Nuclear Engineering and Technology 55 (2023) 814–826

Contents lists available at ScienceDirect

Nuclear Engineering and Technology

journal homepage: www.elsevier.com/locate/net

Original Article

RNN-based integrated system for real-time sensor fault detection and fault-informed accident diagnosis in nuclear power plant accidents

Jeonghun Choi, Seung Jun Lee*

Ulsan National Institute of Science and Technology, 50 UNIST-gil, Ulju-gun, Ulsan, 44919, Republic of Korea

ARTICLE INFO

Article history:
Received 16 May 2022
Received in revised form
11 October 2022
Accepted 25 October 2022
Available online 2 November 2022

Keywords:
Nuclear power plants
Sensor fault detection
Recurrent neural network
Accident diagnosis
Fault-tolerant system
Real-time execution

ABSTRACT

Sensor faults in nuclear power plant instrumentation have the potential to spread negative effects from wrong signals that can cause an accident misdiagnosis by plant operators. To detect sensor faults and make accurate accident diagnoses, prior studies have developed a supervised learning-based sensor fault detection model and an accident diagnosis model with faulty sensor isolation. Even though the developed neural network models demonstrated satisfactory performance, their diagnosis performance should be reevaluated considering real-time connection. When operating in real-time, the diagnosis model is expected to indiscriminately accept fault data before receiving delayed fault information transferred from the previous fault detection model. The uncertainty of neural networks can also have a significant impact following the sensor fault features. In the present work, a pilot study was conducted to connect two models and observe actual outcomes from a real-time application with an integrated system. While the initial results showed an overall successful diagnosis, some issues were observed. To recover the diagnosis performance degradations, additive logics were applied to minimize the diagnosis failures that were not observed in the previous validations of the separate models. The results of a case study were then analyzed in terms of the real-time diagnosis outputs that plant operators would actually face in an emergency situation.

1. Introduction

With the fast progress of machine learning techniques, applications of high computing power and novel machine learning models are being actively investigated in many industrial fields. The high performance of machine learning models has shown high applicability in many industrial and academic fields [1]. In the nuclear field, there have been numerous suggestions for automating the operators' tasks for reducing task loads and human errors. Related research includes autonomous control in the start-up/shutdown phases, automated diagnosis of the plant state, and automated plant parameter predictions. One characteristic of the novel approaches with neural networks in the nuclear field is that they mostly consider diverse transient states of nuclear power plants such as start-up and shutdown, and abnormal and accident situations. These transient states contain nonlinear features with the actuation of automated components, human operation, and peculiar conditions of the plant.

In particular, accidents in nuclear power plants produce quite irregular conditions due to dramatic changes in parameters and unique features following the accident type and its scale. Accidents with the loss of the plant's critical safety functions require an automated reactor trip, which actuates several serial sequences of automatic activations and inactivations of components for maintaining plant integrity. Diverse accident symptoms and multiple status changes of component sets lead to many different plant states according to the accident progression with nonlinear and complex trend changes. Thus, accurately identifying the occurred accident plays an important role in situation awareness and decision-making. In this light, faulty plant sensors can have a significant negative influence on the accident identification process. This was seen in the Three Mile Island accident, in which sensor faults caused operators to wrongly diagnose the accident [2]. In addition to impacting accident identification, sensor faults continuously threaten plant safety by potentially causing human errors during accident responses.

Existing techniques for monitoring sensor states mainly focus on the parameter relations to reconstruct signals and use them to distinguish faulty signals. However, the dynamic features of nuclear

* Corresponding author.
E-mail address: sjlee420@unist.ac.kr (S.J. Lee).

https://doi.org/10.1016/j.net.2022.10.035

accidents make it challenging to construct relations between sensors, where developing parameter relations has different schemes following the accident type. A dilemma arises as to which task is better to perform first, sensor fault monitoring or accident diagnosis. Preceding accident diagnosis results would ensure better fault detection performance, while data validation from preceding fault detection would guarantee accurate diagnosis.

To address the above issue, in a previous work the authors suggested a supervised learning-based sensor fault detection algorithm [3]. By training a neural network with proposed labels, injected sensor faults could be accurately classified. Then in a follow-up study, a sensor fault-tolerant accident diagnosis model was developed [4] to realize fault-informed accident diagnosis, whereby the diagnosis results reflect sensor fault information by removing the influences of faulty sensors.

The goal of sensor fault-tolerant accident diagnosis is for actual inclusion as an operator support system to assist operators in emergency situations involving high stress and workload. Therefore, an integrated system has to be executed in a real-time manner to give updated accident information in a timely manner. In the current study, the previously developed sensor fault monitoring and fault-informed accident diagnosis models are connected into one system for real-time execution in a nuclear power plant accident situation. To treat the issues from delayed fault information transfer between models, improved fault information processing and additional diagnosis decision logics are applied to the integrated system.

2. Sensor fault-tolerant accident diagnosis

2.1. Motivations

Nuclear accidents include diverse malfunctions of critical components that lead to an autonomous reactor trip by exceeding the trip threshold. Examples of typical nuclear power plant accidents include a loss of coolant accident (LOCA), steam generator tube rupture (SGTR), and others. In an accident situation, plant operators judge the occurred accident by the symptoms in terms of the sensor values and trends. Since correctly diagnosing the accident is crucial for executing the appropriate accident responses, *accurate and quick accident diagnosis* has to be achieved from *correct sensor values*. However, large uncertainty exists in accident conditions because of the diverse accident types and the distinct symptoms of each case following the location and size of the break or malfunction and the failure sequences. As shown in Fig. 1, plant parameters have dissimilar patterns even in the same accident type (LOCA in the figure), varying by break location. Considering such dynamic characteristics of accidents, several sensor fault detection methods for monitoring the correct sensor scales [5–8] and accident diagnosis methods [9–11] based on data-driven approaches have been suggested to support operators in an accident situation.

In this circumstance, the pending issue is which task should precede the other between sensor fault monitoring and accident diagnosis. It is evident that the performance of sensor fault detection may differ depending on the scope of the data. A smaller scope of data training and usage, like for one specific accident type, will output more accurate results due to small deviations in the data. On the other hand, accident diagnosis with fault-monitored data will prevent wrong diagnosis cases from trained faults. Between the two options, diagnosis cannot be firstly performed considering the currently low robustness of classification models. While some recurrent neural network (RNN) models have made successful diagnosis results with nuclear power plant accident data [12,13] [Baraldi et al., 2015], showed that machine learning algorithms including RNNs do not assure robustness against artificial faulty data [14]. For this reason, sensor fault detection covering diverse accident conditions has to be the first step toward fault-informed accident diagnosis.

In our prior research, a sensor fault detection model and a sensor fault-tolerant accident diagnosis model were developed [3,4]. Our first study constructed consistency index-based sensor fault monitoring neural networks utilizing the high performance of supervised learning. With the successful fault detection rate of the first study, the results of the fault-informed accident diagnosis model developed in the follow-up study showed effective exclusion of fault influence. Despite the fine performance of each model, the two models were treated separately without direct connections. It is expected that a real-time connection of the two models will result in unexpected failures from the real-time feature. Thus, the purpose of the current study is to integrate the two previous models to confirm the actual consequences of their simultaneous operation.

The overall framework of the current study is shown in Fig. 2. The integrated system is actuated upon automatic reactor trip in the case of an accident involving deviation of sensor parameters. After the reactor trip, a predefined plant parameter set (see Section 4.1.1) is extracted and transferred to the sensor fault monitoring model and accident diagnosis model in a real-time manner. Sensor fault monitoring is firstly executed, after which the fault information is delivered to the accident diagnosis model. The diagnosis model reflects the fault information by lowering the influence of any faulty sensors to achieve accurate diagnosis of the accident.

2.2. Consistency index-based sensor fault detection

Data driven methods are fatally bound with uncertainty according to unpredicted or untrained inputs, which is reflected in the characteristic of robustness. If a model can generate consistent performance even with untrained data, it is considered to have high robustness. Model robustness is the main driver of the existing sensor fault detection methods. As shown in Fig. 3, existing studies on sensor fault detection in the nuclear field, called online monitoring (OLM), are generally based on residual analysis connected with an unsupervised learning-based reconstruction model. The reconstruction model has an auto-associative structure with the same input and output variables. By the training sequence of the auto-associative model, the relations between variables are reflected in the model weights. These extracted features reconstruct the robust output; therefore, a faulty signal is detected through a difference between the original and reconstructed signals. From the early stages of the development of OLM techniques, many data-driven methods such as principal component analysis, neural networks, auto-associative kernel regression, etc., have been applied for the reconstruction model.

