Smart Media Journal / Vol.14, No.12 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2025.14.12.90

Model 최적화를 위한 Pertubation/Gradient 기반 양방향 방법론의 XAI 제안

(XAI Proposal of a Pertubation/Gradient-based Bidirectional Methodology for Model Optimization)

차병래*, 노순국**

(Byung Rae Cha, Sun-Kuk Noh)

요 약

점점 더 많은 기업 또는 조직이 중요한 비즈니스 의사 결정 프로세스에 AI와 ML을 도입하기 시작하고 있다. 딥러닝 모델의 장점은 다양한 작업에 매우 높은 정확도를 보여주며, 노이즈(noise)와 데이터의 변화에도 매우 강한 장점과 대용량 데이터 처리를 위한 확장이 가능하다. 또한, 딥러닝의 단점으로는 데이터 종속성, 계산 비용, 그리고 편향성과 해석의 어려움 등이 존재한다. 본 연구에서는 딥러닝의 단점인 데이터 종속성, 편향성 문제와 해석의 어려움에 추가적인 정보를 제공하는 방법을 제안하고자 한다. 제안하는 XAI 방법론은 학습 데이터(training data) 측면과 ML 모델 측면을 고려하는 사전 예방 분석 단계(precautionary analysis step)와 Perturbation-based 방법과 Gradient-based 방법을 활용한 양방향의 설명 가능성을 제안한다.

■ 중심어 : 딥러닝 ; 설명가능한 AI ; Perturbation 기반 방법 ; Gradient 기반 방법 ; AI Factsheets 서비스

Abstract

More and more enterprises or organizations are starting to introduce AI and ML into their critical business decision-making processes. Despite the many advantages of deep learning models, drawbacks of deep learning include data dependency, computational costs, bias, and difficulty in interpretation. In this study, we propose a method to provide additional information on the shortcomings of deep learning, such as data dependency, bias problems, and difficulty of interpretation. The proposed XAI methodology proposes a precautionary analysis step that considers both the training data and ML model aspects, and bidirectional explainability utilizing Perturbation-based and Gradient-based methods.

■ keywords : Deep Learning ; XAI ; Perturbation-based method ; Gradient-based method ; AI Factsheets Service

Ⅰ. 서 론

많은 산업 분야에서 인공지능(AI, artificial intelligence)이 적용되고 있다[1-3]. 또한 점점 더 많은 기업 또는 조직이 중요한 비즈니스 의사 결정 프로세스에 인공지능(AI, artificial intelligence)를 도입하기 시작하고 있으나 AI 모델은 블랙박스이기 때문에 어떠한 결정에 대해 그 이유를 알 수 없다. 이를 사람이 이해할 수 있는 형태로 설명 형태로 설명하고자 다양한 AI 기술이 나오고 있으며, 이를 설명가능한 인공지능(XAI, eXplainable AI) 라고 한다[4-9]. XAI는 책임 있는 AI를 구현하기 위한 핵심 요구 사항 중 하나이며, 공정성(fairness), 모델 설명 가능성(model explainability) 및 책임성(accountability)을 갖추고 실제 조직에서 AI 방법을 대규모로 구현하기 위한 방법론이다. 또한, XAI 기술의 구성은 예측 정확도. 추적성, 그리고

* 정회원, 조선대학교 IT연구소
** 종신회원, 조선대학교 자유전공학부
이 논문은 2024년도 조선대학교 학술연구비의 지원을 받아 수행된 연구임

접수일자 : 2025년 10월 13일
수정일자 : 2025년 11월 05일
게재확정일 : 2025년 11월 18일
교신저자 : 노순국 e-mail : nsk7078@chosun.ac.kr

의사 결정 이해의 세 가지 주요 방법으로 구성된다. 예측 정확도와 추적성은 기술적 요구 사항을 충족하는 반면, 의사 결정 이해는 인간의 요구를 충족하여야 한다[10].

딥러닝 모델의 장점은 다양한 작업에 매우 높은 정확도를 보여주며, 잡음과 데이터의 변화에도 매우 강한 장점과 대용량 데이터 처리를 위한 확장이 가능하다. 또한, 딥러닝의 단점으로는 데이터 종속성, 계산 비용, 그리고 편향성과 해석의 어려움 등이 존재한다. 특히, 딥러닝 모델은 학습되는 데이터의 변화에 민감할 수 있다. 즉, 데이터가 변경되면 모델을 다시 학습시켜야 할 수 있다. 이러한 문제점에도 불구하고 딥러닝은 많은 현실 문제를 해결할 수 있는 잠재력을 가진 강력한 기술이며, 본 연구에서 딥러닝의 편향성 문제와 해석의 어려움을 해결하기 위해 추가적인 정보를 제공하는 방법을 제안하고자 한다.

II. AI 산업 동향과 XAI

2.1 AI 산업 및 기술 동향

"중소기업 전략기술로드맵 2025~2027 AI[11]"에서 AI 산업은 인공지능 기술개발 및 인공지능 적용 제품 서비스 플랫폼의 생산, 유통, 부가서비스 과정에서 가치를 창출하는 산업을 의미하며 AI 기술 혹은 이를 활용한 제품 서비스를 포함한다. 그림 1은 AI 산업 시장의 성장 추이를 나타낸다.

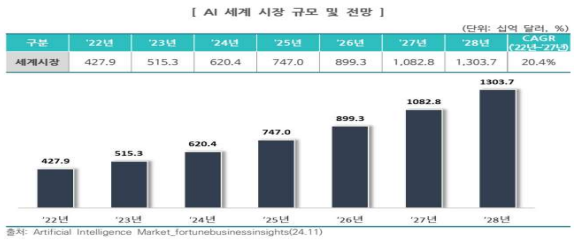

그림 1. AI 서비스의 세계 시장 규모 및 전망
Fig. 1. Market size and outlook of global AI service

2.2 XAI

XAI는 AI와 ML 솔루션이 투명하고, 신뢰할 수 있고, 책임감 있고, 윤리적임을 보장하는 가장 효과적인 실천이며, 알고리즘 투명성, 위험 완화 및 대체 계획에 대한 모든 규제 요구 사항을 효율적으로 처리할 수 있다. 일반적인 ML과 설명 가능한 ML간의 개념과 분류를 그림 1, 2에 나타낸다.

