하이브리드 침입 탐지를 위한 SHAP 기반 설명가능 정책 생성 기법

김건민, 김경백
전남대학교 인공지능융합학과

geonminkim@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

Explainable Policy Generation for Hybrid Intrusion Detection using SHAP

Geonmin Kim, Kyungbaek Kim
Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

본 논문은 기계학습 기반 침입 탐지 시스템의 블랙박스 문제와 규칙 기반 시스템의 낮은 일반화 성능을 동시에 해결하기 위해 설명가능 인공지능(XAI)을 활용한 정책 생성 프레임워크를 제안한다. 제안 기법은 SHAP 기반 피쳐 중요도를 이용하여 공격 탐지에 기여하는 핵심 피쳐를 식별하고, False Positive Rate(FPR) 제약 하에서 공격 탐지율을 최대화하는 임계값 기반 규칙을 생성한다. 또한, 생성된 규칙과 기계학습 모델을 결합한 하이브리드 탐지 구조를 설계하여 상호 보완적인 탐지를 수행한다. 실험 결과, 제안 방법은 약 98.3%의 공격 탐지율을 보였으며, 멀티클래스 분류도 94.2%의 정확도를 달성하였다.

1. 서론

최근 기계학습 기반 침입 탐지 시스템은 높은 탐지 성능을 보이며 다양한 보안 환경에 적용되고 있다. 그러나 대부분의 기계학습 모델은 내부 동작을 해석하기 어려운 블랙박스 형태를 가지며, 실제 보안 운영 환경에서 정책으로 직접 활용하기 어렵다는 한계가 존재한다[1]. 특히 탐지 결과의 근거를 명확히 설명하거나 오탐률을 제어하는 것이 어렵기 때문에, 실무 적용에 제약이 따른다. 한편, 규칙 기반 침입 탐지 방식은 명확한 정책 형태로 해석 가능하고 제어가 용이하다는 장점이 있으나, 고정된 규칙에 의존하기 때문에 다양한 공격 패턴에 대한 일반화 성능이 낮다는 문제가 있다[2]. 기존 연구에서는 기계학습 모델과 규칙 기반 방식을 결합하려는 시도가 있었으나, 대부분 경험적 기준에 의존하거나 설명가능성과 성능을 동시에 고려하지 못한다.

본 논문에서는 이러한 문제를 해결하기 위해, 설명가능 인공지능(XAI)을 활용하여 기계학습 모델의 설명을 정책으로 변환하는 침입 탐지 프레임워크를 제안한다. 제안 기법은 SHAP 기반 피처 중요도를 이용하여 핵심 피처를 식별하고[3], False Positive Rate(FPR) 제약 하에서 공격 탐지율을 최대화하는 임계값 기반 규칙을 생성한다. 또한, 생성된 규칙과 기계학습 모델을 결합한 하이브리드 탐지 구조를 통해 상호 보완적인 탐지를 수행한다. 실험 결과, 제안 방법은 높은 공격 탐지 성능과 함께 해석 가능성과 안정성을 동시에 확보할 수 있음을 확인하였다.

2. 관련 연구

기계학습 기반 침입 탐지 시스템은 다양한 공격 패턴을 효과적으로 탐지할 수 있다는 장점으로 널리 연구되어 왔다[1-2]. RandomForest, SVM, 딥러닝 기반 모델 등은 높은 분류 성능을 보이지만, 모델의 의사결정 과정을 해석하기 어렵다는 블랙박스 문제가 존재한다[4]. 이러한 한계로 인해 실제 보안 정책으로의 적용에는 제약이 따른다.

이를 보완하기 위해 규칙 기반 탐지 방식이 함께 사용되어 왔다. 규칙 기반 시스템은 명확한 정책 형태로 해석 가능하고 제어가 용이하다는 장점이 있으나, 사전에 정의된 규칙에 의존하기 때문에 새로운 공격 패턴에 대한 일반화 성능이 제한적이다. 최근에는 기계학습 모델과 규칙 기반 방식을 결합하는 하이브리드 접근이 제안되고 있으나, 대부분 경험적 기준에 따라 규칙을 설정하거나 모델의 출력값을 단

순 임계값으로 변환하는 방식에 머무르고 있다.

한편, 설명가능 인공지능(XAI)은 기계학습 모델의 의사결정을 해석하기 위한 방법으로 주목받고 있으며[5], SHAP과 같은 기법을 통해 각 피처가 예측에 미치는 영향을 정량적으로 분석할 수 있다. 그러나 기존 연구에서는 이러한 설명 정보를 분석에 활용하는 수준에 그치며, 이를 실제 보안 정책 생성으로 연결하는 연구는 제한적이다. 본 논문은 XAI 기반 설명 정보를 활용하여 정책을 자동으로 생성하고, 이를 기계학습 모델과 결합하는 프레임워크를 제안한다는 점에서 기존 연구와 차별성을 갖는다.

