XAI (Explainable AI)를 이용한 원전 사고진단 기술

나 만 균

조 선 대 학 교

Transactions of the Korean Nuclear Society Spring Meeting
Jeju, Korea, May 18, 2022

CHOSUN UNIVERSITY

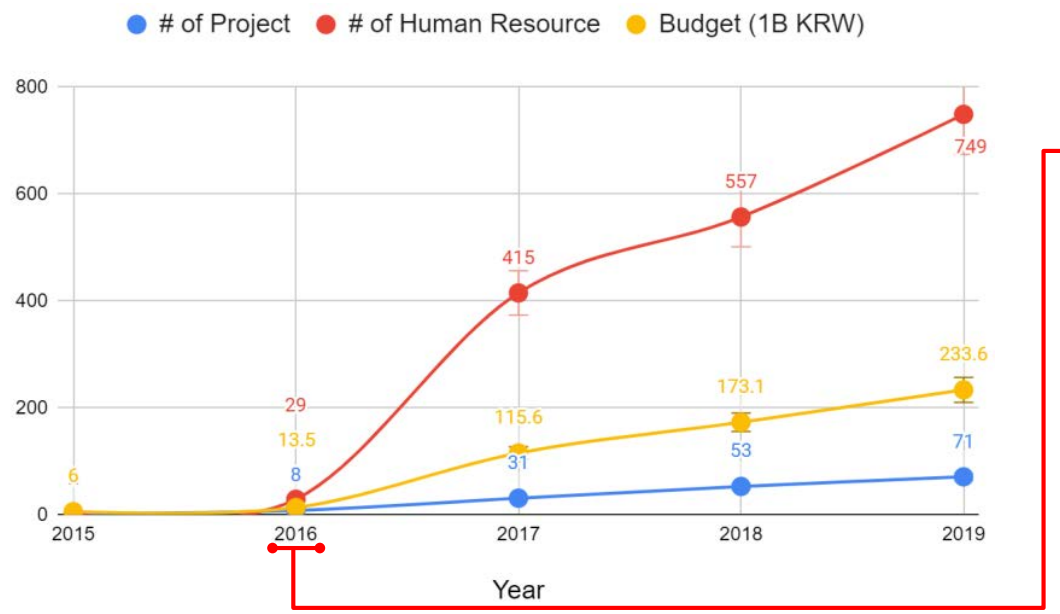
Contents

01 설명가능 인공지능 기술의 필요성

01 설명가능 인공지능 기술의 필요성

▪ 원자력 분야 인공지능 관련 연구 현황

- **인공지능 기술을 활용한 원자력 분야 연구 개발이 활발하게 이루어지고 있음.**
  - 2015년부터 2019년까지 원자력 분야 인공지능 관련 NTIS 등록 연구과제의 수, 투입된 인력과 예산 모두 증가하고 있음.



<2015~2019년 원자력 분야 인공지능 관련 NTIS 등록 연구과제 추이>

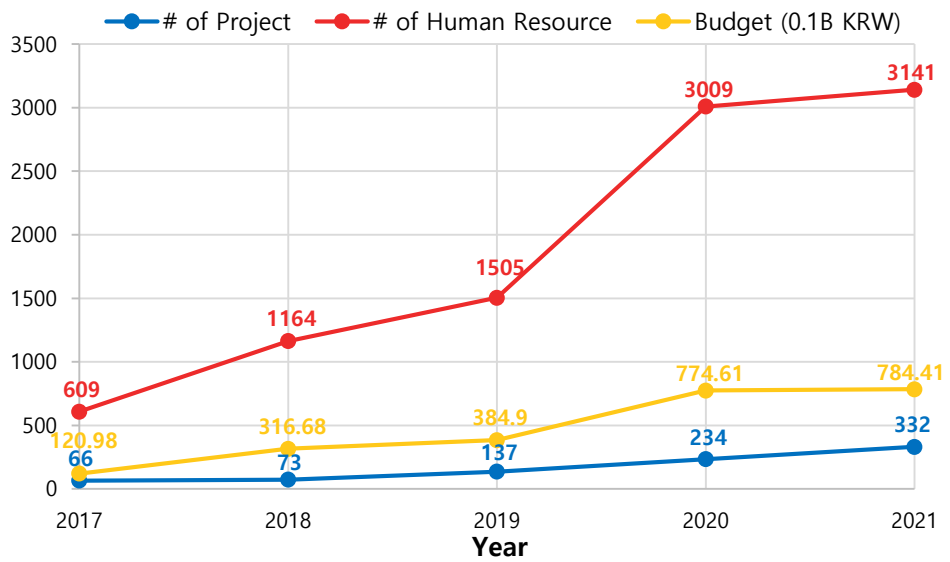
- 2016년 이세돌과 알파고의 대국 이후, 국내에서 딥러닝이 큰 이슈가 되며 2017년 기준 인공지능 관련 과제 예산이 전년 대비 87% 증가하였음.
- 참여연구원 수 또한 2017년 비약적으로 증가하였으며, 관련 인력양성 사업 과제 등의 영향으로 2020년 이후 보다 많은 전문인력이 수급될 것으로 추정됨.

참고문헌: 빅데이터 분석을 통한 원자력 분야 인공지능 관련 연구동향 분석, 원자력연구원 기술현황분석보고서

01 설명가능 인공지능 기술의 필요성

▪ 원자력 분야 설명가능 인공지능 관련 연구 현황

- **설명가능 인공지능 기술**을 활용한 NTIS 등록 연구과제, 인력 및 예산은 연도별로 점차 증가하고 있지만, **원자력 분야 관련 연구는 전무한 실정임**.



<2017~2021년 **전 분야** 설명가능 인공지능 관련 NTIS 등록 연구과제 추이>

- 설명가능 인공지능 연구는 주로 정보/통신, 보건의료, 전기/전자 등 분야에서 수행하고 있으며, 해당 분야는 전체 분야에서 각각 61%, 16%, 6%를 차지하고 있음.
- 그러나, **원자력 분야**의 경우 2017년부터 2021년까지 총 연구과제 수가 4개로 0.5%밖에 되지 않으며, 주로 **방사선치료 및 폐암 진단 등 의학분야와 접목한 연구임**.

참고문헌: NTIS 홈페이지 (www.ntis.go.kr)

01 설명가능 인공지능 기술의 필요성

▪ 인공지능 관련 정책

- 인공지능이 중요작업(Mission Critical)에 사용될 경우 **인공지능의 설명성, 투명성 확보 기술, 기준 정립이 필요함**. 2018년 의사 결정 이유에 대한 설명을 요구하는 EU의 개인정보보호법(GDPR)이 발효되어 의사 결정 이유를 설명할 수 없는 인공지능 기술은 향후 의료, 군사 등 중요작업에는 제한될 것으로 예상됨.

<EU의 일반정보보호규정>

항목	내용
잊혀질 권리	제17조: 정보 주체가 본인의 개인정보 처리를 더 이상 원치 않거나 개인정보를 보유할 법적 근거가 없으면 해당 정보 삭제
자동화된 의사결정 제한	제22조: 자동화된 처리(프로파일링 포함)에만 근거한 결정의 대상이 되지 않을 권리
설명을 요구할 권리	제13-14조: 알고리즘에 의해 행해진 결정에 대해 질문하고, 결정에 관여한 논리에 대해 의미 있는 설명을 요구할 권리
EU 집행력	규정 위반시 해당 기업의 전세계 매출의 최대 4%까지 벌금 부과
발효	2018년 5월 28일

01 설명가능 인공지능 기술의 필요성

▪ 인공지능 관련 정책

- 과학기술정보통신부는 설명가능 인공지능 기술의 중요성을 인식하고 'I-Korea 4.0 실현을 위한 인공지능 R&D 전략(2018.05)'의 R&D 로드맵에 2025년까지 **설명가능 학습 · 추론 기술 개발 추진 계획을 포함**하였음.

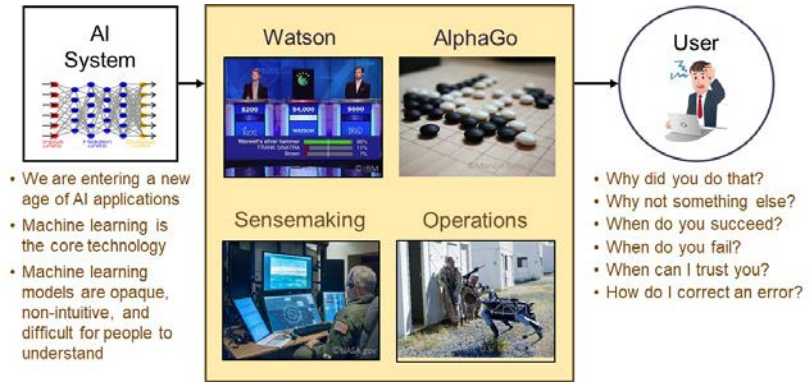
<AI R&D 로드맵>

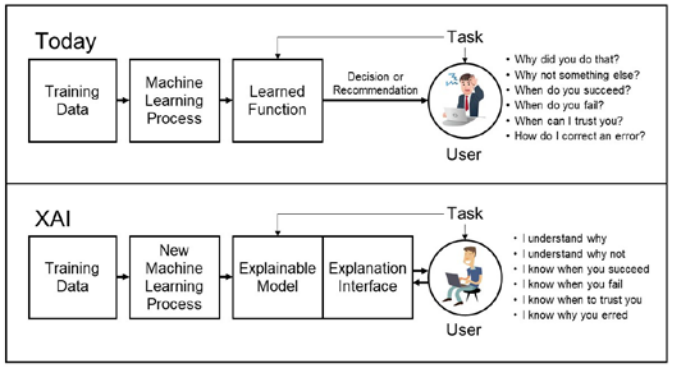
참고문헌: (I-Korea 4.0 실현을 위한) 인공지능(AI) R&D 전략, 과학기술정보통신부, 2018

01 설명가능 인공지능 기술의 필요성

▪ 설명가능 인공지능 기술의 필요성

- 인공지능 기술은 높은 성능에도 불구하고 도출된 결과에 대한 근거가 불분명함.
  - 인공지능 기술의 적용은 시스템의 효율성을 높일 수 있지만, 신뢰성 및 정확성을 확보한 것은 아님.
  - 인공지능 기술의 블랙박스 특성으로 인하여 인공지능 기술의 신뢰성에 대한 의구심이 수반되는 상황임.
- 근거 없이 도출된 인공지능의 결과는 인공지능 기술의 적용성 측면에서도 제한적인 상황임.
  - 다양하고 복잡한 시스템으로 구성된 원전의 경우, 잘못된 결정은 매우 큰 부작용을 초래할 수 있음.
  - 인공지능의 결과에 대한 근거, 도출 과정의 타당성 등 논리적인 설명을 통해 인공지능 기술의 신뢰성 확보가 필요함.