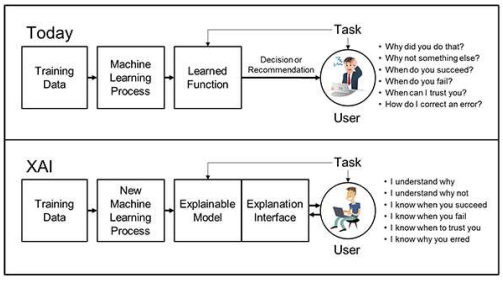

그림 1. XAI 개념 [2]
Fig. 1. XAI Concept

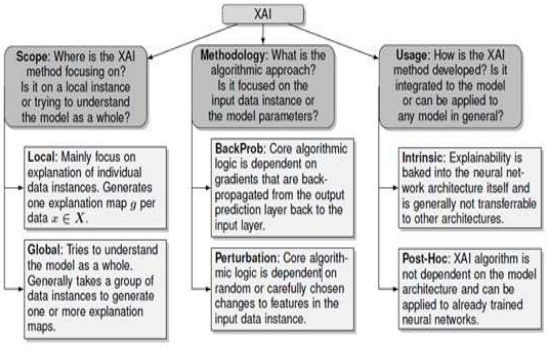

그림 2. 범위, 방법론, 사용법 측면에서 XAI 분류 [9]
Fig. 2. Categorization of XAI in aspect of scope, methology, and usage

XAI의 다양한 알고리즘은 그림 3과 같이 Scope, Methodology, Usage의 기준으로 분류할 수 있으며, scope의 경우 local과 global, methodology의 경우 perturbation과 gradient, usage의 경우 intrinsic, post-Hoc으로 구분하고 있다[12].

또한, Christopher는 XAI의 범위 관점에서 크게 Local model-agnostic 방법(Ceteris Paribus Plots, Individual Conditional Expectation, LIME, Counterfactual Explanations, Scoped Rules, SHAP), Global model-agnostic 방법(Parial Dependence Plot, Accumulated Local

Effects, Feature Interaction, Functional Decomposition, Permutation Feature Importance, Leave One Feature Out Importance, Surrogate Models, Prototypes and Criticisms), 그리고 Neural network interpretation 방법(Learned Features, Saliency Maps, Detecting Concepts, Adversarial Examples, Influential Instances)으로 구분했으며, 그림 4를 참조한다[13].

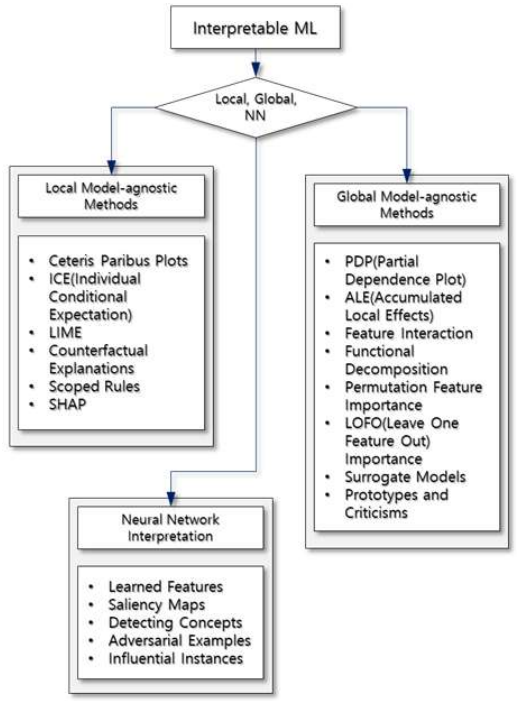

그림 3. Christopher의 Interpretable ML 분류[13]
Fig. 3. Christoper's classification of Interpretable ML[

2.3 XAI의 methodology과 Post-Hoc 활용

XAI 알고리즘의 methodology로 분류하면 크게 Perturbation 기반과 Gradient 기반으로 구분된다. Perturbation 기반 방법론은 모델에 입력 데이터의 다양한 변화를 주며 학습 모델을 반복적으로 조사함으로써 모델을 설명하고자 한다. Gradient 기반 방법론은 신경망 결과에 입력 x가 어떻게 영향을 미치는가를 설명하기 위해 신경망 내에서 정보 흐름의 backward pass에 주목한다.

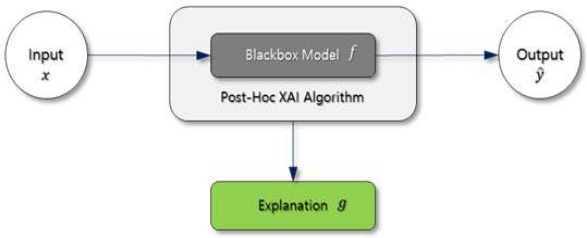

그림 4. XAI의 Post-Hoc 활용
Fig. 4. Post-Hoc usage of XAI

XAI의 Post-Hoc 활용(그림 5 참조)은 ML 모델에 설명 기법이 내재하여 있지 않으며, 이미 훈련된 모델의 결정에 관해 설명하려면 AI를 black 또는 white로 보는 알고리즘이 필요하다. 블랙박스는 내부 operation을 알 수 없으며, 화이트박스에서는 모델의 구조와 레이어 구조를 알고 있어야 한다. 이미 정확도가 높은 모델에 설명가능성까지 부여할 수 있다는 점으로부터 유용하다. XAI 알고리즘은 어떠한 네트워크 구조에서나 쓰일 수 있다는 모델 불가지론(model-agnostic)으로 Deconvolution network, Saliency maps, Most attribution based method 등이 네트워크를 화이트 또는 블랙박스로 여기는 데 쓰일 수 있다.

2.4 ML 모델의 교환 포맷과 AI Factsheets

다양한 ML 모델의 변환과 교환을 위한 NNEF(neural network exchange format)[14]와 ONNX(open neural network exchange)[15]는 딥러닝 프레임워크와 추론 엔진에서 신경망을 표현하고 교환하는 두 가지 유사한 오픈 포맷이다. 핵심적으로 두 포맷 모두 네트워크를 구축할 수 있는 자주 사용되는 연산의 모음을 기반으로 한다[16].

AI 거버넌스 관리를 위한 AI Factsheets를 사용하여 ML 모델의 라이프사이클에 대한 메타데이터와 팩트를 수집하여 ML 모델을 추적한다[16]. AI 솔루션을 개발하기 전에 먼저 비즈니스 유스케이스를 정의한 후 솔루션의 개발, 테스트

및 배치를 관리해야 한다. ML 모델의 목표를 정의하는 모델 유스케이스(model usecase)를 작성하여 정보 플로우(information flow)를 관리하고 통제할 수 있다. 모델이 승인되고 개발이 시작되면 유스케이스의 자산을 추적하여 AI Factsheets를 사용하여 모든 관련 데이터를 캡처하며, 프로덕션에 있는 모델 및 개발이나 변형이 필요한 항목들의 확인이 가능하다.

III. 모델 최적화를 위한 양방향 기법의 XA제안

본 연구에서는 다음과 같은 데이터 측면과 ML 모델 측면의 2가지 제약을 기준으로 XAI 방법을 제안하고자 한다:

- ANI(Artificial Narrow Intelligence) 또는 Weak AI 유형 측면에서 딥러닝의 비정형 데이터(unstructured data) 처리 특성을 반영함.
- 기존 ML 모델의 유지보수와 활용 측면에서 ML의 model-agnostic과 Post-Hoc 활용을 적용하고자 함.

