Abnormal Diagnosis using eXplainable Artificial Intelligence in Nuclear Power Plants

Ji Hun Park, So Hun Yoon, Ye Ji An, Man Gyun Na *

Chosun University

Transactions of the Korean Nuclear Society Virtual Autumn Meeting
Korea, December 17-18, 2020

Contents

01 Introduction

01 Introduction

- **Background**
  - If the Nuclear Power Plants (NPPs) deviates from normal state, the operator must select the corresponding procedure (AOP, EOP) to identify events and perform mitigation actions.

Event identification according to AOP
[Success]

Take mitigation measures
for identified incident
[Success]

Normal state

Event identification according to AOP
[Failure]

Take mitigation measures
for identified incident
[Failure]

The psychological burden of the operators ↑

→ Human Error ↑

CHOSUN UNIVERSITY

01 Introduction

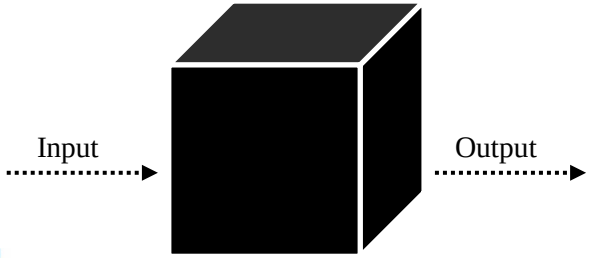
Background

- Diagnosis failure due to human error can adversely affect (safety and economics) NPPs.
- It aims to reduce human error by applying Artificial Intelligence (AI) to the operator's diagnosis task.

Operator's diagnosis task
[Too much information and time pressure]

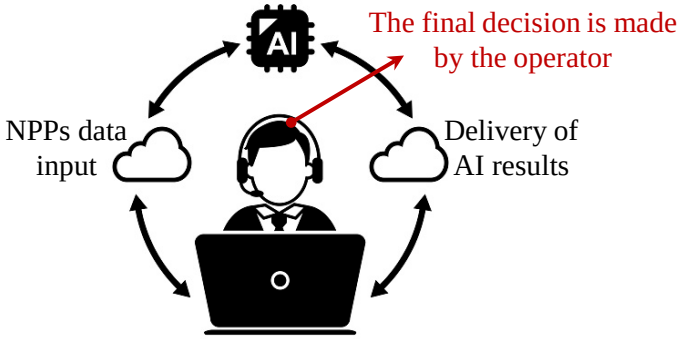
Operator's diagnosis task assistance using AI
→ Reduction of human error

- However, the reliability of AI and the responsibility of the operator become a problem.



Black Box [AI]

Why did the results come out like this?



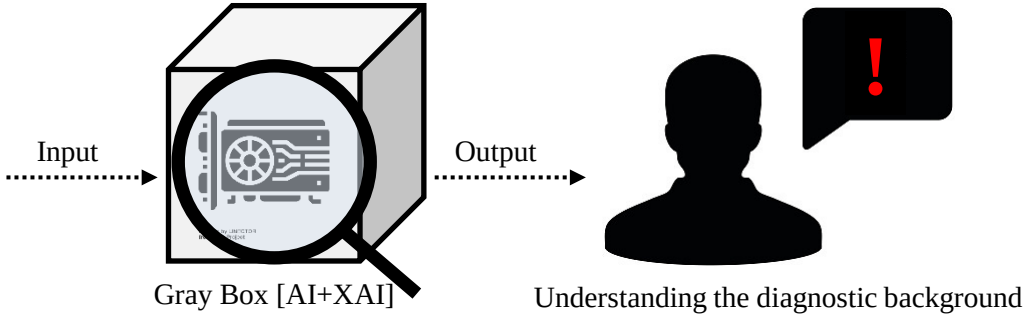
Operator Support System

CHOSUN UNIVERSITY

01 Introduction

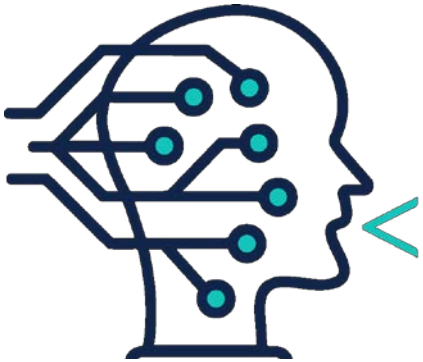
Objective

- Development of abnormal diagnosis support system using AI for abnormal states of NPPs.
- Improving the reliability of AI through the application of eXplainable AI (XAI).



Integrated Abnormal Diagnosis System (IADS)

eXplainable Artificial Intelligence



Integrated Algorithm
[Diagnosis module (based on AI) +
Verification module (based on XAI)]

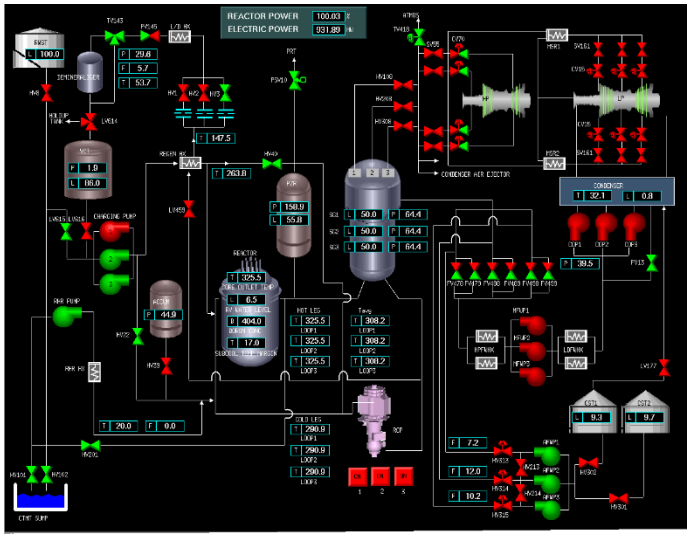
Interpretation of diagnosis results
↓
Derivation of diagnosis evidence → Interface development [Provide efficient information]
↓
Improved AI reliability

02 Data Collection and Pre-processing

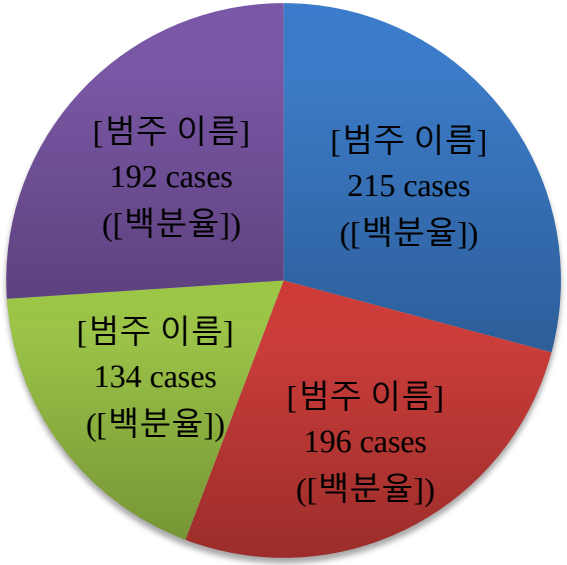
02 Data Collection and Pre-processing

Data Collection using CNS

- Data collection was performed using Compact Nuclear Simulator (CNS) designed based on Westinghouse 930 MWe 3 Loop Pressurized Water Reactor (PWR).
- IADS is aimed at diagnosing abnormal state. Since the abnormal state consists of various scenarios, the criteria for selecting scenarios must be established.
- 1) Automatic logic and instrument errors, 2) Device status abnormality, 3) Pipe leakage



Compact Nuclear Simulator Interface



The Analysis of accident records released OPIS (1978-2018)

CHOSUN UNIVERSITY

02 Data Collection and Pre-processing

List of collected data

- 20 abnormal scenarios and 1 normal scenario were collected, and the malfunction injection was set to 30 seconds. [Total 21 scenarios]

No.	Label	Name	No. of data (train / test)	Entire data matrix (row, column)
		Normal state	20 (15, 5)	(8,604, 2,222)
<Automatic logic and instrument errors>				
1	Ab21-01	PRZ pressure channel failure ‘High’	18 (15 / 5)	(4,698, 2,222)
2	Ab21-02	PRZ pressure channel failure ‘Low’	26 (21 / 5)	(5,212, 2,222)
3	Ab20-01	PRZ water level channel failure ‘High’	6 (0 / 6)	(3,769, 2,222)
4	Ab20-04	PRZ water level channel failure ‘Low’	15 (12 / 3)	(7,954, 2,222)
5	Ab15-07	S/G water level channel failure ‘Low’	40 (35 / 5)	(8,912, 2,222)
6	Ab15-08	S/G water level channel failure ‘High’	40 (35 /5)	(11,384, 2,222)

- PRZ: Pressurizer
- S/G: Steam Generator

CHOSUN UNIVERSITY

02 Data Collection and Pre-processing

List of collected data

- 20 abnormal scenarios and 1 normal scenario were collected, and the malfunction injection was set to 30 seconds. [Total 21 scenarios]

No.	Label	Name	No. of data (train / test)	Entire data matrix (row, column)
<Device status abnormality>				
7	Ab63-04	Control rod fall	40 (40 / 8)	(46,507, 2,222)
8	Ab63-02	Continuous insertion of control rod	8 (6 / 2)	(4,363, 2,222)
9	Ab63-03	Continuous withdrawal of control rod	8 (0 / 8)	(2,689, 2,222)
10	Ab21-12	PRZ PORV opening	52 (45 / 7)	(13,573, 2,222)
11	Ab19-02	PRZ safety valve opening	51 (45 / 6)	(17,370, 2,222)
12	Ab21-11	Faulty opening of PRZ spray valve	50 (45 / 5)	(31,391, 2,222)
13	Ab59-01	Charging pump failure stop	1 (0 / 1)	(678, 2,222)
14	Ab80-02	Stop 2/3 of the main feed water pump turbines	3 (0 / 3)	(3,400, 2,222)
15	Ab64-03	Isolation of main steam line	3 (0 / 3)	(142, 2,222)

- PORV: Power Operated Relief Valve

02 Data Collection and Pre-processing

List of collected data

- 20 abnormal scenarios and 1 normal scenario were collected, and the malfunction injection was set to 30 seconds. [Total 21 scenarios]

No.	Label	Name	No. of data (train / test)	Entire data matrix (row, column)
<Pipe leakage>				
16	Ab60-02	Rupture of the front end of the RHX	50 (45 / 5)	(32,857, 2,222)
17	Ab23-03	Leakage from CVCS to CCW	50 (45 / 5)	(40,498, 2,222)
18	Ab59-02	Leakage at the rear end of the charging water flow control valve	30 (25 / 5)	(20,313, 2,222)
19	Ab23-01	Leakage from RCS to CCW	30 (25 / 5)	(2,900, 2,222)
20	Ab23-06	S/G tube rupture	36 (30 / 6)	(2,738, 2,222)

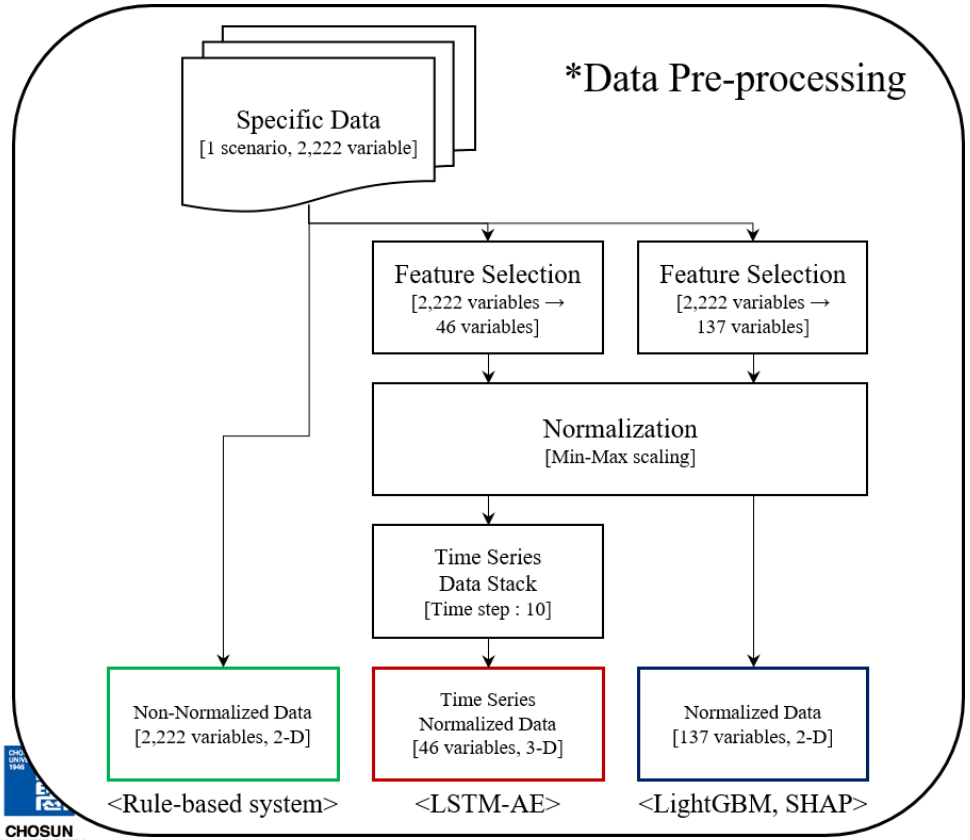
- RHX: Regenerative Heat eXchager
- CVCS: Chemical Volume Control System
- CCW: Component Cooling Water
- RCS: Reactor Coolant System

CHOSUN UNIVERSITY

02 Data Collection and Pre-processing

Data Pre-processing

- IADS is implemented using AI, XAI, and Rule-based system.
  - ✓ Since each methodology has a different type of data required, it is necessary to transform the raw data through data pre-processing.



