Service resource allocation problem in the IoT driven personalized healthcare information platform

1st Ji Fang  
School of Economics and Management  
Southeast University  
Nanjing, China
Faculty of Information Technology  
Monash University  
Melbourne, VIC, Australia
ji.fang@monash.edu

2nd Vincent CS Lee  
Faculty of Information Technology  
Monash University  
Melbourne, VIC, Australia
vincent.cs.lee@monash.edu

3rd Haiyan Wang  
School of Economics and Management  
Southeast University  
Nanjing, China
hywang@seu.edu.cn

Abstract—With real-time monitoring of the personalized healthcare condition, the IoT wearables collect the health data and transfer it to the healthcare information platform. The platform processes the data into healthcare recommendations and then delivers them to the users. The IoT structures in the personalized healthcare information service allows the users to engage in the loop in servitization more convenient in the COVID-19 pandemic. However, the uncertainty of the engagement behavior among the individual may result in inefficient of the service resource allocation. This paper seeks an efficient way to allocate the service resource by controlling the service capacity and pushing the service to the active users automatically. In this study, we propose a deep reinforcement learning method to solve the service resource allocation problem based on the proximal policy optimization (PPO) algorithm. Experimental results using the real world (open source) sport dataset reveal that our proposed proximal policy optimization adapts well to the users’ changing behavior and with improved performance over fixed service resource policies.

Index Terms—personalized healthcare information service, Internet of Things (IoT), value co-creation, user experience in the loop in servitization, dynamic service resource allocation control

I. INTRODUCTION

The COVID-19 pandemic has literally reshaped both business operations and human consciousness regarding the healthcare system [1]. Specifically, due to the lockdown policy, innovative online-based personalized fitness platforms are gradually overtaking their traditional counterparts, since it is difficult for customers to physically attend fitness centers. Besides, people around the world, being required to stay at home for a long time, started paying more attention to their health conditions. These two changing factors together have caused surging trends of both the selling and the usage of digital healthcare devices, yielding more proactive user engagement in the data collection practices [2]. [3]. In light of this new trend, innovative healthcare information platform services, e.g., Peloton and Strave, started shifting focuses on enhancing user engagement experience, by leveraging the Internet of Things (IoT) structures, real-time data collection, and instant transmission. Given the importance of user experience, this work aims to discuss how healthcare service providers can improve their competitiveness via offering interactive health information services, through exploiting the heterogeneous IoTs’ structure based on value co-creation service model.

In our healthcare information service system, we focus on the interaction between the service provider with the assumption that users are independent (i.e., the interactions between them are exogeneous implying that every user is unwilling to share personal data). In our personalized healthcare information platform, the loop in servitization is described in Fig. 1. Users subscribe the specific healthcare services such as building muscle, losing weight, etc to achieve their fitness goals aim to achieve chronic disease management. Then, the platform offers the personalized information service to fit the users’ healthcare needs during the service period. At each interaction between the platform and the user during the service, the platform collects data from several data sources connecting with the user (e.g., wearable devices, treadmill) and delivers a set of healthcare information to the user through data analyzing and information processing. This process requires the service resource in the platform. The value of the healthcare information service is generated when the customers make efforts and perform behaviors using the received information. The value of the healthcare information service is co-created by the service provider and the customer during the service period [4]. [5].

However, a large gap existing in the literature between the service inputs and the health outcomes is that quantifying the service value of the heterogeneous users is based on service-dominant logic (SDL). In the SDL, although the service provider is the key driver of value creation, health outcomes are generated only if the users are actively engaged in the servitization loop. The individual behavior cannot be directly controlled by the service provider in the personalized healthcare information service. Therefore, the problem is that the service resource allocation may result individual’s different health outcomes, depending on the each user’s willingness to participate in the healthcare management.

This paper focuses on the service resource allocation prob-
problem in the healthcare information platform based on IoT structure. The IoT structure is used to monitor and collect personalized data from several physical devices about the physical activities, and share the information between the service provider and the users. During the service period, the service provider controls the service capacity dynamically and pushes the healthcare service to the users, who will follow the health recommendations [11]. The users engagement behavior can be observed by the service provider. The value of the personalized healthcare information service is created when the provider and the user work together. In our resource allocation problem, the platform faces the following questions: (1) How to input the service capacity in every decision epoch? (2) How to allocate the service capacity in the time (serving more users at the same time or providing higher quality of service focus on fewer users)? (3) Which users should be served in the time?