3.1 양방향 기법의 XAI 제안

머신러닝의 최적화(optimization)를 위한 학습 데이터의 Precaution Analysis, XAI, 그리고 하이퍼파라미터 튜닝은 모두 ML 모델의 성능 향상과 신뢰성 확보에 필수적인 요소들이며, 서로 밀접하게 연관되어 있다.

Precaution Analysis는 양질의 학습 데이터셋(training dataset)을 준비하여, 최적화 과정이 올바른 방향으로 진행될 수 있는 기반을 마련하며, 하이퍼파라미터 튜닝은 최적화 알고리즘의 효율성과 최종 모델의 성능을 극대화하여, 주어진 훈련 데이터로부터 최적의 파라미터를 찾도록 돕는다. 그리고 XAI는 최적화된 모델의 결정 과정을 설명하고, 훈련 데이터의 잠재적 문제점(편향 등)이 모델에 미친 영향을 진단하며, 하이퍼파라미터 튜닝의 결과가 모델의 설명력에 어떻게 기여하는지 평가하는 중요한 도구이다.

이 모든 과정은 모델의 성능을 향상시키고, 동시에 모델의 신뢰성, 투명성, 공정성을 확보하여 실제 문제 해결에 성공적으로 적용될 수 있도록 돕는다. 하나의 과정에서 발생한 문제는 다른 과정에 영향을 미치므로, 통합적인 관점에서 접근하고 반복적으로 개선해 나가는 것이 중요하다.

본 연구에서 제안하는 XAI 방법론은 XAI 전략 측면과 학습 데이터(training data) 측면, 그리고 ML 모델 측면을 고려하고자 한다. 학습 데이터 측면에서는 사전 예방 단계(precautionary step)를 제안한다. ML 모델의 분석 측면에서는 Perturbation-based 방법과 Gradient-based 방법을 활용한 양방향의 설명 가능성을 제안하고, ML 모델의 사후 처방으로 Hyperparameter Tuning 단계를 제안한다. 그리고 마지막으로 이를 총괄하기 위한 XAI 전략 단계로 구성되며 그림 6과 같이 플로우챠트로 나타낸다.

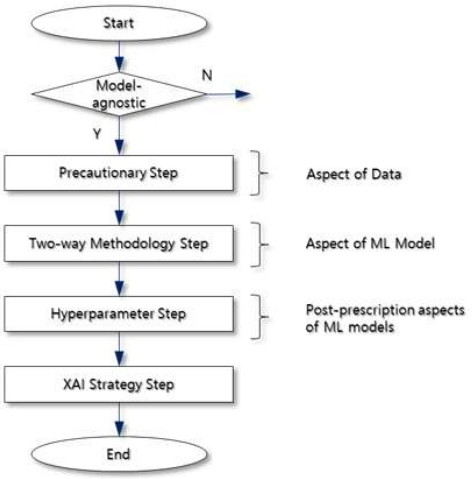

그림 5. 제안하는 XAI 방법론의 플로우챠트.
Fig. 5. Flowchart of proposed XAI methodology

3.2 Precautionary Step for training data analysis

설명 가능성 접근 방식에 대한 이론적 보장에

도 불구하고, 특정 애플리케이션 도메인(domain)이나 사용된 학습 데이터셋으로 인해 속성(features) 중 일부는 구현 과정에서 손실될 수 있다[17].

제안하는 XAI 방법론의 사전 예방 단계로 학습 데이터의 분석을 통한 모델의 표류(data & concept drift, [18])와 미적합/과적합(underfitting/overfitting)을 위한 데이터 증강(data augmentation), 데이터 프로파일(data profile), 적대적 강건성(adversarial robustness), 그리고 데이터 예측 가능성(data forecastability)으로 구성되며, 이를 데이터 중심(data-centric) XAI라고도 한다. 데이터 중심 XAI는 데이터 볼륨(data volume), 데이터 일관성(data consistency), 데이터 순수성(data purity) 측면에서 블랙박스 모델에 대한 설명 가능성을 제공할 수 있으며, 그림 7은 데이터 중심 precaution step의 다양한 기법들을 나타낸다.

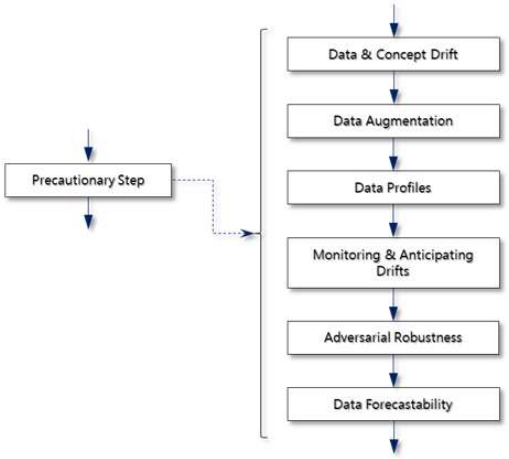

그림 6. 데이터 중심 precaution step의 다양한 기법들.
Fig. 6. Various techniques for data-centric precaution steps.

3.3 Perturbation/Gradient-based 방법론

Perturbation 기반 방법론은 ML 모델에 입력 데이터의 다양한 변화를 주며 학습 모델을 반복적으로 조사함으로써 ML 모델을 설명하고자 한다[20]. 즉, 입력 변수 공간에 변화를 줌으로써 각 변수의 기여도 등을 측정하게 된다. 또한, Gradient 기반 방법론은 Perturbation 기반의 방법론과는 반대로, ML 모델의 결과에 입력 x가 어떻게 영향을 미치는가를 설명하기 위해 ML 모델 내에서 정보 흐름의 backward pass에 주목한다.

본 연구에서 제안하는 양방향의 Perturbation/Gradient-based 방법론은 그림 8과 같이 나타내며, 모든 입력의 변화에 따른 ML 모델의 활동과 개별 특성(feature attribution)이 해당 클래스 출력에 미치는 영향을 이해하려는 Perturbation-based 설명 가능성 방법론과 ML 모델에서 정보 흐름의 역방향 패스를 활용하여 입력 x가 출력에 미치는 영향력과 관련성을 이해하려는 Gradient-based 설명 가능성 방법론을 ML 모델에 양방향으로 적용한다.

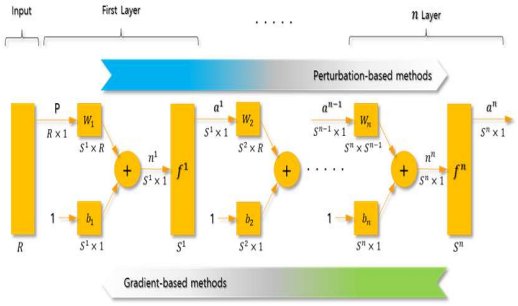

그림 7. ML 모델의 Perturbation/Gradient-based 방법론
Fig. 7. Perturbation/Gradient-based methods of ML model

3.4 Perturbation-based 기법

XAI의 Perturbation-based 기법은 ML 모델의 예측을 설명하기 위해 입력 특성(feature)에 의도적으로 변화(교란, perturbation)를 주었을 때 모델의 출력이 어떻게 변하는지를 관찰하는 방법들을 의미한다. "무엇인가를 건드려보고 결과가 어떻게 바뀌는지 보는" 방식으로, 마치 과학 실험과 유사하다. 그림 9는 Perturbation-based 다양한 기법들을 나타내고 있다.

