

Received October 18, 2020, accepted October 22, 2020, date of publication October 27, 2020, date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034218

Algorithm for Autonomous Power-Increase Operation Using Deep Reinforcement Learning and a Rule-Based System

DAEIL LEE, AWWAL MOHAMMED ARIGI, AND JONGHYUN KIM 

Department of Nuclear Engineering, Chosun University, Gwangju 501-709, South Korea

Corresponding author: Jonghyun Kim (jonghyun.kim@chosun.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning under Grant N01190021-06, and in part by the Korean Government, Ministry of Science and ICT under Grant NRF-2018M2B2B1065651.


ABSTRACT The power start-up operation of a nuclear power plant (NPP) increases the reactor power to the full-power condition for electricity generation. Compared to full-power operation, the power-increase operation requires significantly more decision-making and therefore increases the potential for human errors. While previous studies have investigated the use of artificial intelligence (AI) techniques for NPP control, none of them have addressed making the relatively complicated power-increase operation fully autonomous. This study focused on developing an algorithm for converting all the currently manual activities in the NPP power-increase process to autonomous operations. An asynchronous advantage actor-critic, which is a type of deep reinforcement learning method, and a long short-term memory network were applied to the operator tasks for which establishing clear rules or logic was challenging, while a rule-based system was developed for those actions, which could be described by simple logic (such as if-then logic). The proposed autonomous power-increase control algorithm was trained and validated using a compact nuclear simulator (CNS). The simulation results were used to evaluate the algorithm's ability to control the parameters within allowable limits, and the proposed power-increase control algorithm was proven capable of identifying an acceptable operation path for increasing the reactor power from 2% to 100% at a specified rate of power increase. In addition, the pattern of operation that resulted from the autonomous control simulation was found to be identical to that of the established operation strategy. These results demonstrate the potential feasibility of fully autonomous control of the NPP power-increase operation.

INDEX TERMS Nuclear power plant, autonomous operation, power-increase operation, reinforcement learning, asynchronous advantage actor-critic.

ABBREVIATIONS

A3C	Asynchronous Advantage Actor-Critic
AI	Artificial Intelligence
ASICS	Automatic Start-up Intelligent Control System
C - LSTM	Convolutional Long Short-Term Memory Neural Network
CNN	Convolutional Neural Network
CNS	Compact Nuclear Simulator
CPU	Central Processing Unit

DNN	Deep Neural Network
DQN	Deep Q-learning Network
DRL	Deep Reinforcement Learning
GB	GigaByte
GOP	General Operating Procedure
GPU	Graphics Processing Unit
HVAC	Heating, Ventilation, and Air Conditioning
KAERI	Korea Atomic Energy Research Institute
LSTM	Long Short-Term Memory
MAR	MultiAgent System
NPP	Nuclear Power Plant
PID	Proportional-Integral-Derivative
PWR	Presurizer Water Reactor
RCS	Reactor Coolant System

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang .

RL	Reinforcement Learning
RNN	Recurrent Neural Network
RPM	Revolutions Per Minute
SG	Steam Generator

I. INTRODUCTION

Nuclear power plants (NPPs) are highly automated systems that are designed to increase electricity availability, reduce accident risk, and decrease operating costs [1]. Regulatory authorities require the automation of safety systems because these system functions must be exceptionally reliable and promptly executed to ensure public safety. These safety systems require operator intervention only for high-level decision-making or if the automatic system is not functioning correctly [2].

During the power-increase operation, which is typically conducted manually and is also termed the “start-up operation,” the operators increase the power of the reactor to 100% of its electricity production capacity. Compared to full-power operation, the power-increase operation (which is part of the start-up operation) is more prone to human error due to the following factors:

- A significantly increased need for decision-making such as in selecting the power operation target and determining the control strategy based on guidelines from the operating procedures;
- A huge number of manual actions due to extensive maintenance, tests, and monitoring of plant parameters;
- The manipulation of components for which the automatic system and safety function may be disabled;
- An insufficient or incomplete procedure, which may provide only the operational goal without detailed operator’s tasks.

These situations may provide stressful situations or increase the probability of inappropriate manipulation to the crew. Therefore, during the power-increase operation, the potential for human errors is high due to the operator’s significantly increased workload [3]–[7].

One way to decrease operator’s workload is to improve level of automation with more-advanced artificial intelligence (AI) techniques. Hence, AI is an alternative to develop an intelligent controller for power-increase operations in NPPs.

The utilization of AI is a recent trend in increasingly many industrial fields [8]. AI adoption has grown explosively due to increased data processing, along with developments in hardware designs, graphics processing units (GPUs), and methods [9]. Deep learning is one of the most promising new AI methods for a wide range of uses, e.g., extracting high-level features from raw sensor data with numerous variables and facilitating breakthroughs in computer vision and speech recognition. Most deep learning applications require a range of deep neural network architectures, methodologies for training the neural networks, and vast amounts of labeled training data. These advantages of AI have increased interest in applying them to intelligent controllers that would expand NPP automation capabilities.

This study aimed at developing an algorithm for autonomously increasing the reactor power from 2% to 100%. This algorithm with AI method conduces higher-level NPP similarly to the current operation strategy. For performing similar operator’s operation, this algorithm aims at advancing from existing manual controller to intelligent controller. Proposed algorithm can handle the procedure-based operation (as rule-based system) and the operator’s experience (via AI agent).

II. REVIEW OF RELATED STUDIES

First, this paper reviews of previous studies related to the use of deep reinforcement learning (DRL) for the development and application to advanced control systems, and in practice to improve automation in NPPs. Based on the summarized review, the major gaps of this study are identified.

A. DEEP REINFORCEMENT LEARNING

DRL, which is a method for training deep neural networks, provides a mechanism via AI agents that can optimize their control of an environment to realize a specified objective [10]–[13]. The interaction process between the AI agent and the environment can be represented by a closed-loop, which is very similar to the process of human learning [14], [15]. As a result, an AI agent can also develop its own experiences through trial-and-error, as humans do [16] and can perform tasks that a classic controller cannot do. Such actions may include selecting an operation strategy, operating nonlinear systems, making decisions based on current conditions, and optimizing operations [17]–[20].

Due to these characteristics of DRL, DRL is now an essential technology for the development of AI agents and is being used in many industries. Moreover, DRL is becoming a trend in advanced control systems due to increased safety and efficiency [21]. In the power system field, Zhou *et al.* [22] proposed an AI agent that was based on DRL for handling various operating scenarios for the economic dispatch of a combined heat and power system. In an application to wind turbines [23], DRL has been shown to overcome one of the most important disadvantages of the conventional control strategies, which is the tuning of control parameters and lowering fatigue. In energy management, Esmat Samadi *et al.* proposed the use of decentralized multiagent systems (MASs) for integrated grid-connected microgrids. MASs with DRL have shown not only flexible management while considering customer consumption but also a reduced operating cost [24]. Hussain Kazmi *et al.* optimized the energy efficiency of hot water production by using a DRL controller, which could reduce the energy consumption by almost 20% for a set of 32 Dutch houses [25]. Tianshu Wei *et al.* also significantly reduced the energy cost of an HVAC (heating, ventilation, and air conditioning) system by using DRL instead of rule-based and model-based strategies [26]. In another study [27], DRL was adopted in urban rail transit to effectively improve energy management

compared to the genetic algorithms and to provide dynamic programming.

The advantages of DRL for the development and application of advanced control systems through these research trends are briefly summarized as follows:

- Performance improvement compared to conventional control strategies (e.g., reducing operating costs, reducing failures, and increasing energy efficiency);
- Increased flexibility by adaptable control according to demand and change in practice;
- Optimal control to achieve the required goals.

B. HIGH-LEVEL AUTOMATION IN THE OPERATION OF NPPs

Various methods have been presented for applying AI to the tuning of proportional-integral-derivative (PID) controllers, which are widely used in NPPs [28]–[30]. Upadhyaya *et al.* designed an autonomous control system for a space reactor using a PID controller. The controller's parameters were determined using the genetic algorithm, which is an AI method [31]. Bowen *et al.* proposed a two-level hierarchical controller that consisted of a fuzzy controller and a neural-network-based PID controller for application to a multiunit small modular reactor [32]. Some researchers have proposed the use of AI controllers to manage NPPs. Na *et al.* proposed a neuro-fuzzy controller for managing the power distribution without any residual flux oscillation between the upper and lower halves of the reactor core [33]. Arab-Alibeik and Setayeshi proposed a neural adaptive inverse controller for controlling the core power of a PWR reactor. After an emulation of the inverse dynamics of the reactor was obtained by the multilayer neural networks, it was used as a controller [34]. In [35], an adaptive fuzzy control for power tracking in a research nuclear reactor was proposed. The proposed controller could increase the power in a shorter rise time than the PID controller. In [36] and [37], a fuzzy-PID composite controller was proposed and exploited with direct switching between the fuzzy controller and the PID controller for the core power control of a molten salt reactor. More so, Huang *et al.* proposed a fuzzy-adaptive recursive sliding-mode controller that can perform significantly more mildly with less amplitude power in comparison with a PID controller [38]. Boroushaki *et al.* [39], Hatami *et al.* [40], and Khorramabadi *et al.* [41] proposed an intelligent reactor core controller for a load-following operation that applied AI techniques, namely, recurrent neural network and fuzzy logic systems. Ramazan *et al.* [42] proposed a multi-feedback layer neural network, which is a type of recurrent neural network, and presented the proposed controller that can reduce the power increase time compared to a fuzzy controller.

Studies about start-up operation have investigated the use of knowledge-based technology to automate the power-increase operation and, consequently, reduce the operator's burden. Sekimizu *et al.* [43] developed an automation system for start-up operation and sequential control that executes the operating procedure through if-then logic. An automatic

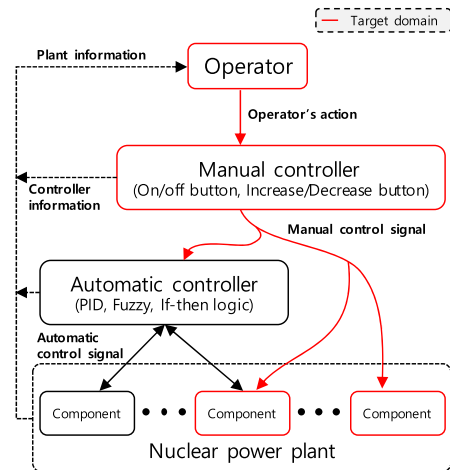


FIGURE 1. Target domain of the proposed power-increase algorithm.

start-up intelligent control system (ASICS), which uses knowledge-based technology and a distributed control system, has also been proposed for controlling a pressurized water reactor (PWR) from the cold shutdown state to 2% reactor power [44].

These studies of practice to improve automation in NPPs are briefly summarized as follows:

- AI techniques had been applied to traditional controllers to tune or identify the optimal parameters of traditional controllers;
- AI techniques tried to replace the traditional control logic at the component level;
- To automatically operate NPPs for start-up, knowledge-based technology was used.

C. SOME GAPS OF RELATED STUDIES

These early studies investigated knowledge-based systems that use if-then logic, which are robust if the logic can be clearly defined. However, for automating power-increase operations in NPPs, these systems have several limitations:

- (1) Transforming many operational tasks into clear if-then logic is challenging, namely, some operating procedure instructions are not sufficiently specific for execution using if-then logic. For instance, an operating procedure may instruct operators to adjust the control rods to increase the reactor power to 20% without specifying the rate of control rod movement;
- (2) Knowledge-based systems poorly handle flexible operations, changes in operating objectives, and nonlinear variables, which may be absent from pre-established knowledge bases. To handle multiple operation object and conduct control functions similar to the operators, advanced AI techniques should be applied in the proposed algorithm;
- (3) Manual controllers, which are operated physically by the plant operators, are not considered in these previous studies. The target domain of this study is the operator's actions with manual controllers as illustrated in Fig. 1.

Therefore, more-advanced artificial intelligence (AI) techniques may be an alternative for the development of an algorithm for power-increase operations in NPPs. In addition, more extensive use of AI techniques must be considered for the realization of autonomous control of higher-level NPP operations [45].