Unsupervised learning-based sensor fault detection methods have been adopted for sensor fault detection; however, nuclear power plant accident data contain diverse symptoms that can be diversely clustered following the type of accident, and thus the development of a neural network model and parameter setting are challenging. Using a supervised learning strategy, [Choi and Lee, 2020] showed accomplished performance of sensor fault classification [3]. The consistency index labeling rule used for supervised learning strategy is illustrated in Fig. 4. The neural network-based fault detection model generates a consistency index of each sensor by processing the multivariate sensor inputs. The fault detection model consists of stacked RNNs, specifically long short-term memory (LSTM) networks. The LSTM is governed by three gate functions, namely input gate (i_t), forget gate (f_t), and output gate (o_t), with the cell state (C_t) carrying the long-term information. The three gate functions regulate the updates of the cell state and

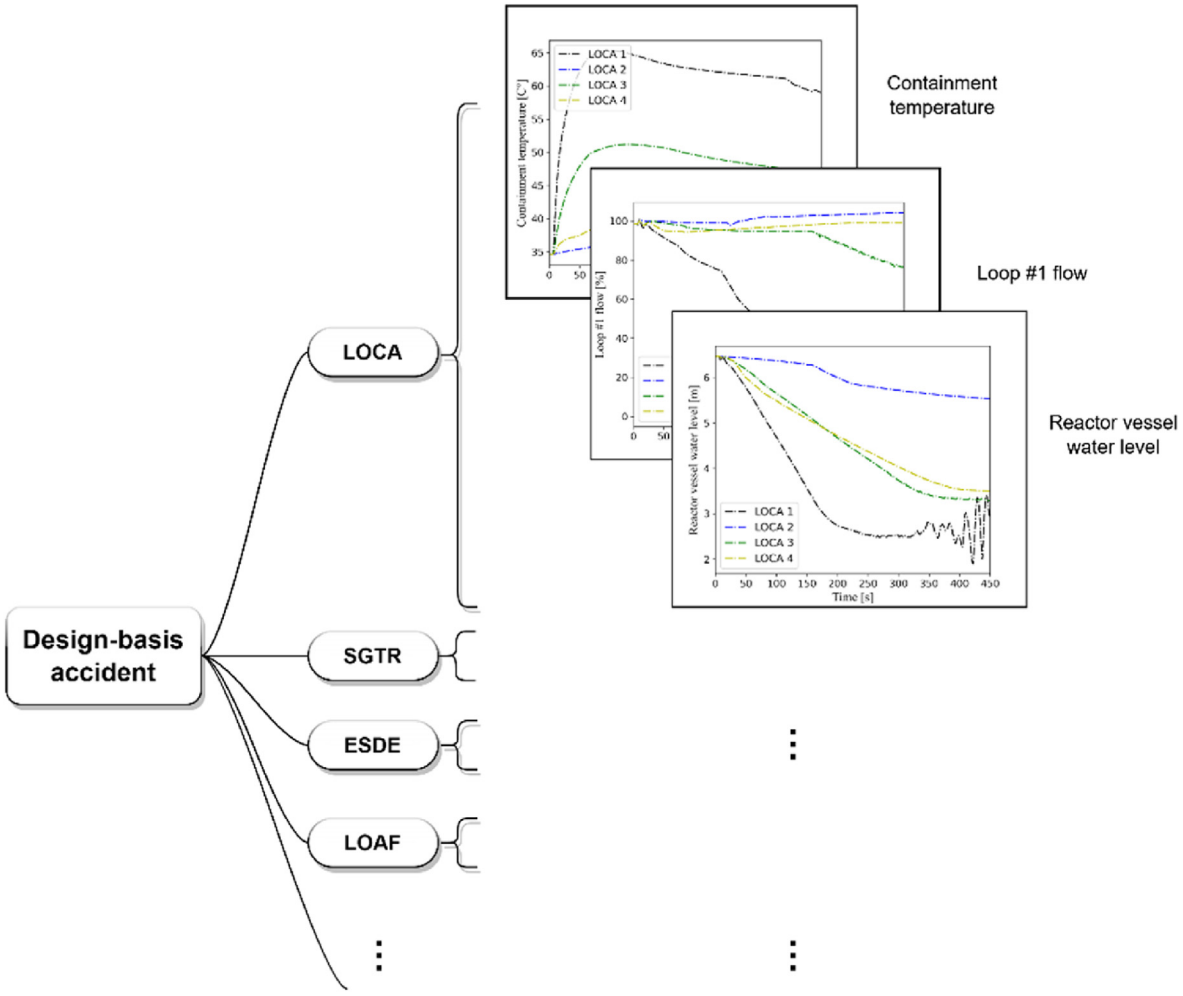

Fig. 1. Diverse accident features following the accident types and malfunction modes.

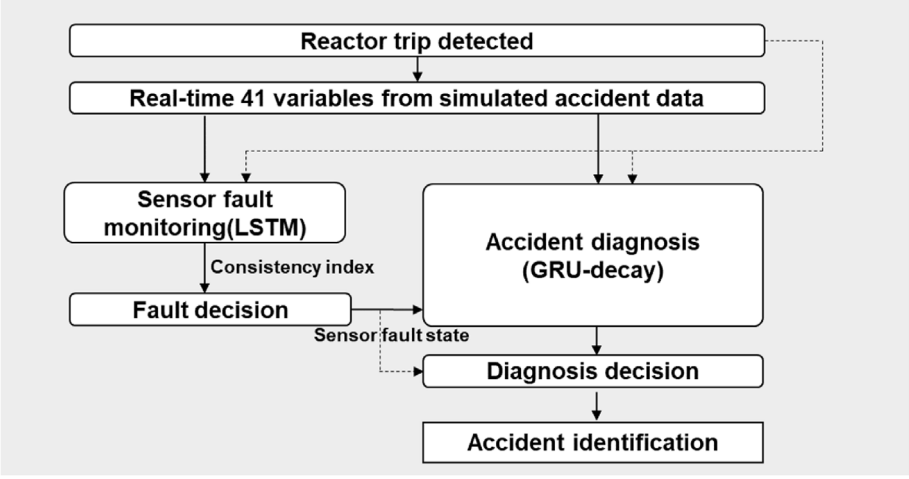

Fig. 2. Integrated system framework.

conclusive hidden state (h_t) with the input (i_t) and the previous hidden state (h_{t-1}).

For the supervised learning of LSTM networks for sensor fault detection, the consistency index was suggested to label each sensor data. The consistency index ($c_{i,t}$), of each sensor is determined based on the squared relative accuracy of the measured value to the real value, $(1-\varepsilon_{i,t})^2$. If the relative error ($\varepsilon_{i,t}$) exceeds the threshold error (ε_{th}), $c_{i,t}$ is fixed to zero. Fig. 4 shows examples of consistency index labeling.

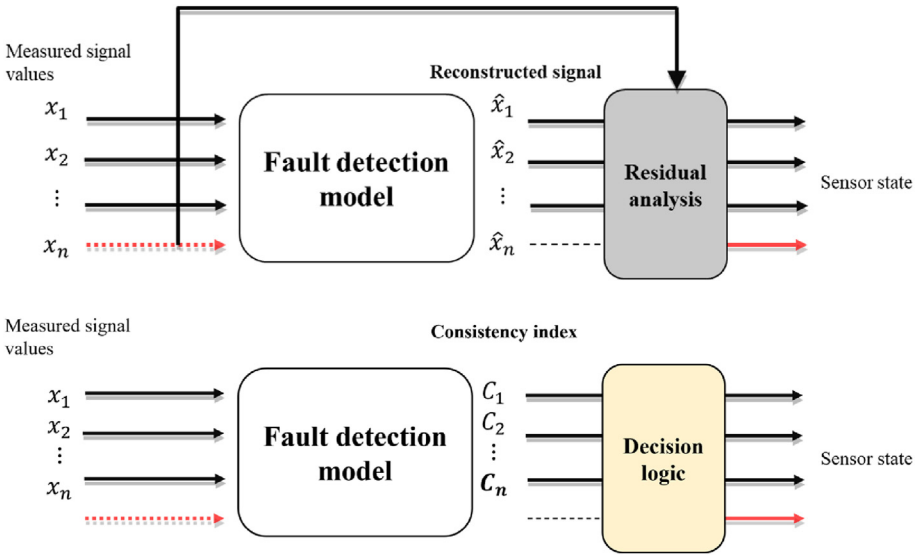

Fig. 3. Unsupervised and supervised learning-based sensor fault detection.

