
3. Set $Q \leftarrow Q^*$
4. For $t = 1,2,3,\ldots,n$
   $\gamma_n(t) = \frac{P_{th}(t), P_{c,n}(t)}{\sum_{j \in c} p_{j,n}(t) p_j(t) + \sigma^2(t)}$, where $P_{th}(t)$ is route to disaster location, $P_{c,n}(t)$ is transportation plan with resources.
   $C_t = \int f_e(P) d, c$, where $f_e(P)$ is density function.
5. while Q is not converged
6. set state
7. $S_{s_t} = \gamma_1(t), \gamma_2(t), \ldots. \gamma_n(t), d_1(t), d_2(t), \ldots d_n(t), b_1(t), b_2(t), \ldots b_n(t)$
8. compute policy
9. compute the reward
10. set next state
11. update weight
12. update Q*
13. compute objective value with final state and reward
14. end while
15. end for
16. return objective value

Figure 4: SARSA Based HL Objective Generation Algorithm
Source: Constructed by authors

RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed framework, Yu *et al*. (2021), Yu *et al*. (2019), and Cao *et al*. (2018) are employed. A SARSA algorithm is implemented in Python 3.7, Windows 10 Professional, 8GB RAM, i7 processor. The principle of proximity is used in distributing the supporting materials to the victims. In this experiment, small-scale instances are employed in order to evaluate the framework's efficiency. In addition, the larger scale instances require more episodes that demand a high-end computation resource. Table 2 outlines the proposed framework's generated objectives at multiple learning rates. The suggested framework achieves an objective

value of 1861.7 at the learning rates of 0.2, 0.4, 0.6, 0.8, and 1, respectively. Figure 5 reveals that the objective values are the same for the learning, 0.2, 0.4, 0.6, 0.8, and 1, respectively.

Table 2: Objective Outcome at Multiple Learning Rates

Learning Rate (α)	1	2	3	4	5	Average
0	5,123.6	5,123.6	5,123.6	5,123.6	5,123.6	5,123.6
0.2	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7
0.4	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7
0.6	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7
0.8	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7
1	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7

Source: Constructed by authors

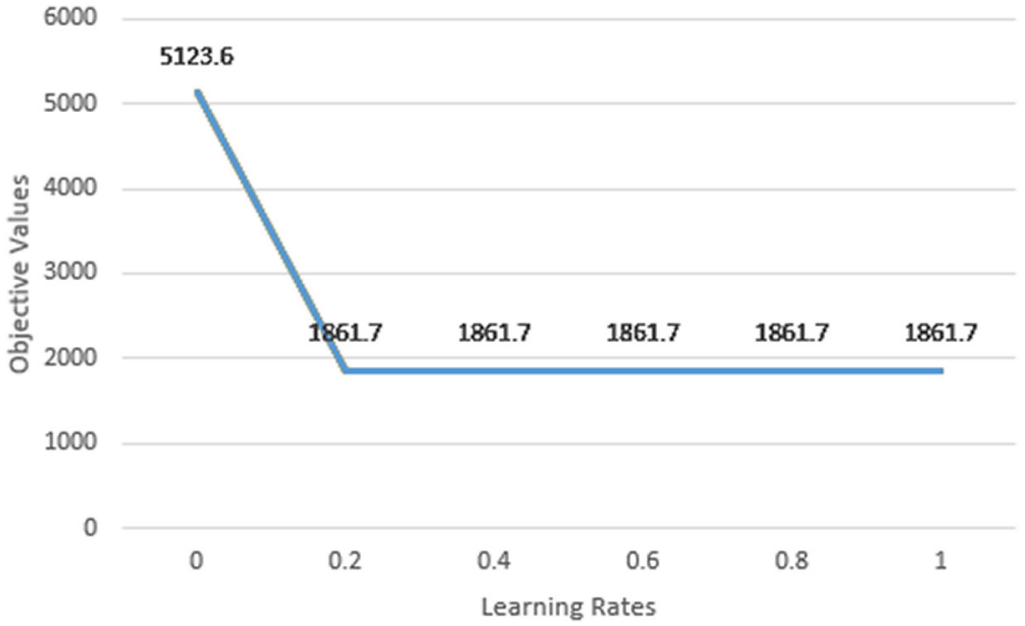

Figure 5: Objective Values at Multiple Learning Rate

Source: Constructed by authors

Table 3 outlines the objective values at multiple discount factors. As with Table 2, the objective values are constant for 0.2, 0.4, 0.6, 0.8, and 1, respectively. Likewise, Figure 6 outlines the objective values of the proposed framework.