Thus, this paper presents the following:

- A power-increase algorithm for conducting higher-level NPP operations similarly to the current operation strategy;
- A classification and analysis of the operator’s tasks (Decision Making, Discrete Control, and Continuous Control) so that the AI agent can conduct actions for increasing the reactor power and electrical power output based on the current operating procedures, the operator’s primary tasks, and the timeline of operations;
- An algorithm that not only enables the procedure-based operation (which is modeled as rule-based system), but also identifies actions that are typically acquired from the operator’s experience (via interaction between the AI agent and NPPs).

First, the current operating procedures, an operator’s primary tasks, and the timeline of the operations for increasing the reactor power and the electrical power output are analyzed. An algorithm for controlling the components as required for increasing the power, which combines Deep reinforcement learning (DRL) and a rule-based system, is proposed. The operator tasks for which the establishment of clear rules was challenging were implemented using an asynchronous advantage actor-critic (A3C), a kind of DRL method, while a rule-based system was applied to tasks for which clear rules could be developed. Then, an algorithm that combines the A3C agent with the long short-term memory (LSTM) network and the rule-based operation is proposed and trained to determine the power-increase operation strategy. Finally, this paper presents validation results, which demonstrate the applicability of the proposed algorithm.

III. ANALYSIS OF THE OPERATIONAL STRATEGY FOR INCREASING POWER IN NPPs

Current NPP operating strategies were considered in the development of an algorithm for increasing the reactor power from 2% to 100% autonomously. This study analyzed the operating procedures and the timeline of the control tasks during the start-up operation of a reference plant, namely, a Westinghouse 900 MWe PWR. The analysis identified the operator’s major tasks, and the tasks were categorized into automatic and manual actions. The manual actions were further divided into discrete and continuous actions. The operational timeline of the main control systems for increasing power was also analyzed.

A. OVERVIEW OF THE POWER-INCREASE OPERATION

The operation for increasing power from 2% to 100% is the part of the start-up operation that increases the temperature and power to the normal conditions for generating

electricity after reactor refueling or shutdown. During the start-up operation, the operators follow general operating procedures (GOPs) for controlling systems and components. There were six GOPs for the reference plant’s start-up operation [9]:

- Reactor coolant system filling and venting,
- Cold shutdown to hot shutdown,
- Hot shutdown to hot standby,
- Hot standby to 2% reactor power,
- Power operation at greater than 2% power,
- Secondary system heat-up and start-up.

Fig. 2 illustrates the trends of six major parameters during the start-up operation, along with the relevant GOPs. These parameters also serve as milestones for operators in the successful performance of the start-up operation.

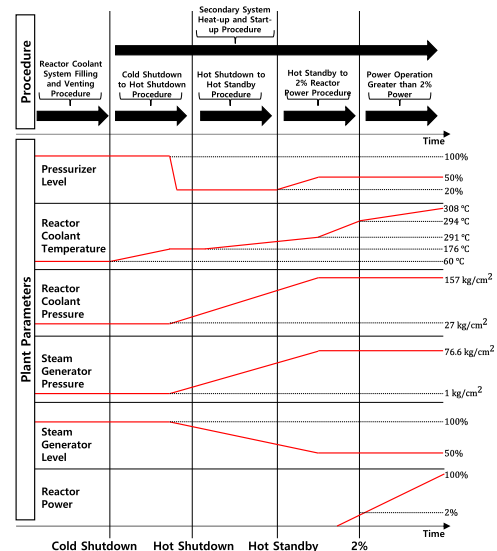


FIGURE 2. The trends of the major parameters for the applicable start-up operation procedures.

To increase the power from 2% to 100%, two GOPs should be applied in the reference plant, namely, “Power operation greater than 2%” and “Secondary system heat-up and start-up”, as presented in Fig. 2. The instructions for increasing the plant load from 2% to 100% are provided in the “Power operation greater than 2%” GOP, while the procedure “Secondary system heat-up and start-up” procedure describes the steps that are necessary for aligning and starting the secondary systems. These GOPs require the operators to operate components, such as the rod controller, turbine load controller, feedwater pumps, condenser pumps, steam generator feedwater valves, and synchronizer, based on the planned rate of power increase. Fig. 3 presents a simplified schematic diagram of the components that are related to the power-increase operation, and the operation’s initial and final conditions are presented in Table 1.

The operators’ tasks in the applicable procedures can be divided into 1) primary system control and 2) secondary system control. When conducting primary system control,

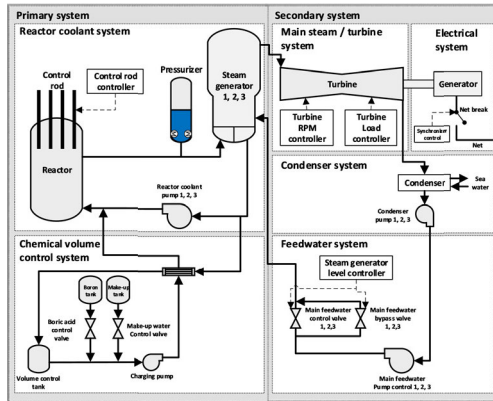


FIGURE 3. Simplified schematic diagram of related components.

TABLE 1. Initial and final conditions of the power-increase operation.

Major parameter	Initial condition	Final condition
Reactor power	2%	100%
Electric power	0 MWe	900 MWe
Reactor coolant system (RCS) average temperature	294 °C	306 °C
Turbine revolutions per minute (RPM)	0	1800 RPM
Turbine load setpoint	0 MWe	900 MWe
Turbine load rate setpoint	0 MWe/min	2 MWe/min
Boron concentration	637 ppm	457 ppm
Rod position	211 Step (A Bank) 95 Step (B Bank) 0 Step (C Bank) 0 Step (D Bank)	228 Step (A Bank) 228 Step (B Bank) 228 Step (C Bank) 220 Step (D Bank)
Rod controller	Manual	Auto
Steam generator controller	Manual	Auto
Feedwater pump 1	On	On
Feedwater pump 2	Off	On
Feedwater pump 3	Off	On
Condenser pump 1	On	On
Condenser pump 2	Off	On
Condenser pump 3	Off	On
Synchronous connection	Disconnected	Connected

the operators withdraw the control rods (reactor coolant system, Fig. 3) and manipulate the boron concentration (chemical volume control system, Fig. 3). At the beginning of the operation for stably increasing the power to 2%, the operators withdraw all control rods to the 100% position, which is the final condition, as specified in Table 1, and subsequently increase the boron concentration to maintain the reactor power at 2%. Once all the control rods have been withdrawn, the operators do not manipulate them further, and they reduce the boron concentration to increase the power from 2% to 100% by increasing the volume of the water from the make-up tank.

The rate of power increase (percent power per hour) is determined by considering the reactor cooling system (RCS)

average temperature and the reference temperature. The reference temperature is the desired RCS temperature, which is predefined based on the current turbine load, while the RCS average temperature is the actual temperature in the primary side [46]. According to the procedure, during the power increase from 2 to 100%, the difference between the reference temperature and the RCS average temperature should be maintained within ± 1 °C. This is only a recommendation and is not mandatory.

Operators must control several components of the secondary system. First, they increase the turbine speed to 1800 revolutions per minute (RPM) using the turbine RPM controller (the main steam/turbine system in Fig. 3). When the turbine and the reactor power reach 1800 RPM and 15%, respectively, the operators close the breaker to connect the generator to the grid and synchronize the frequencies (the electrical system in Fig. 3). In addition, the operators increase the turbine load setpoint, start the feedwater pumps, and start the condenser pumps concurrently with the reactor power increase in the primary system. The primary and secondary systems must be controlled harmoniously to avoid a reactor trip.

B. TASK ANALYSIS OF THE POWER-INCREASE OPERATION

Based on a review of the “Power operation greater than 2%” and “Secondary plant heat-up and start-up” procedures, a task analysis was conducted to identify the tasks that should be automated by the algorithm that is proposed in this study. As presented in Table 2, this analysis identified a total of 21 control actions that are performed by the operators according to these procedures. Only the control-related actions were extracted for the development of the algorithm, although the procedures also provide monitoring actions, e.g., “confirm the RCS temperature is above 200 °C.”

These actions were also categorized into three task types: Decision Making, Continuous Control, and Discrete Control. Decision Making task determines the rate of power increase; the subsequent control actions depend on this rate, although it does not include any control action. The continuous controls in this study adjust component states over a range to realize specified target values for the given parameters, and the rules that govern the necessary adjustments cannot be described with simple logic. For example, the operators adjust the RCS boron concentration to manipulate the power level. In contrast, a discrete control involves the direct setting of a target value based on a binary condition, as in if-then logic. An example of a discrete control is as follows: if the power level is 10%, then the turbine is set to 1800 RPM. The next section proposes an algorithm that can perform these actions.

C. TIMELINE OF THE POWER-INCREASE OPERATION

The timeline of the power-increase operation was analyzed to develop a normative operational strategy. This analysis considered the GOP’s operational rules and the practical operational practices, which were determined from an interview with a senior reactor operator who works at a reference plant.

TABLE 2. Operational tasks for increasing the reactor power.

Step	Task Type	Action
1	Decision Making	Determine the rate of power increase in %/h
2	Continuous Control	Withdraw all control rods to the position of 100% reactor power while maintaining the reactor power at 2% through boration.
3	Continuous Control	If all the control rods are withdrawn, increase the reactor power from 2% to 6%–10% by reducing the boron concentration.
4	Discrete Control	If the reactor power is 10%, the turbine RPM setpoint is 1800 RPM.
5	Discrete Control	If the reactor power exceeds 10%, the acceleration setpoint is 2 MWe/min.
6	Continuous Control	Adjust the boron concentration to increase the reactor power from 10% to 20%.
7	Discrete Control	If the reactor power is between 10% and 20%, the load setpoint is 100 MWe.
8	Discrete Control	If the turbine RPM is 1800 RPM and the reactor power exceeds 15%, push the net-breaker.
9	Discrete Control	If the reactor power is 20%, start condenser pump #2.
10	Continuous Control	Adjust the boron concentration to increase the reactor power from 20% to 100%.
11	Discrete Control	If the reactor power is between 20% and 30%, the load setpoint is 200 MWe.
12	Discrete Control	If the reactor power is between 30% and 40%, the load setpoint is 300 MWe.
13	Discrete Control	If the reactor power is 40%, start main feedwater pump #2.
14	Discrete Control	If the reactor power is between 40% and 50%, the load setpoint is 400 MWe.
15	Discrete Control	If the reactor power is between 50% and 60%, the load setpoint is 500 MWe.
16	Discrete Control	If the reactor power is 50%, start condenser pump #3.
17	Discrete Control	If the reactor power is between 60% and 70%, the load setpoint is 600 MWe.
18	Discrete Control	If the reactor power is between 70% and 80%, the load setpoint is 700 MWe.
19	Discrete Control	If the reactor power is 80%, start main feedwater pump #3.
20	Discrete Control	If the reactor power is between 80% and 90%, the load setpoint is 800 MWe.
21	Discrete Control	If the reactor power is between 90% and 100%, the load setpoint is 900 MWe.

Fig. 3 presents the timeline that was developed, which associates the desired operations with the reactor and electric powers, RCS temperatures and their differences from the

reference temperature, and the control of related systems, such as the steam generator (SG) level, control rods, turbines, valves, and pumps.

The power-increase operation is divided into two operational ranges: 1) maintaining the reactor power at 2% and 2) increasing the reactor power from 2% to 100%. The objective of the first operational range is to adjust the positions of all control rods (Fig. 4 (d)) to 100% while maintaining the reactor power at 2% (Fig. 4 (a)); the average temperature is also maintained because it depends on the reactor power (Fig. 4 (b)). As the control rods are withdrawn, the reactor power increases, and increasing the boron concentration in the RCS reduces the reactor power. To maintain the reactor power at 2%, a boric acid-water solution is injected into the RCS, as illustrated in Fig. 4 (c).

The objective of the second operational range is to increase the reactor power from 2% to 100%, as represented by the red line in Fig. 4 (a). The operators determine the rate of the power increase (%/h); the power is increased by reducing the boron concentration in the RCS using make-up water (Fig. 4 (c)). The electric power is also increased to 100% by following a load setpoint that is increased stepwise. The RCS average temperature increases from 294 °C to 306 °C, as illustrated in Fig. 4 (b). The difference between the RCS average temperature and the reference temperature should be maintained within ± 1 °C, as represented by the gray area in Fig. 4 (b). This condition is applied after the start of the electrical power generation because the reference temperature is calculated based on the electrical power.

