[UCI]I804:24011-200000359914

2021년 2월

석사학위 논문

설명 가능한 인공지능을 활용한 원전 통합 비정상 진단 시스템 개발

조선대학교 대학원

원 자 력 공 학 과

박 지 훈

설명 가능한 인공지능을 활용한 원전 통합 비정상 진단 시스템 개발

Development of an Integrated Abnormal Diagnosis System in Nuclear Power Plants using eXplainable Artificial Intelligence

2021년 2월 25일

조선대학교 대학원

원 자 력 공 학 과

박 지 훈

설명 가능한 인공지능을 활용한 원전 통합 비정상 진단 시스템 개발

지도교수 나 만 균

이 논문을 공학 석사학위신청 논문으로 제출함

2020년 10월

조선대학교 대학원

원 자 력 공 학 과

박 지 훈

박지훈의 석사학위논문을 인준함

위원장 조선대학교 교수 김 종 현 (인)

위 원 조선대학교 교수 김 진 원 (인)

위 원 조선대학교 교수 나 만 균 (인)

2020년 11월

조선대학교 대학원

목차

표 목차

그림 목차

ABSTRACT

Development of an Integrated Abnormal Diagnosis System in Nuclear Power Plants using eXplainable Artificial Intelligence

Ji Hun Park
Advisor : Prof. Man Gyun Na, Ph.D.
Department of Nuclear Engineering
Graduate School of Chosun University

In the Nuclear Power Plants(NPPs), keeping a normal state during operation is very important for safety and economics. If NPPs enters an abnormal state, the operators must take mitigation measures to bring it back to a normal state. Prior to taking mitigation measures, the operators must conduct an incident diagnosis of the current condition. However, in diagnosing an incident, the operators feels stress and psychological burden, which may lead to human error. If a false diagnosis is made due to human error, the NPPs can go out of control. In this thesis, an integrated abnormal diagnosis system in NPPs was developed to perform the reduction of operators' human errors. For the system development, data collection using Compact Nuclear Simulator, data pre-processing, integrated abnormal diagnosis algorithm, and explanation interface development were performed. The system is implemented by applying Artificial Intelligence(AI), eXplainable AI(XAI), and rule-based system, and provides the results of applied methodologies to operators through an interface. By integrating these series of

processes, an integrated abnormal diagnosis system in NPPs was developed. Providing diagnosis results through AI is expected to reduce human error, and diagnosis basis interpreted through XAI is expected to increase the reliability of artificial intelligence. In addition, it is expected that efficient information delivery will be possible by providing the diagnosis result through an intuitively configured explanation interface.

제 1 장 서 론

원자력 발전소는 다양한 원인(인적 실수, 기계 결함, 전기 결함, 계측 결함, 외부 영향 등)으로 인하여 정상 상태를 이탈하게 된다[1]. 이때, 운전원은 원자로 정지(Reactor Trip) 여부에 따라 비정상 상태(알람 2개 이상 발생부터 원자로 정지 이전) 또는 비상 상태(원자로 정지 이후)를 판단한다. 이후 대응하는 비정상 운전 절차서(Abnormal Operating Procedure) 또는 비상 운전 절차서(Emergency Operating Procedure)를 선택하여 절차에 따라 사건 식별 및 완화 조치를 수행해야 한다.

원자력 발전소가 원인을 알 수 없는 사고로 인해 비정상 상태로 진입했을 경우, 운전원은 발생된 사고의 진행을 빠르게 막기 위해서 빠른 시간 내에 정확하게 사건 식별을 완료해야 한다. 빠른 시간 내에 정확한 사건 식별은 운전원의 심리적 부담을 발생시키기 때문에 인적 오류의 확률이 증가하고, 이로 인해 사건 식별 실패를 야기할 수 있다. 사건 식별 실패는 잘못된 완화조치로 직결되며 발생된 비정상 상태는 조기에 완화를 할 수 있음에도 불구하고 원자로 정지 이후의 비상 상태로 악화될 수 있다. 조기 완화 실패는 원자력 발전소의 안정성과 경제성에 매우 큰 악영향을 미치기에 이를 방지하기 위해서는 비정상 상태 발생 초기에 수행되는 사건 식별은 매우 중요하다. 추가적으로, Jung 등[2]은 원자로 정지 이후 노심 손상에 영향을 미치는 인자 중 인적 오류로 인한 노심 손상 확률이 44%(사건 식별 실패 28%, 완화 조치 실패 16%)에 달한다고 제시하였다. 이를 통해 사건 식별은 비상 상태 진입 이후에도 중요하다는 점을 확인할 수 있다.

인적 오류로 인한 사건 식별 실패를 줄이기 위해 인공지능(Artificial Intelligence)을 활용한 사건 식별 연구가 수행되고 있다. 관련 연구로는 Kim 등[3]은 인공지능 방법론 중 Gated Recurrent Units를 2단계로 구성하여 비정상 상태 사건 식별 결과를 심층적으로 운전원에게 제시하는 알고리즘을 제안하였다. Yang 등[4]은 Long Short-Term Memory(LSTM)와 LSTM-Autoencoder (LSMT-AE)을 활용하여 비상 상태 사건 식별 알고리즘을 제안하였으며, Kim 등[5]은 이를 개선한 LSTM과 LSTM-Variational AE를 활용하여 비정상 상태 사건 식별 알고리즘을 제시하였다. Yoo 등[6]은 Support Vector Machine, Cascaded Fuzzy Neural Network를 활용하여 중대사고 상황에서 사건 식별, 냉각재 상실사고 시 파단 크기 예측, 격납건물 내

수소농도 예측, 원자로용기 노심 수위 예측, 골든타임 예측을 수행하여 운전원에게 정보를 제공하기 위한 스마트 지원 시스템을 구현하였다.

하지만, 인공지능을 활용한 원자력 발전소 사건 식별 알고리즘 또는 운전 지원 시스템을 실제 적용하기에는 문제점이 존재한다. 인공지능의 블랙박스 특성으로 인해 정보를 제공받는 운전원은 물론 개발자 또한 어떠한 근거로 인해 특정 결과가 도출되었는지 모른다는 점이다. 이는 인공지능에 대한 신뢰성과 운전원의 책임으로 직결되는 문제로, 운전원은 인공지능으로부터 도출된 진단 결과에 대해서 어떠한 설명도 제공받지 못하고 왜 특정 진단결과가 출력되었는지 알 수 없으며, 출력된 특정 진단이 100% 정확한지에 대한 의문 또한 가지게 된다. 만약 운전원이 사건 식별에 실패한 인공지능의 진단 결과를 바탕으로 완화 조치를 수행할 경우, 원자력 발전소는 제어가 불가능한 상태로 진입할 수 있기 때문에 운전원은 책임 문제에서 자유로워질 수 없다. 이에 도출된 결과만을 전적으로 신뢰할 수 없기 때문에 인공지능이 제시하는 의견은 운전원이 참고할 수 있는 다양한 참고 정보 중 하나로 최종 의사결정은 운전원이 해야 함을 상기시켜 준다[7].

이에 본 논문에서는 최종 결정은 운전원이 수행하고 의사결정을 지원할 수 있는 상세 정보만을 운전원에게 제공하는 운전 지원 시스템 개발을 목표로 하였으며, 대상은 원자력 발전소의 비정상 상태로 국한하였다. 제공되는 정보는 인공지능을 활용한 사건 식별 결과뿐만 아니라 설명 가능한 인공지능(eXplainable AI)을 활용하여 도출된 진단 결과를 해석한 결과로 구성된다. 운전원은 제공되는 정보를 기반으로 인공지능의 판단을 이해할 수 있으며, 기존의 블랙박스의 문제점을 해소(인공지능 결과에 대한 신뢰성 제고)할 수 있을 것으로 기대된다.

논문의 구조는 그림 1과 같이 1) 시뮬레이터를 활용한 데이터 수집, 2) 데이터 전처리, 3) 방법론(인공지능, 설명 가능한 인공지능, 규칙 기반 시스템), 4) 통합 비정상 진단 알고리즘, 5) 설명 인터페이스의 순서로 구성된다.

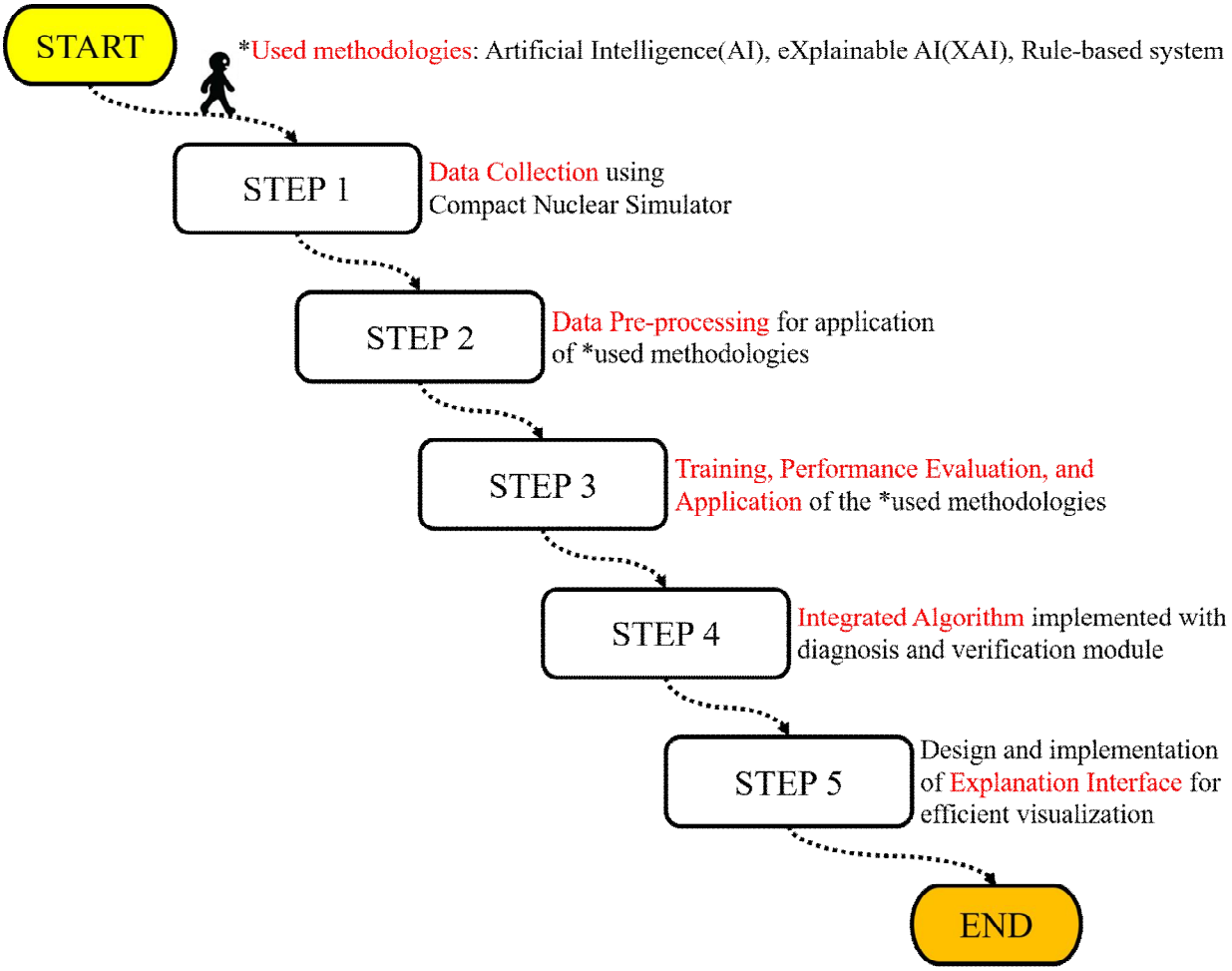


그림 1. 원전 통합 비정상 진단 시스템 개발 로드맵

제 2 장 데이터 수집 및 전처리

해당 장에서는 그림 1에서 제시한 로드맵의 1단계 데이터 수집과 2단계 데이터 전처리에 대해서 다루고자 한다.

제 1 절 데이터 수집

원전 통합 비정상 진단 시스템은 데이터 기반의 다양한 방법론을 기반으로 구성되어 있다. 이때, 방법론에 활용될 데이터 수집을 위해 실제 원자력 발전소의 운전 데이터를 획득해야하지만, 정보 보안 등의 이유로 제한된다. 이를 대체하기 위해서 웨스팅하우스 930 MWe 3 Loop 가압경수로를 기반으로 설계된 Compact Nuclear Simulator(CNS)를 활용하여 데이터 수집을 진행하였다. CNS는 간단한 출력 운전부터 정상 운전, 비정상 및 비상 운전의 데이터를 모사할 수 있으며, 사용자가 원하는 특정 시점에 고장(기기 및 계측기 이상, 누설 등)을 주입할 수 있다.

원전 통합 비정상 진단 시스템은 원자력 발전소의 비정상 상태의 진단을 목표로 설계하였다. 이때, 비정상 상태의 데이터는 다양한 시나리오로 구성되어있기 때문에 기준을 설정하여 수집하여야 한다. 비정상 시나리오 선정 기준을 확립하기 위해서 원전 안전운영정보시스템에 공개된 정보 중 1978년부터 2018년까지 발생한 국내 원전의 사고 및 고장 현황을 발생원인, 계통, 사건 내용별로 분석하였다[1]. 분석 결과 총 737건 중 계측 결함으로 인한 사고는 215건(29%), 기기 이상으로 인한 사고는 196건(27%), 인적 실수로 인한 사고는 134건(18%) 순으로 사고 및 고장의 원인이 되는 것으로 파악되었다. 분석 결과를 기반으로 비정상 시나리오를 1) 자동로직 및 계측기 오류, 2) 기기 상태 이상, 3) 배관 누설의 범주로 분류하였다. 총 20개의 비정상 시나리오와 정상 상태 데이터를 수집하였으며, 비정상 시나리오의 사고 주입은 30초로 설정하였다. 그림 2와 표1은 각각 CNS 인터페이스와 수집된 비정상 시나리오 목록을 나타낸다.

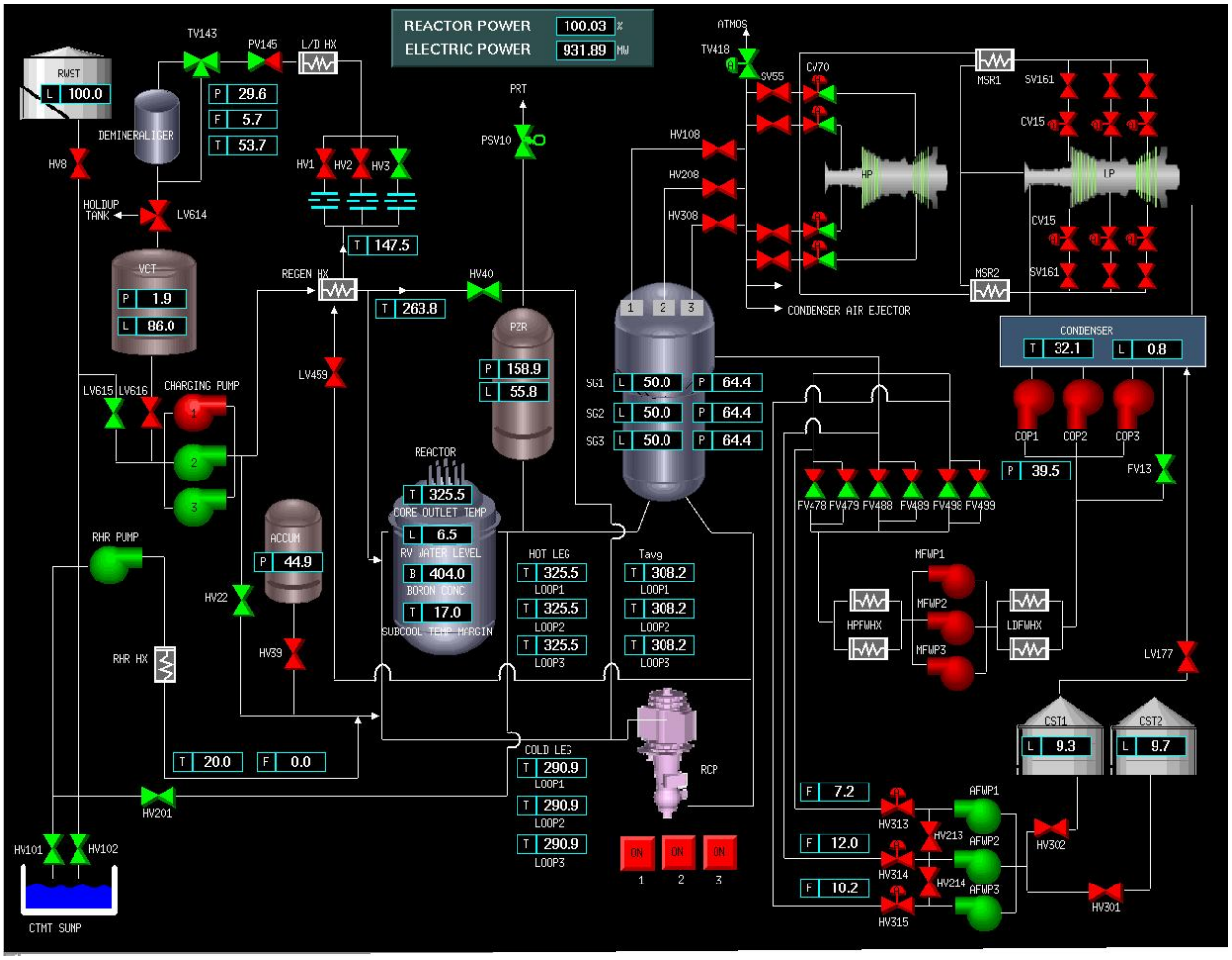


그림 2. CNS 인터페이스

표 1. 수집된 비정상 시나리오 목록

NO	비정상 시나리오 번호	비정상 시나리오 이름	데이터 개수 (훈련, 테스트)	전체 데이터 행렬 (행, 열)
0		정상 상태	20(15, 5)	(8604, 2222)
<자동로직 및 계측기 오류>				
1	Ab21-01	가압기 압력 채널 고장 '고'	18(13, 5)	(4698, 2222)
2	Ab21-02	가압기 압력 채널 고장 '저'	26(21, 5)	(5212, 2222)