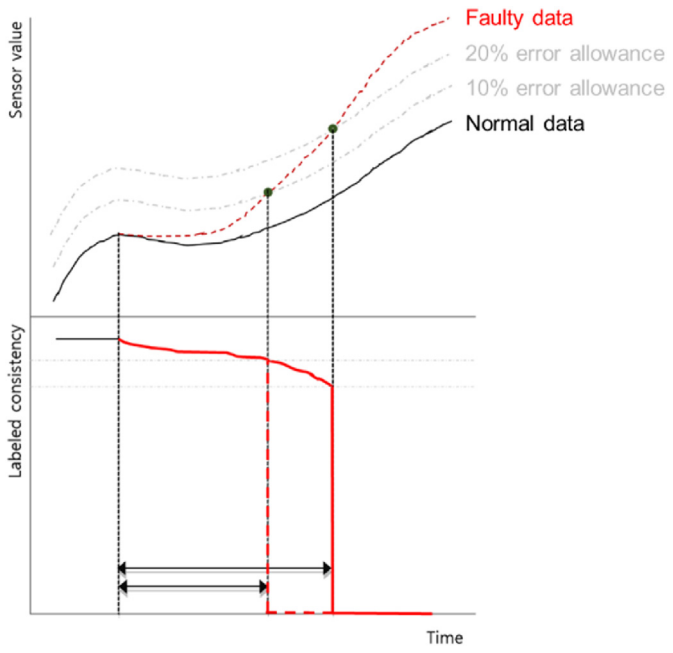

Fig. 4. Consistency index labeling rule.

$$c_{i,t} = \begin{cases} (1-\varepsilon_{i,t})^2 = \left(1 - \left|\dfrac{\tilde{A}_{i,t} - A_{i,t}}{A_{i,t}}\right|\right)^2, 0 \le \varepsilon_{i,t} \le \varepsilon_{th} \\ 0, \varepsilon_{i,t} > \varepsilon_{th} \end{cases} \quad (1)$$

where $\varepsilon_{i,t}$ is the relative error of the obtained signal to the real signal, and ε_{th} is the allowable error.

The sensitivity of fault detection can be modulated by adjusting the error allowance in the labeling rule and the decision logics, which refer to the binary decision from the consistency index output. If the error allowance and decision logics decide the fault state with small margins, the model sensitively outputs faulty sensor states. This can ensure a high true negative rate, which means the proper classification of normal states; however, this would also increase the false positive rate, which means the wrong classification of normal states as faulty states. Accordingly, proper criteria values for both error allowance and decision logics need to be derived from sensitivity studies against useable data for optimizing model performance.

2.3. Fault-informed accident diagnosis

Early phases of nuclear accidents are unclear situations because the accident symptoms differ following various accident types and scales. Unlike normal operation, a nuclear accident is accompanied by a reactor trip, which is automatically actuated with pre-set parameters, and thus sudden changes in the parameters and alarms simultaneously occur. In this situation, correct diagnosis of the accident is essential to plan the appropriate responses. To support the operator tasks related to accident diagnosis, several neural network and knowledge-based methods have been investigated [14–16]. Among the diagnosis algorithms, RNN models have advantages in terms of time contexts in both short- and long-term scales. Studies including [Yang and Kim, 2018] and [Wang et al., 2021] have shown fine performance of RNNs as an accident identification method [13,15].

To achieve fault-informed diagnosis, gated recurrent unit (GRU)-decay (GRUD), which is an improved GRU model, was chosen for the accident diagnosis with fault information model [16]. GRU has a simplified LSTM structure. Two gate functions, reset and update, determine the hidden state from the input and previous hidden state. The hidden state plays the role of both the cell state and hidden state of the LSTM. In the GRUD model, a decay constant (γ) conducts a decay mechanism that affects the input and hidden state.

$$\boldsymbol{z_t} = \boldsymbol{\sigma}(\boldsymbol{W_z x_t} + \boldsymbol{U_z h_{t-1}}) \quad (2)$$

$$\tilde{\boldsymbol{h}}_{\boldsymbol{t}} = \mathbf{tanh}(\boldsymbol{W x_t} + \boldsymbol{U}(\boldsymbol{r_t} \odot \boldsymbol{h_{t-1}})) \quad (3)$$

$$\boldsymbol{h_t} = (\mathbf{1} - \boldsymbol{z_t}) \odot \boldsymbol{h_{t-1}} + \boldsymbol{z_t} \odot \tilde{\boldsymbol{h}}_{\boldsymbol{t}} \quad (4)$$

$$\boldsymbol{r_t} = \boldsymbol{\sigma}(\boldsymbol{W_r x_t} + \boldsymbol{U_r h_{t-1}}) \tag{5}$$

With the basic gate functions above, the update gate (z_t) has the function of the forget and input gates of the LSTM. It determines the degree to which past and present information (past hidden state and present input) is reflected. The reset gate (r_t) determines the extent of resetting the previous hidden state (h_{t-1}). The candidate ($\tilde{h}_t$) is the result of reset gate processing. The present hidden state (h_t), which is the result of the GRU cell, is generated with the update of $\tilde{h}_t$.

$$\boldsymbol{\gamma_t} = \mathbf{exp}(-\mathbf{max}(\mathbf{0}, \boldsymbol{W_\gamma \sigma_t} + \boldsymbol{b_\gamma})) \tag{6}$$

$$\widehat{\boldsymbol{x}}_{\boldsymbol{t}} = \boldsymbol{m_t x_t} + (\mathbf{1} - \boldsymbol{m_t})(\boldsymbol{\gamma_{x_t} x_t} + (\mathbf{1} - \boldsymbol{\gamma_{x_t}})\boldsymbol{x_t}) \tag{7}$$

$$\widehat{\boldsymbol{h}}_{\boldsymbol{t-1}} = \boldsymbol{\gamma_{h_t}} \odot \boldsymbol{h_{t-1}} \tag{8}$$

Originally, the GRUD model was made for the purpose of treating datasets with missing data. The missing information is marked with a masking (m) that activates the decay mechanism. The decay mechanism is driven by a decay constant (γ_t). As seen in Eq. (6), γ_t is determined by the decay weight and bias (W_γ, b_γ), which are jointly trained from all the other parameters in the GRU cell functions. Via negative exponentiated rectifier, the decay mechanism drives a reasonable reduction following the importance of the parameters. If the masking indicates a missing value at time t ($m_t = 0$), x_t is decayed to x_t, which is a predefined imputation value (e.g., mean), and the hidden state is immediately decayed from h_{t-1}. The decay of the hidden state mainly reduces the influence of the missing data, and the model results are acquired from the other parameter calculations. The employed decay mechanism on the GRUD cell is shown in Fig. 5.

Based on the above GRUD structure, a fault-tolerant accident diagnosis model was proposed [4]. The GRUD model has the same function as GRU in the case where the masking always indicates that all parameters are normal ($m_t = 1$), and thus the GRUD model operates in an error-free state with the same high performance as general RNN models. When the fault state is transferred ($m_t = 0$), the algorithm diminishes the influence of the faulty sensors by decaying the hidden state and input, which results in a diagnosis from the other sensor inputs.

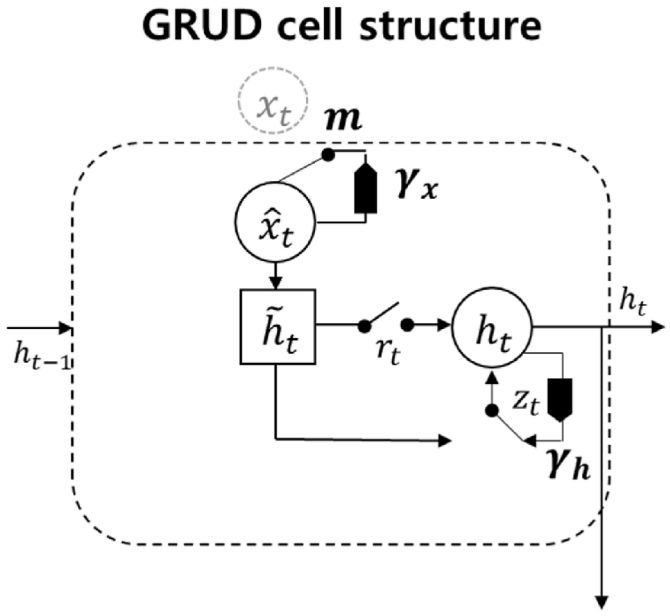

Fig. 5. GRU-decay cell structure with decay mechanism.