Table 3: Objective Outcome at Multiple Discount Factors

Discount Factor (γ)	1	2	3	4	5	Average
0	4,321.81	4,321.81	4,321.81	4,321.81	4,321.81	4,321.81
0.2	2,765.31	2,765.31	2,765.31	2,765.31	2,765.31	2,765.31
0.4	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7
0.6	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7
0.8	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7
1	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7	1,861.7

Source: Constructed by authors

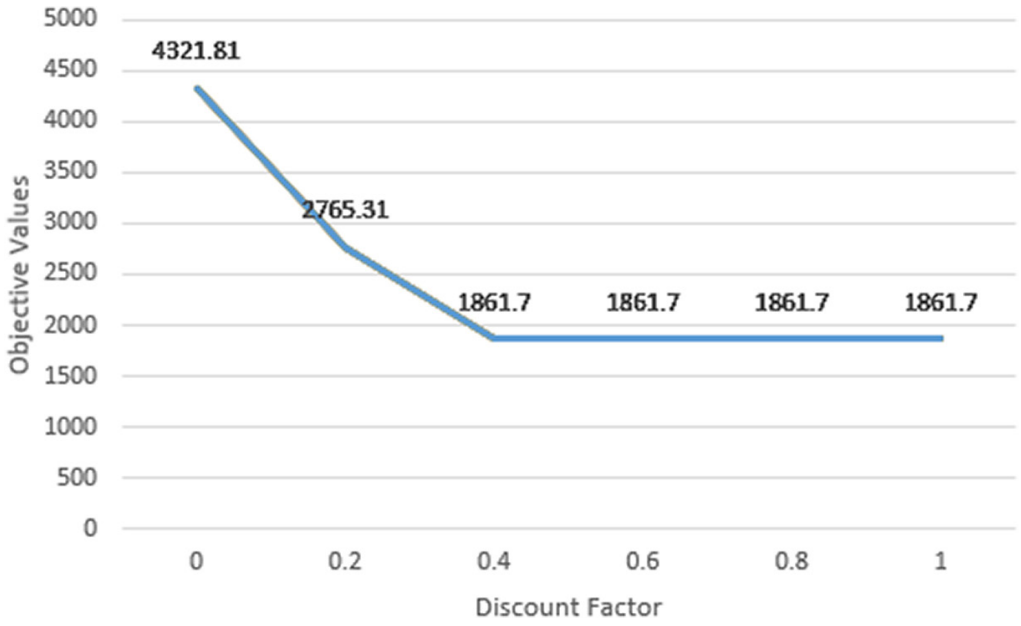

Figure 6: Objective Values at Different Discount Factor

Source: Constructed by authors

The proposed framework achieves an objective value of 1,861.7. The number of episodes can be increased to gain a better accurate weight in a limited time. Instances are denoted in {N,U,T}, where N indicates the number of disaster locations, U is a unit of resources, and T is the total time. The initial set of experiments is conducted for a different value of N and T, while the unit is set to 1. The size of the state is increased exponentially with the higher value of N and T. Table 4 presents the comparative analysis outcome of the HL frameworks. The proposed framework achieved a superior outcome comparing to the state-of-the-art frameworks. Figure 7 shows the computation time of the HL frameworks. The proposed framework outperforms the recent HL frameworks.

For context, a TP (1), (4,2,2) indicates that the resource centre deploys a manned vehicle to transport two units of supporting materials to four disaster locations in 2 C period. The reward function is unique for different states. It is calculated based on the specific environment. The proposed framework minimises the cost and maximises the reward.