3	Ab20-01	가압기 수위 채널 고장 '고'	6(0, 6)	(3769, 2222)
4	Ab20-04	가압기 수위 채널 고장 '저'	15(12, 3)	(7954, 2222)
5	Ab15-07	증기발생기 수위 채널 고장 '저'	40(35, 5)	(8912, 2222)
6	Ab15-08	증기발생기 수위 채널 고장 '고'	40(35, 5)	(11384, 2222)
<기기 상태 이상>				
7	Ab63-04	제어봉 낙하	48(40, 8)	(46507, 2222)
8	Ab63-02	제어봉의 계속적인 삽입	8(6, 2)	(4363, 2222)
9	Ab63-03	제어봉의 계속적인 인출	8(0, 8)	(2689, 2222)
10	Ab21-12	가압기 *PORV 열림	52(45, 7)	(13573, 2222)
11	Ab19-02	가압기 안전밸브 고장	51(45, 6)	(17370, 2222)
12	Ab21-11	가압기 살수밸브 고장 '열림'	50(45, 5)	(31391, 2222)
13	Ab59-01	충전펌프 고장 정지	1(0, 1)	(678, 2222)
14	Ab80-02	주급수펌프 터빈 2/3 대 정지	3(0, 3)	(3400, 2222)
15	Ab64-03	주증기관 차단	3(0, 3)	(142, 2222)
<배관 누설>				
16	Ab60-02	재생열교환기 전단부위 파열	50(45, 5)	(32857, 2222)
17	Ab23-03	*CVCS → *CCW 누설	50(45, 5)	(40498, 2222)
18	Ab59-02	충전수 유량조절밸브 후단누설	30(25, 5)	(20313, 2222)
19	Ab23-01	*RCS → *CCW 누설	30(25, 5)	(2900, 2222)
20	Ab23-06	증기발생기 전열관 누설	36(30, 6)	(2738, 2222)

*PORV: Power Operated Relief Valve
*CVCS: Chemical Volume Control System
*CCW: Component Cooling Water
*RCS: Reactor Coolant System

제 2 절 데이터 전처리

원전 통합 비정상 진단 시스템에 적용된 방법론은 특성에 따라 요구하는 데이터의 형태가 다르다. 수집된 데이터를 방법론에 적용하기 위해서는 데이터 전처리 과정이 필요하다. 그림 3은 데이터 형태 변형을 위한 데이터 전처리 과정을 나타낸다. 특징 추출(Feature selection), 정규화(Normalization), 시계열 데이터 처리(Time series data stack)를 활용하여 Non-Normalized data, Time Series Normalized data, Normalized data와 같은 3가지 형태로 전처리된다.

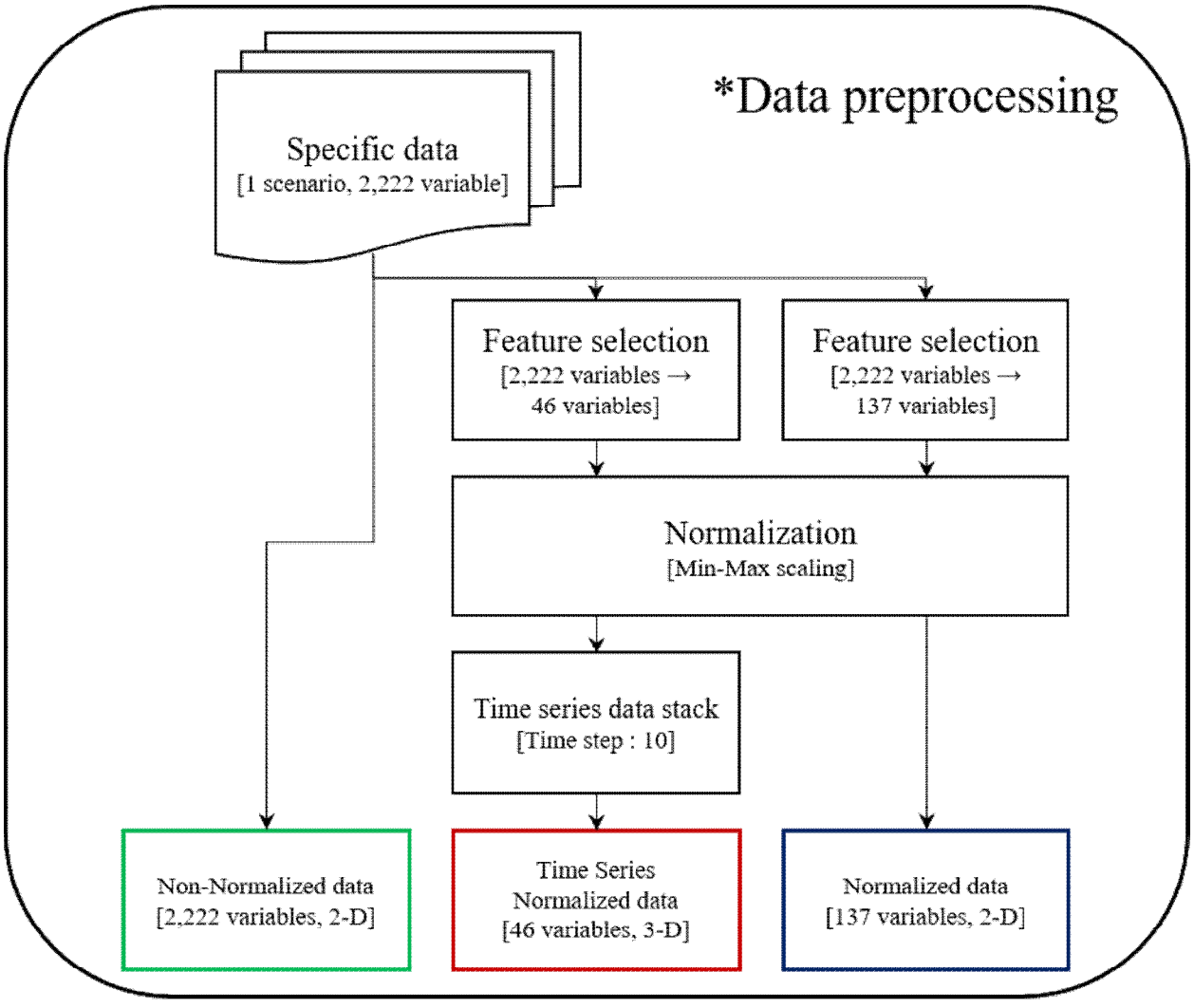


그림 3. 데이터 전처리 과정

특징 추출은 수집된 데이터로부터 특정 변수를 선택한다. 변수에 따라 성능이 바뀌는 방법론의 특성을 반영하기 위해 수행된다. 수집된 데이터의 2,222개 변수 중 방법론에서 활용되는 2,222개, 46개, 137개로 선택된다. 변수 선정 기준은 각 방법론마다 상이하므로 3장에서 설명한다.

정규화는 데이터를 일정한 규칙에 따라 변형하는 작업으로 데이터의 각 변수를 0과 1사이에 위치시키는 것을 의미한다. 인공지능 학습의 데이터 편향을 방지하기 위해 수행해야 하는 작업이다. 데이터 편향이란 실제 고려해야할 데이터의 패턴이 아닌 데이터의 수치에 의존하는 현상을 뜻한다. 정규화에는 여러 방법들이 존재하지만, 데이터의 최솟값과 최댓값을 기준으로 정규화를 수행하는 Min-Max Normalization을 활용하였다. Min-Max Normalization은 식 (1)과 같이 표현된다.

$$Xscale = (X - Xmin)/(Xmax - Xmin) \quad (1)$$

시계열 데이터 처리는 데이터의 시계열성을 가지고 학습하는 인공지능 방법론을 위해 수행된다[8]. 예를 들어, 시간 간격을 10초로 설정했을 경우 현재 시점으로부터 10초 전의 데이터를 쌓아 3차원의 시계열 데이터로 시간 흐름에 따른 변수의 거동을 학습에 활용한다.

제 3 장 인공지능 방법론

해당 장에서는 그림 1의 로드맵 중 3단계인 알고리즘 구현을 위해 활용되는 방법론에 대해 설명한다. 방법론은 인공지능 방법론인 LSTM-AE, Light Gradient Boosted Machine(LightGBM)과 설명 가능한 인공지능 방법론인 SHapley Additive exPlanations(SHAP), 그리고 규칙 기반 시스템이 적용되었다.

인공지능은 흔히 기계학습(Machine Learning)이라고도 불리며, 이러한 기계학습은 데이터를 기반으로 기계가 학습한다하여 붙여진 이름이다. 이러한 학습 방법은 크게 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 분류된다[9]. 지도학습은 쉽게 말해 정답을 알려주면서 학습을 진행하는 것이다. 이러한 정답은 라벨링(Labeling) 작업을 통해 입력 데이터에 대응하여 정답을 붙이며, 이를 통해 훈련된 인공지능이 정답을 잘 맞혔는지 아닌지 쉽게 확인할 수 있다. 지도학습은 크게 분류(Classification)와 회귀(Regression)로 나뉘며, 본 논문에서는 인공지능을 활용한 진단을 수행하기에 분류에 대해서만 다룬다. 분류는 다시 이진 분류와 다중 분류로 구분된다. 이진 분류는 어떤 데이터에 대해 두 가지 중 하나로 분류하는 것이며, 다중 분류는 어떤 데이터에 대해 여러 값 중 하나로 분류하는 것을 의미한다. 비지도학습은 정답을 따로 알려주지 않고, 비슷한 데이터들을 군집하여 학습하는 것이다. 이러한 비지도학습은 별도로 라벨링 작업을 할 필요가 없기 때문에 데이터 처리에 있어서 시간은 절약되지만 정확도가 지도학습에 비해 낮다는 단점을 가지고 있다.

인공지능 방법론 중 LSTM-AE는 1절에서 다루며, LightGBM은 2절에서 다룬다. 설명 가능한 인공지능은 3절에서, 규칙 기반 시스템은 4절에서 다룬다.

제 1 절 LSTM-AE

LSTM-AE는 원자력 발전소 데이터의 시계열 특성을 효율적으로 활용하기 위한 LSTM 방법론과 입력된 데이터를 재구성(복사)하여 출력하는 AE 방법론을 합성한 인공지능 방법론이다. LSTM-AE는 주로 오류 탐지[10], 데이터 생성 등에서 활발히 연구되고 있으며, 원전 통합 비정상 진단 시스템의 경우 오류 탐지에 대해서 다룬다. 오류 탐지 분야에서 출력 값은 이진 분류 형태인 정상 상태 또는 오류 상태로 출력된다.

그림 4는 LSTM-AE의 구조를 보여주며, 시계열 데이터 활용을 위해 입력층(Input Sequence)과 출력층(Reconstructed Sequence)에 LSTM 방법론 내의 시퀀스 구조를 배치하였으며, 입력된 시계열 데이터를 재구성하여 출력하고자 중간에 AE 구조가 삽입되어 있음을 확인할 수 있다.

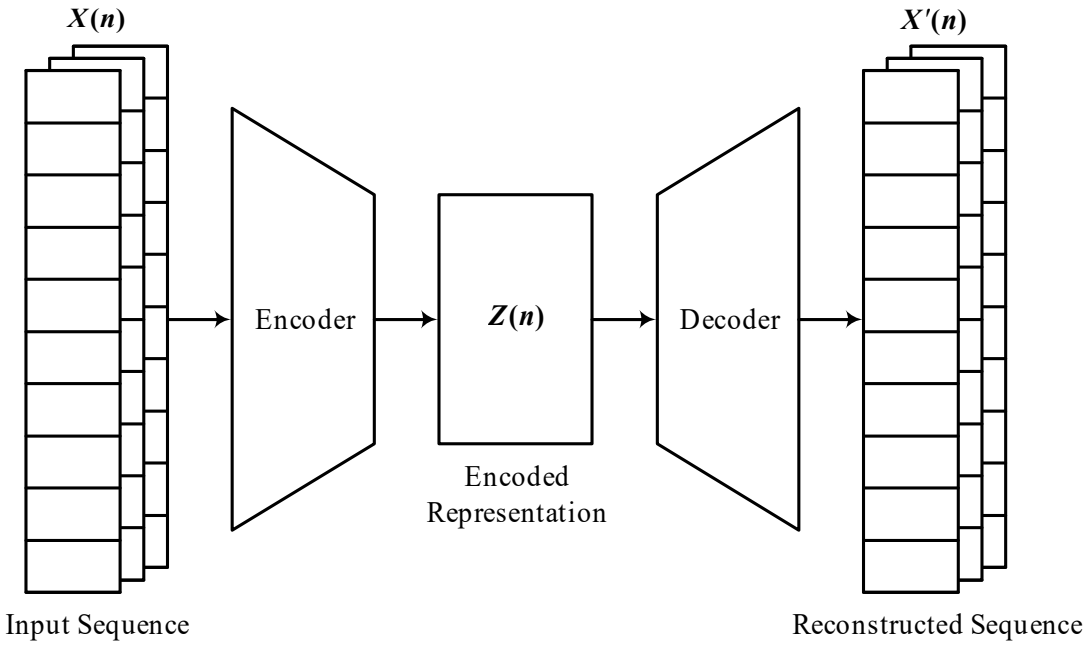


그림 4. LSTM-AE 구조

LSTM-AE는 두 가지의 특징이 존재한다. 먼저, LSTM-AE는 입력 변수가 적을 경우, 학습 효율이 증가하는 특징을 가지고 있다. 이러한 특징을 위해서 데이터 전처리 과정에서 수행된 46개 변수만을 활용하였다. 선정된 46개 변수는 수집된 비정상 시나리오의 예상되는 증상 중 주요 변수만을 고려하여 선택되었다. 또한, 입력층과 출력층에서 시계열 데이터를 요구하기 때문에 데이터 전처리 과정의 시계열 데이터 처리도 사용하였으며, 인공지능 학습 시 데이터 편향의 방지를 위해 정규화 또한 수행하였다. 두 번째로, 데이터 불균형을 해소할 수 있다는 점이다. 데이터 불균형이란 각 데이터의 범주가 가지고 있는 데이터의 양에 차이가 큰 경우를 의미한다. 이는 제시된 표1로부터 확인할 수 있다. 예를 들어, CNS를 활용하여 수집한 데이터의 양은 각 범주에 따라 상이하며 이는 데이터 불균형을 의미한다. 실제로 대부분의 산업 환경 내에서 비정상 데이터보다 정상 데이터가 압도적으로 많은 경우가 대다수이다. 이렇게 수집된

데이터는 데이터 불균형을 내포하고 있으며, 이를 인공지능 방법론에 학습하게 될 경우 압도적으로 많은 정상 데이터에 편향적으로 반응하는 인공지능 모델이 형성되기 때문에 실제로 적용하는데 있어서 많은 제약이 존재한다. 하지만, LSTM-AE는 전체 데이터 범주를 훈련시키지 않고 한가지의 데이터 범주만을 훈련시켜 이진분류를 수행할 수 있는 방법론이다. 한 가지 데이터 범주만을 갖는 데이터를 학습할 경우, 훈련된 LSTM-AE 모델은 학습된 데이터에 한해 높은 정확도로 재구성을 수행하며, 학습되지 않은 데이터를 입력할 경우 재구성에 실패하게 된다. 하지만, 출력으로 재구성되는 데이터는 약간의 오차(Error)가 포함된 형태로 출력되게 된다. 이러한 오류는 재구성 오차(Reconstruction Error)로 정의되며 식 (2)와 같이 표현된다.

$$L(x,x') = \| x - x' \|^2 \tag{2}$$

이 때, x는 입력 값을 의미하며, x'는 재구성되어 출력되는 출력 값을 의미한다. 이러한 재구성 오차의 평균값과 표준 편차를 활용하여 이진 분류 지표로 활용되는 문턱값(Threshold)을 산출하며, 식 (3)과 같이 계산된다.

$$Threshold = \mu \pm 3\sigma \tag{3}$$

이때, μ는 재구성 오차의 평균값이며, σ는 재구성 오차의 표준 편차를 의미한다. 산업계에서 많이 쓰이는 3시그마를 사용하였으며, 이때의 신뢰도는 99.7%이다. 비정상 상태를 탐지하는 것이 목표이므로 99.7%까지는 정상 상태로 판단하며 그 이외의 구간에서는 비정상 상태로 판단하게 된다. 이러한 원리를 활용하여 이진분류를 수행할 수 있으며, 문턱 값을 기준으로 아래에 재구성 오차 값이 존재하면 정상 상태로 판단하며, 위에 존재할 경우 비정상 상태로 판단한다.

제 2 절 LightGBM

Gradient Boosting Decision Tree(GBDT)는 효율성, 정확도, 해석 가능성이 높아

널리 활용되는 기계 학습 알고리즘으로 다중 분류, 예측과 같이 다양한 기계 학습에서 높은 성능을 보여준다. 하지만, 최근 빅데이터의 등장으로 인해 GBDT는 정확도와 효율성 사이에서 상충되는 문제점이 발생하였다. 이는 GBDT의 기존 구현이 입력되는 각 변수마다 가능한 모든 분할점(노드)에 대해 정보 획득을 평가하기 위해 데이터 개체 모두를 확인해야하기 때문에 시간 소모가 매우 많기 때문이다. 이러한 문제 해결을 위해 Ke 등 [11]은 LightGBM 방법론을 소개하였으며, 두 가지 핵심 기술인 Gradient-based One-Side Sampling(GOSS)과 Exclusive Feature Bundling(EFB)를 제시하였다.

