Dynamic physical activity recommendation on personalised mobile health information service: A deep reinforcement learning learning approach

Ji Fang\textsuperscript{1,2}, Vincent CS Lee\textsuperscript{2}, Haiyan Wang\textsuperscript{1,*}

\textsuperscript{1}School of Economics and Management, Southeast University, Nanjing, 210096, P R China
\textsuperscript{2}Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University (Clayton Campus), Melbourne, VIC, 3800, Australia

\textsuperscript{1}First author: ji.fang@monash.edu
\textsuperscript{2}Second author: vincent.cs.lee@monash.edu
\textsuperscript{1,*}Corresponding and third author: hywang@seu.edu.cn
Abstract

Mobile health (mHealth) information service makes healthcare management easier for users, who want to increase physical activity and improve health. However, the differences in activity preference among the individual, adherence problems, and uncertainty of future health outcomes may reduce the effect of the mHealth information service. The current health service system usually provides recommendations based on fixed exercise plans that do not satisfy the user’s specific needs. This paper seeks an efficient way to make physical activity recommendation decisions on physical activity promotion in personalised mHealth information service by establishing data-driven model. In this study, we propose a real-time interaction model to select the optimal exercise plan for the individual considering the time-varying characteristics in maximising the long-term health utility of the user. We construct a framework for mHealth information service system comprising a personalised AI module, which is based on the scientific knowledge about physical activity to evaluate the individual's exercise performance, which may increase the awareness of the mHealth artificial intelligence system. The proposed deep reinforcement learning (DRL) methodology combining two classes of approaches to improve the learning capability for the mHealth information service system. A deep learning method is introduced to construct the hybrid neural network combining long-short term memory (LSTM) network and deep neural network (DNN) techniques to infer the individual exercise behavior from the time series data. A reinforcement learning method is applied based on the asynchronous advantage actor-critic algorithm to find the optimal policy through exploration and exploitation. We tested our DRL methodology using real-world data from a runner program, and our policy was validated by comparison to other fixed exercise plans. The results show that the proposed physical activity recommendation system for the personalised mHealth information service, represented as a personalized AI module, can adapt to the users' changing behaviour, boost the users' exercise performance, and promote a pareto optimal healthy lifestyle.

Keywords: OR in health services; mHealth information service system; personalised practice process optimization; real-time interaction model; A deep reinforcement learning methodology
1. Introduction

Mobile Health (mHealth) information services combine the users' physiology data and daily exercise information with their fitness demands, such as Fitbit, Pelton. These services provide recommendations of short-term physical activity plans for their users, improving their performance of the physical activities and enabling them to live a healthy life. For decades, wearable activity trackers and fitness applications have been widely adopted in health self-management, making it low cost and convenient for users to exercise under the guidance of a scientific training system (Liu & Avello, 2020; Nahum-shani et al., 2014). Typically, mHealth information services record and visualize the individual exercise activities through wearable trackers for their users (Kong et al., 2020), then deliver a set of recommendations for daily exercise via mobile applications. The mHealth information service has been used to support health promotion programs in physical activity (Hardeman et al., 2019; Schoeppe et al., 2017), weight management (Aswani et al., 2019; Gasparetti et al., 2020). Perhaps a growing number of people will manage their health via mHealth information service in the coming years with the health risk increasing caused by the prevalence of obesity and sedentary lifestyle habits.

Although the mHealth information services have great potential for their users' promoting health without limitation by time and location using mobile devices, the challenge problem is evaluating and keeping the performance of the physical activity recommendations (Blackman et al., 2013). In these mHealth information services, the reason for the inefficient may be threefold. First, most popular mHealth services push non-personalized and fixed physical activity plans to their users. There is increasing evidence that such static daily exercise plans may result in unchanged or even reduced physical activity to the individual (Zhou et al., 2018). Hence, choosing effective recommendations for the individual daily exercise is difficult because of the heterogeneity of physical characteristics and fitness demands. Second, the long duration of mHealth information service with the uncertainty of future health outcomes is another difficult issue for physical activity recommendations. Physical activity recommendations have impacts on the users' exercise base levels in the long term associated with the changing of the "stock of fitness" and the "stock of fatigue" in each workout (Roels, 2020; Sedighi Maman et al., 2020). Third, health improvement depends on every effort made by the individual to implement daily exercise plans (Jain, 2009). However, the individual has self-control problems which may cause failure to complete the daily tasks (O'Donoghue & Rabin, 2011). The pre-designed exercise program does not work for the individual uncertain behavior. In conclusion, the heterogeneities of physical characteristics, fitness goals, and exercise preferences across the individuals make the physical activity recommendation problem challenging. Moreover, the individual's changing preference and uncertainty of future health outcomes further amplifies the difficulty in this long-term service. Besides, the gaps between planning and implementation on the
user's behavior lead to uncertain results of the physical activity performance.

According to our investigation, a whole range of different approaches has been developed in the healthcare information service, including mathematical models (Mcgregor et al., 2009; Roels, 2020) and data-driven models (Sánchez-Ruiz et al., 2020). However, these approaches cannot solve the physical activity recommendation problems considering the personal difference, process optimization in the long-term service, and the individual uncertain behavior. On the one hand, mathematical optimization methods (e.g., dynamic programming, decision tree models) are ineffective to adapt the ongoing feedback in our problem. Because these methods rely on exact models to characterize the environment (e.g., system states, human behavior). On the other hand, data-driven approaches cannot solve the mHealth information service recommendation problem using the knowledge-based system because there have not been apparent methods to integrate them into an optimization model (Aswani et al., 2019), such as artificial neural networks (Galán-Mercant et al., 2019; Sánchez-Ruiz et al., 2020), support vector machines (Akay et al., 2009; Oztekin et al., 2018; Vijayarajeswari et al., 2019; H. Wang et al., 2018; Xu et al., 2021). Hence, a state-of-art learning method is needed for solving the personalized mHealth physical activity recommendations problem, which can establish a framework to achieve knowledge-based mHealth information recommendation system, and enhance learning capability to train decision policy to adapt to the individual’s time-varying characteristics and the dynamic health improvement process.

In this situation, the AI system for mHealth information service is used to monitor and collect structured information about users' physical activities, which could allow the mHealth services to modify the user's exercise plans appropriately in the dynamic environment. To achieve dynamic physical activity recommendations, we propose an artificial intelligence (AI) system for the mHealth information service that integrates the IoT module (Dong et al., 2020; Nweke et al., 2018; Pan et al., 2022; Sharavana Kumar & Sarma Dhulipala, 2020; Y. Wang et al., 2019), AI module (Afzal et al., 2018; Sedighi Maman et al., 2020), and user communication module (Adjerid et al., 2021) in figure 1. The basic process of the mHealth information service system is an interaction between the system and the user in the just-in-time adaptive intervention. At each interaction, the system offers health intervention services via user communication module by formulating short-term physical activity recommendation controlled by AI module to help their users control exercise intensity and improve health. Then, the system receives feedback on the user's physical performance through the IoT module, meanwhile, adjusts physical activity plans based on individual behavior until the end time of the service.
This study presents a knowledge-based framework based on an improved fitness-fatigue model for the personalized mHealth physical activity recommendation problem and designs an advanced deep reinforcement learning (DRL) methodology to improve the learning capability for the mHealth information service system, enabling it to respond to the input information from wearable devices immediately. Each user's mHealth practice progress is modeled as a discrete-time, finite-horizon Markov decision process (MDP), which can dynamically control the personalized exercise plans involving a sequential progression of decisions. The dependency among the exercise plans over time is associated with changing behavioral preferences of the individual, which can be recorded through the interaction traces by mobile devices. To solve the MDP problem, we propose a DRL methodology to improve learning capability on the mHealth information service system. The methodology combines
a hybrid neural network and the model-free reinforcement learning algorithm to help improve the
decision-making for mHealth information service by using observational data in real-world situations.

This work proposes a personalized artificial intelligence (AI) module for the personalized mHealth
information service. Firstly, a quantitative model is built for the individual with the user’s specific
requirements and health goals. Then, we design a dynamic expert system based on the specific
quantitative model. In contrast with the conventional expert system which has no learning capability,
our proposed dynamic expert system is a DRL-based recommender, which can learn from input
temporal data to predict the individual current and future behavior and the physical activity state using
historical data. This dynamic expert system at the same time, utilises the estimated model to generate
actionable physical activity recommendations to maximise the individual's total utility, i.e. healthy life
style. The advantages of the proposed personalized AI module are in three aspects: (1)We construct
this methodology based on the widely accepted model utilising professional practice knowledge to
describe performance in exercise, improving outcome interpretability for the machine learning
approaches. (2) Being different from the traditional operation research methods in health service, this
methodology does not require a well-represented mathematical model of the decision systems. Instead,
it learns from previous experiences to develop an optimal execution strategy directly. In fact, it is often
difficult to construct an accurate model for the individual reaction of the mHealth information service
because of the differences in activity preference, time-varying characteristics, and uncertainty of future
health outcomes (Zhou et al., 2018). (3) Compared with the practice process optimization research
presented by Roels (2020), this study focus on the gaps between plans and actions. In our model, these
gaps can be bridged when the service provider can modify the exercise plans during the process. We
then provide experiments on how our policy would have performed in real life. The model is calibrated
using published data, the parameters of individual characteristics can be estimated from the real-world
data, and then our DRL methodology generates the individual-specific practice plans. The results show
that our DRL policy is superior to the fixed practice plans. Although adopting an artificial intelligence
system for the mHealth service providers requires an initial investment, these investments pay off with
higher users’ satisfaction during the long-term service. The more satisfied the customers are, the
service repurchase rate is higher, meanwhile, the goodwill of the mHealth service is increasing to
appeal to the new customer, a more competitive advantage business model.

