Fault Detection and Diagnosis of Photocouplers in Reactor Protection Systems Using AI and PHM Approaches

Ho Jun Lee, Hye Seon Jo, Dong Geon Jang, Man Gyun Na*
Department of Nuclear Engr., Chosun Univ., 10, Chosundae 1-gil, Dong-gu, Gwangju, 61452
**Corresponding author: magyna@chosun.ac.kr*

Keywords: Reactor Protection System, Input Data Diagnosis, Fault Detection, Fault Diagnosis, AI

1. Introduction

The Reactor Protection System (RPS) in nuclear power plants (NPPs) plays a crucial role in protecting the reactor core and coolant system, as well as supporting accident mitigation under limiting conditions. Additionally, it monitors safety-related parameters and provides trip actuation signals. Since the 2000s, NPPs in South Korea have adopted digital RPS, replacing analog systems. The digital RPS consists of four channels, each comprising a bistable processor, coincidence processor, interface and test processor, and maintenance and test panel. These processors use the POSAFE-Q programmable logic controller (PLC). To ensure the safe operation of RPS, maintenance is performed through self-diagnosis functions, online testing during operation, and scheduled maintenance. However, these methods present certain challenges. In particular, scheduled maintenance is conducted regardless of actual component failure, leading to limitations in assessing failures during operation.

To optimize RPS maintenance, fault detection and diagnosis (FDD), a core component of prognostics and health management (PHM) technology, is applied. This study employs artificial intelligence (AI)-based RPS condition diagnosis for a photocoupler, which is one of the critical components that significantly affect RPS performance. Before applying AI, the National Institute of Standards and Technology (NIST) and the European Union Agency for Cybersecurity (ENISA) raised concerns regarding data integrity concerns, including manipulation, contamination, and environmental variations, emphasizing their potential impact on system reliability and security [1,2]. Therefore, diagnosing input data is necessary to ensure robust AI-based fault detection. In this study, we propose a framework that integrates input data diagnosis (IDD) and FDD. To achieve this, unsupervised learning models, such as long short-term memory-autoencoder (LSTM-AE), and supervised learning models, such as long short-term memory (LSTM), which demonstrated strong performance in previous studies, were utilized [3]. Additionally, bidirectional LSTM-autoencoder (BiLSTM-AE) and bidirectional LSTM (BiLSTM) were used for performance evaluation. Finally, future research directions are discussed.

2. Framework

To ensure the efficient and safe operation of the RPS, a new framework was designed by incorporating IDD and FDD. The framework utilizing IDD and FDD is shown in Fig. 1.