To solve these problem, we propose a novel algorithm for the healthcare information platform to allocate the service resource automatically adapt with the customers’ engagement behavior. The experiment results show that the presented algorithm may increase the efficiency of the service because the resources are allocated to individuals who are willing to engage in healthcare management at that time. Our contributions are three-folds,

- Exploit an Internet of things (IoT) driven personalized healthcare information service, described as service-dominant logic (SDL) service for enhancing the communication between the service provider and the customer, and delivering a seamless turnkey users experience in the service.
- Describe the value co-creation process for the healthcare information service as illustrated by the data-driven sequential decision making process, reflecting the dynamic and heterogeneous nature of the personalized health outcomes.
- Formulate a Proximal Policy Optimization for the personalized healthcare management embedded in the healthcare information service, which can meet the individual healthcare need adaptively and improve the users service experience.

II. RELATED WORK

A whole range of different approaches has been developed in the healthcare information service, including mathematical models [12] and data-driven models [13]. However, these approaches are necessary but insufficient to solve the service resource allocation problem due to the personal differences, process optimization in the long-term service, and the individual changing engagement behavior. Hence, an advanced learning method is needed for solving the service resource allocation problem in the personalized healthcare information platform enhancing learning capability to train decision policy to adapt to the personalized time-varying behavior and the dynamic health outcome in value co-creation process. Deep reinforcement learning has been used to solve the resource allocation problems in some other service systems [14].

III. THE RESOURCE ALLOCATION PROBLEM IN THE HEALTHCARE INFORMATION PLATFORM

The healthcare information service is a kind of subscription business in which the user pays a recurring price at regular intervals for access to the service. We focus on the service operation of the platform during one service interval assuming no changes in both the market size and total income. We assume that the platform serves a finite set of customers, \( N = \{1, 2, \ldots, n\} \). Without loss of generality, all the users subscribe for \( T \) time service interval at the same time. We consider a series of pre-chosen decision points \( t \in \{1, \ldots, T\} \) where the user-platform interaction happens, for example, the platform delivers the live courses in the fixed time every day in online coaching services.

We model the resource allocation problem using the sequential-decision-making model considering the users’ dynamic behavior. Fig. 2 describes the agent-environment interaction process in the healthcare information service. At each interaction, the agent chooses an action \( a_i \) among all possible actions, including the service capacity control \( c_t \) and the service allocation decision \( y_i \) (i.e., determine which individuals have received the healthcare recommendation message). The users make efforts to implement the healthcare improvement planning and have changes in their health states which can be observed through the IoT devices. The environment yields a numerical reward \( R_t \) for the chosen action. The learning agent adjusts its action in each decision epoch with either the highest reward value based on the historical data (exploitation) or a random action (exploration). The key elements of the resource allocation problem are described as follows.

1) Agent and action space: The learning agent represents the healthcare information platform that needs to make the service capacity allocation decision to serve \( n \) users. The learning agent takes action \( a_i \) to the environment in time \( t \) in the deep reinforcement learning algorithm. Each action \( a_i \) will result in a change in the service outcome \( h_i^t \) with a reward \( r \) after the user \( i \) receives the healthcare information service and makes an effort \( z_i^t \) to improve the health. The action \( a_i \in A \), is a vector of actions including two components: the service capacity input \( c_t \) and the service provision actions.
over $n$ users $y_i$, where $a_t = [c_t, y_t]$. For the service capacity input $c_t$, the action space is defined as $\{C_1, C_2, \ldots, C_m\}$. We assume that there are $m$ kinds of alternative service capacity plans. The service resource can be controlled by arranging the training experts’ working time in the platform, maintaining the servers, and so on. For the service provision actions over $n$ users $y_t$, the allowed actions on the user $i$ include not provide service (i.e., $y_i^t = 0$) or provide service (i.e., $y_i^t = 1$). Therefore, the size of the entire action space is $m \cdot 2^n$. To overcome the “curse of dimensionality” for the resource allocation problem, we use a deep reinforcement learning algorithm to solve this problem by representing the learned functions as a neural network.