The main contributions of this study are highlighted as follows.

(1) This study is among the first to investigate a cutting-edge Markov Decision Process (MDP)
model based on the fitness-fatigue theory (Roels, 2020) to describe the interaction between the
mHealth information service system and user considering the time-varying characteristics of
the individual’s exercise process.
A knowledge-based framework is designed to recommend physical activity plans for the mHealth information service system, reflecting the individual's adherence behavior to the exercise plans, the changing exercise preferences and the dynamic health outcome during the personalised mHealth information service.

We develop a novel end-to-end DRL methodology for the mHealth information service system to improve the learning capability by integrating the deep learning method and reinforcement method. The deep learning method combines the LSTM network and the fully connected neural network overcoming the limitations of the high-dimensional state space of individual behaviour and time-dependency of the individual's exercise base level. The reinforcement learning algorithm is designed to train the objective function value and improve the learning performance based on an asynchronous advantage actor-critic algorithm, which can adapt to control the exercise plans to meet the individual's fitness need adaptively.

2. Related work

Recently, there has been a growing concern about improving physical activity recommendation effect in personalized healthcare service (Aswani et al., 2019; Mintz et al., 2017). In a similar vein, we consider how to design the physical activity recommendation policy for personalized mHealth information service when the individual's adherence behavior is stochastic. We review the existing literature concerning: (1) health-related physical fitness and personalized mHealth service, (2) processing optimization problem for the physical activity, and (3) health recommender systems and artificial intelligence technologies.

2.1 Health-related physical fitness and personalized mHealth service

mHealth service has the potential to transform the way of health services delivering by quicker facilitation of health information. Recently, the literature concerning health-related physical fitness using the mHealth service. Health-related physical fitness is defined as people's ability to perform physical activity, which could reduce the risk of developing diseases related to physical inactivity (Liu & Avello, 2021).

Up to now, several systematic reviews and empirical studies have confirmed the effectiveness of mHealth service in physical activity interventions (Adjerid et al., 2021; Bohanec et al., 2021; Degroote et al., 2020; Domin et al., 2021; Marcolino et al., 2018; Milne-Ives et al., 2020). By using real-time situation triggered reminders, pushes, and notifications, the mHealth service can assist the individual to improve the effectiveness of daily exercise (Marcolino et al., 2018). Moreover, mHealth service makes it possible to implement just-in-time adaptive interventions for the users and collect automated
data (e.g., passive tracking of wearable devices usage behavior, movement data) as well as manual data (e.g., mental state) (Rabbi et al., 2019).

Many studies have begun to design artificial intelligence decision systems to optimize healthcare management (Mezei & Nikou, 2018; Mintz et al., 2020; Rabbi et al., 2019). However, mHealth is a complex domain that requires diverse expertise knowledge of health science and data science (Messner et al., 2019). Designing the mHealth information service system need to integrate much knowledge, such as data analysis and physical activity. We introduce widely used physical activity performance evaluating theory in our mHealth information service system to understand the fluctuation of practice performance, which could help mHealth service push suitable exercise plans to their users more scientifically.

2.2 The processing optimization problem for the physical activity

The physiology literature concerning evaluating exercise performance is vast. Analytical research on practice performance was proposed by Calvert et al. (1976). In their study, practice performance is associated with training intensity by a positive effect ("fitness") and a negative effect ("fatigue"). The fitness-fatigue model has been used to predict athletes' performance in a variety of endurance sports, including running (Mcgregor et al., 2009; Roels, 2020), swimming (Calvert et al., 1976), soccer (Jaspers et al., 2017), among others. The existing literature on practice performance is extensive and focuses particularly on the process optimization perspective (Schmidt & Bjork, 1992). Topol (2019) presented a dynamic model of the practice process to maximize performance. His study offered some important insights and guidance on the optimal amount of spacing or optimal intensity of each practice in practice process configurations.

mHealth service can only provide persuasive recommendations to engage users in physical activity, leading to the practice performance optimization model not being applied directly in mHealth service. To the best of our knowledge, practice process optimization has not been investigated considering people implement behavior.

2.3 Health recommender systems and artificial intelligence technologies

Health recommender systems are useful in changing behavior and health outcomes by selecting among multiple potential practice plans for the user. With the purpose of providing low-effort to procedure outcomes equivalent to a high level of intervention, artificial intelligence technologies are needed, which should enable actual adoption by the users (del Carmen Rodríguez-Hernández & Ilarri, 2021; Gasparetti et al., 2020). Reinforcement learning can constantly monitor the user activities and adjust recommendation policy throughout the intervention period (Forman et al., 2019). To test the
effects of reinforcement learning technology in health recommender systems, Yom-Tov et al. (2017) experimented with a reinforcement learning approach on 27 sedentary type 2 diabetes patients to encourage them to increase the level of their physical activity. The results showed that the reinforcement learning improved gradually in predicting which intervention messages would lead patients to exercise. Forman et al. (2019) evaluated the feasibility, acceptability, cost savings, and effectiveness of reinforcement learning for weight loss intervention in behavioral weight-loss trials. They observed that the reinforcement learning system can considerably lower costs in the optimized conditions while still producing a satisfying intervention effect.

Reinforcement learning has already been considered for the physical activity recommendation service. However, no research focuses on the impact of exercise intensity among potential practice plans on the outcomes in health recommender systems (e.g., over exercise). We thus introduce practice theory to evaluate the performance of physical activities in health recommender systems. In the mHealth service, we fill a gap because there is limited guidance on the optimal intensity throughout the intervention period from the practice process optimization perspective.

Overall, this study provides new insights into physical activity promotion in the healthcare management. Specifically, the scientific training principle is introduced into mHealth service to improve professional support efforts that are linked. The scientific training principle is applied to evaluate the performance, where the people who participate in physical activity are not considered as fixed entities but focus on the interaction between people and the physical activity. The recommendations policy for personalized physical activity promotion considering the exercise intensity is obtained based on the scientific training principle using data from the wearable trackers, which allows sequential evolution to tailor the individual's changing behavior. Close to our work is that by Roels (2020), who modeled the practice process and optimized its profile to maximize performance for the individual. Although he improved the practice process optimizing model on people-centric operations, our work primarily differs in considering a real-time interaction between the decision maker with the people who participate in physical activity through the intelligent devices, whereas the scope of his study is under the deterministic environment. Our decision maker's challenge is to design an adaptive policy to solve the mHealth optimization problem by modeling it into the Markov decision process, whereas his decision maker involves multiple phases of the optimal practice process. Beyond this, our work considers the incentive effect of mHealth interventions.

3. Personalised modelling

The personalized mHealth information service contains two stages of exercise: exercise skill acquisition stage and the exercise skill retention stage. According to Roels (2020), repeated practices
have a compounding effect on memories through a "reconsolidation" process, while early skill acquisition is needed before practice can have a compounding effect. Therefore, the additive model is considered to fit the exercise skill acquisition stage, and the multiplicative model is considered to fit the exercise skill retention stage (Roels, 2020).

On top of that, we consider that the decision maker has no prior knowledge of the user’s exercise behavior at the beginning of each decision epoch. The user is supposed to follow the exercise plan precisely during the current epoch by the planer. This assumption is significantly different from the model presented by Roels (2020) where he asserts that the practice process could be controlled entirely.

In addition, the utility function for the individual in each practice session is considered to have two components: (1) the performance of exercise effort made by the user, and (2) the impact related to the mHealth intervention. According to Roels (2020), the practice performance contains the impacts of every workout on the "stock of fitness" and the "stock of fatigue", which are measured relative to an exercise base level. Besides, both stocks decay exponentially between practice sessions. On the other hand, the formulation of the mHealth intervention service is motivated by Aswani et al. (2019), considering the impact of achieving exercise plans and failing exercise plans. We combine the two studies to establish our performance evaluation model for the mHealth information service.

