This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

Deep Reinforcement Learning Based Predictive Maintenance Model for Effective Resource Management in Industrial IoT

Kevin Shen Hoong Ong, *Member, IEEE,* Wang Wenbo, *Member, IEEE,* Dusit Niyato, *Fellow, IEEE* and Thomas Friedrichs

***Abstract*—Unplanned breakdown of critical equipment interrupts production throughput in Industrial IoT (IIoT), and data-driven Predictive Maintenance (PdM) becomes increasingly important for companies seeking a competitive business advantage. Manufacturers, however, are constantly faced with the onerous challenge of manually allocating suitably competent manpower resources in the event of an unexpected machine breakdown. Furthermore, human error has a negative rippling impact on both overall equipment downtime and production schedules. In this paper, we formulate the complex resource management problem as a resource optimisation problem to determine if a model-free Deep Reinforcement Learning (DRL) based PdM framework can be used to automatically learn an optimal decision-policy from a stochastic environment. Unlike the existing PdM frameworks, our approach considers PdM sensor information and the resources of both physical equipment and human as part of the optimisation problem. The proposed DRL-based framework and Proximal Policy Optimisation Long Short Term Memory (PPO-LSTM) model are evaluated alongside baselines results from human participants using a maintenance repair simulator. Empirical results indicate that our PPO-LSTM efficiently learns the optimal decision-policy for the resource management problem, outperforming comparable DRL methods and human participants by 53% and 65% respectively. Overall, the simulation results corroborate the proposed DRL-based PdM framework's superiority in terms of convergence efficiency, simulation performance and flexibility.**

Index Terms- **Industrial Internet of Things, Resource Management, Predictive Maintenance, Deep Reinforcement Learning, Decision-Support Systems.**

I. Introduction

INFREQUENT equipment maintenance results in erratic production of defective goods, wastes resources, and results in significant revenue losses. Predictive Maintenance (PdM), enabled by the Industrial Internet of Things (IIoT), seeks to minimise unscheduled downtime of equipment through early identification of potential failures from online sensor data. However, owing to the heterogeneity of manufacturing equipment, maintenance expense, and resource constraints, establishing a generic PdM framework that learns to manage resources (e.g. physical equipment and human) remains a major challenge. Therefore, PdM-based resource management is a promising maintenance strategy for advancing toward a generic PdM framework [1].

According to [2], the majority of existing works use Deep Learning (DL) techniques in equipment failure prognostics. For example, DL-based solutions are used to improve equipment Remaining Useful Life (RUL) estimation and anomaly detection [3]–[6]. Some researchers have recently examined the applicability of Deep Reinforcement Learning (DRL) to PdM, with promising findings. To diagnose and classify faults in time-series based equipment health sensor data, the authors [7] and [8] propose using Deep Q Network (DQN). Similarly, [9] investigates the use of Temporal Difference (TD) Learning for RUL forecasting. Finally, [10] proposes the use of Double DQN to learn the optimal replacement point for equipment based on its health index value, with good generalisation performance across similar equipment. But PdM encompasses more than equipment maintenance [1]. Instead, considerable emphasis should also be placed on enabling maintenance automation to optimise maintenance strategy, particularly in the area of human resource management.

Motivated by this research gap, the improved PdM framework must also consider human resource management, ensuring that maintenance tasks are allocated to the most competent technician. Although [11]–[13] make specific references to predictive maintenance and human resource management in their work, their assumptions are mainly limited to simpler and ideal maintenance scenarios. Note that the highlighted papers represent a fraction of the PdM-related works, and we refer readers to surveys [1]–[3] for details. Ultimately, both businesses and decision-makers will greatly benefit from the data-driven maintenance recommendations, since they will be able to take the best possible action for any physical equipment.

To address the stochastic resource optimisation problem, we propose a Deep Reinforcement Learning (DRL) based model framework that leverages the Proximal Policy Optimisation (PPO) Long Short Term Memory (LSTM) (i.e. PPO-LSTM) model. Firstly, we introduce the edge-powered PdM

This research is carried out under the collaboration program between Computer Networks and Communications Lab, Nanyang Technological University of Singapore and Robert Bosch (SEA) Pte Ltd. This research is also supported, in part, by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), National Research Foundation, Singapore, under AI Singapore Programme (AISG Award No: AISG-GC-2019-003), WASP/NTU grant M4082187 (4080), and Singapore Ministry of Education (MOE) Tier 1 (RG16/20). (Corresponding author: Wenbo Wang.)

Kevin Shen Hoong Ong and Dusit Niyato are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (email: ongs0129@ntu.edu.sg, dniyato@ntu.edu.sg).

Wenbo Wang is with the Faculty of Engineering, Bar Ilan University, Ramat Gan, Israel 5290002 (email: wangwen@biu.ac.il).

Thomas Friedrichs is with the Robert Bosch (SEA) Pte Ltd, Singapore 573943 (e-mail:thomas.friedrichs@sg.bosch.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

framework architecture, which enables PdM for a network of equipment within a generic production facility. Secondly, we formulate pertinent PdM and resource management elements into a Markov Decision Process framework to discover the optimal decision policy. Furthermore, we coin the term "equipment severity rating", which quantifies the probability of equipment failure in relation to the widely used equipment health indicator value in PdM literature. Thirdly, a model-free PPO-LSTM model is presented to address the reward sparsity issue in stochastic settings and discover the optimal decision policy given the stochastic resource optimisation problem. Specifically, the LSTM module is used to capture pertinent spatial-temporal information before further processing by the PPO model. Fourthly, we conduct extensive simulation experiments using a Maintenance Repair Simulator (MRS) to train the PPO-LSTM agent on the relationship between the equipment severity rating, the maintenance cost model, and the technician competence level. Additionally, we undertake IRB-approved real-world experiments with two groups of working professionals to provide a human-level benchmark against which performance is compared to. Consequently, we demonstrate the efficacy of our proposed approach as a decision-support tool. Finally, our DRL approach is also extended to the NASA C-MAPSS dataset, a well-known time-series PdM dataset, and observed positive results. Our **main contributions** in this paper are as follows:

1) We formulate the PdM manpower allocation into a resource optimisation problem and present a DRL-based PdM Model framework. The overall optimisation objective is to increase production revenues while maximising the cumulative equipment uptime in an IIoT network powered by edge computing. With the proposed data-driven framework in place, the routine but challenging task of manpower resource allocation is now automated, resulting in increased productivity for both production and maintenance teams.
2) We propose the PPO-LSTM model for automating the decision-making process associated with PdM-based resource management. LSTM is used to improve the performance of PPO in stochastic settings, while PPO is responsible for selecting the best state-action to perform.
3) We perform extensive simulation experiments using a Maintenance Repair Simulator (MRS) to evaluate the performance of PPO-LSTM, and we undertake IRB-approved real-world experiments, comprised of working professionals, to provide a human-level baseline for performance comparison. Empirically, the proposed PPO-LSTM outperforms both comparable DRL methods and human participants by 53% and 65%, respectively.
4) We also expand the PPO-LSTM model to the NASA C-MAPSS dataset, a well-known time-series PdM dataset. Interestingly, PPO alone reports a 73% improvement in learning efficiency relative to prior work. Overall, these empirical results demonstrate the effectiveness of our proposed DRL-based PdM Maintenance Model framework as a decision-support tool.

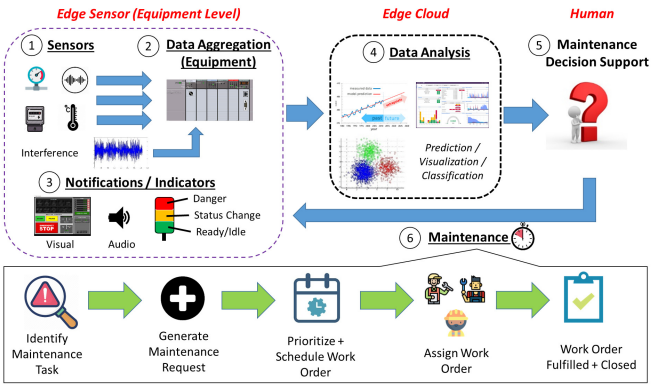


Fig. 1. Predictive maintenance model overview: equipment data generation (Steps 1,2), equipment data notification and indicators (Step 3), equipment data analysis (Step 4), maintenance decision-support (Step 5), and maintenance task workflow (Step 6).

II. Preliminary and Related Work

In this paper, Predictive Maintenance serves two purposes, namely equipment health assessment, and maintenance resource management. In this section, we briefly describe each function and summarise the symbiotic relationship in Figure 1. Thereafter, we describe and critically analyse the limitations of related work and existing PdM-based system models to motivate our research question.

A. Equipment Health Assessment

In situ sensors in modern manufacturing equipment monitor numerous process parameters for signs of process deviations that can affect product quality and provide early warning of equipment failure, see Figure 1 (Steps 1 and 2). However, numerous factors can corrupt sensor data, resulting in anomalous data, in the form of false system alarms and alert notifications, see Figure 1 (Step 3). An anomaly is a statistical concept that refers to data points which vary greatly from other measurements, and previous studies [14]–[17] suggest that the fault evolution of various system degradation models may be modelled as an exponential decay function following:

$$F(t) = e^{-\lambda t}. \tag{1}$$

In addition, it is proposed in [18] that the initial degradation conditions, stress, and wear-rate of individual components can also be generalised into a normalised time-varying health index (H_t) with the following mathematical expression:

$$H_t = 1 - g - e^{at^b}, \tag{2}$$

where the initial degradation condition of g is non-zero; a and b are generalised wear-rates that correspond to the effects of temperature and relevant sub-system stress terms [18].

Overall, the health index models the various component sub-systems, and the trained machine-learning model can be deployed on the equipment either in situ or retrofitting of edge sensors [10]. At the same time, decoupling the model training and re-training processes needed for each equipment

type opens up a conceptual paradigm shift that significantly reduces the risk of re-training a monolithic machine-learning model. In this work, we re-formulate the time-varying health index in [10] to better reflect the maintenance context, with formal definitions described in Section IV-A.

B. Equipment Maintenance Resource Management

One of PdM's primary functions is to forecast the RUL, which may actually lower the economic cost of maintenance. When the cost model-derived maintenance interval T_i is less than the equipment's remaining useful life T_{RUL}, the cost model result may be used for maintenance decision-making. On the contrary, $T_i > T_{RUL}$ implies that equipment failure is impending prior to the next monitoring point. Therefore, maintenance decisions are made using RUL data received from the PdM model, as shown in Figure 1 (Step 4).

The proposed secondary function of PdM is the allocation and management of the maintenance resources, see Figure 1 (Steps 5 and 6). Maintenance resources, in this sense, refer to a group of maintenance staff comprising of engineers and technicians. Effective management of maintenance resources is considered an NP-hard combinatorial problem [19], and conventional mathematical programming techniques are incapable of providing exact solutions within a reasonable time frame. Consider the conventional maintenance routine, where varying complexity of maintenance tasks are allocated to a fixed-shift of technicians with different responsibilities. Simultaneously, the maintenance engineer is accountable for manually allocating maintenance tasks to suitably competent technicians. In this regard, the maintenance engineer may easily have neglected other maintenance factors, such as the cost of technician assigned, complexity of equipment maintenance, and expected mean time to repair. As a result, more equipment downtime is unintentionally incurred relative to the first effort at assigning competent maintenance technicians. Due to the fast-paced nature of the work environment, technicians may perform internal ad hoc task re-scheduling, which increases the likelihood of longer-than-planned equipment downtime.

C. Machine Learning for PdM Applications

Deep-Learning: The broad applications and impacts of Deep-Learning can be found in industry applications such as RUL estimation of aircraft engine [16], RUL of ball-bearings [5], RUL of Battery Life [14], Signal Feature Extraction [6], and Anomaly Detection [20]. Some commonly used models include Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Deep Belief Networks (DBN) and Auto Encoder (AE). These works are often targeted on specific equipment components diagnosis, and no existing work has attempted to extend their proposed solution to multi-component systems, atypical of real-world equipment [1].