3. Real-time model connection

Each previously developed RNN model for sensor fault detection and accident diagnosis was separately evaluated with simulation data and showed fine performance. However, the real-time connection of the two models might present performance degradations, such as from an uncertain or random time of fault information transfer due to the time required for sensor fault detection. The required time for fault detection can vary by the different degrees of sensor faults, accident types, fault injection times, and so on. In this section, the real-time integration of the two models, namely sensor fault detection and fault-informed accident diagnosis, was carried out as a pilot study. Based on the initial results, additional logics were analyzed to minimize the effect of the observed performance degradations.

3.1. Overview of the integrated system with real-time inputs

The first step for the real-time execution of the integrated system is time window (t_{window}) extraction from the multivariate time series data. Two-dimensional accident data are generated with time-steps and sent to the sensor fault detection LSTM and accident diagnosis GRUD models. The LSTM model generates the estimated consistency index output, which is close to zero if the sensor is in a fault state and otherwise close to one. The consistency outputs are processed by the sensor fault decision logic, so that the fault information brings about a binary sensor fault state. The binary sensor state can be used as-is as a masking input of GRUD; more specifically, an index value of 1, meaning a normal sensor, is regarded as a useable input in the diagnosis model, while an index value of 0, meaning a faulty sensor, is excluded from calculation in the diagnosis model. The GRUD model generates the probability of each accident label with the time window input and masking. The accident is diagnosed by the probabilities of the accident labels processed by the diagnosis decision logic. Fig. 6 describes the data flow from the sensor signals in an accident situation to the diagnosis output.

3.2. Pilot study of real-time execution

To observe any potential unexpected negative effects from the integration of the models, a pilot study of the developed system was tested with accident simulation data containing sensor faults. Here, sensor fault data produces diverse diagnosis outputs. The trends of the diagnosis outputs can be classified into four cases, as shown in Table 1 and the examples in Fig. 7. Case 1 involves unchanging true trends, or in other words, in this case the output with the maximum probability accurately indicates the occurred accident, as shown in Fig. 7(a). Case 1 takes up 91.8% of the overall data, meaning that the faulty data were not overly influential or that the faults were isolated before any fault effect appeared. Case 2, possessing 1.8% of the total cases, reflects a negative effect of sensor isolation. These cases occurred when a highly weighted parameter was isolated, e.g., the secondary radiation sensor faults in the SGTR accident. From such an isolation, the stable true trends of the diagnosis output rapidly deteriorated, as in Fig. 7(b). After the true trend interruption though, the output gradually recovered over time, meaning the accident diagnosis was sufficiently generated based on the other sensors. Case 3 represents the appropriate functioning of sensor fault detection and isolation. By the sensor fault effect, the true diagnosis output initially deteriorates, but after

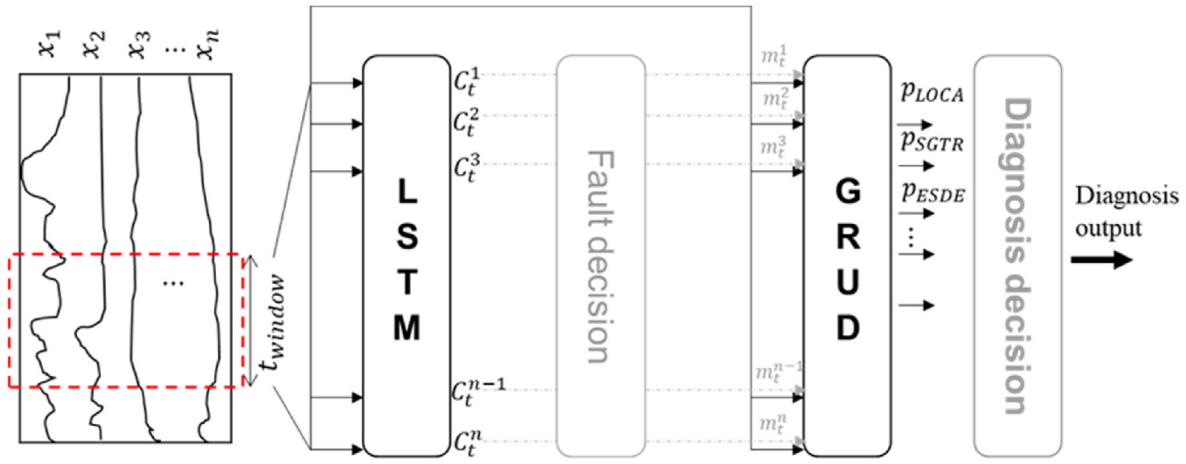

Fig. 6. Data flow and fault information transfer.

Table 1
Pilot study of the integrated system with faulty data.

#		Percentage
Case 1	Unchanging true trends	91.3%
Case 2	True trend interrupted by sensor isolation but recovered after a period of time	1.8%
Case 3	False trend from sensor fault occurred but recovered by faulty sensor isolation	6.4%
Case 4	False trend from sensor fault occurred and not recovered	0.5%

sensor fault isolation, the diagnosis result is dramatically recovered to the true results. This case takes up 6.4% of the total cases. The last case, Case 4 taking up 0.5% of the total cases, shows the result of when the sensor fault effect is not removed via sensor isolation. This case may imply the existence of high-importance sensors that cannot be isolated in accident diagnosis, i.e., their isolation cannot be recovered by other sensors like in Case 2.

3.3. Additive logics in the integrated system

Among the observed cases, Cases 1 and 3 demonstrated the successful execution of the suggested model. As Case 4 reflected a situation in which sensor isolation does not support accident diagnosis, this case is excluded from further analysis. On the other hand, Case 2 of the pilot study revealed a negative aspect of the suggested system, and therefore additive processing measures need to be applied. The issue in Case 2 is that the isolation of an important sensor interrupts the stable true diagnosis trend. The reasons for this negative effect are believed to be that the isolated sensor is a decisive source for accident diagnosis and that there are no other features that sufficiently represent the occurred accident. But as the accident symptoms accumulate over time, the diagnosis result is recovered to the true accident label, as in Fig. 7(b).

In Case 2, one fact to notice is the GRU result, for which the sensor is not isolated with the masking maintained as 1 even if a sensor fault is detected (see Section 2.3), indicates the true diagnosis result. For improving the diagnosis performance against Case 2, additive logic processing of the diagnosis output would be helpful. We recognized that the failed cases from the sensor isolation showed high instability and no specific label was maintained at a high value due to the insufficient basis for ensuring the diagnosis. Taking this observation into account, we applied two additive diagnosis decision logic strategies as below, each with multiple logics, and then conducted a case study with the integrated system applying the additive logics in the next section.

3.3.1. Output processing logic

The first strategy was to apply an additional accident decision logic by processing the probabilities of each label. As shown in Fig. 7(b), there is a region in the results after sensor isolation in Case 2 with unstable peaks of false labels. In Case 3, there is also a region of diagnosis failure from an accumulated fault effect. To design a logic effective for Cases 2 and 3, a strategy considering the probabilities within the time interval was suggested. At each time step, the diagnosis label is chosen from the summation of probabilities within the defined time interval. The two logics employed for this are as follows.

- Cumulative: Diagnosis is determined by the accumulation of the SoftMax output from the reactor trip point to the present time.

$$R_t^{acc} = \boldsymbol{argmax}\left(\sum_0^t \boldsymbol{f}^{\boldsymbol{GRUD}}(\boldsymbol{Y_t})\boldsymbol{dt}\right) \tag{9}$$

- Moving average: Diagnosis is determined by the average value of the SoftMax output for the defined time interval.

$$R_t^{acc} = \boldsymbol{argmax}\left(\sum_{t-n}^t \boldsymbol{f}^{\boldsymbol{GRUD}}(\boldsymbol{Y_t})\boldsymbol{dt}\right) \tag{10}$$

Where, Y_t is the diagnosis outputs in $M \times 1$ matrix form, where M is the number of accident labels. The argmax function refers to the generation of an identical matrix marked with an index of 1 on the maximum coordinate and an index of 0 on the other input vectors. With the argmax logic, the label having the maximum probability accumulation is highlighted as the one-hot vector, R_t^{acc}, which is the final diagnosis result.