A subscript $t$ denotes the value of a variable on the $t$-th decision epoch ($t = 0, 1, \ldots, T$). We adopt $s_t = [e_t, b_t, f_t, g_t]$ to represent the state variables of the individual containing all the necessary information for decision making. The $e_t \geq 0$ is a random variable that represents the intensity of the workout in time $t$. The $b_t > 0$ is defined as the base level of past practices. The $f_t, g_t \geq 0$ represent the fitness and the fatigue at time instance, $t$ respectively involving the physical activities performance. We refer to $\chi = (\alpha, \beta, \lambda, \mu, \delta, k_f, k_g)$ as the specific characteristics of the individual. The parameters $\alpha, \beta \in (0, 1)$ represent the decay rates of fitness and fatigue respectively caused by forgetting (for fitness) or recovery (for fatigue), $\lambda, \mu > 0, \lambda \leq 1, \mu \geq 1$ represent the nonlinearities in the response of fitness and fatigue to exercise effect, including the concave effect on fitness and the convex effect on fatigue introduced by Roels (2020), and $0 < \delta \leq 1$ measures the amount of decay in the base level. Another set of the parameters is used in the utility function. These include $k_f, k_g > 0$ which represent the marginal utilities of $f_t$ and $g_t$, $m \geq 0$ which captures the impact of completing an exercise plan, and $l \geq 0$ which represents the marginal disutility of failing an exercise plan. Finally, the decision maker decides the exercise plan $a_t$ among potential exercise plans in time $t$ in the mHealth information service. Using these quantities, we define the following utility functions and dynamics.

(1) Skill acquisition stage (additive model)

$$r_t = r_0 + k_f \tilde{f}_t - k_g \tilde{g}_t + m \cdot \mathbb{I}(e_t \geq a_t) + l \cdot (e_t - a_t)^-,$$
\[ f_t = \alpha f_{t-1} + \left( \frac{a_t}{b_t} \right)^\lambda, \]
\[ g_t = \beta g_{t-1} + \left( \frac{a_t}{b_t} \right)^\mu, \]
where the estimated values of fitness \( f_t \) and fatigue \( g_t \) are updated as:
\[ f_t = \alpha f_{t-1} + \left( \frac{e_t}{b_t} \right)^\lambda, \]
\[ g_t = \beta g_{t-1} + \left( \frac{e_t}{b_t} \right)^\mu. \] (2)

(2) Skill retention stage (multiplicative model)
\[ r_t = r_0 + \frac{(k_f \hat{f}_t)}{(k_g \hat{g}_t)} + m \cdot I(e_t \geq a_t) + l \cdot (e_t - a_t)^-, \]
\[ \hat{f}_t = f_{t-1} \cdot \left( \alpha + \left( \frac{a_t}{b_t} \right)^\lambda \right), \]
\[ \hat{g}_t = g_{t-1} \cdot \left( \beta + \left( \frac{a_t}{b_t} \right)^\mu \right), \] (3)
where the estimated values of fitness \( f_t \) and fatigue \( g_t \) are updated as:
\[ f_t = f_{t-1} \cdot \left( \alpha + \left( \frac{e_t}{b_t} \right)^\lambda \right), \]
\[ g_t = g_{t-1} \cdot \left( \beta + \left( \frac{e_t}{b_t} \right)^\mu \right). \] (4)

Note that in both exercise stages, the planer can predict the practice performance based on the predicted fitness \( \hat{f}_t \) and the predicted fatigue \( \hat{g}_t \) by pushing an exercise plan \( a_t \) for the user. It can avoid overestimating or underestimating the current level of exercise intensity in the assumption. After this decision, the planer will revise the estimated fitness value \( f_t \), the estimated fatigue value \( g_t \), and the estimated exercise base level \( b_t \) for the user according to the actual exercise intensity \( e_t \) collected by the devices and then update existing plans, where \( b_t = \delta \cdot \max(e_{t-1}, b_{t-1}) \). \( I(.) \) is the indicator function.

The purpose of the decision maker is to provide effective mHealth service, improving the individual's utility during the service period. The objective function can be expressed as follows:
\[ \max_{a_t} \sum_{t=1}^{T} r_t \] (5)

4. Dynamic mHealth information service system
We describe the personalized mHealth information service as a dynamic system communication by wearable trackers, where the sequential decision for the practice plans is considered as a Markov Decision Process (MDP). At the beginning of each decision epoch, the mHealth service monitors the
user exercise behavior through wearable devices, then chooses an exercise plan among potential exercise plans based on the calculation of exercise intensity and the evaluation of the user’s utility as a reward. After this decision, the user makes effort to implement the exercise plan, then the individual’s health state is transited according to the underlying Markov chain. The decision maker will then receive the signal of the user’s utility and an observation for the individual actual exercise intensity. The interaction process is described in figure 2.

![Diagram of the dynamic mHealth information service system](image)

**Fig.2:** The process of the dynamic mHealth information service system

Compared with Roels (2020), we focus on the ongoing personalized practice plans decision process supported by real-time communication between the planner and the executor. Specifically, this study has improved an individual’s health state from three aspects. First, the individual’s changing time dependent preference may impact the exercise behavior and thereby influences the mHealth recommendation policy. Second, we observe that the mHealth service provider only can make decisions on exercise plans rather than control the realized behavior of the individual in reality. Third, we notice that the same intensity activity may cause varying levels of pressure and different challenges for different people. It is necessary to consider the individual's motivation effect for the exercise plans when evaluating the impact of mHealth intervention service. To achieve these targets, we train a learning agent who can sense the individual's states through the wearable trackers and take actions adaptively to guide the exercise with individual exercise preference continuing to change. The wearable trackers can observe and record the individual's exercise state in time. The objective is to learn the optimal policy rule to maximise the individual's practice performance as well as the mHealth service provider’s recommendation effect in the long term. The body of sports healthcare literature has
published different quantifying models for various sports to analyze the individual's physiological data. For example, the APP named STRAVA published a feature that quantifies effort during riding using heart rate data. Daniels (2014) created scientific evaluation criteria for running.

4.1 Markov decision process

Due to the gaps between the health plan and the individual physical exercise behavior, we describe this mHealth optimization problem as a sequential decision process in an uncertain environment using the MDP model. Our goal is to maximise the accumulated practice performance of the individual during the mHealth service. Mathematically, it can be expressed as a discrete online constraint optimization problem, subject to various quantitative and/or qualitative constraints. At each decision period, the MDP can take the amount of actual daily exercise and the time-dependent estimate values of the fitness, the fatigue as inputs and then output the next activity plan to induce and encourage him to exercise. The individual receives the exercise plan from the mHealth service delivered by the wearable device and then makes efforts to do it, resulting in an immediate reward and a stochastic transition to a new state observed by the sensors. For model tractability, we assume the wearable trackers could measure the individual's exercise intensity accurately. The estimated values related to practice history can be calculated according to the method provided by Roels (2020). The motivation effect of mHealth intervention is inspired by Aswani et al. (2019). Next, we formulate the various MDP elements, which leads us to present the optimality recommendation policies expression.

(1) Decision epochs: The decision epochs in our problem are denoted by \( t = 0, 1, \ldots, T \), where \( T \) represents the length of service. Practice recommendations are considered to be delivered periodically by the mHealth service in time \( t \).

(2) State space: The state at period \( t \), denoted by \( s_t \in S \), where \( s_t = [e_t, b_t, f_t, g_t] \). As mentioned in Section 3.1, the state \( s_t \) contains the individual's exercise status and the accumulated impacts on practice performance.

The individual's exercise state at each period is denoted by \( e_t \in \mathcal{E} \), where \( \mathcal{E} \) contains discrete exercise intensity levels, which is determined by the precision of wearable devices. The precision of mHealth system in our experiment is 0.01, so \( \mathcal{E} = \{ E, E + 0.01, E + 0.02, \ldots, \bar{E} \} \). \( E \) and \( \bar{E} \) are the minimum and the maximum exercise intensity levels recorded by the wearable devices, respectively.

For the individual's accumulated impacts on practice performance, the impact of the previous exercise is included as part of the state definition because the previous exercise often has long-term, lasting effects. Therefore, it must be incorporated into the state definition to retain the Markov property. The individual's state at period \( t \) will be changed following his/her exercise \( e_t \) in time \( t \). We take the exogenous information into account to estimate the accumulated impacts on practice performance.
using the stochastic exercise state $e_t$ and the previous state $s_{t-1}$, where the accumulated impacts on practice performance contain three parts: the exercise base level $b_t > 0$; the stock of fitness $f_t \geq 0$, and the stock of fatigue $g_t \geq 0$. Besides, the exogenous information is related to the service content, which contains the information of sports events and the individual's sports skills. Hence, the individual's exercise state can be observed by wearable devices while the individual's accumulated impacts on practice performance will be estimated according to equation (2) or equation (4) for the skill acquisition stage or the skill retention stage respectively.

(3) Action space: At each decision epoch $t$, the mHealth service decides the exercise plan with intensity $a_t \in \mathcal{A}$, assumed to be unidimensional, where $\mathcal{A} = \{0\} \cup \{A_1, A_2, \ldots, A_{n-1}\}$ contains no exercise plan, and discrete exercise intensity.