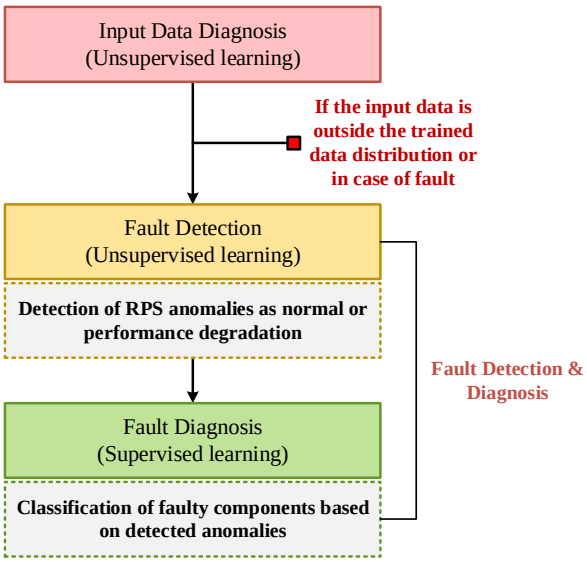


Fig. 1. IDD and FDD framework

The IDD component assesses input data to prevent incorrect data entry or situations where the AI model encounters data it has not been trained on. The FDD framework consists of fault detection and fault diagnosis. Fault detection identifies anomalies within the RPS, enabling a rapid assessment of whether the system is in a normal state or undergoing performance degradation. Fault diagnosis, on the other hand, focuses on identifying which specific component is responsible for the detected anomaly. This framework follows an FDD approach as a part of PHM, ensuring the efficient and safe operation of the RPS. Although this approach is intended to incorporate various critical electronic components within the RPS such as photocouplers and their peripheral circuits, DC-DC converters, and transient voltage suppressors this study focuses solely on the photocoupler. In this context, fault diagnosis was also implemented as a binary classification task for the photocoupler, in the same manner as fault detection.

3. Methodology

This section describes the AI methods used in this study. The supervised learning models used LSTM and BiLSTM, while the unsupervised learning models used LSTM-AE and BiLSTM-AE.

3.1 Long Short-Term Memory

LSTM is an RNN-based model developed to overcome the vanishing gradient problem, which limits the learning of long-term dependencies in RNN [4]. LSTM consists of an input gate, forget gate, output gate, and cell state, which regulate the flow of information. The input gate determines how much new information is stored, the forget gate removes unnecessary information, and the output gate controls the final output. The cell state retains important information over time, allowing LSTM to effectively learn sequential patterns. Due to this structure, LSTM is widely used in natural language processing, time-series analysis, speech recognition, and video analysis. The structure of LSTM is shown in Fig. 2.

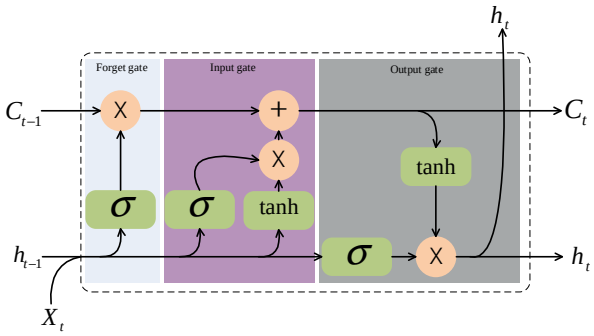


Fig. 2. Structure of LSTM

3.2 BiLSTM

BiLSTM is an advanced version of LSTM that processes both forward and backward sequences in time-series data. In contrast to standard LSTM, which processes data in a single direction, BiLSTM consists of two LSTM layers, one for forward processing and the other for backward processing. This structure enables BiLSTM to capture both past and future dependencies, enhancing its capability to capture relationships in sequential data. BiLSTM is particularly effective in tasks requiring bidirectional context, such as natural language processing, speech recognition, and time-series anomaly detection. By leveraging bidirectional information, BiLSTM improves pattern recognition accuracy and enhances anomaly detection performance compared to standard LSTM. The structure of BiLSTM is shown in Fig. 3.

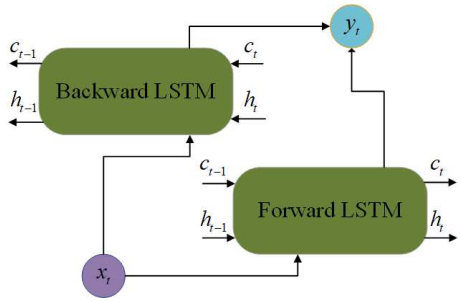


Fig. 3. Structure of BiLSTM

3.3 LSTM-Autoencoder

LSTM-AE is an unsupervised learning model that integrates LSTM with an AE structure to learn and reconstruct sequential data. AE compresses input data using an encoder and reconstructs it through a decoder. LSTM-AE follows this structure but replaces all units with LSTM layers, enabling it to capture long-term dependencies more effectively. It is primarily used for dimensionality reduction, feature extraction, and anomaly detection in time-series data. The structure of LSTM-AE is shown in Fig. 4.