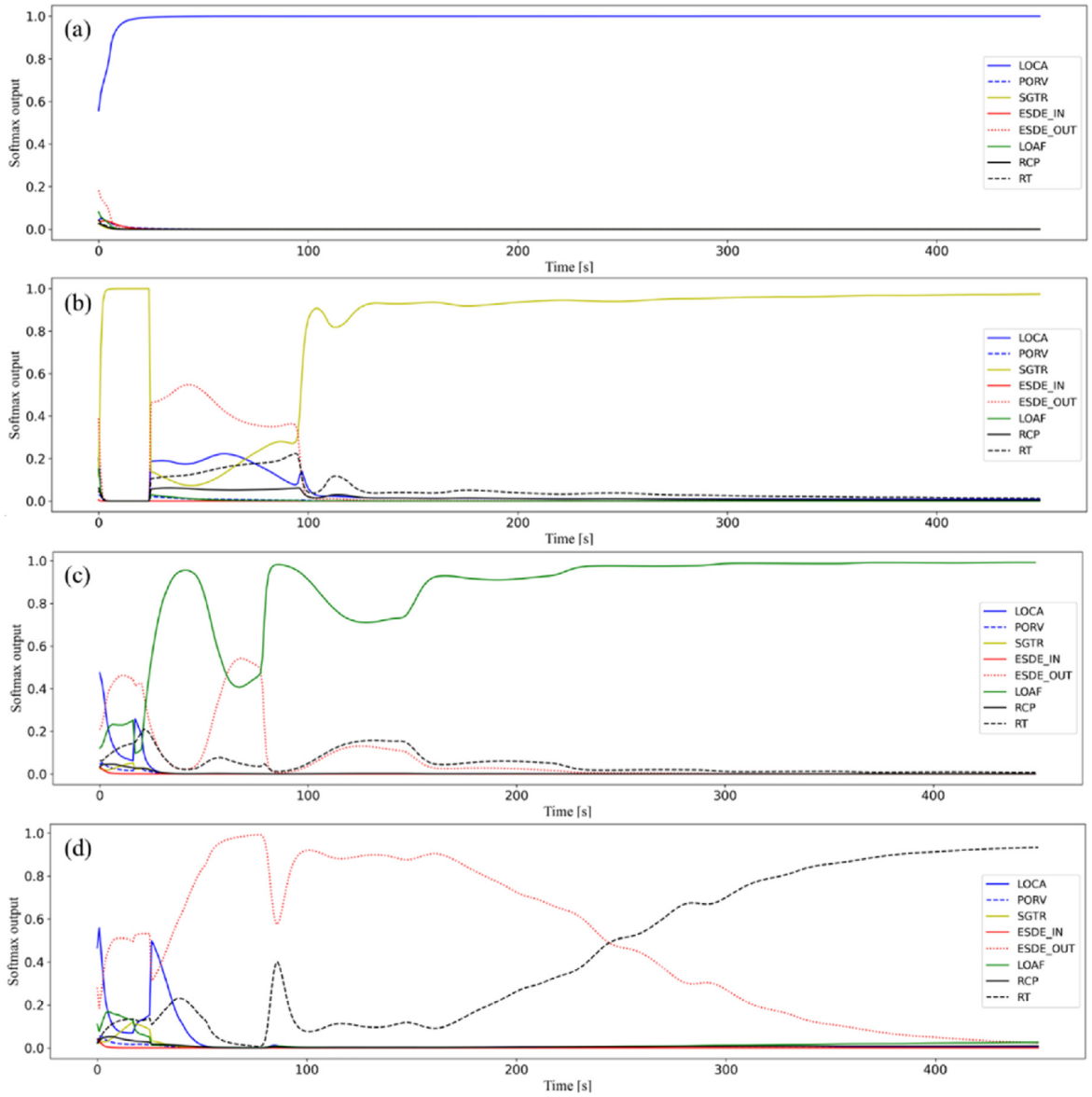

Fig. 7. Diagnosis results examples: (a): Case 1, (b): Case 2, (c): Case 3, (d): Case 4.

3.3.2. Output stability logic

The second strategy involved selective sensor isolation. As previously mentioned, we observed very unstable and inconsistent diagnosis output in Case 2, where it is believed that the isolation of an important sensor causes diagnosis failure due to the resulting lack of essential symptom information. In most cases, the diagnosis was rather successful when the sensor fault was not isolated. By an additive logic for selective sensor isolation, we intended the model to generate an output either with sensor isolation or no isolation. Before and after the sensor fault detection (isolation) points, the stabilities of the diagnosis outputs were compared. Three measures for output stability were evaluated as below.

- Variance: Variance of the SoftMax output at the time of diagnosis
- Complexity (Information theory): Information entropy of the SoftMax output at the time of diagnosis
- Hand-crafted threshold: Average of the maximum labels before and after the diagnosis point in the time interval

$$R_t^{acc} = \begin{cases} \boldsymbol{argmax}\left(\boldsymbol{f}^{\boldsymbol{GRUD}}(\boldsymbol{X_t})\right), \boldsymbol{H}(\boldsymbol{X_{t-1}}) > \boldsymbol{H}(\boldsymbol{X_{t+1}}) \\ \boldsymbol{argmax}\left(\boldsymbol{f}^{\boldsymbol{GRU}}(\boldsymbol{X_t})\right), \boldsymbol{H}(\boldsymbol{X_{t-1}}) \leq \boldsymbol{H}(\boldsymbol{X_{t+1}}) \end{cases} \quad (11)$$

4. Case study

4.1. Data description

The data source for the case study was a compact nuclear simulator developed by the Korea Atomic Energy Research Institute (KAERI) [17,18]. The compact nuclear simulator has a one-dimensional nodalization design based on the SMABRE thermal-hydraulic code for fast computation and is based on a three-loop

pressurized water reactor design by Westinghouse [19,20]. Such CNSs have been used for the data generation of several data-driven machine learning applications in the nuclear field. They can simulate emergency or accident data with typical malfunction options including several representative scenarios, and the clear symptoms for diagnosis have been precisely confirmed. Referring to existing emergency operating procedures, five accident scenarios were selected for data generation; Table 2 shows the eight accident labels with detailed break (malfunction) location and severities.

Each accident can have distinct failure features and severity. The LOCA, SGTR, and excess steam demand events are distinguished by several break locations and break sizes. The loss of all feedwater and reactor trip accidents do not have different severities from the break but vary by the timing of component failures. For example, LOCA includes the total failure of the main feedwater pumps and auxiliary feedwater pumps, where the malfunctions of the six pumps can generate various combinations of failure sequences and time intervals between failures. Based on Table 2, 1397 training data and 453 test data were generated.

4.1.1. Fault injection taxonomy

Fault data were artificially generated by injecting expected fault behaviors into the simulated data. Fault modes include sensor drifts and stuck faults, which are generally considered in prior sensor fault monitoring studies [21–23], and the faults are expected to arbitrarily produce similar signal trends with consistent sensor signals. Among the formulas below, the fault signal, $f(t)$, is generated from the original signal, $s(t)$, with the fault behavior reflected by $d(t)$.

4.1.1.1. Drift faults. A sensor drift refers to the continual and accumulative deviation of a sensor value from the correct value, as shown in Fig. 8. The fault scale, $\boldsymbol{\delta}$, determines the degree of deviation at each time step. Drift faults can be injected with diverse scales and directions. Slow and rapid drifts are divided by the degree of the fault, and upward and downward drifts are classified by the direction of the deviation. The deviated (or drifted) value is a scaled change of the present sensor value from the prior one.

$$\boldsymbol{f}(\boldsymbol{t}) = \boldsymbol{s}(\boldsymbol{t}) + \boldsymbol{d}(\boldsymbol{t}), \tag{12}$$

$$\boldsymbol{d}(\boldsymbol{t}) = \boldsymbol{\delta}^*(\boldsymbol{s}(\boldsymbol{t}) - \boldsymbol{s}(\boldsymbol{t}-\mathbf{1})) \tag{13}$$

To implement an upward drift, which refers to a positive accumulated deviation, $\boldsymbol{\delta}$ is defined as below.

$$\boldsymbol{\delta\_up} = \begin{cases} \boldsymbol{\delta}, \boldsymbol{if}, \boldsymbol{s}(\boldsymbol{t}) \geq \boldsymbol{s}(\boldsymbol{t}-\mathbf{1}) \\ \dfrac{\mathbf{1}}{\boldsymbol{\delta}}, \boldsymbol{if}, \boldsymbol{s}(\boldsymbol{t}) < \boldsymbol{s}(\boldsymbol{t}-\mathbf{1}) \end{cases} \tag{14}$$