CHOSUN UNIVERSITY

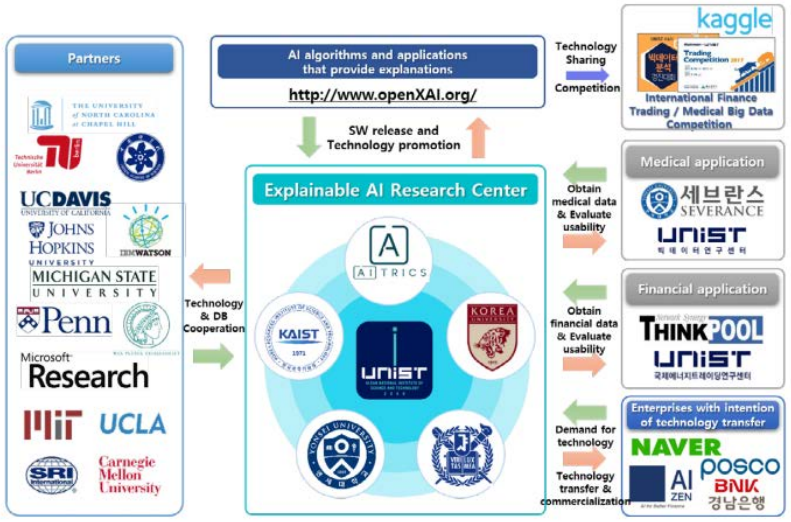
참고문헌: Explainable Artificial Intelligence (XAI), DARPA

02 국내외 설명가능 인공지능 기술 동향

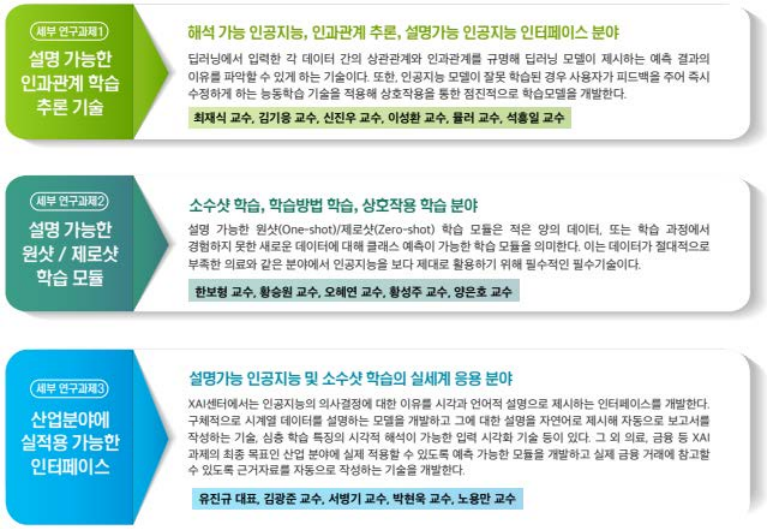
02 국내외 설명가능 인공지능 기술 동향

▪ 국내 설명가능 인공지능 기술 동향

- 설명가능 인공지능 연구센터
  - 한국과학기술원, 울산과학기술원, 서울대학교 등 다양한 학계와 국외기관이 연계하여 금융, 게임, 의료 등 다양한 분야의 설명가능 인공지능 모델 개발에 참여중임.
  - 설명가능 인공지능을 적용한 인공지능 기술을 만들어 사용자가 새로운 세대의 인공지능 시스템을 이해하고 적절하게 신뢰하며 효과적으로 관리할 수 있도록 하는 것을 목표로 함.



<설명가능 인공지능 연구센터 참여 기관>



<설명가능 인공지능 연구센터 주요 연구개발 기술>

<설명가능 인공지능 연구센터 주요 실적>

참고문헌: 설명가능 인공지능 연구센터 뉴스레터 창간호(2020.09)

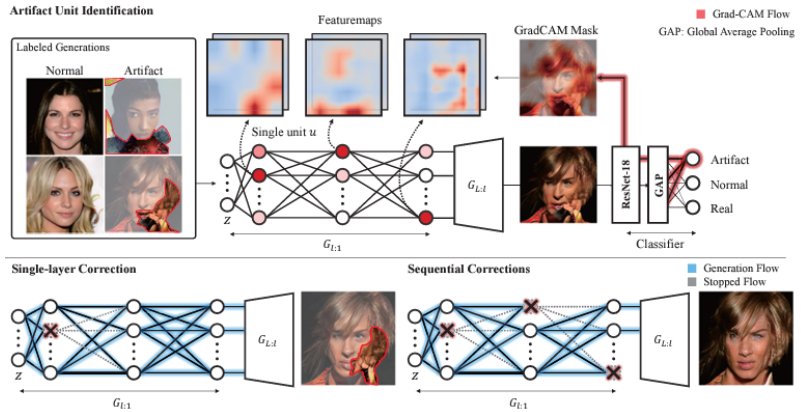
02 국내외 설명가능 인공지능 기술 동향

▪ 국내 설명가능 인공지능 기술 동향

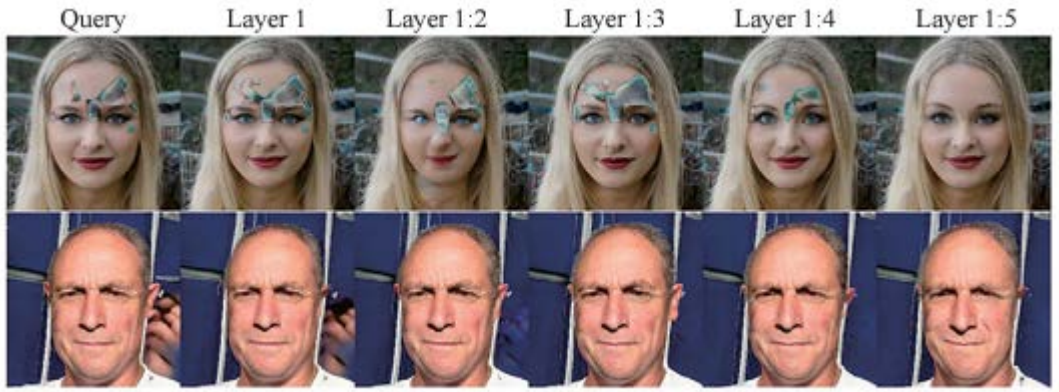
• 설명가능 인공지능 연구센터 – 연구사례 ①

▸ Automatic Correction of Internal Units in Generative Neural Networks

- 딥러닝 기반 이미지 생성모델의 결함 이미지 생성 문제를 해결하기 위해 설명가능 인공지능 기술을 활용함.
- 해당 연구는 (1) 정상 생성과 오류 생성을 구분할 수 있는 분류기를 학습하고, 설명가능 인공지능 기술을 적용하여 오류 영역을 추출하는 과정, (2) 오류 영역과 생성 모델 내부 뉴런 사이의 공유 영역을 계산하여 오류 유발 뉴런 검출, (3) 검출된 뉴런의 계층별 제거를 통한 결함 수리를 수행함.
- 해당 기술은 모델 구조에 대한 의존성이 적어 다양한 딥러닝 기반 생성 모델에 적용 가능함.



<개발 기술 모식도>



<오류 유발 뉴런의 계층별 제거에 따른 결함 수리 결과>

참고문헌: Tousi, Ali, et al. "Automatic correction of internal units in generative neural networks."

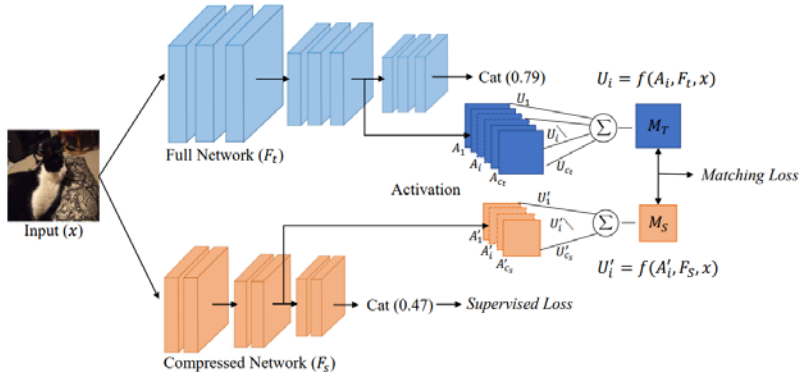
02 국내외 설명가능 인공지능 기술 동향

▪ 국내 설명가능 인공지능 기술 동향

• 설명가능 인공지능 연구센터 – 연구사례 ②

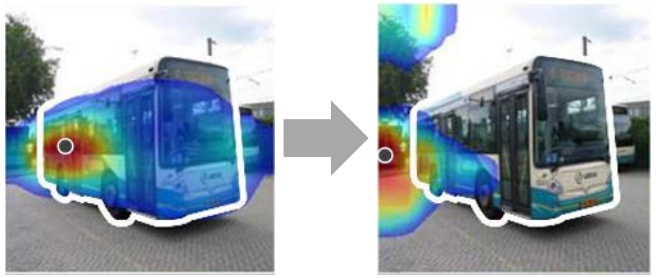
▸ Attribution Preservation in Network Compression for Reliable Network Interpretation

- 인공지능 모델을 소형 반도체 칩에 탑재하여 활용하기 위해 최근 딥러닝 모델의 신경망 크기와 연산량을 줄이는 압축 알고리즘들이 사용되고 있음.
- 기존 압축 알고리즘은 해당 예측에 대한 성능은 보존하나 **예측에 대한 설명은 손실될 수 있음**.
- 해당 연구에서는 기존 네트워크와 압축된 네트워크의 특성을 일치시켜 의사결정 설명을 보존하는 프레임워크를 제시함.
- 결과적으로, 개발한 프레임워크는 **네트워크의 설명을 보존**할 뿐만 아니라 **예측성능의 향상**을 확인함.