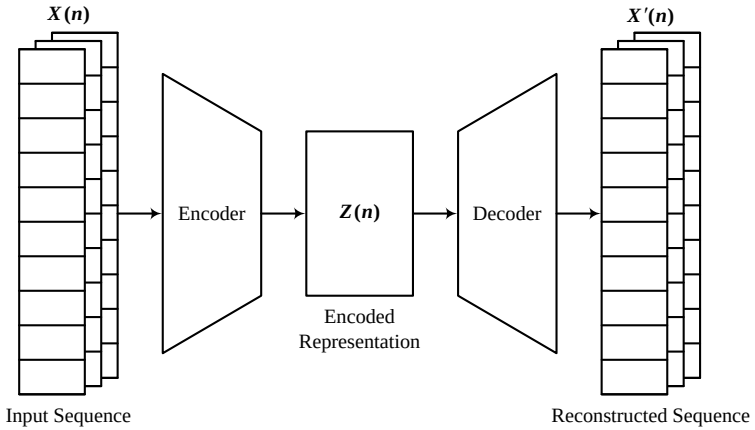
- Feature Selection
  - Technique to extract only the desired data out of the total 2,222 variables
- Normalization
  - Technique to place variables between 0 and 1 according to certain rules → Avoid data bias
  - Use Min-Max Normalization

  * Eq. 1) $Xscale = (X - Xmin)/(Xmax - Xmin)$
- Time Series Data Stack
  - Technique that generates 3-D sequence data by accumulating data according to time intervals

03 Method

03 Method

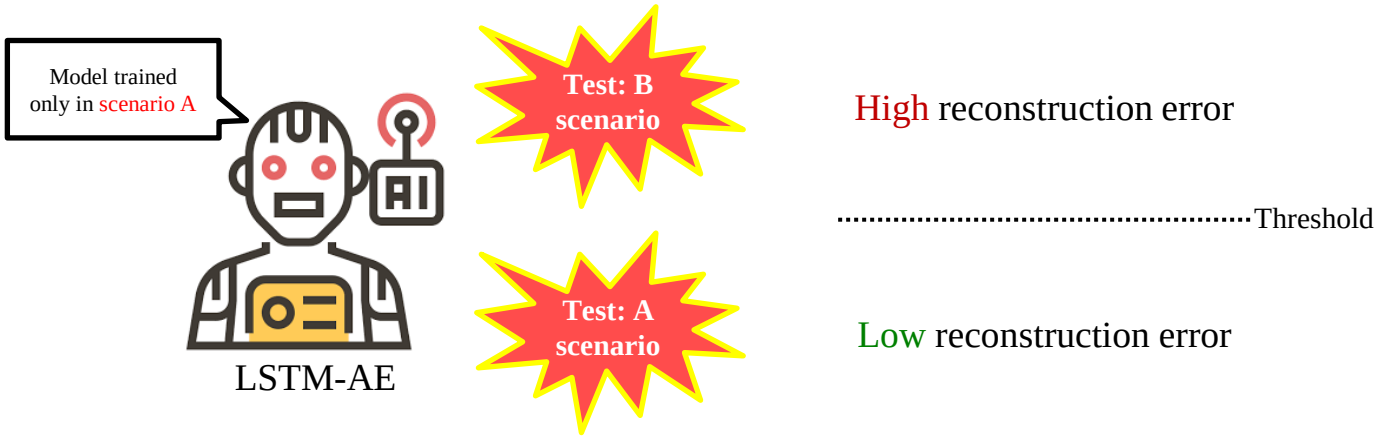
- **LSTM-AE**
  - Synthesis of LSTM method to utilize time series characteristics and AE method to reconstruct input data
    - ✓ Mainly used for error detection or data generation
  - There are two characteristics.
    - ✓ Learning efficiency increases when there are few input variables.
      - 46 input variables were selected by synthesizing symptoms for each scenario.
    - ✓ Easy to learn unbalanced data (The amount of data by category is different)
      - There is a phenomenon in which the amount of data is biased toward a large amount. → Solve by training only one category [One-Class Classification]



<LSTM-AE structure>

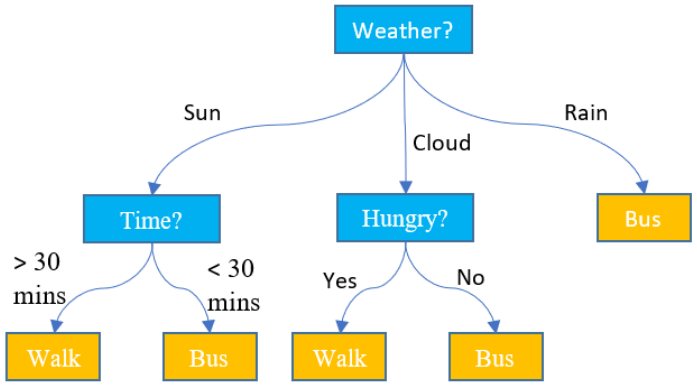
03 Method

- **LSTM-AE reconstruction error and threshold**
  - LSTM-AE reconstruction error
    - ✓ The reconstruction error is defined as the result of calculating the input data and output data through Eq. 2.
      - * Eq. 2) $L(x, x') = \|x - x'\|^2$
  - LSTM-AE threshold
    - ✓ The threshold value is the result of using 3 sigma for the reconstruction error calculated through Eq. 2 for input data and output data [using train data].
    - ✓ Based on the mean value and standard deviation of the reconstruction error, the threshold is calculated as shown in Eq. 3.
      - * Eq. 3) Threshold $= \mu \pm 3\sigma$



03 Method

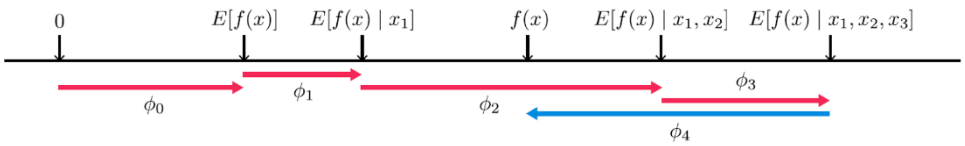
- **LightGBM**
  - It is a model that developed high-performance Gradient Boosting Decision Tree (GBDT) and solved the problem caused by big data that is emerging recently. (Time consuming ↑)
    - ✓ Gradient-based One-side Sampling (GOSS)
      - Technique to reduce the number of training data: Excluding a large part of data with small gradient
    - ✓ Exclusive Feature Bundling (EFB)
      - Technique for reducing variables in training data: Bundling mutually exclusive variables
  - 137 input variables were selected using the feature importance method that calculates the importance of variables supported by the LightGBM model.



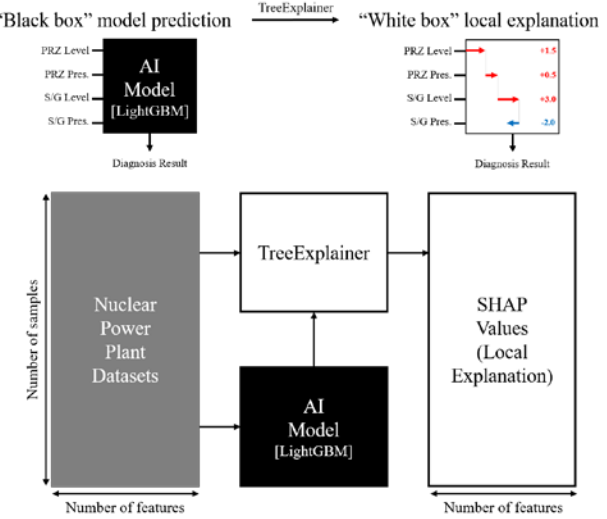
<Decision Tree structure>

03 Method

- **SHAP**
  - The key idea is the independence between variables, which allows to calculate numerically how much each variable contributed to creating to overall outcome.
    - ✓ When the contribution of each variable is excluded, the degree of change in overall performance can be calculated. (Refer to Eq. 4)
      - * Eq. 4) $\phi_i(v) = \sum_{S \in N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!} (v(S \cup \{i\}) - v(S))$



<SHAP application result form>

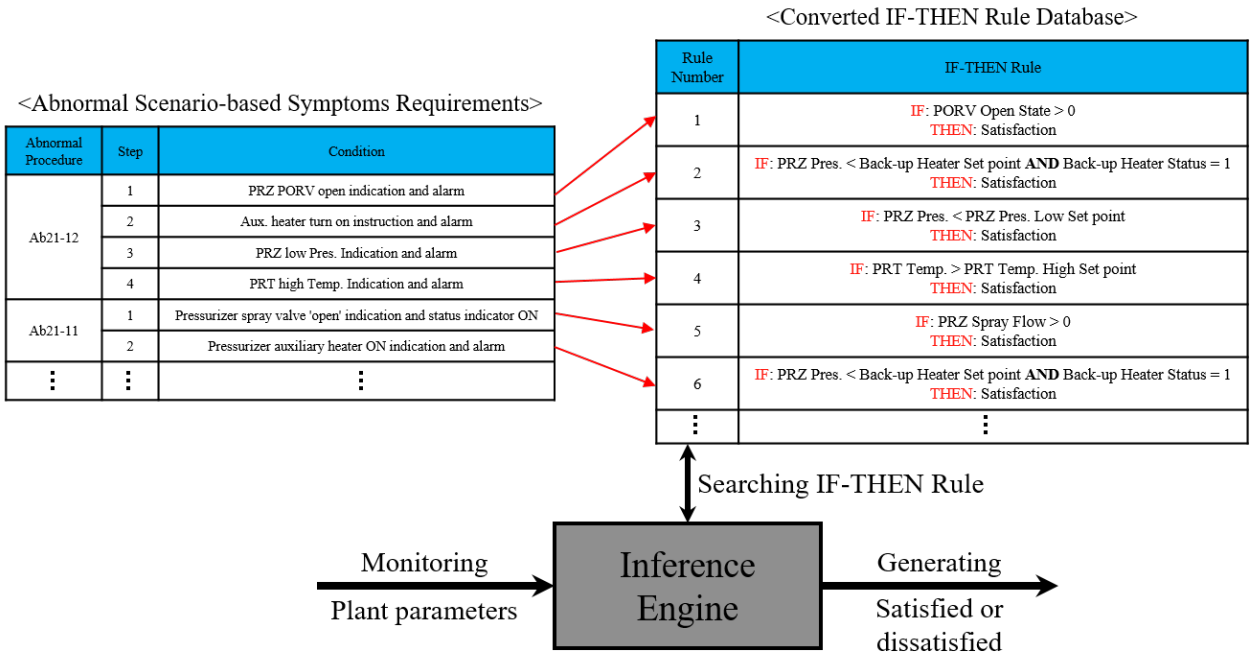


<SHAP application structure>

03 Method

Rule-based system

- System that deal with problems based on clear rules (also called expert system).
  - ✓ Rule-based system components
    - Knowledge base: knowledge needed to solve the problem
    - Database: Comparing to rules stored in the knowledge base
    - Inference engine: Connect knowledge base and database to perform inference