Table 4: Comparative Analysis Outcome with Constant Unit Size

Instances	State Size	Yu *et al.* (2021)		Yu *et al.* (2019)		Cao *et al.* (2018)		Proposed Framework	
		Objective Values	Time (seconds)	Objective Values	Time (seconds)	Objective Values	Time (seconds)	Objective Values	Time (seconds)
{2,1,4}	275	2,365.2	1.25	2,863.4	2.56	1,986.4	1.86	1,861.7	0.85
{2,1,6}	754	4,569.6	18.4	5,264.5	5.89	3,457.8	11.36	2,634.5	2.6
{2,1,8}	1,925	8,966.5	124.23	9,675.2	96.87	7,845.9	89.4	4,869.2	94.7
{3,1,4}	563	3,256.1	3.65	3,124.6	3.78	2,156.4	3.5	2,145.7	2.4
{3,1,6}	2,145	6,985.3	23.4	6,325.1	18.5	5,236.8	25.4	4,896.5	18.6
{3,1,8}	5,264	9,635.1	134.6	7,894.5	114.5	7,564.8	96.8	6,897.4	99.4
{4,1,4}	956	3,452.3	7.4	3,654.7	5.78	2,896.4	6.7	1,956.4	5.4
{4,1,6}	2,356	7,698.1	16.8	5,689.1	21.4	5,687.4	12.5	5,634.2	26.4
{41,8}	5,698	9,899.2	89.4	8,936.1	96.5	7,963.8	69.4	8,964.5	104.1
{5,1,4}	987	3,658.4	14.2	2,896.4	4.36	2,993.8	3.9	2,314.5	6.5
{6,1,6}	2,687	8,963.5	36.8	5,647.1	28.4	5,361.4	34.5	4,521.2	34.2
{5,1,8}	5,996	12,326.1	153.3	7,567.7	102.3	9,645.8	125.3	8,965.6	112.3

Source: Constructed by authors

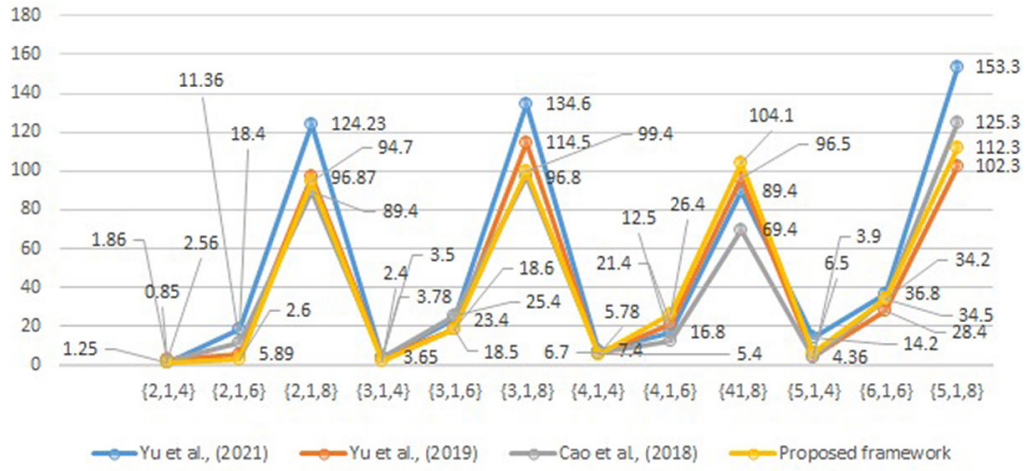

Figure 7: Computation Time of HL Frameworks with Fixed Unit Size

Source: Constructed by authors

Table 5 highlights the comparative analysis outcome with different unit sizes. The proposed framework achieved a superior objective value at different unit sizes. SARSA and Double Q-learning algorithm support the proposed framework to allocate optimum resources with an effective transportation plan. Figure 8 outlines the computation time for the HL frameworks.

Table 5: Comparative Analysis Outcome with Different Unit Size

Instances	State Size	Yu *et al.* (2021)		Yu *et al.* (2019)		Cao *et al.* (2018)		Proposed Framework	
		Objective Values	Time (seconds)	Objective Values	Time (seconds)	Objective Values	Time (seconds)	Objective Values	Time (seconds)
{2,1,4}	275	2,365.2	1.25	2,863.4	2.56	1,986.4	1.86	1,861.7	0.85
{2,2,6}	986	4,589.5	26.4	4,369.4	36.4	3,265.4	24.7	2,364.6	12.4
{2,3,8}	2,456	7,896.3	78.9	6,896.4	104.5	5,689.4	123.4	4,563.7	78.36
{3,1,4}	563	3,256.1	3.65	3,124.6	3.78	2,156.4	3.5	2,145.7	2.4
{3,2,6}	1,086	5,468.5	15.6	5,856.4	45.6	4,125.6	34.7	5,896.4	15.7
{3,3,8}	3,124	8975.6	96.4	8,964.5	112.5	6,325.4	89.7	7,893.4	89.4
{4,1,4}	956	3,452.3	7.4	3,654.7	5.78	2,896.4	6.7	1,956.4	5.4
{4,2,6}	2,456	5,468.1	18.9	5,369.8	37.9	3,964.6	36.9	3,654.5	18.4
{4,3,8}	4,286	7,896.1	106.7	8,594.2	98.7	6,457.8	104.6	7,263.1	93.4
{5,1,4}	987	3,658.4	14.2	2,896.4	4.36	2,993.8	3.9	2,314.5	6.5
{5,2,6}	2,364	5,426.1	68.4	4,561.5	29.4	4,216.8	39.8	4,526.4	25.4
{5,3,8}	3,565	7,589.4	123.4	8,968.7	99.4	5,968.7	105.6	8,756.1	101.3