(4) Reward functions: The immediate reward function $r_t(s_t, a_t)$ captures the one-period mHealth service performance and the mHealth intervention effect on utility when the individual takes the exercise intensity level $e_t$ after receiving the mHealth recommendation $a_t$ in the state $s_t$. For different practice programs, we define the reward functions according to the equation (1) and equation (3), as follows:

The reward function in the sports acquisition stage can be expressed as:

$$r_t(s_t, a_t) = r_0 + k_f \left( \alpha f_{t-1} + \left( \frac{a_t}{b_t} \right)^3 \right) - k_g \left( \beta g_{t-1} + \left( \frac{a_t}{b_t} \right)^5 \right) + m \cdot \mathbb{I}(e_t \geq a_t) + l \cdot \min(0, e_t - a_t).$$

The reward function in the sports retention stage can be expressed as:

$$r_t(s_t, a_t) = r_0 + \frac{k_f f_{t-1} \left( \alpha + \left( \frac{a_t}{b_t} \right)^3 \right)}{k_g g_{t-1} \left( \beta + \left( \frac{a_t}{b_t} \right)^5 \right)} + m \cdot \mathbb{I}(e_t \geq a_t) + l \cdot \min(0, e_t - a_t).$$

4.2 The optimality recommendation policies

The goal of the mHealth service is to learn an optimal adaptive decision rule to improve the mHealth service performance. The policy $\pi$ is a computable function that outputs for each state $s_t \in S$ an action $a_t \in \mathcal{A}$. Therefore, a policy $\pi$ is a distribution over actions given states:

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

We can now formulate the value function for the practice optimization problem by putting all the above pieces together. Let $v_\pi(s_t)$ be the total expected value for the individual from a given state $s_t$ following policy $\pi$ in period $t$. Then, we decompose the state-value function $v_\pi(s_t)$ into immediate reward plus the successor state value. The Bellman expectation equations can be written as:

$$v_\pi(s_t) = \mathbb{E}_\pi[r_t + v_\pi(s_{t+1})|s_t], \forall s_t \in S.$$
We denote the action-value function $q_\pi(s_t, a_t)$ as the expected value for the individual from a given state $s_t$, taking action $a_t$, and then follows policy $\pi$ in period $t$. The Bellman expectation equations can be written as:

$$q_\pi(s_t, a_t) = \mathbb{E}_t[r_t + q_\pi(s_{t+1}, a_{t+1})|s_t, a_t], \forall s_t, a_t \in \mathcal{A}. \tag{10}$$

Solving the personalized mHealth optimization problem need to find optimal policies $\pi^*$ that achieves maximum reward during the service. Therefore, there exists an optimal policy $\pi^* \geq \pi, \forall \pi$, where

$$\pi^* = \text{argmax}_\pi v_\pi(s_t), \forall s_t \in \mathcal{S}. \tag{11}$$

The optimal policies $\pi^*$ share the same state-value function and action-value function, denoted the optimal state-value function $v^*$ and the optimal action-value function $q^*$ respectively, and defined as:

$$v^*(s_t) = \max_\pi v_\pi(s_t), \forall s_t \in \mathcal{S},$$

$$q^*(s_t, a) = \max_\pi q_\pi(s_t, a), \forall s_t \in \mathcal{S}, a_t \in \mathcal{A}. \tag{12}$$

This function gives the expected return for taking action $a_t$ in the state $s_t$ and thereafter following an optimal policy for the state-action pair $(s_t, a_t)$. Thus, we obtain:

$$q^*(s_t, a_t) = \mathbb{E}[r_t + v^*(s_{t+1})|s_t, a_t]. \tag{13}$$

We backup the spans of future states and actions from state-action pair $(s, a)$ based on the Bellman optimality equation, the state-action value function denotes as $q^*(s, a)$, can be formulated as follows:

$$q^*(s_t, a_t) = \mathbb{E}[r_{t+1} + \max_a q^*(s_{t+1}, a')|s_t, a_t], \tag{14}$$

where $a'$ represents the successor action of action $a_t$.

The optimal policy $\pi^*$ can be obtained through the Bellman optimality equation (9) when all information about the individual exercise behavior is known to the mHealth service. However, the individual exercise behavior is dynamic and uncertain, as well as the state space of the MDP increases exponentially to the number of components in the personalized mHealth optimization problem. We learn the individual exercise behavior from experience based on sampling and design the intelligent algorithm to solve the personalized mHealth optimization problem.

### 5. Realisation of DRL-based recommender system

In the personalised mHealth optimization problem, no assumption is made on the individual exercise behavior considering the unexpected effects of external influences. Besides, the individual's state space and the mHealth alternative recommendation action spaces are discrete and quite large, requiring a lot of computation time and storage space to solve this practice plans optimization problem.
Furthermore, the individual's exercise preference often changes randomly according to the person's state of mind and mood. Hence, the resulting optimization problem cannot be readily resolved by the traditional methods, such as dynamic programming, genetic algorithm that with a time independent fitness (i.e. objective) function.

As an alternative, deep reinforcement learning (DRL) is a promising tool to overcome problems in a computationally efficient manner. The DRL methodology can learn from the experience rather than rely on the prior knowledge of the individual's changing behavior. This method samples sequences of states, actions, and rewards from interaction with an actual environment. In addition, we improve the DRL methodology by constructing the hybrid neural network, which can capture temporal information from the individual's exercise history by the LSTM neural network and improve the fitting effects by the fully connected neural network.

Inspired by the asynchronous advantage actor-critic reinforcement learning (Mnih et al., 2016), the stochastic decision policy in this study can be trained by combing the generalization method (actor network) and parameterized policy method (critic network). The two major techniques, including asynchronous training agents and advantage function in policy gradient, are used to train our hybrid neural network in the DRL algorithm. The main structure of one of the training agents is illustrated in figure 3. There are two separate modules in the training agent, namely "actor" and "critic", which control the action selection and criticize the actor's actions, respectively. The influence of the practice recommendations tends to be nonlinear and time-dependent in the individual's behavior system, and therefore difficult to predict. We propose a novel neural network model for learning the value function for the current policy \( \pi \) in period \( t \). The actor network with weights \( \theta \) and the critic network with weights \( \phi \) 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 \( v(s_t;\phi) \), respectively. In addition, the advantage function is introduced to assess how much better it is to take action than the average, general action at the state \( s_t \).
There are two main essential elements that need to be elaborated in our DRL methodology. The first one is the structure of the critic network and the actor network, which need to infer the individual’s exercise behavior effectively. The second one is the computation processing of our DRL methodology, which can learn recommendation policy automatically.

5.1 The networks architecture

The training processes of the critic network and actor network in our DRL methodology are shown in figure 3. The parameters $\varphi$ of the critic network can be trained through gradient descent to minimize the mean square error. Then, the "Actor" updates the policy distribution in the direction suggested by the "Critic". The complete architecture network aims to realise the dual evolution: exploitation and exploration principle cum concept embedded in the Deep Reinforcement Learning cycle.

(1) The critic network: The critic network model is a hybrid forecaster by an LSTM layer and a deep neural network with multiple fully connected hidden layers. The main structure of the critic network is presented in figure 4. The critic network consists of an input layer where the input data are fed into one LSTM layer. In our personalised mHealth optimisation problem, the input data is a three-dimensional vector: state characteristics, the stack of the previous state, and time. The state characteristics are the individual's exercise intensity $e_t$, the practice base level $b_t$, the fitness level $f_t$, and the fatigue level $g_t$ in period $t$. The stack of the previous state is constructed because the reward
relies on the current state and the previous state directly. The time refers to the discrete-time, which equals 1. The LSTM has feedback connections that can learn and model the sequential relations in the time series data. Then, we construct a deep neural network connecting with the LSTM, which can learn from the previous layer and develop new representations for the action-value function approximation. Our critic network has $l$ hidden layers, where the first hidden layer is with the LSTM cells, and the others are fully connected layers. The number of neural cells in the hidden layer $i$ is $y_i$, $i = 1, \ldots, l$. The $h_{k(t)}$ represent the outputs of the LSTM cell in time $t$, where $k = 1, \ldots, y_1$, and $C(t)$ represents the long-term memory, where $t = 0, 1, \ldots, T$. Besides, the outputs of neural cells in the hidden layer $i$ are denoted as $h^i_k$, where $k = 1, \ldots, y_i$, $i = 2, \ldots, l$. The parameters are used in this critic network, including the weight parameters of the neural network $\phi$, and the bias parameters $\xi$. The $\sigma(.)$ represent the activation functions in the critic network.