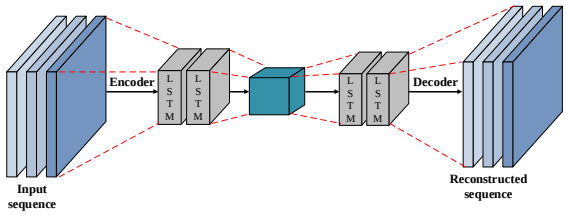


Fig. 4. Structure of LSTM-AE

3.4 BiLSTM-Autoencoder

BiLSTM-AE is an unsupervised learning model that integrates BiLSTM into the AE structure. The structure of BiLSTM-AE is shown in Fig. 5[5].

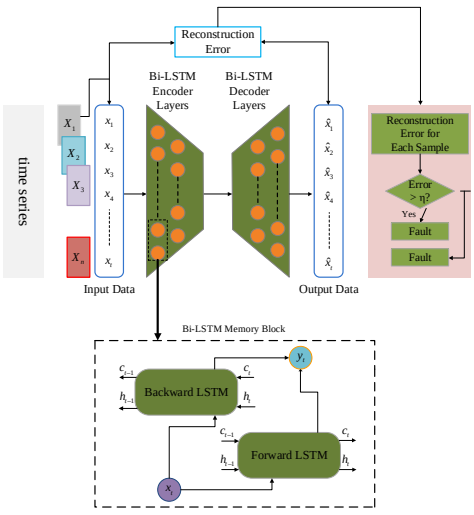


Fig. 5. Structure of BiLSTM-AE

Similar to LSTM-AE, BiLSTM-AE follows the AE structure but uses BiLSTM units in the encoder, enabling it to learn patterns more effectively and detect anomalies in time-series data. BiLSTM-AE is particularly useful for tasks that require bidirectional information.

4. Data Preparation

This section describes the data used in this study. The data consists of photocoupler accelerated aging data. Photocouplers were selected because they are electronic components within RPS that can critically impact RPS failures.

4.1 Data Acquisition

This study utilized accelerated aging data from photocouplers, critical electronic components in RPS, to analyze performance degradation. Since direct collection of photocoupler failure data in NPPs is challenging due to long maintenance cycles, Korea Atomic Energy Research Institute conducted an accelerated aging experiment, exposing 40 photocouplers to 130°C for 92 days. Data acquisition was performed every 5 seconds, measuring input and output voltages.

4.2 Data Preprocessing

To assess degradation, the Current Transfer Ratio (CTR) was used as a performance indicator. However, due to the identical input and output resistances, the input-output voltage ratio was adopted as a practical alternative. The degradation point was defined as 95% of the nominal CTR, with performance degradation observed starting on the 88th day. Since electronic components such as photocouplers are designed for long-term operation and do not exhibit a clear failure point, a conservative and safety-oriented approach was applied. Accordingly, the 20th percentile value of the CTR distribution was used as the threshold for anomaly detection. The distribution of the data used is shown in Fig. 6.

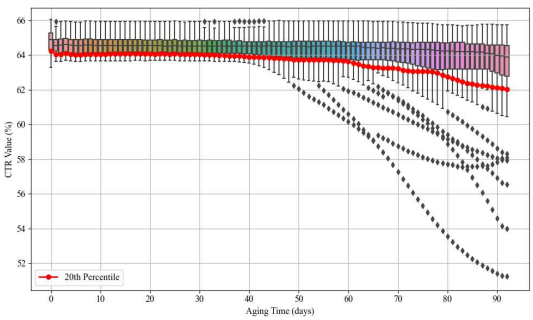


Fig. 6. Data distribution and 20th percentile trend of 40 photocouplers

Due to more severe experimental conditions compared to actual NPPs environments, data transformation was performed based on the Arrhenius equation, resulting in 10 overlapping temperature cases (30°C–72°C in 6°C increments with 2°C overlap). Table I presents the temperature dataset by range.