2) Environment and state space: The $n$ users constitute the environment in the deep reinforcement learning algorithm, which interacts with the learning agent and co-creates the health outcome. State $s_t = [b_t, q_t, x_t, h_t, E_t]$, where $s_t \in S$ is a vector at time $t$. Each component is defined as follows:

- $b_t \in \mathbb{R}_+:$ available operation cost balance at time $t$, which depends on the action $a_t$ and last state. Non-negative balance $b_t \geq 0$ is considered as a constraint in the deep reinforcement learning algorithm.
- $q_t \in \mathbb{R}_+:$ the quality of service received by the user at time $t$, which depends on the action $a_t$ directly. We assume the quality of service is distributed equally among the users who accessed the platform. The quality of service received by the user $q_t$ is calculated by $q_t = \frac{\sum_{i=1}^{n} y_i^t}{\sum_{i=1}^{n} y_i^t}$ based on the action $a_t$.
- $x_t \in \mathbb{R}_+:$ the engagement behavior of each user depending on the inner behavior motivation of the individual, where $n$ denotes the number of users.
- $h_t \in \mathbb{R}_+:$ the health outcome of each user, represents the co-creation value by the platform and the user, where $n$ denotes the number of users.
- $E_t \in \mathbb{R}_+:$ other related information of each user observed by the heterogeneous IoT devices. The motivation of human engagement behavior is complex. When predicting the state of the users, various information needs to be considered, e.g., the fatigue level and sleep state. We assume that $\xi$ kinds of indicators can be collected in the healthcare information service system.

Hence, the state space is a $(\xi n + 2n + 2)$-dimensional vector. The neural network structure is used in the deep reinforcement learning algorithm to deal with the discrete and quite large state space.

3) State transition: The state transition of the service process is shown in Fig. [2] At each state, action $a_t$ is chosen including the service capacity input $c_t$, and the service providing actions over $n$ users $y_t$. $y_i^t$ represents the service action on user $i$ ($i = 1, \ldots, n$). The available operation cost balance $b_{t+1}$ will change along with the service input in time $t$, where $b_{t+1} = b_t - c_t \geq 0, b_0 = B$. If the $b_{t+1} < 0$, the next actions $c_{t+1}$ will equal 0 until the end of the service period. Sequentially, the transition of the health outcome $h_t^i$ for the user $i$ is defined as $h_{t+1}^i = h_t^i + y_t^i (x_t)\alpha_1 q_t^{\alpha_2}$, where $\alpha_1, \alpha_2$ are the given parameters in the healthcare information service system. The health output is considered as a Cobb-Douglas function, which is used to model the relationship between production output and production inputs. This function is in fact a value co-creation function which is also adapted in Demirezen et al. [6]. In our resource allocation problem, $\alpha_1, \alpha_2$ are the service-specific parameters, determined by the importance between the effort produced by the user and the effort produced by the platform in the value co-creation, where $\alpha_1, \alpha_2 \in (0, 1), \alpha_1 + \alpha_2 < 1$ implies a decreasing return to scale so that the value co-creation service cannot produce an infinite amount of output. For example, the online healthcare consultation service is more dependent on the platform effort than the fitness APP service with a higher value of $\alpha_2$.

4) Reward function: We define the reward function as the change of the health outcome value when action $a_t$ is taken at state $s_t$ and arriving at new state $s_{t+1}$, denoted as follows.

$$R(s_t, a_t, s_{t+1}) = h_{t+1} - h_t = y_t^i (x_t)\alpha_1 \left( \frac{c_t}{\sum_{i=1}^{n} y_t^i} \right)^{\alpha_2}.$$ (1)

The objective of the resource allocation problem and the constraints can be written as follows:

$$\max \sum_{i=1}^{n} H_t^i, \text{subject to } \sum_{t=1}^{T} \beta c_t \leq B. \quad (2)$$

The performance of the healthcare information service system is considered as the all the users’ health outcomes at the end of service because each user experiences the healthcare information service by the collaboration and develops a sense of trustworthiness toward the service platform based on the health improvement. In addition, considering the sustainable operations in the service system, the service operation cost must be lower than the service operation budget, for a given constant $B$, to ensure profitability throughout the service interval, hence $\sum_{i=1}^{T} \beta c_t \leq B$, where the constant $\beta$ represents the cost elasticity of the service capacity input.
IV. A DEEP REINFORCEMENT LEARNING ALGORITHM: A PROCESS OF PPO

The deep reinforcement learning algorithm aims to learn an optimal adaptive service resource allocation decision rule that maximizes the total health outcome improvement for $n$ users. The policy $\pi$ is a computable function that outputs for each state $s_t \in S$ an action $a_t \in A$. Therefore, a policy $\pi$ is a distribution over actions given states:

$$\pi(a_t|s_t) \equiv \mathbb{P}[a_t|s_t], \forall a_t \in A, s_t \in S, t = 0, 1, \ldots, T. \quad (3)$$

The value function under policy $\pi$ describes the expected return for the agent from a given state $s_t$ in period $t$, and is defined as $v_{\pi}(s_t) = \mathbb{E}_\pi[r_t + v_{\pi}(s_{t+1})|s_t]$, $s_t \in S$, where $\mathbb{E}_\pi[\cdot]$ denotes an expected value, $v_{\pi}(s_{t+1})$ denotes the value function for the next state $s_{t+1}$.