먼저, GOSS는 학습 데이터의 개수를 줄이기 위해서 작은 기울기를 갖는 데이터 인스턴스의 상당 부분을 제외하고, 나머지 부분만을 이용하여 정보 획득을 추정한다. 이는 더 큰 기울기를 가진 데이터 인스턴스가 정보 획득을 계산 하는데 더 중요한 역할을 수행한다고 증명되었기 때문이며, GOSS가 훨씬 더 작은 데이터 크기를 가지고 다소 정확한 정보 획득 추정치를 얻을 수 있기 때문이다.

두 번째로, EFB는 변수의 개수를 감소시키기 위해서 상호 배타적인 변수를 묶는 기법이다. 변수가 많고 데이터가 산발적(Sparse)인 경우, 변수들이 동시에 0의 값을 갖는 경우는 드물다. 이러한 변수들이 서로 배타적이거나 배타적인 특성을 갖는 경우, 각 변수들이 가지고 있는 정보 손실을 최소화하면서 하나의 변수로 결합(Bundling)할 수 있다. 이를 통해 정확도 손실 없이 GBDT의 훈련을 상당히 가속화할 수 있다. 이러한 핵심 기술 적용을 통해 기존의 GBDT와 비슷한 성능을 내며, 빠른 계산 속도와 적은 메모리 사용을 가능하게 하였다.

이에 본 논문에서는 비정상 시나리오 진단 및 설명 가능한 인공지능의 활용을 위해서 높은 성능과 해석 가능성을 갖는 LightGBM 방법론을 사용하였다. 해석 가능성은 추후 설명 가능한 인공지능을 활용하기에 매우 용이하며, 빠른 시간 내에 근거를 도출할 수 있다. 모델 훈련을 위해서 수집된 데이터를 활용하였으며, 각 비정상 시나리오에 번호를 할당하여 라벨링을 한 후 지도학습을 통해 다중 분류 모델을 생성하였다. 또한, 출력 값을 확률로 표현하기 위해서 softmax 함수를 사용하였으며, softmax 함수는 출력으로 0과 1사이의 값으로 모두 정규화하며 출력 값들의 총합은 항상 1이 되는 특성을 가진 함수이다. softmax 함수는 식 (4)와 같이 표현되며, 0과 1사이로 출력되는 값에 100을 곱하여 확률로서 결과 값을 출력하도록 하였다. 구체적인 적용 결과

는 다음 장에서 설명한다.

$$f(\vec{x})_i = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \text{ for } i = 1,...,K \tag{4}$$

LightGBM 모델은 비정상 시나리오 진단을 위해서만 사용되는 것이 아닌 설명 가능한 인공지능을 통해 진단 근거에 대한 해석 또한 수행한다. 이러한 해석은 입력 변수를 기반으로 수행되기 때문에 입력 변수가 10개가 존재할 경우, 진단된 이유를 입력된 변수 내에서 찾게 된다. 그렇기에 양질의 해석 정보를 출력하기 위해서는 입력 변수를 적절하게 선정하는 것이 매우 중요하다. 해당 모델 학습을 위해서 데이터 전처리 과정이 수행된 137개의 변수를 사용하였다. 137개의 변수는 Feature Importance 기능을 활용하여 모델 구성에 있어 중요도가 높은 변수만을 선택하였다.

제 3 절 설명 가능한 인공지능

앞서 설명한 인공지능 방법론에는 두 가지 측면의 문제점이 있다. 먼저, 연구자 입장에서는 세부적으로 해당 모델의 연산 과정을 파악할 수 없다는 문제를 가지고 있다. 이는 모델 내 신경망의 가중치 전달 과정이 동시 다발적으로 발생하는 복잡성을 가지고 있고, 깊은 구조를 가질수록 이는 더 심화되기 때문이다. 두 번째로는 사용자 입장에서 인공지능으로부터 도출된 결과가 어떠한 방식으로 진단되었는지에 대한 인과적 설명을 제공받지 못하기 때문에 결과에 대해 만족하지 못하고 신뢰하지 못하게 된다는 문제를 가지고 있다. 이러한 문제는 많은 연구자들이 고민하고 있는 문제이다[12].

이를 해결하기 위해서 미국 국방부 산하 방위 고등 연구 계획국(DARPA; Defense Advanced Research Projects Agency)에서 David Gunning 주도로 설명 가능한 인공지능 프로젝트를 수행하고 있다[13]. 해당 프로젝트의 연구 분야는 크게 예측과정에 대한 추적을 통해 딥러닝 과정을 이해할 수 있도록 도와주는 설명 가능한 모델에 대한 개발과 예측 결과에 대한 설명을 어떠한 방식으로 제공할 것인지에 대한 설명 인터페이스 개발 연구로 나뉜다. 해당 프로젝트의 계획에 따르면, 기존의 인공지능의 블랙박

스 문제점을 개선할 수 있다고 한다. 그림 5에서 훈련데이터를 인공지능 모델에 학습을 시킨 후 결과로부터 93%의 확률로 고양이라는 출력 값을 얻을 수 있다. 하지만, 사용자 입장에서는 어떤 부분 때문에 고양이가 진단이 되었는가에 대한 의문이 들며 인공지능에 대한 신뢰성이 낮아지게 된다. 이를 해결하기 위해서 설명 가능한 인공지능을 적용하면 그림 5에서 고양이의 털, 수염, 발톱 등으로부터 고양이라고 사용자에게 알려주는 것을 확인할 수 있다. 현재의 인공지능으로는 진단 결과가 왜 이런지, 다른 진단 결과는 왜 출력되지 않는지, 진단이 성공했는지, 진단이 실패했는지 등 사용자의 의문을 해소할 수 없다. 하지만, 설명 가능한 인공지능을 활용하면 왜 이런 진단 결과가 나왔는지, 왜 다른 진단 결과는 선택받지 못했는지, 믿을 수 있는지, 성공인지 실패인지 등의 의문을 해소시킬 수 있을 것으로 기대된다. 이에 비정상 시나리오 진단 결과에 대한 의문을 해소하고자 설명 가능한 인공지능 방법론인 SHAP을 사용하였다.

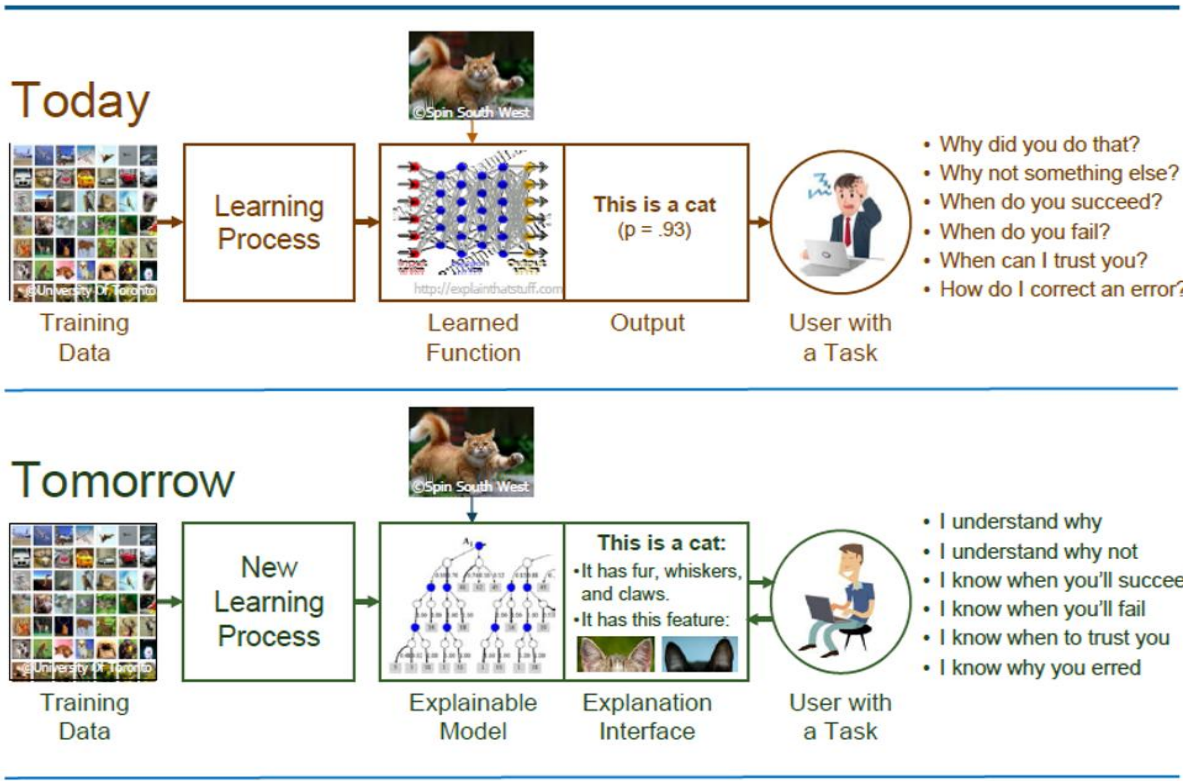


그림 5. 설명 가능한 인공지능의 필요성[13]

SHAP은 로이드 섀플리가 만든 이론 위에 변수 사이의 독립성을 근거로 덧셈이 가

능하게 활용도를 넓힌 방법으로 논문을 통해 공개되었다[14, 15]. SHAP은 shapley value와 변수 간 독립성을 핵심 아이디어로 사용하며, 전체성과를 창출하는데 있어서 각 변수가 얼마나 공헌했는지를 수치적으로 계산할 수 있다. 계산식은 식 (5)와 같이 표현되며, 이를 간략히 표현하자면 각 변수의 기여도는 해당 변수의 기여도를 제외했을 때 전체성과의 변화 정도로 계산이 가능하다.

$$\phi_i(\nu)=\sum_{S\in N\setminus\{i\}}\frac{|S|!(n-|S|-1)!}{n!}(\nu(S\cup\{i\})-\nu(S)) \tag{5}$$

이때, ϕ_i는 i 변수에 대한 shapley value이며, n은 전체 변수 개수, S는 전체 그룹에서 i 변수를 제외한 모든 집합, $|S|$는 S의 원소 개수, $\nu(S)$는 i 변수를 제외하고 나머지 부분 집합이 결과에 공헌한 기여도, $\nu(S\cup\{i\})$는 i 변수를 포함한 전체 기여도를 의미한다. 즉, i 변수가 기여하는 정도는 전체 기여도에서 i 변수가 제외된 기여도의 합을 뺀 값이다. 예를 들어, x_1, x_2, x_3, x_4의 4가지 변수로 구성된 그룹이 있다고 가정할 때 전체 그룹에 대해서 x_3 변수가 기여한 가치 S는 식 (6)과 같이 계산할 수 있다.

$$\begin{aligned}\phi_3(\nu)&=\frac{1}{4!}\sum_R\left[\nu\left(P_3^R\cup\{x_3\}\right)-\nu\left(P_3^R\right)\right]\\&=\frac{1}{4!}\sum_R\left[\nu\left(\{x_1,x_2,x_4\}^R\cup\{x_3\}\right)-\nu\left(\{x_1,x_2,x_4\}^R\right)\right]\end{aligned} \tag{6}$$

이러한 식 (6)은 전체 그룹에서 표현 가능한 모든 조합과 변수 x_3를 제외한 그룹의 조합을 빼서 평균을 내는 방식이다. 또한 식 (6)과 같이 단일 변수에 대한 shapley value 이외에 $\{x_1,x_2\}$ 변수의 shapley value는 $\phi_1(\nu), \phi_2(\nu)$의 합으로 구할 수 있다.

그림 6은 각각의 변수를 기준으로 shapley value를 계산한 결과를 보여주며, $\phi_{0,1,2,3}$은 선택된 결과에 긍정적으로 기여한 변수이며 ϕ_4는 부정적으로 기여한 변수를 의미한다.

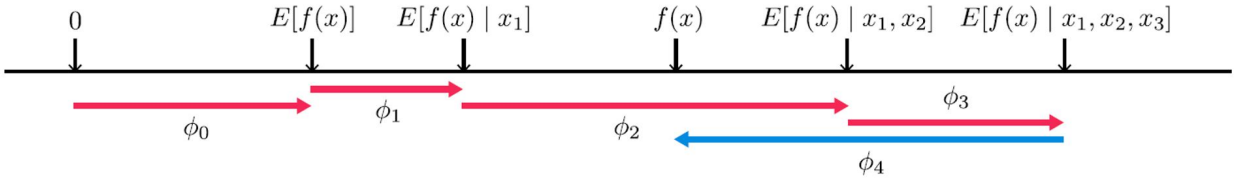


그림 6. 그룹 f의 출력을 각각의 ϕ_i로 계산한 결과[14]

이러한 SHAP 방법론은 그림 7의 구조를 활용하여 인공지능을 해석한다. 해석 대상인 인공지능의 입력 변수를 기여도로 산출하는 방식을 활용한다.

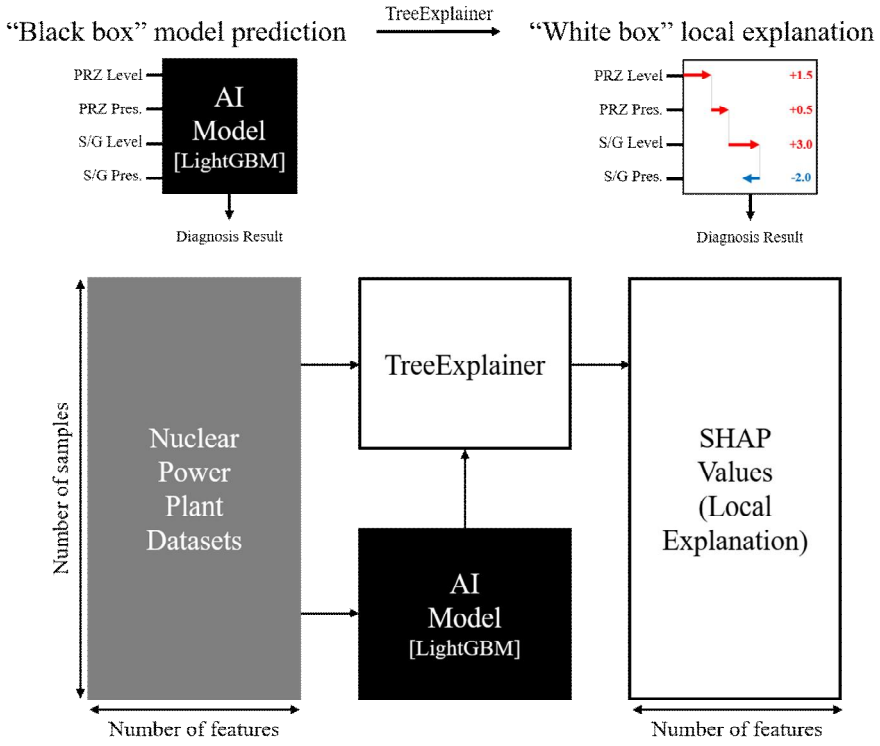


그림 7. 설명 가능한 인공지능의 적용 구조[16]

제 4 절 규칙 기반 시스템

규칙 기반 시스템은 규칙 기반 전문가 시스템이라고도 불리며, 특정 분야나 주제에 대한 지식이 풍부하고, 관련 문제를 푸는데 능숙한 사람으로 주제에 대한 전문 지식을 갖춘 전문가에 의해 만들어진다. 이러한 규칙 기반 시스템은 크게 기반지식, 데이터베이스, 추론 엔진으로 구성된다. 기반지식은 전문가에 의해 IF-THEN rule의 형태로 작성되며, 데이터베이스는 입력되는 데이터를 의미한다. 추론 엔진은 전문가에 의해 작성된 기반지식과 데이터베이스를 결합하여 입력되는 데이터가 작성된 IF-THEN rule과 부합하는지를 확인한다. 부합할 경우와 부합하지 않을 경우 또한 전문가에 의해 작성된 기반 지식에 따라 출력 값 변경이 가능하다. 그림 8은 실제 적용된 규칙 기반 시스템의 예제로 입력되는 발전소 변수와 IF-THEN rule의 비교를 통해 출력 값을 형성하는 것을 확인할 수 있다. 이때, 입력되는 데이터는 기반지식의 확장성을 넓히기 위해 2,222개 변수를 갖는 Non-Normalized Data가 활용된다.