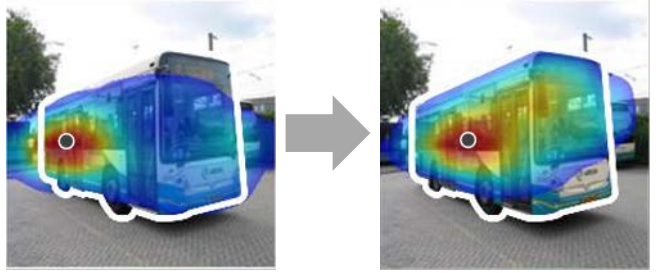
<프레임워크 개요>

<연구 전>



기존 압축 알고리즘 적용 시 **설명 손실**
→ 버스 개체 설명 손실

<연구 후>



제시한 프레임워크 적용 시 **설명 보존**

참고문헌: Park, Geondo, et al. "Attribution Preservation in Network Compression for Reliable Network Interpretation."

02 국내외 설명가능 인공지능 기술 동향

▪ 국내 설명가능 인공지능 기술 동향

- **설명가능 인공지능 연구센터 – 산업 적용 사례**
  - **제조, 의료, 금융, IT 산업 등 다양한 산업에서 설명가능 인공지능 기술을 적용하고 있음.**
    - **제조: 포항제철소(인공지능을 활용한 용광로 제어)에서 설명가능 인공지능 적용을 통해 용광로의 제어 행동의 이유를 설명함.**
    - **의료: 환자의 상태가 호전 또는 악화될 경우, 이에 대한 이유를 설명함.**
    - **IT: Bixby 및 Siri와 같은 지능형 개인비서 학습 시 발생한 오류에 대한 원인을 설명함.**

<의료산업에 설명가능 인공지능 시스템 적용 사례>

<설명가능 인공지능을 이용한 무인 원스톱 시설물 점검시스템(서울시 적용사례)>

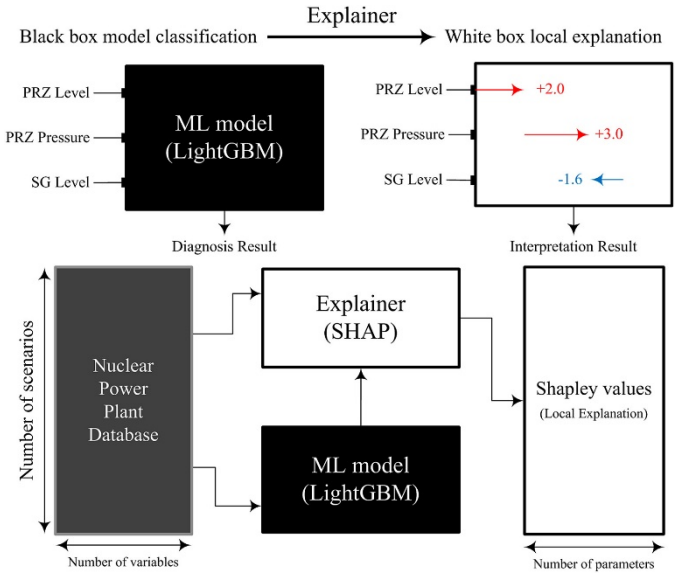
참고문헌: 설명가능 인공지능 연구센터 홈페이지, 미래경제뉴스 발췌

02 국내외 설명가능 인공지능 기술 동향

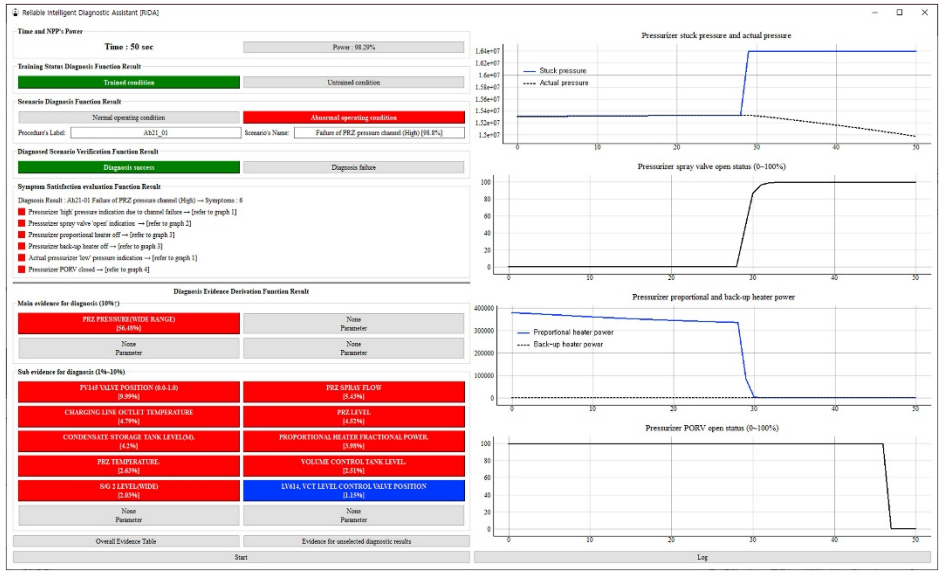
- ❖ LightGBM (Light Gradient Boosting Model)
- ❖ SHAP (SHapley Additive exPlanations)

▪ 국내 설명가능 인공지능 원자력 분야 기술 동향

- **A Reliable Intelligent Diagnostic Assistant for Nuclear Power Plants using Explainable Artificial Intelligence**
  - ▸ **설명가능 인공지능 기술을 활용한 원전 비정상 운전 지원 시스템을 제안함.**
  - ▸ **사용한 인공지능 및 설명가능 인공지능 방법론은 각각 LightGBM과 Tree SHAP임.**
  - ▸ **사용한 데이터는 CNS를 활용하여 수집한 21가지의 시나리오(비정상 20건, 정상 1건)임.**



<설명가능 인공지능 적용 구조>



<가압기 압력채널 고장 '고' 시나리오 적용 결과 >

- ▪ **가압기 압력채널 고장 '고' 시나리오를 98.8%의 정확도로 진단하고 있으며, 진단 근거로 가압기 압력을 56.48%의 기여인자로 제시함.**
- ▪ **진단 결과의 신뢰도 향상을 위해 비정상 절차서의 증상 요건을 규칙 기반 시스템으로 구현하여 만족 여부를 확인함.**

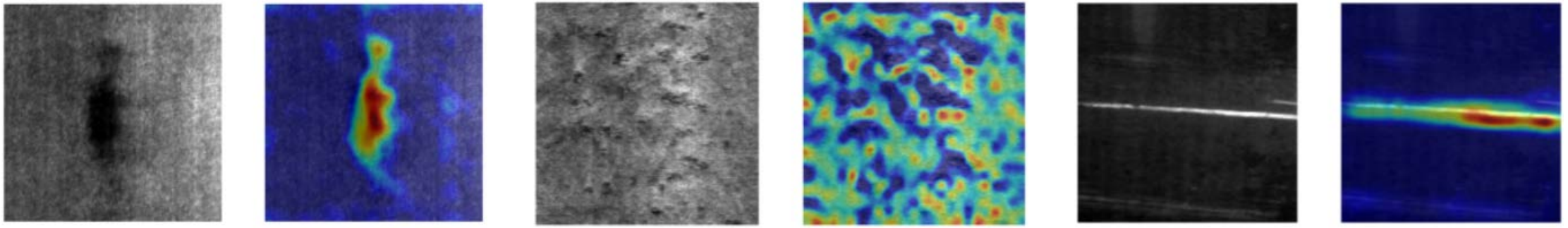
CHOSUN UNIVERSITY

참고문헌: Park, Ji Hun, et al. "A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP."

❖ DCNN (Deep Convolution Neural Network)
❖ CAM (Class Activation Map)

▪ 국내 설명가능 인공지능 원자력 분야 기술 동향

- **Steel Surface Defect Diagnostics Using Deep Convolutional Neural Network and Class Activation Map**
  - 원전의 강철 표면에 대한 이미지 데이터를 활용하여 설명가능 인공지능 기술을 적용함으로써 **강철 표면 결함 진단**을 연구함.
  - 강철 표면 결함 진단 결과의 시각적 의사결정 프로세스를 지원하기 위해 사용한 인공지능 및 설명가능 인공지능 방법론은 각각 **DCNN**과 **CAM**임.
  - 아래 그림은 설명가능 인공지능을 적용한 결과로 강조된 부분이 CAM을 통해 시각적으로 설명함으로써 인간의 의사결정과 인공지능 기술 간의 상호작용이 가능하게 함.

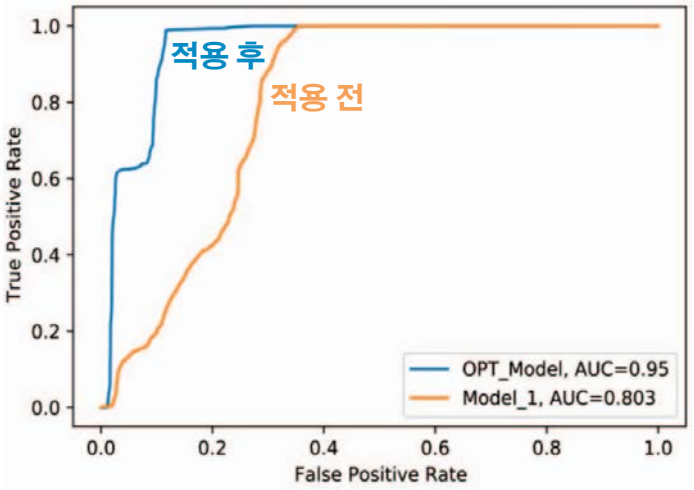


<강철 표면 결함을 대상으로 설명가능 인공지능을 적용한 결과>

참고문헌: Lee, Soo Young, et al. "Steel surface defect diagnostics using deep convolutional neural network and class activation map."