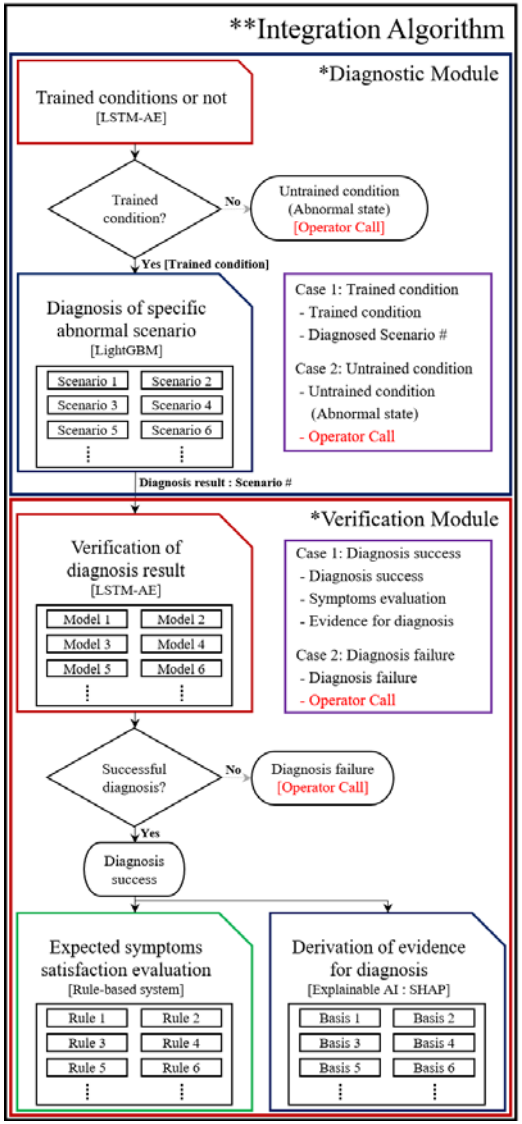


<Rule-based system structure>

04 Integrated Abnormal Diagnosis Algorithm

04 Integrated Abnormal Diagnosis Algorithm

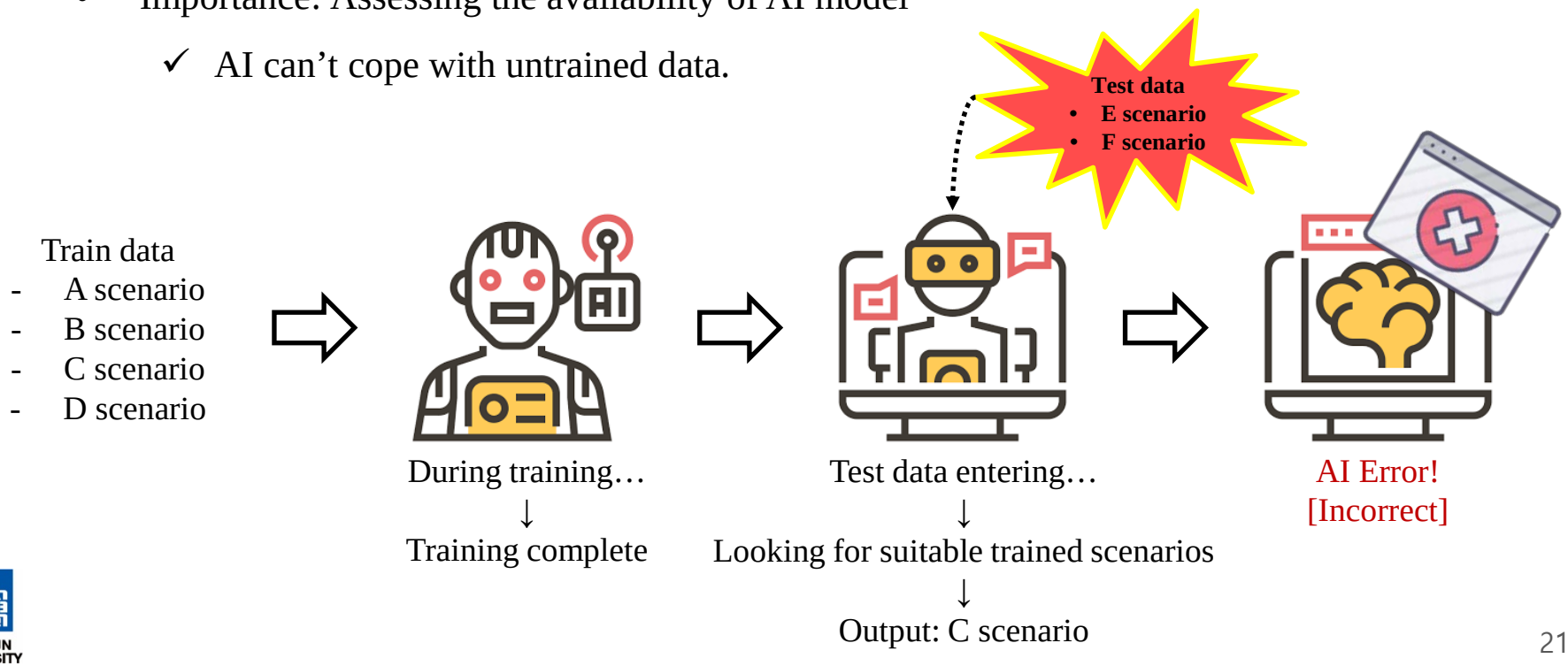
Integrated Abnormal Diagnosis Algorithm Overview



- ❖ Diagnostic Module (based on AI)
  - ✓ Training status diagnosis function
    - Trained condition → next function
    - Untrained condition → operator call
  - ✓ Abnormal scenario diagnosis function
    - Specific abnormal scenario → next module
- ❖ Verification Module (based on XAI)
  - ✓ Diagnosis result verification function
    - Diagnosis success → next function
    - Diagnosis failure → operator call
  - ✓ Expected symptoms satisfaction evaluation function
    - Symptom satisfaction or dissatisfaction
  - ✓ Diagnosis evidence deriving function
    - Diagnosed and undiagnosed evidence

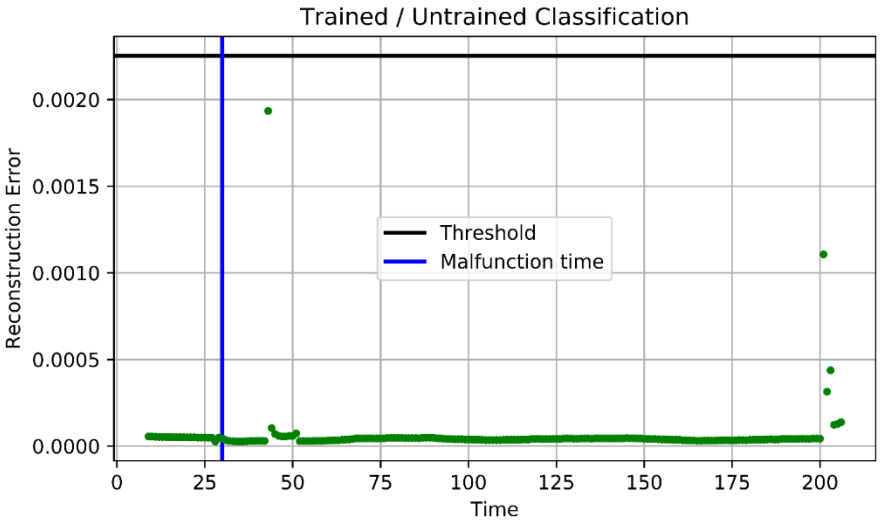
04 Integrated Abnormal Diagnosis Algorithm

- **Diagnostic Module – Training status diagnosis function**
  - Used method: LSTM-AE (*AI-unsupervised learning)
  - Used data: train data (15 abnormal and 1 normal scenarios), test data (5 abnormal scenarios)
    - ✓ Refer to pages 9-11 [Table]
  - Output form: Trained condition or Untrained condition (Abnormal state and operator call)
  - Importance: Assessing the availability of AI model
    - ✓ AI can't cope with untrained data.

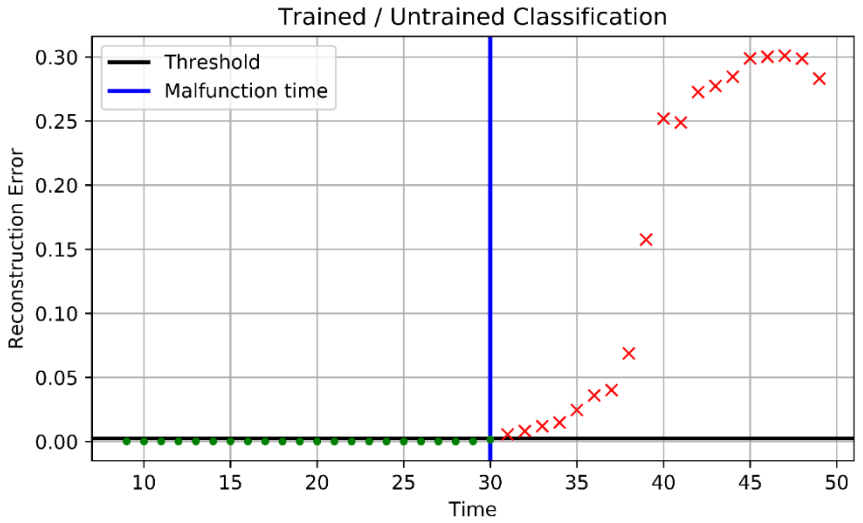


04 Integrated Abnormal Diagnosis Algorithm

- **Diagnostic Module – Training status diagnosis function result**
  - If the reconstruction error is below the threshold, it is judged as a trained condition.
    - ✓ Test: Faulty opening of PRZ spray valve scenario → Trained condition
  - If the reconstruction error is above the threshold, it is judged as an untrained condition.
    - ✓ Test: Isolation of main steam line scenario → Untrained condition



<Training status diagnosis function result: Trained condition>



<Training status diagnosis function result: Untrained condition>

CHOSUN UNIVERSITY

04 Integrated Abnormal Diagnosis Algorithm

- **Diagnostic Module – Abnormal scenario diagnosis function**
  - Used method: LightGBM (*AI-supervised learning)
  - Used data: Target output as trained condition in the training status diagnosis function
    - ✓ 15 abnormal and 1 normal scenarios [Total 16 scenarios]
  - Output form: One of the 16 trained condition scenarios
  - Importance: Assistance in operator's diagnosis tasks
    - ✓ The target function of deriving diagnosis evidence using XAI

Increased human error
due to psychological burden

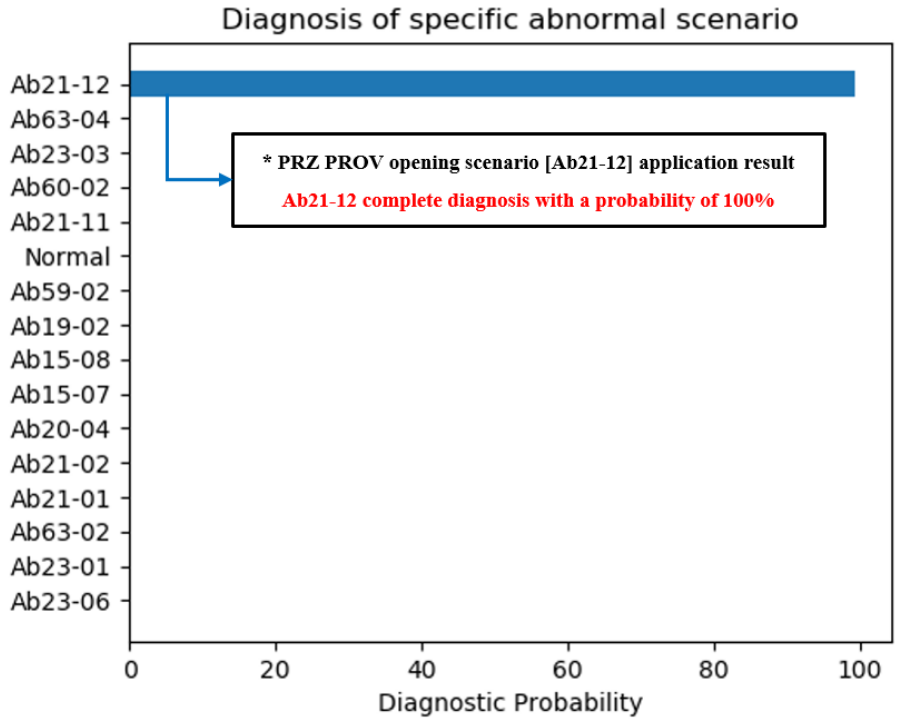
Assistance of the
operator's diagnosis tasks
[Reduction of human error]