To increase the reactor power, the operators manipulate seven systems, as illustrated in Fig. 4 (e). As described in Table 2, they withdraw the control rods and manipulate the boron concentration continuously, which corresponds to Steps 2, 3, 6, and 10. At 10% reactor power, in Steps 4, 5, and 7, the turbine RPM, acceleration setpoint, and load setpoint are adjusted to 1800 RPM, 2 MWe/min, and 100 MWe, respectively. Subsequently, the operators adjust the load setpoint with every 10% increase in the reactor power (Steps 11, 12, 14, 15, 17, 18, 20, and 21). At 15% reactor power, the plant and the grid are synchronized (Step 8). At 20% reactor power, condenser pump #2 is started (Step 9); condenser pump #1 is already running. Condenser pump #3 is started at 50% reactor power (Step 16). Main feedwater pumps #2 and #3 are started at reactor powers of 40% (Step 13) and 80% (Step 19), respectively; main feedwater pump #1 is already running. This study applies the pre-established automatic control algorithm for the SG level control.

IV. DEVELOPMENT OF AN ALGORITHM FOR POWER-INCREASE CONTROL

This paper presents an algorithm that employs a rule-based system and deep reinforcement learning to facilitate the autonomous increase of NPP power from 2% to 100% by controlling several systems. Fig. 5 illustrates the structure of the proposed algorithm, which consists of two

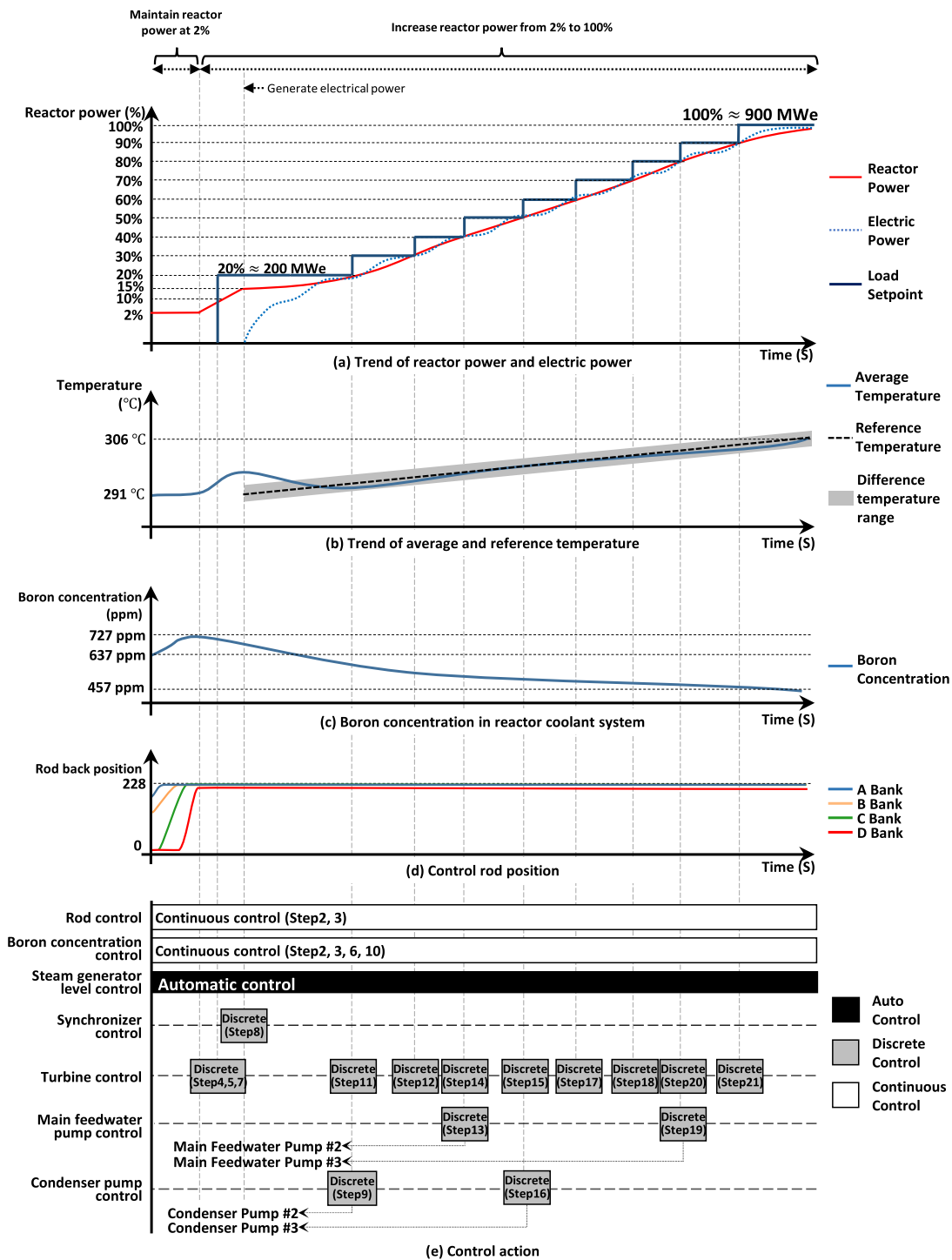


FIGURE 4. Timeline for increasing the reactor power from 2% to 100%.

modules: 1) a discrete control module and 2) a continuous control module. The discrete control module directs the synchronization, turbine, main feedwater pump, and condenser pump controls, for which rule-based systems can be developed based on the operating procedures.

The continuous control module dictates the adjustment of the control rods and the RCS boron concentration. The associated procedures do not specify rules for the operators; e.g., they do not specify the number of steps in which the control rod should be withdrawn or the volumes of

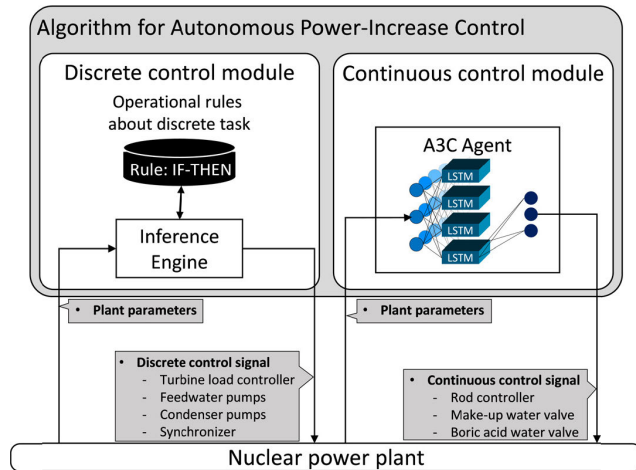


FIGURE 5. Overview of the algorithm for the power-increase operation.

make-up or boric acid water that should be added. The procedures specify only the objective of the control activity, e.g., “increase the power to 20% by altering the control rod position or RCS boron concentration.”

Deep reinforcement learning was deemed suitable for use as the continuous control module. A neural network and a training algorithm are selected by considering the characteristics of the operational steps in NPPs. The types of control for NPPs are regulatory control (e.g., adjustment of valve position) and discrete control (e.g., on/off control). For discrete control, the set-point and operating conditions are specified in detail in the operating procedure. Operators can conduct discrete control according to rules that are specified in the operating procedures. In contrast, only operational target values are provided for regulatory control. Accordingly, regulatory control is based on the operator’s experience, which includes monitoring previous and current plant conditions. The target of the continuous control module is the regulatory control. Thus, this study attempted to implement controls in accordance with the operator’s behavioral pattern through trial and error using a long short-term memory (LSTM) and an asynchronous advantage actor-critic (A3C) algorithm.

(1) This study used a LSTM network, a kind of recurrent neural network (RNN), by considering the characteristics of the plant parameters. The trends of the plant parameters are well known to be the same as that of time series data. To extract and analyze meaningful information, e.g., the timing of an AI agent’s action, from time-series data, it is important to identify the correlations between previous and current data. The output of an LSTM can be calculated by considering previous data, in contrast to other neuronal networks such as convolutional neural networks and vanilla feedforward neural networks. Moreover, LSTM not only stores the values that are calculated from the previous time data in the LSTM cell but also considers previously saved values when calculating the next time data. The author’s previous studies showed that the LSTM can support well the operation of nuclear systems [3], [5] as well as the diagnosis of events [47], [48].

Moreover, to better support the selection of the LSTM neural network, this study added Appendix to compare the performance of other neural networks such as deep neural network (DNN), convolutional neural network (CNN), LSTM, and C-LSTM(CNN + LSTM).

(2) An asynchronous advantage actor-critic (A3C) algorithm was quickly trained in the specified domain. The A3C algorithm is well known for fast training due to parallel actor-learners that are based on the central processing unit’s (CPU’s) multiple threads and the asynchronous network update. This study used a nuclear simulator to test and train an AI agent. This simulator does not recommend calculation acceleration with a stable calculation performance. As a result, the AI agent takes more than 14 hours per episode to train the entire power increase operation. To solve this problem, we not only built multiple environments but also applied a parallel training algorithm, namely, A3C.

The goal of the continuous control module is to select actions necessary to meet the operational goals of the sequential plant states. The continuous control module with A3C algorithm can find an operational path in parallel. An operational path is a set of actions for controlling a component to achieve flexible operating goals that are assigned by the operators. A reward algorithm was developed for training the agent, and an LSTM network was used for selecting the actions necessary to meet the operational goals of the sequential plant states.

A. DESIGN OF THE DISCRETE CONTROL MODULE USING IF-THEN LOGIC

A rule basis for discrete control was developed for the synchronizer, turbine, main feedwater pump, and condenser pump controls by transforming the operating procedures into if-then rules, which are presented in Table 3.

The tasks that are identified as discrete controls in Table 2 were analyzed and categorized into four functions based on the controlled system, and the applicable rules were extracted from the procedures’ task instructions. The inputs and outputs that were required for the module to control the tasks were identified. An input is a plant parameter that must be obtained to correctly determine the control action that is needed for accomplishing a task, while an output is the control action that will be performed as a result.

B. DESIGN OF THE CONTINUOUS CONTROL MODULE USING THE A3C AGENT

The A3C agent for continuous control aims at managing the reactor power by manipulating the control rods and boron concentration, and, if fully trained, can manage the reactor power based on a specified rate of power increase and the obtained plant parameters. The A3C agent’s strategies relate to three operational strategies: increase power, decrease power, and stay.

Fig. 6 illustrates the overall structure of the A3C agent for continuous control, which consists of a reward algorithm and an LSTM network model. The reward algorithm evaluates

TABLE 3. Discrete control module if-then rules for increasing the reactor power from 2% to 100%.