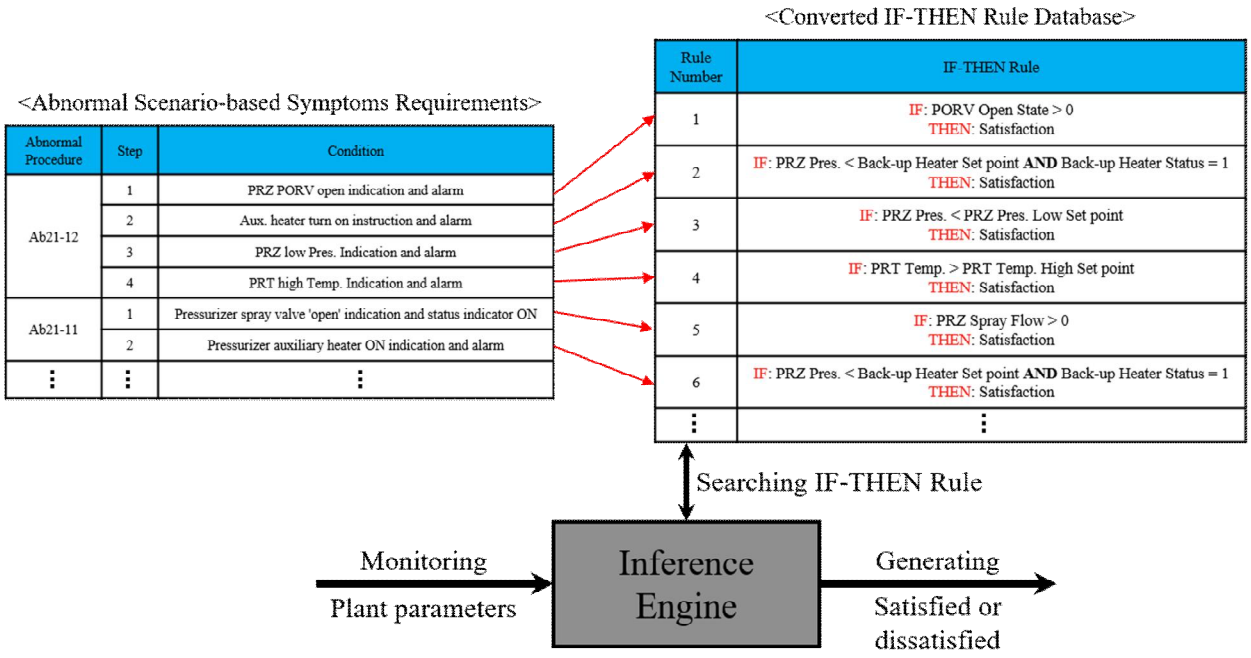


그림 8. 규칙 기반 시스템 구조

제 4 장 통합 비정상 진단 알고리즘

4장에서는 그림 1의 로드맵 중 4단계 통합 비정상 진단 알고리즘에 대해서 소개한다. 설계된 알고리즘은 크게 알고리즘-모듈-기능으로 순으로 체계화 되어 있으며, 모듈은 진단 모듈과 검증 모듈로 구성되어 있다. 각 모듈의 기능 설계를 위해서 Yang과 Kim의 알고리즘 일부를 참고하여 구성하였다[4, 5]. 진단 모듈은 1) 훈련 여부 진단 기능, 2) 비정상 시나리오 진단 기능으로, 검증 모듈은 1) 진단 결과 검증 기능, 2) 예상 증상 만족 평가 기능, 3) 진단 근거 도출 기능으로 구성되어있다. 표 2는 통합 비정상 진단 알고리즘의 개요를 나타나며, 그림 9는 통합 비정상 진단 알고리즘 구조를 보여준다. 알고리즘 구현을 위해 3장에서 소개된 인공지능 방법론인 LSTM-AE와 LightGBM, 설명 가능한 인공지능인 SHAP, 그리고 규칙 기반 시스템을 활용하였다.

표 2. 통합 비정상 진단 알고리즘 개요

모듈	기능	방법론	입력 값	출력 값
진단 모듈	훈련 여부 진단 기능	LSTM-AE	Time Series Normalized data	훈련된 상태 또는 훈련되지 않은 상태
	비정상 시나리오 진단 기능	LightGBM	Normalized data	비정상 시나리오
검증 모듈	진단 결과 검증 기능	LSTM-AE	Time Series Normalized data	진단 성공 또는 진단 실패
	예상 증상 만족 평가 기능	규칙 기반 시스템	Non-Normalized data	증상 만족 또는 증상 불만족
	진단 근거 도출 기능	SHAP	Normalized data	진단 근거

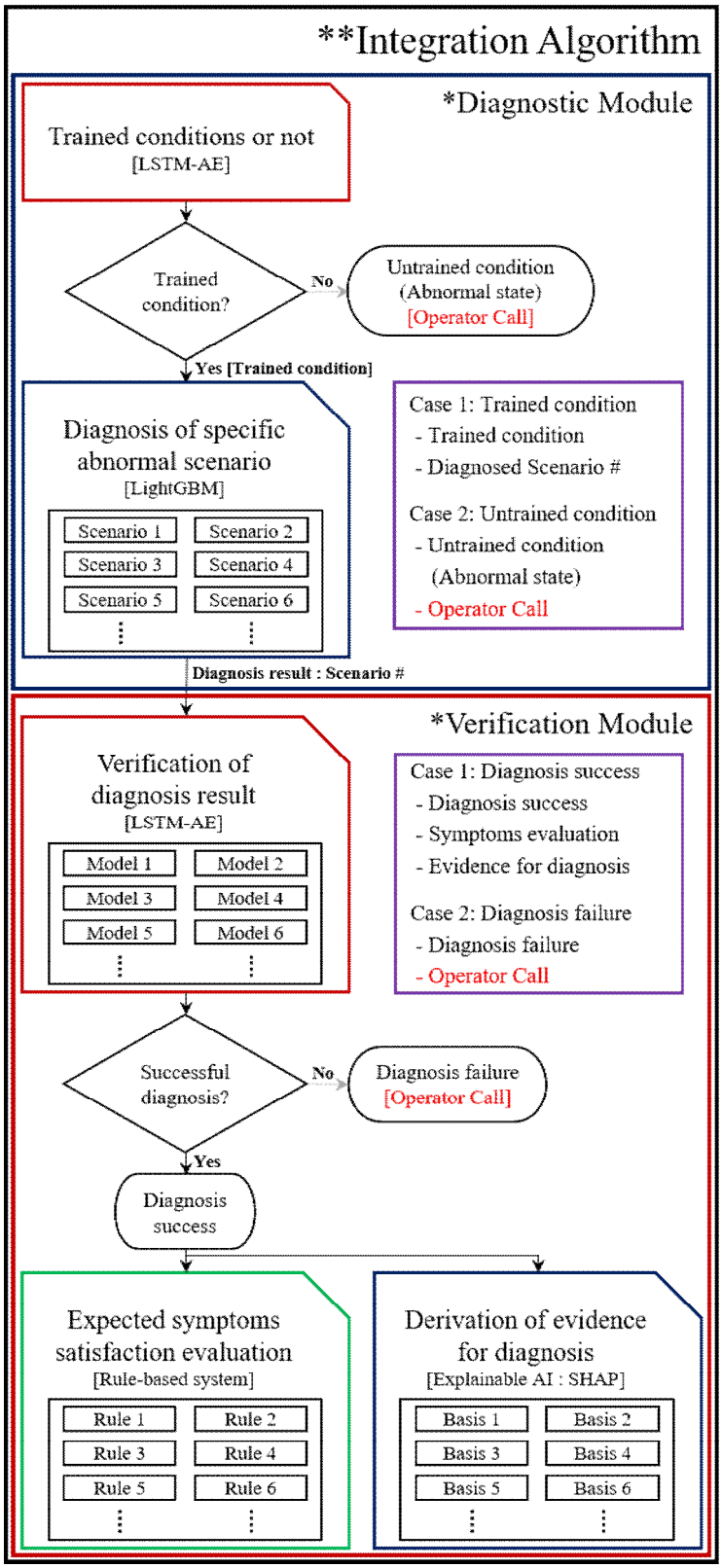


그림 9. 통합 비정상 진단 알고리즘 구조

제 1 절 진단 모듈

진단 모듈은 인공지능 방법론인 LSTM-AE와 LightGBM을 사용하여 구현하였으며, 1) 훈련 여부 진단 기능, 2) 비정상 시나리오 진단 기능으로 구성되어 있다. 각 기능은 절차적으로 연결되어 있으며, 특정 조건이 만족되어야 다음 기능으로 진행된다.

1. 훈련 여부 진단 기능

훈련 여부 진단 기능은 LSTM-AE를 활용하여 구현하였다. 데이터 전처리 과정을 통해 전처리된 데이터(Time Series Normalized data)를 입력으로 받아 훈련된 상태 또는 훈련되지 않은 상태로 출력된다. 훈련된 상태로 진단된 경우 비정상 시나리오 진단 기능으로 진입하지만, 훈련되지 않은 상태로 진단된 경우에는 인공지의 특성으로 인해 정확한 진단이 제한되기 때문에 운전원 호출을 수행한다. 인공지능으로 진단이 제한되는 이유는 인공지능은 제공되는 데이터의 패턴 등을 기반으로 학습을 진행하기 때문에 학습되지 않은 데이터가 입력될 경우 학습된 패턴 내에서 어떻게든 적합한 패턴을 도출하기 때문에 부정확한 진단을 수행할 수 있기 때문이다. 이러한 과정은 인공지능으로 입력된 데이터를 진단할 수 있는가, 진단할 수 없는가에 대한 내용이므로 인공지능 기반의 통합 비정상 진단 알고리즘에서 해당 기능은 매우 중요하다. 기능 구현을 위해 활용되는 LSTM-AE 모델 훈련을 위해서 수집된 데이터를 훈련 데이터와 테스트 데이터로 분류하였다. 훈련 데이터는 15가지 비정상 시나리오와 정상 시나리오를 활용하였으며, 테스트 데이터는 5가지 비정상 시나리오를 활용하였다(표 1 참조). 또한, 훈련 데이터는 정상 시나리오를 포함하고 있기 때문에 훈련되지 않은 상태로 출력된 경우는 비정상 상태의 의미를 내포하고 있다. 훈련 여부 진단 기능의 결과는 그림 10과 11과 같이 표현된다. 그림 10은 훈련된 상태로 진단된 경우로 가압기 살수밸브 고장 열림 시나리오를 적용한 결과이다. 설정된 문턱 값을 기준으로 재구성 오차가 아래에 위치해있기 때문에 훈련된 상태로 판단되는 것을 확인할 수 있다. 그림 11은 훈련되지 않은 상태로 진단된 경우로 실제 미훈련 시나리오로 활용한 주증기관 차단 시나리오를 적용한 결과이다. 마찬가지로 기설정된 문턱 값을 기준으로 31초에 재구성 오차가 위에 위치해있기 때문에 훈련되지 않은 상태로 판단되는 것을 확인할 수 있으며 고장이 주입된 30초 이후 바로 훈련되지 않은 상태로 정확하게 진단하는 성능 또한 보여준다.

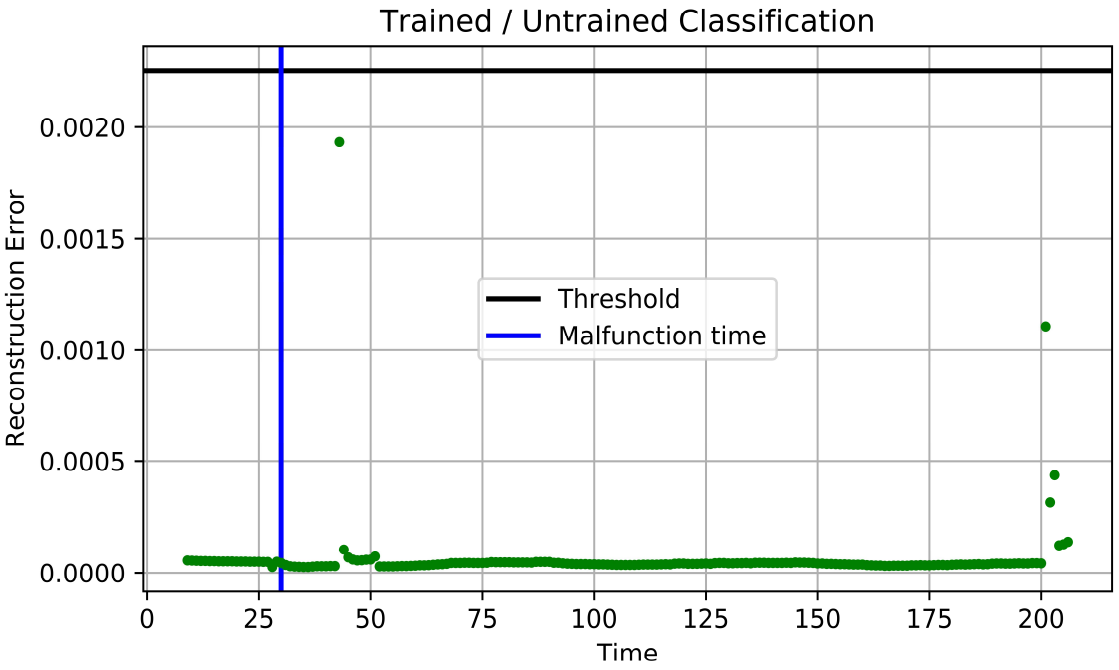


그림 10. 훈련 여부 진단 기능 결과: 훈련된 상태

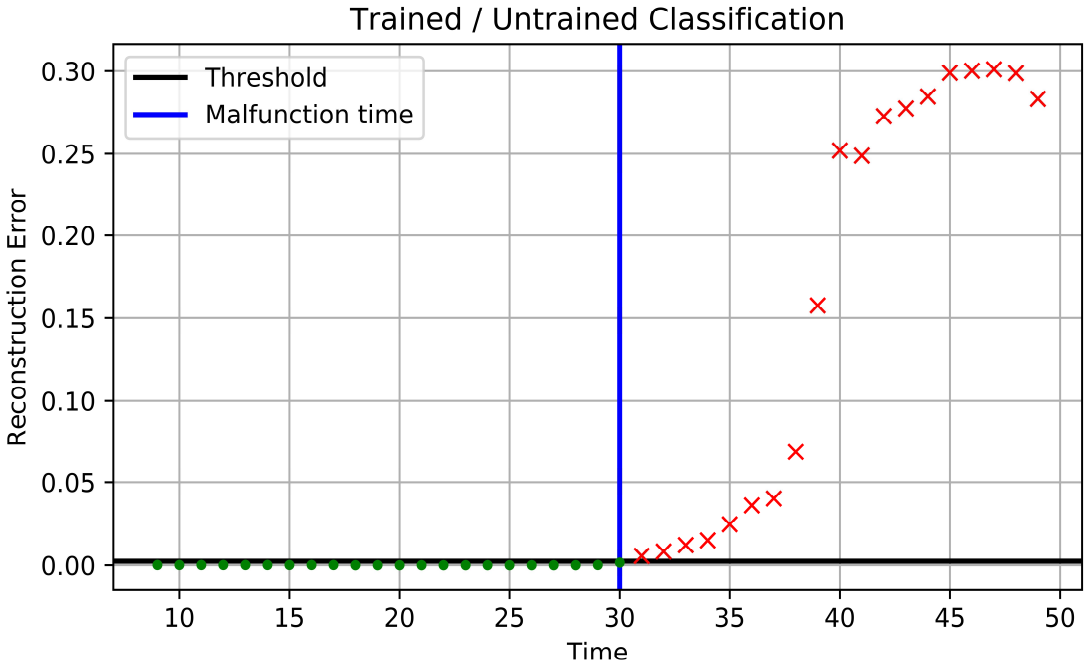


그림 11. 훈련 여부 진단 기능 결과: 훈련되지 않은 상태

2. 비정상 시나리오 진단 기능

비정상 시나리오 진단 기능은 인공지능 방법론 중 지도 학습 모델인 LightGBM을 활용하여 구현되었다. 다중 분류 형태로 출력 값이 형성되며, 수집된 비정상 시나리오 16가지(정상 시나리오, 15가지 비정상 시나리오) 중 하나의 형태로 출력된다. 또한, 해당 기능은 훈련 여부 진단 기능에서 훈련된 상태로 진단된 경우에 한해 수행된다. 기능 구현을 위해 활용되는 LightGBM 방법론 훈련을 위해서 수집된 데이터를 훈련 데이터와 테스트 데이터로 분류하였다. 훈련 데이터는 앞선 훈련 여부 진단 기능에서 테스트 데이터로 활용된 5가지 비정상 시나리오는 제외하고 15가지 비정상 시나리오와 정상 시나리오를 활용하였다. 테스트 데이터의 경우에는 훈련 데이터 중 일부 시나리오를 추출하여 사용하였다. 이때, 훈련 데이터와 테스트 데이터는 서로 독립적인 데이터를 갖는다(표 1 참고). 진단 결과는 그림 12와 같이 표현된다. 그림 12는 테스트 데이터인 가압기 PORV 열림 시나리오(Ab21-12)를 적용한 결과로 100%의 확률로 진단되었음을 확인할 수 있다.