Perturbation-based 기법의 기본 아이디어는 인과 관계 파악과 "Black-box" 모델에 적용이

가능이다. 인과 관계 파악은 특정 입력 특성을 변경했을 때 모델의 예측이 크게 변한다면, 해당 특성이 모델의 예측에 중요한 영향을 미친다고 추론할 수 있다. 그리고 "Black-box" 모델 적용 가능은 모델의 내부 구조를 알 필요 없이, 단순히 입력과 출력만을 사용하여 설명력을 도출할 수 있다는 큰 장점이 있으며, 이는 Gradient-based 기법과 대조된다.

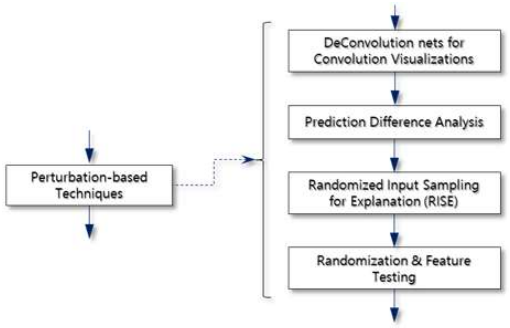

그림 8. Perturbation-based 다양한 기법들

Fig. 8. Perturbation-based various techniques

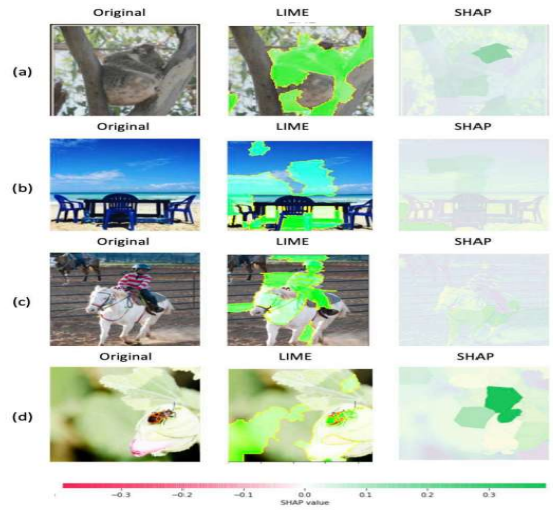

그림 9. Perturbation-based 방법론의 LIME과 SHAP 기법의 예제[9]

Fig. 9. Examples of LIME and SHAP techniques in Perturbation-based method [9]

Perturbation-based 방법론(그림 10 참조)은 주로 개별 feature의 속성(output으로의)을 설명하기 위해 input feature space의 변화에 초점을 두며, LIME(Local Interpretable Model-agnostic Explanations)은 모델이 현재 데이터의 어떤 영역에 집중했고, 어떤 영역을 분류의 근거로 사용했는지 알려주는 XAI 기법이다[19] SHAP(Shapley Additive exPlanations)은 협력 게임 이론(cooperative game theory)의 shapley 값 개념에 기반을 두고 있으며, 가산적 특성 중요도를 고려하며, shapley 값은 특징 공간에서 가능한 모든 값에 걸쳐 각 특징값의 평균 경계 기여도(mean marginal contribution)로 정의된다.

3.5 Gradient-based 기법

XAI의 Gradient-based 기법은 딥러닝 모델의 예측을 설명하기 위해 모델의 출력에 대한 입력 특성(feature)의 기울기(gradient) 정보를 활용하는 방법들을 의미한다. 즉, 모델이 특정 예측을 내릴 때 어떤 입력 부분이 얼마나 중요하게 작용했는지를 밝히는 데 사용된다. Gradient-based 기법의 기본 아이디어는 기울기와 특성 중요도이다. 기울기는 어떤 함수의 변화율을 나타낸다. 딥러닝 모델에서 기울기는 입력 특성이 조금 변했을 때 모델의 출력이 얼마나 변하는지를 보여준다. 특성 중요도는 기울기의 절댓값이 크다면 해당 입력 특성이 모델의 예측에 큰 영향을 미친다는 것을 의미한다. 이를 통해 어떤 입력 부분이 모델의 결정에 중요하게 기여했는지 파악할 수 있게 된다. 그림 11은 Gradient-based 다양한 기법들을 나타내고 있다.

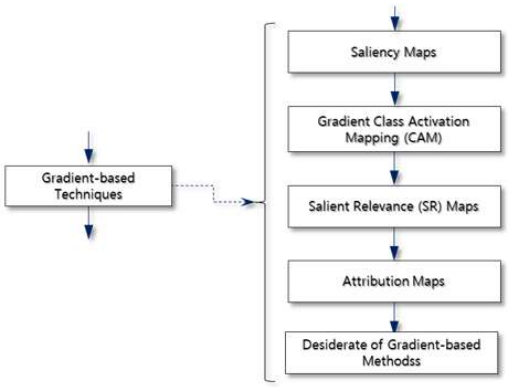

그림 10. Gradient-based 다양한 기법들

Fig. 10. Gradient-based various techniques

Gradient-based 방법론(그림 12 참조)으로 LRP(Layer-wise Relevance backPropagation)는 모델의 결과를 역추적하기 위해 타당성 전파(relevance propagation)와 분해(decomposition) 방법을 사용해 모델을 해부한다. 출력값에서부터 시작해 타당성 점수 또는 기여도라 불리는 relevance score를 입력 계층 방향으로 계산해 나가며 그 비중 분배를 역추적하는 방법이다. CNN에서는 타당성이 계층 간 전파되며, RNN에서는 타당성이 은닉 상태와, 메모리 셀에 전파되기 때문에 모든 모델에 적용할 수 있다.

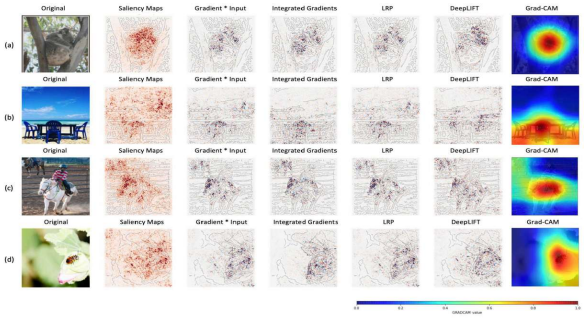

그림 11. Gradient-based 방법론의 다양한 기법들의 예제[9]
Fig. 11. Examples of various techniques in Gradient-based method [9]

3.6 인터페이스 측면의 XAI 시각화와 하이퍼파라미터 튜닝

XAI 시각화(visualization)는 데이터와 ML 모델 간의 복잡한 패턴과 관계를 직관적으로 이해하는 게 매우 중요하다. 특히, 시각화 도구는 학습 데이터와 ML 모델의 복잡성을 이해하고, 모델 가정을 검증하고, 모델 성능을 평가하는 등의 간결한 방법을 제공하며, 다양한 플롯을 통해 데이터와 ML 모델 간의 유용하고 중요한 메타 정보(meta information)를 인식하는 것이 매우 중요하다.