Function	Rule Number(s)	If-then Rule	Input(s)	Output(s)
Synchronizer control	1	If the turbine RPM is 1800 RPM and the reactor power is greater than 15%, push the net-breaker button.	Reactor power, Turbine RPM	Net-breaker button control
	2	If the reactor power is 10%, the turbine RPM setpoint is 1800 RPM.	Reactor power, Turbine RPM	Turbine RPM setpoint control
Turbine control	3	If the reactor power is greater than 10%, the acceleration setpoint is 2 MWe/min.	Turbine acceleration	Turbine acceleration setpoint control
	4	If the reactor power is between 10% and 20%, the load setpoint is 100 MWe.	Reactor power, Load setpoint	Load setpoint control
	5–11
	12	If the reactor power is between 90% and 100%, the load setpoint is 900 MWe.	Reactor power, Load setpoint	Load setpoint control
Main feedwater pump control	13	If the reactor power is 40% and the state of the main feedwater pump 1 is “activated,” start main feedwater pump 2.	Reactor power, Main feedwater pumps 1 and 2 states	Main feedwater pump 2 control
	14	If the reactor power is 80% and the state of main feedwater pump 2 is “activated,” start main feedwater pump 3.	Reactor power, Main feedwater pumps 2 and 3 states	Main feedwater pump 3 control
Condenser pump control	15	If the reactor power is 20% and the state of condenser pump 1 is “activated,” start condenser pump 2.	Reactor power, Condenser pumps 1 and 2 states	Condenser pump 2 control
	16	If the reactor power is 50% and the state of condenser pump 2 is “activated,” start condenser pump 3.	Reactor power, Condenser pumps 2 and 3 states	Condenser pump 3 control

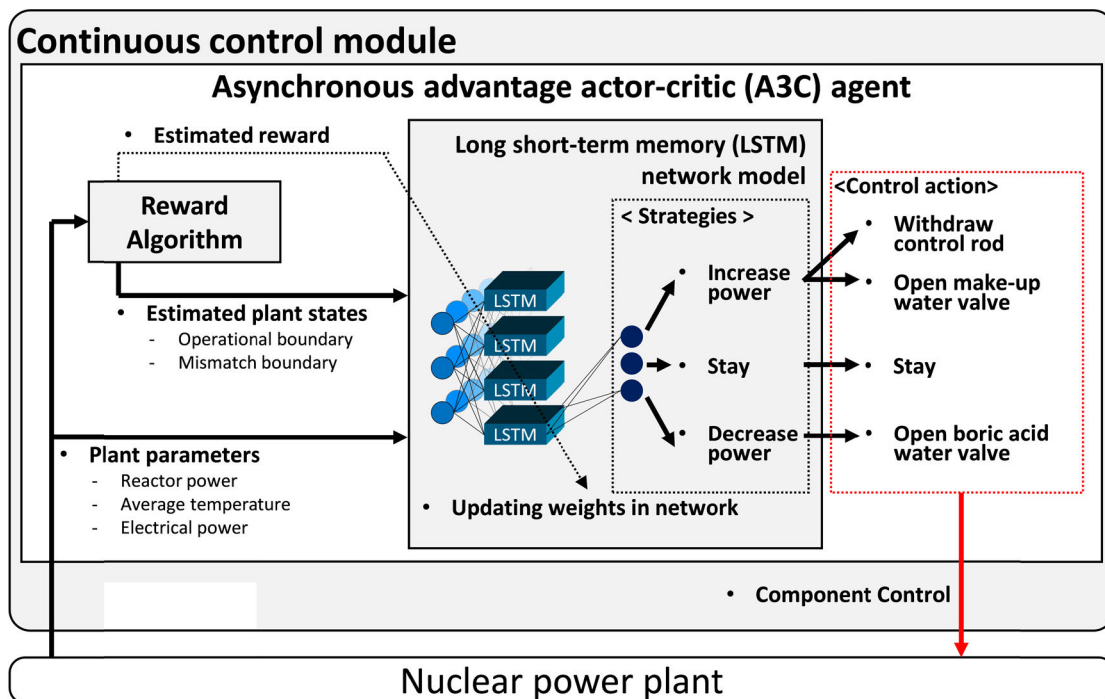


FIGURE 6. Overview of the continuous control module.

the obtained plant parameters to determine whether and the degree to which the prior operation or action of the A3C

agent was successful, and this reward is used to update the weights in the LSTM network model. The LSTM network

model generates an operational strategy using the obtained and evaluated plant parameters. Then, the A3C agent selects the option that is associated with the highest probability value from among the available outputs of the LSTM network: increase, decrease, or stay.

The operational strategies comprise the control actions that are required for realizing the objective of each strategy. For example, for the “stay” strategy, the A3C agent stops manipulating components, and the boric acid water valve is opened to increase the boron concentration and, therefore, decrease the reactor power. The strategies for “power increase” consist of two control actions; the A3C agent withdraws the control rods and changes the control action to the opening of the make-up water control valve to reduce the boron concentration.

1) BACKGROUND OF THE A3C

Reinforcement learning (RL) is a method for training an agent through its interaction with the environment [10], [49]–[51]. The agent interacts with the environment in a series of independent episodes, each of which comprises a sequence of turns. One episode consists of several discrete time steps, $t=0,1,2,3,\dots$. At each time step (t), the agent receives a state (s_t) from the environment. Then, the agent selects an action (a_t) from a set of possible actions based on its policy (π). The policy is a mapping from states (s_t) to actions. The environment provides the next state (s_{t+1}) and a reward (r_t) for the action (a_t) of the agent. Through this interaction with the environment, the agent is trained to maximize the returned reward that is associated with the specified state (s_t) from the environment. Through this trial-and-error process, the agent determines the optimum policy for realizing the specified operational objective.

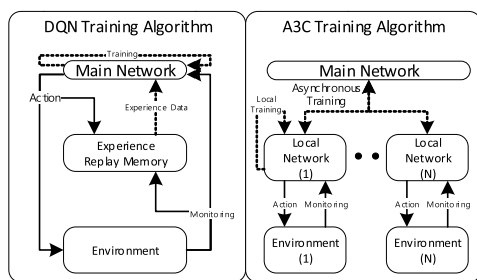


FIGURE 7. DQN and A3C training algorithms.

This study utilizes A3C, which is a type of DRL method, to reduce the agent training time for the continuous control module. Although the deep Q-learning network (DQN) is a well-known basic model of DRL, slow training speed and biased actions are problematic. To address these issues, A3C utilizes parallel actor-learners that are based on the central processing unit’s (CPU’s) multiple threads and the asynchronous network update, while DQN utilizes one agent on one CPU. Fig. 7 illustrates the A3C and DQN training algorithms. A3C replaces the experience memory with the local network memory to reduce the interactions

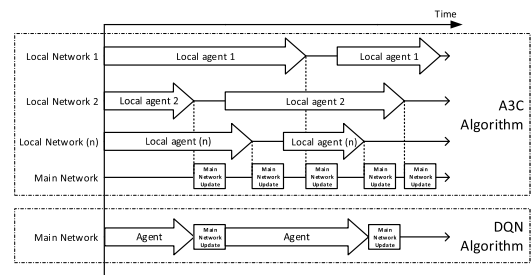


FIGURE 8. Agent’s weight update process.

between the collected training datasets. In addition, A3C utilizes multiple agents in the multiple simulations for training an agent that has a local neural network [52]. In A3C, each local network asynchronously updates the main network at regular intervals. In this asynchronous approach, after collecting a short memory (which is called a mini-batch) of data points, each of the local networks computes gradients and uses them to update the weights [53]. This update process increases the training speed by providing training datasets that consist of pairs of various actions that correspond to similar states. As illustrated in Fig. 8, the A3C agent updates the network’s weights more frequently than the DQN agent.

2) DESIGN OF THE REWARD ALGORITHM

In DRLs, the reward is an essential element that is used to update the weights of the A3C agent; learning by the agent is associated with updating the weights of the network to maximize the accumulative reward [13]. The reward algorithm evaluates the agent’s behavior based on a specified state in the environment to determine the reward. Therefore, the reward algorithm guides the agent to obtain a high accumulative reward in the target domain [54]. To find the best operational path, the use of operational guidelines or boundaries is a suggested for designing a reward algorithm [55]. Furthermore, if the operational goal is more than one, like in the multi-objective problems, Garduno-Ramirez and Lee [56] proposed defining the upper and lower boundaries for each operational goal. In this study, the specified operational objectives were used to design the reward algorithm for increasing the reactor power.

This study proposes a reward algorithm that is designed for training the proposed A3C agent to increase the reactor power. It has two reward criteria, which are based on the reactor power and the average temperature. Fig. 9 presents the criteria for providing a reward via the proposed reward algorithm. The first reward criterion is related to the reactor power. As illustrated in Fig. 8, two bandwidths were applied. While maintaining the reactor power at 2% (the blue area in Fig. 9), the reward boundary was defined as $\pm 1\%$ of the reactor power, namely, 1% to 3%. During the power increase after reaching 2% reactor power, the bandwidth was determined by the following linear equations that were based on the pre-determined rate of power increase (the red area in Fig. 9). The upper boundary was 3% at 2% reactor power and 110% at 100% reactor power, while the lower boundary was

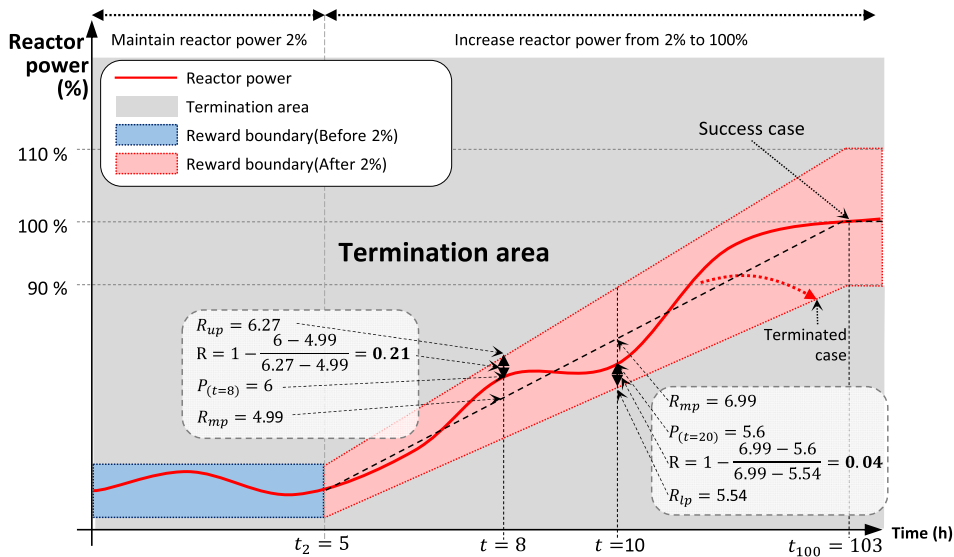


FIGURE 9. Power reward for the A3C agent.

1% at 2% reactor power and 90% at 100% reactor power.

$$\begin{aligned} \text{End of operation time}(t_{100}) \\ = t_2 + \frac{100 - 2}{Pr} \end{aligned} \quad (1)$$

$$\begin{aligned} \text{Upper boundary} \\ = \begin{cases} 3 & (t_2 \geq t) \\ \frac{100 - 3}{t_{100} - t_2}(t - t_2) + 3 & (t_{100} \geq t > t_2) \\ 110 & (t > t_{100}) \end{cases} \end{aligned} \quad (2)$$

$$\begin{aligned} \text{Lower boundary} \\ = \begin{cases} 1 & (t_2 \geq t) \\ \frac{90 - 1}{t_{100} - t_2}(t - t_2) + 1 & (t_{100} \geq t > t_2) \\ 90 & (t > t_{100}) \end{cases} \end{aligned} \quad (3)$$

- Pr: Predefined rate of power increase (% /h)
- t: Time
- t₂: Time when all rods are 100% withdrawn
- t₁₀₀: End of operation time

The power reward was calculated as the difference between the current power at time t and the most desirable power, which was the predefined power at that time and is represented by the dashed line in the center of the reward boundary in Fig. 9. The power reward was calculated via Eq. 4 by using a normalized value of the distance. The reward was maximal, namely, 1, when the current power was equal to the predefined power, while it was 0 when the current power was located on the upper or lower boundary. For instance, at t = 8 h in Fig. 8, when the reactor power increased from 2% at 5 h to 100% at 103 h at a 1%/h rate of increase, the reactor power, the predetermined power that was based on the rate of power increase, and the upper boundary were 6%, 4.99%, and 6.27%, respectively. The resulting reward was

0.21 by $R=1 - (6 - 4.99)/(6.27 - 4.99)$. Similarly, at t = 10 h and P = 5.6%, the reward was 0.04, as presented in Fig. 9.

If the power moved outside the boundary, the training was terminated. In addition, the agent stopped the training when it realized the objective of the operation, namely, when the reactor power was 100%.

$$\begin{aligned} \text{Power reward}(0 \sim 1) \\ = \begin{cases} 0 & (P > R_{up}) \\ 1 - \frac{P - R_{mp}}{R_{up} - R_{mp}} (R_{up} \geq P > R_{mp}) & \\ 1 & (P = R_{mp}) \\ 1 - \frac{R_{mp} - P}{R_{mp} - R_{lp}} (R_{mp} > P \geq R_{lp}) & \\ 0 & (P < R_{lp}) \end{cases} \end{aligned} \quad (4)$$

- P: Current power at time t (%)
- R_{mp}: Middle of power reward boundary, i.e., predetermined power at time t
- R_{up}: Upper power reward boundary
- R_{lp}: Lower power reward boundary

The second reward criterion relates to the difference between the average temperature and the reference RCS temperature that is provided by the GOP. This reward represents that the rule that the average RCS temperature should be controlled by the agent to within ± 1 °C of the reference RCS temperature (the gray area in Fig. 10). Since the reference temperature is calculated based on the current turbine load (MWe), the upper and lower limits of this reward boundary are calculated after the electrical power generation has begun.