And for a downward drift, referring to a negative accumulated deviation, $\boldsymbol{\delta}$ is defined as below.

$$\boldsymbol{\delta\_down} = \begin{cases} \dfrac{\mathbf{1}}{\boldsymbol{\delta}}, \boldsymbol{if}, \boldsymbol{s}(\boldsymbol{t}) \geq \boldsymbol{s}(\boldsymbol{t}-\mathbf{1}) \\ \boldsymbol{\delta}, \boldsymbol{if}, \boldsymbol{s}(\boldsymbol{t}) < \boldsymbol{s}(\boldsymbol{t}-\mathbf{1}) \end{cases} \tag{15}$$

4.1.1.2. Stuck faults. A stuck fault refers to a falsely constant sensor signal that differs from the original sensor value, as shown in Fig. 9. The implemented stuck faults include stuck constant and stuck zero faults. The stuck constant fault maintains a fixed sensor value at the time of the fault occurrence as below.

$$\boldsymbol{f}(\boldsymbol{t}) = \boldsymbol{s}(\boldsymbol{t_0}) \tag{16}$$

Stuck zero faults have a fixed zero value from when the fault occurs.

$$\boldsymbol{f}(\boldsymbol{t}) = \mathbf{0} \tag{17}$$

4.2. Sensor fault detection performance – sensitivity studies

Even though supervised learning-based sensor fault detection performance has been fully evaluated in a prior study [3,4], the performance of the model should be reevaluated with the newly

Table 2
Data description.

Accident type	Accident labels and break (malfunction) location	Severity
Loss of coolant accident (LOCA)	1. **LOCA**- Loop #1 cold leg - Loop #2 cold leg - Loop #3 cold leg - Loop #1 hot leg - Loop #2 hot leg - Loop #3 hot leg - Vessel top - Vessel bottom	5–150 cm^2
	2. **PORV LOCA**- Pressurizer (top)	
Steam generator tube rupture (SGTR)	3. **SGTR**- Loop #1 S/G - Loop #2 S/G - Loop #3 S/G	4–100 cm^2
Excess steam demand event (ESDE)	4. **ESDE in containment**(Main steam line break inside the containment) - Loop #1 inside-containment - Loop #2 inside-containment - Loop #3 inside-containment	In-containment: 100–1000 cm^2 Outside-containment: 360–2000 cm^2
	5. **ESDE outside containment**(Main steam line break outside the containment) - Loop #1 outside-containment - Loop #2 outside-containment - Loop #3 outside-containment	
Loss of all feedwater (LOAF)	6. **LOAF**- All feedwater pumps trip in Loop #1, 2, and 3 - All auxiliary FW pumps trip in Loop #1, 2, and 3	
Reactor trip	7. **Reactor coolant pump failure**- Reactor coolant pump trip	
	8. **Reactor protection system failure**- Spurious reactor trip from a reactor protection system failure	

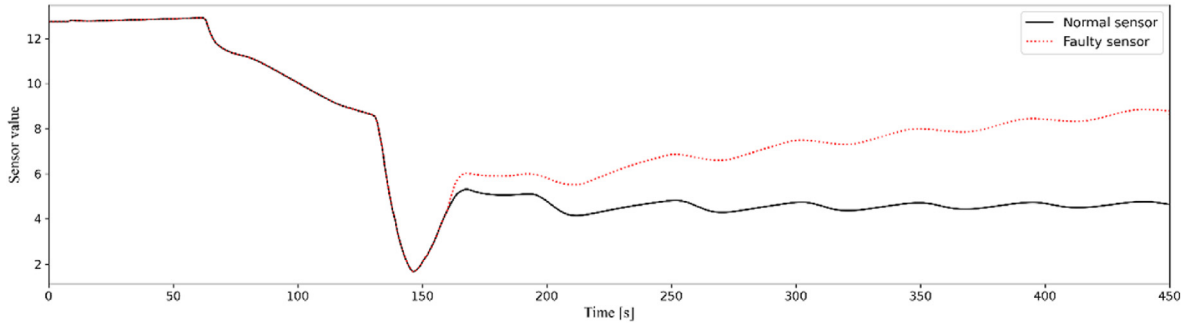

Fig. 8. Example of a sensor drift fault.

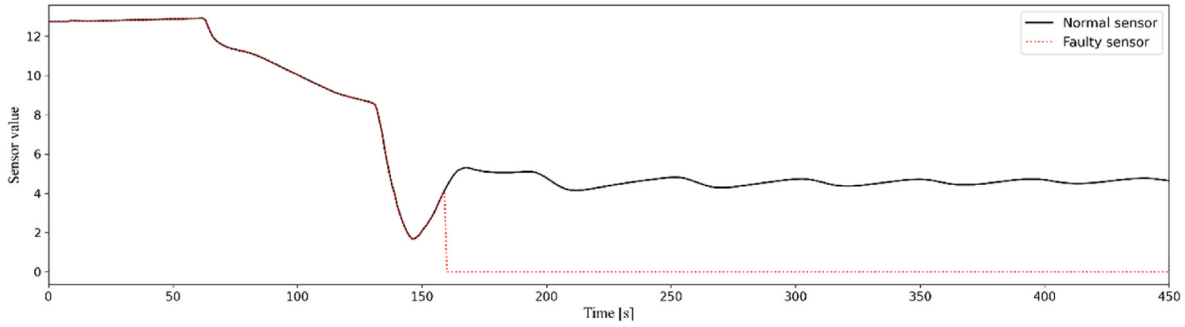

Fig. 9. Example of a sensor stuck fault.

extracted dataset. In the prior evaluation, the fault detection model achieved complete classification of both normal and fault data. However, the test results in the present study with novel data showed notably unstable consistency outputs, especially for the normal data. To prevent false cases from normal data, the model sensitivity was modulated by adjusting the labeling rule and consistency output processing logic. As explained in Section 2.2, the labeling rule of the consistency index affects the sensitivity with an error allowance that sets the consistency index directly to zero. Output processing with constrained fault decision will also prevent false positive cases. False positive cases need to be thoroughly eliminated because the false indication of a sensor fault would disable a healthy sensor and negatively impact plant safety. To confirm the true negative cases, a sensitivity study on the error allowance and output processing was conducted, with results shown in Table 3.

Overall, higher error allowance and more restricted output processing logic achieved high specificity except for the combination of 30% error allowance and 0.7 output consistency threshold. We selected several options of error allowance and output processing that satisfy the complete specificity; four options were tested with fault data prioritizing the lower criteria that achieve high sensitivity, as shown in Table 4. Among the four options, the one with 10% error allowance and 0.5 consistency threshold with 3 s averaging showed the optimal results with a complete classification of the fault data within the shortest time.

4.3. Diagnosis performance with additive decision logics

Based on the adjusted sensor fault detection performance from the sensitivity studies, the additional diagnosis decision logics suggested in Section 3.3 were applied. The logics are actuated with the transferred fault information from the fault detection model. The diagnosis accuracy of each logic is measured with the

Table 3
Specificity (true negative rate) of the sensor fault monitoring model with fault decision logic and error allowance of consistency labeling.

Consistency threshold	Error allowance	30%	20%	10%	5%
	Moving average				
0.4	1	99.96%	99.87%	99.65%	98.68%
0.5	**1**	99.96%	99.82%	99.51%	98.15%
	2	100.00%	100.00%	99.91%	99.21%
	3	100.00%	100.00%	100.00%	99.56%
	5	100.00%	100.00%	100.00%	99.78%
0.6	**5**	97.22%	96.20%	97.35%	96.60%
	10	98.63%	98.23%	98.76%	98.45%
	15	99.51%	99.34%	99.74%	99.38%
	20	100.00%	99.96%	100.00%	99.91%
0.7	**10**	92.14%	93.82%	94.08%	91.57%
	25	95.01%	99.60%	99.38%	96.91%
	30	95.45%	99.82%	99.60%	97.79%
	40	96.16%	100.00%	100.00%	98.76%

Table 4
Sensitivity (true positive rate) of sensor fault monitoring model with fault decision logic and error allowance of consistency labeling.