Table I: Temperature Dataset by Range

No.	Temperature range (°C)
1	30 - 36
2	34 - 40
3	38 - 44
4	42 - 48
5	46 - 52
6	50 - 56
7	54 - 60
8	58 - 64
9	62 - 68
10	66 - 72

For AI model training, min-max normalization was used to scale the data between 0 and 1, and the sliding window technique was applied for time-series pattern recognition. The IDD model was trained until the performance degradation point, while the FDD model was trained until 30 days prior to the degradation point. For the input variables of the AI model, only operating time and temperature were used for training, considering variables that can be directly utilized in the field.

5. Results

This section describes the results of IDD and FDD. Methodologies based on previous research, including LSTM and LSTM-AE, were used. additionally, BiLSTM and BiLSTM-AE were employed to compare and analyze both supervised and unsupervised learning models.

5.1 Hyperparameter Optimization of AI Models

Grid search was employed to optimize the performance of the AI models. The LSTM and BiLSTM models, as well as the LSTM-AE and BiLSTM-AE models, share the same architecture, with the only difference being the use of bidirectional layers in the BiLSTM variants. Table II summarizes the structure and configuration of each model based on the optimal hyperparameters identified through grid search.

Table II: Optimized Model Configurations by Grid Search

Parameter	(Bi)LSTM-AE	(Bi)LSTM
Number of layers	8	7
Hidden units	512 → 256 → 256 → 512	512 → 256 → 128
Activation	-	Softmax

Optimizer	Adam	Adam
Learning Rate	0.001	0.001
Batch size	512	1024
Loss	MSE	Categorical crossentropy
Input shape	(10, 2)	(10, 2)

5.2 Performance Evaluation Metrics

The performance evaluation metrics included Accuracy (ACC), F1-score, and area under the curve (AUC). ACC indicates the overall ACC of the model, while Receiver Operating Characteristic (ROC)-AUC assesses the model's overall performance. F1-score is used to address data imbalance. Both ROC-AUC and F1-score are particularly effective for evaluating anomaly detection and binary classification performance. Table III presents the variables used to calculate each performance metric.

Table III: Variables for Performance Metrics Calculation

Variables	Description
True Positive (TP)	Correctly predicted positive samples
True Negative (TN)	Correctly predicted negative samples
False Positive (FP)	Incorrectly predicted positive samples
False Negative (FN)	Incorrectly predicted negative samples

ACC represents the proportion of correctly classified samples among all samples, where values close to 1 indicate higher model performance. ACC is defined in Eq. (1).

$$(1)\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

ROC-AUC is the area under the ROC curve, which measures the model ability to differentiate between positive and negative classes based on various thresholds. Values close to 1 indicate high model performance, while values near 0.5 indicate random classification. For a given model and threshold, the corresponding coordinate points (X = false positive rate (FPR), Y = true positive rate (TPR)) are derived from the true labels and predicted values for all samples. ROC-AUC is computed using Eqs. (2)-(3).

The TPR, also known as recall, denotes the ratio of correctly predicted positive samples to the total actual positive samples:

$$(2)\ TPR = \frac{TP}{TP+FN}$$

The FPR denotes the ratio of actual negative samples that are misclassified as positive:

$$(3)\ FPR = \frac{FP}{FP+TN}$$

F1-score is the harmonic mean of precision and recall, providing a balanced evaluation in imbalanced datasets. A higher F1-score reflects the model ability to capture both precision and recall, making it suitable for scenarios requiring a balance between false positives and false negatives. F1-score is computed using Eqs. (4)-(6).