Solving the resource allocation problem needs to find optimal policies $\pi^*$ that achieve maximum reward during the service. Therefore, there exists an optimal policy $\pi^*$ better than or equal to all other policies, $\pi^* \geq \pi, \forall \pi$, where

$$\pi^* \equiv \arg \max_{\pi} v_{\pi}(s_t). \quad (4)$$

Our algorithm to learn the optimal policy $\pi^*$ to solve the resource allocation problem is combined deep learning method and reinforcement learning method. On the one hand, the deep learning method with a multilayer perceptron neural network is used to overcome the limitations of the high-dimensional state space in the problem. Besides, the proximal policy optimization (PPO) algorithm, a kind of reinforcement learning method, is employed to train the agent considering the multi discrete action space. PPO is based on an actor-critic structure combining the generalization method (actor network) and parameterized policy method (critic network) \cite{7, 8}. There are two separate modules in the training agent, namely "actor" and "critic", which control the action selection and criticize the actions in the actor, respectively. The idea of the trust region policy optimization is used to improve the learning performance of the actor in the PPO. The main structure of one of the training agents is illustrated in \cite{3}. The actor network with weights $\theta$ and the critic network with weights $\varphi$ are designed to learn the parameterized policy $\pi(a_t|s_t; \theta)$ and to evaluate the current parameterized policy $\pi(a_t|s_t; \theta)$ by computing the state-value function $v(s_t; \varphi)$, respectively. In addition, the actor network stores the old policy with weights $\theta_{old}$ before an update, denotes as $\pi(a_t|s_t; \theta_{old})$, which can measure the difference between the new policy and the old policy.

The parameters $\varphi$ of the critic network can be trained through gradient descent to minimize the mean square error of the TD-error $L(\varphi)$. Then, the "Actor" updates the policy distribution $\pi(a_t|s_t; \theta)$ in the direction suggested from the "Critic" by maximizing an objective function $L^{clip}(\theta)$ with respect $\theta$.

The value-function error for critic network training is:

$$L(\varphi) = (R_t - v(s_t; \varphi))^2. \quad (5)$$

On the other hand, the actor network training uses the objective function as follows.

$$L^{clip}(\theta) = \mathbb{E}_t \left[ \min \left\{ r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t \right\} \right], \quad (6)$$

where $\mathbb{E}_t[\cdot]$ denotes the empirical expectation over timesteps (i.e., time $t$). $r_t(\theta)$ is the probability ratio between old policy and the new policy, which is described as $r_t(\theta) = \frac{\pi(a_t|s_t; \theta)}{\pi(a_t|s_t; \theta_{old})}$. $\hat{A}_t$ is the estimated advantage at time $t$, considering the state-value function $v_{\pi}(s_t)$ as a baseline to evaluate an action for current policy $\pi$ is given as follows.

$$\hat{A}_t = q_{\pi}(s_t, a_t) - v_{\pi}(s_t). \quad (7)$$

The function clip $(r_t(\theta), 1 - \epsilon, 1 + \epsilon)$ clips the ratio $r_t(\theta)$ falling within the interval $[1 - \epsilon, 1 + \epsilon]$, where $\epsilon$ is a hyperparameter. The objective function $L^{clip}(\theta)$ takes the minimum of the normal and the clipped objective avoiding large deviation from the old policy, which can enable the PPO algorithm to improve stability of the policy networks training. The pseudocode of the proposed PPO algorithm is provided as follows. The input data includes: the architecture of the actor network, and the critic network with the initial weights $\theta_{old}, \varphi$. clip range $\epsilon > 0$; Total training steps $T$; Step counter per update $n$. The result is outputting the trained $\theta, \varphi$.

V. PERFORMANCE EVALUATION

A. Data source

We use the lifelog data of the daily activities collected from 16 participants from 01/11/2019 to 30/03/2020. The dataset combines input from Fitbit Versa 2 smartwatch wristbands, the PMSys sports logging smartphone allocation, and Google forms. Refer to \cite{9} for more details.