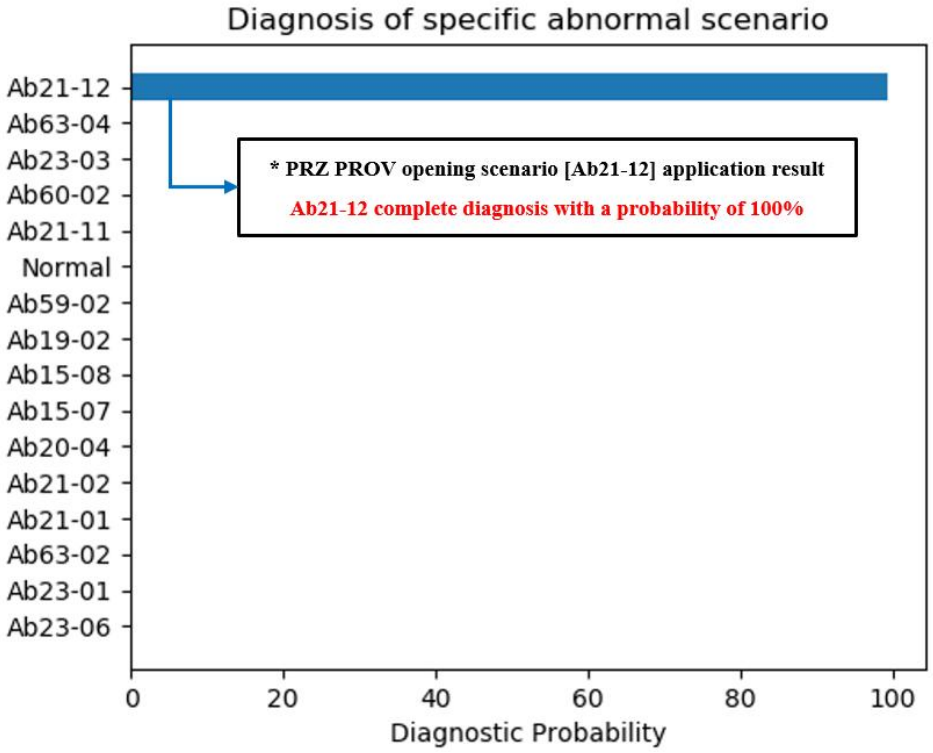


그림 12. 비정상 시나리오 진단 기능 결과

제 2 절 검증 모듈

검증 모듈은 진단 모듈로부터 도출된 결과에 대해서 검증하고 신뢰도 제고를 위한 목적으로 설계 및 구현하였다. 모듈 구현을 위해서 인공지능 방법론인 LSTM-AE와 설명 가능한 인공지능 방법론인 SHAP, 그리고 규칙 기반 시스템을 활용하였다. 검증 모듈은 3개의 세부 기능인 1) 진단 결과 검증 기능, 2) 예상 증상 만족 평가 기능, 3) 진단 근거 도출 기능으로 구성되어 있다. 각 기능은 절차적으로 연결되어 있으며, 특정 조건 만족시 다음 기능으로 진행되는 구조를 가지고 있다. 또한, 검증 모듈은 진단 모듈 내의 비정상 시나리오 진단 기능의 결과 값을 입력 값으로 활용한다.

1. 진단 결과 검증 기능

먼저, 진단 결과 검증 기능은 인공지능 방법론 중 비지도 학습 모델인 LSTM-AE를 활용하여 구현되었다. 이진 분류 형태로 출력 값이 형성되며, 이는 진단 성공 또는 진단 실패로 표현된다. 출력 값이 진단 성공으로 도출된 경우 다음 단계인 증상 만족 평가 기능과 진단 근거 도출 기능으로 진입하지만, 진단 실패인 경우에는 통합 비정상 진단 알고리즘 종료와 함께 운전원 호출을 수행하게 된다. 기능 구현을 위해 활용되는 LSTM-AE 방법론 훈련을 위해서 수집된 데이터를 훈련 데이터와 테스트 데이터로 분류하였다. 이때, 입력 값으로 활용되는 비정상 시나리오 진단 기능의 결과를 기반으로 훈련 데이터와 테스트 데이터는 형성된다. 만약, 입력 값으로 가압기 PORV 열림 시나리오로 진단된 결과가 입력된 경우 훈련 데이터는 가압기 PORV 열림 시나리오를 활용하고 테스트 데이터로는 훈련 데이터 이외의 20가지 시나리오를 활용한다. 이러한 형태로 입력 값이 16가지 시나리오로 입력될 경우의 수가 있으므로, 각 입력 값에 대응하여 16가지 LSTM-AE 모델을 형성하여 대응하도록 설계되었다.

진단 결과 검증 기능의 결과는 그림 13과 14와 같이 표현된다. 그림 13은 진단 성공으로 진단된 경우로 입력되는 가압기 PORV 열림 시나리오와 대응되는 LSTM-AE 모델을 적용하여 도출된 결과이다. 설정된 문턱 값을 기준으로 재구성 오차가 아래에 위치해있기 때문에 진단 성공으로 판단되는 것을 확인할 수 있다. 그림 14는 진단 실패로 진단된 경우로 입력되는 가압기 PORV 열림 시나리오와 대응되지 않는 LSTM-AE 모델을 적용하여 도출된 결과이다. 기설정된 문턱 값을 기준으로 재구성

오차가 위에 위치해있기 때문에 진단 실패로 판단되는 것을 확인할 수 있다. 하지만, 고장 주입(30초) 150초 이후에 진단 실패로 진단되며, 이는 적용된 LSTM-AE 모델이 진단된 시나리오와 증상이 유사한 가압기 안전밸브 고장 시나리오에 대응되는 모델을 활용하였기 때문에 다소 진단이 늦음을 확인할 수 있다.

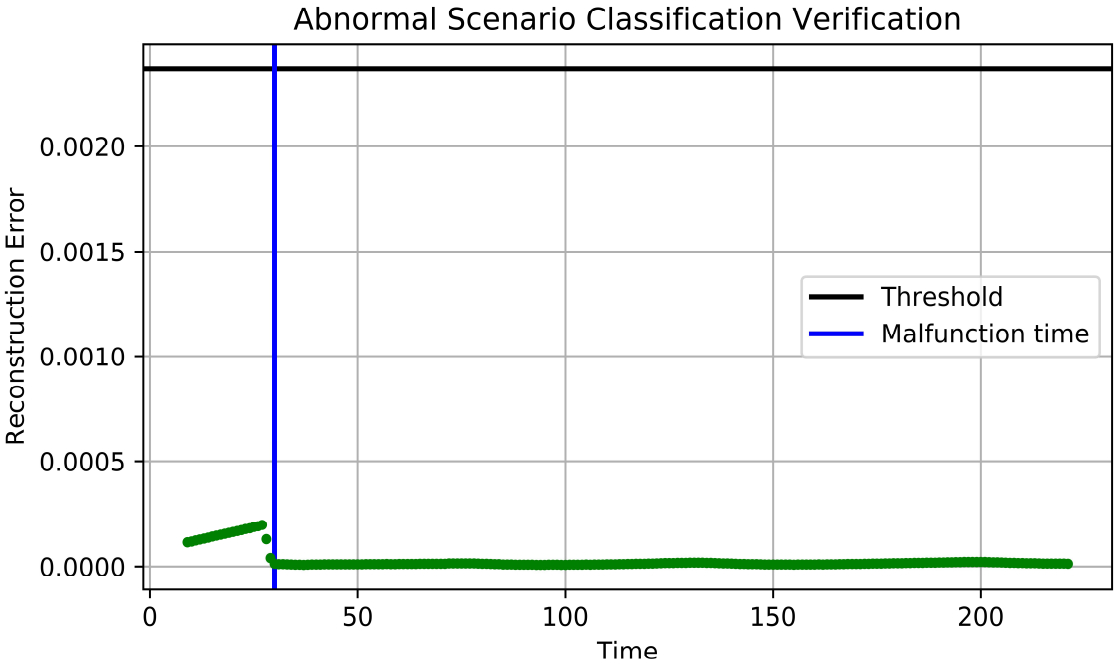


그림 13. 진단 결과 검증 기능 결과: 진단 성공

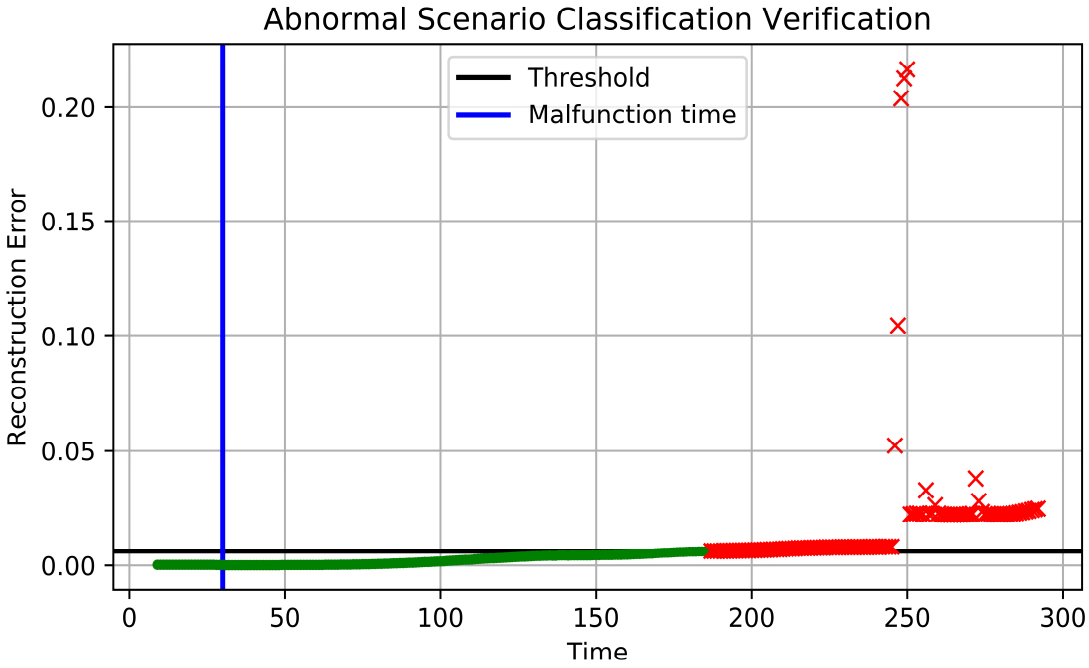


그림 14. 진단 결과 검증 기능 결과: 진단 실패

2. 예상 증상 만족 평가 기능

예상 증상 만족 평가 기능은 규칙 기반 시스템을 활용하여 구현되었다. 이진 분류 형태로 출력 값이 형성되며, 이는 증상 만족 또는 증상 불만족으로 표현된다. 기능 구현을 위해서 IF-THEN Rule을 작성하였으며, 작성을 위해서 15가지 비정상 시나리오의 예상되는 증상을 활용하였다. 예를 들어, 예상되는 증상은 급격한 가압기 수위의 증가 또는 특정 알람의 발생 등이 존재한다. 이를 기반으로 조건을 만족한 경우에는 증상 만족으로 평가하였으며, 조건을 불만족한 경우 증상 불만족으로 평가하였다. 하지만, 각 비정상 시나리오별로 증상이 다르며 증상의 수 또한 상이하다. 이를 정량적으로 평가하여 운전원에게 정보를 제공하고자 전체 증상의 수와 만족한 증상의 수를 백분율로 환산하였다.

예상 증상 만족 평가 기능의 결과는 그림 15와 같이 표현된다. 그림 15는 가압기 PORV 열림 시나리오를 적용한 결과이며, 가압기 PORV 열림 시나리오(ab21_12)와 CVCS 누설 시나리오(ab23_03)의 증상이 100% 만족했음을 확인할 수 있다. 하지만, 해당 결과는 단순 수치로만 표시되며 어떤 조건이 어떻게 만족되었는지는 확인할 수 없으며, 분모가 다르기 때문에 정확한 평가가 어렵다는 단점이 존재한다. 이러한 단점은 다음 장인 설명 인터페이스에서 보완하였다.

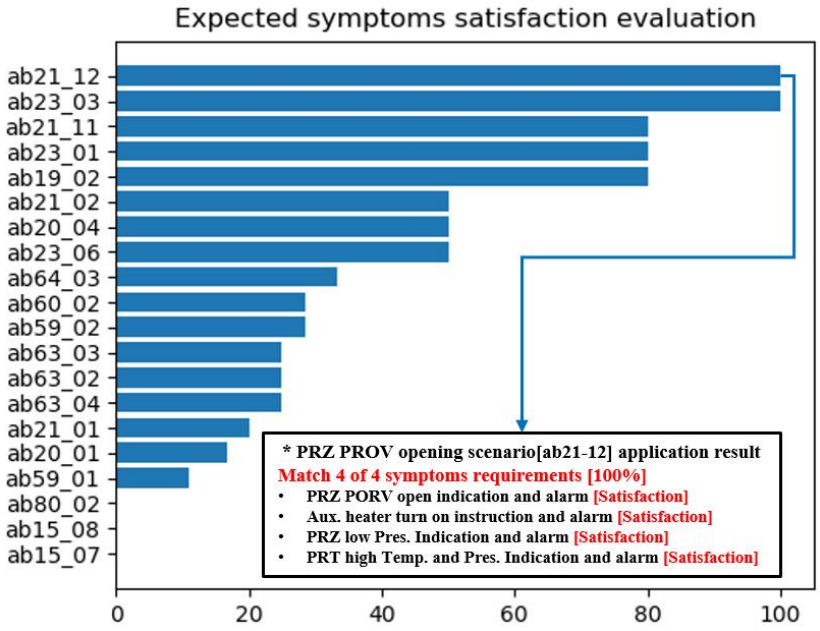


그림 15. 예상 증상 만족 평가 기능 결과

3. 진단 근거 도출 기능

진단 근거 도출 기능은 설명 가능한 인공지능 방법론인 SHAP을 활용하여 구현되었다. 출력 값으로는 진단 모듈의 비정상 시나리오 진단 기능에서 출력 가능한 15가지 비정상 시나리오와 정상 시나리오에 대해 진단된 시나리오의 경우 진단 근거를 도출하고 진단되지 않은 시나리오의 경우에는 진단되지 않은 근거에 대해서 출력한다.

진단 근거 도출 기능의 결과는 그림 16과 17과 같이 표현된다. 그림 16은 비정상 시나리오 진단 기능의 출력 값이 가압기 살수밸브 고장 열림 시나리오로 출력되었을 경우를 산정하여 해석한 결과이다. 해석 결과로 가압기 살수밸브 유량(ZINST66)이 진단에 있어서 가장 크게 기여했음을 확인할 수 있다.

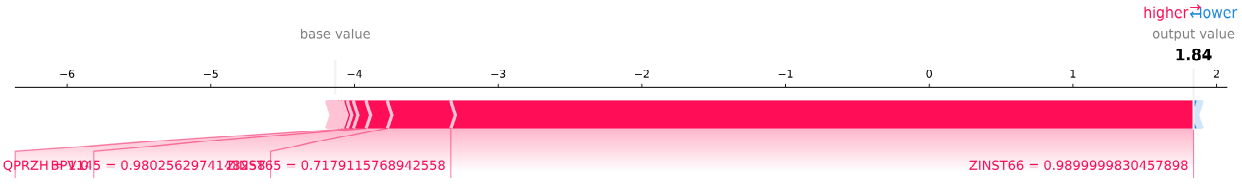


그림 16. 진단된 가압기 살수밸브 고장 열림 시나리오 해석 결과

그림 17은 그림 16에서 적용된 가압기 살수밸브 고장 열림 시나리오와 증상이 유사한 가압기 PORV 열림 시나리오가 진단되지 않은 이유에 대해서 보여준다. 그림 17의 base value 보다 output value가 더 왼쪽에 위치하기 때문에 선택되지 않았음을 확인할 수 있으며, 원인으로는 PORV 개방 상태(BPORV)가 가장 크게 부정적인 영향을 미쳤음을 알 수 있다. 실제 가압기 PORV 열림 시나리오에서는 밸브가 열림으로 인해 발생하는 비정상 사고이므로 닫혀있는 상태에서는 발생할 수 없는 시나리오이다.

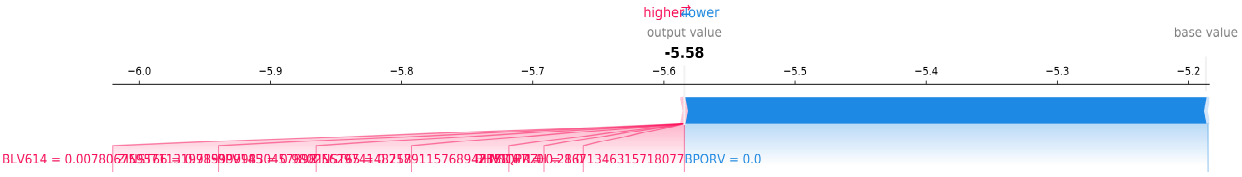


그림 17. 진단되지 않은 가압기 PORV 열림 시나리오 해석 결과

비정상 시나리오 진단 결과를 해석하여 보여준 그림들은 많은 정보가 내포되어 있으나, 이를 운전원에게 그대로 제공하기에는 다소 복잡한 구조로 되어있다. 이를 해결하고자 그래프에서 정보를 추출하여 변수명과 변수의 기여도를 계산하여 그래프 형태로 변형하였으며, 이는 그림 18과 같다. 그림 18은 가압기 PORV 열림 시나리오가 진단된 근거에 대한 내용으로, PORV 개방 상태가 가장 큰 영향을 미쳤음을 확인할 수 있다. 하지만, 그래프 형식의 그림 또한 해당 변수의 설명을 알 수 없으며 단순히 PORV 개방 상태가 기여를 했다는 사실만을 확인할 수 있다. 구체적으로 PORV 개방 상태가 증가해서인지, 감소해서인지, 열려있기 때문인지, 닫혀있기 때문인지는 알 수 없다. 이러한 문제점을 다음 장인 설명 인터페이스에서 해결하고자 한다.