▪ 국외 설명가능 인공지능 기술 동향

- **Using Kernel SHAP XAI Method to Optimize the Network Anomaly Detection Model**
  - **최적화된 네트워크 이상 탐지 모델을 개발하기 위해 설명가능 인공지능 기술을 변수 선정에 활용함.**
  - **해당 연구는 네트워크 공격에 대한 데이터 세트인 CICIDS2017를 활용함.**
  - **사용한 인공지능 및 설명가능 인공지능 방법론은 각각 Autoencoder와 Kernel SHAP임.**
  - **Kernel SHAP을 활용함으로써 네트워크 이상탐지 근거에 대한 설명을 제공하고, 각 변수에 대한 기여도를 산출함으로써 40개의 입력변수를 선정함.**



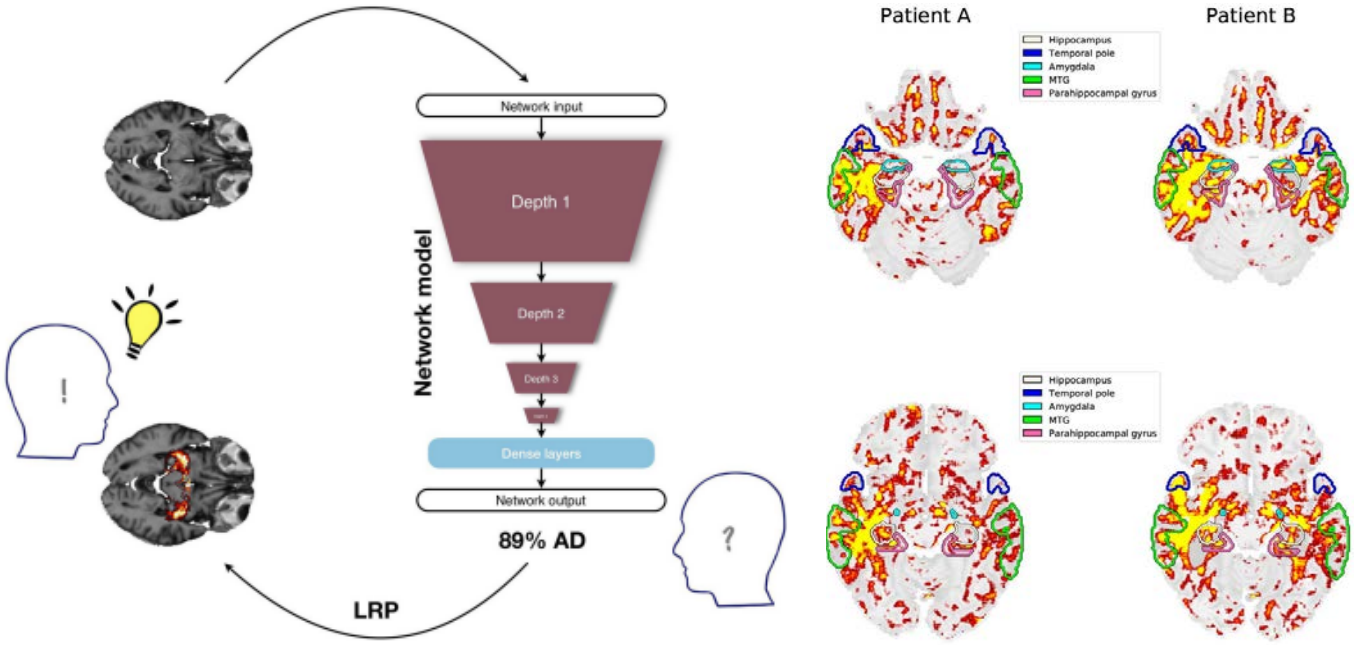
<설명가능 인공지능 적용 여부에 따른 성능평가 결과>

<네트워크 이상탐지 모델에 설명가능 인공지능을 적용한 결과>

참고문헌: Roshan, Khushnaseeb, and Aasim Zafar. "Using Kernel SHAP XAI Method to Optimize the Network Anomaly Detection Model."

▪ 국외 설명가능 인공지능 기술 동향

- **Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification**
  - 의료분야에서 설명가능 인공지능 기술을 활용하여 **알츠하이머 진단 근거를 제공**하는 연구를 진행함.
  - ADNI 데이터 세트를 활용하여 학습한 인공지능 및 설명가능 인공지능 방법론은 **CNN**과 **LRP**임.



<설명가능 인공지능 적용 구조와 적용 결과>

- 설명가능 인공지능 방법론인 LRP를 적용하여 출력된 히트맵은 인공지능 결과에 대한 **신뢰성을 높이고 진단도구로써 큰 잠재력**을 확인할 수 있음.
- 이를 통해 설명가능 인공지능 기술이 사용자에게 설명성을 제공하여 유용한 도구가 될 수 있어 **인공지능 기반 진단에 대한 신뢰를 높일 수 있음**.

참고문헌: Böhle, Moritz, et al. "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification."

03 설명가능 인공지능 기반 원전 사고진단

03 설명가능 인공지능 기반 원전 사고진단

❖ LightGBM (Light Gradient Boosting Model)
❖ SHAP (SHapley Additive exPlanations)
❖ LIME (Local Interpretable Model-agnostic Explanation)

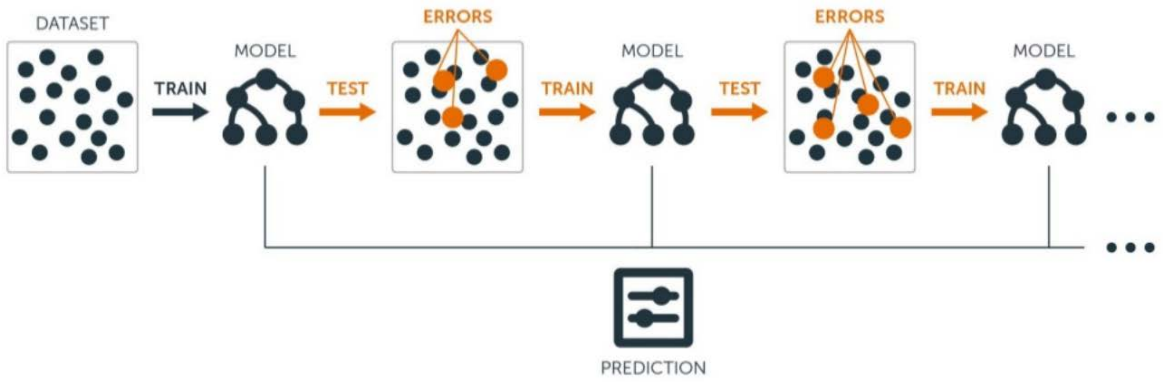
▪ 연구 개요

- 설명가능 인공지능 기술을 활용한 원전 비정상 단일 및 복합사고 진단
- 인공지능 방법론: **LightGBM**
- 설명가능 인공지능 방법론: **Tree SHAP, LIME**
- 데이터: CNS를 활용한 단일 비정상 14건, 복합 비정상 65건 수집 (**전체 80건**; 정상 포함)
  - 복합 비정상 시나리오의 경우, 단일 비정상 시나리오를 조합하여 구성함.

03 설명가능 인공지능 기반 원전 사고진단

▪ 인공지능 방법론 – LightGBM

- Tree 계열의 방법론은 일반적으로 딥러닝 방법에 비해 성능이 낮은 것으로 알려져 있지만, 높은 해석성을 가지고 있음.
- 기존 Tree 계열의 낮은 성능을 향상시키기 위해 Gradient Boosting 알고리즘을 사용하는 Gradient Boosting Decision Tree (GBDT)가 제안됨.
- LightGBM 방법은 GBDT가 가진 긴 학습시간을 해결하기 위해 Gradient-based One-side Sampling (GOSS)와 Exclusive Feature Bundling (EFB) 방법을 적용하였음.



<GBDT 학습 원리>

- GBDT는 이전에 학습된 트리의 오류가 다음 반복에서 새 의사결정 트리에 추가되는 시퀀스로 훈련된 의사 결정 트리의 앙상블임.
- 이는 모든 후속 모델이 실제 출력과 예측 간의 차이를 학습하는 것을 의미함.