적절한 하이퍼파라미터 조합을 찾는 것은 일반적으로 방대한 검색 공간을 고려할 때 매우 어려운 작업이며, 하이퍼파라미터 튜닝의 접근 방식은 간단한 알고리즘을 구현할 때 활용할 수 있는 수동 실험과 여러 머신러닝 프레임워크(framework)와 호환되는 전문 라이브러리를 이용한 최적화이다. 하이퍼파라미터 튜닝에 전문 라이브러리를 사용하면 작업 확장성이 훨씬 높아지고 시간과 노력을 절약할 수 있다. 해당 도구의 기능, 강점, 단점을 고려하여 특정 ML 모델의 프로젝트에 가장 적합한 도구를 선택함과 동시에 비침습적 옵션(Non-invasive options), 모델링 작업에 완벽하게 통합 가능한 도구, 그리고 AutoML 하이퍼파라미터 최적화 접근 방식 등의 기능이 제공되어야 적합할 것이다.

Ⅳ. P/G 기반 XAI 프레임워크

AI 설계자나 개발자는 수백 개의 머신러닝과 딥러닝 모델에 대해 단 하나의 설명 가능한 AI 솔루션을 설계할 수 없다[19,20]. AI 통찰력(insight)을 이해관계자에게 효과적으로 전달하려면 개별적인 계획, 설계 및 시각화 선택이 필요하다. 또한, 개발자는 실제 구현에서 엄청난 양의 데이터와 결과를 처리하여 적절한 도구 없이는 설명을 찾는 것이 거의 불가능한 상황이다.

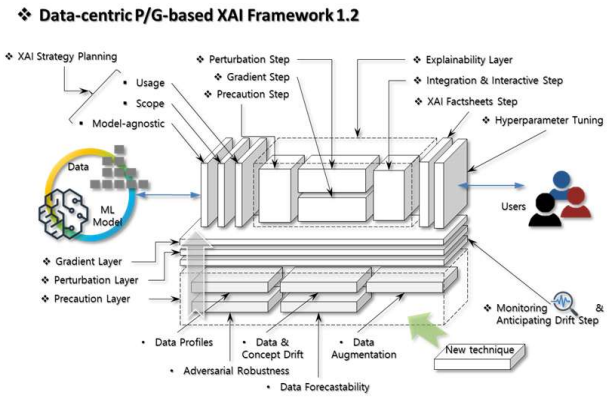

그림 12. 제안하는 데이터 중심의 P/G 기반 XAI 프레임워크의 개념도
Fig. 12. Concept diagram of proposed data-centric P/G-based XAI framework

본 연구에서 제안하는 데이터 중심의 Perturbation /Gradient 기반 XAI 프레임워크의

개념도를 그림 13과 같이 나타내며, XAI 계획, 다양한 기법들의 모듈이 적재된 Precaution Analysis 계층, Perturbation-based 계층, 그리고 Gradient-based 계층을 기반 위에서 각각의 XAI 방법의 기법들이 적용되며, 마지막에는 다양한 기법들의 시각화와 통합된 XAI 보고서가 제공될 것이다.

4.1 모델 최적화의 관계성

제안하는 데이터 중심 Perturbation/Gradient 기반 XAI 프레임워크에서 ML 모델의 최적화, 학습 데이터의 Precaution Analysis, XAI, 그리고 하이퍼파라미터 튜닝은 모두 모델의 성능 향상과 신뢰성 확보에 필수적인 요소들이며, 서로 밀접하게 연관되어 있으며, 각 개념 간의 관계를 자세히 살펴본다.

(1) 모델 최적화

머신러닝 모델의 최적화는 모델의 예측 오차를 최소화하고, 주어진 데이터에 가장 잘 맞는 모델 파라미터(가중치 및 편향)를 찾는 과정이다. 이는 손실 함수(loss function)를 정의하고, 이 손실 함수의 값을 최소화하는 방향으로 파라미터를 업데이트하는 알고리즘을 사용하는 것을 포함하며, 모델 최적화 측면의 학습 데이터, Precaution Analysis, XAI, 그리고 하이퍼파라미터 튜닝 간의 관계성은 다음과 같다:

- 학습 데이터 – 머신러닝 모델은 초기에 학습 데이터를 사용하여 구축된다. 학습 데이터로부터 특징 패턴을 학습하여 모델 파라미터를 설정하고, 점증적으로 최적의 상태에 도달하게 된다.
- Precaution Analysis – 최적화 과정에서 발생할 수 있는 문제점(과적합(overfitting), 불균형 데이터 문제, 이상치(anomaly) 영향)을 사전에 파악하고 대비하는 것이 Precaution Analysis이다. 이는 최적화가 올바른 방향으로 이루어지도록 돕는다.
- XAI – 최적화된 모델이 어떻게 특정 예측을 내리는지 이해하는 데 XAI가 사용된다. 최적화 과정에서 모델이 특정 특징에 과도하게 의존하게 되는 등의 문제가 발생했다면, XAI를 통해 이를 진단하고 최적화 전략을 수정할 수 있다.
- 하이퍼파라미터 튜닝 – 최적화 알고리즘 자체에도 학습률(learning rate), 배치 크기 등과 같은 하이퍼파라미터가 존재한다. 이 하이퍼파라미터들은 최적화 과정의 효율성과 최종적인 모델 성능에 큰 영향을 미치므로, 하이퍼파라미터 튜닝을 통해 최적화 과정을 더욱 효과적으로 만들게 된다.

(2) 학습 데이터의 Precaution Analysis

Precaution Analysis는 모델 학습에 사용될 학습 데이터를 미리 분석하여 잠재적인 문제점(데이터 불균형, 결측치, 이상치, 편향, 특성 간 상관관계, 도메인 드리프트 가능성)을 파악하고, 이에 대한 예방 조치를 취하는 과정이며, 데이터 전처리, 증강, 샘플링 등이 포함되며, Precaution Analysis 측면의 관계성은 다음과 같다:

- 최적화 – 깨끗하고 대표성 있는 학습 데이터는 최적화 과정의 성공에 필수적이다. 불량 데이터는 최적화가 잘못된 방향으로 이루어지거나, 수렴하지 못하게 만들 수 있다. Precaution Analysis를 통해 최적화가 효과적으로 진행될 수 있는 환경을 사전에 조성한다.
- XAI – 데이터 편향은 모델의 예측에 편향을 유발할 수 있으며, 이는 XAI를 통해 드러날 수 있다. Precaution Analysis 단계에서 데이터 편향을 사전에 식별하고 보정하면, 나중에 XAI를 통해 편향된 예측을 진단하는 수고를 덜 수 있다. XAI는 Precaution Analysis가 충분했는지 검증하는 도구로도 활용될 수 있다.