Quick mitigation action
due to quick diagnosis

CHOSUN UNIVERSITY

04 Integrated Abnormal Diagnosis Algorithm

- **Diagnostic Module – Abnormal scenario diagnosis function result**
  - Calculate the diagnosis probability for the trained scenarios
    - ✓ Choose 99% ↑ or the highest probability of diagnosis
    - ✓ Test: PRZ PORV opening scenario [Label: Ab21-12] → Diagnosed with 100% probability

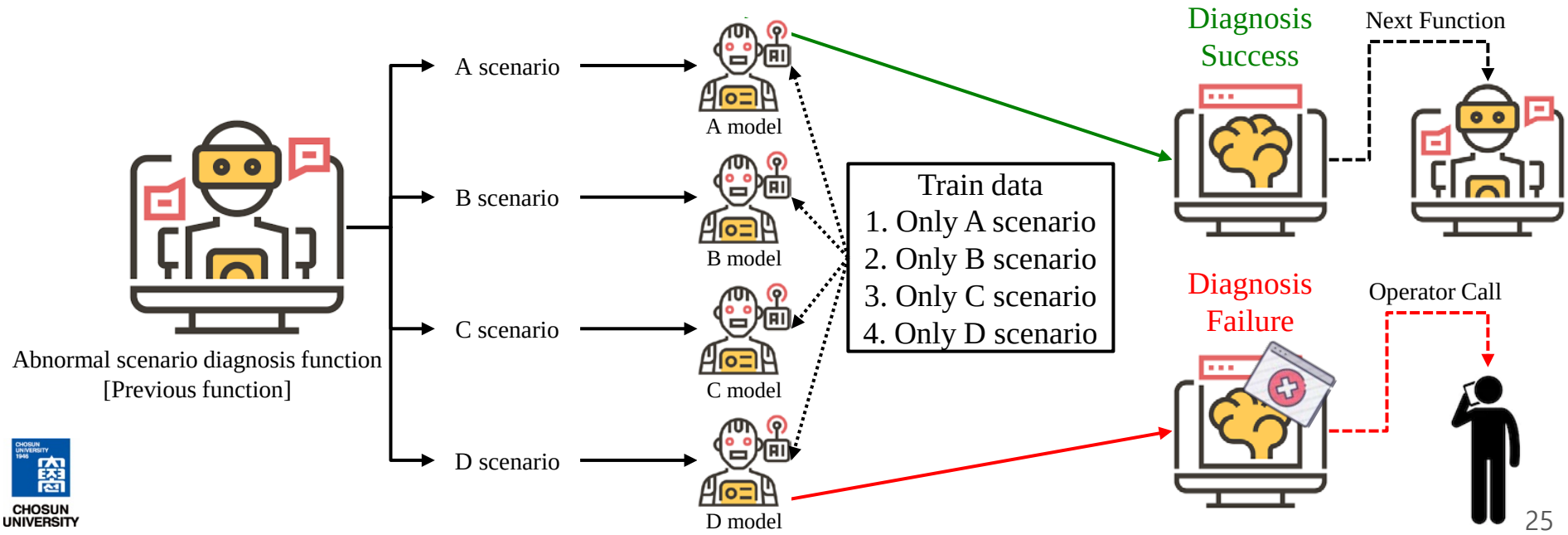


<Abnormal scenario diagnosis function result:
PRZ PORV opening scenario [Ab21-12]>

CHOSUN UNIVERSITY

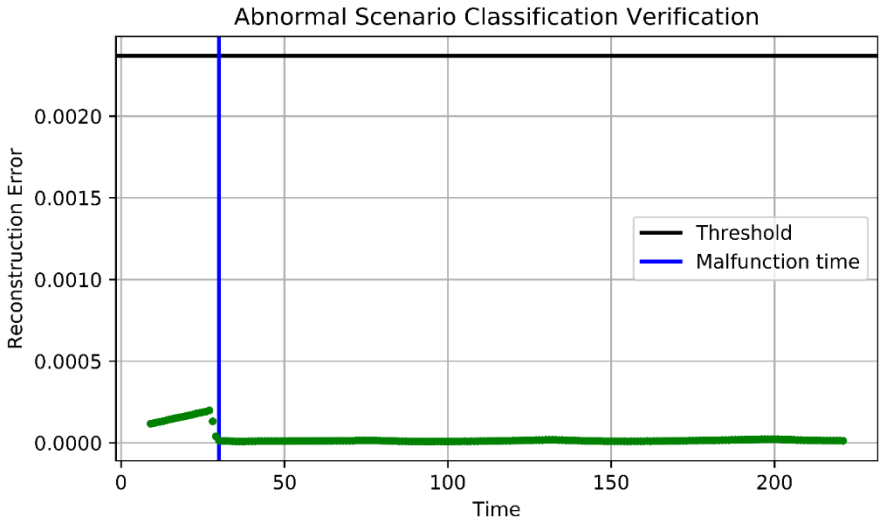
04 Integrated Abnormal Diagnosis Algorithm

- **Verification Module – Diagnosis result verification function**
  - Used method: LSTM-AE (*AI-unsupervised learning)
  - Used data: Target output as trained condition in the training status diagnosis function
    - ✓ Creation of each AI model for each scenario (Total 16 models)
  - Output form: Diagnosis success or failure (operator call)
  - Importance: Re-diagnose and verify the result of AI model
    - ✓ Increased reliability by performing re-validation

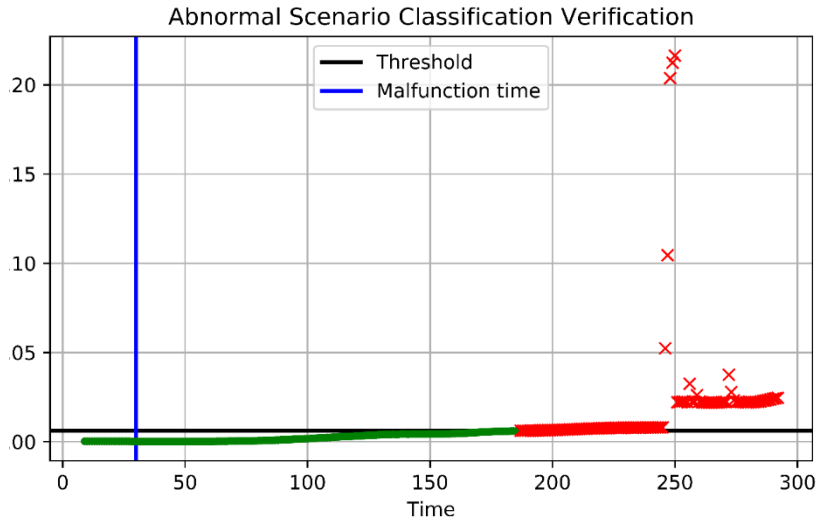


04 Integrated Abnormal Diagnosis Algorithm

- **Verification Module – Diagnosis result verification function result**
  - If the reconstruction error is below the threshold, it is judged as a diagnosis success.
    - ✓ Test: PRZ PORV opening scenario → Diagnosis success (Proper AI model application)
  - If the reconstruction error is above the threshold, it is judged as a diagnosis failure.
    - ✓ Test: PRZ PORV opening scenario → Diagnosis failure (Improper AI model application)
    - ✓ Improper model: PRZ safety valve opening model (similar symptoms)



<Diagnosis result verification function result: Diagnosis success>

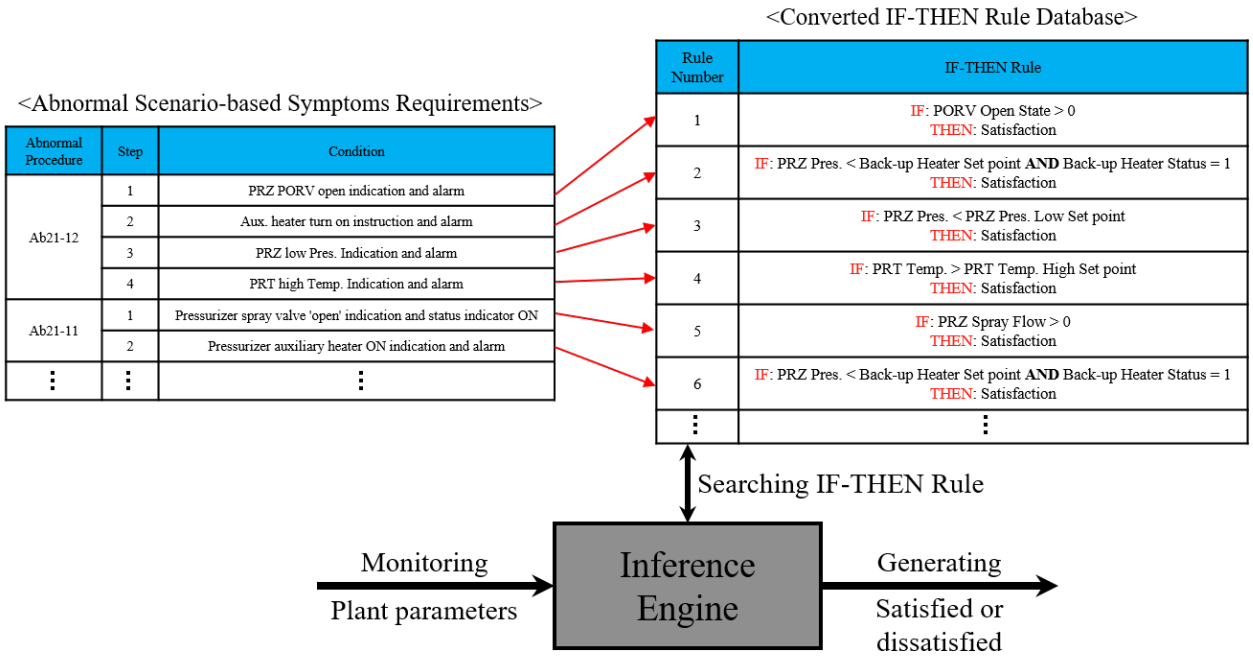


<Diagnosis result verification function result: Diagnosis failure>

CHOSUN UNIVERSITY

04 Integrated Abnormal Diagnosis Algorithm

- **Verification Module – Expected symptoms satisfaction evaluation function**
  - Used method: Rule-based system
  - Used data: Entire dataset entered (2,222 variables)
  - Output form: Symptom satisfaction or dissatisfaction
  - Importance: Evaluate whether the input value matches the expected symptoms
    - ✓ Improved reliability of diagnosed results

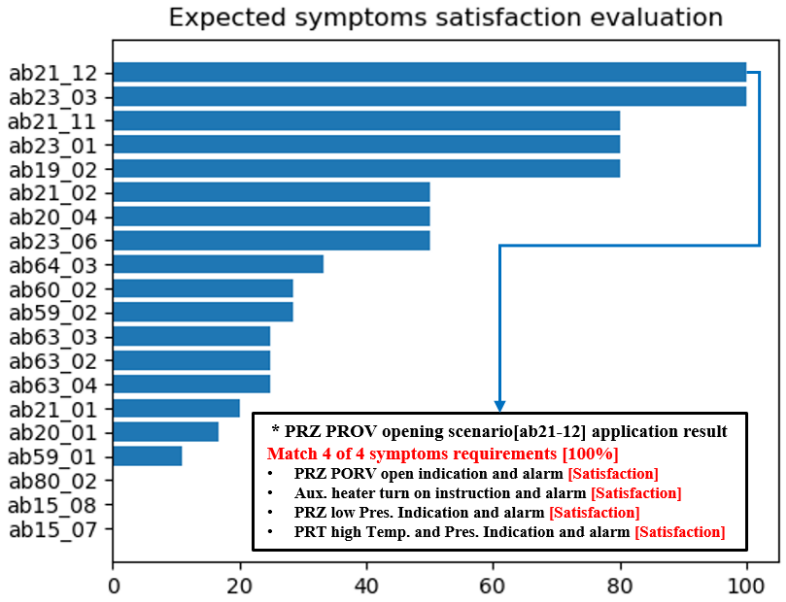


04 Integrated Abnormal Diagnosis Algorithm

- **Verification Module – Expected symptoms satisfaction evaluation function result**
  - For each scenario, symptom evaluation is performed by constructing an IF-THEN rule.
    - ✓ It is calculated as a probability by synthesizing the satisfied symptoms of the entire scenario.
  - Information presented by probability has the following problems. → Interface
    - ✓ The number of symptoms is different for each scenario.
    - ✓ The details of specific symptoms can't be confirmed.
    - ✓ Similar scenario symptoms exist.