03 설명가능 인공지능 기반 원전 사고진단

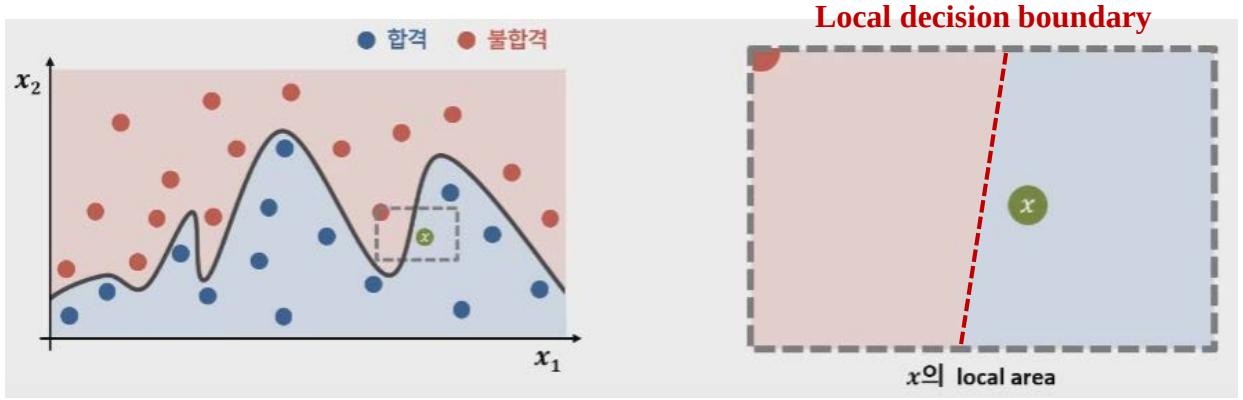
▪ 설명가능 인공지능 기술 분류 기준

- **모델 복잡성 (Complexity)**
  - **Intrinsic: 모델 자체를 해석 가능한 구조로 만드는 기법**
  - **Post-hoc: 모델 자체가 설명력을 지니지 않을 경우, 사후 해석 기법**
- **해석 범위 (Scope)**
  - **Global: 모든 예측 결과에 대해서 항상 설명력을 갖는 전역적인 기법**
  - **Local: 하나 또는 일부 예측 결과만 설명 가능한 국소적인 기법**
- **모델 의존성 (Dependency)**
  - **Model-specific: 특정 종류의 모델만 적용할 수 있는 설명 기법**
  - **Model-agnostic: 모델에 관계 없이 범용적으로 적용할 수 있는 설명 기법**
- **각 기준은 서로 다른 관점을 취하고 있으며, 세 가지 기준 중 하나에 귀속시키는 것이 아님.**

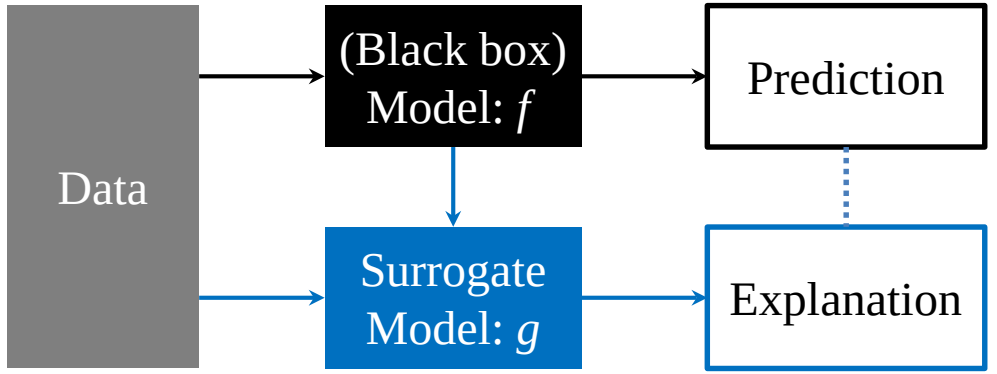
03 설명가능 인공지능 기반 원전 사고진단

▪ 설명가능 인공지능 방법론 – LIME

- LIME은 단일 관측치(혹은 데이터셋 일부분)에 대한 모델 예측 값 해석에 초점을 둠. (Local scope)
- Surrogate Model: 원래 모델 자체로 해석하기 어려울 때 외부에 구조가 간단한 대리 모델로 해석
- LIME은 다음의 두 가지 아이디어를 기반으로 개발된 방법임.
  - 복잡한 데이터에 적합된 복잡한 모델의 전역적인 해석(Global Interpretation)은 어려움.
  - 국소적(local)으로는 비교적 해석이 간단한 모델(Surrogate Model)로 근사 시킬 수 있다고 가정하면, 국소적인 해석(Local Interpretation)으로 설명 가능할 것임.



<Global and Local interpretation>



<LIME 적용 구조>

참고문헌: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier.“, An Overview of Model-Agnostic Interpretation Methods, DMQA.korea.ac.kr

▪ 설명가능 인공지능 방법론 – LIME

- LIME은 관측치 x에 대한 좋은 설명 모델 g (Surrogate Model)을 만들기 위한 두 가지 기준이 존재함.
  - 해석할 수 있는 단순한 모델이여야 함.
  - 해석하고자 하는 모델의 예측과 유사하게 예측할 수 있어야 함.
- $\xi(x) = argmin\, L(f, g, \pi_x) + \Omega(g); g \in G$
  - G: 해석력이 좋은 모델들의 집합
  - $\Omega(g)$: 설명 모델 g의 복잡한 정도 → 해석할 수 있는 단순한 모델인지를 평가
  - $L(f, g, \pi_x) = \sum_{z,z' \in Z} \pi_x(z)(f(z) - g(z'))^2$ → 해석하고자 하는 모델의 예측과 유사한지를 평가
    - f: 해석이 어려운 복잡한 모델 (해석하고자 하는 모델)
    - z: 관측치 x 주변의 근접 데이터
    - z': 차원 축소된 z
    - π_x: z와 x 사이의 거리를 기반으로 계산된 각 z의 가중치
      - ✓ 관측치 x와 가까운 z에 높은 가중치 부여

CHOSUN UNIVERSITY

참고문헌: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier."

▪ 설명가능 인공지능 방법론 – LIME

- **장점**
  - Global한 해석이 아닌 개별 데이터 인스턴스에 대한 Local 해석력을 제공함.
  - Perturbation의 방식을 다르게 함으로써 Model-Agnostic하게 해석할 수 있는 도구를 제공함.
  - 이후에 나오는 SHAP보다 계산량이 적음.
- **단점**
  - 데이터 분포가 국소(Local)적으로도 매우 비선형일 경우, Local에서 선형성을 가정하는 LIME 또한 설명력에 한계를 갖게 됨.
  - 하이퍼 파라미터에 따라서 샘플링 성능이 불안정(Inconsistent)함.

참고문헌: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier."

03 설명가능 인공지능 기반 원전 사고진단

▪ 설명가능 인공지능 방법론 – SHAP

- **SHAP 방법론은 Shapley Values를 기반으로 예측 값에 대해 각 feature가 미치는 기여도를 측정하여 예측에 대한 해석을 제공하는 방법임.**
- **Shapley Values란 1951년 Lloyd Shapley가 제안하였으며 가장 합리적인 분배 이론 중 하나로, 게임 이론을 바탕으로 게임에서 각 플레이어의 기여도에 따라 상금을 공정하게 할당하기 위한 방법임.**

1. 각 플레이어에게 할당되는 상금의 총합은 게임의 상금과 동일해야 함.

2. 플레이어의 기여도가 높을수록 높은 상금이 할당되어야 함.

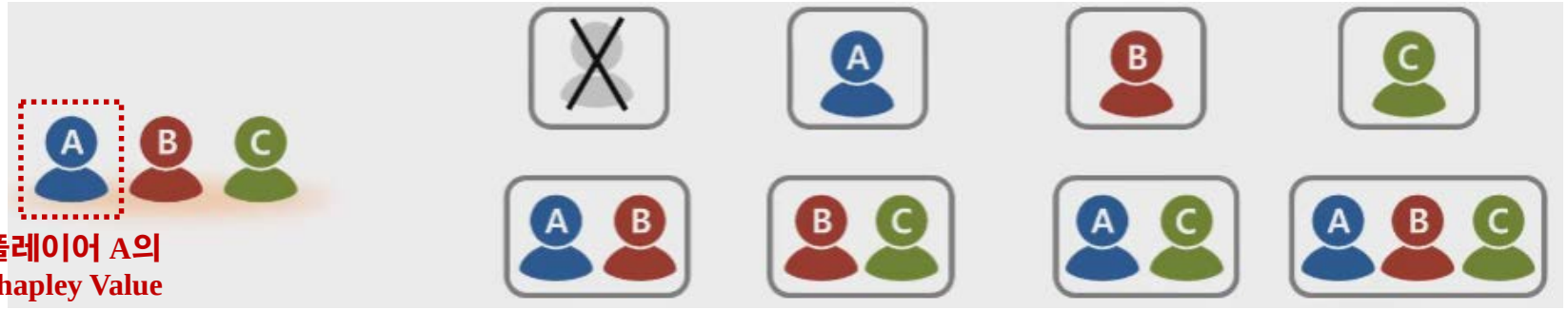
<공정한 상금 할당을 위한 두 가지 조건>

참고문헌: Scott Lundberg et al. “From local explanations to global understanding with explainable AI for trees”, An Overview of Model-Agnostic Interpretation Methods, DMQA.korea.ac.kr

03 설명가능 인공지능 기반 원전 사고진단

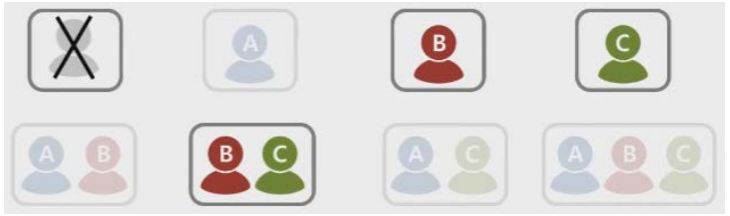
▪ 설명가능 인공지능 방법론 – SHAP

- Shapley values는 플레이어의 Marginal contributions를 계산하여 가중 평균한 값을 의미함.
  - Marginal contributions란 **플레이어 전체 집합**에 대해 가능한 **모든 부분 집합**마다 **특정 플레이어 존재 여부**에 따른 상금 변화량을 의미함.