- 하이퍼파라미터 튜닝 – 데이터의 특성(잡음의 양, 특징(feature)의 수)은 모델의 복잡성(complexity)과 최적의 하이퍼파라미터 값에 영향을 미칠 수 있다. 예를 들어, 잡음이 많은 데이터는 과적합을 방지하기 위해 더 작은 모델의 복잡도(낮은 학습률, 적은 은닉층)가 필요할 수 있다. Precaution Analysis를 통해 데이터 특성을 이해하면 하이퍼파라미터 튜닝 범위를 효율적으로 설정할 수 있다.

(3) XAI

XAI는 복잡한 ML 모델이 어떻게 특정 결정을 내렸는지, 그 과정과 이유를 인간이 이해할 수 있는 형태로 설명하려는 노력과 기술을 의미한다. 이는 모델의 투명성(transparency), 신뢰성(reliability), 공정성(fairness)을 높이는 데 기여하며, XAI 측면의 관계성은 다음과 같다:

- 최적화 – XAI는 최적화된 모델이 왜 특정 예측을 내리는지 분석하여, 최적화가 의도된 대로 이루어졌는지 검증한다. 만약 모델이 비합리적인 이유로 예측을 내린다면 (특정 잡음에만 반응), 이는 최적화 과정에 문제가 있었음을 시사하며, 이에 따라 최적화 전략을 재고할 수 있다.
- Precaution Analysis – XAI는 모델의 예측이 데이터의 어떤 특성(혹은 편향된 특성)에 의해 주도되는지 보여준다. 만약 모델이 특정 민감 정보에 과도하게 의존하여 예측한다면, 이는 훈련 데이터에 잠재적인 편향이 있었음을 시사하며, Precaution Analysis가 미흡했음을 알려준다. XAI는 Precaution Analysis의 결과가 모델에 미치는 영향을 평가하는 데 유용하다.
- 하이퍼파라미터 튜닝 – XAI는 하이퍼파라미터 튜닝의 결과를 평가하는 데 사용될 수 있다. 예를 들어, 특정 하이퍼파라미터 조합이 모델의 설명력을 떨어뜨리거나, 특정 특징에 대한 과도한 의존성을 유발한다면, 이는 하이퍼파라미터 튜닝 전략을 재조정해야 함을 의미한다. XAI를 통해 다양한 하이퍼파라미터 설정이 모델의 "이해 방식"에 어떻게 영향을 미치는지 분석할 수 있다.

(4) 하이퍼파라미터 튜닝

하이퍼파라미터 튜닝 과정은 모델 학습 과정에서 직접 학습되지 않고, 사용자가 사전에 설정해야 하는 값(학습률, 은닉층 수, 뉴런 수, 정규화 강도, 배치 크기 등)들을 최적의 성능을 내도록 조정하는 과정이며, 그리드 서치(grid search), 랜덤 서치(random search), 베이지안 최적화(bayesian optimization) 등의 방법이 사용된다. 하이퍼파라미터 튜닝 측면의 관계성은 다음과 같다:

- 최적화 – 하이퍼파라미터는 최적화 알고리즘의 동작 방식과 효율성에 직접적인 영향을 미친다. 예를 들어, 너무 큰 학습률은 발산을 유발할 수 있고, 너무 작은 학습률은 수렴이 매우 느려지게 한다. 따라서 하이퍼파라미터 튜닝은 최적화가 성공적으로 이루어지도록 돕는다.
- Precaution Analysis – 데이터의 특성에 따라 최적의 하이퍼파라미터는 달라질 수 있다. Precaution Analysis를 통해 데이터의 복잡성, 잡음 수준 등을 파악하면, 하이퍼파라미터 튜닝 시 탐색 공간을 효율적으로 설정하거나 특정 하이퍼파라미터에 더 집중할 수 있다.
- XAI – 튜닝된 하이퍼파라미터 설정이 모델의 설명력에 어떤 영향을 미치는지 XAI를 통해 분석할 수 있다. 예를 들어, 특정 정규화 강도가 모델을 너무 단순하게 만들어 중요하지 않은 특징까지 무시하게 만들거나, 너무 복잡하게 만들어 특정 잡음에 과도하게 반응하게 만들 수 있다. XAI는 하이퍼파라미터 튜닝의 목표가 단순히 성능 지표를

넘어서 모델의 이해 가능성과 신뢰성을 확보하는 데 있음을 보여준다.

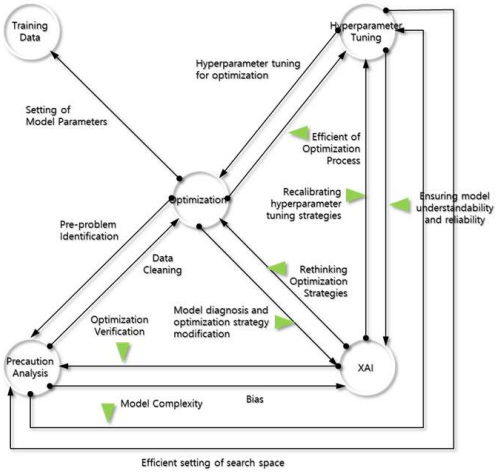

그림 13. 모델 최적화의 관계성
Fig. 13. Relationship for model optimization

이 네 가지 개념은 상호 의존적이며, 성공적인 머신러닝 모델 개발을 위해 유기적으로 연결되어야 하며, 그림 14와 같이 나타낸다.

Precaution Analysis는 양질의 훈련 데이터를 준비하여, 최적화 과정이 올바른 방향으로 진행될 수 있는 기반을 마련하며, 하이퍼파라미터 튜닝은 최적화 알고리즘의 효율성과 최종 모델의 성능을 극대화하여, 주어진 훈련 데이터로부터 최적의 파라미터를 찾도록 돕는다. XAI는 최적화된 모델의 결정 과정을 설명하고, 훈련 데이터의 잠재적 문제점(편향 등)이 모델에 미친 영향을 진단하며, 하이퍼파라미터 튜닝의 결과가 모델의 설명력에 어떻게 기여하는지 평가하는 중요한 도구이다. 결론적으로, 이 모든 과정은 모델의 성능을 향상시키고, 동시에 모델의 신뢰성, 투명성, 공정성을 확보하여 실제 문제 해결에 성공적으로 적용될 수 있도록 돕는다. 하나의 과정에서 발생한 문제는 다른 과정에 영향을 미치므로, 통합적인 관점(integrated perspective)에서 접근하고 반복적으로 개선해 나가는 것이 중요하다.