Error allowance	20%	10%	10%	10%
Consistency threshold	0.5	0.5	0.6	0.7
Moving average	2	3	20	40
Sensitivity	99.94% (10876/11325)	100.00% (11325/11325)	100.00% (11325/11325)	100.00% (11325/11325)
Time for detection	31.5 s	19.7 s	33.4 s	49.2 s

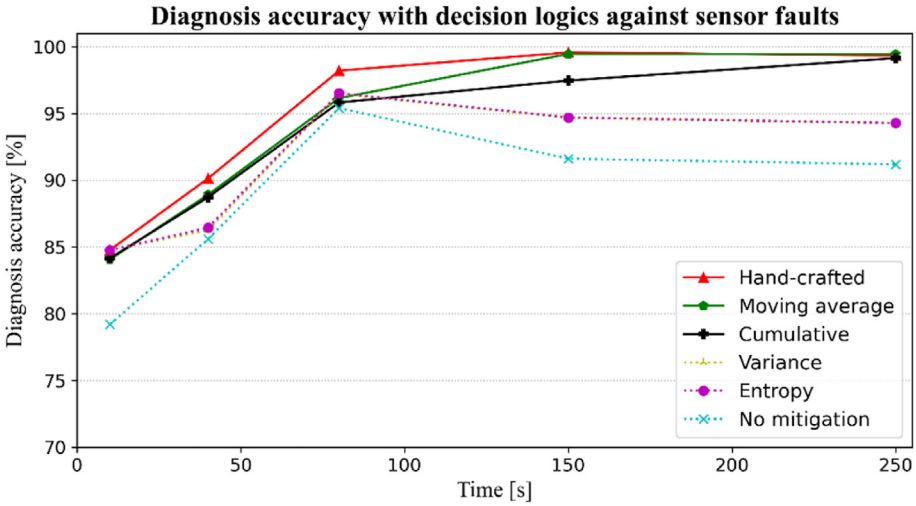

Fig. 10. Diagnosis accuracy with decision logics.

Table 5
Results of the integrated system with determined diagnosis decision logic.

#		Percentage
Case 1	Unchanging true trends	92.6%
Case 2	True trend interrupted by sensor isolation but recovered after a period of time	0.4%
Case 3	False trend from sensor fault occurred but recovered by faulty sensor isolation	6.4%
Case 4	False trend from sensor fault occurred and not recovered	0.5%

percentage of exact matches between the diagnosed label and the true accident label per time step. Fig. 10 shows the measured accuracy at the early phases of the accidents.

The first noteworthy aspect of the results is the degraded accuracy of GRU diagnosis without the fault mitigation feature. The first observations (10 s) showed a considerably lower performance of GRU due to the injected fault coincidence with the reactor trip (0 s). As the accident features accumulate, the overall accuracies including the GRU results showed increasing trends of accuracy. But after the temporary improvement (from 80 s), the diagnosis accuracy of GRU fell, which is estimated to result from the accumulation of fault severities causing more diagnosis failures in the latter

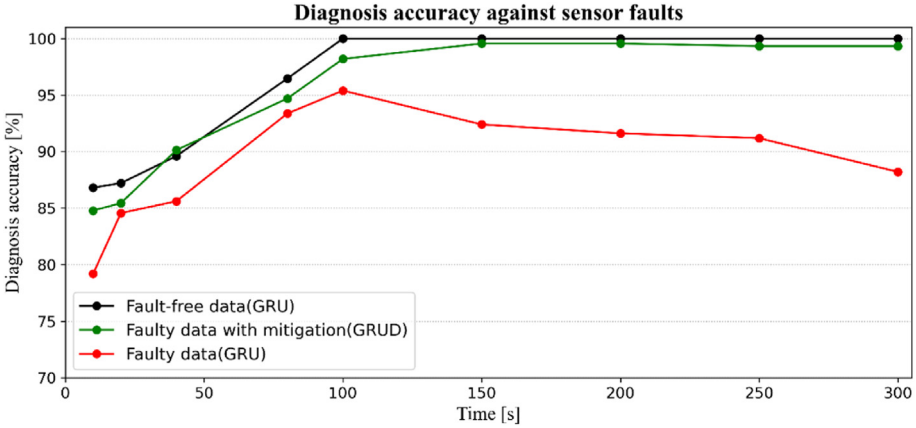

Fig. 11. Diagnosis accuracy results.

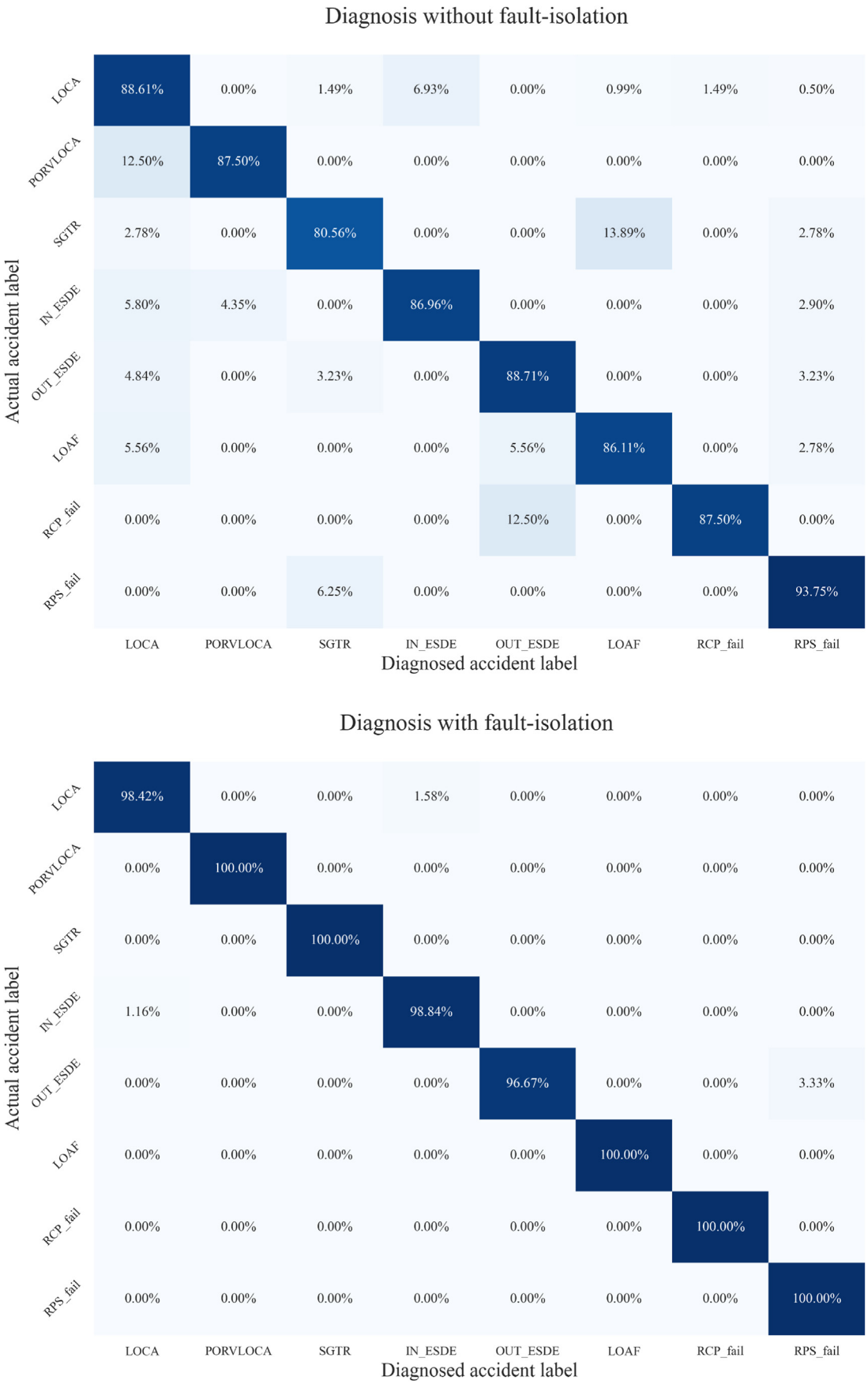

Fig. 12. Confusion matrix of diagnosis with sensor fault data without fault-isolation (upper) and with fault isolation (lower).

phases. On the other hand, the diagnosis accuracy of the GRUD models with additional decision logics showed steadily increasing trends except for the entropy-based logic. Among the compared results, the hand-crafted logic exhibited the best performance improvement. As shown in Table 5, by maintaining the proper role of fault-isolated diagnosis, the integrated system with hand-crafted logic minimizes Case 2, in which the system application causes an adverse effect by an interruption of the true diagnosis.