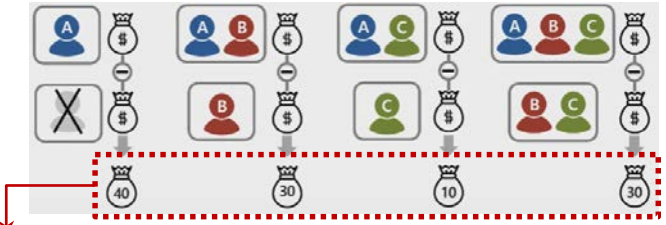
플레이어 A의
Shapley Value

<플레이어 전체 집합에 대해 가능한 모든 부분 집합 예시>



<플레이어 A가 소속된 집합 제외>

<각 부분 집합에 대한 상금 계산>



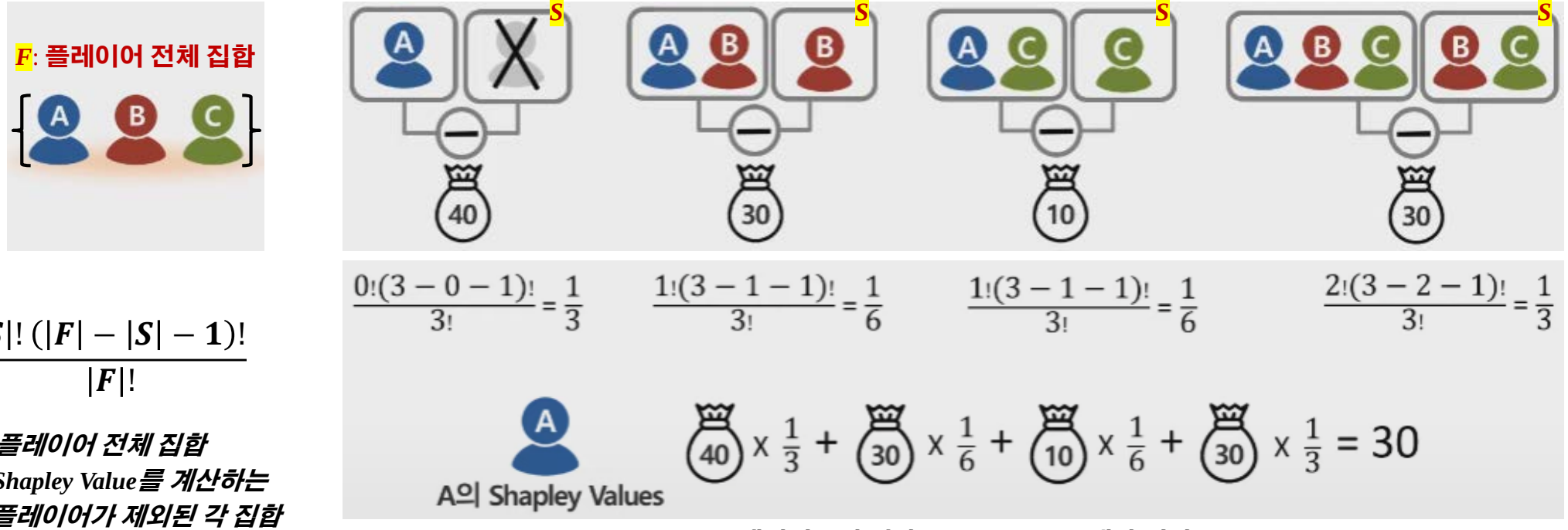
Marginal contributions

<A에 대한 상금만 계산>

참고문헌: Scott Lundberg et al. “From local explanations to global understanding with explainable AI for trees”, An Overview of Model-Agnostic Interpretation Methods, DMQA.korea.ac.kr

▪ 설명가능 인공지능 방법론 – SHAP

- Shapley values는 플레이어의 Marginal contributions를 계산하여 가중 평균한 값을 의미함.
  - 플레이어 A, B, C는 모델의 input features이며, 상금은 모델의 예측을 의미함.



$$\frac{|S|!\,(|F|-|S|-1)!}{|F|!}$$

❖ F: *플레이어 전체 집합*
❖ S: *Shapley Value를 계산하는 플레이어가 제외된 각 집합*

<플레이어 A에 대한 Shapley Values 계산 절차>

참고문헌: Scott Lundberg et al. “From local explanations to global understanding with explainable AI for trees”, An Overview of Model-Agnostic Interpretation Methods, DMQA.korea.ac.kr

▪ 설명가능 인공지능 방법론 – SHAP

- SHAP은 게임 이론을 바탕으로 하는 Shapley Values를 사용하여 예측에 대한 각 feature의 기여도를 계산함.
- Shapley Values 계산 시 수많은 경우에 대해 연산이 이루어지므로 오랜 시간이 걸린다는 단점이 존재함.
- Shapley Values의 단점을 보완하기 위해 Kernel SHAP, Deep SHAP, Tree SHAP 등 다양한 구조의 방법이 제안됨.

$$\phi_i = \sum_{S \in F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]$$

❖ ϕ_i : *i 데이터에 대한 Shapley Value*
❖ F: *전체 집합*
❖ S: *전체 집합에서 i 번째 데이터가 제외된 나머지 모든 부분 집합*
❖ $f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right)$: *i 번째 데이터를 포함한(=전체) 기여도*
❖ $f_S(x_S)$: *i 번째 데이터가 제외된 나머지 부분 집합의 기여도*

CHOSUN UNIVERSITY

참고문헌: Scott Lundberg et al. “From local explanations to global understanding with explainable AI for trees”, An Overview of Model-Agnostic Interpretation Methods, DMQA.korea.ac.kr

▪ 설명가능 인공지능 방법론 – SHAP

- **장점**
  - **Model-Agnostic 방법론 중에서 Explanation model이 가져야할 좋은 특성들이 이론적으로 잘 증명된 방법론임.**
  - **각 관측치에 대한 Local Explanation뿐만 아니라, 각 feature 별 SHAP mean으로 Global Explanation도 얻을 수 있음.**
- **단점**
  - **Kernel SHAP의 경우, 계산 속도가 느림.**
  - **자칫하면 SHAP value를 원인/결과로 해석할 여지가 있음.**

참고문헌: Scott Lundberg et al. “From local explanations to global understanding with explainable AI for trees”.

03 설명가능 인공지능 기반 원전 사고진단

▪ 데이터베이스

- CNS를 활용하여 단일 비정상 14건, 복합 비정상 65건을 수집함. (정상 상태를 포함하여 전체 80건)
  - 복합 비정상 시나리오의 경우, 단일 비정상 시나리오를 조합하여 구성함.
- 아래 표는 단일 비정상 시나리오에 대한 정보를 보여줌.

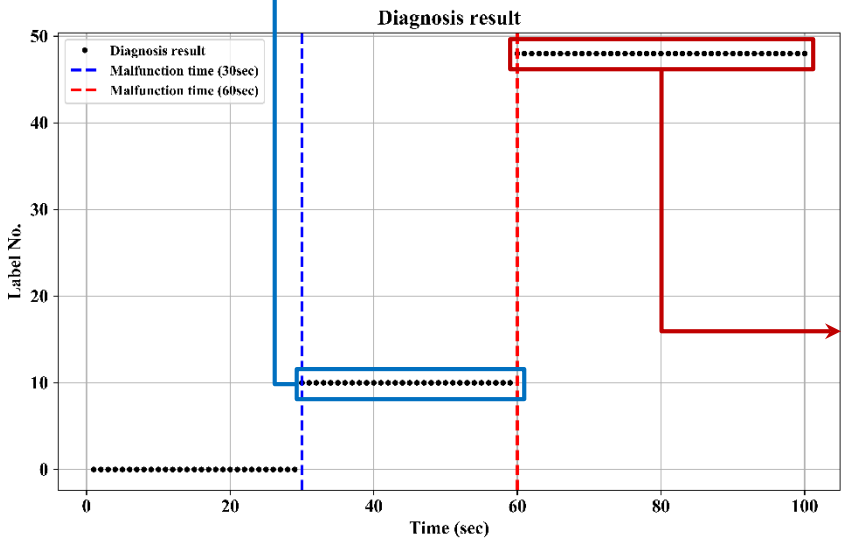
No.	Label	Name	Data No./Length
1	ab20_04	가압기 수위채널 고장 '저'	15/17,042
2	ab20_01	가압기 수위채널 고장 '고'	6/10,626
3	ab21_01	가압기 압력채널 고장 '고'	15/2,539
4	ab15_07	증기발생기 수위채널 고장 '저'	40/70,840
5	ab15_08	증기발생기 수위채널 고장 '고'	40/3,901
6	ab19_02	가압기 안전밸브 고장	50/24,753
7	ab21_12	가압기 PORV 고장 '열림'	51/29,693

No.	Label	Name	Data No./Length
8	ab21_11	가압기 살수밸브 고장 '열림'	70/47,119
9	ab63_02	제어봉의 계속적인 삽입	8/11,144
10	ab23_01	1차기기 냉각수(CCW)계통으로 누설 시 (RCS에서)	42/1,290
11	ab23_03	1차기기 냉각수(CCW)계통으로 누설 시 (CVCS에서)	50/88,550
12	ab60_02	재생열교환기 전단부위 파열	50/88,550
13	ab59_02	충전수 유량조절밸브 후단 누설	42/74,382
14	ab23_06	증기발생기 전열관 누설 시	36/1,467

03 설명가능 인공지능 기반 원전 사고진단

▪ 원전 사고진단 결과 ①

- 복합 비정상 시나리오: 가압기 살수밸브 고장 '열림' (30초) + 증기발생기 수위채널 고장 '저' (60초)



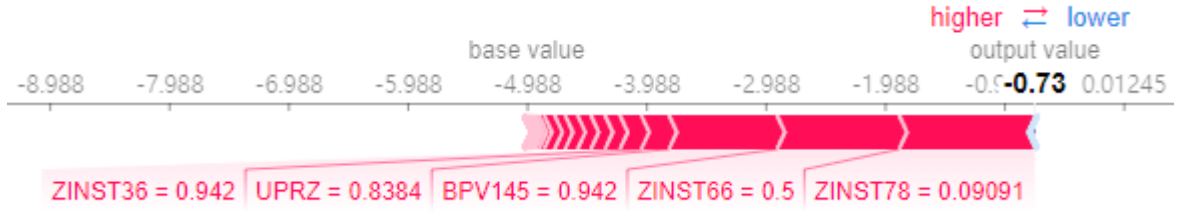
<LightGBM 적용 결과>

가압기 살수밸브 고장 '열림' 진단 근거 (단일 비정상)

<SHAP 적용 결과(35초)>

1) 가압기 살수 유량, 2) Letdown 출구 온도, 3) Letdown 배압조절밸브 개도 상태

가압기 살수밸브 고장 '열림' + 증기발생기 수위채널 고장 '저' 진단 근거 (복합 비정상)