Source: Constructed by authors

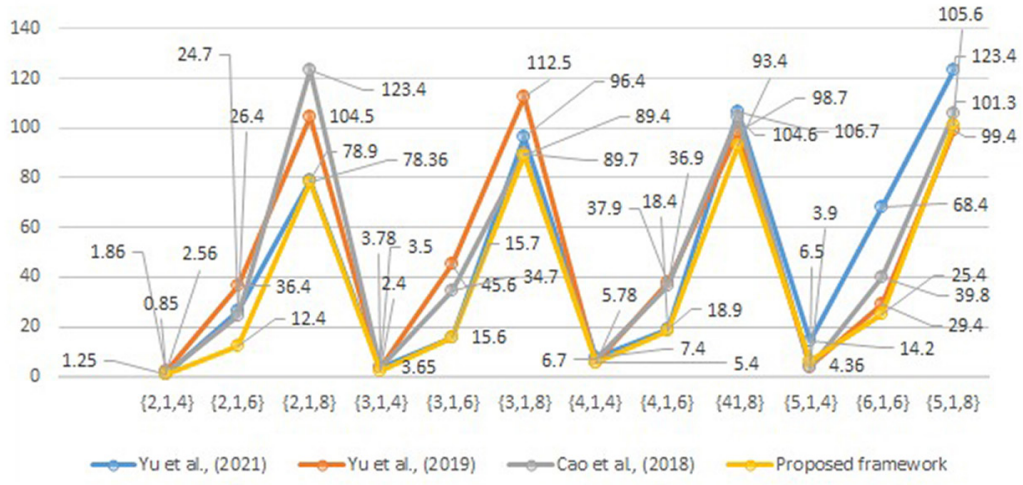

Figure 8: Computation Time of HL Frameworks with Different Unit Size

Source: Constructed by authors

CONCLUSIONS

In this study, we develop an humanitarian logistics resource allocation model to reduce the delivery cost, start the state deprivation cost, and end the state penalty cost. A unique approach is devised based on the Q-learning algorithm, a reinforcement learning technique to guarantee the efficacy, efficiency, and fairness of the allocation processes. The authors provide an in-depth explanation of the SARSA algorithm's core components and operating principles, including the environment, action space, and reward functions. The experimental section also covers adjusting the parameters of the suggested method. To prove the correctness and effectiveness of the suggested algorithm, numerical experiments are given and compared to the outcomes of a straightforward approach and a greedy heuristic algorithm. Based on the data, the SARSA algorithm is superior to the dynamic programming approach in terms of processing speed and to the greedy algorithm regarding final objective values. In scenarios when the unit size is 1, the suggested approach can get the best answer in a limited time if the episode is large enough. In scenarios with different unit sizes, the SARSA method converges into a viable approach in a limited time. Therefore, the suggested framework performs excellently when applied to the nonlinear, multi-period, multi-objective resource allocation issue. In particular, the higher number of episodes and pre-algorithm training may be used to improve the SARSA algorithm for use in actual settings.

ACKNOWLEDGEMENT

The support offered by AlMaarefa University is appreciated by authors.

REFERENCES

Agafonov, A. and Myasnikov, V. (2021): Traffic Signal Control: a Double Q-learning Approach. In *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)* (pp.365-369). IEEE.

Agafonov, A., Yumaganov, A. and Myasnikov, V. (2022): An Algorithm for Cooperative Control of Traffic Signals and Vehicle Trajectories. In *2022 4th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)* (pp.675-680). IEEE.

Agafonov, A.A., Yumaganov, A.S. and Myasnikov, V.V. (2022): Adaptive Traffic Signal Control Based on Neural Network Prediction of Weighted Traffic Flow. *Optoelectronics, Instrumentation and Data Processing*, Vol. 58, No. 5, pp.503-513.