- PRZ PORV opening scenario [Ab21-12] → 100%
- Leakage from CVCS to CCW [Ab23-03] → 100%

- Faulty opening of PRZ spray valve [Ab21-11] → 80%
- Leakage from RCS to CCW [Ab23-01] → 80%
- PRZ safety valve opening [Ab19-02] → 80%



CHOSUN UNIVERSITY

04 Integrated Abnormal Diagnosis Algorithm

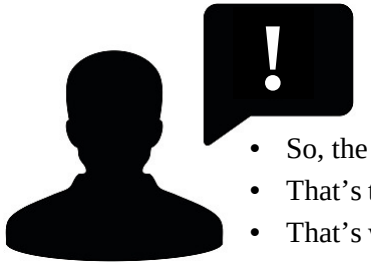
- **Verification Module – Diagnosis evidence deriving function**
  - Used method: SHAP (*XAI)
  - Used data: Same as the data used in LightGBM model
  - Output form: Diagnosed and Undiagnosed evidence
  - Importance: Derivation of evidence by interpreting AI model that provide only simple result
    - ✓ Increased reliability of AI result

Abnormal scenario diagnosis function

A scenario

- I know the result, but why was it diagnosed like that?
- Can I believe that?
- I think it's like that, why isn't that diagnosed?

SHAP Application

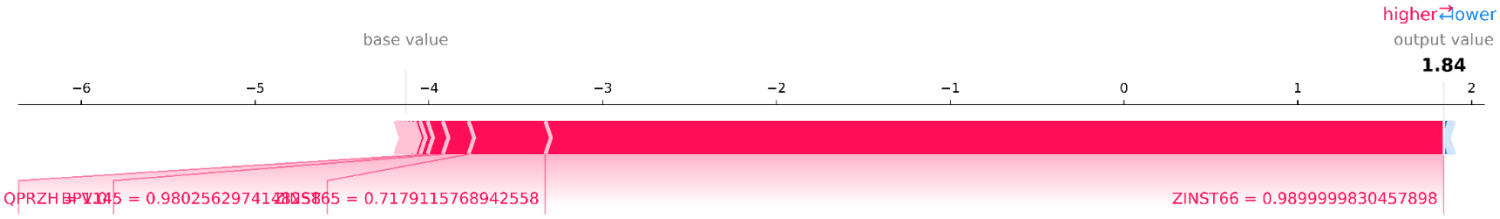


- So, the diagnosed result is that!
- That's the reason. I can believe it!
- That's why it wasn't diagnosed!

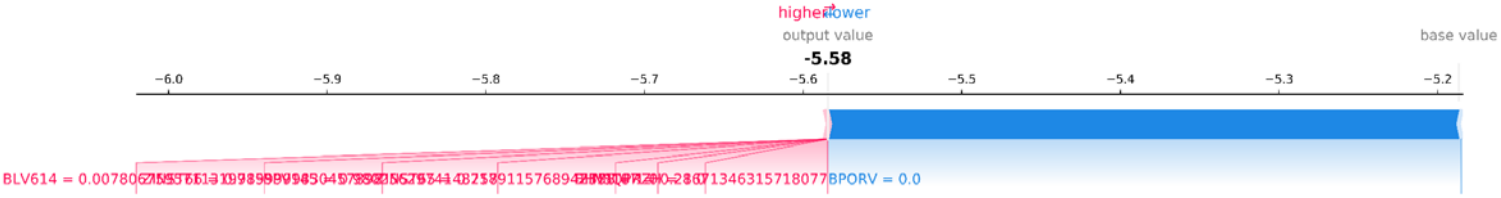
CHOSUN UNIVERSITY

04 Integrated Abnormal Diagnosis Algorithm

- **Verification Module – Diagnosis evidence deriving function result**
  - Interpretation result of diagnosed scenario
    - ✓ Test: Faulty opening of PRZ spray valve scenario [ZINST66: PRZ spray flow]
  - Interpretation result of undiagnosed scenario
    - ✓ Test: PRZ PORV opening [BPORV: PORV opening state]
  - The information displayed in the figure below has the following problems. → Interface
    1. It is difficult to check the variable names and values in the figure.
    2. The meaning of the variable name is unknown.



<Derive the evidence for the diagnosed scenario>

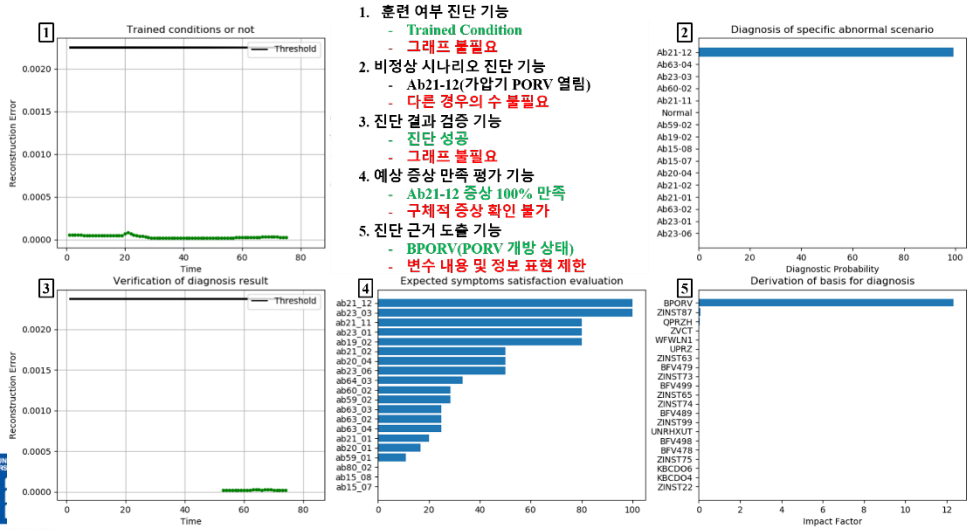


<Derive the evidence for the undiagnosed scenario>

04 Integrated Abnormal Diagnosis Algorithm

▪ Algorithm result problems

- Need to understand graph logic
- Structure problem of adding graphs for additional information
- Difficulty confirming specific details
- Different interpretation criteria for different operators



- Explanation Interface
  - ✓ Intuitive structure
  - ✓ Information conciseness
  - ✓ Visual effect

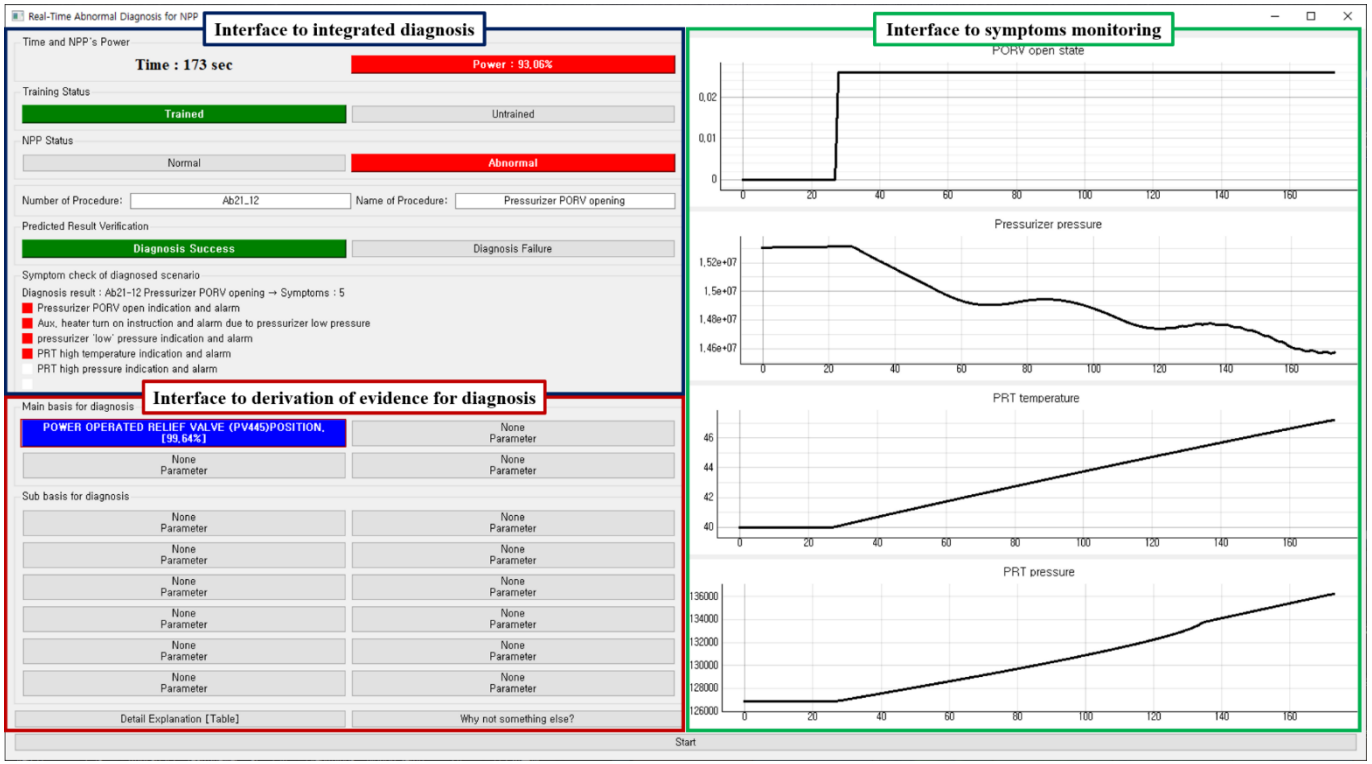
CHOSUN UNIVERSITY

05 Interface

05 Interface

Interface Overview

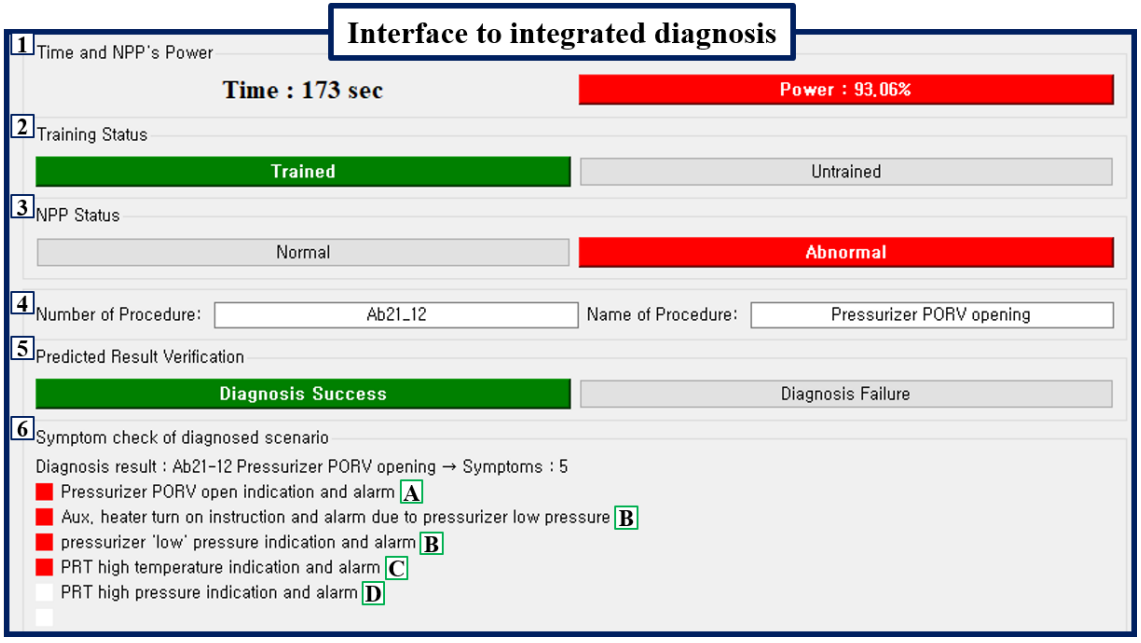
- The interface is designed and implemented to efficiently provide information on the results derived from the algorithm. [Algorithm linkage]
- The interface consists of 1) interface to integrated diagnosis, 2) interface to symptoms monitoring, and 3) interface to derivation of evidence for diagnosis.



05 Interface

Interface to integrated diagnosis

- The integrated diagnosis interface includes all functions except for the function of deriving diagnosis evidence within the verification module.