3. SHAP 기반 설명가능 정책 생성 및 하이브리드 탐지 기법

본 논문에서는 설명가능 인공지능 기반 정책 생성과 하이브리드 탐지를 결합한 침입 탐지 프레임워크를 제안한다. 전체 구조는 기계학습 모델, 설명 기반 피처 중요도 분석, 임계값 기반 규칙 생성, 그리고 하이브리드 탐지로 구성된다.

먼저 입력 데이터 X와 라벨 y를 이용하여 RandomForest 기반 분류 모델 $M(x)$를 학습한다. 이후 SHAP를 활용하여 각 피처의 중요도를 계산하고, 이를 기반으로 피처의 우선순위를 결정한다. 다음으로, 상위 피처에 대해 입계값 기반 규칙을 생성한다. 각 규칙 $r_j(x)$는 수식 (1)과 같이 정의된다.

$$r_j(x) = 1\left(\bigwedge_{((f,op,\theta)\in C_j} (x_f \; op \; \theta) \right), op \in \{>, <\} \quad (1)$$

여기서 C_j는 규칙을 구성하는 피처-연산자-임계값의 집합을 의미하며, 단일 조건 또는 다중 조건으로 구성될 수 있다. 전체 규칙 기반 탐지 함수 $R(x)$는 수식 (2)와 같이 정의된다.

$$R(x) = 1\left(\bigvee_{j=1}^{K} r_j(x) = 1 \right) \quad (2)$$

즉, 여러 규칙 중 하나라도 만족할 경우 공격으로 판단한다. 각 임계값 θ는 수식 (3)과 같은 제약 최적화 문제를 통해 결정된다.

$$\theta^* = \arg\max_\theta TPR(\theta) \;\; s.t. \;\; FPR(\theta) \le \alpha \quad (3)$$

이를 통해 오탐률을 일정 수준 이하로 유지하면서 공격 탐지율을 최대화하는 규칙을 생성할 수 있다. 마지막으로, 기계학습 모델과 규칙 기반 탐지를 결합한 하이브리드 탐지 함수는 수식 (4)와 같이 정의된다.

$$H(x) = M_{attack}(x) \vee R(x) \quad (4)$$

여기서 $M_{attack}(x)$는 모델이 입력 x를 공격으로 분류한 경우를 의미하며, 이준 분류에서는 직접적인 공격 여부를, 멀티클래스 분류에서는 정상이 아닌 클래스로 분류된 경우를 의미한다. 이와 같은 구조를 통해 모델과 규칙이 상호 보완적으로 작용하여 전체 탐지 성능을 향상시킬 수 있다.

4. 탐지 성능 및 설명가능성 평가

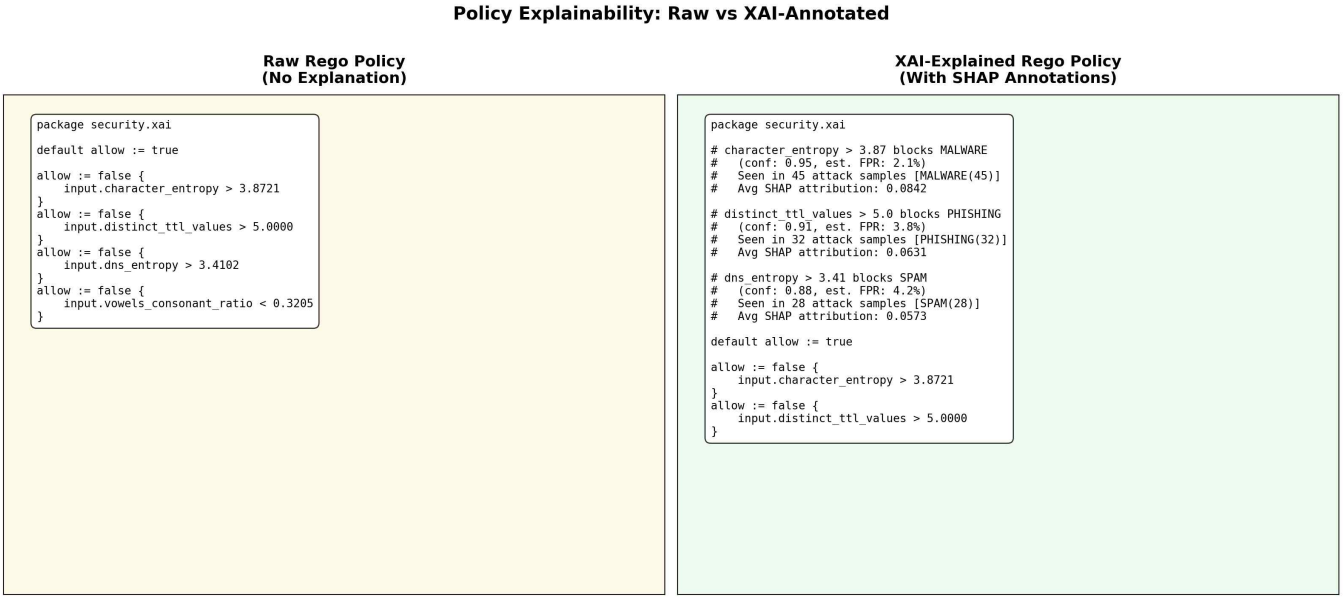
본 실험에서는 CIC-IDS2017 데이터셋의 flow 기반 CSV 데이터를 사용하여 제안 기법의 성능을 평가하였다. 총 8개의 CSV 파일을 활용하였으며, 각 파일에서 정상 및 다양한 공격 트래픽이 포함된 데이터를 사용하였다. 데이터 전처리는 수치형 피처와 DNS 구조 기반 피처를 포함하여 구성하고, 이후 상수 피처, 저분산 피처, 그리고 높은 상관관계를 가지는 중복 피처를 제거하여 데이터 누설을 방지하고, 모든 피처는 StandardScaler를 통해 정규화하였다.