![LSTM Network Diagram](image)

**Fig. 4:** The structure of the critic network of the DRL methodology

An illustration of the LSTM cell is given in figure 5, where the black circles represent pointwise operations while the grey boxes are learned neural network layers. The LSTM can remove or add information to the cell state by a hidden state $h_{i(t)}$ to store the short-term memory and a cell state $C(t)$ to store the long-term memory in time $t$. Specifically, the LSTM cell is regulated by three gates. The forgot gate $f$ controls the range to which a value remains in the LSTM cell. The input gate $i$ controls...
the recording of the latest information in the individual's state. The output gate \( o \) objectively selects information to compute the output activation of the LSTM unit.

For instance, the calculation process of the critic network in time \( t \) is defined as follows. The architecture of the critic network consists of one LSTM layer and \( l - 1 \) fully connected layers.

In the LSTM cell \( k \) \((k = 1, 2, \ldots, y_1)\), the first step is to transform the three-dimensional input data into a matrix \( X_t \). Then, decide what information will be thrown away from the cell state controlled by the forgot gate \( j_t \), taking the previous hidden state \( h_{k,(t-1)} \) and the current information \( X_t \) as input and generating a number between 0 and 1 for each number in the cell state \( C_{k,(t-1)} \) as output by the sigmoid activation function \( \sigma_{\text{Sigmoid}}(\cdot) \).

\[
j_t = \sigma_{\text{sigmoid}}(\varphi_j \cdot [h_{k,(t-1)}, X_t] + \xi_j).
\] (15)

Fig.5: Illustration of the LSTM cell in the critic network

The next step is to decide what new information will be stored in the cell state controlled by the input gate \( i_t \). In particular, a neural network layer \( i_t \) with the sigmoid activation function decides what values will be updated in the cell state. Then, a neural network layer with a hyperbolic tangent activation function \( \sigma_{\text{tanh}}(\cdot) \) generates a vector of new candidate values \( \tilde{C}_{k,(t)} \), which could be added to the cell state \( C_{k,(t)} \). Finally, the old cell state \( C_{k,(t-1)} \) will be updated into the new cell state \( C_{k,(t)} \) by multiplying \( C_{k,(t-1)} \) by \( j_t \), forgetting the information predetermined by the forgot gate, and then adding new information with the candidate values \( \tilde{C}_{k,(t)} \) scaling by \( i_t \).

\[
i_t = \sigma_{\text{sigmoid}}(\varphi_i \cdot [h_{k,(t-1)}, X_t] + \xi_i),
\]

\[
\tilde{C}_{k,(t)} = \sigma(\varphi_c \cdot [h_{k,(t-1)}, X_t] + \xi_c)_{\text{tanh}}
\]

\[
C_{k,(t-1)} = j_t \times C_{k,(t-1)} + i_t \times \tilde{C}_{k,(t)}.
\] (16)
After that, a neural network layer $o_t$ with the sigmoid activation function decides what parts of the LSTM cell will be output. Then, the cell state $c_{k,(t)}$ can be normalised between -1 and 1 through hyperbolic tangent function and multiply it by the $o_t$ to obtain the output $h_{k,(t)}$ of the LSTM cell.

$$o_t = \sigma_{\text{sigmoid}}(\phi_o[h_{k,(t-1)}, X_t] + \xi_o),$$

$$h_{k,(t)} = o_t \times \tanh(c_{k,(t)}).$$ (17)

Then, the outputs from the LSTM layer $h_{k,(t)}, k = 1, \ldots, y_1$ in time $t$ are connected to a fully connected network of the hidden layers. Mathematically, the output of the neural cell $k$ in the layer $i$ is given by the following equation:

$$h^i_k = \sigma_{\text{ReLU}}\left(\sum_{j=1}^{y_1} \varphi_{h_j} h_{j,(t)} + \xi_{h^i_k}\right), k = 1, 2, \ldots, y_2,$$

$$h^i_k = \sigma_{\text{ReLU}}\left(\sum_{j=1}^{y_1} \varphi_{h_j} h_{j}^{i-1} + \xi_{h^i_k}\right), k = 1, 2, \ldots, y_i, i = 3, \ldots, l. \quad (18)$$

Finally, the critic network produces the output of the estimated value of the value function:

$$v_{\pi}(s_t; \varphi) = \sum_{j=1}^{y_l} \varphi_{h_j} h_{j}^{i} + \xi_{h^i_j}. \quad (19)$$

The activation function $\sigma(.)$ includes sigmoid function $\sigma_{\text{sigmoid}}$, hyperbolic tangent function $\sigma_{\text{tanh}}$, and ReLU function $\sigma_{\text{ReLU}}$. The activation function can be formulated as follows:

$$\sigma_{\text{sigmoid}}(x) = \frac{1}{1 + e^{-x}},$$

$$\sigma_{\text{tanh}}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

$$\sigma_{\text{ReLU}}(x) = \max(x, 0). \quad (20)$$

(2) The actor network: In our DRL methodology, an actor will learn policy to select an action in a given state $s_t$. The actor network is constructed to learn the parameterized policy $\pi(a_t|s_t; \theta) = P[a_t|s_t; \theta]$. The actor network is similar to the critic network combing an LSTM layer and a deep neural network with fully connected layers.

The actor network has the same architecture as the critic network with the weight parameters of the neural network $\theta$. Because the daily personal practice behavior may be subject to random disturbance from external, the stochastic policy is better than the determined policy in personalized practice optimization problem. The only difference between the actor network and critic network is computing in the output layer. We introduce the Softmax function as the activate function in the neural cells in the output layer.

$$\pi_t(a^i|s_t; \theta) = \exp(h^i_t)/\exp(\sum_{j=1}^{n} h^j_t), i = 1, \ldots, n, \quad (21)$$
where \( n \) represents the number of action space \( A \).

### 5.2 Learning process of the DRL methodology

Since the DRL methodology has two networks, including the critic network and the actor network. The critic network is constructed to estimate the state-value function \( v(s_t; \phi) \) and \( v(s_{t+1}; \phi) \) for current policy \( \pi \), which can be trained by the mean square of the advantage function. The advantage function considering the state-value function \( v_\pi(s_t) \) as a baseline to evaluate an action for current policy \( \pi \) is given as follows:

\[
A_\pi(s_t, a_t) = q_\pi(s_t, a_t) - v_\pi(s_t)
\]

We adopt the sampling method according to the equation (3.2.8) to evaluate the action-value \( q_\pi(s_t, a_t) \). Therefore, the advantage function can be express as follows:

\[
A_\pi(s_t, a_t) = r_t + v_\pi(s_{t+1}) - v_\pi(s_t) \approx r_t + v(s_{t+1}; \phi) - v_\pi(s_t; \phi)
\]

Therefore, the accumulated gradient of the weight \( \phi \) of the critic network is obtained as follows.

\[
d\phi \leftarrow d\phi + \frac{\partial (A_\pi(s_t, a_t))^2}{\partial \phi} \phi
\]

On the other hand, the actor updates the policy distribution in the direction suggested by the critic. According to the policy gradient theorem (Sutton & Barto, 2018), using the sampling method to instantiate the generic stochastic gradient ascent algorithm yields the accumulated gradient of the parameters \( \theta \).

\[
d\theta \leftarrow d\theta + \nabla_\theta \log \pi (a_t | s_t; \theta) A_\pi(s_t, a_t)
\]

We use the forward view of \( n \)-step method and asynchronous technique to train the networks. The forward view of \( n \)-step method could improve convergence speed in the DRL methodology. The asynchronous approach is to adopt multiple independent agents with their own weights to interact with a different copy of the environment in parallel, which could explore a more significant part of the state-action space in much less time. In the asynchronous actor-critic algorithm, the agents are trained in parallel, and update periodically a global network, which holds shared parameters. After each update, the agents reset their parameters to those of the global network and continue their independent exploration and training for \( n \) steps until they update themselves again. The pseudo-code of the proposed asynchronous advantage actor-critic algorithm is described below.

**Algorithm** Asynchronous advantage actor-critic (A3C) for the personalised mHealth problem

**Input:** The architecture of the global actor network with the weights \( \theta \); The architecture of the global critic network with the weights \( \phi \); The architecture of actor networks with the weights \( \theta^i \) for thread \( i \), where \( i = 1, \ldots, N \); The architecture of critic networks with the weights \( \phi^i \) for thread \( i \), where \( i = 1, \ldots, N \);
Action set $\mathcal{A}$;

**Parameters:** Upper bound of global shared step counter $T_{\text{max}}$; Learning rate $\nu > 0$, $\omega > 0$; Thread step counter $n$ in each thread.