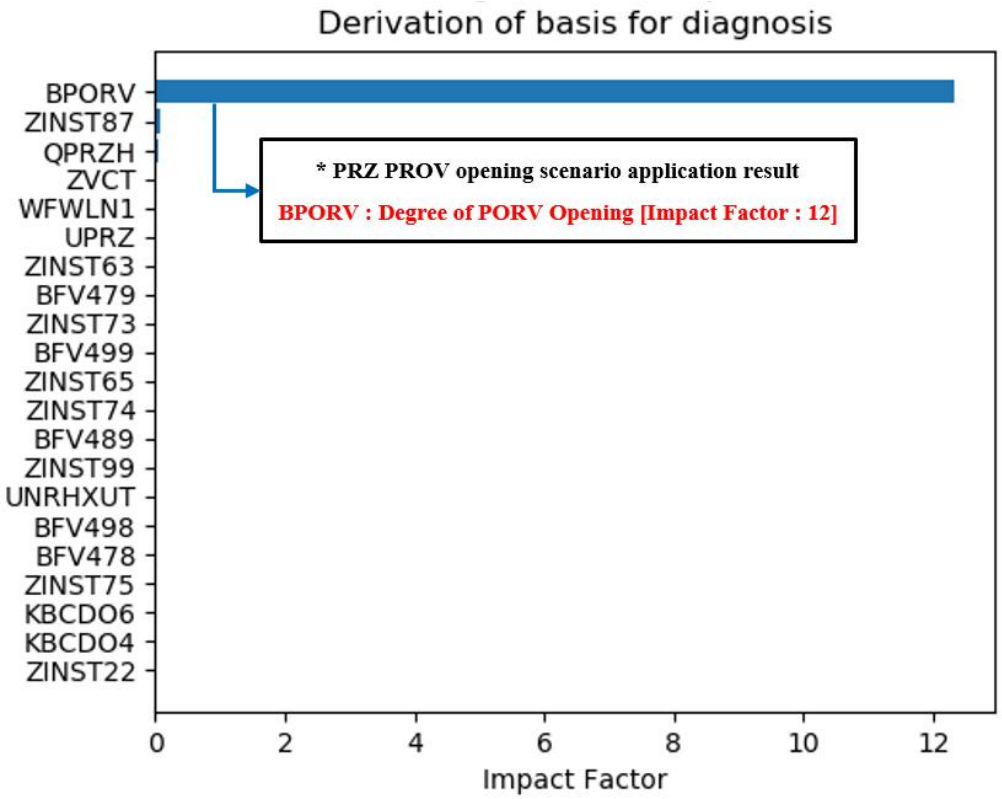


그림 18. 진단 근거 도출 기능 결과

제 5 장 설명 인터페이스

각 방법론을 적용한 통합 비정상 진단 알고리즘의 결과는 그림 19와 같이 1) 훈련 여부 진단 기능, 2) 비정상 시나리오 진단 기능, 3) 진단 결과 검증 기능, 4) 예상 증상 만족 평가 기능, 5) 진단 근거 도출 기능의 결과를 보여준다. 제공되는 정보는 함축적으로 표현되어 운전원의 세부적인 해석이 필요하고 직관적이지 않기에 효율적으로 정보를 전달하기에는 제한된다. 세부적인 해석이란 제공되는 정보 중 그래프의 색, 심볼의 형태, 문턱 값 등을 기준으로 현재 상태에 대해 진단 결과를 파악하는 것을 의미한다. 운전원에게 추가적인 정보를 제공하고자 할 경우, 더 많은 그래프를 추가해야하는 구조적인 문제점이 발생하여 운전원이 한눈에 파악하기에는 다소 제한된다. 이에 본 논문에서는 운전원에게 제공되는 정보의 제한 사항을 해결하고자 직관적인 형태의 설명 인터페이스를 설계하였다.

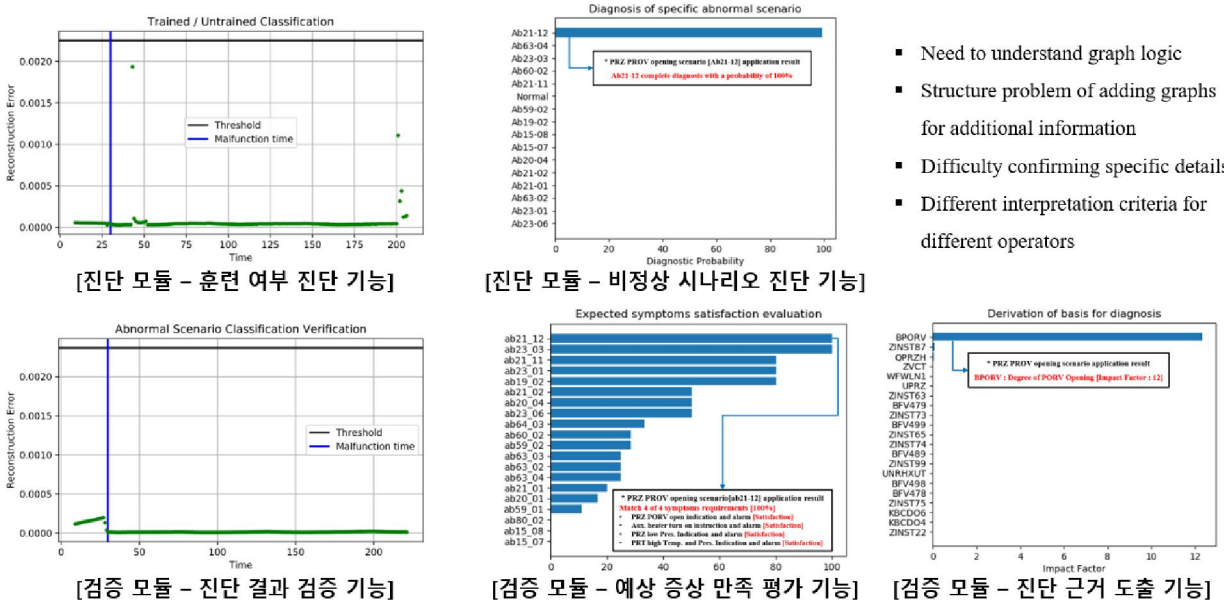


그림 19. 통합 비정상 진단 알고리즘 결과 제공의 문제점

제 1 절 설명 인터페이스 개요

설명 인터페이스는 그림 20과 같이 통합 진단 인터페이스, 증상 감시 인터페이스, 진단 근거 도출 인터페이스로 구성되어 있다. 각 인터페이스는 통합 비정상 진단 알고리즘에서 산출된 결과 값을 토대로 시각화되며, 각 요소는 내부적으로 연계되어있다. 구체적으로 통합 진단 인터페이스에서는 통합 비정상 진단 알고리즘의 훈련 여부 진단 기능, 비정상 시나리오 진단 기능, 진단 결과 검증 기능, 예상 증상 만족 평가 기능이 사용되었으며, 증상 감시 인터페이스는 예상 증상 만족 평가 기능과 연계되어 있다. 진단 근거 도출 인터페이스는 진단 근거 도출 기능과 연계되어 있으며, 각각의 버튼에는 팝업창 기능을 추가하였다. 알람 기능 구현을 위한 색상은 초록색과 빨간색을 구성되며, 초록색은 운전 중인 원자력 발전소의 상태가 안정적일 경우 또는 긍정적인 경우에 사용되며, 빨간색은 불안정할 경우 또는 부정적인 경우에 사용된다. 구현된 팝업창을 모두 실행했을 경우, 그림 21과 같이 표현되며 버튼과의 연관 관계는 화살표를 통해 표시하였다.

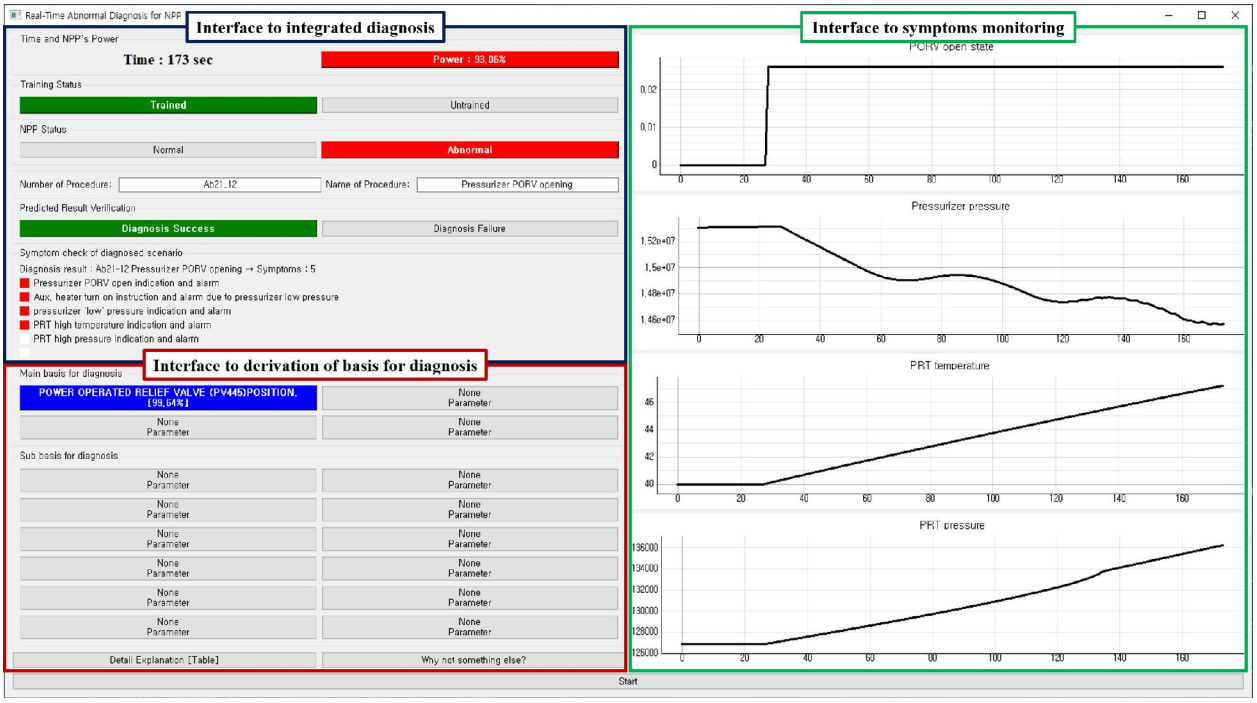


그림 20. 설명 인터페이스 개요

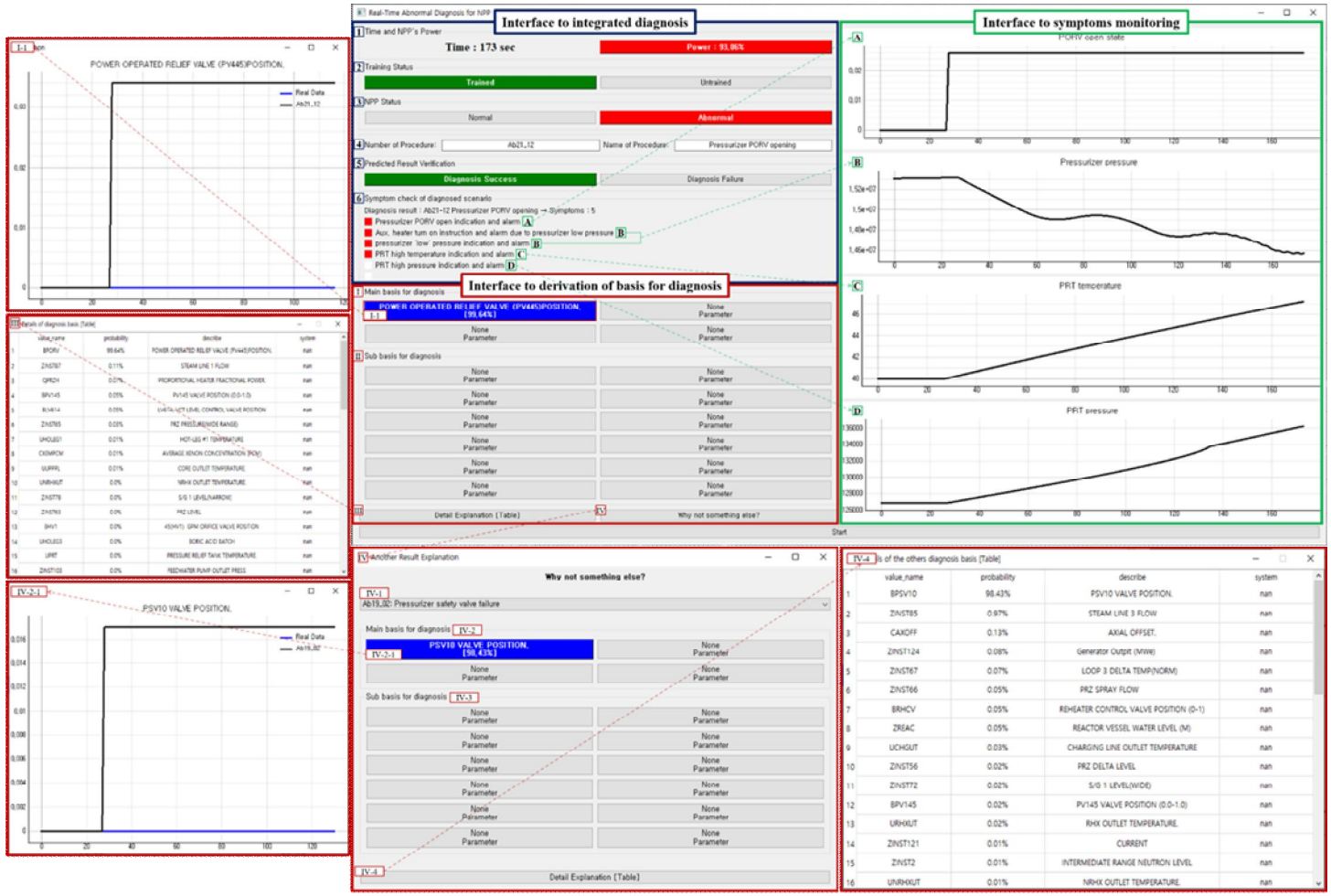


그림 21. 설명 인터페이스 전체 구성

제 2 절 통합 진단 인터페이스

설명 인터페이스의 통합 진단 인터페이스에서는 그림 22와 같이 크게 6개의 기능으로 구분하여 시각화된다. 또한, 각 기능은 통합 비정상 진단 알고리즘의 입출력 관계와 연계되어 있다. 통합 진단 인터페이스의 기능은 그림 22에 표시된 숫자를 기준으로 다음과 같이 설명된다.

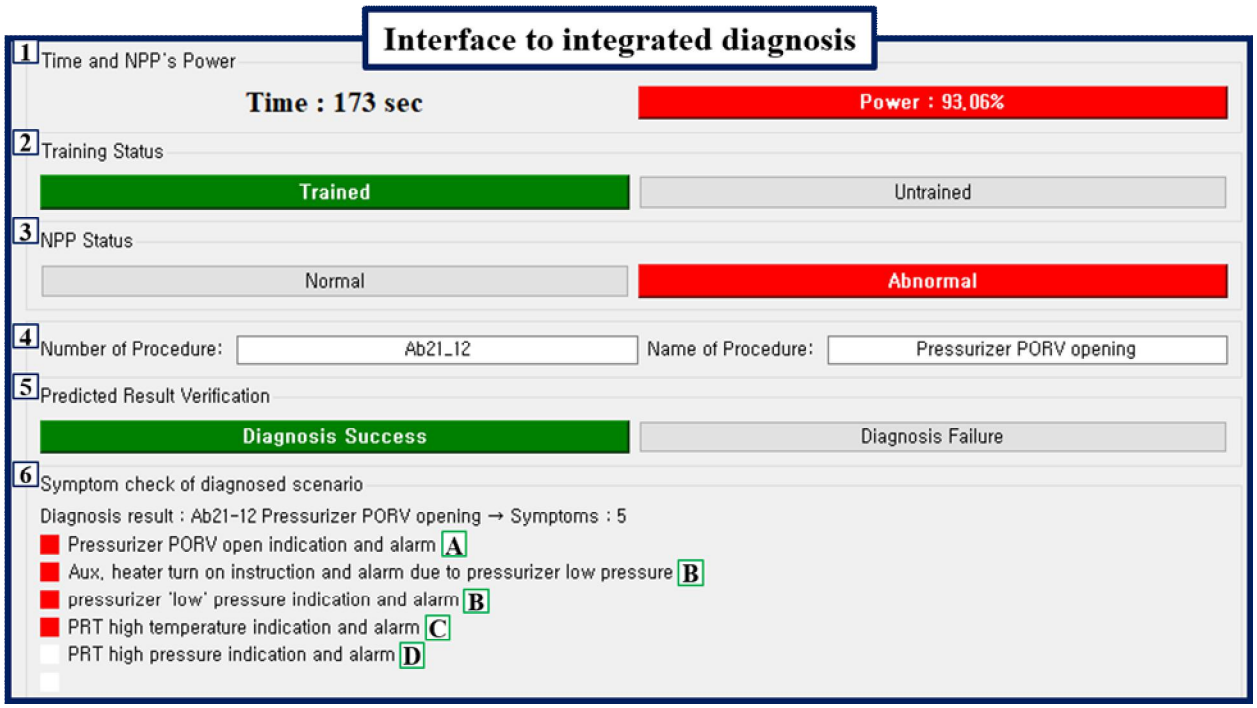


그림 22. 통합 진단 인터페이스 개요

1번 기능은 입력되는 원자력 발전소 데이터의 운전 시간 및 출력에 대해서 시각화를 수행한다. 이때, 출력 변수가 95% 미만일 경우 원자력 발전소의 상태의 불안정을 표현하기 위해서 빨간색으로 표시되도록 알람 기능을 구현하였다. 1번 기능의 실제 적용 결과로는 현재 발전소의 시점은 173초이며, 출력은 93.06%로 알람 설정치인 95% 미만의 값을 가지므로 빨간색의 알람이 활성화되었음을 확인할 수 있다.

2번 기능은 진단 모듈의 훈련 여부 진단 기능에 대한 출력 값을 시각화하였다. 훈련 여부 진단 모듈의 출력 값은 훈련된 상태와 훈련되지 않은 상태의 이진 분류 형태의

결과 값으로 생성되기 때문에 Trained, Untrained의 버튼 2개로 구현하였다. 훈련된 상태(Trained)의 경우 다음 단계의 기능을 구현할 조건이 만족되므로 긍정을 표현하기 위해서 초록색으로 알람 기능을 구현하였다. 훈련되지 않은 상태(Untrained)의 경우에는 인공지능을 활용한 진단이 제한되기 때문에 부정을 표현하기 위한 빨간색으로 알람 기능을 구현하였다. 구체적인 구현 방식은 그림 10, 11의 문턱 값을 기준으로 재구성 오류가 아래에 위치할 경우 Trained 버튼이 초록색으로 활성화되며, 위에 위치할 경우 Untrained 버튼이 빨간색으로 활성화되도록 설계하였다. 2번 기능의 실제 적용 결과로는 입력된 데이터가 훈련된 상태로 진단되었으며, 이에 Trained 버튼이 초록색으로 활성화되었음을 확인할 수 있다.