<SHAP 적용 결과(65초)>

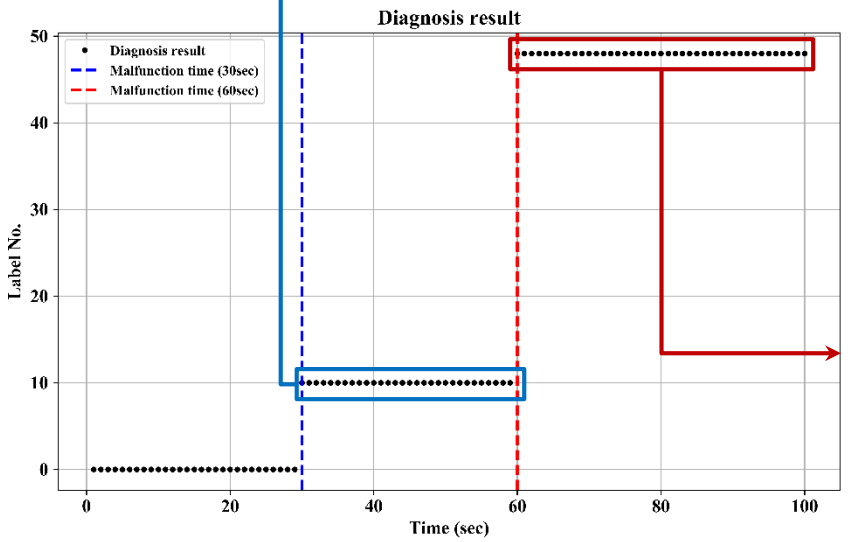
1) 증기발생기 수위, 2) 가압기 살수 유량, 3) Letdown 배압조절밸브 개도 상태

CHOSUN UNIVERSITY

03 설명가능 인공지능 기반 원전 사고진단

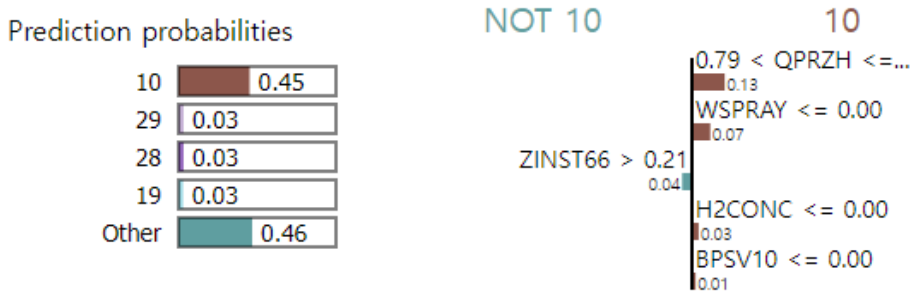
▪ 원전 사고진단 결과 ①

- 복합 비정상 시나리오: 가압기 살수밸브 고장 '열림' (30초) + 증기발생기 수위채널 고장 '저' (60초)



<LightGBM 적용 결과>

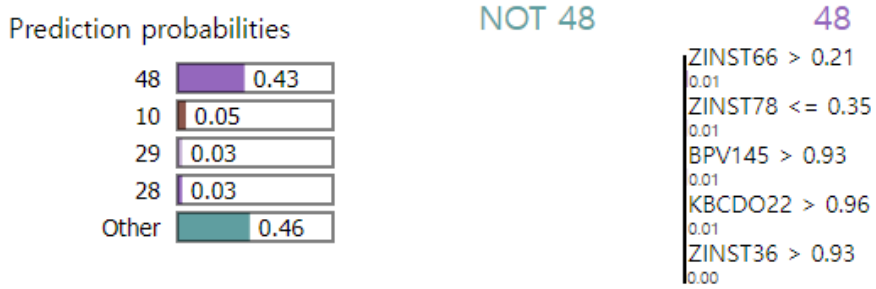
가압기 살수밸브 고장 '열림' 진단 근거 (단일 비정상)



<LIME 적용 결과(35초)>

1) 비례전열기 출력, 2) 가압기 살수 유량

가압기 살수밸브 고장 '열림' + 증기발생기 수위채널 고장 '저' 진단 근거 (복합 비정상)



<LIME 적용 결과(65초)>

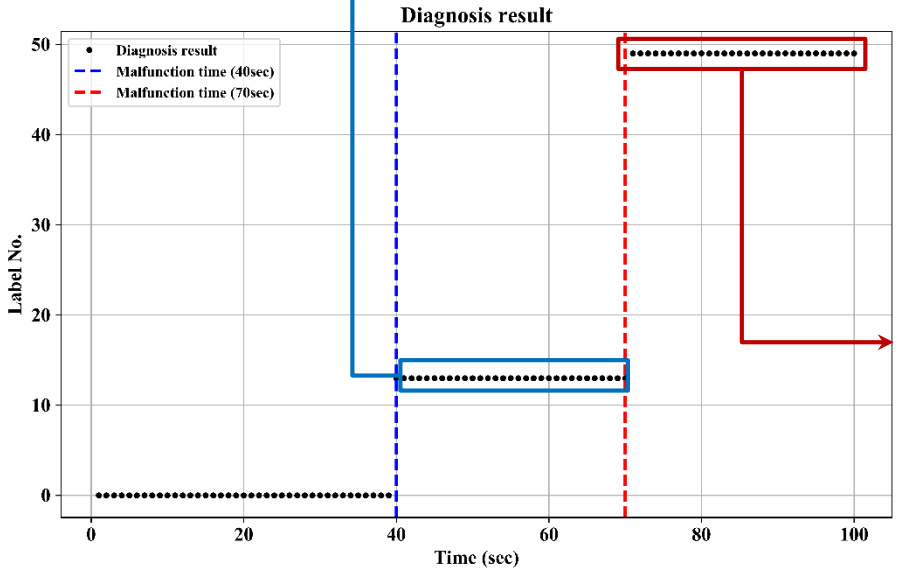
1) 가압기 살수 유량, 2) 증기발생기 수위, 3) Letdown 배압조절밸브 개도 상태

CHOSUN UNIVERSITY

03 설명가능 인공지능 기반 원전 사고진단

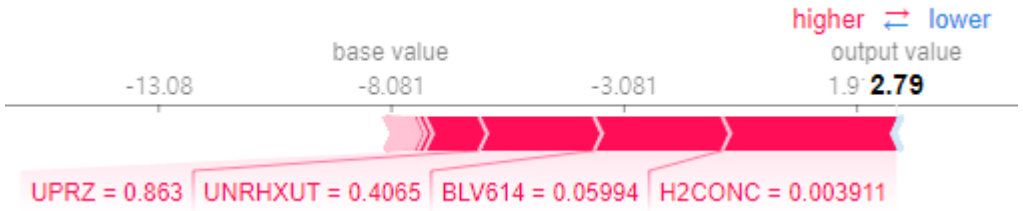
▪ 원전 사고진단 결과 ②

- 복합 비정상 시나리오: 1차기기 냉각수(CCW)계통으로 누설 시 (RCS에서)(40초) + 가압기 안전밸브 고장 (70초)



<LightGBM 적용 결과>

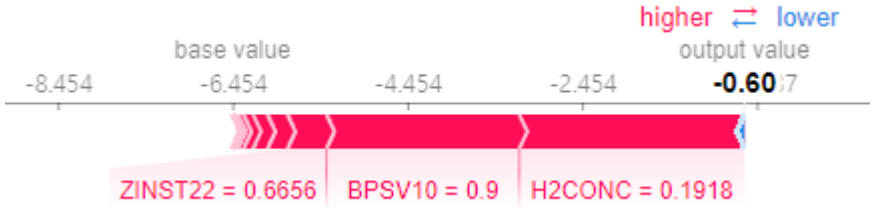
1차기기 냉각수(CCW)계통으로 누설 시 (RCS에서) 진단 근거 (단일 비정상)



<SHAP 적용 결과(45초)>

1) 수소 농도, 2) VCT 수위 조절 밸브 개도 상태, 3) Letdown 출구 온도

1차기기 냉각수(CCW)계통으로 누설 시 (RCS에서) + 가압기 안전밸브 고장 진단 근거 (복합 비정상)



<SHAP 적용 결과(75초)>

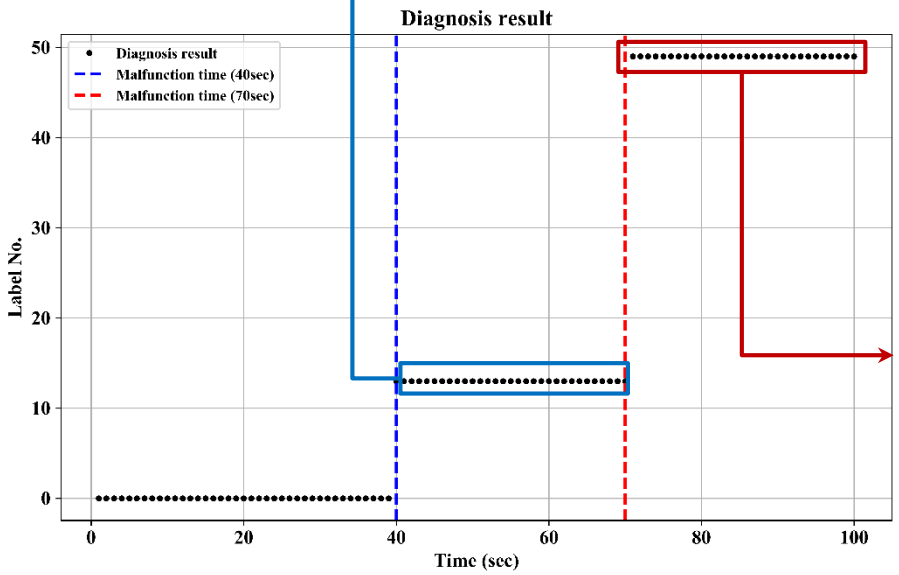
1) 수소농도, 2) 가압기 안전밸브 개도 상태, 3) 격납용기 방사능

CHOSUN UNIVERSITY

03 설명가능 인공지능 기반 원전 사고진단

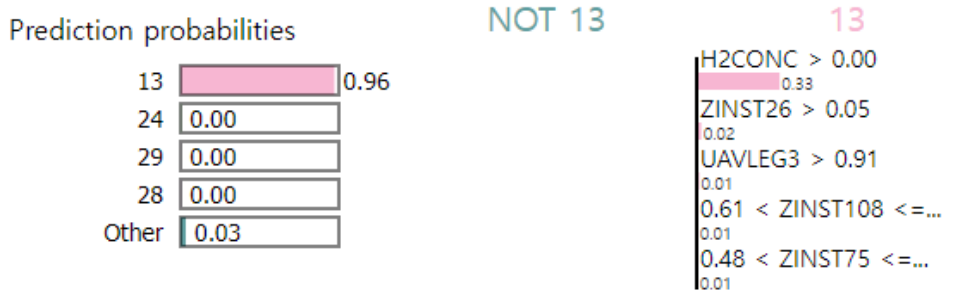
▪ 원전 사고진단 결과 ②

- 복합 비정상 시나리오: 1차기기 냉각수(CCW)계통으로 누설 시 (RCS에서)(40초) + 가압기 안전밸브 고장 (70초)