Reinforcement Learning: DRL continues to be appealing to companies since it eliminates the labour-intensive data labelling process, and is slowly gaining traction in PdM applications. In [8], the fault diagnosis problem is formulated as a diagnostic game, which is then applied to roller bearings and hydraulic pumps with very high classification accuracy results achieved. Similarly, the work in [7] utilises Deep Q-Network (DQN) for data classification with the UCR dataset [21]. The study in [9] suggests the use of temporal difference learning to derive the equipment health state degradation, and is evaluated on the NASA C-MAPSS dataset. From these examples, DRL usage appears to have been mostly limited to fault classifications and RUL estimation until recently. [10] formulates the equipment health indicator as a function of equipment run-time with the aim of recommending an appropriate replacement point based on the equipment's health. Their proposed Double-DQN based solution is evaluated using the NASA C-MAPSS dataset, and has been observed to achieve good learning performance and sampling efficiency. Despite these encouraging results, DRL-based solutions for PdM applications remains largely under-explored.

Maintenance Recommendation Systems and Framework: [11] propose a general IoT-based framework for businesses to utilise when selecting appropriate Fleet Management Systems (FMS) platform solutions for real-world use. SERENA [22] is a platform for predictive analytic that runs on a lightweight hybrid architecture that combines cloud and edge computing. The SERENA-enabled predictive analytic service, however, provides only equipment condition monitoring using a variety of established machine learning algorithms: Random Forest, Decision Tree, and Gradient Boosted Tree. On the contrary, [23], [24] focus on actionable maintenance recommendations. Both works, which are based on the PHM 2013 Data Challenge, utilise collaborative filtering and Bayesian inference methodology to achieve accurate RUL estimate results for maintenance recommendation. [4] focuses on RUL estimation and proposes the use of Genetic Programming to fuse data from various sensor sources into a composite Health Index, as stated in (2). More recently, performance degradation modelling and threshold-based monitoring approach are investigated in [25] to alert users proactively. However, their approach suffers from a high incidence of false alarms as the human resource component is ignored. As a result, [12] proposes the use of Genetic Algorithm (GA) for asset management, which encompasses both the machine and human component, in order to reduce equipment failure. Unfortunately, GA-based solutions may not scale well with increasing problem complexity, learning convergence may not be as efficient as other optimisation methods, and the solution may be inferior to human-level performance. [13] examines the potential of DRL for decision-support maintenance management and compares the performance to those of working professionals. In this paper, we extend the findings in [13] by considering additional maintenance and human resource parameters in our proposed PdM Model Framework.

D. Critical Analysis

PdM enables maintenance team to address issues prior to equipment failure, and Table I summarises the existing PdM and resource management techniques. As IIoT-enabled PdM is still in its infancy [2], the majority of existing work focuses solely on improving the RUL estimation accuracy and ignores the manpower resource management [27], [28].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

References	Resource		Predictive Equipment Maintenance	Fog / Cloud / Edge / Local	Resource Performance Metrics				Method
	Physical	Human			Time	Cost	Competency Levels	Maintenance Complexity	
[16]	✓	-	✓	Local	-	-	-	-	Attention-based LSTM
[5]	✓	-	✓	N/A	-	-	-	-	LSTM
[14]	✓	-	✓	N/A	-	-	-	-	Deep Belief Network
[6]	✓	-	✓	Local	-	-	-	-	Stacked AutoEncoder
[20]	✓	-	✓	N/A	-	-	-	-	Cumulative Sum-based LSTM
[23]	✓	-	✓	N/A	-	-	-	-	Collaborative Filtering
[24]	✓	-	✓	N/A	-	-	-	-	Bayesian Networks
[7], [8]	✓	-	✓	N/A	-	-	-	-	DQN
[4]	✓	-	✓	Local	-	-	-	-	Genetic Programming
[9]	✓	-	✓	N/A	-	-	-	-	Temporal Difference Learning
[10]	✓	-	✓	Edge	-	-	-	-	DDQN-PER-PN
[11]	✓	✓	-	N/A	-	-	-	-	IoT Asset & Fleet Management Framework
[22]	✓	-	✓	Fog	-	-	-	-	Random Forest, Decision Tree, Gradient Boosted Tree
[26]	✓	✓	✓	N/A	-	-	-	-	Maintenance 4.0 System Architecture
[12]	✓	✓	✓	Fog	✓	✓	-	-	Genetic Algorithm
[13]	✓	✓	✓	Local	✓	✓	1	-	Recurrent Advantage Actor-Critic
Proposed work	**✓**	**✓**	**✓**	**Edge**	**✓**	**✓**	**3**	**✓**	**PPO, PPO-LSTM and DRL-based PdM Framework**

TABLE I: Comparison of PdM-based Resource Management methods with existing works.

Except for [13], none of the existing work addresses all four performances metrics: time, cost, competency levels and maintenance complexity when proposing a resource management method that encompasses both physical and human resources. For example, operational and maintenance costs, such as skilled technician's man hours, mean time to repair, parts replacement and maintenance budget, are often overlooked in existing PdM framework. Moreover, [13] considers only a Junior technician level with a constant repair probability, while the proposed work takes into account several levels of competence and varying repair probabilities.

Given the aforementioned issues, it is essential to investigate and propose an alternative PdM framework. Notably, the research question that our paper address is: *how to manage effectively the symbiotic relationship between the technicians, equipment and AI in a complex environment using sequential decision-making methods while simultaneously optimise revenue as a function of production performance under uncertainty.* In this paper, the proposed PdM framework can include complementary AI-based decision support to facilitate effective resource management across both physical and human resources. In addition, the four performance metrics should be taken into consideration. The IIoT sensor data is fully-utilised for PdM model training, and the utilisation of edge-based sensors will alleviate IIoT network congestion. Owing to these features, this paper addresses the challenges of existing resource management technique, and further details are described in Section III.

III. System Model and PdM Framework

To put Figure 1 into perspective, we consider a generic production facility for predictive equipment maintenance in IIoT (Figure 2), which comprises an Edge Cloud (EC), Edge Sensor (ES), and Manpower Resources. At the equipment level, the system model consists of a network of ESs with direct connection to the EC. Due to computational resource constraints, each ES aggregates in situ time-series sensor data and transmits it to EC for storage and analysis, as shown in Figure 1 (Steps 1 and 2). Meanwhile, a predictive model (i.e. agent) monitors the incoming data for anomalous behaviour and triggers notifications or alarms based on preset parameter threshold.

A. Proposed Predictive Maintenance Framework Architecture

Design Objective: The pandemic's widespread effects continues to be felt as businesses struggle to remain financially viable, and low-wage employees like technicians are perceived as replaceable by automation or rehiring. Given the commercial confidentiality on staffing capacity information, we will assume in this paper that the maintenance personnel to critical equipment ratio increases from 1:1 to 1:10. In particular, our proposed framework considers the real-world constraints of the maintenance budget and technician actions within the decision-making framework, which are both challenging to model and under-explored in similar literature. The proposed framework is briefly illustrated in Figure 2 for reference purposes.

Edge Sensor (Equipment): A static ES symbolises an intelligent sensor predictive model that generates the metadata of the monitored equipment. Metadata is considered to be an abstract representation of the complex relationship between multiple components, where sensors constantly monitor essential component parameters. Examples of sensor data include pump pressure, oil temperature, and milling speed. Generally, a condition-based approach establishes the minimum and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

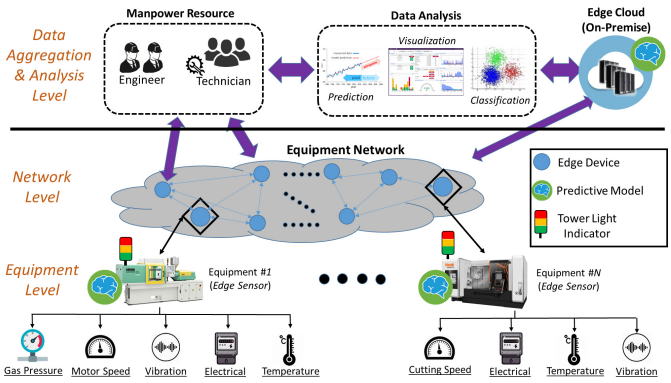


Fig. 2. System overview of the proposed predictive maintenance framework for IIoT networks with edge computing capability per equipment, with on-device predictive models.

maximum limits for each sensor in order to estimate the current health state of the equipment on the basis of domain knowledge. Such an approach is likely to result in a significant number of false alarms, reduces the overall equipment runtime, and wastes precious network bandwidth. Alternatively, an in situ PdM model may be deployed on an ES device with adequate computational resource, and generates metadata upon detection of abnormal sensor signals, such as probabilistic based alarm information and maintenance action recommendations, whilst preserving precious network bandwidth.

Here, we build on [10]'s work, and the proposed list of available actions is based on the data from multiple sensors observed: [*Severity*, *Repair*, *Replace*, *Hold*]. *Severity* denotes the suspected abnormality pertaining to the operational state of the equipment, raw sensor data inputs, and the severity ratings can either be software or hardware based. Depending on the severity level, the PdM model may recommend either *Repair* or *Replace* to the maintenance team. With adequate data and model training, the PdM model is able to assess the optimal action to be taken via a threshold-free approach, and can be remotely deployed on the ES device in order to achieve the outset objective. To be clear, both *Severity* Level 1 and *Hold* are the default states and actions. Besides, the generated metadata is transmitted to the Data Aggregation and Analysis (DAA) layer via a network of Edge Sensor Network (ESN), which can either be a wired or wireless network. The ESDs need only communicate with the Edge Cloud with minimal network latency following the use of high-speed communication technologies such as Gigabit Ethernet, Wi-Fi 6 (IEEE 802.11ax), and 5G.

Edge Cloud (On-Premise): In general, the on-premise EC has orders of magnitude more processing power and memory resources than ES. EC offers several functionalities: (1) Storage for industrial big data; (2) Machine learning pipeline analytic such as data pre-processing, model training and testing, prediction and model deployment; (3) Real-time monitoring and analysis of production planning; (4) Resource management for application-specific decision-making. These resources include but are not limited to human, autonomous robots, and departments. In this paper, we are only concerned with the human manpower resource.

Manpower Resource: Recent survey [29] highlights that a wide variety of competent technicians are expected to retire within the next decade, causing a wider gap in labour's skill/knowledge. To close the gap and achieve effective resource management, we propose that the manpower resource layer in Figure 2 offers multiple functions: (1) Database of relevant employee information, such as experience levels; (2) Estimated manpower cost; (3) Behavioural Risk Profile. We focus on the issues of maintenance management, which are essential to optimising production time of the machinery. One of the decision-making priorities is to assign the correct maintenance technician, for any given equipment severity rating, in order to maximise the probability of fixing the equipment at the first time round, thereby minimising the equipment downtime. Furthermore, each technician wears a mobile wireless device that functions as a PdM model feedback mechanism, and the user inputs are captured by means of feedback ratings through the mobile device's touch-screen interface.

IV. Problem Formulation

Our main objective is to optimise the total production throughput with minimum cost by leveraging the current manpower resource. Given the cost constraints, achieving such trade-off is challenging. We first consider the ESN[1] and DAA layers, and formalise their objectives separately. Then, we merge both layers' definitions and formulate the two-layer interactions. Finally, we conduct problem transformation to contextualise the system parameters and obtain the optimal task assignment strategy. For readers' convenience, we provide a summary of notations used across this paper in Table II.