The performance of the value co-creation is calculated according to engagement behavior of the participant in the healthcare management. The indicator “readiness” in PMSys represents how ready the participant is to engage in the healthcare activity, ranging from 0 (not ready at all) to 10 (fully ready). Four other kinds of indicators related to the
individual daily exercise are also considered, including “calories” collected by the Fitbit Versa 2 smartwatch, the “fatigue” and “mood” extracted from the PMSys, and the “training load of perceived exertion (sPRE)”, which is the product of the training load and the reported rating of perceived exertion. We use Pearson correlation to analyse the correlation between “readiness” and other four indicators. Results confirm that there exists statistically correlations between “readiness” and other indicators, especially “fatigue” and “mood”.

B. Performance of the deep reinforcement learning algorithm

The model is built with Python and Tensorflow, and run on a PC with Intel i9 CPU and 32.00 GB of RAM. All results are reported over an average of 100 simulations. Besides, the dataset is partitioned into training and test set with a date-independent setting, where the former contains data from the first 121 days and the latter contains the last 30 days.

In reinforcement learning algorithm, the agent is trained to learn the participants’ engagement behavior from the time series life logging data. Then, we compare our adaptive policy with three kinds of fixed resource allocation plans, including the service capacity equal distribution in every time (denoted as the “fixed plan 1”), delivering the healthcare information service to every participant in each time (denoted as the “fixed plan 2”), and the service capacity equal distribution and delivering the healthcare information service to every participant in every time (denoted as the “fixed plan 3”). To evaluate the performance, we test our deep reinforcement learning algorithm in the back-testing experiments. We assume that the healthcare information platform prepares 16 alternative service capacity decision plans with a range 0-15, where level 0 is not inputting service capacity, level 1 is the lowest service capacity, and 15 is the highest service capacity level.

The neurons number in 1st hidden layer of the proposed model is set to 16, and 2nd layer is set to 32. Besides, we set clipping rate $\epsilon$ to 0.2, batch size to 256 and learning rate to 0.01. Model converges at $\sim$18,000 iterations, taking $\sim$80 seconds.

As shown in Fig. 4, our algorithm policy compares with other healthcare platform resource allocation policies for the fixed service capacity input, delivering the service to every participant at each time and pushing the fixed service capacity per day with serving every participant each day. Figure 4a shows the total rewards in the 30 days service period with 100 times running, and Fig. 4b shows the daily reward for one time running. The results indicate that the resource policy produced from our deep reinforcement learning algorithm is superior to the other fixed service resource allocation plans. Especially, the plans of delivering the service to every participant are not performed well in the healthcare information platform operation. The main reason is that providing the healthcare recommendation service to the inactive participant cannot produce the service value in the healthcare information service system. On the other hand, the dynamic service capacity input can improve the service performance for the healthcare information platform, which can adaptively control the service capacity according to the marketing activeness. Therefore, our deep reinforcement learning algorithm in the business logic tier in the healthcare information platform improves service resource utilization.

C. Benchmarking test

We compare our deep reinforcement learning algorithm with two kinds of algorithms with the same neural network architecture as used in our deep reinforcement learning algorithm (i.e., two layers with 16 and 32 neurons respectively).

Algorithm 1: PPO for the resource allocation problem

while $t \leq T$

$\begin{align*}
\text{Initialize thread step counter } t &\leftarrow 1; \\
\text{Reset accumulate gradients: } d\theta &\leftarrow 0, d\varphi \leftarrow 0; \\
\text{Initialize state } s_t; \\
\text{repeat} & \\
\text{Take action } a_t \text{ according to policy } \\
& \pi(a_t|s_t; \theta_{old}); \\
\text{Compute reward } R_t \text{ using (1)}; \\
\text{Observe new state } s_{t+1}; \\
&s_t \leftarrow t + 1; \\
\text{until } s_t \text{ is terminal or } t - t_{\text{start}} \leq n; \\
q(s, t) &= \begin{cases} 
0, & \text{for terminal } s_t, \\
v(s_t; \varphi), & \text{for non-terminal } s_t.
\end{cases} \\
\text{while } i \leq t_{\text{start}} \text{ do} & \\
& i \leftarrow t - 1; \\
q(s, i) &= R_t + q(s, i + 1); \\
\text{Compute advantage estimates } \hat{A}_i \text{ using (7)}; \\
\text{Accumulate gradients wrt } \varphi: \\
d\varphi &\leftarrow d\varphi + \frac{\partial(q(s, i; \varphi))}{\partial \varphi}; \\
i &\leftarrow i + 1; \\
\text{end} & \\
\text{Optimize the PPO clip objective function } L_{\text{clip}}(\theta) \text{ using (5)}; \\
\text{Perform } \theta \text{ update via gradient ascent with Adam}; \\
\text{Perform } \varphi \text{ update via gradient descent with Adam}; \\
\text{end}
\end{align*}$