Similar to the power reward, the temperature reward was also calculated via Eq. 5 based on the difference between the current temperature at time t and the most desirable temperature, namely, the reference temperature.

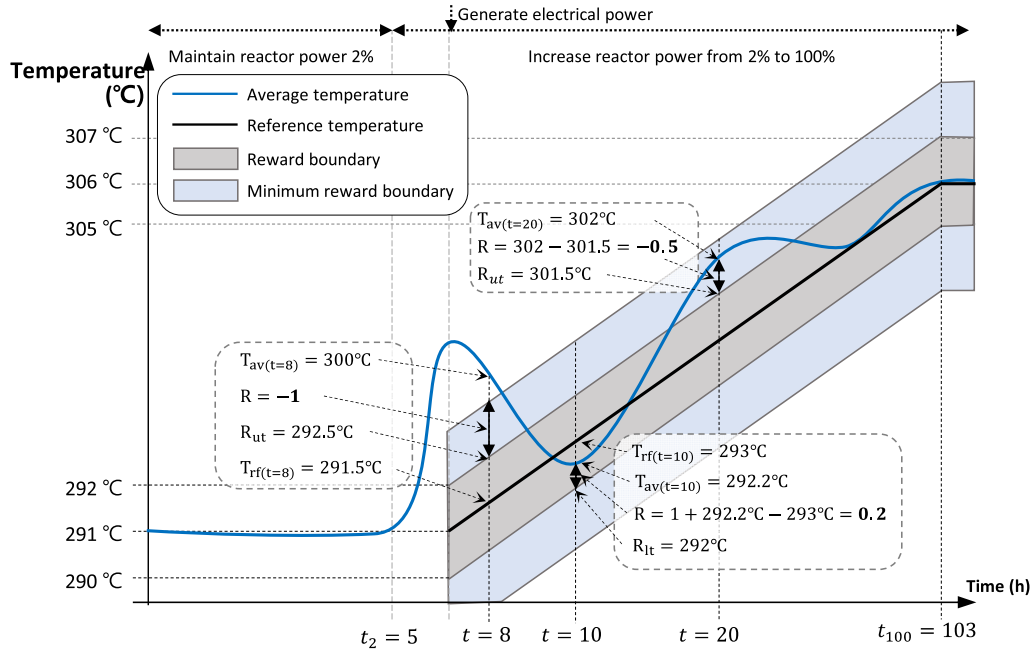


FIGURE 10. Temperature reward for the A3C agent.

The maximal reward, namely, 1, was obtained when the average RCS temperature was equal to the reference temperature. In contrast to the power reward, if the average RCS temperature moved outside the boundary, the training was not terminated; instead, the reward had a negative value that was proportional to the distance from the closest boundary, with -1 being the lowest possible value.

Temperature reward ($-1 \sim 1$)

$$= \begin{cases} -1 & (R_{ut} + 1 < T_{av}) \\ -T_{av} + R_{ut} & (R_{ut} + 1 \geq T_{av} > R_{ut}) \\ 1 - T_{av} + T_{rf} & (R_{ut} \geq T_{av} > T_{rf}) \\ 1 & (T_{av} = T_{rf}) \\ 1 + T_{av} - T_{rf} & (T_{rf} > T_{av} \geq R_{lt}) \\ T_{av} - R_{lt} & (R_{lt} - 1 \leq T_{av} < R_{lt}) \\ -1 & (R_{lt} - 1 > T_{av}) \end{cases} \quad (5)$$

- T : Average RCS temperature at time t
- T_{rf} : Middle of temperature reward boundary, i.e., the reference temperature at time t
- R_{ut} : Upper temperature reward boundary ($T_{rf} + 1$) at time t
- R_{lt} : Lower temperature reward boundary ($T_{rf} - 1$) at time t

As shown in Fig. 10, when the average RCS temperature was between the upper and lower boundaries, a positive reward was returned and was inversely proportional to the distance from the reference temperature (as shown at $t = 10$ h in Fig. 10). Outside this boundary and up to a difference of $\pm 2^\circ\text{C}$, a negative reward was given proportional to the distance to the closest boundary (as shown at $t = 20$ h in Fig. 9). If the

temperature difference was greater than 2°C , the reward was -1 .

The total reward was calculated as the arithmetic mean of the power and temperature rewards, as expressed in Eq. 6. The agent conducted the training to obtain the largest total reward for each episode and, in the process, was incentivized to shift the reactor power and the average RCS temperature to the middle values of the reward boundaries. The episode continued until the reactor power reached 100% or moved outside the reward boundary.

$$\text{Total reward}(-1 \sim 1) = \frac{\text{power reward} + \text{temperature reward}}{2} \quad (6)$$

3) LSTM NETWORK MODELING

This study used LSTM cells to generate the operational strategies of the continuous control module. LSTM cells are an advanced form of RNNs and can calculate time-series data [47], [48], [57]. An RNN can represent a dynamic system naturally, capture the dynamic behavior of the system, and extract the information features that are related to the dynamic system in its hidden layer [58]. However, when the network has five or more layers, an RNN may encounter a gradient vanishing problem [59], wherein the gradient value becomes too large or vanishes at an exponential rate to zero when updating the weights in many layers. This imposes limitations on the dataset for the long-term memory within an RNN; LSTM cells have been proposed to address this problem.

Fig. 11 illustrates the structure of an LSTM cell. Each LSTM cell is composed of units, namely, “constant error

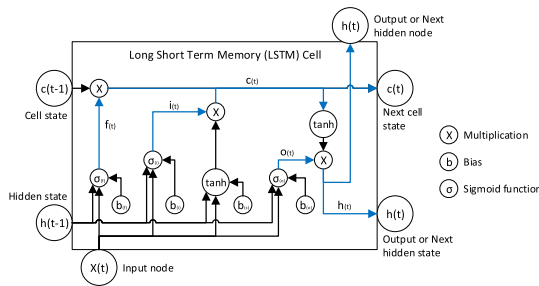


FIGURE 11. Structure of an LSTM cell.

carousels” (CECs), that retain the state across time-steps and three types of specialized gate units (input, output, and forget gates) [60]. Eq 7–11 describe the output from each gate unit in an LSTM cell:

$$i_t = \sigma(x^t W_{xi} + h_{t-1} W_{hi} + b_i) \quad (7)$$

$$f_t = \sigma(x^t W_{xt} + h_{t-1} W_{hf} + b_f) \quad (8)$$

$$o_t = \sigma(x^t W_{xo} + h_{t-1} W_{ho} + b_o) \quad (9)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(x^t W_{xc} + h_{t-1} W_{hc} + b_c) \quad (10)$$

$$h_t = o_t \circ \tanh(c_t) \quad (11)$$

where x^t is the input to the LSTM cell, and i_t , f_t , o_t , c_t , and h_t are the input gate, forget gate, output gate, cell state, and output of the LSTM cell, respectively, at the current time step t . W_{xi} , W_{xf} , and W_{xo} are the weights between the input layer and the input gate, between the input layer and the forget gate, and between the input layer and the output gate, respectively. W_{hf} , W_{hi} , and W_{ho} are the weights between the hidden recurrent layer and the forget gate, between the hidden recurrent layer and the input gate, and between the hidden recurrent layer and the output gate, respectively, of the memory block. Finally, b_i , b_f , and b_o are the additive biases of the input, forget, and output gates, respectively. The set of activation functions consists of the sigmoid function, elementwise multiplication (the inner product of a vector, \circ), and the hyperbolic activation function. At time step 0, o_0 and h_0 are initialized as zero matrices.

Fig. 12 illustrates the proposed LSTM network of the continuous control module’s A3C agent for producing an operational strategy (increase, decrease, or stay). The final control action of the continuous control module is selected based on the reactor power and the operational strategy. Each operational strategy maps to the required control action. For example, the decrease strategy is mapped to the opening of the boric acid water valve. If the output strategy of the LSTM network is “stay,” the A3C agent does not control the component. In the increase strategy, the A3C agent selects a control according to the current operational objective:

- Withdraw the control rod (when maintaining the reactor power at 2%) or

- Open the make-up water valve (when increasing the reactor power from 2% to 100%).

The proposed LSTM network model consists of an input layer, an LSTM layer, and an output layer. The sizes of the input and output layers can be defined based on the numbers of plant parameters and control actions, respectively. The number of LSTM cells is determined by the time window.

The input layer of the investigated LSTM network had a 10-step time window, which considered the trends of plant parameters by exploiting the collected historical data. The historical data were sampled from the simulator every 30 s to optimize the dataset size; the trends that were observed when the data were collected every second did not differ significantly. The A3C agent used the current and previous states as a two-dimensional array and as a training dataset, which included the plant parameters for 300 s. At each time window, the LSTM network used eight input parameters, namely, four plant parameters (reactor power, average temperature, reference temperature, and electric power) and four variables that represented the distances of the current power and average RCS temperature from their upper and lower boundaries.

At the LSTM network’s output layer, the probability of each operational strategy was generated using a softmax function, which can map a network’s output to a probability distribution between 0 and 1; the sum of the generated output values is one. If the A3C agent selected the strategy with the highest probability among the operational strategies, it received a large reward or realized the operational objective. Finally, the A3C agent selected a control action based on the selected operational strategy. The detailed structure and hyperparameters of the LSTM network were determined as illustrated in Fig. 12 through an experimental optimization.

V. EXPERIMENTS

A. TRAINING ENVIRONMENT

A compact nuclear simulator (CNS) was used as a real-time testbed for training and validating the proposed autonomous power increase algorithm. The CNS was originally developed by the Korea Atomic Energy Research Institute (KAERI) [61] using a Westinghouse 900 MWe, three-loop PWR as the reference. Fig. 13 shows the display for the chemical and volume control system in the CNS.

Fig. 14 shows the A3C agent training environment structure, which consists of four desktop computers—one main computer and three sub-computers. One main agent and sixty local agents for implementing the proposed algorithm were installed on the main computer. The CNS was installed on the three sub-computers, which had Intel Core™ i7-8700K processors and 16 GB of memory. Each sub-computer could run 20 CNS simulations at a time; therefore, a total of 60 simulations could be conducted simultaneously. The A3C global

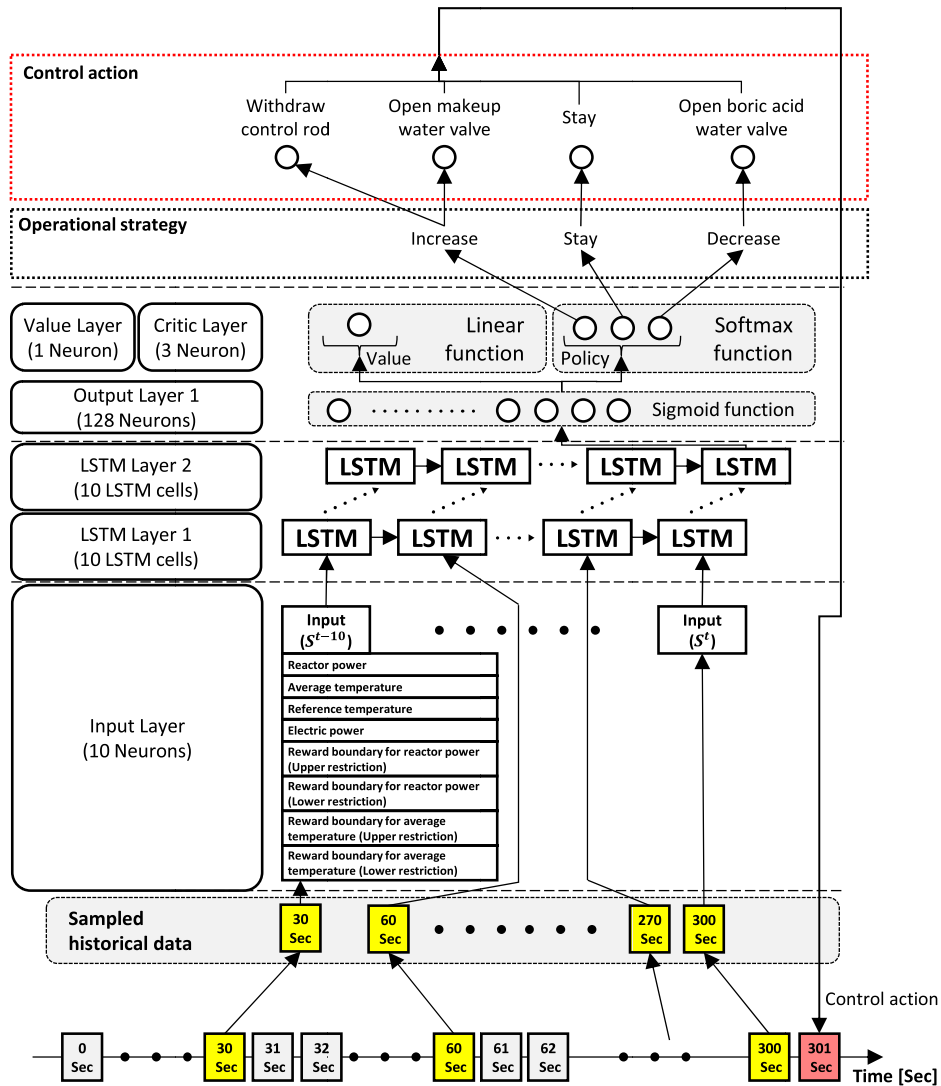


FIGURE 12. The structure of the LSTM network for the A3C agent.