A. ESN - Sensor MetaData

The objective of the ESN is defined in (3), with the aim to maximise the cumulative uptime (ρ_{E}) of a network of equipment, where g_q denotes the individual uptime of equipment q ($q \in \{1, 2, \ldots, N\}$). Without loss of generality, a single ES is assumed and communicates directly with the on-premise EC.

$$\max \rho_{\mathrm{E}} = \sum_{q=1}^{N} g_q \tag{3}$$

Random ambient noise can affect sensor measurements, and a higher incident rate of false alarms is to be anticipated if the noise characteristics are not quantifiable or observable. Considering external noise as a hidden feature and ES data acquisition is Markovian [9], we propose modelling the complex environment as a sequential decision-making problem using the Partially Observable Markov Decision Process (POMDP) framework. Within the ESN layer, the predictive maintenance module is denoted as an agent for modelling purposes.

Sensor ($\mathcal{S}_{\mathbf{E}}^{x}$): At each equipment, an agent observes at every time-step t the sensor state information, denoted by x_t^i, where

[1]This model was introduced in [10] with simple assumptions, which does not consider the significant implications of severity rating in a resource management problem within the broader maintenance framework. Furthermore, the use of raw sensor data with an unknown noise function necessitates the use of a different problem formulation, which is not considered in [10], either.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

TABLE II: List of Important Notations

Symbol	Definition
ρ_{E}	(Objective) Maximise equipment network uptime
ρ_{D}	(Objective) Optimise resource management
ρ_{H}	(Objective) Optimise user-rating
g_q	Equipment q's uptime
ι	Overall Factory Revenue
N	Number of equipment
M	Number of manpower resource (e.g. technician) set
Γ	Manpower cost w.r.t. experience-level set
τ	Maintenance Budget
$\mathcal{C}$	Cost constraint w.r.t. Maintenance Action
U	User-ratings received w.r.t. contextual situation set
x_t^i	ith sensor's data value at t time-step
H_t	Normalised equipment health value at t time-step
ψ, ψ_t^i	Normalised Equipment Severity Rating set, scalar value with i ratings at time-step t
$\mathcal{S}_{\mathrm{E}}^{\psi}$, $\mathcal{S}_D^{\psi}$	(State) Discrete severity level at ESN and DAA layers respectively
S^Z	(State) Human emotion
n	Number of human emotion categories/levels
Υ	Equipment priority
χ	Array or group of repair actions
κ	(Action) Idle/Hold
η	(Action) Repair
ϵ	(Action) Replace
y_η	Number of repair actions taken until successful fix
y_ϵ	Number of replace actions taken until successful fix
ω_{k}	Positive constants
π^*	Optimal policy
α	Learning Rate
γ	Discounting factor
$\hat{A}_t$	Estimation of Advantage Function at t time-step
G	Number of actors in actor-critic network
d_{target}	Target value of KL Divergence
β	Weighted factor for KL-Divergence
f_t, i_t, o_t	Forget, Input and Output Gates at t time-step
h_t	State of hidden layer at t time-step
b_f, b_i, b_C, b_o	Bias vectors
$c_t, \tilde{c}_t$	Cell state memory, Cell state candidate at t time-step
σ	Sigmoid activation function

$i \in \{1, 2, \ldots, \varrho\}$; ϱ denotes the upper-bound constraint on the number of sensors per equipment for monitoring and practicality purposes. Considering that the sensor data (i.e. state) is sampled at every time-step, the cumulative states theoretically become a continuous state space, and providing exact solutions within a reasonable time period becomes impractical [30]. We handle this limitation by grouping the data samples into a series of time slices (φ) of constant length μ. Namely, each φ value is a discrete representation of the arithmetic mean sensor data ($\bar{x}_t^i$), such that $\bar{x}_t^i \in \left\{x_1^i, x_2^i, \ldots, x_\mu^i\right\}$. Without loss of generality, $\bar{x}_t^i$ is formalised within the environmental state ($\mathcal{S}_{\mathrm{E}}^x$) as $\mathcal{S}_{\mathrm{E}}^x \leftarrow \bar{x}_t^i \ ; \ \forall \varphi$.

Operating Condition ($\mathcal{S}_{\mathbf{E}}^L$): Owing to the repeated use of equipment and complex environmental conditions, the overall performance of the equipment steadily deteriorates over time and a concave-like exponential decay trend [18] is observed before eventual equipment failure, see Figure 3a. Likewise, the sensor's life expectancy decreases over time where some sensors fail faster than others due to ageing and environmental exposure, such as operating temperature. For modelling purposes, the environmental state ($\mathcal{S}_{\mathrm{E}}^L$) can be influenced by the operating temperature condition (L) and binary operating status is utilised. For instance, normal operating status ($\mathcal{S}_{\mathrm{E}}^L = 0$) is user-defined and conditional on $L \in [25, 70]$. Otherwise, abnormal operating status ($\mathcal{S}_{\mathrm{E}}^L = 1$) is assumed.

Severity Rating ($\mathcal{S}_{\mathbf{E}}^{\psi}$): Given N equipment, each ESN-level equipment outputs a severity rating (ψ_t^i), where $i \in [0, \delta]$ at every time-step (t), and the range of severity rating is arbitrarily defined. We normalise i, where $\delta = 1$, and ψ_t^i is formalised within environment state ($\mathcal{S}_{\mathrm{E}}^{\psi}$) as $\mathcal{S}_{\mathrm{E}}^{\psi} \leftarrow \psi_t^i$. The observed severity rating sequentially increases in accordance to the equipment's operational status. For example, we can assume that $\psi_t^{0.2}$ indicates normal operational state while ψ_t^1 indicates critical operational state.

Thus, the overall ESN-based state space $s \in \mathcal{S}_{\mathrm{E}}$ is summarised as an N-set Cartesian product using our system model and POMDP framework as follows:

$$\mathcal{S}_{\mathrm{E}} = \mathcal{S}_{\mathrm{E}}^x \times \mathcal{S}_{\mathrm{E}}^L \times \mathcal{S}_{\mathrm{E}}^{\psi} \tag{4}$$

Action ($\mathcal{A}_{\mathbf{E}}$): The action space comprises of *Hold*(κ), *Repair*(η) and *Replace*(ϵ), and a maintenance budget (τ) is managed by the agent. In order to simulate the maintenance decision-making process, an action-based cost constraint ($\mathcal{C}$) helps to ensure the agent establishes a reasonable maintenance strategy, where $\mathcal{C} \in \{\mathcal{C}_\kappa, \mathcal{C}_\eta, \mathcal{C}_\epsilon\}$ actions are available. By assuming the equipment's health degradation follows (1), the agent is then tasked with selecting the sequence of actions to take at each state without violating τ. By default, κ action incurs zero maintenance cost and the list of action cost relationship is defined as follows:

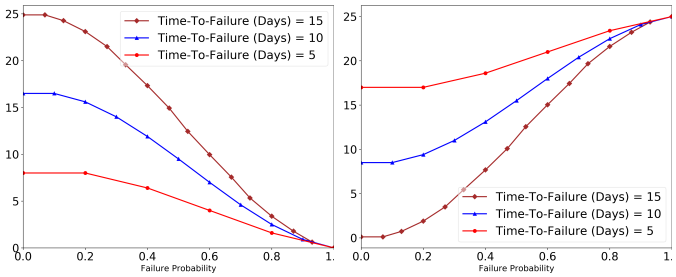
$$\mathcal{A}_{\mathrm{E}} = \begin{cases} (\epsilon, \eta, \kappa) \mid \\ \sum_{q=1}^{N} \mathcal{C}_\epsilon^q \geq 2 \sum_{q=1}^{N} \mathcal{C}_\eta^q \ \text{ and } \ \beta - \sum_{q=1}^{N} \mathcal{C}_\epsilon^q \geq 0, \\ \sum_{q=1}^{N} \mathcal{C}_\epsilon^q \leq \sum_{q=1}^{N} (\mathcal{C}_\eta^q / 2) \ \text{ and } \ \beta - \sum_{q=1}^{N} \mathcal{C}_\eta^q \geq 0. \end{cases} \tag{5}$$

Belief State Transition ($\mathcal{O}$): Compared to a new sensor, the damaged sensor can record values of the sensor state values that are likely different, and such behaviour is detected externally due to multiple state skipping. Assuming that the multiple state skipping phenomenon can be used to infer the existence of an indirectly observable interference, the environmental state (s) is therefore deemed partially observable (o). For convenience, we set o to be s, and the observation-transition probability ($\mathcal{O}$) is instead used to identify the true value of s. In a sense, the environmental state in POMDP is encoded, and the agent relies on a set of *beliefs* (b) for a probability distribution over s, where $b_t(s) = P(s_t = s | o_t, a_{t-1}, \ldots, b_o)$ and b_o is the initial belief vector. The Bayes rule is then used to calculate the belief state transitions, defined as follows:

$$\begin{aligned} b'(s) &= P(s|o, a, b) \\ &= \frac{\Omega(o|s', a) \sum_{s \in S} T(s'|s, a)\, b(s)}{\sum_{s' \in S} \Omega(o|s', a) \sum_{s \in S} T(s'|s, a)\, b(s)}. \end{aligned} \tag{6}$$

For an equipment under monitoring, the belief state transitions for each equipment sensor are iteratively calculated and analysed by the agent. A single set of complementary metadata information is comprised of *Equipment Health* (H) and *Severity* (ψ), defined in (2) and (7) respectively. Recall,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal



(a) Relative Equipment Health. (b) Relative Severity.

Fig. 3. H_t in (2) is obtained by dimension reduction of the multiple sensor data acquired using Principal Component Analysis, and follows the decay pattern in (1). We then slice H_t over an arbitrary time-interval (in days) and apply the Markov chain rule to obtain the state-transition values as H_t approaches 0, indicative of equipment failure. The corresponding range of severity ratings is based on (7).

H presumably follows the exponential decay trend in (2). Intuitively, we expect ψ to increase with decreasing values of H_t. Formally, $\psi \in \psi_t^i$ is a relative complement of $H \in H_t$, where $H_t \in [0,1]$ and $\psi \in [0,1]$. For reference, we visually describe the complementary relationship in Figure 3 and is mathematically expressed as follows:

$$\psi = 1 - H. \tag{7}$$

Given the stochastic environment, the agent will randomly choose actions from (5) to perform based on the observed belief state changes and transitions. Consequently, an equipment's H_t can be restored with ϵ to an almost new condition, while η regresses H_t to a previously observed state by y_η states. Furthermore, the quality of repairs varies and the equipment health state change $\phi(\mathcal{S}_\text{E})$ will differ depending on the $\mathcal{A}_\text{E}$. For example, performing η at $\mathcal{S}_\text{E} = y - 1$ induces a belief state transition from $\mathcal{S}_\text{E} = y$ to $\mathcal{S'}_\text{E} = y - |\phi(\mathcal{S}_\text{E})|$. Such behaviour is concisely represented as $y_\eta \in \{\phi(\mathcal{S}_\text{E}) | \mathcal{S}_\text{E} \in \mathcal{A}_\text{E}\}$.

Reward ($\mathcal{R}_\mathbf{E}$): To contextualise (3) within the POMDP framework, we propose for ρ_E be re-expressed as $\mathcal{R}_\text{E}$, where $\rho_\text{E} \rightarrow \mathcal{R}_\text{E}$. Based on the observed values of $\mathcal{A}_\text{E}$ and $\mathcal{S}_\text{E}$, the agent learns to make better decisions, and this behaviour is motivated by the following reward function:

$$r_t(s_t, a_t) = \begin{cases} \mathcal{R}_\epsilon, & \text{if } \mathcal{S}_\text{E}^x > 0, \ \ \beta > 0, \\ \mathcal{R}_\eta, & \text{if } \mathcal{S}_\text{E}^x > 0, \ \ \beta > 0, \\ \mathcal{R}_\text{Exp}, & \text{if } \mathcal{S}_\text{E}^x > 0, \\ \mathcal{R}_\text{Frug}, & \text{if } \mathcal{S}_\text{E}^x > 0, \ \ \beta > 0, \\ -1, & \textit{Otherwise}, \end{cases} \tag{8}$$

where $r_t \in \mathcal{R}_\text{E}$, $s_t \in \mathcal{S}_\text{E}$ and $a_t \in \mathcal{A}_\text{E}$ at every time-step (t). Given current state s_t and $(\tau - \mathcal{C}_\epsilon > 0) \wedge (\tau - \mathcal{C}_\eta > 0)$, the agent selects either Replace (r_ϵ) or Repair (r_η) actions and values of s_{t+1} and r_t are received. Otherwise, the agent is penalised where $(s_t = H_t = 0) \vee (a_t = \kappa)$. In order to mitigate the reward sparsity problem in real-world applications, the agent is encouraged to explore the problem state space using $\mathcal{R}_\text{Exp}$. Similarly, $\mathcal{R}_\text{Frug}$ is defined to positively reinforce the agent's frugal behaviour when learning the optimal action to take, emulating real-world human decision-making parameter.