Fig. 4: Comparisons with fixed resource allocation plans ($\alpha_1 = 0.5, \alpha_2 = 0.4, B = 100, \beta = 0.9$)
The parameters related to the resource allocation problem in the healthcare information platform are the important factors between users' effort and the platform effort in the value co-creation. These parameters are determined by the service operation cost $\alpha$ and the cost elasticity of the service $\beta$. These exogenous factors can be evaluated through behavioral experiments. Other parameters include the characteristics of the service itself, which can be assessed according to the financial information of the platform. Because of the conditions limit, we analyze the sensitivity of the resource allocation policies to the financial information of the platform. Because the more important aspect of the important factors in the healthcare information platform is the impact of the important factors on the users' effort, the more important aspect of the important factors on the platform effort in the value co-creation is the impact of the important factors on the service operation policies. An extension of the current study is to analyze the factor of the network effect on the healthcare information service, for example, the users can engage with others by liking in some social fitness applications.

VI. CONCLUSION AND FUTURE WORK

This study has led to conclude that our proposed service resource allocation mechanism, i.e., using the reinforcement learning algorithm to realize Proximal Policy Optimization and the theoretical approaches and practical implications, can help and the sense-and-respond framework, are valuable. The sense-and-respond framework can be valuable to update the system for the personalized service fusion approach based on microservice architecture.

TABLE I: Average results of the deep reinforcement learning algorithm and the competition algorithms

| Algorithm | Running time | Convergence steps | Total rewards |
|-----------|--------------|-------------------|---------------|
| TRPO      | 66.59        | 1539              | 325.24        |
| A2C       | 138.74       | 3585              | 611.45        |
| PPO       | 87.50        | 2372              | 510.80        |
| Expert    | 45.65        | 1023              | 387.74        |

REFERENCES

[1] T. M. Chen, "Doing more with less: a managerial analysis of Value Co-creation in online social networks," Journal of Online Science, vol. 39, no. 3, pp. 2057–2076, Sep. 2020.
[2] A. D. Joachims, "Value Co-creation in service-fusion approach based on microservice architecture," International Journal of Web Intelligence and Knowledge Management, vol. 39, no. 3, pp. 121–135, Apr. 2018.
[3] D. Joachims, "Learning to rank using gradient descent," in ICWS 2017, Jul. 2019, pp. 57–61.
[4] C. Joachims, "Beyond Convergence: The Power of Multilevel Optimization Methods in Machine Learning," in ICWS 2017, Jul. 2019, pp. 57–61.
[5] H. Saarijärvi, P. K. Kannan, and H. Kuusela, "Value co-creation: Theoretical approaches and practical implications," European Journal of Information Management, vol. 39, no. 3, pp. 2057–2076, Sep. 2020.
[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” Jul. 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[8] Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, and Anssi Kanervisto, “Stable baselines,” 2018. https://github.com/hill-a/stable-baselines. (accessed Mar. 21, 2022).

[9] V. Thambawita et al., “PMData: A sports logging dataset,” in MMSys 2020, May 2020, pp. 231–236.

[10] Carlos Natalino and Paolo Monti, “The Optical RL-Gym: An open-source toolkit for applying reinforcement learning in optical networks,” in ICTON 2020, 2020, pp. 1–5.

[11] S. He, B. Cheng, Y. Huang, L. Duan, and J. Chen, “Proactive Personalized Services in Large-Scale IoT-Based Healthcare Application,” ICWS 2017, Sep. 2017, pp. 808–813.

[12] G. Roels, “High-performance practice processes,” Management Science, vol. 66, no. 4, pp. 1509–1526, 2020.

[13] B. Sayin et al., “Crowd-Powered Hybrid Classification Services: Calibration is all you need,” ICWS 2021, 2021, pp. 42–50.

[14] S. P. Panda, A. Banerjee, and A. Bhattacharya, “User Allocation in Mobile Edge Computing: A Deep Reinforcement Learning Approach,” ICWS 2021, 2021, pp. 447–458.