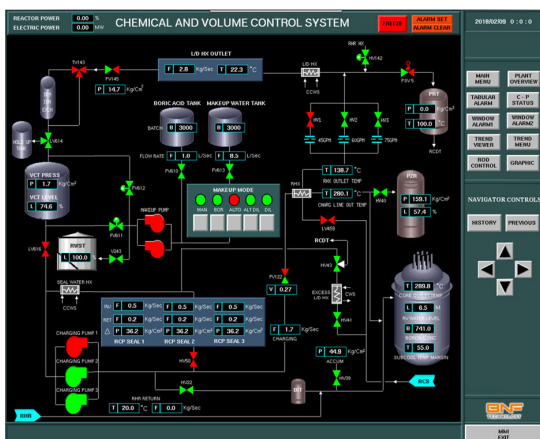


FIGURE 13. Chemical and volume control system in the CNS.

network was trained using two Nvidia GeForce GTX 1080 Ti graphics cards, while the A3C training algorithm was trained using 60 threads of Intel Core X-Series™ i7-7820X CPUs.

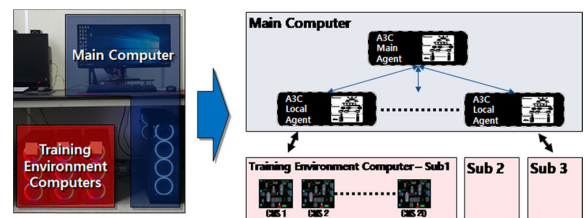


FIGURE 14. Structure of the training environment for the A3C agent.

The A3C agent was developed based on the Python programming language with the TensorFlow and Keras machine learning libraries.

B. TRAINING AND STABILITY FOR THE ENTIRE POWER-INCREASE OPERATION

For a complete (from 0% to 100%) power-increase operation at a rate of 3%/h, the A3C agent was trained in 8800 episodes. The A3C agent training was complete when the average

maximum probability converged to a specified value or when the value became stable.

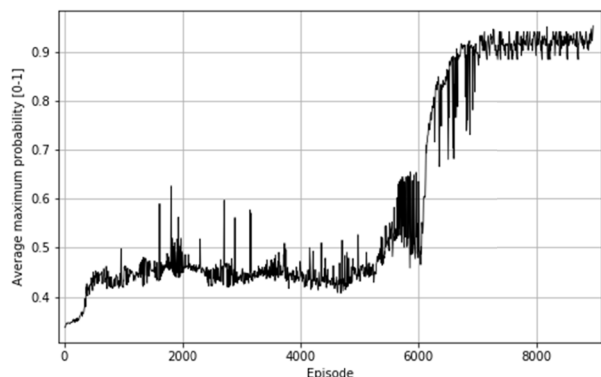


FIGURE 15. Average maximum probability per episode for the A3C LSTM network.

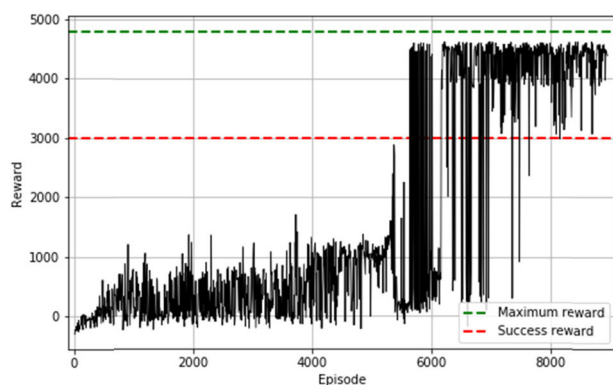


FIGURE 16. Rewards obtained by the A3C LSTM network.

Fig. 15 presents the trend in the average maximum output probability per episode over time. The A3C network approached a stable probability (larger than 0.9) after approximately 7500 episodes. Fig. 16 shows the trend of the rewards that were obtained by the A3C agent as the number of episodes increased. In one episode, the theoretical maximum cumulative reward during the entire power-increase operation was 4800 (the green dashed line in Fig. 16); this is because the largest reward for a training dataset was 1, and the total number of datasets that increased the reactor power to 100% over 144 000 s at the rate of 3%/h, plus an additional margin of 4000 s, was 4800. The maximum practicably feasible reward for power-increase operation success was observed to be 3000.

C. EXPERIMENTAL RESULTS

After the algorithm for autonomous power increase control was trained, an experiment was conducted to demonstrate that the proposed algorithm could autonomously increase the power at a specified rate. The continuous control module was implemented using an A3C and an LSTM network, while the discrete control module was implemented with a rule-based system. Fig. 17 (a–h) presents the experimental results for a

3.0%/h rate of power increase, which demonstrate that the proposed algorithm can increase the power at the intended rate within the operational boundary (Fig. 17 (a)). In addition, Fig. 17 (b) shows that the proposed algorithm managed the average temperature within the mismatch boundary from the reference temperature over the reactor power of 30% and could effectively restore an increased or decreased average temperature to within the mismatch operation range. The changes in the average temperature that were observed at approximately 40 000 s were due to connecting to the grid and starting a condenser pump, which impacted the overall plant state.

The continuous control module also managed the boron concentration during the power increase; the results are presented in Fig. 17 (c) and (d). To maintain the power at 2%, the boron concentration was increased to compensate for the effect of the control rod withdrawal, which occurred at approximately 22 000 s, as shown in Fig. 17 (e). Then, the controller decreased the boron concentration by increasing the volume of the make-up water to increase the reactor power from 2% to 100%.

The discrete control model operated the system's synchronous connection to connect to the electrical grid at a reactor power of 15%. The discrete control module also selected the turbine load (Fig. 17 (f)) and RPM setpoints (Fig. 17 (g)) based on the reactor power. Additional actions that were performed by the discrete control module during the power-increase operation are presented in Fig. 17 (h) and include starting feedwater pumps 2 and 3 and condenser pumps 2 and 3 to circulate feedwater in the secondary part of the plant. The control module started these pumps in sequence according to the general operating procedure.

VI. DISCUSSION

The experimental results demonstrated that the proposed algorithm successfully controlled the components to increase the reactor power and generate electrical power at the intended rate of power increase. The performance of this algorithm was also compared with that of the established operation strategy, as presented in Section 2. According to Fig. 18, the proposed algorithm had a pattern of operation that was nearly identical to that of the established operational strategy. Therefore, it is concluded that the proposed algorithm, which combines a rule-based system and reinforcement learning, can successfully control the complicated power-increase operation.

In this algorithm, the discrete control module operated the synchronizer controller, turbine controller, main feedwater pump, and condenser pump according to the operational steps that are clearly stated in the GOPs. The continuous control module adjusted the valves to manage the boron concentration and manipulated the rod controller. The continuous control module can provide experiential control of these inputs, thereby gradually affecting the power increase, based on the parameter trends, the predetermined rate of power increase, and the current operational boundaries. In addition,

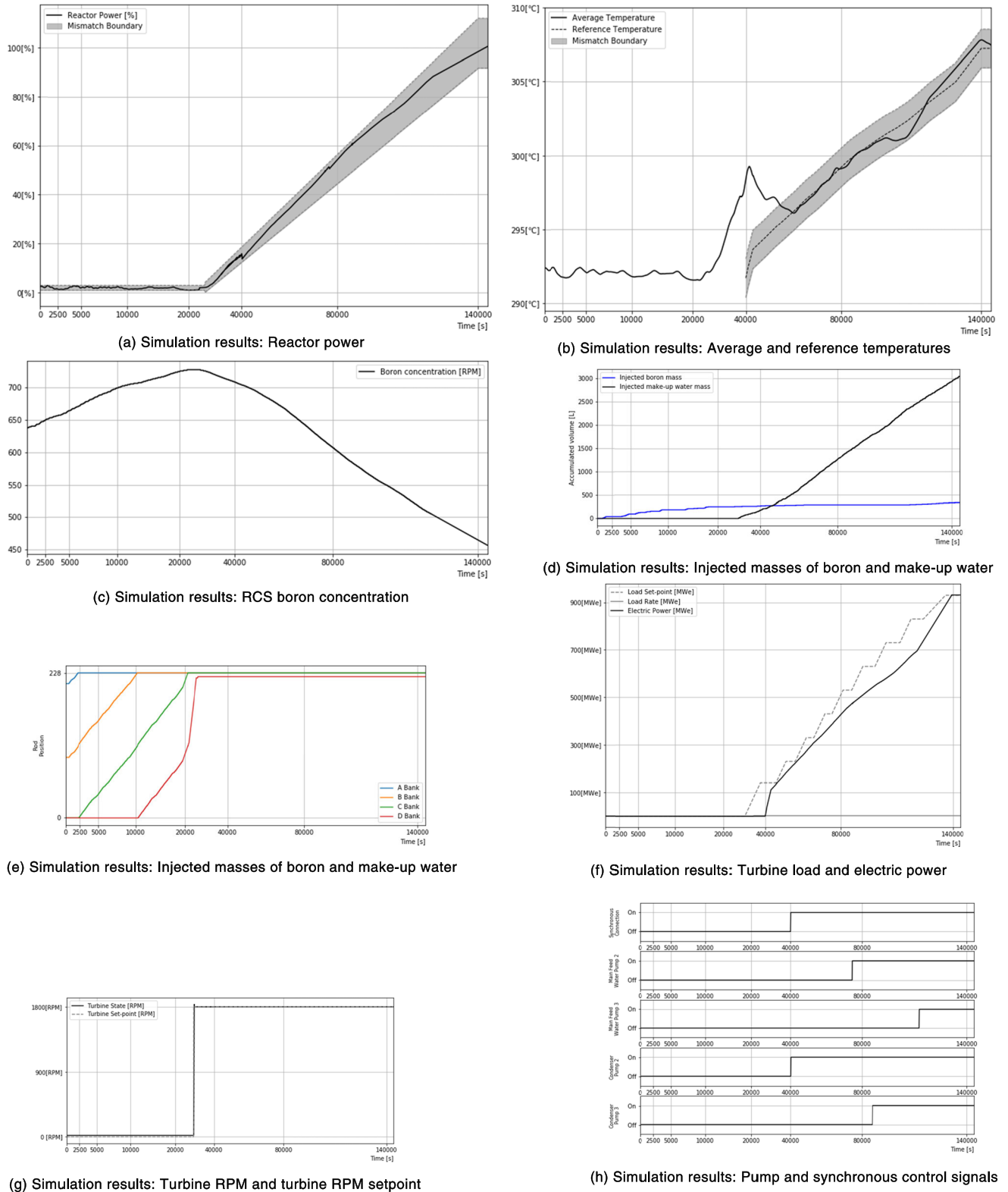


FIGURE 17. Simulation results for a 3%/h autonomous power-increase operation.

the results demonstrate that the continuous control module effectively managed the boron concentration (Fig. 17 (c))

such that the difference between the average temperature and the reference temperature was maintained within $\pm 1^\circ\text{C}$.