3번 기능은 진단 모듈의 비정상 시나리오 진단 기능에 대한 출력 값을 시각화하였다. 비정상 시나리오 진단 기능의 출력 값은 정상 시나리오와 15개의 비정상 시나리오로 분류 가능하며 큰 범주에서 이진 분류 형태의 결과 값(정상 시나리오 또는 비정상 시나리오)으로 출력될 수 있기 때문에 Normal, Abnormal의 버튼 2개로 구현하였다. 정상 상태(Normal)의 경우 원자력 발전소의 안정되어 있기에 초록색으로 알람 기능을 구현하였으며, 비정상 상태(Abnormal)의 경우 불안정한 상태를 나타내기 때문에 빨간색으로 알람 기능을 구현하였다. 3번 기능의 실제 적용 결과로는 입력된 데이터가 비정상 상태로 진단되었기에 Abnormal 버튼이 빨간색으로 활성화되었음을 확인할 수 있다.

4번 기능은 진단 모듈의 비정상 시나리오 진단 기능에 대한 출력 값을 시각화하였다. 앞선 2번과 3번 기능은 이진 분류의 형태로 결과가 출력되어 2개의 버튼을 구현하여 알람 기능을 추가하였지만, 비정상 시나리오 진단 기능의 출력 값이 16가지의 비정상 시나리오로 출력되기 때문에 각각의 버튼을 구현하여 알람 기능을 추가하기에는 인터페이스의 복잡성이 증가하기 때문에 단순히 진단된 비정상 시나리오 번호와 이름을 표시하도록 설계하였다. 구체적인 구현 방식은 그림 12에서 가장 높은 확률을 갖는 비정상 시나리오의 번호와 이름을 출력한다. 4번 기능의 실제 적용 결과로는 입력된 데이터가 Ab21_12의 번호로 할당된 가압기 PORV 열림 시나리오로 진단되었음을 확인할 수 있다.

5번 기능은 검증 모듈의 진단 결과 검증 기능에 대한 출력 값을 시각화하였다. 진단 결과 검증 기능의 출력 값은 진단 성공과 진단 실패의 이진 분류 형태의 결과 값으로

출력되기에 Diagnosis Success, Diagnosis Failure의 버튼 2개로 구현하였다. 진단 성공(Diagnosis Success)의 경우 비정상 시나리오의 진단이 성공했음을 알리는 긍정적인 신호이기에 초록색으로 알람 기능을 구현하였다. 진단 실패(Diagnosis Failure)의 경우 비정상 시나리오의 진단이 실패했으며 이는 원전 통합 비정상 진단 시스템의 시스템과 직결되는 부정적인 사항이기에 빨간색으로 알람 기능을 구현하였다. 구체적인 구현 방식은 4번 기능의 출력 값에 대응하여 인공지능 모델이 선택되며, 그림 13, 14의 문턱 값을 기준으로 재구성 오류가 아래에 위치할 경우 Diagnosis Success 버튼이 초록색으로 활성화되며, 위에 위치할 경우 Diagnosis Failure 버튼이 빨간색으로 활성화되도록 설계하였다. 5번 기능의 실제 적용 결과로는 입력된 데이터가 진단 성공으로 진단되었기 때문에 Diagnosis Success 버튼이 초록색으로 활성화되었음을 확인할 수 있다.

마지막 6번 기능은 검증 모듈의 예상 증상 만족 평가 기능에 대한 출력 값을 시각화하였다. 예상 증상 만족 평가 기능의 출력 값은 증상 만족과 증상 불만족이며, 이는 각 조건에 대한 결과로써 출력된다. 앞선 3번 기능 출력 값인 비정상 시나리오에 대응하여 조건이 목록화되며, 각 조건이 만족했을 경우에는 빨간색으로 알림 기능을 구현하였으며 불만족일 경우에는 하얀색으로 알람 기능을 구현하였다. 6번 기능의 실제 적용 결과 전체 조건 5가지가 목록화되었으며, 이 중 4개가 증상 만족으로 빨간색으로 활성화되었음을 확인할 수 있다.

제 3 절 증상 감시 인터페이스

설명 인터페이스의 증상 감시 인터페이스에서는 그림 23과 같이 각 변수에 대해 그래프로 도식화하는 작업을 수행한다. 증상 감시 인터페이스는 통합 진단 인터페이스의 6번 기능과 연계되어있다. 6번 기능을 통해 목록화되는 각 증상들 중 주요 변수를 시간에 따라 모니터링 하는 기능을 수행한다. 이는 운전원에게 진단된 비정상 시나리오의 예상 증상과 부합하는지 확인하기 위해서 제공한다. 6번 기능에서는 가압기 PORV 개방 상태, 가압기 저 압력에 따른 보조전열기 가동 상태, 가압기 저 압력 경보 발생 여부, PRT 고온 및 고압 상태에 대한 증상이 목록화되었다. 목록화된 증상을 증상 감시 인터페이스에서는 시간에 따라 실시간으로 가압기 PORV 개방 상태, 가압기 압력, PRT 온도 및 압력을 그래프로 도식화한다.

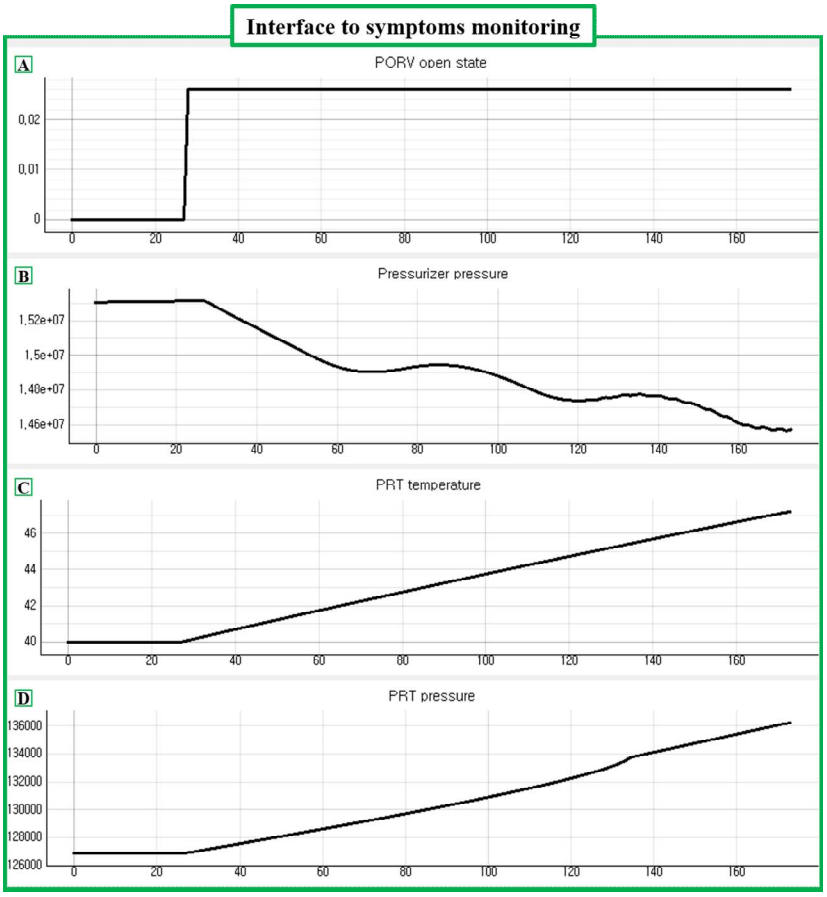


그림 23. 증상 감시 인터페이스 개요

제 4 절 진단 근거 도출 인터페이스

설명 인터페이스의 진단 근거 도출 인터페이스에서는 그림 24와 같이 크게 4개의 기능으로 구현하였다. 진단 근거 도출 인터페이스는 검증 모듈의 진단 근거 도출 기능의 출력 값을 활용하여 시각화하였다. 기존 도식화한 그림 19의 결과는 진단된 비정상 시나리오의 선정 근거를 기여도(Impact factor)를 기반으로 입력 변수를 제시하였다. 하지만, DARPA[13]에서는 왜 해당 시나리오가 선택되었는지 뿐만 아니라 다른 시나리오가 선택되지 않은 이유에 대해서도 제시하고자 하였다. 이에 해당 인터페이스에는 진단된 이유뿐만 아니라 진단되지 않은 이유에 대해서도 제시하고자 한다.

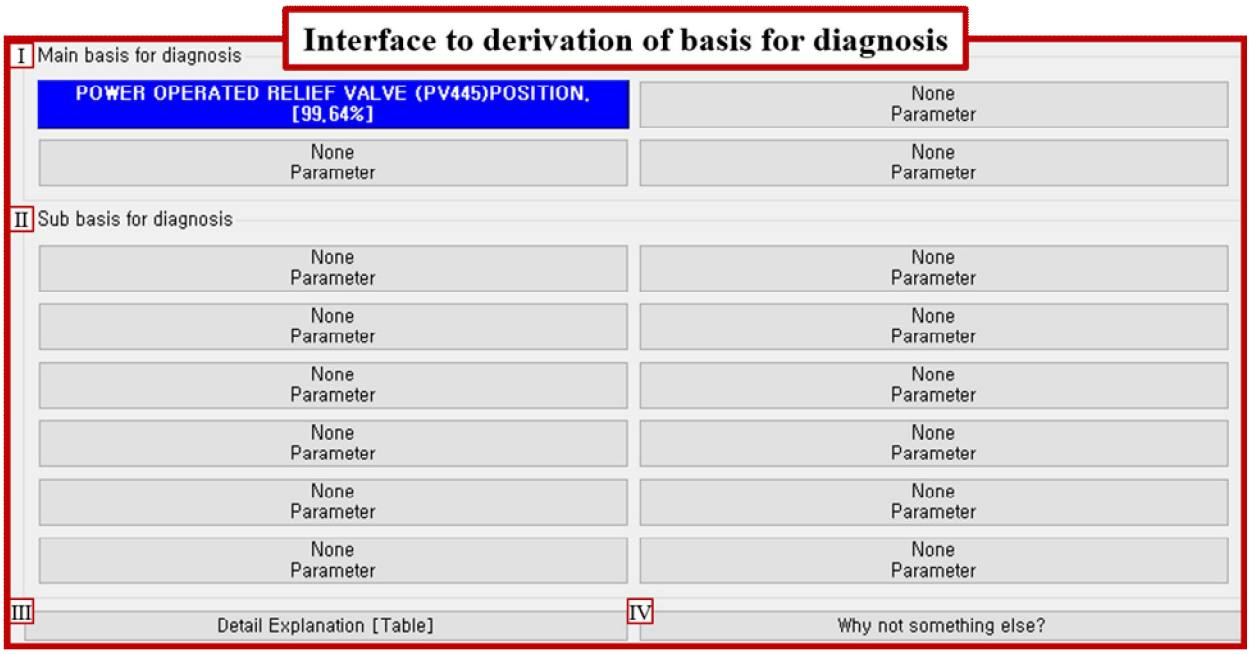


그림 24. 진단 근거 도출 인터페이스 개요

I 번 기능은 통합 진단 인터페이스에서 진단된 비정상 시나리오를 입력 값으로 받아 진단된 근거에 대해서 입력 변수를 기반으로 기여도가 높은 순서로 시각화를 수행한다. 이때, 기여도가 높은 주요 근거만을 시각화하며 주요 근거는 해당 진단에 기여한 기여도가 10%이상인 경우를 말한다. 또한, 기여도가 10% 이상인 경우 중요도를 표현하기 위해서 파란색으로 알람 기능을 구현하였다.

앞선 비정상 시나리오 진단 모듈에서 가압기 PORV 열림 시나리오가 진단되었으며, 진단 근거 도출 인터페이스에서는 PORV의 개방 상태가 99.64%의 기여도로 주요 진단 근거임을 확인할 수 있다.

Ⅰ번 기능에서 기여도가 10% 이상인 경우만을 대상으로 시각화를 하였지만, Ⅱ번 기능에서는 입력 변수의 기여도가 1% 이상부터 10% 미만까지의 경우를 대상으로 시각화를 수행하였다. 이때, Ⅰ번 기능에 비해 중요도가 떨어지므로 별도의 알람 기능은 구현하지 않았다. 실제 적용된 결과에는 설정된 범위 안의 기여도를 가진 입력 변수가 존재하지 않기 때문에 None Parameter로 출력됨을 확인할 수 있다. 100%를 기준으로 Ⅰ번 기능의 PORV 개방 상태가 99.64%의 기여도를 가지고 있기에 나머지는 0.36%를 나누어 가지게 된다. 하지만, 1% 미만의 기여도를 갖는 입력 변수는 인터페이스에 표현되지 않기에 이를 3번 기능을 통해 보여주고자 한다.

Ⅰ번과 Ⅱ번 기능에서 표시된 변수의 이름 버튼을 클릭할 경우, 그림 25와 같이 정상 상태 데이터와 입력된 데이터를 비교하는 그래프를 보여준다. 이를 통해 비정상 상태로 판단된 데이터를 정상 상태에서의 데이터와 비교함으로써 진단된 이유를 보다 쉽게 판단할 수 있을 것으로 기대된다.

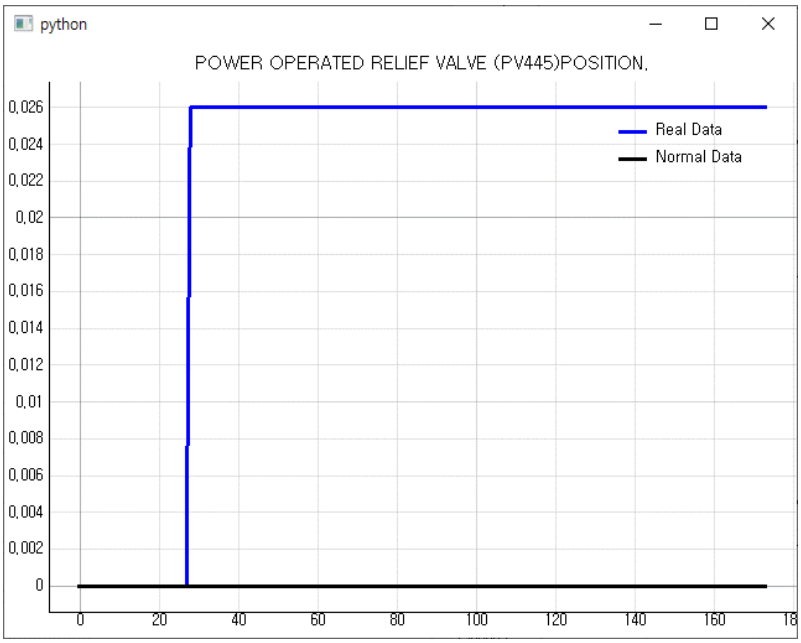
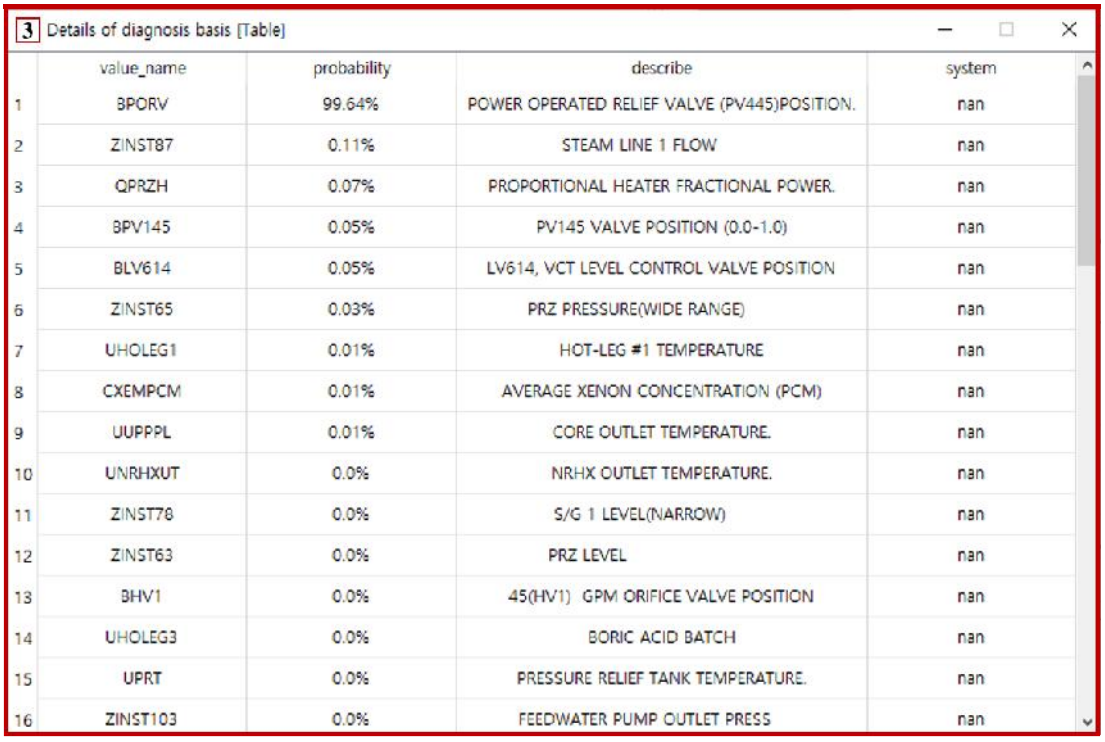


그림 25. 진단 근거 도출 인터페이스 – 주요 근거 변수 비교 (Ⅰ, Ⅱ번 기능)

그림 26은 99.64%의 기여도를 갖는 PORV 개방 상태를 나타내는 변수를 클릭했을 때 나오는 팝업 기능으로, 정상 데이터는 밸브 상태가 0으로 닫혀있지만 입력되는 데이터는 0.026으로 열려있는 것을 확인할 수 있다. 이때, 밸브 상태는 0에서 1의 값으로 0.026은 100% 기준으로 2.6% 열려있다는 의미이다. 이를 통해, 정상 상태에서는 닫혀있는 밸브 상태이지만, 비정상 시나리오가 발생함에 따라 해당 밸브가 열려있음을 운전원이 확인할 수 있다. 실제 가압기 PORV 열림 시나리오는 가압기 PORV가 개방됨에 따라 가압기 압력이 감소하는 등의 특징을 갖는 비정상 시나리오이다.