1) Time and NPP's Power

2) Training Status

3) NPP Status

4) Diagnosed scenario

5) Diagnosed scenario verification

6) Symptom check of diagnosed scenario

CHOSUN UNIVERSITY

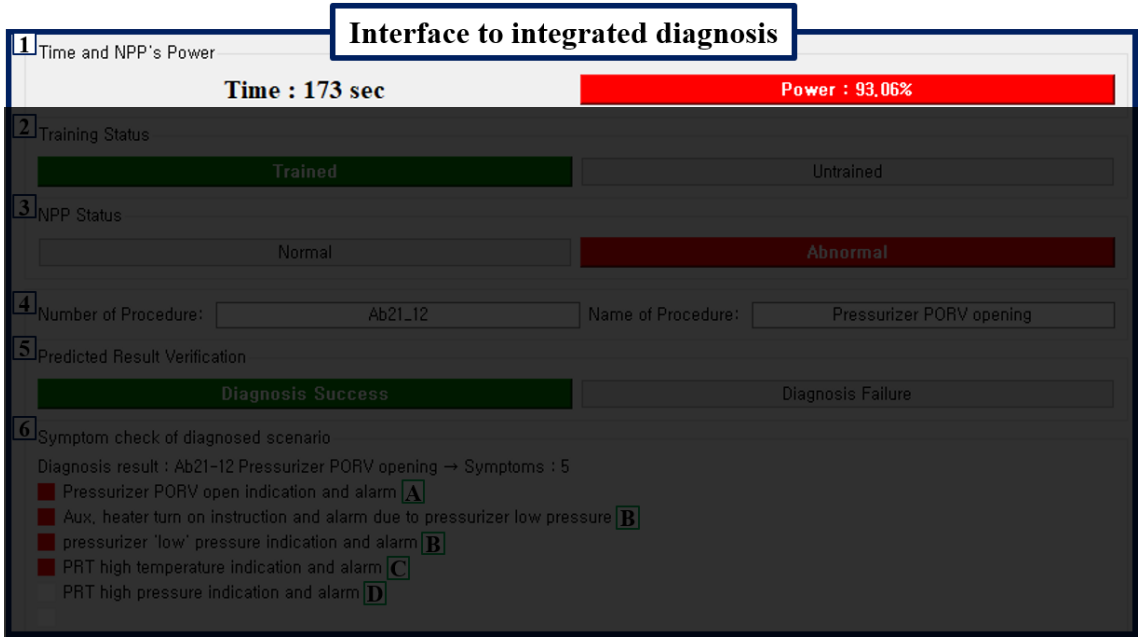
- **Interface to integrated diagnosis**

1. Time and NPP's Power

   ✓ Shows input time and NPP's power value

   - Result: Time 173 seconds, NPP's Power 93.06% [Red]

     - IF NPP's Power < 95% / THEN Red Alarm



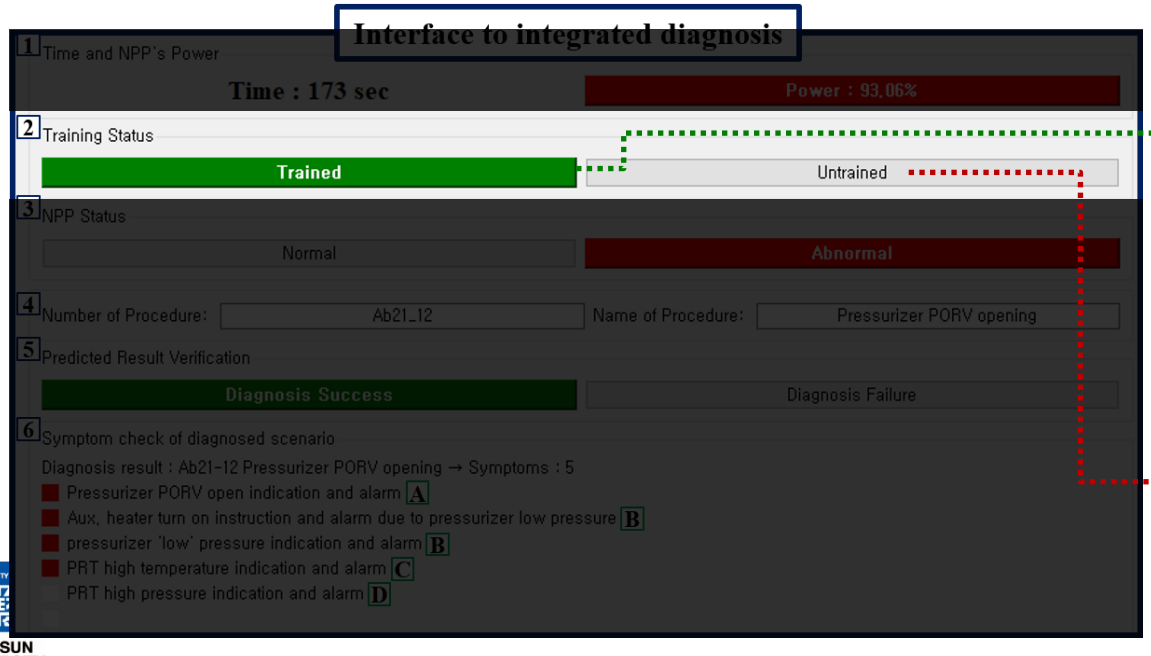
05 Interface

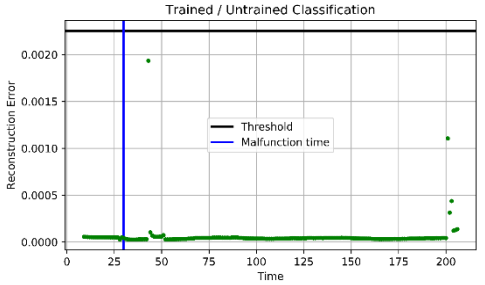
- **Interface to integrated diagnosis**

2. Training Status

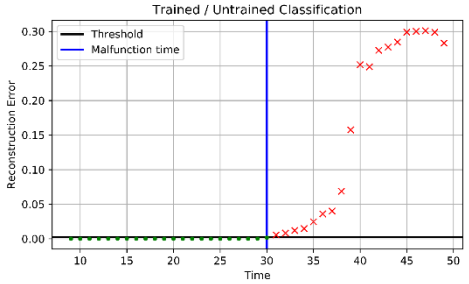
✓ Display the results of the training status diagnosis function

- Result: Trained condition [Green]
  - IF Reconstruction error ≦ Threshold / THEN Trained condition [Green Alarm]
  - IF Reconstruction error > Threshold / THEN Untrained condition [Red Alarm]





<Training status diagnosis function result: Trained condition>



<Training status diagnosis function result: Untrained condition>

CHOSUN UNIVERSITY

05 Interface

- **Interface to integrated diagnosis**

3. NPP Status

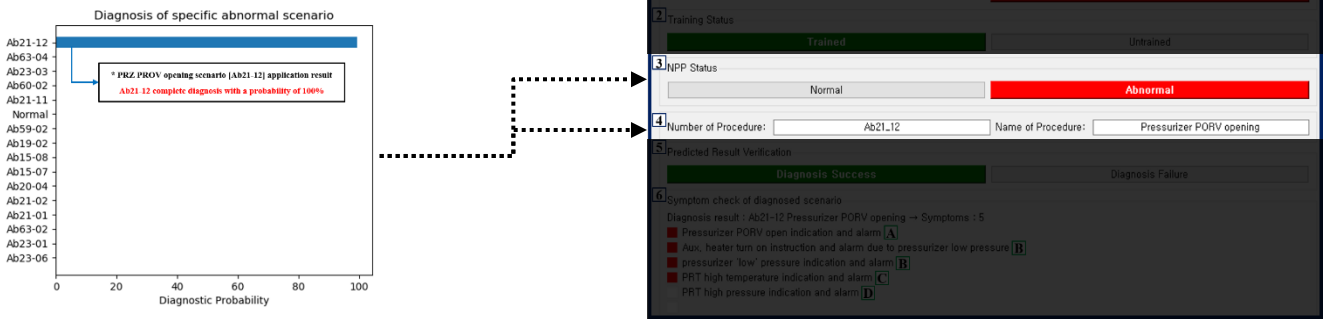
   ✓ Using the results of the abnormal scenario diagnosis function

   - Result: Abnormal state [Red]
     - IF Diagnosis result = normal / THEN Normal state [Green Alarm]
     - IF Diagnosis result ≠ normal / THEN Abnormal state [Red Alarm]

4. Diagnosed scenario

   ✓ Using the results of the abnormal scenario diagnosis function

   - Result: PRZ PORV opening [Ab21-12]
     - Prints the label and name of the abnormal scenario with the highest diagnosis probability



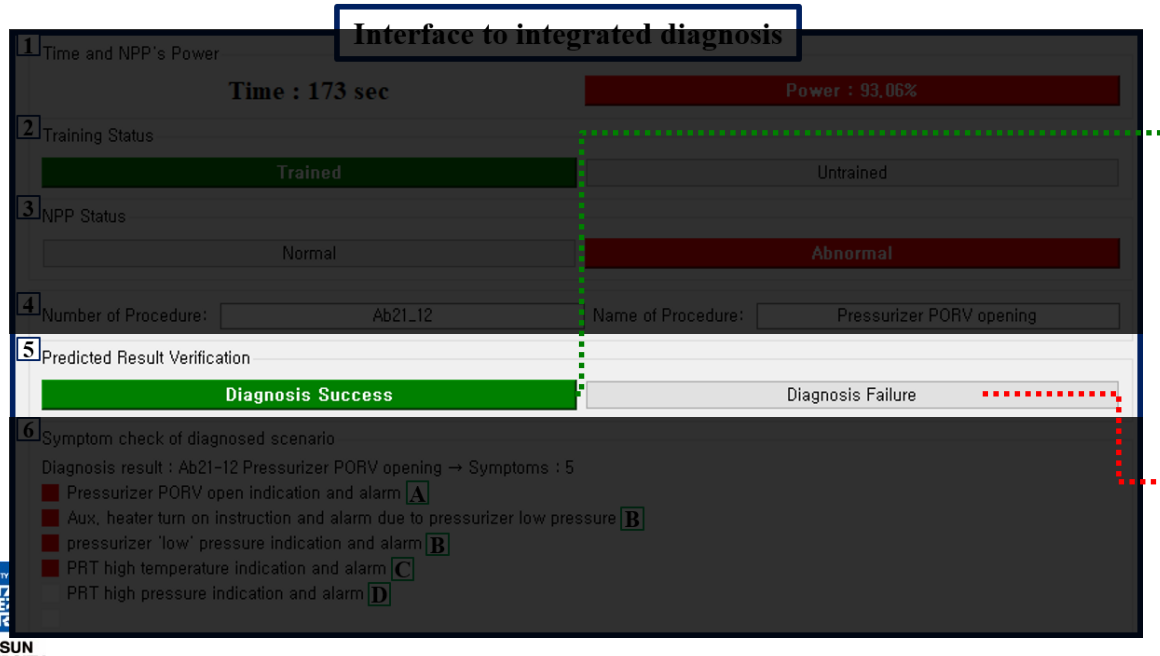
05 Interface

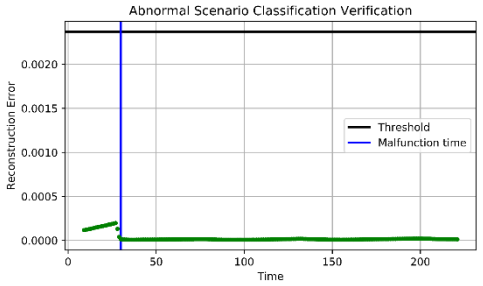
- **Interface to integrated diagnosis**

5. Diagnosed scenario verification

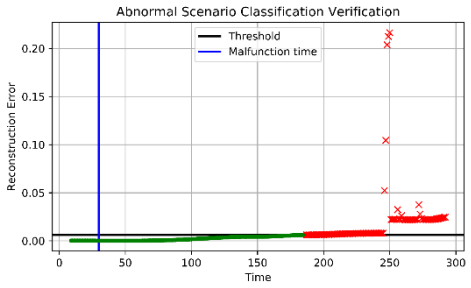
✓ Display the results of the diagnosis result verification function

- Result: Diagnosis success [Green]
  - IF Reconstruction error ≦ Threshold / THEN Diagnosis success [Green Alarm]
  - IF Reconstruction error > Threshold / THEN Diagnosis failure [Red Alarm]