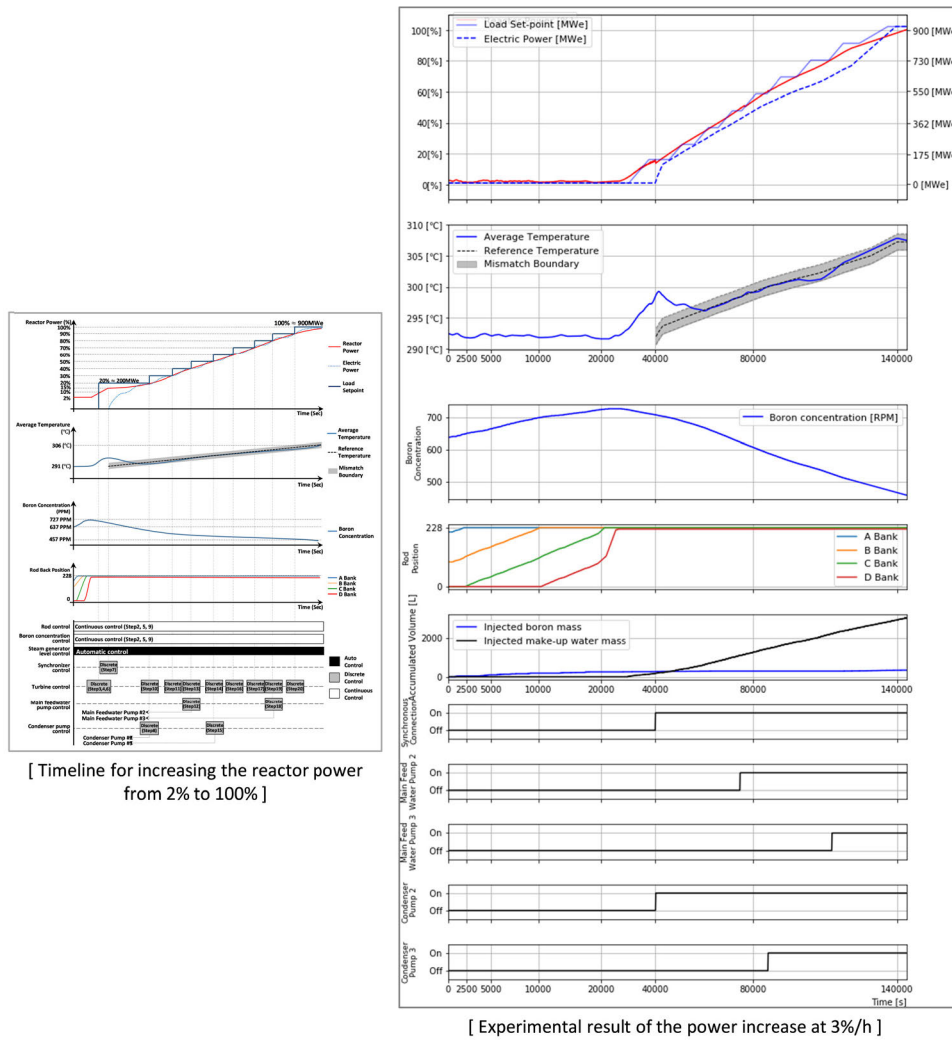


FIGURE 18. Comparison between the existing operational strategy and the simulation results.

Since this rule is not mandated in the GOPs, the control module allowed average temperatures that were outside the mismatch boundary. However, based on interviews with senior operators who work at the reference plant, this restriction can be satisfied after the reactor power reaches 30%; in the earlier stages of the power-increase operation, the start-up of large components results in system disturbances that complicate temperature control. Therefore, these results demonstrate that the A3C agent in the continuous control module can effectively conduct experience-based control after training with the simulator and the discrete control module can control components according to rules that are based on the operating procedures.

VII. CONCLUSION

This study proposed an algorithm for the power-increase operation that uses AI techniques. The power increase algorithm was designed through an analysis of the current operational strategy, which considered the operation staffing and

operating procedures. To train the continuous control, the proposed algorithm used an A3C agent and an LSTM network and applied a rule-based system for the discrete control components. A compact nuclear simulator was used to determine whether the proposed algorithm could effectively and autonomously control the power-increase operation at a 3%/h rate of power increase. Based on the simulation results, the power increase algorithm was proven capable of identifying an acceptable operation path for increasing the reactor power from 2% to 100% at a specified rate of power increase.

The suggested approach seems to be applicable to other operational modes in NPPs, if the reward algorithm is adjusted according to the operation objectives, strategies, methods, and required procedure steps for each operating range. Future studies may suggest developing an agent that can select and control a contextual operating strategy, either in the entire operation range or in part. Future studies may also consider emergency as well as abnormal situations during power-increasing operation. More so, to realize a fully

automated NPP, an autonomous control system should be capable of: automatic operation of the NPP, fault detection, diagnosis (identifying the causes of component failures or incidents), simulation, forecasting the status of the plant, identifying the possible control options, and recommending the best option for optimizing the plant performance. This autonomous control is expected to be a key technology in small modular reactors that are under development.

Several aspects should be further considered regarding the practical application of this algorithm: 1) Since the power-increase operation is only a small part of the overall plant operation, to cover the entire plant operation, the proposed reward algorithm should be changed according to the operation objectives, strategies, operational methods, and required procedural steps for each operating range. Moreover, the AI agent should be capable of selecting and controlling an operating strategy based on the context. 2) To further improve the safety of NPPs, an AI agent requires additional functions (e.g., fault detection, diagnosis, forecasting the status of the plant, identifying the possible control options, and recommending the best option) to address emergency, abnormal, and normal situations. 3) The signal noise in a plant should be an additional consideration; signals in NPPs contain noise, while the simulator does not. Therefore, a technique that can mitigate the signal noise, e.g., signal validation or noise tolerance, must be developed. 4) Another issue is the differences in behavior between the simulator model and actual power plants, which mandates a thorough validation of the practical application.

**APPENDIX
AN EXPERIMENT ON IDENTIFYING A FAST-TRAINING
NETWORK**

This study identifies a network that can be quickly trained in the specified domain since the A3C network requires more than 14 hours per episode to train the entire power increase operation. In this study, the considered networks are DNN (deep neural network), CNN (convolutional neural network), LSTM (long short-term memory), and C-LSTM (CNN + LSTM). DNN is a typical feed-forward neural network that contains many hidden layers of nonlinear hidden units and a very large output layer. In CNN, the hidden layers have fewer connections and parameters because filters that perform convolution operations are utilized. CNN has been demonstrated to outperform DNN in feature extraction from input data. LSTM can calculate time-sequential input data for units that are called constant error carousels. It can facilitate the memorization of important events or long-term data. C-LSTM is a combined model of CNN and LSTM. This network has been proposed for extracting features of data and for handling time-sequential data.

To train these networks under the same conditions, they should have the same number of parameters. The parameters at each layer of the network model are arranged with a normal distribution (mean = 0.0 and standard deviation = 1.0), which supports stable training under the same conditions.

TABLE 4. Architectures of the compared networks.

Network	Network layer	Layer type	Time-sequence	Node	Parameter
DNN	Common	Input layer	-	8	0
		Dense	-	32	224
		Dense	-	64	2112
	Actor	Dense	-	70	4550
		Dense	-	64	4544
		Output layer	-	3	195
		Dense	-	32	2272
Critic	Output layer	-	1	33	
	Dense	-	32	2272	
CNN	Common	Input layer	10	8	0
		Conv1D	10	10	190
		Max pooling	3	10	0
		Flatten	-	30	0
		Dense	-	64	1984
	Actor	Dense	-	70	4550
		Dense	-	64	4544
		Output layer	-	3	195
		Dense	-	32	2272
	Critic	Output layer	-	1	33
		Dense	-	32	2272
LSTM	Common	Input layer	10	8	0
		LSTM	-	32	4992
		Dense	-	64	2112
	Actor	Dense	-	64	4160
		Output layer	-	3	195
		Dense	-	32	2080
		Output layer	-	1	33
C-LSTM	Common	Input layer	10	8	0
		Conv1D	10	10	190
		Max pooling	3	10	0
		LSTM	-	32	5504
	Actor	Dense	-	60	1900
		Dense	-	64	3904
Critic	Output layer	-	3	195	
	Dense	-	32	1952	
		Output layer	-	1	33

Table 4 describes the architectures of the networks that are used in the A3C algorithm for the experiment. Each network consists of three layers: common, actor, and critic. The actor and critic layers are linked to the common layer.

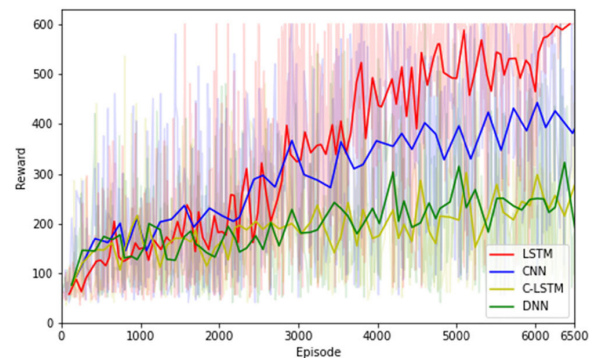


FIGURE 19. Average duration of each network.

Before training on the entire power increase operation, the A3C agent is trained between 2% and 15% power to identify the optimal network. Each network has been trained by 6500 episodes. Fig. 19 shows the trend of the duration

of each network versus the number of episodes. Each line represents the average duration over 10 episodes. The agent's objective is to increase the power within the operational boundary, which is the power reward boundary in this paper, for 600 seconds. For strict comparison of these networks, an operation with a duration of less than 600 seconds is regarded as a failed operation. These networks are trained until the average duration is 600 seconds. In Fig. 19, the LSTM network is the best performing network as it realized an average duration of 600 seconds in 6500 episodes. The second-best performing network is CNN, which realized a duration of approximately 400 seconds in 6500 episodes. C-LSTM and DNN show poor performance (durations of less than 250 seconds). The results of this experiment demonstrate that the LSTM network can realize the operational objective in fewer training episodes than the other networks.