$$(4)\ Precision = \frac{TP}{TP+FN}$$

$$(5)\ Recall = \frac{TP}{TP+FN}$$

$$(6)\ F1-score = 2\times\frac{Precision\times Recall}{Precision+Recall}$$

5.3 Input Data Diagnosis and Fault Detection Results

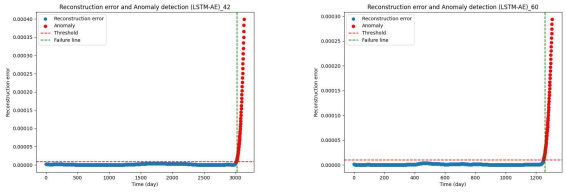
The IDD model was developed using unsupervised learning methods, while in the FDD framework, the fault detection component was also implemented using unsupervised learning models. As described in the data preprocessing section, these models were designed with different failure time points. For model development, a multi-model approach was implemented, where separate models were constructed for different temperature ranges using a tree-based structure. This approach ensures that each model is tailored to the specific temperature range, enabling more accurate diagnosis. 10 models were developed, corresponding to the temperature datasets in Table I. Additionally, to mitigate instability in overlapping temperature regions, the model boundaries were intentionally overlapped, enhancing robustness.

Both the IDD model and the fault detection component in the FDD framework were evaluated by comparing the performance of LSTM-AE and BiLSTM-AE. The models were trained using a window size of 10. Both models demonstrated good performance, with LSTM-AE demonstrating slightly superior performance. The performance evaluation results for the IDD model, showing the average performance of the 10 models, are given in Table IV. The performance results of the IDD model are visualized in Fig. 7.

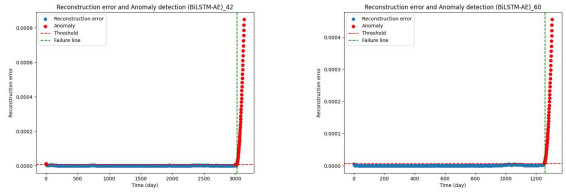
Table IV: Performance evaluation results of IDD

Method	Data	ACC	AUC	F1-score
LSTM-AE	Train	0.9964	0.9995	0.9513
	Validation	0.9962	0.9996	0.9494
	Test	0.9964	0.9995	0.9516
BiLSTM-AE	Train	0.9958	0.9995	0.9428
	Validation	0.9959	0.9995	0.9436

	Test	0.9958	0.9995	0.9432



(a) Result of LSTM-AE



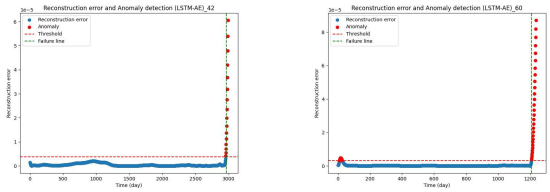
(b) Result of BiLSTM-AE

Fig. 7. Performance Visualization of IDD for 42°C and 60°C

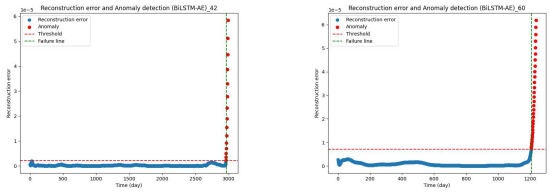
Table V presents the performance evaluation results for the fault detection model, which presents the average performance across 10 models. Training was conducted with a window size of 10. Both models demonstrated good performance, with LSTM-AE demonstrating slightly better performance. Fig. 8 illustrates the performance results of the fault detection model.

Table V: Performance Evaluation of LSTM-AE Based Fault Detection

Method	Data	ACC	AUC	F1-score
LSTM-AE	Train	0.9987	0.9997	0.9029
	Validation	0.9970	0.9997	0.9089
	Test	0.9969	0.9996	0.9056
BiLSTM-AE	Train	0.9966	0.9999	0.9032
	Validation	0.9967	0.9999	0.9079
	Test	0.9966	0.9999	0.9054



(a) Result of LSTM-AE



(b) Result of BiLSTM-AE

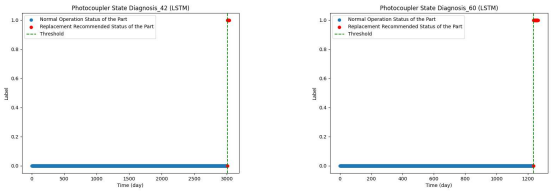
Fig. 8. Performance Visualization of Fault Detection for 42°C and 60°C