1. Initialize global shared step counter $T \leftarrow 0$
2. Repeat
3. For each thread $i \leftarrow 1$ to $N$ do
4. Initialize thread step counter $t \leftarrow 1$
5. Reset accumulate gradients: $d\theta \leftarrow 0$, $d\phi \leftarrow 0$
6. Synchronize thread-specific parameters: $\theta^i \leftarrow \theta$, $\phi^i \leftarrow \phi$
7. $t_{\text{start}} \leftarrow t$
8. Initialize state $s_t$
9. Repeat
10. Take action $a_t$ according to policy $\pi(a_t|s_t; \theta^i)$
11. Compute reward $r_t$ using equation (6) or equation (7)
12. Observe new state $s_{t+1}$
13. $t \leftarrow t + 1$, $T \leftarrow T + 1$
14. Until $s_t$ is terminal or $t - t_{\text{start}} == n$
15. $R = \begin{cases} 0 & \text{for terminal } s_t \\ v(s_t; \phi^i) & \text{for non-terminal } s_t \end{cases}$ (Bootstrap form last state)
16. For $j \leftarrow t - 1$ to $t_{\text{start}}$ do
17. $R = r_j + R$
18. Accumulate gradients wrt $\theta^i$: $d\theta \leftarrow d\theta + \nabla_{\theta^i} \log \pi(a_j|s_j; \theta^i)(R - v(s_j; \phi^i))$
19. Accumulate gradients wrt $\phi^i$: $d\phi \leftarrow d\phi + \partial \left( (R - v(s_j; \phi^i))^2 / \partial \phi^i \right)$
20. End for
21. Perform asynchronous update $\theta, \phi$: $\theta \leftarrow \theta - \nu d\theta$, $\phi \leftarrow \phi - \omega d\phi$
22. End for
23. Until $T > T_{\text{max}}$

### 6. Simulation of exercise program

For our simulation, we use five years of training data from a runner (Cannon et al., 2018). Then, we estimate the parameters of the personalized practice optimization model on running data. We conclude by demonstrating the ability of our DRL methodology to make decisions on the personalized mHealth promotion programs by competition experiments and sensitivity analysis.

#### 6.1 Data preprocessing

We use the individual training dataset from a runner over a five year period from 2002-2006. The data with 1128 observations is freely available for the textbook STAT2: Modeling with Regression and ANOVA (second edition) downloaded from [https://rdrr.io/cran/Stat2Data/man/Marathon.html](https://rdrr.io/cran/Stat2Data/man/Marathon.html).
The professional knowledge of running is applied in running data preprocessing inspired by Roels (2020). We estimate the runner's exercise intensity by Training IMPulses (TRIMPs) provided by Daniels (2014). Each practice session is characterized as easy running (E), marathon-pace running (M), threshold running (T), interval training (I), and repetition training (R). The recording points associated with various intensities of training can be calculated by the following rough estimates: 0.2 points/min for E running, 0.4 points/min at M pace, 0.6 points/min at T pace, 1.0 points/min at I pace, and 1.5 points/min at R pace (Daniels, 2014). The speeds of different types of training can be estimated using the VDOT training tables, where the VDOT value is based on the runner's most recent marathon time using Table 5.1 in Daniels (2014). The records at a Marathon distance of 26.2 miles are observed on Days 225, 359, 615, 745, and 1107, corresponding to VDOT of 48, 41, 39, 46, and 42, respectively. We divide the training data into five years with the benchmark paces (minutes/mile) corresponding to VDOT values, as shown in Table 1. Hence, we account for both training load and intensity. The practice process is displayed in figure 6, practice intensity varies from 0 to 122.5 TRIMPs.

Table 1: The benchmark paces (minutes/mile) to VDOT

| VDOT | E     | M     | T     | I     | R     | Period          |
|------|-------|-------|-------|-------|-------|-----------------|
| 48   | 8:13 - 9:15 | 7:32  | 7:02  | 6:25  | 5:56  | Day 1 - Day 365 |
| 41   | 9:21 – 10:28 | 8:39  | 8:02  | 7:20  | 6:48  | Day 366 - Day 730 |
| 39   | 9:44 – 10:53 | 9:02  | 8:22  | 7:36  | 7:04  | Day 731 - Day 1096 |
| 46   | 8:31 – 9:34  | 7:49  | 7:17  | 6:40  | 6:08  | Day 1097 - Day 1461 |
| 42   | 9:10 – 10:17 | 8:28  | 7:52  | 7:12  | 6:40  | Day 1462 - Day 1825 |

| Point (points/min) | 0.2 | 0.4 | 0.6 | 1 | 1.5 |

Data source: Table 5.2 in Daniels (2014)

Fig.6: Practice intensity of the runner
The running performance without the mHealth intervention can be measured on the raw data by using the runner's VO$_2$max, which is the maximum oxygen consumption (in milliliters) per minute per kilogram of body weight. As Roels (2020) states: 'the VO$_2$max is considered as a good, though imperfect, predictor of performance'. Typically, the VO$_2$max is collected by the runner's wearable devices at the end of practice sessions. We use the VDOT to replace the VO$_2$max as the predictor of performance due to the lack of available data. Daniels (2014) has established the VDOT running calculator gathered from years of testing many runners of various ability levels related to the VO$_2$max. The performance time series is displayed in figure 7.

![Time series of performance](image)

**Fig.7**: Practice performance of the runner

### 6.2 Model parameterization

The individual-specific parameters can be estimated using the processed training data with practice intensity and practice performance while the runner practiced without the mHealth intervention. We split these time series data for in-sample and out-of-sample as 80%:20% for estimation and prediction, respectively. Using the estimation data set, we fitted the two models outlined in Section 3.1 (where is it!) by optimising their parameters without any consideration of mHealth intervention effects. We use the nonlinear least-squares method in Python to minimize the residual sum between the estimated practice performance by the model with the practice intensity data and the observed practice performance data. The estimated parameters of individual characteristics on the running data are reported in Table 2.
Table 2: The estimated parameters of practice optimization models without mHealth intervention

| Model                           | Acquisition stage | Retention stage |
|---------------------------------|-------------------|-----------------|
| In-sample RMSE                  | 5.88              | 3.188           |
| Out-of-sample RMSE              | 7.77              | 2.42            |
| Initial value of base level     | $b_0$             | 1.5             | 2.5             |
| Decay rate in base level        | $\delta$         | 0.9             | 0.5             |
| Initial value of practice       | $r_0$             | 40.23           | 38.7            |
| performance                     |                   |                 |                 |
| Marginal utility rate of fitness| $k_f$             | 1.4             | 50              |
| Marginal utility rate of fatigue| $k_g$             | 0.5             | 56              |
| Decay rate of fitness           | $\alpha$         | 0.895           | 0.1             |
| Nonlinearities effect of fitness| $\lambda$        | 0.9             | 0.995           |
| Decay rate of fatigue           | $\beta$          | 0.895           | 0.1             |
| Nonlinearities effect of fatigue| $\mu$            | 1.25            | 1.045           |

6.3 Learning performance of the AI module on the mHealth information service

To evaluate the performance, we test our DRL methodology using the test data, which are randomly generated three-month running data from the preprocessed real data. The mHealth service prepares ten alternative exercise plans on different exercise intensity levels (TRIMPs) of 1.8, 7, 12.2, 17.4, 22.6, 27.8, 33, 38.2, 43.4, 48.6 according to the histogram of the dataset to push to the user or not send an exercise plan. The DRL methodology aims to learn the optimal policy to push the exercise plan for the individual practice process problem. Firstly, the agent is trained to learn the environment from the time series training data. Then, we compare our adaptive policy with the fixed exercise plans and without mHealth intervention.

All simulations are run in Python using TensorFlow and run on a PC with the Intel (R) Core (TM) i7 CPU with 32.00 GB of RAM. Each simulation runs 100 times for a statistical purpose.

For the proposed DRL methodology, the number of hidden layers and the number of neurons in the layer of the critic network and actor network are determined by experience and experiments because there is no more mature theory available for selecting the parameters of the hidden layers. On the other hand, the parameters in the asynchronous advantage actor-critic algorithm impact the speed of convergence and the performance of learning, which are adjusted by a set of experiments. We test the learning performance of the proposed DRL model with the number of neurons in the hidden layer, and determined the parameters of the model. The parameters of the proposed DRL methodology are given in Table 3.
Table 3: The parameters of the proposed DRL methodology

| Parameters                                      | value |
|------------------------------------------------|-------|
| The number of units in the 1st LSTM hidden layer | $y_1$ 64 |
| The number of neurons in the 2nd fully connected hidden layer | $y_2$ 32 |
| Upper bound of global shared step counter       | $T_{max}$ 25000 |
| Thread step counter                             | $n$ 5 |
| Learning rate                                   | $\nu$ 0.0007 |
| Learning rate                                   | $\omega$ 0.0007 |

We observe that each simulation converges in 5100 steps, and take less than 355 seconds. Figure 8 shows the performance of our algorithm.