B. DAA - Resource Management

Similar to ESN in (3), the DAA layer (ρ_D) aims to optimise the total equipment run-time (g_q), based on the equipment severity ratings (ψ) and maintenance budget constraints (τ), and is denoted as follows:

$$\max \rho_\text{D} = \max \rho_\text{E} \mid (\psi, \tau) = \max \sum_{q=1}^{N} g_q \mid (\psi, \tau). \tag{9}$$

Given the similarity to ESN's objective, and that DAA's environmental constraints are fully-observable, we exploit this information and re-express the objective function as a fully-observable Markov Decision Process (MDP). Within the DAA layer, the resource management model or human participant is abstracted as an agent for modelling purposes. Every equipment is uniquely represented as $\mathcal{S}_\text{E}^i$, where $1 \leq i \leq N$ denotes the i^th equipment in the equipment network.

Equipment Severity Rating ($\mathcal{S}_\text{D}^\psi$): Previously in Section IV-A, the initial values of $\mathcal{S}_\text{E}^\psi$ are considered to be partial observations o in the POMDP context. Subsequently, the o values are believed to be accurate because data processing using traditional machine learning technique is utilised to calculate and acquire accurate equipment health H metadata as defined in (7). As a result, the o values of ψ in (7) can be regarded as the true equipment severity rating state within the fully-observable MDP context. In other words, based on the received severity rating information from $\mathcal{S}_\text{E}^\psi$ in (4), the observed data is assumed accurate and requires no further data processing. Thus, we can mathematically represent $\mathcal{S}_\text{E}^\psi$ within DAA's environment state ($\mathcal{S}_\text{D}^\psi$) as $\mathcal{S}_\text{D}^\psi \leftarrow \mathcal{S}_\text{E}^\psi$.

User Emotion ($\mathcal{S}_\text{D}^Z$): Unlike existing works that assume a completely rational agent, the human action is a function of multiple parameters, such as emotional states, age, and risk-aversion attitude [31]. While the human emotional states are stochastic, we employ the Markov chain [32], [33] to model human emotions (S_D^Z) into n emotional states, such as *Calm*, *Cautious*, and *High Alert* with conditional constraints in (10). In addition, other factors may affect the equipment's severity level to behave stochastically with the similar behaviour as observed over N equipment.

$$\mathcal{S}_\text{D}^Z = \begin{cases} \text{Calm}, & \text{if } Z \in [0, 1/(n)] \mid \{n \in \mathbb{R}\}, \\ \text{Cautious}, & \text{if } Z \in [0.01 + 1/n, 2/n] \mid \{n \in \mathbb{R}\}, \\ \text{High Alert}, & \text{if } Z \in [0.01 + 2/n, 1.0] \mid \{n \in \mathbb{R}\}. \end{cases} \tag{10}$$

With reference to our system model and MDP framework, we integrate the S_D^Z into the DAA-based state space $s \in \mathcal{S}_\text{D}$ as an N-set Cartesian product as follows:

$$\mathcal{S}_\text{D} = \mathcal{S}_\text{D}^{(1,\,\psi,\,Z)} \times \mathcal{S}_\text{D}^{(2,\,\psi,\,Z)} \times \cdots \times \mathcal{S}_\text{D}^{(N,\,\psi,\,Z)}, \tag{11}$$

where $s \in \mathcal{S}_\text{D}$; $N \in [1, \infty]$ represents the equipment index; $Z \in [0,1]$ represents the normalised emotional state; n denotes the user-defined levels of human emotional states; ψ denotes the equipment's fully observable severity level state.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

Although the degradation behaviour of no two identical equipment is alike, continuous equipment usage, after some time, inherently leads to an exponential increase in occurrences of non-operational states, where $\psi_t^i > 1$. In the event where $N > 1$ equipment reports $\psi_t^i > 1$ values, a human operator psychologically associates and combines present state information [34], prioritises the equipment information received, before focusing on the subsequent course of action.

Action ($\mathcal{A}_\mathbf{D}$): The similarities between $\mathcal{A}_\mathrm{E}$ and the proposed action space ($\mathcal{A}_\mathrm{D}$) is $\mathcal{A}_\mathrm{E} \subset \mathcal{A}_\mathrm{D}$, where $\mathcal{A}_\mathrm{D}$ comprises two additional independent action groups: *Equipment Priority* (Υ) and *Repair Type* (χ). An array of $\mathcal{S}_\mathrm{D}^{(N,\,i,\,Z)}$ values, from (11) are stored within Υ, and *heuristic* is used to recommend the appropriate equipment sequence to action upon. The scalar actions of *Hold*(κ), *Repair*(η) and *Replace*(ϵ) can be compacted as $\chi \in \{\kappa, \eta, \epsilon\}$. The frequencies of actions ϵ and η are finite, constraint by τ, so as to imitate real-world decision-making. We propose to characterise the maintenance action constraint as $\mathcal{C}$, where $\mathcal{C} \in \{\mathcal{C}_\kappa, \mathcal{C}_\eta, \mathcal{C}_\epsilon\}$ actions are available. Otherwise, κ remains the default action and zero cost is incurred. Labour costs (Γ) in particular take into account the aforementioned maintenance action as well as the skill levels of the dispatched technician. For example, the skill levels can be classified as: *0 to 5 years* ($\Gamma_{l=1}$), *6 to 15 years* ($\Gamma_{l=2}$) and *≥15 years* ($\Gamma_{l=3}$), where $\Gamma \in \Gamma_l \mid \Gamma_l \in \{\Gamma_1, \Gamma_2, \Gamma_3\}$. The action space, which includes the corresponding maintenance action and manpower costs, can then be defined as follows:

$$\mathcal{A}_\mathrm{D} = \begin{cases} (\overbrace{\epsilon, \eta, \kappa,}^{\chi} \Gamma, \Upsilon) \mid \\ \sum_{q=1}^{N} \mathcal{C}_\epsilon^q \geq 2 \sum_{q=1}^{N} \mathcal{C}_\eta^q \;\text{ and }\; \tau - \Gamma - \sum_{q=1}^{N} \mathcal{C}_\epsilon^q \geq 0, \\ \sum_{q=1}^{N} \mathcal{C}_\epsilon^q \leq \sum_{q=1}^{N} (\mathcal{C}_\eta^q / 2) \;\text{ and }\; \tau - \Gamma - \sum_{q=1}^{N} \mathcal{C}_\eta^q \geq 0. \end{cases} \tag{12}$$

State-Action Transition ($\mathcal{T}$): Consider that the following $N = 3$ equipment states are observed: $\mathcal{S}_\mathrm{D}^{(1,\,0.75)}$, $\mathcal{S}_\mathrm{D}^{(2,\,1)}$, $\mathcal{S}_\mathrm{D}^{(3,\,0.25)}$. The optimisation algorithm first chooses action Υ and heuristically determines the equipment order priority as: $\mathcal{S}_\mathrm{D}^{(2,\,1)}$, $\mathcal{S}_\mathrm{D}^{(1,\,0.75)}$. Thereafter, the agent will determine an appropriate action to perform, based on the current value i, and notify the appropriate maintenance technician accordingly. However, in the real world, the maintenance technician is unable to perform maintenance on multiple equipment at the same time, and the severity rating of each unattended equipment will remain at current levels, with κ continuously invoked, until the maintenance personnel invokes either ϵ or η. For example, ϵ is performed on $N = 2$ equipment index, and the observed changes in severity rating state ($\phi(\mathcal{S}_\mathrm{D})$) transitions from $\mathcal{S}_\mathrm{D}^{(2,\,1)}$ to $\mathcal{S}_\mathrm{D}^{(2,\,0.25)}$, which is the default equipment severity rating state, see Figure 4.

Likewise, η generally reverts the current severity rating state towards $\mathcal{S}_\mathrm{D}^{(1,\,0.25)}$. Besides, repair quality is likely to vary, and additional χ actions may be required to adjust an equipment's $\phi(\mathcal{S}_\mathrm{D})$ by y_η times. For reader's convenience, we summarise the aforementioned state-transitions in (13) and (14) respectively.

$$y_\epsilon \in \phi(\mathcal{S}_\mathrm{D}), \;\text{ if }\; \mathcal{S}_\mathrm{D}^{(N,\,i)} \in \chi, \;\; 0.75 \leq i \leq 1, \tag{13}$$

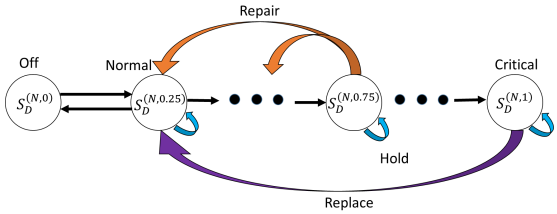


Fig. 4. An example of severity rating state transition for N^{th} equipment w.r.t. types of maintenance action repair performed by the maintenance agent (e.g. human technician or optimisation algorithm.

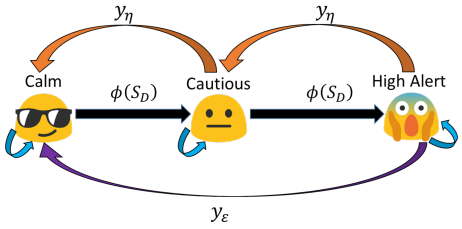


Fig. 5. Proposed human emotional state-transition for maintenance resources based on emotional and mental state transition network models [37], [38] in response to external stimuli [36], such as equipment severity rating.

$$y_\eta \in \phi(\mathcal{S}_\mathrm{D}), \;\text{ if }\; \mathcal{S}_\mathrm{D}^{(N,\,i)} \in \chi, \;\; 0.25 < i \leq 1. \tag{14}$$

According to [31], the Wundt curve model [35] is widely used to model the underlying human behavioural trend with respect to increasing rewards and increasing stimulus intensity. Notably, it is possible to re-interpret the Wundt curve model by splitting it into two partial systems, i.e. a primary reward system and a risk-aversion system with respect to increasing external stimuli [35]. Intuitively, we can correlate external stimuli with equipment severity rating and behavioural based human actions with emotional states. Thus, we can use the state-action transition diagram in Figure 5 to visually describe these correlations, and the actions under consideration includes both action groups Υ and χ. [36] empirically validated the plausibility that human emotions undergo temporal state transitions that typically follow exponential decay, with the decay rate also being situational dependent. In the absence of relevant literature for our maintenance-based research problem, we propose following [36]'s state-transition assumptions and evaluate it against the state-transitions from real-world human participant experimental data.