Ⅲ번 기능은 1번과 2번 기능에서 보여주는 1% 이상의 기여도를 갖는 입력 변수만이 아닌 전체 입력 변수의 기여도를 대상으로 표로 정리하여 해당 버튼을 누를 경우 팝업창으로 정보를 제공하고자 설계하였다. 구현된 표는 그림 27과 같으며 해당 표에는 변수 이름, 기여도, 변수 설명 등이 포함되어있으며, 기여도를 기준으로 오름차순하여 기여도가 높은 순서대로 확인할 수 있다.


3 Details of diagnosis basis [Table]

	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

그림 26. 진단 근거 도출 인터페이스 – 전체 기여 변수 표 (Ⅲ번 기능)

앞선 기능들 모두 진단된 비정상 시나리오를 대상으로 기여도를 평가하여 선택 근거

를 도출하지만, 마지막 Ⅳ번 기능에서는 그림 5의 DARPA 프로젝트 내용 중 'Why not something else?'를 차용하여 선택되지 않은 비정상 시나리오에 대해서 왜 선택되지 않았는지 대한 근거를 보여준다[13]. Ⅳ번 기능을 클릭할 경우 Ⅲ번 기능과 동일하게 팝업창이 출력되며, 그림 27은 Ⅳ번 기능에 해당하는 팝업창을 보여준다. 해당 팝업창은 설명 인터페이스의 진단 근거 도출 인터페이스와 유사한 디자인으로 설계하였으며, 기능 또한 동일하게 작용한다. 한 가지 차이점은 진단되지 않은 근거를 제시하기 위해서 진단되지 않은 비정상 시나리오를 선택하여 결과를 출력하기 위해서 시나리오를 고를 수 있는 칸이 존재한다는 것이다. 그림 28은 Ⅳ번 기능의 팝업창에서 Detail Explanation [Table] 버튼을 클릭할 경우 나오는 표이며, Ⅲ번 기능과 동일한 구현 방식이 적용된다. 그림 27은 비정상 시나리오 진단 모듈에서 진단된 가압기 PORV 열림 시나리오가 아닌 가압기 안전밸브 고장 시나리오가 왜 선택되지 않았는지에 대해서 보여준다. 주요 근거로는 PSV10 밸브의 개방 상태가 98.43%의 기여도를 갖는 것으로 제시되었으며, 버튼을 클릭할 경우 그림 29를 보여준다. 그림 29는 현재 입력되는 데이터는 값이 0으로 밸브가 닫혀있는 상태에 비해 가압기 안전밸브 고장 시나리오의 데이터 값은 0.017로 100% 기준으로 1.7% 열려있는 상태임을 확인할 수 있다. 이를 통해 가압기 PORV 열림 시나리오는 PSV10 밸브가 개방되지 않았기 때문에 선택되지 않았음을 확인할 수 있다.

가압기 PORV 열림 시나리오와 가압기 안전밸브 고장 시나리오는 가압기에 위치한 밸브가 원인을 모르는 채 갑자기 열리는 시나리오로 가압기 압력, 수위, 온도 등에 영향을 미친다. 하지만, 설명 가능한 인공지능을 통해 도출된 결과로부터 구체적인 밸브의 이름 등의 장치(Component)까지도 확인할 수 있다.

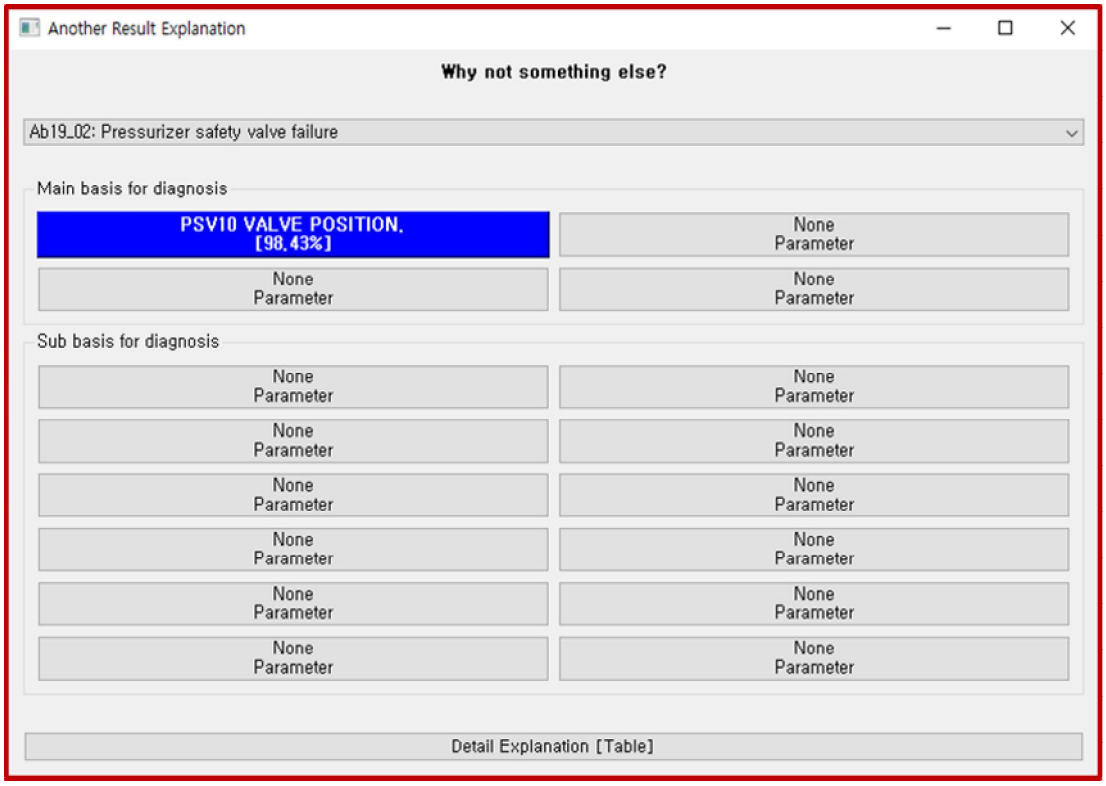
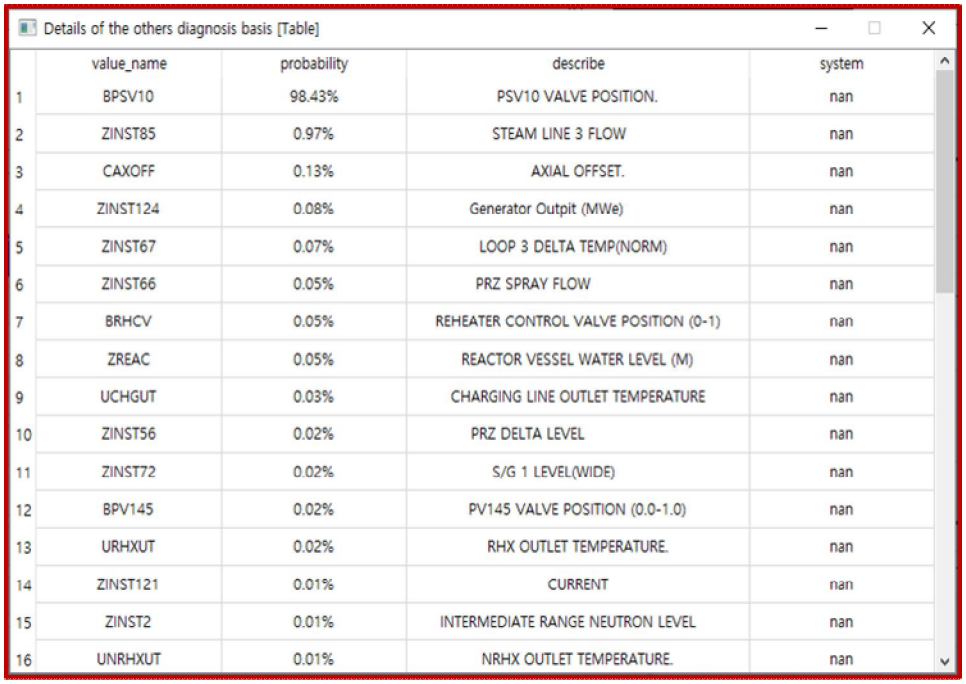


그림 27. 진단 근거 도출 인터페이스 - 진단되지 않은 이유 (Ⅳ번 기능)



	value_name	probability	describe	system
1	BPSV10	98.43%	PSV10 VALVE POSITION.	nan
2	ZINST85	0.97%	STEAM LINE 3 FLOW	nan
3	CAXOFF	0.13%	AXIAL OFFSET.	nan
4	ZINST124	0.08%	Generator Outpit (MWe)	nan
5	ZINST67	0.07%	LOOP 3 DELTA TEMP(NORM)	nan
6	ZINST66	0.05%	PRZ SPRAY FLOW	nan
7	BRHCV	0.05%	REHEATER CONTROL VALVE POSITION (0-1)	nan
8	ZREAC	0.05%	REACTOR VESSEL WATER LEVEL (M)	nan
9	UCHGUT	0.03%	CHARGING LINE OUTLET TEMPERATURE	nan
10	ZINST56	0.02%	PRZ DELTA LEVEL	nan
11	ZINST72	0.02%	S/G 1 LEVEL(WIDE)	nan
12	BPV145	0.02%	PV145 VALVE POSITION (0.0-1.0)	nan
13	URHXUT	0.02%	RHX OUTLET TEMPERATURE.	nan
14	ZINST121	0.01%	CURRENT	nan
15	ZINST2	0.01%	INTERMEDIATE RANGE NEUTRON LEVEL	nan
16	UNRHXUT	0.01%	NRHX OUTLET TEMPERATURE.	nan

그림 28. 진단되지 않은 이유 (Ⅳ번 기능) - 전체 기여 변수 표

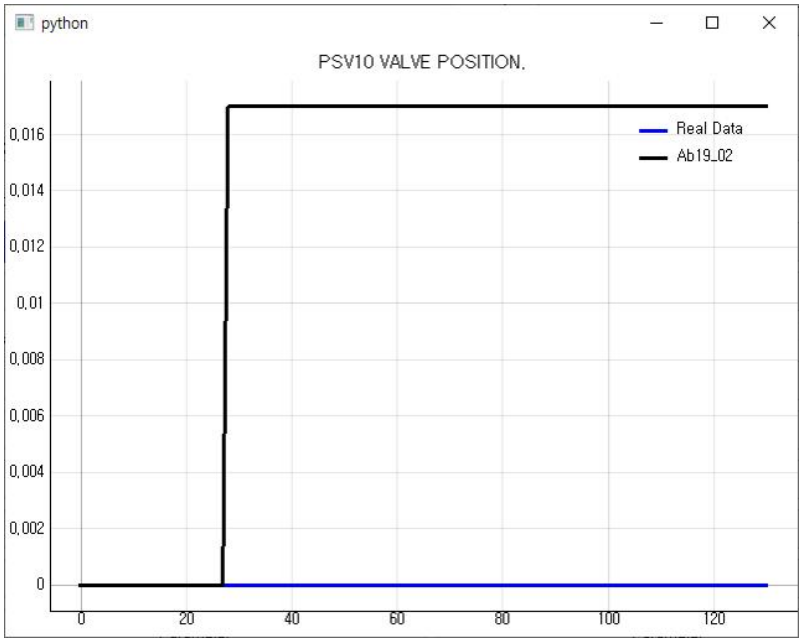


그림 29. 진단되지 않은 이유 팝업창 (Ⅳ번 기능) - 주요 근거 변수 비교

제 6 장 결 론

본 논문에서 소개한 원전 통합 비정상 진단 시스템은 원자력 발전소에서 비정상 사고가 발생할 경우, 인공지능을 활용한 사고 진단 및 설명 가능한 인공지능을 통한 진단 근거 도출을 통해 운전원의 의사결정을 지원하고자 개발하였다. 원전 통합 비정상 진단 시스템 개발을 위해 Compact Nuclear Simulator를 활용한 데이터 수집, 사용된 방법론인 인공지능, 설명 가능한 인공지능, 규칙 기반 시스템의 학습 및 구현을 위한 데이터 전처리, 명확하고 효율적인 비정상 진단을 위한 통합 비정상 진단 알고리즘 설계, 직관적인 정보 제공을 위한 설명 인터페이스 설계를 수행하였다.

원전 통합 비정상 진단 시스템을 통해 운전원은 실시간으로 원자력 발전소 상태에 대한 진단 정보를 제공받으며, 구체적으로 인공지능을 활용한 해당 데이터의 훈련 여부, 비정상 여부, 비정상 시나리오 진단 결과, 진단 결과 검증, 규칙 기반 시스템을 활용한 예상 증상 만족 평가, 설명 가능한 인공지능을 활용한 진단 근거 도출 결과를 설계한 GUI(Graphical User Interface)인 설명 인터페이스를 통해 제공받는다. 설명 인터페이스는 통합 진단 인터페이스, 증상 감시 인터페이스, 진단 근거 도출 인터페이스로 세분화되며, 효과적인 정보 제공을 위해 알람 기능을 제공한다.

원전 통합 비정상 진단 시스템은 원자력 발전소에서 비정상 사고가 발생한 경우, 한정된 시간 내에 정확하게 사고 진단을 수행하는 운전원의 심리적 부담 및 인적 오류를 줄이고자 한다. 하지만, 인공지능을 활용한 사고 진단 결과는 100% 정확하지 않으며 잘못된 진단 결과를 운전원이 활용하여 완화 조치를 수행한다면 이에 대해 운전원은 책임을 져야 한다. 이에, 단순히 인공지능을 활용하여 제공하는 진단 결과만이 아닌 설명 가능한 인공지능을 통해 도출된 진단 근거를 제공함으로써 운전원이 인공지능으로부터 출력된 결과에 대한 신뢰성을 제고하고자 한다.

참고문헌

[1] "원전사고·고장현황." OPIS원전안전운영정보시스템. 2020년 10월 5일 접속, https://opis.kins.re.kr/opis.

[2] W. D. Jung, J. K. Park, J. W. Kim, J. J. Ha, Analysis of an Operators' Performance Time and Its Application to a Human Reliability Analysis in Nuclear Power Plants, IEEE transactions on nuclear science 54(5), 2007, pp. 1801-1811.

[3] J. M. Kim, G. Lee, C. Lee, S. J. Lee, Abnormality diagnosis model for nuclear power plants using two-stage gated recurrent units, Nucl. Eng. Technol. 52(9), 2020, pp. 2009-2016.

[4] J. M. Yang, J. H. Kim, Accident diagnosis algorithm with untrained accident identification during power-increasing operation, Reliability Engineering & System Safety 202, 2020, p. 107032.

[5] H. J. Kim, J. H. Kim, Algorithm of Abnormal Event Diagnosis with the Identification of Unknown Events and Output Confirmation, Transactions of the Korean Nuclear Society Virtual Spring Meeting July 9-10, 2020.

[6] K. H. Yoo, J. H. Back, M. G. Na, S. Hur, H. M. Kim, Smart support system for diagnosing severe accidents in nuclear power plants. Nucl. Eng. Technol. 50(4), 2018, pp. 562-569.

[7] 김동우, 한혜영. 2019 학술정보 글로벌 동향 Vol. 10. 한국교육학술정보원, 2019.

[8] Francois Chollet. 케라스 창시자에게 배우는 딥러닝. 길벗, 2018.

[9] 솔라리스. 텐서플로로 배우는 딥러닝. 영진닷컴, 2018.

[10] Ali H. Mirza, Selin Cosan, Computer network intrusion detection using sequential LSTM neural networks autoencoders, In: 2018 26th signal processing and communications applications conference (SIU), IEEE, 2018, pp. 1-4.

[11] Guolin Ke et al, LightGBM: A highly efficient gradient boosting decision tree, In: Advances in neural information processing systems, 2017, pp.

3146-3154.

[12] 이상원. 사용자 중심의 User-centered 설명가능한 인공지능 Explainable Artificial Intelligence 설명 인터페이스 Explanation Interface, 한국경영과학회 학술대회논문집 , 2019, pp. 1983-1999.

[13] D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).

[14] Scott M. Lundberg, Gabriel G. Erion, Su-In Lee, Consistent individualized feature attribution for tree ensembles, arXiv preprint arXiv:1802.03888, 2018.

[15] 안재현. XAI 설명 가능한 인공지능, 인공지능을 해부하다. 위키북스, 2020.

[16] Scott M. Lundberg, et. al, From Local Explanations to Global Understanding with Explainable AI for Trees, Nature machine intelligence 2(1), 2020, pp. 2522-5839.