<Diagnosis result verification function result: Diagnosis success>



<Diagnosis result verification function result: Diagnosis failure>

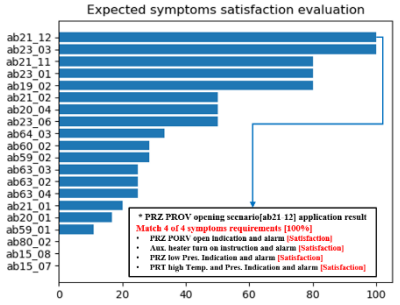
05 Interface

- **Interface to integrated diagnosis**

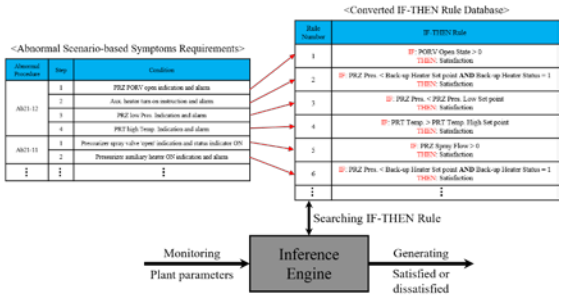
6. Symptom check of diagnosed scenario

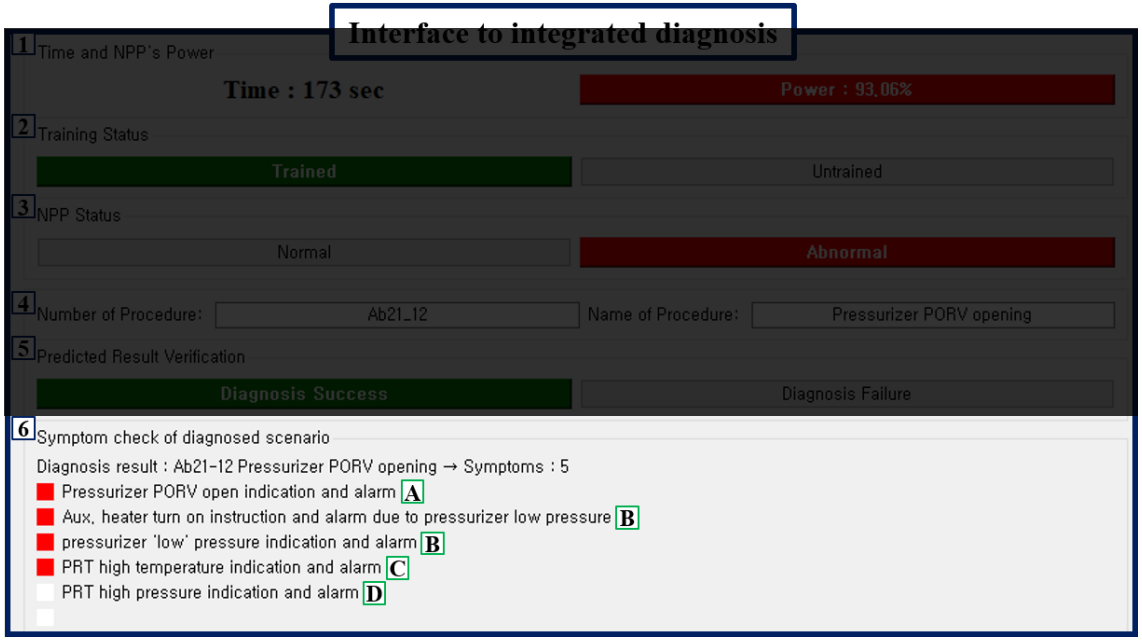
✓ Shows symptom list and satisfaction evaluation result for only diagnosed scenario

- Result: Satisfied with 4 out of 5 symptoms [Satisfaction: Red / Dissatisfaction: White]



<Expected symptoms satisfaction evaluation result>

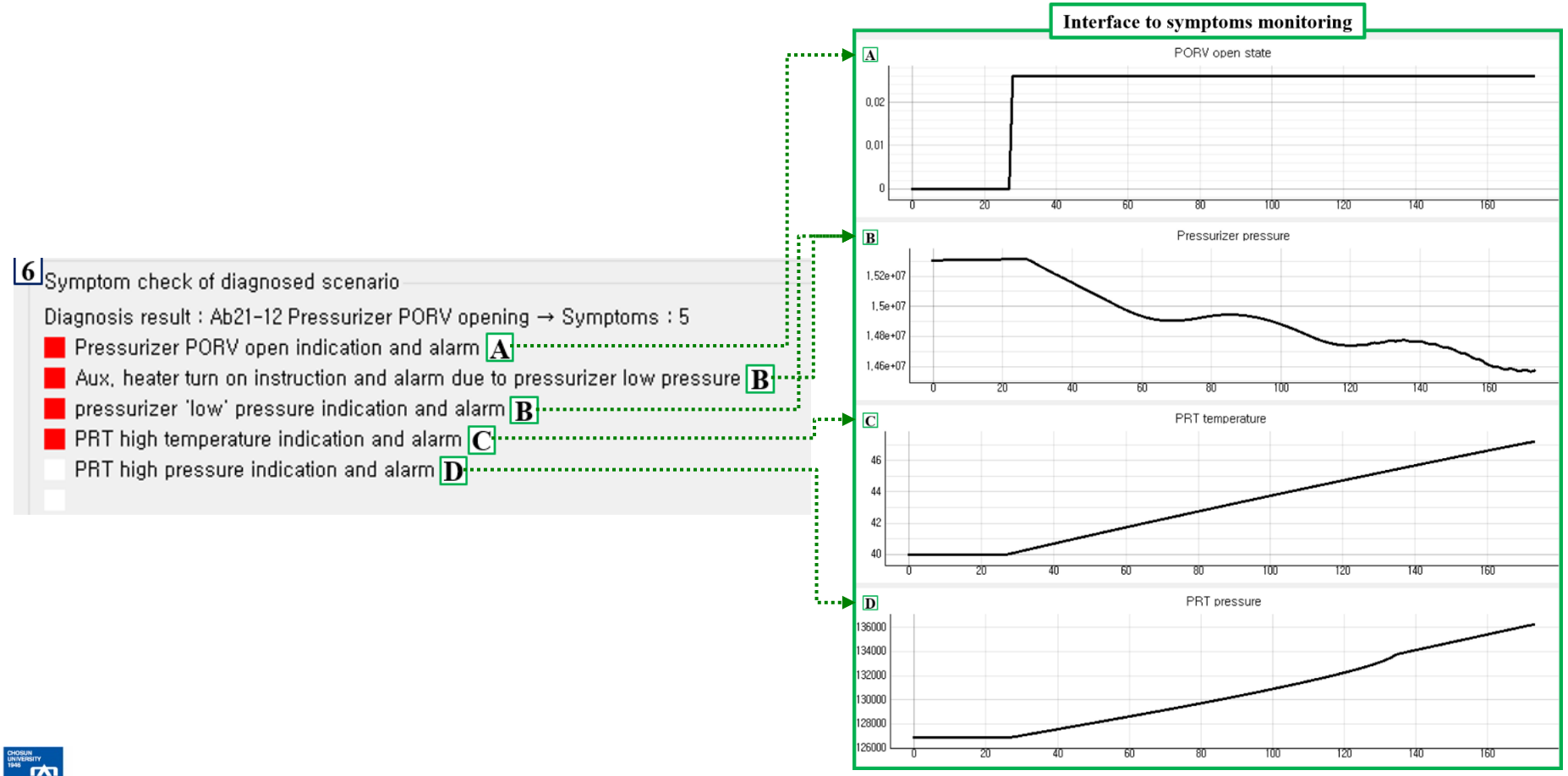




CHOSUN UNIVERSITY

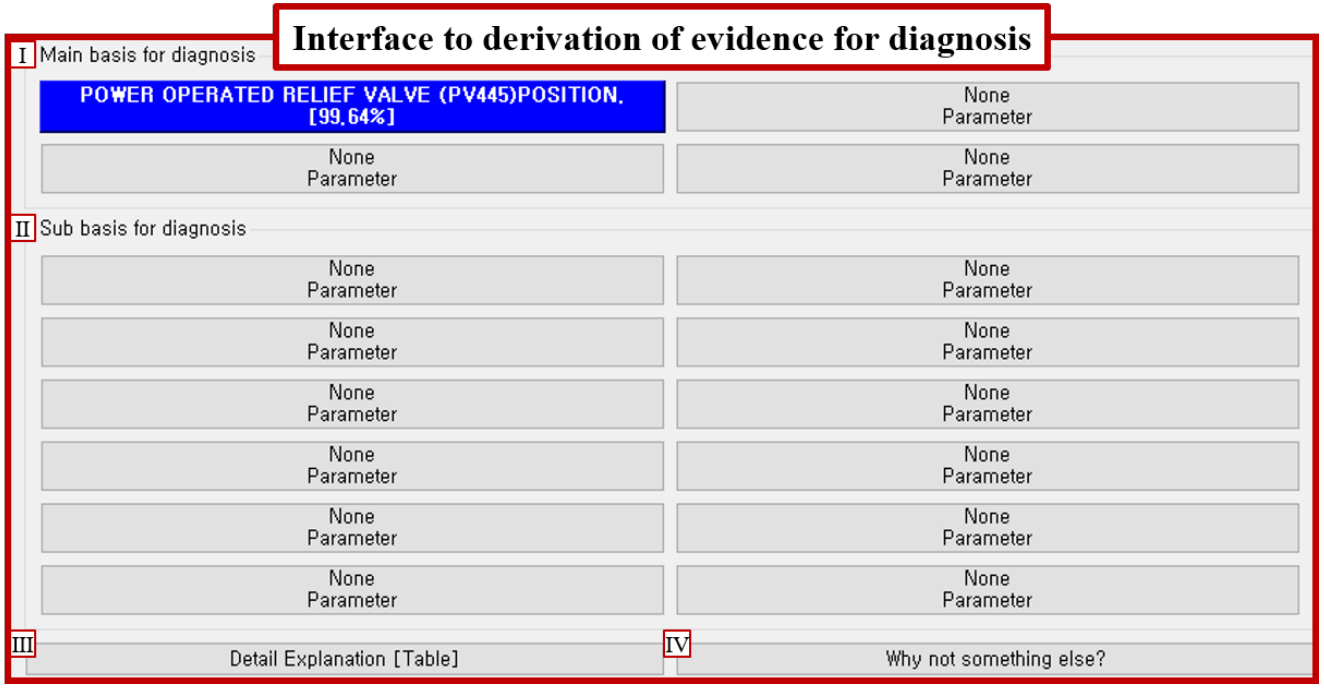
05 Interface

- **Interface to symptoms monitoring**
  - Perform variables monitoring of [6. Symptom check of diagnosed scenario]



05 Interface

- **Interface to derivation of evidence for diagnosis**
  - Interface to present diagnosis evidence deriving function results of verification module
    I. Main basis for diagnosis
    II. Sub basis for diagnosis
    III. Detail Explanation [Table]
    IV. Why not something else?

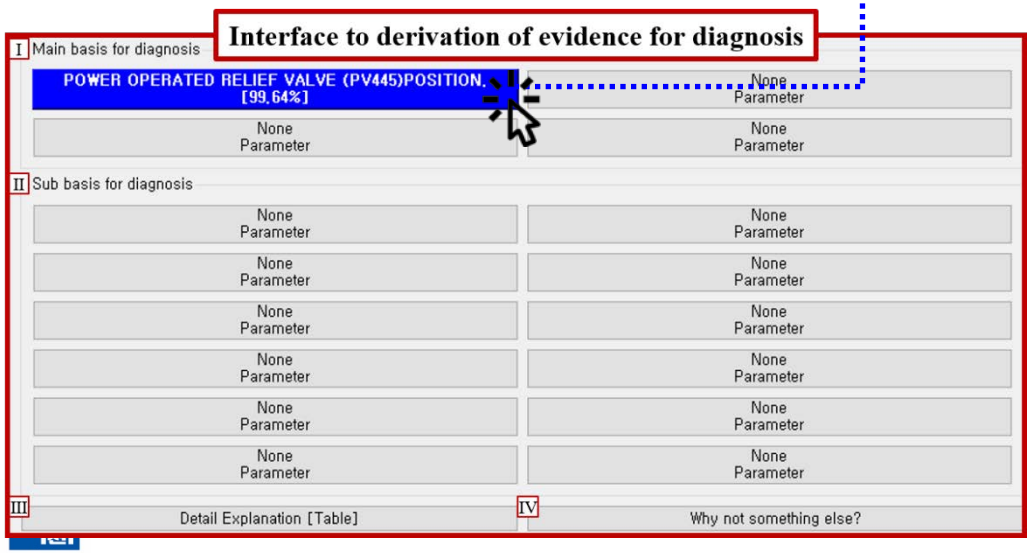


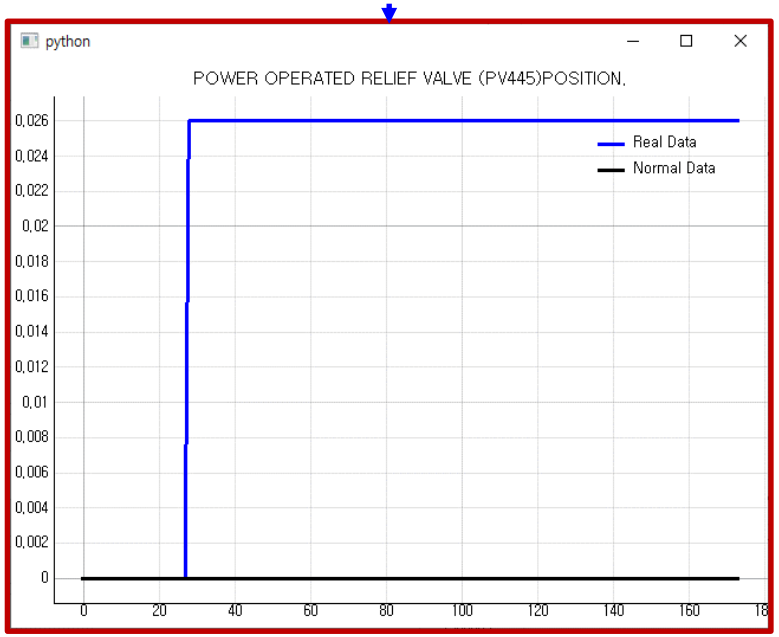
05 Interface

- **Interface to derivation of evidence for diagnosis**

  I. Main basis for diagnosis

  - ✓ Express more than 10% contribution (Present in ascending order)
  - ✓ When the button is clicked, a graph pop-up window is displayed comparing the diagnosed abnormal and normal scenarios.
  - ✓ It can be confirmed that the PORV, which will be closed in the normal scenario, is open in the diagnosed scenario → PRZ PORV opening scenario