Considering the maintenance context problem, we assume a non-linear correlation exists between $\mathcal{S}_\mathrm{D}^{\psi}$ and $\mathcal{S}_\mathrm{D}^{Z}$ for N equipment. When, for example, the severity rating state of equipment index $N = 1$ rises from $\mathcal{S}_\mathrm{D}^{(1,\,\psi=0.2)}$ to $\mathcal{S}_\mathrm{D}^{(1,\,\psi=0.3)}$, a similar user emotional transition is predicted following (10). Given $Z \in \mathcal{S}_\mathrm{D}^{Z} \mid \{n = 3\}$, no emotional state transition occurs until $\psi > 1/n$, which consequently triggers $\mathcal{S}_\mathrm{D}^{Z}$ state transition from *Calm* to *Cautious*, as shown in Figures 4 and 5 respectively. Formally, we can express this state-transition correlation as:

$$\phi(\mathcal{S}_\mathrm{D}) \Rightarrow \left\{\phi(\mathcal{S}^\mathrm{Z}) | Z \in \chi, \; Z \in [0, 1]\right\}. \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

Reward ($\mathcal{R}_\mathbf{D}$): When invoking an action from $\mathcal{A}$ given severity rating of state $\mathcal{S}_\mathrm{D}$, the quality of the decision-making policy can be improve via the reward function $\mathcal{R}_\mathrm{D}(s_t, a_t, s_{t+1})$. Recall (9), the cumulative sum of equipment runtime can be re-defined as the cumulative rewards received from N equipment, based on the values of $\mathcal{S}$ and $\mathcal{A}$ taken at each time-step. In consideration of the time-variant repair action, we let the success probability of the η action $P(\varsigma_j = 1 \mid \varsigma_{j-1} = 0)$ be uniformly distributed within Ω time-steps. Hence, a successful repair is denoted as $\varsigma_j = 1$ and the reward function is defined as follows:

$$r_t(s_t, a_t) = \begin{cases} r_\epsilon, & \text{if } \xi \wedge a_t = \epsilon, \\ r_\eta = \Omega - j + 5, & \text{if } \xi \wedge a_t = \eta \wedge \varsigma_j = 1, \\ r_{\eta-1} = +5, & \text{if } \xi \wedge a_t = \eta \wedge , \\ & (0 < \varsigma_j < 1), \\ -0.05, & Otherwise, \end{cases} \tag{16}$$

where $r_t \in \mathcal{R}_\mathrm{D}$ at every time-step (t) and $\xi \Rightarrow (\mathcal{S}_D^\psi > 0, \tau > 0)$; $j \in [1, \ldots, \Omega]$ represents time-to-repair, defining at which time-step the equipment repair is successful. Given the current value of $\mathcal{S}_\mathrm{D}$, the reward signal from r_ϵ and r_η corresponds to the Replace and Repair actions respectively. Furthermore, we enforce a negative reward to encourage state-space exploration at every time-step regardless of the state-action pair selection. For the purpose of contextualising (9) within the MDP framework, we re-express ρ_D as the maintenance resource reward function $\mathcal{R}_\mathrm{D}$, where $\rho_\mathrm{D} \rightarrow \mathcal{R}_\mathrm{D}$.

C. DAA - User-Rating

Acquiring user-ratings is often challenging, and the analytical process is complicated by the lack of dependent variables. As in the maintenance scenario, we define the user-ratings ($U \in U_p$) as an interplay of multiple situational factors, and the user-ratings function (ρ_U) is defined as follows:

$$\max \rho_\mathrm{U} = U_p \times \psi_t^i \times \Gamma_l, \tag{17}$$

where Γ_l and ψ_t^i refers to the skill level of the technician and the equipment severity rating respectively. In other words, both ψ_t^i and Γ_l can be loosely represented as state $s \in \mathcal{S}_\mathrm{U} \mid \{\mathcal{S}_\mathrm{U} \leftarrow \psi_t^i\}$ and action $a \in \mathcal{A}_\mathrm{U} \mid \{\mathcal{A}_\mathrm{U} \leftarrow \Gamma_l\}$. M represents the total number of equipment technicians and the user-defined user-ratings are normalise to $U_p \in [0, 1]$, where $p \in [1, 2, \ldots, M]$. For example, let us consider a user-rating range between 0 (worst) to 10 (best), and the user 2 inputs a feedback rating of 8. Therefore, $U_p = 0.8$ (i.e. $\frac{8}{10}$).

Next, we cast ρ_U into the MDP framework by re-expressing $\rho_\mathrm{U} \rightarrow \mathcal{R}_\mathrm{U}$, where $\mathcal{R}_\mathrm{U}$ denotes the reward function for user-ratings. Consequently, an optimisation algorithm (i.e. user-rating agent) learns to select the optimal action ($\mathcal{A}_\mathrm{U}$) with respect to the $\mathcal{S}_\mathrm{U}$ in order to maximise $\mathcal{R}_\mathrm{U}$ received, which generally leads to improved values of U_p.

D. Overall Problem Formulation

Recall, the aim of this paper is to collectively maximise overall factory revenue (ι) while optimising the factory resources, as described in Sections IV-A, IV-B and IV-C. Taking these multiple objectives into account, the sub-objectives be integrated into the MDP Framework as: $\mathcal{S} \in \{\mathcal{S}_\mathrm{E},\ \mathcal{S}_\mathrm{D},\ \mathcal{S}_\mathrm{U}\}$, $\mathcal{A} \in \{\mathcal{A}_\mathrm{E},\ \mathcal{A}_\mathrm{D},\ \mathcal{A}_\mathrm{U}\}$ and $\mathcal{R} \in \{\mathcal{R}_\mathrm{E},\ \mathcal{R}_\mathrm{D},\ \mathcal{R}_\mathrm{U}\}$.

Considering the optimisation problem for resource management, the resource management algorithm (i.e. agent) interacts with the environment state $\mathcal{S}$ at t time-step, selects an action $\mathcal{A}$ to perform, and receives new values of $\mathcal{S}$ and $\mathcal{R}$ from the environment respectively. As a result, the agent will continue to interact with the environment iteratively, optimising the actions taken based on each state value in order to maximise the cumulative rewards received $\mathcal{R}$. In retrospect, we let $\mathcal{R} \in \mathcal{R}_\iota$ by re-expressing $\iota \rightarrow \mathcal{R}_\iota$, and the new factory revenue reward function is designed as follows:

$$\begin{aligned} &\max \mathcal{R}_\iota = (\omega_1 \mathcal{R}_\mathrm{E} + \omega_2 \mathcal{R}_\mathrm{D} + \omega_3 \mathcal{R}_\mathrm{U}) \mid \mathcal{C} \\ &\text{s.t. } (a):\ \omega \in \{\omega_1, \omega_2, \omega_3\}, \\ &\qquad (b):\ \omega_1 + \omega_2 + \omega_3 = 1, \\ &\qquad (c):\ \mathcal{C} \in \{\mathcal{C}_\kappa, \mathcal{C}_\eta, \mathcal{C}_\epsilon\}, \end{aligned} \tag{18}$$

where $\mathcal{C}$ refers to the maintenance action cost constraints from (12), and the remaining parameter constraints are defined in (18a, 18b, 18c). Positive weight constants $\omega_1, \omega_2,$ and ω_3 are necessary for balancing the three sub-rewards. For page economy and research scope reasons, we set $\omega_1 = 0.1,\ \omega_2 = 0.9,\ \omega_3 = 0$, where ω_3 is considered for future work.

V. Problem Transformation based on RL

The model optimisation problem in (18) is challenging to solve as the optimisation objective requires to manage maintenance resource effectively in the absence of limited or lack of information for modelling purposes. Moreover, the selection of an insufficiently skilled technician to conduct maintenance repair on an equipment with critical severity status potentially leads to a sub-optimal solution (i.e. revenue reduction due to longer equipment downtime). Likewise, learning a hidden Markov model from noisy or stochastic environments is challenging for the Reinforcement Learning (RL) agent, particularly when incomplete and temporal-dependent environment information are critical to obtaining the optimal solution of the model optimisation objective. In this context, traditional model-based dynamic programming tools are also unsuitable due to the significance of time-critical maintenance and an agent's inability to anticipate what the next state will be before the chosen action is taken.

Therefore, in this section, we apply the MDP framework to address our maintenance resource management problem and adopt model-free RL as a solution tool. In what follows, we describe how the MDP framework can be used to achieve optimal decision-making policy. For clarity and brevity, standard RL notations are used for the parameters of *state*, *action*, and *rewards*. The limitations of related RL approaches are briefly highlighted to motivate our proposed DRL solution.

The RL agent's objective is to learn an optimal policy from (18) through *trial-and-error* interactions within the stochastic environment. For every interaction with the environment, the agent receives information about the next state s_{t+1} in addition to the current state reward r_t received. Then, the agent attempts to maximise the long-term cumulative expected reward values

of being in state s_t recursively, and the optimal state-value policy ($V_*(s)$) is achieved by maximising the value function, in (19), over all existing decision policies, where $\gamma \in [0,1]$ is the discounting factor.

$$V_\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma r_{t+1} | s_t = s]. \tag{19}$$

The state transition probability $P(s^{'}, r|(s,a))$ of any stochastic environment is both dynamic and unknown. Hence, the RL agent's strategy is to search recursively for an optimal decision policy $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$ that maps the state $s_t \in \mathcal{S}$ to action $a_t \in \mathcal{A}$. The Q-learning algorithm can learn the optimal policy by maximising the action-value function ($Q_\pi(\mathcal{S}, \mathcal{A})$) over all Q-value policies in (20). The Q-value is recursively updated using Temporal Difference (TD) Learning [30], and the off-policy transitions (s_t, a_t, r_t, s_{t+1}) is learnt.

$$Q_\pi(s,a) = \mathbb{E}_\pi[r_{t+1} + \gamma Q_\pi(s_{t+1}, a_{t+1}) | s_t = s, a_t = a], \tag{20}$$

$$Q^*(s,a) = (1-\alpha) Q(s,a) + \alpha Q_{\text{obs}}(s,a). \tag{21}$$

The optimal Q-function is obtainable as $Q^*(s,a) = \max_\pi V_\pi(s,a)$. The Q-function is updated following (21), where α is the learning rate and $Q_{\text{obs}}(s,a) = r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q'(s',a')$. With $Q^*(s,a)$, the optimal policy is

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} . Q^*(s,a). \tag{22}$$

Reward sparsity is a well-known RL problem and with policy gradient based RL methods, the effect of multiple actions makes it challenging to identify the series of optimal actions to take given a sequence of steps/states, especially in an online setting. In the context of our maintenance problem, the state-of-the-art actor-critic RL approach is more adept and is able to converge to an optimal decision policy in both online and offline policy settings. For instance, the Generalised Advantage Estimator (GAE) [39] approach calculates the advantage of taking an action by using a weighted average of individual advantages over n-steps so as to reduce the variance of the estimator whilst minimising bias. However, the use of multiple actors with GAE trade-off learning efficiency with high variance in subsequent policy network update intervals [40], which can result in sub-optimal decision policy convergence too. A potential solution would be to integrate a shared Long-Short Term Memory (LSTM) module into the actor-critic network in order to reduce policy network variance in estimating actions based on current environment state, at the expense of increased GPU memory resource requirements and longer training time.

In this work, we propose using the Proximal Policy Optimisation (PPO) method to improve the policy network variances of actor-critic solutions, which is responsible for action estimation and is further discussed in the next section. Briefly, PPO imposes penalty-like restrictions to further reduce the variance between policy network updates, at the expense of adding some bias while reducing the occurrence of the agent taking sub-optimal actions. Likely benefits include quicker learning convergence and attaining optimal performance for our maintenance resource management problem. Recently, Furthermore, we would also like to investigate whether it is possible for a non-memory-based actor-critic solution to outperform the LSTM-based actor-critic solutions before extending the comparison to even more challenging equipment maintenance problems.

VI. Proximal Policy Optimisation for Effective Maintenance Resource Management

Policy gradient methods operate by calculating an estimation of a policy gradient and optimising it by using the stochastic gradient ascent algorithm. The estimator $\hat{g}$ can be obtained by differentiating the objective

$$\mathcal{L}^{PG}(\theta) = \hat{\mathbb{E}}_t \left[\log \pi_\theta (a_t|s_t) \hat{A}_t \right], \tag{23}$$

where $\hat{\mathbb{E}}_t [\cdots]$ denotes an empirical average expectation of a batch of finite samples within a sampling and optimisation algorithm; π_θ denotes a stochastic policy and $\hat{A}_t$ is an estimation of the advantage function at time-step t.