4.2 XAI 전략

소프트 데이터 관점은 계획 수립 과정에서 적극적으로 활용해야 한다. 하드 데이터에서 지식을 얻지만, 지혜를 얻을 수 있는 곳은 소프트 데이터이다. 소프트 데이터를 분석하기가 쉽지 않지만, 종합적 판단을 내리기 위해서는 필수불가결한 요소다. 특히, XAI 전략은 설명 가능성 프로세스의 초기 단계이다. 이는 전반적인 XAI 수행 계획과 설명 가능성을 완료하는 기술을 수립하는 데 도움이 된다. XAI 전략은 설명 가능성 프로세스를 보다 체계적이고 객관적으로 만든다.

4.3 XAI의 Integration & Interactive 단계

제안하는 프레임워크의 Integration & Interactive 단계는 XAI 결과의 시각화(Visualization)를 수행하며, 다양한 XAI 시각화 도구(SHAP, LIME, WIT, Saliency Maps/Heatmaps, InterML, AI Explainability 360 등)의 통합 기능과 분석 결과에 대한 Interactive 기능을 제공한다.

제안하는 프레임워크의 XAI 시각화 도구 통합은 AI 모델 유형(분류, 회귀, 이미지, 텍스트 등), 설명 수준(개별 예측의 상세한 설명 또는 전체 모델의 동작 방식을 이해), 사용자(개발자, 도메인 전문가, 일반 사용자), 통합 용이성 등을 고려해야 하며, 이러한 도구들의 통합은 AI 시스템의 투명성을 높이고, 개발자와 사용자 간의 신뢰를 구축하며, 궁극적으로 더 책임감 있는 AI 개발에 기여하게 된다.

사용자와의 제안하는 프레임워크 간의 능동적 상호작용할 수 있게 됨에 여러 가지 장점(이해도와 분석 효율성 증대, 사용자 참여와 몰입도 향상, 문제 해결과 의사 결정 지원, 오류 발견 및 디버깅 용이성, 그리고 접근서와 유용성 증대)을 제공한다:

4.4 XAI Factsheets 단계

제안하는 프레임워크의 XAI Factsheets 단계는 Report 기능을 수행하며, IBM의 AI Factsheets[16]와 유사하게 XAI 프레임워크에 의한 ML 모델 라이프 사이클에 대한 메타데이터 및 팩트를 수집과 테스트한 ML 모델의 추적한 결과를 통합하고 제공한다. XAI Factsheets는 XAI Governance의 핵심이며, 생성된 AI 자산 및 머신러닝 모델을 위한 통합 거버넌스 솔루션의 일부로 제공할 수 있다[20]. 또는 IBM 및 외부 머신러닝 모델을 거버넌스하기 위해 다른 서비스와 함께 작동하는 독립 서비스로 설치하거나, 거버넌스를 구현하려는 방식에 따라 설치 옵션을 선택할 수 있어야 한다.

XAI Factsheets 서비스는 모델 데이터 및 이벤트 추적을 위한 자체 솔루션을 구현하는 데 편리한 대안을 제공할 수 있다. XAI Factsheets 서비스는 머신 러닝 모델과 각 모델 및 배포에 대한 세부 정보를 추적하는 데 사용할 수 있는 팩트시트를 구성하는 데 사용할 수 있는 모델 인벤토리에 연결된다. XAI Factsheets를 사용하면 AI 모델 거버넌스 및 규정 준수에 필요한 필수 모델 세부 정보를 캡처하고 모델 개발 라이프사이클 전체에서 추적할 수 있다.

4.5 Hyperparameter Tuning Report

하이퍼파라미터는 알고리즘이 ML 모델을 생성할 때 동작 방식을 조절하는 데 사용되는 매개변수이다. 최고의 성능을 내는 최적의 하이퍼파라미터 조합을 선택하는 과정을 하이퍼파라미터 최적화(hyperparameter optimization)라고 한다[21]. 적절한 하이퍼파라미터 조합을 찾는 것은 일반적으로 방대한 검색 공간을 고려할 때 매우 어려운 작업이다.

모델에 가장 적합한 하이퍼파라미터 조합을 찾는 데에는 일반적으로 두 가지 접근 방식이 있다. 그리드 서치나 랜덤 서치와 같은 간단한 알고리즘을 구현할 때 활용할 수 있는 수동 실험과 여러 머신러닝 프레임워크와 호환되는 전문 라이브러리를 이용한 최적화이다[22]. 모델 하이퍼파라미터 최적화를 달성하는 데 도움이 되는 Python 라이브러리로는 Bayesian Optimization Library[23], Scikit-Optimize[24], GPyOpt[25], Hyperopt[26], SHERPA[27], Optuna[28], Ray Tune[29], Microsoft's NNI[30], MLMachine[31], Talos[32] 등이 있다. 특히, Talos는 하이퍼파라미터 최적화를 위한 30개 이상의 유틸리티를 제공하며, 그림 15의 다이어그램은 Talos와 Keras를 통합하여 처음부터 프로덕션 환경에 바로 적용 가능한 모델까지 빠르게 구축하는 반자동화 워크플로우(semi-auto workflow)를 나타낸 것이다.

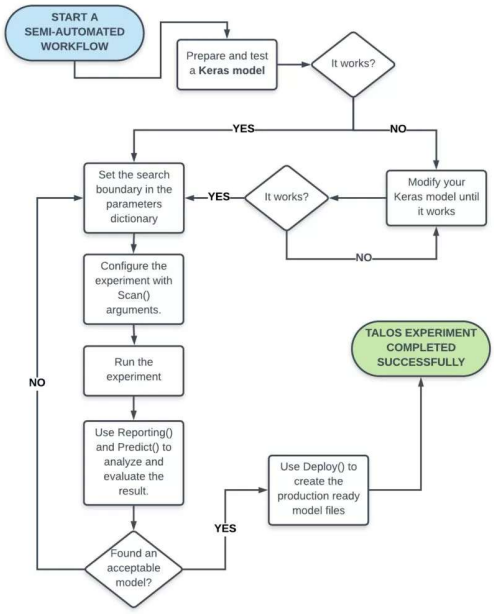

그림 14. 하이퍼파라미터 튜닝 도구 Talos의 반자동 워크플로
Fig. 14. Semi-auto workflow of hyperparamet tuning tool Talos

V. 결 론

점점 더 많은 기업 또는 조직이 중요한 비즈니스 의사 결정 프로세스에 AI와 ML을 도입하기 시작하고 있으며, AI 또는 ML 도입을 늘리기 위해 블랙박스 알고리즘(black-box algorithms)을 즉각적으로 해석하고 이해하기 쉽게 설명하는

것을 기대하고 있다. 딥러닝의 단점으로 데이터 종속성, 편향성 문제와 해석의 어려움에 추가적인 정보를 제공하는 방법을 제안하고자 한다. 본 연구에서 제안하는 XAI 방법론은 학습 데이터(training data) 측면과 ML 모델 측면을 고려하는 사전 예방 단계(precautionary step)와 Perturbation-based 방법과 Gradient-based 방법을 활용한 양방향의 설명 가능성을 제안하였다. 머신러닝의 최적화, 학습 데이터의 Precaution Analysis, XAI(설명 가능한 인공지능), 그리고 하이퍼파라미터 튜닝, 이 네 가지 개념은 상호 의존적이며, 성공적인 머신러닝 모델 개발을 위해 유기적으로 연결되어 있다. 그러나, 본 연구는 XAI 프레임워크의 정의와 기능 등의 간략한 제안의 한계점을 갖고 있으며, 추가 연구를 통해 XAI 프레임워크에 맞는 XAI 도구를 개발하고자 한다. 또한, 향후 연구 내용으로 기존 연구의 연장 측면으로 XAI의 자동화(automation) 연구와 이를 통한 ML 모델의 하이퍼파라미터의 최적화(hyperparameter optimization) 연구로 확장하고자 한다.