Applying the hand-crafted logic to the diagnosis model, Case 2 decreased to 0.4% from 1.8%. The difference goes to the Case 1 because the 1.4% formerly in Case 2 had a steady true value without the decay mechanism following the addition of the logic. The other rates were maintained.

With the optimized diagnosis output decision logic, real-time operation of the sensor fault detection and fault-tolerant accident diagnosis system was tested. Fig. 11 shows the diagnosis accuracies of GRU with fault and fault-free data as well as the constructed GRUD-based system with fault data. The results of GRU with fault-free data show a gradual increase from the start point and reach complete diagnosis before 100 s. This means that the simulated accident data contained enough symptoms for successful diagnosis, and that the GRU model was aware of the accident symptoms and generated proper accident diagnosis results. Even though the diagnosis accuracy was remarkably worsened from artificial sensor faults, the integrated system removed the influence of the sensor faults and recovered the degraded diagnosis accuracy to almost the same level as the fault-free data.

From the diagnosis results at 300 s where the diagnosis performance reached its full recovery, confusion matrices were derived as in Fig. 12 with and without the fault-isolated diagnosis strategy in the sensor fault state. The most cases were observed in the LOCA label, but the sensor faults had the biggest effect on the misdiagnosis of SGTR considering the proportions of each label. Sensor faults in the secondary sensor, which is a crucial parameter to distinguish the SGTR accident, still largely affected the diagnosis results. The other labels had an even diagnosis accuracy degradation from the sensor faults. After the fault mitigation, only 1.6% cases in the LOCA label and 3.3% of OUT_ESDE label were observed, while the other cases were completely recovered from the application of fault isolation. It is believed that these results came from the data imbalance, the largest uncertainty in the LOCA data came from having the most diverse break sizes and locations, and indistinctive symptom of OUT_ESDE accident.

5. Conclusion

The integrated sensor fault and fault-tolerant diagnosis system was applied to simulated real-time accident data from a nuclear power plant simulator. The sensor fault detection model, which has an LSTM basis, was evaluated with the newly generated dataset, and the model training and output processing were optimized via sensitivity studies. To observe the effect of a delayed transfer of fault information to the diagnosis model and the effect of feature isolation, a pilot study was conducted. Two cases showed unintended diagnosis failure due to the isolation of an important feature. Additional diagnosis decision logics in the GRUD model were compared to generate accurate diagnosis based on the optimal diagnosis output processing and whether to apply the decay mechanism. In the tested situation with sensor faults, the sensor-fault tolerant diagnosis system with added logic presented comparable diagnosis performance with senor fault-free diagnosis.

As pointed out regarding the execution of the integrated system comprising two RNNs, the real-time transfer of sensor fault information has an inevitable time delay to detect the faults, during which the diagnosis model is exposed to the influence of any faults before the related sensor is isolated by the received sensor fault information. Even though the diagnosis accuracy of the integrated system with fault data recovered to almost the same level as fault-free cases over some elapsed time, its performance at the very early phases of the simulated accidents was lower due to the delayed detection of faults and the delay from the additional decision logic of the diagnosis model. To further improve the diagnosis accuracy of the suggested system for sensor faults, the following three points should be considered: (1) improvements to the diagnosis performance of the GRU model itself, (2) upgrades of the sensor fault detection model for quicker detection of sensor faults to reduce the time delay, and (3) more sophisticated decision logics to make up for the inefficient diagnosis performance in early phases.

In the nuclear field, diverse neural network-based operator support systems targeting emergency situations have been developed, with many using RNNs. A nuclear emergency situation urgently requires prompt actions, and thus providing accurate information in a real-time manner is essential. In addition to the presented integrated system, the developed sensor fault monitoring technique can be unified with plant parameter predictions, autonomous operation, and other support systems [24–26] to prevent model failures from sensor faults.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.NRF-2018M2B2B1065653 and No.RS-2022-00144042).

References

[1] Juan Pablo Usuga Cadavid, et al., Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0, J. Intell. Manuf. 31 (2020) 1531–1558, 6.
[2] A.C. Cilliers, Benchmarking an expert fault detection and diagnostic system on the Three Mile Island accident event sequence, Ann. Nucl. Energy 62 (2013) 326–332.
[3] Jeonghun Choi, Seung Jun Lee, Consistency index-based sensor fault detection system for nuclear power plant emergency situations using an LSTM network, Sensors 20 (2020) 1651, 6.
[4] Jeonghun Choi, Seung Jun Lee, A sensor fault-tolerant accident diagnosis system, Sensors 20 (2020) 5839.
[5] Wei Li, Minjun Peng, Qingzhong Wang, Improved PCA method for sensor fault detection and isolation in a nuclear power plant, Nucl. Eng. Technol. 51 (2019) 146–154, 1.
[6] Ting-Han Lin, Shun-Chi Wu, Sensor fault detection, isolation and reconstruction in nuclear power plants, Ann. Nucl. Energy 126 (2019) 398–409.
[7] Jamie B. Coble, et al., A Review of Sensor Calibration Monitoring for Calibration Interval Extension in Nuclear Power Plants, 2012.
[8] Younhee Choi, Gyeongmin Yoon, Jonghyun Kim, Unsupervised learning algorithm for signal validation in emergency situations at nuclear power plants, Nucl. Eng. Technol. 54 (4) (2022) 1230–1244.
[9] Victor Henrique Cabral Pinheiro, Roberto Schirru, Genetic programming applied to the identification of accidents of a PWR nuclear power plant, Ann. Nucl. Energy 124 (2019) 335–341.
[10] Silvia Tolo, et al., Robust on-line diagnosis tool for the early accident detection in nuclear power plants, Reliab. Eng. Syst. Saf. 186 (2019) 110–119.
[11] Abiodun Ayodeji, Yong-kuo Liu, Hong Xia, Knowledge base operator support system for nuclear power plant fault diagnosis, Prog. Nucl. Energy 105 (2018) 42–50.
[12] Alexandre Evsukoff, Sylviane Gentil, Recurrent neuro-fuzzy system for fault detection and isolation in nuclear reactors, Adv. Eng. Inf. 19 (2005) 55–66, 1.
[13] Hang Wang, et al., Advanced fault diagnosis method for nuclear power plant based on convolutional gated recurrent network and enhanced particle swarm optimization, Ann. Nucl. Energy 151 (2021), 107934.
[14] Piero Baraldi, et al., Comparison of data-driven reconstruction methods for fault detection, IEEE Trans. Reliab. 64 (2015) 852–860, 3.

[15] Jaemin Yang, Jonghyun Kim, An accident diagnosis algorithm using long short-term memory, Nucl. Eng. Technol. 50 (2018) 582–588, 4.
[16] Zhengping Che, et al., Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (2018) 1–12, 1.
[17] Kee-Choon Kwon, et al., Compact Nuclear Simulator and its Upgrade Plan, Korea Atomic energy institute, 1997.
[18] Jae Chang Park, et al., Equipment and Performance Upgrade of Compact Nuclear simulator." No. KAERI/RR-1856/98, Korea Atomic Energy Research Institute, Daejeon, Korea, 1998.
[19] J. Miettinen, et al., Oscillations of single-phase natural circulation during overcooling transients, in: Anticipated and abnormal transients in nuclear power plants., ANS, 1987.
[20] J. Miettinen, Development and assessment of the SBLOCA code SMABRE, in: Proceedings of the CSNI Specialists' Meeting on Small Break LOCA Analyses in LWRs, 1985. Pisa, Italy.
[21] Sana Ullah Jan, Young Doo Lee, Soo Koo, A distributed sensor-fault detection and diagnosis framework using machine learning, Inf. Sci. 547 (2021) 777–796.
[22] Eliahu Khalastchi, Meir Kalech, Lior Rokach, Sensor fault detection and diagnosis for autonomous systems, in: Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems, 2013.
[23] Ansari Ahmad, Dennis S. Bernstein, Aircraft sensor fault detection using state and input estimation, in: 2016 American Control Conference (ACC), IEEE, 2016.
[24] Jung Sung Kang, Seung Jun Lee, Concept of an intelligent operator support system for initial emergency responses in nuclear power plants, Nucl. Eng. Technol. 54 (7) (2022) 2453–2466.
[25] Junyong Bae, Geunhee Kim, Seung Jun Lee, Real-time prediction of nuclear power plant parameter trends following operator actions, Expert Syst. Appl. 186 (2021), 115848.
[26] Jonghyun Kim, et al., Conceptual design of autonomous emergency operation system for nuclear power plants and its prototype, Nucl. Eng. Technol. 52 (2020) 308–322, 2.