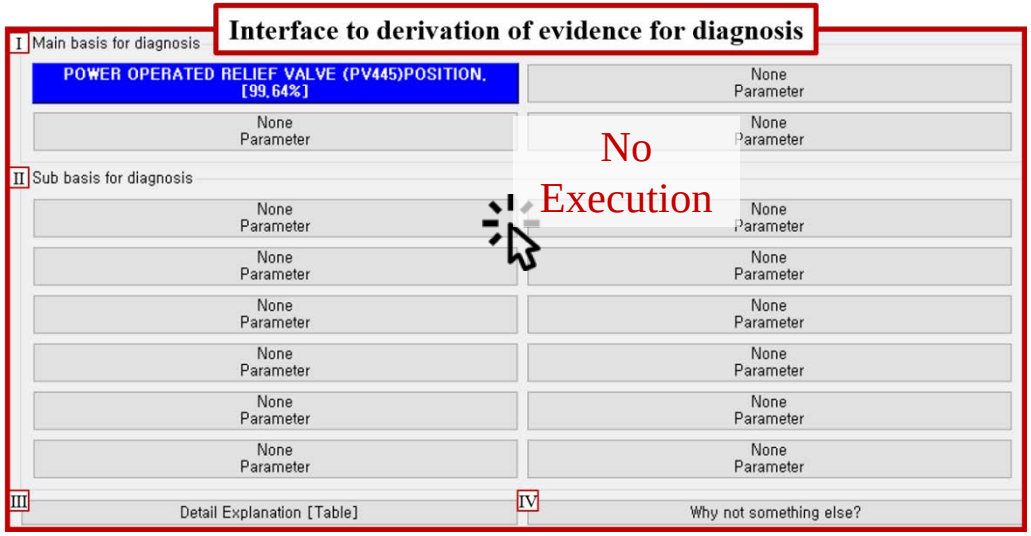


05 Interface

- **Interface to derivation of evidence for diagnosis**

  II. Sub basis for diagnosis

  - ✓ Express 1~10% contribution (Present in ascending order)
  - ✓ Like the main basis for diagnosis, when clicked, a graph pop-up window is displayed.
  - ✓ However, it can't be executed because there is no variable in the set section (1~10%).

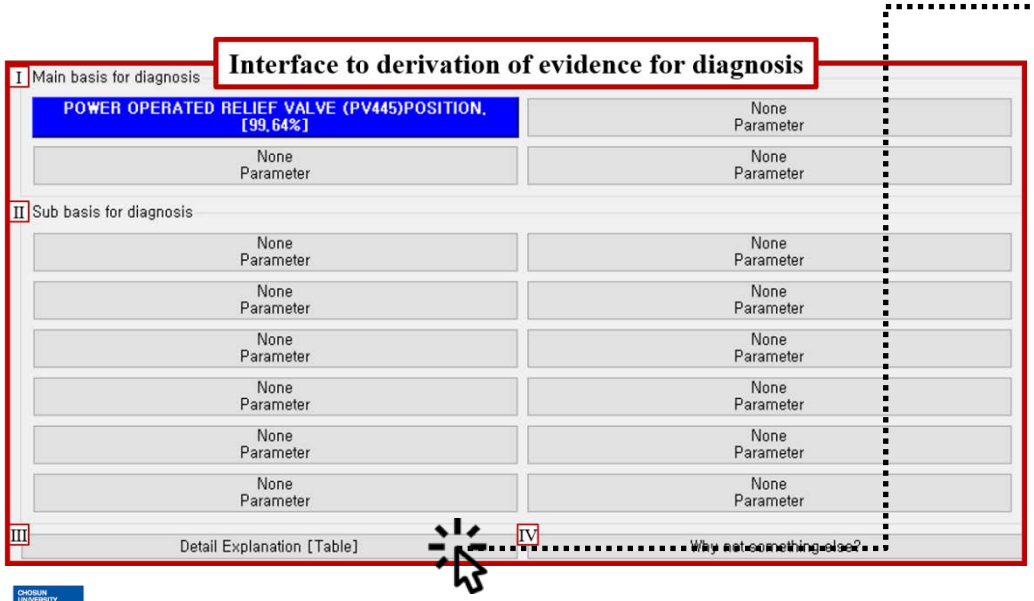


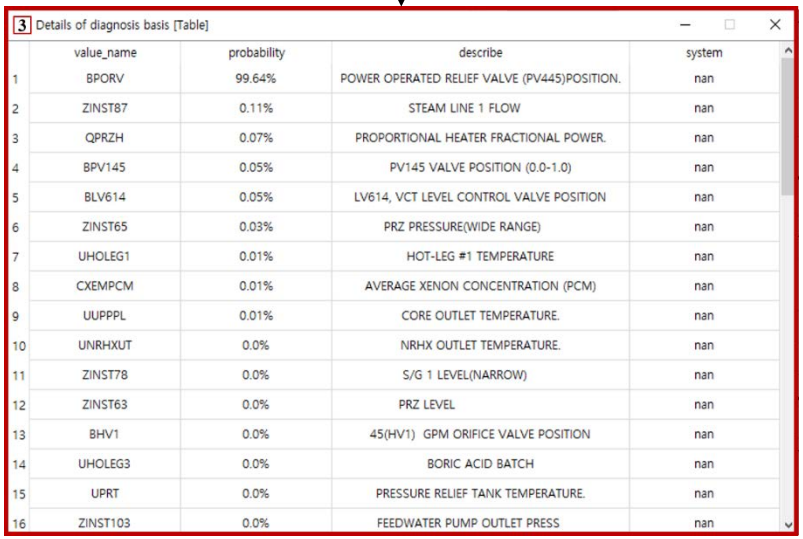
05 Interface

- **Interface to derivation of evidence for diagnosis**

III. Detail Explanation [Table]

- ✓ Total contribution expressed in tabular format
- ✓ When clicked, a table pop-up window consisting of all variable name, probability, and description is displayed.
- ✓ Ascending order based on the value of probability





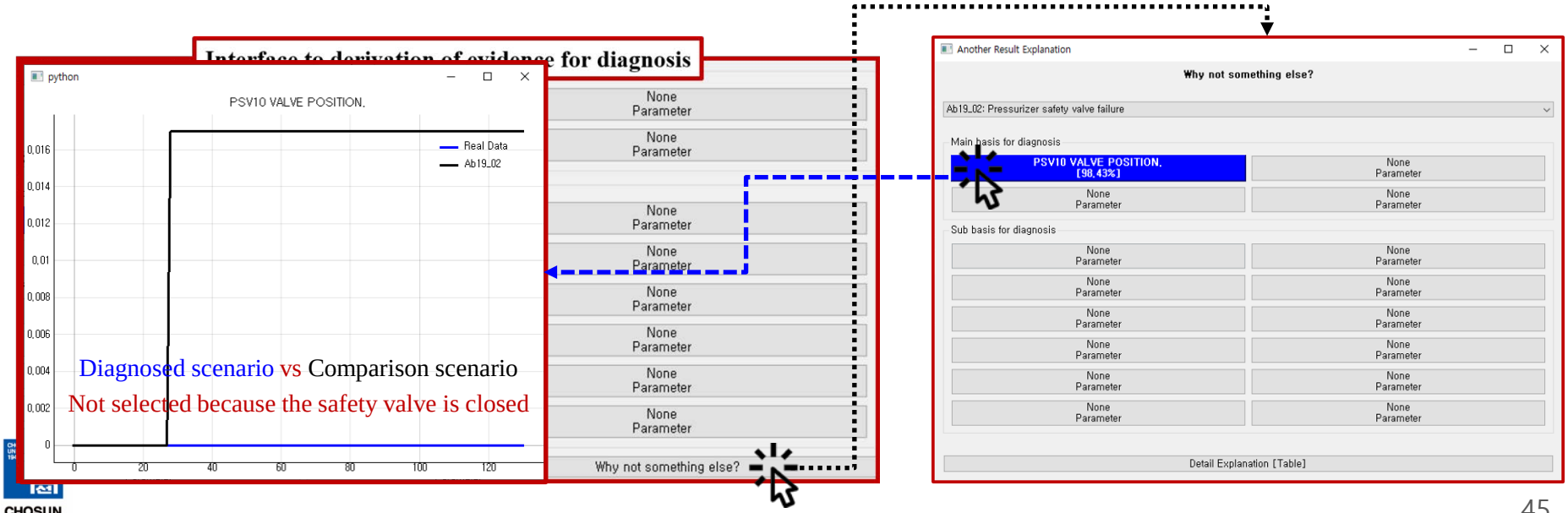
	value_name	probability	describe	system
1	BPORV	99.64%	POWER OPERATED RELIEF VALVE (PV445)POSITION.	nan
2	ZINST87	0.11%	STEAM LINE 1 FLOW	nan
3	QPRZH	0.07%	PROPORTIONAL HEATER FRACTIONAL POWER.	nan
4	BPV145	0.05%	PV145 VALVE POSITION (0.0-1.0)	nan
5	BLV614	0.05%	LV614, VCT LEVEL CONTROL VALVE POSITION	nan
6	ZINST65	0.03%	PRZ PRESSURE(WIDE RANGE)	nan
7	UHOLEG1	0.01%	HOT-LEG #1 TEMPERATURE	nan
8	CXEMPCM	0.01%	AVERAGE XENON CONCENTRATION (PCM)	nan
9	UUPPPL	0.01%	CORE OUTLET TEMPERATURE.	nan
10	UNRHXUT	0.0%	NRHX OUTLET TEMPERATURE.	nan
11	ZINST78	0.0%	S/G 1 LEVEL(NARROW)	nan
12	ZINST63	0.0%	PRZ LEVEL	nan
13	BHV1	0.0%	45(HV1) GPM ORIFICE VALVE POSITION	nan
14	UHOLEG3	0.0%	BORIC ACID BATCH	nan
15	UPRT	0.0%	PRESSURE RELIEF TANK TEMPERATURE.	nan
16	ZINST103	0.0%	FEEDWATER PUMP OUTLET PRESS	nan

CHOSUN UNIVERSITY

- **Interface to derivation of evidence for diagnosis**

IV. Why not something else?

- ✓ Present the results of deriving evidence for undiagnosed scenarios
- ✓ When clicked, a sub pop-up window like the interface to derivation of evidence for diagnosis is displayed.
- ✓ It performs the same function, and as a difference, it is possible to confirm the unselected evidence for the entire diagnosable scenario.



05 Conclusion

05 Conclusion

Conclusion

- Diagnosis research using AI is being conducted to reduce the increase in human error caused by operator's psychological burden.
- However, AI is concerned about a decrease in reliability due to the problem that the evidence for judgment can't be confirmed.
- Therefore, the following research was conducted in this paper.
  - ✓ Derivation of diagnosis evidence using XAI
  - ✓ Development of explanation interface for efficient information delivery

  → Integrated Abnormal Diagnosis System (IADS) development

Improved reliability considering the application of AI to perform diagnostic tasks for NPPs

CHOSUN UNIVERSITY

Thank you for your attention