<LightGBM 적용 결과>

1차기기 냉각수(CCW)계통으로 누설 시 (RCS에서) 진단 근거 (단일 비정상)



<LIME 적용 결과(45초)>

1) 수소농도, 2) 격납용기 압력

1차기기 냉각수(CCW)계통으로 누설 시 (RCS에서) + 가압기 안전밸브 고장 진단 근거 (복합 비정상)

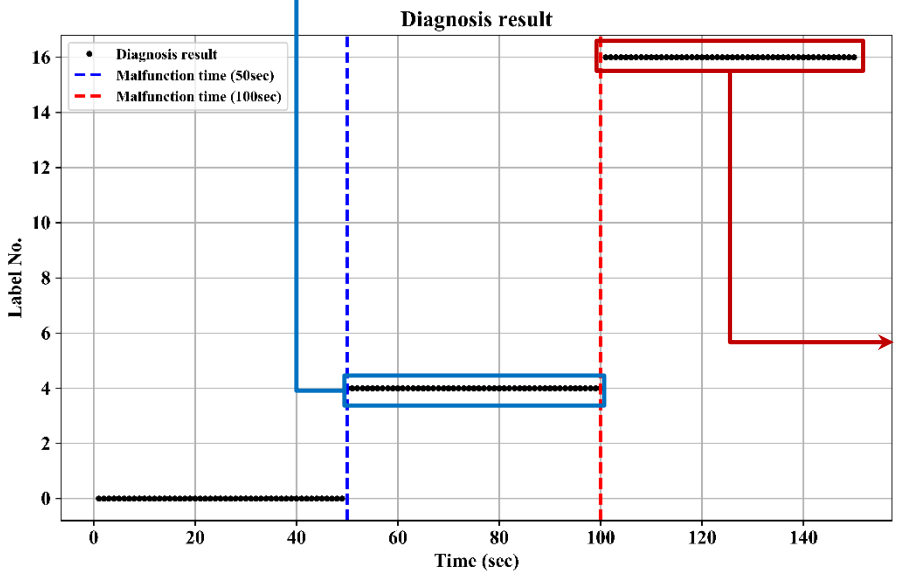
<LIME 적용 결과(75초)>

1) 수소농도, 2) 격납용기 온도, 3) 격납용기 압력

03 설명가능 인공지능 기반 원전 사고진단

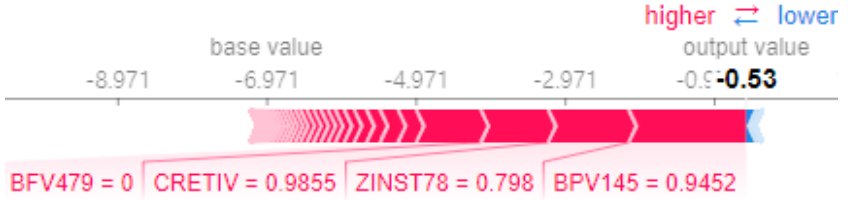
▪ 원전 사고진단 결과 ③

- 복합 비정상 시나리오: 증기발생기 수위채널 고장 '고' (50초) + 증기발생기 전열관 누설 시 (100초)



<LightGBM 적용 결과>

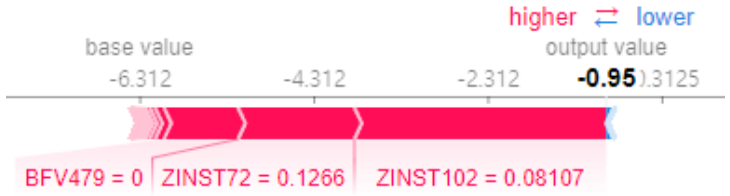
증기발생기 수위채널 고장 '고' 진단 근거 (단일 비정상)



<SHAP 적용 결과(55초)>

1) Letdown 배압조절밸브 개도 상태, 2) 증기발생기 수위, 3) 반응도

증기발생기 수위채널 고장 '고' + 증기발생기 전열관 누설 시 진단 근거 (복합 비정상)



<SHAP 적용 결과(105초)>

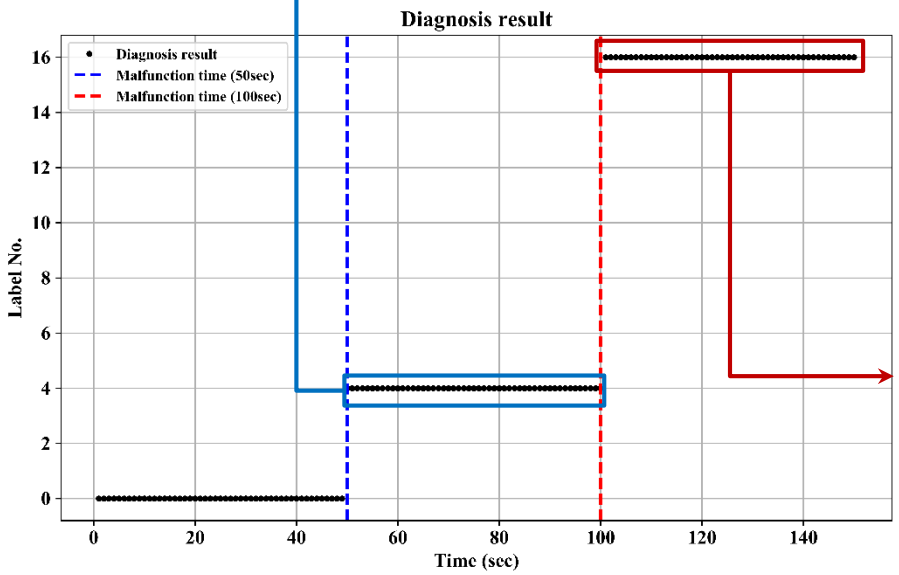
1) 2차측 방사선, 2) 증기발생기 수위, 3) 주급수 우회 밸브 개도 상태

CHOSUN UNIVERSITY

03 설명가능 인공지능 기반 원전 사고진단

▪ 원전 사고진단 결과 ③

- 복합 비정상 시나리오: 증기발생기 수위채널 고장 '고' (50초) + 증기발생기 전열관 누설 시 (100초)



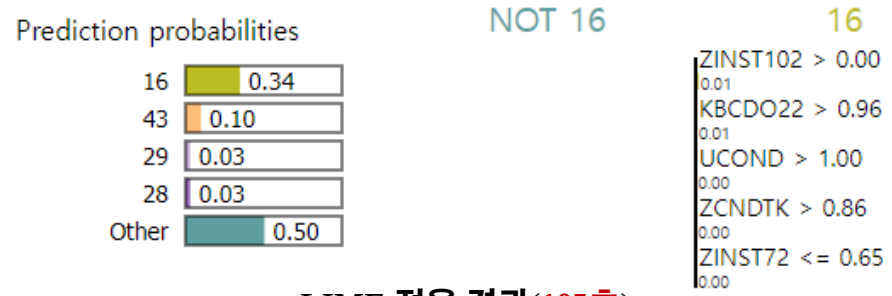
<LightGBM 적용 결과>

증기발생기 수위채널 고장 '고' 진단 근거 (단일 비정상)

<LIME 적용 결과(55초)>

1) 증기발생기 수위

증기발생기 수위채널 고장 '고' + 증기발생기 전열관 누설 시 진단 근거 (복합 비정상)



<LIME 적용 결과(105초)>

1) 2차측 방사선, 2) 전기 출력, 3) 컨덴서 유량 온도

CHOSUN UNIVERSITY

04 기대효과 및 활용방안

04 기대효과 및 활용방안

▪ 설명가능 인공지능 기술의 원전 적용에 따른 기대효과 및 활용방안

- 설명가능 인공지능 기술을 활용함에 따라 기존 인공지능의 블랙박스 특성을 해소함으로써 **신뢰도를 강화**할 수 있음.
- 4차 산업혁명에 따른 인공지능 고도화에 비해 적용성 측면에서 미비한 현 상황에 설명가능 인공지능 기술을 적용함으로써 **원전 산업에 인공지능의 적용 시점을 앞당길 수 있음**.
- 설명가능 인공지능 기술을 통해 인공지능 기술을 활용한 의사결정에 대한 **사회적 신뢰를 구축하고, 인공지능 기반 시스템에 대한 명확성을 보장**할 수 있음.
- 기존에 중점을 두었던 성능 개선을 비롯하여 인공지능 기술의 신뢰성, 품질 개선 방안을 바탕으로 **원전 적용을 위한 인공지능 기술 활용 전략 수립 및 경쟁력 확보가 가능함**.
- 인공지능 기술의 신뢰성, 사용성, 이식성 등 인공지능 소프트웨어 품질을 향상시키기 위한 방안을 제안함으로써 인공지능 기술 도입을 통한 **원전의 안전과 효율 증진에 대한 효과성을 제고**할 수 있음.

Thank you for your attention