![Fig.8: Performance of the proposed DRL methodology on the training data](image)

As shown in figure 9, our DRL policy compares with the current mHealth recommendation policies for the fixed exercise plans and without exercise plans. The fixed exercise plans on different TRIMPs values of 1.8, 7, 12.2, 17.4, 22.6, 27.8, 33, 38.2, 43.4, 48.6. The results indicate that our DRL policy is superior to the other fixed plans and without mHealth service on average. However, fixed plans with low exercise intensity perform better than our DRL policy with some probability. This result may be explained by the fact that the individual is physically inactive in some special situations, where the exercise plans with low intensity are more likely to be completed. In addition, it may cause the individual to be discouraged when the individual challenges and difficulties to complete the high-intensity exercise plans.
6.4 Benchmarking test

We compare our DRL methodology with a wide variety of algorithms. Specifically, we compare our methodology with 6 other algorithms shown in Table 4. Below is a summary of the competitors from each category. Each simulation runs 100 times for a statistical purpose.

**Table 4: Competitors**

| Category                                                                 | Competitor         |
|-------------------------------------------------------------------------|--------------------|
| Our A3C algorithm on different network architectures                    | A3C-MLPs           |
|                                                                         | A3C-LSTM           |
| Our hybrid network architecture but different reinforcement learning algorithm | DQN-hybrid         |
|                                                                         | ACKTR-hybrid       |
|                                                                         | PPO-hybrid         |
|                                                                         | GAIL-hybrid        |

**Our Asynchronous advantage actor-critic algorithm on different network architectures.** The approaches in this category use the same A3C reinforcement learning algorithm used by our model but employ a different network architecture. Specifically, we compare with three different models using some of the most popular network architectures for deep reinforcement learning methodology (Mousavi Seyed Sajad and Schukat, 2018), namely multi-layer perceptrons (MLPs), long short term memory (LSTM) recurrent neural networks. The number of neurons in the hidden layers of the A3C-MLPs, A3C-LSTM are determined by empirical experiments: the A3C-MLPs has two hidden layers with 64 neurons, the A3C-LSTM has a hidden layer with 32 neurons.
Our hybrid network architecture but different reinforcement learning algorithm. In this category, all reinforcement learning approaches use exactly the same hybrid network architecture as used in our approach. However, instead of using the A3C reinforcement learning algorithm, these models use different learning algorithms namely DQN, ACKTR, PPO, and GAIL (Chen et al., 2021).

Table 5 shows the results for all the algorithms. It can be observed that our DRL methodology consistently outperforms all competitors in terms of computing complexity and average speed. Below, we provide a more detailed comparison of DRL methodology with the competitors.

Consider using the A3C learning method on different network architectures (MLP, and LSTM). Our DRL methodology outperforms these algorithms significantly in averaging speed. This shows that the hybrid neural network architecture is useful for learning the user’s exercise behavior in the personalized mHealth service system.

Next, we explain the results for the models that use the same network architecture as our DRL methodology but different reinforcement learning algorithms (DQN, ACKTR, PPO, and GAIL). Our DRL methodology outperforms all these models. This shows the effectiveness of the A3C algorithm used to train our hybrid network, and find the optimization recommendations for the mHealth information service.

**Table 5: Results of the proposed DRL methodology and the competitor algorithms**

| Algorithms  | Model                      | Average running time (seconds) | Average covering steps |
|-------------|----------------------------|--------------------------------|------------------------|
| **DRL methodology** | Skill acquisition model | 351.09                          | 4500                   |
|             | Skill retention model     | 367.46                          | 5100                   |
| **A3C-MLP** | Skill acquisition model   | 426.71                          | 5400                   |
|             | Skill retention model     | 410.38                          | 5340                   |
| **A3C-LSTM**| Skill acquisition model   | 431.08                          | 4960                   |
|             | Skill retention model     | 405.91                          | 5250                   |
| **DQN-hybrid**| Skill acquisition model  | 422.13                          | 9600                   |
|             | Skill retention model     | 449.12                          | 10000                  |
| **ACKTR-hybrid**| Skill acquisition model | 361.61                          | 6800                   |
|             | Skill retention model     | 374.27                          | 6100                   |
| **PPO-hybrid**| Skill acquisition model  | 349.56                          | 6400                   |
|             | Skill retention model     | 348.62                          | 6260                   |
| **GAIL-hybrid**| Skill acquisition model  | 368.43                          | 6100                   |
|             | Skill retention model     | 380.19                          | 5940                   |
6.5 Sensitivity analysis

The parameters related to the mHealth intervention service include the impacts of completing the exercise plan $m$ and the marginal disutility of failing the exercise plan $l$. These exogenous factors can be assessed through a scale or behavioral experiments. Because of the conditions limit, we analyze the sensitivity of the parameters of the effect of mHealth intervention service.

Figure 10 shows that increasing the impacts of completing the exercise plan $m$ will increase the average performance of the individual's utility and may cause a more significant difference in the individual's utility during the service period for both the skill acquisition and retention stages. Besides, only when the motive for completing the exercise plan is beyond a certain degree, could the performance of the individual who uses mHealth information service is better than that who doesn't use service. It is turning now to the experimental evidence on the sensitivity analysis of the marginal disutility of failing the exercise plan $l$ in figure 11. The average performance of the individual's utility decrease with $l$. The most surprising aspect of the result is that the individual's utility declines sightly or even increases with $l$ in low values. A possible explanation for this might be that the mHealth service has no motivation to improve the quality of the practice recommendation service when the user's perceived loss value from failing to achieve the exercise plans is very low.

![Fig.10: Total rewards for different values of $m$ ($l = 2$)](image-url)
7. Conclusion and recommendation of direction for further research

Mobile health information service is a new tread in self-health management, especially during the COVID-19 pandemic. In the mHealth information service, the recommender system record and visualize the individual health conditions through wearable trackers for the individual, then deliver a set of activity plans for daily exercise with motivational messages. This dynamic artificial intelligence system can adapt to the user’s heterogeneous characteristics, ongoing exercise preferences, and uncertain health outcomes during the service. The adaptiveness, learning capability of the health management system have been improved than the current health system, which pushes the fixed exercise recommendations to the user. In particular, this study presents the performance evaluation for mHealth service based on training theory in sports science. Given the parameters in the performance evaluation model, the DRL methodology can be tailored to fit individual characteristics. Moreover, the interaction between the mHealth service provider and the individual through wearable devices enable the sequential evolution of recommender policy using observed information from wearable devices.

These online adaptations of the agent increase the service provider's awareness of the individual exercise behavior. Through better awareness, the service provider can push better exercise recommendations to the users and increase their service satisfaction. For example, exercise equipment and media companies such as Peloton could provide customized bike Bootcamp services for the digital membership through a wireless device installed in Pelton bike (record the member exercise progress). The customized bike Bootcamp classes could potentially be scheduled using the proposed approach to improve the users' utility and thus increase users' dependency and engagement. Hence, this study can contribute to service operation for healthcare. Although, the artificial intelligence system may cost the fixed investment at the beginning. It will improve the user’s satisfaction during the healthcare service, improve customer loyalty. Over the long term, it will be valuable to update the system for the mHealth information service.
A limitation of this study is the lack of measurement of parameter values for mHealth interventions, including the impacts of completing the exercise plan and the marginal disutility of failing the exercise plan. Notwithstanding this limitation, the study suggests that the incentives for completing the exercise plans can improve the expected average performance of mHealth service. Meanwhile, low incentives may cause the users' satisfaction less than those who don't use mHealth service. On the other hand, the study suggests that low punishment for failing the exercise plan does not improve the individual's satisfaction during the service. Therefore, the service provider needs to use appropriate rewards and punishments to encourage the individual to participate in physical activity. An extension of the current study is to employ the behavioral experiments to design the optimal setting of rewards and punishments. Another extension of the study is to combine more characteristics of the individual into the recommender system, i.e. sleeping monitor data.

References

Adjerid, I., Loewenstein, G., & Purta, R. (2021). Gain-Loss Incentives and Physical Activity: The Role of Choice and Wearable Health Tools. *Management Science, 1909*, 1–26.

Afzal, M., Ali, S. I., Ali, R., Hussain, M., Ali, T., Khan, W. A., Amin, M. B., Kang, B. H., & Lee, S. (2018). Personalization of wellness recommendations using contextual interpretation. *Expert Systems with Applications, 96*, 506–521.

Akay, M. F., Inan, C., Bradshaw, D. I., & George, J. D. (2009). Support vector regression and multilayer feed forward neural networks for non-exercise prediction of VO2max. *Expert Systems with Applications, 36*(6), 10112–10119.

Aswani, A., Kaminsky, P., Mintz, Y., Flowers, E., & Fukuoka, Y. (2019). Behavioral modeling in weight loss interventions. *European Journal of Operational Research, 272*(3), 1058–1072.

Blackman, K. C. A., Zoellner, J., Berrey, L. M., Alexander, R., Fanning, J., Hill, J. L., & Estabrooks, P. A. (2013). Assessing the internal and external validity of mobile health physical activity promotion interventions: A systematic literature review using the RE-AIM framework. *Journal of Medical Internet Research, 15*(10), 1–15.

Calvert, T. W., Banister, E. W., Savage, M. V., & Bach, T. (1976). A Systems Model of the Effects of Training on Physical Performance. *IEEE Transactions on Systems, Man and Cybernetics, SMC-6*(2), 94–102.