Policy gradient methods suffer from two main problems: *Unstable Policy Updates* and *Data Inefficiency* [30]. Policy changes are unpredictable for policy gradient methods because of their large step updates leading to poor policy updates, which consequently lead to learning bad policies. On the contrary, smaller step updates lead to slower learning. It is also preferable for these learning methods to learn from recent experience and exploit. However, current policy gradient methods discard this experience following gradient changes and this exacerbates the learning process, as a neural network requires a large amount of data to learn effectively. In this section, we propose that these issues be mitigated through the Proximal Policy Optimisation (PPO) algorithm [40]. Furthermore, to cope with stochastic environments with long-term dependencies, we suggest supplementing PPO with LSTM, which we explain in this section.

A. Objective Clipping

Based on the idea of importance sampling and a neural network's preference for normalised data, PPO requires to maintain two policy networks. The first policy network $\pi_\theta (a_t|st)$ is used to refine the policy updates based on the previous policy $\pi_{\theta_{\text{old}}} (a_t|st)$, in which this ratio is clipped and the minimum of the both policy actions will instead be considered. In doing so, large policy network updates will be restricted by the clipping threshold (ϵ), and the clipped objective function is described as follows:

$$\mathcal{L}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip} \left(r_t(\theta), 1-\epsilon, 1+\epsilon \right) \right) \hat{A}_t \right], \tag{24}$$

where the probability ratio $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t$ and $r_t(\theta_{\text{old}}) = 1$. Depending on the value of $\hat{A}_t$, the choice of clipping ratio, clip$(r_t(\theta), 1-\epsilon, 1+\epsilon)$, can be either $1-\epsilon$ or $1+\epsilon$ interval range. The pseudocode is shown in Algorithm 1.

B. Adaptive Kullback-Liebler Penalty Coefficient

With reference to Trusted Region Policy Optimisation (TRPO) [41], we can assume that the optimal policies calculated in the trust region are always better, with some upper

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

Algorithm 1 PPO with Clipped Objective

Input: Initial policy parameters θ_0, clipping threshold ϵ
for k=0,1,2,... **do**
  Collect set of partial trajectories D_k on policy $\pi_k = \pi(\theta_k)$
  Estimate advantages $\hat{A}_t^{\pi_k}$ using GAE algorithm [39]
  Compute policy update:
    $\theta_{k+1} = \underset{\theta}{argmax}\, L_{\theta_k}^{CLIP}(\theta)$
  by taking K steps of minibatch SGD (via Adam), where
  $\mathcal{L}_{\theta_k}^{CLIP}(\theta)$
  $= \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\right)\hat{A}_t\right]$
end for

bound guarantee, over the old policy. Thus, the objective function can be calculated as:

$$\max_{\theta} \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}\hat{A}_t\right] - \beta KL\left[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)\right], \tag{25}$$

where β induces a weighted factor to the Kullback-Liebler (KL) KL-divergence penalty, to penalise or incentivize some target value of KL-divergence (d_{target}) during policy updating. In other words, the KL-penalised objective, after several policy updates by stochastic gradient descent, can be written as:

$$\mathcal{L}^{KLPEN}(\theta) = \hat{\mathbb{E}}_t\left[r_t(\theta) - \beta KL\left[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)\right]\right]. \tag{26}$$

Likewise, the KL-divergence, denoted as d in (27), is also computed after every policy updates such that if $d < d_{\text{targ}}/1.5, \beta \leftarrow \beta/2$; $d > d_{\text{targ}} \times 1.5, \beta \leftarrow \beta \times 2$. Then, the updated β value is used in the next policy update interval.

$$\hat{\mathbb{E}}_t\left[r_t(\theta) - \beta KL\left[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)\right]\right]. \tag{27}$$

As a result, PPO is able to inherit TRPO performance, and is optimised by gradient descent methods. For reference, the pseudocode algorithm is described in Algorithm 2.

Algorithm 2 PPO with Adaptive KL Penalty Coefficient

Input: Initial policy parameters θ_0, initial KL penalty β_0, target KL-divergence δ
for k=0,1,2,... **do**
  Collect set of partial trajectories D_k on policy $\pi_k = \pi(\theta_k)$
  Estimate advantages $\hat{A}_t^{\pi_k}$ using generalised advantage estimation algorithm
  Compute policy update:
    $\theta_{k+1} = argmax L_{\theta_k}(\theta) - \beta_k D_{KL}(\theta||\theta_k)$
  by taking K steps of minibatch SGD (via Adam)
  if $D_{KL}(\theta_{k+1}||\theta_k \geq 1.5\delta)$ **then** $\beta_{k+1} = 2\beta_k$
  else if $D_{KL}(\theta_{k+1}||\theta_k \leq \delta/1.5)$ **then** $\beta_{k+1} = \beta_k/2$
  end if
end for

C. Recurrent Neural Network

Long Short Term Memory (LSTM) [42] is a variant of recurrent neural network, and is often used in DRL literature for spatial-temporal feature learning. Individual cells can extract feature states across a recurrent network while preserving temporal information within each cell state, and LSTM uses a gated-like structure for selective transmission of sequential information. Each LSTM cell comprises of forget gates f_t, input gates i_t and output gates o_t. The gate structures are formally described as follows:

$$\begin{aligned} f_t &= \sigma_l(W_f \cdot [h_{t-1}, x_t] + b_f), \\ i_t &= \sigma_l(W_i \cdot [h_{t-1}, x_t] + b_i), \\ \tilde{c}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \\ c_t &= f_t * c_{t-1} + i_t * \tilde{c}_t, \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \\ h_t &= o_t * \tanh(c_t). \end{aligned} \tag{28}$$

From (28), we denote $W_f, W_i, W_c,$ and W_o as weight matrices and $b_f, b_i, b_c,$ and b_o as bias vectors for input vector of sensor data x_t at time-step t; c_t denotes the cell state memory at t time-step, h_{t-1} represents the state of the hidden layer at time-step $t-1$ whereas h_t represents the state of the hidden layer at time-step t; $\tilde{c}_t$ represents a candidate for some cell state at time-step t; $*$ denotes the element-wise multiplication of the vectors and the gated structure behaviour follows a sigmoid activation function σ.

D. PPO Algorithm

For our maintenance simulation problem, we consider an image-based state space (i.e. 500 x 500 pixels) with state representation at the pixel level. Therefore, a Convolutional Neural Network(CNN) is warranted for PPO's policy and value function to learn via shared parameters, hereby termed PPO-CNN. In addition, a loss function is required to back-propagate the policy target and value function gradients for optimisation purposes. In PPO, we also consider an entropy bonus S to manage the state-space exploration and exploitation trade-off in a similar way to the epsilon-greedy strategy of Deep Q Network. Following [40], S can be combined with (26), (24) and (23) at each iteration to optimise an overall objective function, defined as follow:

$$\mathcal{L}^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t\left[\mathcal{L}^{CLIP}(\theta) - c_1\mathcal{L}^{VF} + c_2 S[\pi_\theta](s_t)\right]. \tag{29}$$

From (29) we denote $c_1,$ and c_2 as regularisation coefficients, S denotes an entropy bonus; $\mathcal{L}^{VF}$ is a compact representation of the squared error loss between the learned state-value function $V(s)$ and the target state-value function as $\left(V_\theta(s_t) - V_t^{\text{target}}\right)^2$.

One major drawback of parameter sharing, of both value and policy networks, is performance instability due to the simultaneous backpropagation of gradients across the network during the model learning phase. As such, we plan to empirically identify suitable hyperparameters to manage this issue. In this work, the PPO algorithm to be implemented utilises a fixed-length trajectory where G actors will each collect T time-steps of training data in parallel. Then, the losses for each corresponding objectives on GT time-steps of data will be optimised using a gradient descent-based methods for K

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

epochs, such as the Adaptive Moment Estimation Optimiser (Adam).

The above-mentioned solution, however, is suitable for stochastic environments with short spatial-temporal dependencies (i.e. state-action value pair) and may not achieve optimal results, as it requires the PPO agent to retrieve older state-action sequence information. For example, given the same equipment and technician, the mean-time-to-repair of the equipment can vary greatly due to the different rate of equipment degradation and environmental factors. To address this issue, we propose to modify existing PPO-CNN architecture by inserting an LSTM layer between the Convolutional and Feedforward layers. This model variant is termed PPO-LSTM (Figure 6b), and the cells in the LSTM layer with $h_t = \text{LSTM}\,(o_t,\ h_{t-1})$ are thus used to estimate the $Q\,(h_t, a_t)$ instead of $Q\,(s_t, a_t)$. To be clear, o_t refers to the current observation o_t and may not necessarily correspond to the current environment state s_t while h_{t-1} represents the state of the hidden layer at time-step $t-1$.

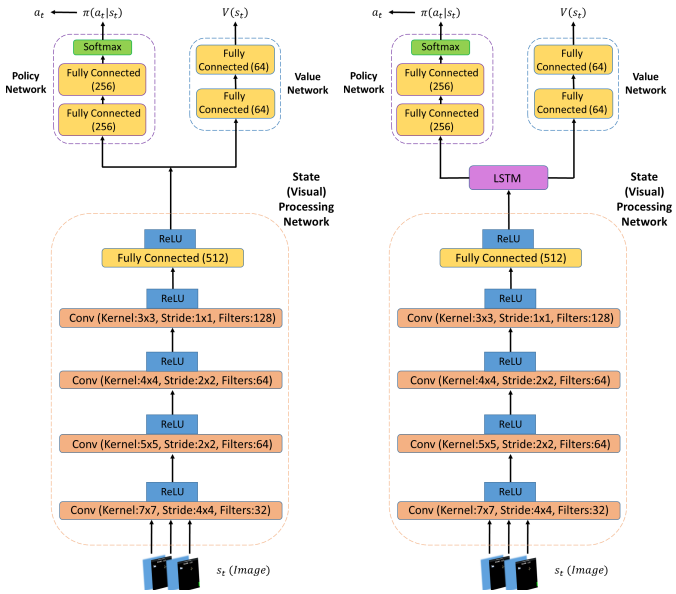
In summary, the proposed policy network constraints of *objective clipping*, *adaptive KL divergence penalty* and *entropy bonus* will be validated with an actor-critic based neural network solution. For reader's understanding, the proposed actor-critic based PPO architecture variants are shown in Figure 6 and the pseudocode is given in Algorithm 3.

VII. Experiment Setup

Due to the scope of the proposed DRL framework, we propose and describe three experiments in this section.

A. Maintenance Repair Simulator

We create a Maintenance Repair Simulator (MRS) that is adequately versatile for a range of purposes, such as model training and validation for the DRL agent and data collection



(a) CNN based PPO (PPO-CNN).

(b) CNN+LSTM based PPO (PPO-CNNLSTM).

Fig. 6. PPO-based actor-critic architecture variants.

Algorithm 3 PPO, Actor-Critic Style

for iteration=1,2,... **do**
  for actor=1,2,...,N **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T time-steps
    Compute advantage estimates $\hat{A}_1, \ldots, A\hat{} + t$
  end for
  Optimise surrogate L w.r.t. θ, with K epochs and minibatch size $M \leq GT$
  $\theta_{\text{old}} \leftarrow \theta$
end for

from human participants for benchmark purposes. As a baseline, we set $\Omega = 120$ in MRS and at each simulation step, the machine repair success probability decreases linearly as $\frac{1}{120}$, $\frac{1}{119}$ at $i = 2$, and so on to emulate real-world scenario in which repair times differ with varying equipment severity rating. Similarly, we set 5 severity levels and generate 3 technician skill levels (eg. Junior, Senior, Expert). Through iterative interaction with MRS, the rewards received by the DRL model also takes into account the different cost constraints and will, based on the skill level of the selected technician, learn an optimal decision-making policy to satisfy (18). We then report the corresponding results for all severity levels, technician skill levels, and PPO respectively.