REFERENCES

- [1] R. T. Wood, J. S. Neal, C. R. Brittain, and J. A. Mullens, "Autonomous control capabilities for space reactor power systems," in *Proc. AIP Conf.*, 2004, pp. 631–638.
- [2] J. Kim, D. Lee, J. Yang, and S. Lee, "Conceptual design of autonomous emergency operation system for nuclear power plants and its prototype," *Nucl. Eng. Technol.*, vol. 52, no. 2, pp. 308–322, Feb. 2020.
- [3] D. Lee and J. Kim, "Autonomous algorithm for start-up operation of nuclear power plants by using LSTM," in *Proc. Int. Conf. Appl. Hum. Factors Ergonom.* Orlando, FL, USA: Springer, 2018, pp. 465–475.
- [4] S. J. Lee and P. H. Seong, "Development of automated operating procedure system using fuzzy colored Petri nets for nuclear power plants," *Ann. Nucl. Energy*, vol. 31, no. 8, pp. 849–869, May 2004.
- [5] D. Lee, P. H. Seong, and J. Kim, "Autonomous operation algorithm for safety systems of nuclear power plants by using long-short term memory and function-based hierarchical framework," *Ann. Nucl. Energy*, vol. 119, pp. 287–299, Sep. 2018.
- [6] Y. Kima and J. Park, "Envisioning human-automation interactions for responding emergency situations of NPPs: A viewpoint from human computer interaction," in *Proc. Trans. Korean Nucl. Soc. Autumn Meeting*, Yeosu-si, South Korea, Oct. 2018.
- [7] A. R. Kim, J. Park, J. T. Kim, J. Kim, and P. H. Seong, "Study on the identification of main drivers affecting the performance of human operators during low power and shutdown operation," *Ann. Nucl. Energy*, vol. 92, pp. 447–455, Jun. 2016.
- [8] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
- [9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [10] M. Aggarwal, A. Arora, S. Sodhani, and B. Krishnamurthy, "Improving search through A3C reinforcement learning based conversational agent," in *Proc. Int. Conf. Comput. Sci.* Wuxi, China: Springer, 2018, pp. 273–286.
- [11] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
- [12] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 3, Apr./May 2004, pp. 2619–2624.
- [13] A. Y. Ng, A. Coates, M. Diel, V. Ganapathi, J. Schulte, B. Tse, E. Berger, and E. Liang, "Autonomous inverted helicopter flight via reinforcement learning," in *Experimental Robotics IX*. Berlin, Germany: Springer, 2006, pp. 363–372.
- [14] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
- [15] T. Yang, L. Zhao, W. Li, and A. Y. Zomaya, "Reinforcement learning in sustainable energy and electric systems: A survey," *Annu. Rev. Control*, vol. 49, pp. 145–163, Apr. 2020.
- [16] A. G. E. Collins, "Reinforcement learning: Bringing together computation and cognition," *Current Opinion Behav. Sci.*, vol. 29, pp. 63–68, Oct. 2019.
- [17] M. De Paula, L. O. Ávila, and E. C. Martínez, "Controlling blood glucose variability under uncertainty using reinforcement learning and Gaussian processes," *Appl. Soft Comput.*, vol. 35, pp. 310–332, Oct. 2015.
- [18] M. M. Noel and B. J. Pandian, "Control of a nonlinear liquid level system using a new artificial neural network based reinforcement learning approach," *Appl. Soft Comput.*, vol. 23, pp. 444–451, Oct. 2014.
- [19] A. Kamiya, H. Kimura, M. Yamamura, and S. Kobayashi, "Power plant start-up scheduling: A reinforcement learning approach combined with evolutionary computation," *J. Intell. Fuzzy Syst.*, vol. 6, no. 1, pp. 99–115, 1998.
- [20] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, and S. Petersen, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [21] A. Viitala, R. Boney, and J. Kannala, "Learning to drive small scale cars from scratch," 2020, *arXiv:2008.00715*. [Online]. Available: <http://arxiv.org/abs/2008.00715>
- [22] S. Zhou, Z. Hu, W. Gu, M. Jiang, M. Chen, Q. Hong, and C. Booth, "Combined heat and power system intelligent economic dispatch: A deep reinforcement learning approach," *Int. J. Electr. Power Energy Syst.*, vol. 120, Sep. 2020, Art. no. 106016.
- [23] A. Saenz-Aguirre, E. Zulueta, U. Fernandez-Gamiz, J. Lozano, and J. Lopez-Guede, "Artificial neural network based reinforcement learning for wind turbine yaw control," *Energies*, vol. 12, no. 3, p. 436, Jan. 2019.
- [24] E. Samadi, A. Badri, and R. Ebrahimpour, "Decentralized multi-agent based energy management of microgrid using reinforcement learning," *Int. J. Electr. Power Energy Syst.*, vol. 122, Nov. 2020, Art. no. 106211.
- [25] H. Kazmi, F. Mehmood, S. Lodeweyckx, and J. Driesen, "Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems," *Energy*, vol. 144, pp. 159–168, Feb. 2018.
- [26] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building HVAC control," in *Proc. 54th Annu. Design Autom. Conf.*, Jun. 2017, pp. 1–6.
- [27] Z. Yang, F. Zhu, and F. Lin, "Deep-reinforcement-learning-based energy management strategy for supercapacitor energy storage systems in urban rail transit," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 10, 2020, doi: 10.1109/TITS.2019.2963785.
- [28] S. Bennett, "The past of pid controllers," *Annu. Rev. Control*, vol. 25, pp. 43–53, Jan. 2001.
- [29] S. M. H. Mousakazemi, "Computational effort comparison of genetic algorithm and particle swarm optimization algorithms for the proportional–integral–derivative controller tuning of a pressurized water nuclear reactor," *Ann. Nucl. Energy*, vol. 136, Feb. 2020, Art. no. 107019.
- [30] S. M. H. Mousakazemi, "Control of a PWR nuclear reactor core power using scheduled PID controller with GA, based on two-point kinetics model and adaptive disturbance rejection system," *Ann. Nucl. Energy*, vol. 129, pp. 487–502, Jul. 2019.
- [31] B. R. Upadhyaya, K. Zhao, S. Perillo, X. Xu, and M. Na, "Autonomous control of space reactor systems," Univ. Tennessee, Knoxville, TN, USA, Tech. Rep., 2007.
- [32] B. Zhang, M. Peng, S. Cheng, and L. Sun, "Novel fuzzy logic based coordinated control for multi-unit small modular reactor," *Ann. Nucl. Energy*, vol. 124, pp. 211–222, Feb. 2019.
- [33] M. Gyun Na and B. R. Upadhyaya, "A neuro-fuzzy controller for axial power distribution in nuclear reactors," *IEEE Trans. Nucl. Sci.*, vol. 45, no. 1, pp. 59–67, Feb. 1998.
- [34] H. Arab-Alibeik and S. Setayeshi, "Adaptive control of a PWR core power using neural networks," *Ann. Nucl. Energy*, vol. 32, no. 6, pp. 588–605, Apr. 2005.
- [35] E. Rojas-Ramírez, J. S. Benítez-Read, and A. S.-D.-L. Ríos, "A stable adaptive fuzzy control scheme for tracking an optimal power profile in a research nuclear reactor," *Ann. Nucl. Energy*, vol. 58, pp. 238–245, Aug. 2013.
- [36] Q. Jiang, Y. Liu, W. Zeng, and T. Yu, "Study on switching control of PWR core power with a fuzzy multimodel," *Ann. Nucl. Energy*, vol. 145, Sep. 2020, Art. no. 107611.
- [37] W. Zeng, Q. Jiang, J. Xie, and T. Yu, "A fuzzy-PID composite controller for core power control of liquid molten salt reactor," *Ann. Nucl. Energy*, vol. 139, May 2020, Art. no. 107234.
- [38] Z. Huang, R. M. Edwards, and K. Y. Lee, "Fuzzy-adapted recursive sliding-mode controller design for a nuclear power plant control," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 1, pp. 256–266, Feb. 2004.

- [39] S. S. Khorramabadi, M. Boroushaki, and C. Lucas, "Emotional learning based intelligent controller for a PWR nuclear reactor core during load following operation," *Ann. Nucl. Energy*, vol. 35, no. 11, pp. 2051–2058, Nov. 2008.
- [40] M. Boroushaki, M. B. Ghofrani, C. Lucas, and M. J. Yazdanpanah, "An intelligent nuclear reactor core controller for load following operations, using recurrent neural networks and fuzzy systems," *Ann. Nucl. Energy*, vol. 30, no. 1, pp. 63–80, Jan. 2003.
- [41] E. Hatami, N. Vosoughi, and H. Salarieh, "Design of a fault tolerated intelligent control system for load following operation in a nuclear power plant," *Int. J. Electr. Power Energy Syst.*, vol. 78, pp. 864–872, Jun. 2016.
- [42] R. Coban, "Power level control of the TRIGA mark-II research reactor using the multifeedback layer neural network and the particle swarm optimization," *Ann. Nucl. Energy*, vol. 69, pp. 260–266, Jul. 2014.
- [43] K. Sekimizu, T. Araki, and S. Kawakami, "Knowledge representation for automated boiling water reactor plant startup," *Nucl. Technol.*, vol. 100, no. 3, pp. 295–309, Dec. 1992.
- [44] J.-T. Kim, K.-C. Kwon, I.-K. Hwang, D.-Y. Lee, W.-M. Park, J.-S. Kim, and S.-J. Lee, "Development of advanced I&C in nuclear power plants: ADIOS and ASICS," *Nucl. Eng. Des.*, vol. 207, pp. 105–119, Jul. 2001.
- [45] H. Basher and J. Neal, *Autonomous Control of Nuclear Power Plants*. Oak Ridge, TN, USA: Oak Ridge National Laboratory, 2003.
- [46] M. G. Na, S. H. Shin, and W. C. Kim, "A model predictive controller for nuclear reactor power," *Nucl. Eng. Technol.*, vol. 35, no. 5, pp. 399–411, 2003.
- [47] J. Yang and J. Kim, "Accident diagnosis algorithm with untrained accident identification during power-increasing operation," *Rel. Eng. Syst. Saf.*, vol. 202, Oct. 2020, Art. no. 107032.
- [48] J. Yang and J. Kim, "An accident diagnosis algorithm using long short-term memory," *Nucl. Eng. Technol.*, vol. 50, no. 4, pp. 582–588, May 2018.
- [49] Y. Li, "Deep reinforcement learning: An overview," 2017, *arXiv:1701.07274*. [Online]. Available: <https://arxiv.org/abs/1701.07274>
- [50] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
- [51] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1998.
- [52] V. Mnih, A. P. Badia, A. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.
- [53] I. Adamski, R. Adamski, T. Grel, A. Jędrych, K. Kaczmarek, and H. Michalewski, "Distributed deep reinforcement learning: Learn how to play Atari games in 21 minutes," in *Proc. Int. Conf. High Perform. Comput.* Bengaluru, India: Springer, 2018, pp. 370–388.
- [54] X. Guo, "Deep learning and reward design for reinforcement learning," Ph.D. dissertation, Dept. Comput. Sci. Eng., Michigan Univ., Ann Arbor, MI, USA, 2017.
- [55] T. Chen, W. Niu, Y. Xiang, X. Bai, J. Liu, Z. Han, and G. Li, "Gradient band-based adversarial training for generalized attack immunity of A3C path finding," 2018, *arXiv:1807.06752*. [Online]. Available: <https://arxiv.org/abs/1807.06752>
- [56] R. Garduno-Ramirez and K. Y. Lee, "Multiobjective optimal power plant operation through coordinate control with pressure set point scheduling," *IEEE Trans. Energy Convers.*, vol. 16, no. 2, pp. 115–122, Jun. 2001.
- [57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [58] S. Şeker, E. Ayaz, and E. Türkcan, "Elman's recurrent neural network applications to condition monitoring in nuclear power plant and rotating machinery," *Eng. Appl. Artif. Intell.*, vol. 16, nos. 7–8, pp. 647–656, Oct. 2003.
- [59] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, Edinburgh, U.K., 1999, pp. 850–855.
- [60] T. T. Trinh, R. Yoshihashi, R. Kawakami, M. Iida, and T. Naemura, "Bird detection near wind turbines from high-resolution video using LSTM networks," in *Proc. World Wind Energy Conf. (WVEC)*, 2016, p. 6.
- [61] KAERI, "Advanced compact nuclear simulator textbook," KAERI, Nucl. Training Center Korea At. Energy Res. Inst., Daejeon, South Korea, Tech. Rep., 1990.



DAEIL LEE received the B.Sc. and M.Sc. degrees in nuclear engineering from Chosun University, Gwangju, South Korea, where he is currently pursuing the Ph.D. degree in autonomous operations of nuclear systems with artificial intelligence. His research interests broadly include machine learning, deep learning, reinforcement learning, and the applications of artificial intelligence for the nuclear power plant industry. Most of his recent works are related to the development of operational supporting algorithms for autonomous systems.



AWWAL MOHAMMED ARIGI received the bachelor's degree in electrical and computer engineering from the Federal University of Technology Minna, Nigeria, the master's degree in nuclear power plant (NPP) engineering from the KEPCO International Nuclear Graduate School, South Korea, and the Ph.D. degree in nuclear systems engineering from Chosun University, South Korea. His research interests include cognitive engineering, including human-in-the-loop experiments, development of methodologies for human reliability analysis, human factors analysis in the digitized operation of NPPs, and artificial intelligence for enhancing human cognition in safety-critical infrastructures.



JONGHYUN KIM received the B.Sc., M.Sc., and Ph.D. degrees in nuclear engineering from the Korea Advanced Institute of Science and Technology (KAIST), South Korea. He worked as a Researcher with KEPCO ENC, Korea Hydro & Nuclear Power Company Ltd., and Paul Scherrer Institute, from 2004 to 2011. He was also an Associate Professor with the KEPCO International Nuclear Graduate School, from 2011 to 2015. He is currently the Head of the Department of Nuclear Engineering, Chosun University, Gwangju, South Korea. His research interests include human factors engineering and artificial intelligence application for nuclear power plants.

...