Anuar, W.K., Moll, M., Lee, L.S., Pickl, S. and Seow, H.V. (2019): Vehicle routing optimization for humanitarian logistics in disaster recovery: A survey. In *Proceedings of the International Conference on Security and Management (SAM)* (pp.161-167). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Benkacem, A., Kamach, O., Chafik, S. and Frichi, Y. (2022): Supervised machine learning to allocate emergency department resources in disaster situations. In *2022 14th International Colloquium of Logistics and Supply Chain Management (LOGISTIQUA)* (pp.1-6). IEEE.

Cao, C., Li, C., Yang, Q., Liu, Y. and Qu, T. (2018): A novel multi-objective programming model of relief distribution for sustainable disaster supply chain in large-scale natural disasters. *Journal of Cleaner Production*, Vol. 174, pp.1422-1435.

Castellanos, C.L., Marti, J.R. and Sarkaria, S. (2018): Distributed reinforcement learning framework for resource allocation in disaster response. In *2018 IEEE Global Humanitarian Technology Conference (GHTC)* (pp.1-8). IEEE.

Chamola, V., Hassija, V., Gupta, S., Goyal, A., Guizani, M. and Sikdar, B. (2020): Disaster and pandemic management using machine learning: a survey. *IEEE Internet of Things Journal*, Vol. 8, No. 21, pp.16047-16071.

Dubey, R., Bryde, D.J., Foropon, C., Tiwari, M., Dwivedi, Y. and Schiffling, S. (2021): An investigation of information alignment and collaboration as complements to supply chain agility in humanitarian supply chain. *International Journal of Production Research*, Vol. 59, No. 5, pp.1586-1605.

Hachiya, D., Mas, E. and Koshimura, S. (2022): A Reinforcement Learning Model of Multiple UAVs for Transporting Emergency Relief Supplies. *Applied Sciences*, Vol. 12, No. 20, p.10427.

Jaiswal, A., Kumar, S. and Dohare, U. (2020): Green Computing in Heterogeneous Internet of Things: Optimizing Energy Allocation Using SARSA-based Reinforcement Learning. In *2020 IEEE 17th India Council International Conference (INDICON)* (pp.1-6). IEEE.

Khadilkar, H. (2018): A scalable reinforcement learning algorithm for scheduling railway lines. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 2, pp.727-736.

Khan, I.H. and Javaid, M. (2020): Automated COVID-19 emergency response using modern technologies. *Apollo Medicine*, Vol. 17, Suppl. 1, p.S58-S61.

Kim, Y., Kim, S. and Lim, H. (2019): Reinforcement learning based resource management for network slicing. *Applied Sciences*, Vol. 9, No. 11, p.2361.

Klaine, P.V., Nadas, J.P., Souza, R.D. and Imran, M.A. (2018): Distributed drone base station positioning for emergency cellular networks using reinforcement learning. *Cognitive Computation*, Vol. 10, No. 5, pp.790-804.

Lonkar, Y.S., Bhagat, A.S. and Manjur, S.A.S. (2019): Smart disaster management and prevention using reinforcement learning in IoT environment. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp.35-38). IEEE.

Lopez, C. (2019): Distributed reinforcement learning in emergency response simulation (Doctoral dissertation, University of British Columbia).

Lopez, C., Marti, J.R. and Sarkaria, S. (2018): Distributed reinforcement learning in emergency response simulation. *IEEE Access*, Vol. 6, pp.67261-67276.

Lu, S., Christie, G.A., Nguyen, T.T., Freeman, J.D. and Hsu, E.B. (2022): Applications of artificial intelligence and machine learning in disasters and public health emergencies. *Disaster Medicine and Public Health Preparedness*, Vol. 16, No. 4, pp.1674-1681.

Mohammed, A., Nahom, H., Tewodros, A., Habtamu, Y. and Hayelom, G. (2020): Deep reinforcement learning for computation offloading and resource allocation in blockchain-based multi-UAV-enabled mobile edge computing. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp.295-299). IEEE.

Munawar, H.S., Hammad, A.W. and Waller, S.T. (2021): A review on flood management technologies related to image processing and machine learning. *Automation in Construction*, Vol. 132, p.103916.

Nadi, A. and Edrissi, A. (2016): A reinforcement learning approach for evaluation of real-time disaster relief demand and network condition. *International Journal of Economics and Management Engineering*, Vol. 11, No. 1, pp.5-10.