B. Human Participants

A total of 26 working professionals participated in our IRB-approved experiment which consists of both white-collar and blue-collar workers. The mix-gender participants, aged between 20 to 50 are from Singapore and China, and the overall average participant age falls within the range of 30 to 40 years old. Each participant is presented with a set of instruction and the game objective, which is to maximise the total game rewards received. Relevant game data is automatically captured and every human participant utilises the keyboard to navigate the MRS game environment. Additionally, a warm-up game followed by two games is permitted for each participant, where each game comprises of 30 rounds of gameplay, for example. Yet unbeknownst to all participants, they will play the same game environment under four different game difficulty, where game difficulty is synonymous with equipment severity rating.

The experimental data for all participants is aggregated, pre-processed, and we perform the statistical analysis. Example of data collected for each human participant are user's score, time taken to game completion, and a snapshot of actions taken to complete each game. Thereafter, we demonstrate the efficacy of the AI-based solution for maintenance decision-making by comparing the human participant results to the proposed DRL variants. For disclaimer purposes, the anonymity of all human participant is strictly enforced for data security and privacy purposes throughout the course of experiments.

C. Turbofan Engine Dataset and Data Preparation

The NASA Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset [18] is generated from a commercial degradation simulator for turbofan engines. It includes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

measurements that simulates failure under various operating conditions for several turbofan engines. A quick overview of the engine datasets FD001 and FD003 are shown in Table III as well as the respective fault conditions across multiple sensor measurements.

Dataset	FD001	FD003
Training Set	100	100
Test Set	100	100
Operating Conditions	1	1
Fault Conditions	1	2

TABLE III: C-MAPSS Dataset[2]under test

Each dataset consists of 26 data columns. Columns 1 and 2 refer to engine cycle for specific engines; Columns 3, 4, and 5 denote the sensor measurements, such as temperature and pressure; the remaining columns reflect the simultaneous condition monitoring of 21 sensors. We apply standard data normalisation [10] on the sensor data and assume equipment degradation behaviour following (2). The objective of this experiment is to learn and consistently recommend an effective replacement action ϵ before imminent failure of turbofan engines, based on varying states of equipment health degradation.

VIII. Results and Discussion

We shall briefly highlight the organisation of results for reader's convenience, with details in the respective subsections. Firstly, we present the results from multiple experiments in order to highlight the benefits of PPO, based on MRS Game-1, and articulate the performance contributing factors for the policy network components in comparison to the baseline models and existing work. Secondly, we present results from the human participants and highlight important statistical findings. Table IV is then compiled to aggregate both the human participant and DRL results to illustrate the potential significance of augmenting human technicians in complex decision-making situations. In doing so, we demonstrate the applicability of PPO at the DAA level. Finally, we discuss the effectiveness of our proposed DRL system by presenting key results from C-MAPSS, which is utilised to mimic in situ decision-making at the equipment or ES layer. Furthermore, all model results reported in this paper are the median average over 5 independent runs.

A. Performance Evaluation of Proposed Algorithm

Empirically, a deeper CNN design (i.e. PPO-CustomCNN) benefits from an increase of 7% in learning efficiency and 2% improvement in mean score deviations, when compared to the 3-layer CNN PPO algorithm (i.e. PPO-CNN). For baseline purposes, we hereby denote the non-clipped form of PPO to be implicitly convergent (I.C.), and benchmark PPO's performance against the A2C variants, in Figure 8. Notably, the proposed PPO variants are able to achieve up to 42% increase in learning efficiency. Besides, PPO reliably achieves an optimal score for Game-1, unlike the A2C variants which are merely near optimal. Besides, the performance of PPO

[2]https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#turbofan

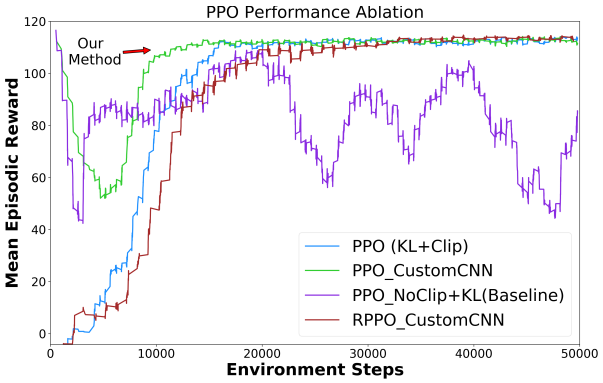


Fig. 7. Performance of our proposed PPO solutions.

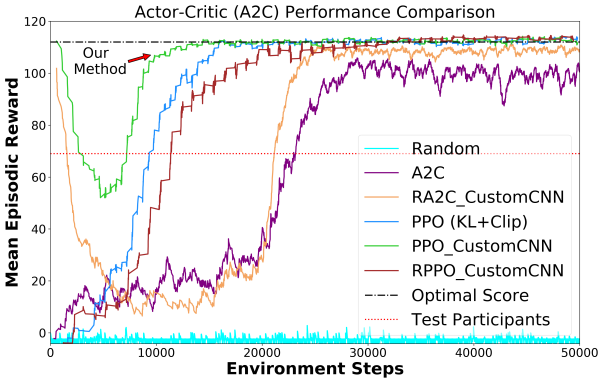


Fig. 8. Performance comparison between A2C variants, PPO variants, and human participants.

with clipping is almost empirically identical to previous state-of-the-art results (i.e. A2C-LSTM). For completeness, we also include the PPO-CNNLSTM variant as part of our performance comparison, and the results are shown in Figures 7 and 8.

Compared to atypical A2C networks, PPO is more robust to hyperparameter tuning. On the contrary, the PPO-CNNLSTM variant (i.e. RPPO_CustomCNN) requires a reasonably comprehensive hyperparameter tuning in order to achieve learning convergence. In other words, we discover that increasing the number of recurrent cells in the LSTM network to twice the step size of PPO yields best results and convergence. One interesting insight from PPO-CNNLSTM variant is that it appears to trade-off learning efficiency with higher overall scores as opposed to non memory-augmented PPO policies. A beneficial side-effect of the LSTM network inclusion is the overall reduction of policy variance losses, and learning convergence variability in reward scores, as shown in Figure 7. Across all 4 games, the PPO-CNNLSTM design on average outperforms both PPO-CNN and A2C-CNN-LSTM by 4% and 3% respectively. For reader's convenience, we highlight the top performers per game, in bold, and the experimental results are listed in Table IV.

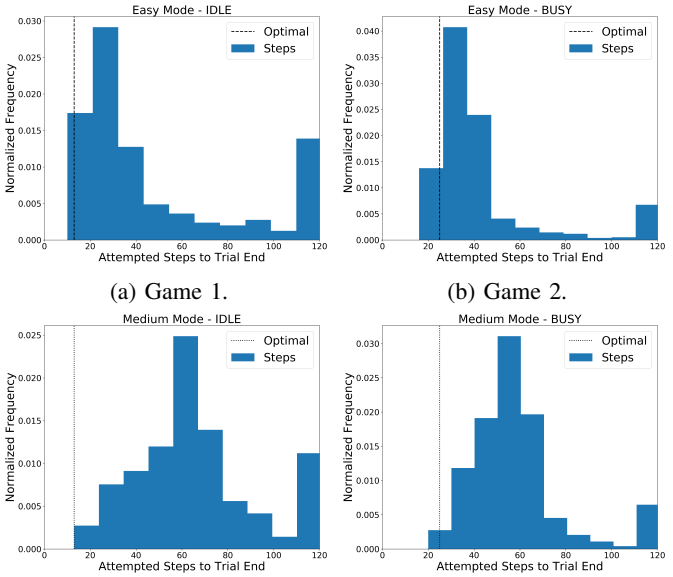
B. Human Participant Analysis

In relation to Figure 8 and Table IV, the human participant scores are clearly sub-optimal across all 4 games. By performing statistical analysis, we obtain insights into the human

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3109955, IEEE Internet of Things Journal

Game Modes	Time-steps to Learning Convergence (10^3)					Mean Reward Received					Human Participants Score
	A2C	PPO				A2C	PPO				
	CNN + LSTM	CNN	(w/CustomCNN)			CNN + LSTM	CNN	(w/CustomCNN)			
		w/ Clipping	PG No Clipping	Clipping	Clipping + LSTM		w/ Clipping	PG No Clipping	Clipping	Clipping + LSTM	
Game 1 (P(Fix)=1.0)	26	16	I.C.	**15**	32	108 ± 9.7	112 ± 6.2	88 ± 38.5	112 ± 6.1	**113 ± 5.5**	66 ± 47
Game 2 (P(Fix)=0.9)	27.5	**19**	I.C.	24	32	95 ± 9.5	98 ± 6.5	81 ± 28.8	**98 ± 6.2**	97 ± 7.1	72 ± 42
Game 3 (P(Fix)=0.6)	28	**12.5**	I.C.	15.5	21	143 ± 84.7	134 ± 55.1	92 ± 68.8	138 ± 46.8	**154 ± 48.3**	70 ± 52
Game 4 (P(Fix)=0.5)	31.5	**18.5**	I.C.	**18.5**	26	121 ± 63.4	115 ± 39	61 ± 57.9	114 ± 36.1	**118 ± 32.4**	88 ± 51

TABLE IV: Sample efficiency (lower is better) for A2C and PPO on 4 game environments are shown with and without the proposed optimisations. Performance results (higher is better) of each test with the mean rewards obtained for each benchmark. For comparison, the human participants score are also shown.



(a) Game 1. (b) Game 2.

(c) Game 3. (d) Game 4.

Fig. 9. Histogram analysis of human participant results (bins=10).

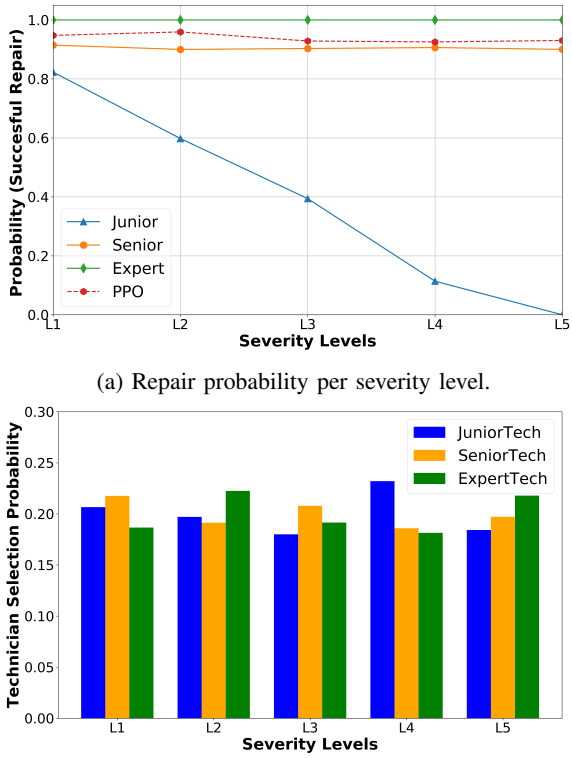
participant group's mean performance distribution as well as relative individual performance metrics in every game. For reference, the histogram analysis for all human participant performance, in all 4 games, is described in Figure 9.

C. Equipment Severity Rating w.r.t. Technician Skill Level

With the encouraging results from Table IV, the PPO-CNNLSTM model evaluations are also conducted on the MRS simulator with all technician skill levels being effected across the range of equipment severity ratings. Mean-Time-To-Repair (MTTR) information can be derived from the rewards earned by each technician and, by normalising the MTTR information, the probability of successful repairs can be obtained, see Figure 10a. Notably, these findings are reasonably consistent with our hypothesised risk-based state-transition relationship in Section IV-B.

The overall performance of PPO-CNNLSTM is empirically consistent between the senior and expert technicians, and its performance is characterised by considerations of the cost and availability of the manpower resources at any given time. For instance, once PPO-CNNLSTM identifies and dispatches a technician to repair the equipment, further resource substitution is disallowed, and the repair job must be completed by the dispatched technician, i.e., the DRL agent in this simulation case. Through iterative interaction with MRS, PPO-LSTM learns to dispatch the optimal skilled technician so as to maximise its overall reward received at all severity ratings. Accordingly, it is sub-optimal to select the expert level technician for the majority of severity ratings and its performance is justified by the severity level based technician selection probabilities in Figure 10b.