지금까지 XAI의 기술 발전은 기계 학습이나 통계학을 중심으로 진행되어 왔으며, AI 사용자에게 정말로 가치가 높은 설명을 실현하기 위해서는 범분야적이고 폭넓은 협력과 발전이 필요하다. 향후 XAI의 이상적인 모습은 XAI에 지식을 활용하는 방법(설명에 지식을 추가 또는 구조적 지식의 활용)과, 심리학이나 인지공학, 행동사회학 등 폭넓은 연구분야와 연계가 필요하다.

감사의 글

"이 논문은 조선대학교 학술연구비의 지원을 받아 연구되었음(2024년도)."

REFERENCES

[1] 신병춘, 차윤석, 김채윤, 차병래, "학습된 머신러닝의 표류 현상에 관한 고찰," *스마트미디어저널,* 제11권, 제7호, 61-69쪽, 2022년 08월

[2] 이소영, 정혜선, 최윤성, 이충권, "딥러닝을 이용한 의류 이미지의 텍스타일 소재 분류," *스마트미디어저널,* 제12권, 제7호, 43-51쪽, 2023년 8월

[3] 정도윤, 최광미, 김남호, "GAN기반의 Semi Supervised Learning을 활용한 이미지 생성 및 분류," *스마트미디어저널,* 제13권 제3호, 27-35쪽, 2024년

[4] A. Adadi and M. Berrada. Peeking Inside the Black-Box: "A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access,* vol. 6, pp. 52138-52160, 2018.

[5] Defense Advanced Research Projects Agency, DARPA "Explainable Artificial Intelligence (XAI)," *DARPA presentation,* DARPA. Retrieved 17 Jul. 2017.

[6] D.E. Lee, C.S. Park, J.-W. Kang, M.W. Kim, "A review of Explainable AI Techniques in Medical Imaging," *Journal of Biomedical Engineering Research,* vol. 43, no 4, pp. 259-270, 2022.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?': Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* pp. 1135-1144, 2016.

[8] Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access,* vol. 6, pp. 52138-52160, 2018.

[9] Sun-Kuk Noh, "Review of medical imaging systems, medical imaging data problems and XAI in the medical imaging field," *J. Internet Comput. Serv.,* vol. 25, no. 5, pp. 53-65, Oct. 2024.

[10] Aditya Bhattacharya, "Applied machine learning explainability techniques," *Packt publishing,* 2022.

[11] IBM, *"What is explainable AI?,"* https://www.ibm.com/think/topics/explainable-ai (accessed July, 8, 2025).

[12] Arun Das, Paul Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," Computer Vision and Pattern Recognition, *IEEE,* Jun. 2020.

[13] Christoph Molnar, *"Interpretable Machine Learning: A Guide For Making Black Box Models Explainable," Self-published Christoph Molnar 3rd Edition,* 2025.

[14] NNEF, https://www.khronos.org/nnef (accessed July, 15, 2025).

[15] ONNX, [Internet] https://onnx.ai/ (accessed July, 15, 2025).

[16] Kang Dae-gi, “Trends in Artificial Neural Network Standard Formats for Deep Learning,” *TTA Jornal,* Vol. 179, pp. 85-90, 2018.
[17] Kacper Sokol, Peter Flach, “Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches,” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* pp. 56-67, 2020.
[18] Max Bennett, “A brief history of Intelligence - Evolution, AI, and the five breakthroughs that made our brains,” *Gilbut Publishing,* 2025.
[19] IBM, *“AI Factsheets,”* https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/factsheets-model-inventory.html?context=cpdaas (accessed Aug., 5, 2025).
[20] Jaehyun Ahn, “AI Explainable Artificial Intelligence, Dissecting Artificial Intelligence - Understanding the Black Box and Improving the Reliability of the System,” *WikiBooks,* 2020.
[21] Louis Owen, “Hyperparameter Tuning with Python - Boost your machine learning model’s pereformance via hyperparameter tuning,” *Packt Publishing,* 2022.
[22] Nicolas Bohorquez, *“Top 10 Tools For Hyperparameter Optimization In Python,”* May 18, 2023. https://www.activestate.com/blog/top-10-tools-for-hyperparameter-optimization-in-python/
[23] Bayesian Optimization Library, https://github.com/bayesian-optimization/BayesianOptimization (accessed Aug., 10, 2025)
[24] Scikit-Optimize, https://scikit-optimize. github.io/stable/ (accessed Aug., 10, 2025).
[25] GPyOpt, https://sheffieldml.github.io/GPyOpt/ (accessed Aug., 15, 2025).
[26] Hyperopt, https://hyperopt.github.io/hyperopt/ (accessed Aug., 15, 2025).
[27] SHERPA, https://github.com/sherpa-ai/sherpa (accessed Aug., 15, 2025).
[28] Optuna, https://optuna.org/ (accessed Aug., 18, 2025).
[29] Ray Tune, https://docs.ray.io/en/latest/tune/index.html (accessed Aug., 18, 2025).
[30] Microsoft’s NNI, https://github.com/microsoft/nni (accessed Aug., 18, 2025).
[31] MLMachine, https://ml-machine.org/ (accessed Aug., 20, 2025).
[32] Talos, https://github.com/autonomio/talos (accessed Aug., 20, 2025).

저 자 소 개

차병래(정회원)

2004년 목포대학교 대학원 컴퓨터공학 공학박사
2005년 ~ 2009년 호남대학교 컴퓨터공학과 전임강사
2009년 ~ 2023년 광주과학기술원 AI대학원 연구부교수
2023년~현재 조선대학교 IT연구소 연구교수
<주관심분야 : 정보보안, 빅데이터, SDS, AI 등>

노순국(종신회원)

1995년 조선대학교 전자공학 공학사
1997년 조선대학교 대학원 전자공학 공학석사
2000년 조선대학교 대학원 전자공학 공학박사
2002년 ~ 2004년 전북대학교 BK기금교수
2004년 ~ 2009년 호남대학교 전파공학과 전임강사
2009년 ~ 2011년 호남대학교 이동통신공학과 조교수
2012년 ~ 2018년 조선이공대학교 전자과 조교수
2018년 ~ 2024년 조선대학교 SW중심대학사업단 부교수
2024년 ~현 재 조선대학교 자유전공학부 부교수
<주관심분야 : 무선이동통신, 전파전파, IoT시스템, 인공지능응용 등>