Nassar, A. and Yilmaz, Y. (2019): Reinforcement learning for adaptive resource allocation in fog RAN for IoT with heterogeneous latency requirements. *IEEE Access*, Vol. 7, pp.128014-128025.

Vereshchaka, A. and Dong, W. (2019): Dynamic resource allocation during natural disasters using multi-agent environment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (pp.123-132). Springer, Cham.

Wang, H., Liu, C.H., Dai, Z., Tang, J. and Wang, G. (2021): Energy-efficient 3D vehicular crowdsourcing for disaster response by distributed deep reinforcement learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp.3679-3687).

Wu, C., Ju, B., Wu, Y., Lin, X., Xiong, N., Xu, G., Li, H. and Liang, X. (2019): UAV autonomous target search based on deep reinforcement learning in complex disaster scene. *IEEE Access*, Vol. 7, pp.117227-117245.

Yan, Y., Chow, A.H., Ho, C.P., Kuo, Y.H., Wu, Q. and Ying, C. (2022): Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 162, p.102712.

Yang, Z., Nguyen, L., Zhu, J., Pan, Z., Li, J. and Jin, F. (2020): Coordinating disaster emergency response with heuristic reinforcement learning. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp.565-572). IEEE.

Yu, L., Yang, H., Miao, L. and Zhang, C. (2019): Rollout algorithms for resource allocation in humanitarian logistics. *IISE Transactions*, Vol. 51, No. 8, pp.887-909.

Yu, L., Zhang, C., Jiang, J., Yang, H. and Shang, H. (2021): Reinforcement learning approach for resource allocation in humanitarian logistics. *Expert Systems with Applications*, Vol. 173, p.114663.

Yu, L., Zhang, C., Yang, H. and Miao, L. (2018): Novel methods for resource allocation in humanitarian logistics considering human suffering. *Computers & Industrial Engineering*, Vol. 119, pp.1-20.

Zou, R., Bury, N., Hasegawa, H., Jinno, M. and Subramaniam, S. (2021): December. DeepDRAMA: Deep Reinforcement Learning-based Disaster Recovery with Mitigation Awareness in EONs. In *2021 IEEE Global Communications Conference (GLOBECOM)* (pp.1-6). IEEE.

BIOGRAPHY

Dr Khalid I. Alqumaizi is a highly trusted healthcare executive leader in KSA with over 20 years of experience in the health profession. His experience is equally balanced in strategic transformation of both healthcare management and healthcare academic education and training. He is an ambitious pioneer leader who supports MOH 2030 Vision National Transformation and Corporatization Program. He is a Board member of several academic and national cluster boards. He was the National Model of Care Lead, contributing to strategy design. He currently is an Advisor to AlMaarefa University, President and Dean of College of Medicine, leading the transformation in the College of Medicine and Clinical Phase of the University. He was the Medical School Dean of IMSIU. He has published many research papers in reputable international journals, and contributed to national and international conferences. He has received outstanding leadership training in multiple reputable management schools.

Prof. Dr Ashit Kumar Dutta works as a Full Professor in the Department of Computer Science and Information Systems at AlMaarefa University. He specialises in the fields of artificial intelligence and cyber security. According to the Stanford Ranking, he is among the top 2% of scientists in the world, and has 21 years' experience in both national and international levels of education. He has published many research papers in ISI journals and received various meritorious awards at both national and international level. He has completed eight research projects at local and international levels. He received the best researcher award by the Nature Science Foundation in 2022. He is a member of the editorial board of several international journals. He is one of the committee members for scientific research, curriculum development, and quality management of educational institutions. He is a certified ethical hacker by the EC Council, USA.

Dr Sultan Alshehri has a PhD from the School of Pharmacy, University of Mississippi, USA, 2015. Following his return to Saudi Arabia, he was appointed as an Assistant Professor at the College of Pharmacy, King Saud University. Since then, he has been involved in teaching multiple courses for undergraduate and graduate students. His research interests include solid dosage forms, solid dispersion, formulation design, bioequivalence analyses and industrial pharmacy. On the administrative level, Dr Alshehri was actively involved early on in many committees and councils in and outside the college, leading him to be nominated as Vice Dean for Academic Affairs and Quality, School of Pharmacy, AlMaarefa University, in addition to his aforementioned academic and research activities. Moving forward, Dr Alshehri has been promoted to Associate Professor and continues to expand his knowledge base in the pharmaceutical industry field.