5.4 Fault Diagnosis

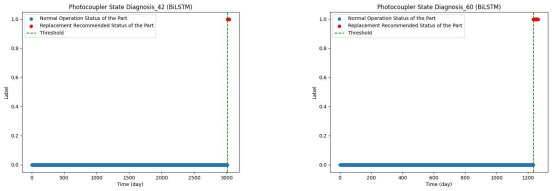
The fault diagnosis model was developed using a supervised learning approach. In contrast to the fault detection model, which employed multiple models for different temperature ranges, the fault diagnosis model used a single-model approach, where one model was trained on data from all temperature ranges. The fault diagnosis model was implemented using LSTM. In this supervised learning approach, the performance of BiLSTM was slightly superior to that of LSTM. Table VI presents the performance evaluation results for the fault diagnosis model. Fig. 9 illustrates the performance results of the fault diagnosis model.

Table VI: Performance Evaluation Results of Fault Diagnosis

Method	Data	ACC	AUC	F1-score
LSTM	Train	0.9988	0.9999	0.9677
	Validation	0.9989	0.9999	0.9713
	Test	0.9987	0.9999	0.9650
BiLSTM	Train	0.9990	0.9999	0.9729
	Validation	0.9990	0.9999	0.9782
	Test	0.9991	0.9999	0.9741



(a) Result of LSTM



(b) Result of BiLSTM

Fig. 9. Performance Visualization of Fault Diagnosis for 42°C and 60°C

6. Conclusion

This study proposes a framework that combines IDD and PHM to improve the maintenance and operational efficiency of the RPS. The goal is to ensure the safety of AI-based systems by addressing data integrity issues, while optimizing RPS maintenance through FDD, an important component of PHM. To develop AI models for IDD and FDD, this study used accelerated aging data from photocouplers, which are critical electronic components in the POSAFE-Q PLC. For IDD and fault

detection, the performance of unsupervised learning models, including LSTM-AE and BiLSTM-AE, was compared. For fault diagnosis, the performance of supervised learning models, such as LSTM and BiLSTM, was evaluated.

The results showed that LSTM-AE performed slightly better than BiLSTM-AE in IDD and fault detection, while BiLSTM outperformed LSTM in fault diagnosis. the results of this study imply that IDD enhances AI model safety by verifying input data. In addition, AI-based FDD methods can improve RPS maintenance strategies and reduce unnecessary operational costs during scheduled overhaul periods. The proposed framework allows for rapid fault detection and accurate fault classification, enabling a proactive maintenance approach that enhances the safe and efficient operation of RPS. future research will extend the model to incorporate additional critical electronic components beyond photocouplers and develop an RPS FDD model that considers temperature variations.

Acknowledgment

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-00144239) and the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government (MOTIE) (20224B10100120, Development of commercialization technology for failure diagnosis of reactor control and digital I&C systems).

REFERENCES

[1] National Institute of Standards and Technology (NIST), “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” U.S. Department of Commerce, Washington, DC, USA, pp. 7-8, 2023.
[2] European Union Agency for Cybersecurity (ENISA), “AI Cybersecurity Challenges,” ENISA, Greece, 2020.
[3] H. J. Lee, S. H. Lee, and M. G. Na, “Preliminary Modeling and Applicability Evaluation for Condition Diagnosis and Failure Detection in Reactor Protection System,” Transactions of the Korean Nuclear Society Spring Meeting, Jeju, Korea, May 9-10, 2024.
[4] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, Vol. 9, No. 8, pp. 1735-1780, 1997.
[5] A. S. Raihan and I. Ahmed, “A Bi-LSTM Autoencoder Framework for Anomaly Detection - A Case Study of a Wind Power Dataset,” Proceedings of the 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE), Auckland, New Zealand, Aug. 26–30, pp. 1–6, 2023.