Cannon, A. R., Cobb, G. W., Hartlaub, B. A., Legler, J. M., Lock, R. H., Moore, T. L., Rossman, A. J., & Witmer, J. A. (2018). *STAT2: Modeling with Regression and ANOVA*. W.H. Freeman; 2nd edition (December 21, 2018).

Chen, K., Guo, S., Zhang, T., Xie, X., & Liu, Y. (2021). Stealing Deep Reinforcement Learning Models for Fun and Profit. *ASIA CCS 2021 - Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security, 307–319.

Daniels, J. (2014). Daniels’ Running Formula (3rd edition). Champaign, IL: Human Kinetics.

Degroote, L., Van Dyck, D., De Bourdeaudhuij, I., De Paepe, A., & Crombez, G. (2020).
Acceptability and feasibility of the mHealth intervention “MyDayPlan” to increase physical activity in a general adult population. *BMC Public Health*, 20(1), 1–12.

del Carmen Rodríguez-Hernández, M., & Ilarri, S. (2021). AI-based mobile context-aware recommender systems from an information management perspective: Progress and directions. *Knowledge-Based Systems*, 215, 106740.

Domin, A., Spruijt-Metz, D., Theisen, D., Ouzzahra, Y., & Vögele, C. (2021). Smartphone-based interventions for physical activity promotion: Scoping review of the evidence over the last 10 years. *JMIR MHealth and UHealth*, 9(7).

Dong, J., Du, Y., & Cai, Z. (2020). Gender recognition using motion data from multiple smart devices. *Expert Systems with Applications*, 147.

Forman, E. M., Kerrigan, S. G., Butryn, M. L., Juarascio, A. S., Manasse, S. M., Ontañón, S., Dallal, D. H., Crochiere, R. J., & Moskow, D. (2019). Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of Behavioral Medicine*, 42(2), 276–290.

Galán-Mercant, A., Ortiz, A., Herrera-Viedma, E., Tomas, M. T., Fernandes, B., & Moral-Munoz, J. A. (2019). Assessing physical activity and functional fitness level using convolutional neural networks. *Knowledge-Based Systems*, 185, 104939.

Gasparetti, F., Aiello, L. M., & Quercia, D. (2020). Personalized weight loss strategies by mining activity tracker data. *User Modeling and User-Adapted Interaction*, 30(3), 447–476.

Hardeman, W., Houghton, J., Lane, K., Jones, A., & Naughton, F. (2019). A systematic review of just-in-time adaptive interventions (JITIs) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 16(1).

Jain, S. (2009). Self-control and optimal goals: A theoretical analysis. *Marketing Science*, 28(6), 1027–1045.

Jaspers, A., Brink, M. S., Probst, S. G. M., Frencken, W. G. P., & Helsen, W. F. (2017). Relationships Between Training Load Indicators and Training Outcomes in Professional Soccer. *Sports Medicine*, 47(3), 533–544.

Kong, D., Bao, Y., & Chen, W. (2020). Collaborative learning based on centroid-distance-vector for wearable devices. *Knowledge-Based Systems*, 194, 105569.

Liu, Y., & Avello, M. (2020). Status of the research in fitness apps: A bibliometric analysis. *Telematics and Informatics, September*, 101506.

Liu, Y., & Avello, M. (2021). Status of the research in fitness apps: A bibliometric analysis. *Telematics and Informatics*, 57(July 2020), 101506.

Marcolino, M. S., Oliveira, J. A. Q., D’Agostino, M., Ribeiro, A. L., Alkmim, M. B. M., & Novillo-Ortiz, D. (2018). The impact of mHealth interventions: Systematic review of systematic reviews. *JMIR MHealth and UHealth*, 6(1).

Mcgregor, S. J., Weese, R. K., & Ratz, I. K. (2009). Performance Modeling in an Olympic 1500-M Finalist: A Practical Approach. *Journal of Strength and Conditioning Research*, 23(9), 2515–2523.

Messner, E.-M., Probst, T., O’Rourke, T., Stoyanov, S., & Harald, B. (2019). mHealth Applications: Potentials, Limitations, Current Quality and Future Directions. In *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics* (pp. 235–248).
Mezei, J., & Nikou, S. (2018). Fuzzy optimization to improve mobile health and wellness recommendation systems. *Knowledge-Based Systems, 142*, 108–116.

Milne-Ives, M., LamMEng, C., de Cock, C., van Velthoven, M. H., & Ma, E. M. (2020). Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: Systematic review. *JMIR MHealth and UHealth, 8*(3), 1–16.

Mintz, Y., Aswani, A., Kaminsky, P., Flowers, E., & Fukuoka, Y. (2017). Non-stationary bandits with habituation and recovery dynamics. *ArXiv, April 2021.*

Mintz, Y., Aswani, A., Kaminsky, P., Flowers, E., & Fukuoka, Y. (2020). Non-stationary bandits with habituation and recovery dynamics. *Operations Research, 68*(5), 1493–1516.

Mnih, V., Badia, A. P., Mirza, L., Graves, A., Harley, T., Lillicrap, T. P., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *33rd International Conference on Machine Learning, ICML 2016, 4*, 2850–2869.

Mousavi Seyed Sajad and Schukat, M. and H. E. (2018). Deep Reinforcement Learning: An Overview. In S. and B. R. Bi Yaxin and Kapoor (Ed.), *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016* (pp. 426–440). Springer International Publishing.

Milne-Ives, M., LamMEng, C., de Cock, C., van Velthoven, M. H., & Ma, E. M. (2020). Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: Systematic review. *Knowledge-Based Systems, 142*, 108–116.

Nahum-shani, I., Smith, S. N., Witkiewitz, K., Collins, L. M., Spring, B., & Murphy, S. A. (2014). Just-in-time adaptive interventions (JITAs): An organizing framework for ongoing health behavior support. *The Methodology Center Technical Report, 073975*(14), 1–37.

Nweke, H. F., Teh, Y. W., Al-garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. In *Expert Systems with Applications* (Vol. 105, pp. 233–261). Elsevier Ltd.

O’Donoghue, T., & Rabin, M. (2011). Doing it now or later. *Advances in Behavioral Economics, 89*(1), 223–251.

Oztekin, A., Al-Ebbini, L., Sevkli, Z., & Delen, D. (2018). A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. *European Journal of Operational Research, 266*(2), 639–651.

Pan, Q., Brulin, D., & Campo, E. (2022). Wrist movement analysis for long-term home sleep monitoring. *Expert Systems with Applications, 187.*

Rabbi, M., Klasnja, P., Choudhury, T., Tewari, A., & Murphy, S. (2019). Optimizing mHealth interventions with a bandit. In *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics* (pp. 277–291). Springer.

Roels, G. (2020). High-performance practice processes. *Management Science, 66*(4), 1509–1526.

Sánchez-Ruiz, D., Olmos-Pineda, I., & Olvera-López, J. A. (2020). Automatic region of interest segmentation for breast thermogram image classification. *Pattern Recognition Letters, 135*, 72–81.

Schmidt, R. A., & Bjork, R. A. (1992). New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychological Science, 3*(4), 207–217.

Schoeppe, S., Alley, S., Rebar, A. L., Hayman, M., Bray, N. A., Van Lippevelde, W., Gnam, J. P., Bachert, P., Direito, A., & Vandelanotte, C. (2017). Apps to improve diet, physical activity
and sedentary behaviour in children and adolescents: A review of quality, features and behaviour change techniques. *International Journal of Behavioral Nutrition and Physical Activity, 14*(1), 1–10.

Sedighi Maman, Z., Chen, Y. J., Baghdadi, A., Lombardo, S., Cavuoto, L. A., & Megahed, F. M. (2020). A data analytic framework for physical fatigue management using wearable sensors. *Expert Systems with Applications, 155*.

Sharavana Kumar, M. G., & Sarma Dhulipala, V. R. (2020). Fuzzy allocation model for healthcare data management on IoT assisted wearable sensor platform. *Measurement: Journal of the International Measurement Confederation, 166*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.

Vijayarajeswari, R., Parthasarathy, P., Vivekanandan, S., & Basha, A. A. (2019). Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. *Measurement: Journal of the International Measurement Confederation, 146*, 800–805.

Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research, 267*(2), 687–699.

Wang, Y., Cang, S., & Yu, H. (2019). A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications 137*, 167–190.

Xu, P., Wei, G., Song, K., & Chen, Y. (2021). High-accuracy health prediction of sensor systems using improved relevant vector-machine ensemble regression. *Knowledge-Based Systems, 212*, 106555.

Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M., & Hochberg, I. (2017). Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *Journal of Medical Internet Research, 19*(10), e338.

Zhou, M., Fukuoka, Y., Mintz, Y., Goldberg, K., Kaminsky, P., Flowers, E., & Aswani, A. (2018). Evaluating machine learning–based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial. *JMIR MHealth and UHealth, 6*(1).