To summarise the first two experiments, the potential benefits of DRL augmented human decision-making for predictive maintenance action recommendation are clearly shown. Be-



(a) Repair probability per severity level.

(b) Normalised PPO action probability.

Fig. 10. Evaluating the decision-making effects of technician selection w.r.t. severity levels.

sides, PPO-CNNLSTM's improved learning efficiency results are due to the use of multiple actors, which utilise modern Edge Computing resources, such as multi-core CPUs and GPUs, to realise a practical reduction in model training wall time when compared to similar DRL approaches.

D. C-MAPSS

When the original $N = 256$ learner hyperparameter is applied, poor performance convergence is observed because the multiple actors execute conflicting updates with our 256-step trace updates hyperparameter, causing PPO to behave like a Monte Carlo process [30]. Furthermore, by replacing the CNN module with two fully-connected layer with 64 neurons each, the standalone PPO model achieves comparable performance to [10] with the added benefit of learning efficiency improvement of 73% (i.e. reduction from 12×10^3 to 3.2×10^3 time-steps). Notably, the main hyperparameter for attaining learning convergence is to reduce the number of PPO actors to a single learner and environment.

IX. Conclusion and Future Work

In this paper, we presented a deep reinforcement learning framework for an edge computing-based predictive maintenance model, to effectively manage the dynamic decision-making process involving equipment maintenance, maintenance cost model, and manpower resource. We formulated the complex resource management as a deep reinforcement learning problem for learning an optimal decision policy given a stochastic environment and time-series data. We evaluated the performance of the proposed PPO-LSTM using a maintenance repair simulator, and the findings are compared to those of human participants. The simulation results verify the efficacy of our framework and PPO-LSTM approach in addressing the challenging maintenance resource management problem, outperforming both human participants and the baselines in terms of convergence rate and performance. For future work, we plan to increase the learning efficiency of reinforcement learning using knowledge transfer approaches, such as offline-to-online policy learning, continual learning and transfer learning.

References

[1] Y. Ran, X. Zhou, P. Lin, Y. Wen, and R. Deng, "A survey of predictive maintenance: Systems, purposes and approaches," *arXiv preprint arXiv:1912.07383*, 2019.

[2] M. Compare, P. Baraldi, and E. Zio, "Challenges to iot-enabled predictive maintenance for industry 4.0," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4585–4597, 2019.

[3] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[4] P. Wen, Y. Li, S. Chen, and S. Zhao, "Remaining useful life prediction of iiot-enabled complex industrial systems with hybrid fusion of multiple information sources," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 9045–9058, 2021.

[5] L. Guo, N. Li, F. Jia, Y. Lei, and J. Lin, "A recurrent neural network based health indicator for remaining useful life prediction of bearings," *Neurocomputing*, vol. 240, pp. 98–109, 2017.

[6] P. Lin and J. Tao, "A novel bearing health indicator construction method based on ensemble stacked autoencoder," in *2019 ICPHM*. IEEE, 2019, pp. 1–9.

[7] C. Martinez, G. Perrin, E. Ramasso, and M. Rombaut, "A deep reinforcement learning approach for early classification of time series," in *2018 26th EUSIPCO*. IEEE, 2018, pp. 2030–2034.

[8] Y. Ding, L. Ma, J. Ma, M. Suo, L. Tao, Y. Cheng, and C. Lu, "Intelligent fault diagnosis for rotating machinery using deep q-network based health state classification: A deep reinforcement learning approach," *Advanced Engineering Informatics*, vol. 42, p. 100977, 2019.

[9] C. Zhang, C. Gupta, A. Farahat, K. Ristovski, and D. Ghosh, "Equipment health indicator learning using deep reinforcement learning," in *Joint ECML PKDD*. Springer, 2018, pp. 488–504.

[10] K. S. H. Ong, D. Niyato, and C. Yuen, "Predictive maintenance for edge-based sensor networks: A deep reinforcement learning approach," in *2020 IEEE 6th WF-IoT*. IEEE, 2020, pp. 1–6.

[11] J. Backman, J. Väre, K. Främling, M. Madhikermi, and O. Nykänen, "Iot-based interoperability framework for asset and fleet management," in *2016 21st Int. Conf. on ETFA*. IEEE, 2016, pp. 1–4.

[12] Y. K. Teoh, S. S. Gill, and A. K. Parlikad, "Iot and fog computing based predictive maintenance model for effective asset management in industry 4.0 using machine learning," *IEEE Internet Things J.*, 2021.

[13] K. S. H. Ong, W. Wang, T. Friedrichs, and D. Niyato, "Augmented human intelligence for decision making in maintenance risk taking tasks using reinforcement learning," in *2021 IEEE Int. Conf. on Syst, Man, and Cybern. (SMC)*. IEEE, 2021, p. To appear in late 2021.

[14] L. Li, Y. Peng, Y. Song, and D. Liu, "Lithium-ion battery remaining useful life prognostics using data-driven deep learning algorithm," in *2018 PHM-Chongqing*. IEEE, 2018, pp. 1094–1100.

[15] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—reviews, methodology and applications," *Mechanical systems and signal processing*, vol. 42, no. 1-2, pp. 314–334, 2014.

[16] Z. Chen, M. Wu, R. Zhao, F. Guretno, R. Yan, and X. Li, "Machine remaining useful life prediction via an attention based deep learning approach," *IEEE Trans. on Industrial Electronics*, 2020.

[17] G. Zhao, G. Zhang, Y. Liu, B. Zhang, and C. Hu, "Lithium-ion battery remaining useful life prediction with deep belief network and relevance vector machine," in *2017 ICPHM*. IEEE, 2017, pp. 7–13.

[18] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *2008 ICPHM*. IEEE, 2008, pp. 1–9.

[19] S. Bouajaja and N. Dridi, "A survey on human resource allocation problem and its applications," *Operational Research*, vol. 17, no. 2, pp. 339–369, 2017.

[20] G. Aydemir and B. Acar, "Anomaly monitoring improves remaining useful life estimation of industrial machinery," *Journal of Manufacturing Systems*, vol. 56, pp. 463–469, 2020.

[21] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, G. Shaghayegh, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, and H.-M. Batista, Gustavo, "The ucr time series classification archive," October 2018, https://bit.ly/3CjHcDj.

[22] S. Panicucci, N. Nikolakis, T. Cerquitelli, F. Ventura, S. Proto, E. Macii, S. Makris, D. Bowden, P. Becker, N. O'Mahony *et al.*, "A cloud-to-edge approach to support predictive analytics in robotics industry," *Electronics*, vol. 9, no. 3, p. 492, 2020.

[23] S. Das, "Maintenance action recommendation using collaborative filtering," *International Journal of Health Policy and Management*, vol. 4, no. 2, pp. 7–12, 2013.

[24] V. Katsouros, V. Papavassiliou, and C. Emmanouilidis, "A bayesian approach for maintenance action recommendation," *International Journal of Prognostics and Health Management*, 2013.

[25] A. K. Farahat, C. Gupta, and H.-k. Tang, "System for maintenance recommendation based on maintenance effectiveness estimation," Oct. 23 2018, uS Patent 10,109,122.

[26] A. Cachada, J. Barbosa, P. Leitño, C. A. Gcraldcs, L. Deusdado, J. Costa, C. Teixeira, J. Teixeira, A. H. Moreira, P. M. Moreira *et al.*, "Maintenance 4.0: Intelligent and predictive maintenance system architecture," in *2018 23rd Int. Conf. on ETFA*, vol. 1. IEEE, 2018, pp. 139–146.

[27] J. Wang, L. Ye, R. X. Gao, C. Li, and L. Zhang, "Digital twin for rotating machinery fault diagnosis in smart manufacturing," *International Journal of Production Research*, vol. 57, no. 12, pp. 3920–3934, 2019.

[28] L. Decker, D. Leite, L. Giommi, and D. Bonacorsi, "Real-time anomaly detection in data centers for log-based predictive maintenance using an evolving fuzzy-rule-based approach," in *2020 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–8.

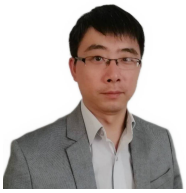
[29] B. Weber, "Predict the unpredictable," https://www.pwc.nl/nl/assets/documents/pwc-predictive-maintenance-4-0.pdf, accessed: 2020-08-04.

[30] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[31] J. Beckmann and H. Heckhausen, *Situational Determinants of Behavior*. Cham: Springer International Publishing, 2018, pp. 113–162.

[32] P. Xiaolan, X. Lun, L. Xin, and W. Zhiliang, "Emotional state transition model based on stimulus and personality characteristics," *China Communications*, vol. 10, no. 6, pp. 146–155, 2013.
[33] M. A. Thornton and D. I. Tamir, "Mental models accurately predict emotion transitions," *Proceedings of the National Academy of Sciences*, vol. 114, no. 23, pp. 5982–5987, 2017.
[34] J. E. Korteling, A.-M. Brouwer, and A. Toet, "A neural network framework for cognitive bias," *Frontiers in Psychology*, vol. 9, p. 1561, 2018.
[35] D. Berlyne, "The vicissitudes of aplopathematic and thelematoscopic pneumatology (or the hydrography of hedonism)," *Pleasure, reward, preference: Their nature, determinants, and role in behavior*, pp. 1–33, 1973.
[36] M. Sudhof, A. Goméz Emilsson, A. L. Maas, and C. Potts, "Sentiment expression conditioned by affective transitions and social forces," in *Proceedings of the 20th ACM SIGKDD*, 2014, pp. 1136–1145.
[37] H. Xiang, P. Jiang, S. Xiao, F. Ren, and S. Kuroiwa, "A model of mental state transition network," *IEEJ Trans. on Electronics, Information and Systems*, vol. 127, no. 3, pp. 434–442, 2007.
[38] P. B. Henryranu, T. Hiroki, and T. Koichi, "Deep time-delay markov network for prediction and modeling the stress and emotions state transition," *Scientific Reports (Nature Publisher Group)*, vol. 10, no. 1, 2020.
[39] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.
[40] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[41] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *ICML*. PMLR, 2015, pp. 1889–1897.
[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Kevin Shen Hoong Ong (M'19) received the B.Eng. honours degree in Electronics Engineering with Management from University of Dundee, United Kingdom, in 2007, and the MEng. degree from School of Computer Science and Engineering, Nanyang Technological University, Singapore, in 2014. He is currently pursuing a Ph.D. degree under the BOSCH-NTU Industrial Ph.D. program at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests are in the areas of Internet of Things (IoT), deep reinforcement learning in predictive maintenance, resource allocation, and edge computing.



Wenbo Wang (S'13-M'17) received his B.S. and M.S. degrees from the School of Automation, Beijing Institute of Technology, Beijing, China. He received the Ph.D. degree in Computing and Information Sciences from Rochester Institute of Technology, Rochester, NY, USA, in 2016. He is currently a Research Fellow with the Faculty of Engineering, Bar Ilan University, Israel. Before that, he was with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include machine learning and mechanism design for multimedia wireless networks and Internet of Things (IoT).

Dusit Niyato (M'09-SM'15-F'17) is currently a professor in the School of Computer Science and Engineering, at Nanyang Technological University, Singapore. He received B.Eng. from King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand in 1999 and Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada in 2008. His research interests are in the areas of Internet of Things (IoT), machine learning, and incentive mechanism design.

Thomas Friedrichs received the Ph.D. degree in Nuclear Physics from Technische Universität Braunschweig, Germany in collaboration with Institute Laue-Langevin - Grenoble, France, in 1998. He is currently the Director of the IT Strategy and Innovation Asia Pacific, Robert Bosch (SEA) Pte Ltd, Singapore.