모델은 RandomForest 분류기를 사용하였으며, 설명가능 인공지능 기법으로 SHAP을 적용하여 피처 중요도를 계산하였다. 이를 기반으로 주요 피처를 선정하고, 각 피처에 대해 임계값 기반 규칙을 생성하였다. 임계값은 오탐률(FPR)이 사전에 정의된 값 α 이하를 만족하는 조건에서 공격 탐지율을 최대화하도록 선택하였다.

평가는 within-file holdout 방식으로 수행되었으며, 각 실험은 3개의 서로 다른 random seed에 대해 반복 수행되었다. 성능 평가지표로는 공격 탐지율(Detection Rate)과 오탐률(False Positive Rate)을 사용하였으며, 규칙 기반 탐지(Rule-Only), 모델 기반 탐지(Model-Only), 그리고 두 방법을 결합한 하이브리드 탐지(Hybrid)의 성능을 각각 측정하였다.

4.1 이진 분류 결과 분석

표 1은 이진 분류(정상과 공격) 성능 비교 결과를 나타낸다. Model-Only는 높은 탐지율과 매우 낮은 오탐률을 보이며 안정적인 성능을 나타낸다. 반면, Rule-Only는 낮은 탐지율과 높은 오탐률을 보이며,



(그림 1) SHAP 기반 Feature Attribution을 적용한 XAI-Explained 정책과 기존 raw Rego 정책 비교

<표 1> 이진 분류 성능 비교(3 seeds)

방법	Attack Recall (Block Rate)	FPR	F1
Rule-Only	0.538	0.194	0.600
Model-Only	0.868	0.001	**0.930**
Hybrid	**0.983**	0.194	0.879

이는 단일 임계값 기반 규칙이 복잡한 공격 패턴을 충분히 표현하지 못함을 의미한다. 반면 Hybrid 구조는 탐지율 0.983을 달성하여 이는 Model-Only 대비 11.5% 향상된 성능으로, 모델이 포착하지 못한 공격 패턴을 규칙이 보완적으로 탐지함으로써, 탐지 영역이 확장되었음을 의미한다.

4.2 다중 분류 결과 분석

표 2 와 표 3은 다중 분류 성능을 나타낸다. 전체 정확도는 0.942, macro-F1은 0.735로 나타났으며, 이는 클래스 간 성능 편차가 존재함을 의미한다. DDoS와 WEB 공격은 각각 0.996, 0.9847의 재현률을 보이며 안정적으로 탐지되었다. 이는 해당 공격 유형이 특정 피처 공간에서 명확한 분포 특성을 가지기 때문으로 해석된다.

<표 2> 다중 분류 전체 지표(3 seeds)

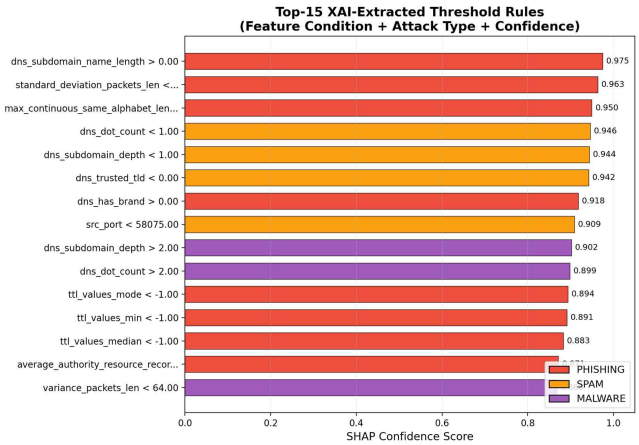
지표	값
Accuracy	**0.942**
Macro F1	0.735
Weighted F1	0.913

<표 3> 다중 분류 클래스별 Recall

클래스	Recall	Support
DDOS	**0.9996**	821
WEB Attack	**0.9847**	783
Normal	0.9999	2427

4.3 설명가능성 및 정책 특성 분석

그림 1은 생성된 정책의 설명 가능성을 시각적으로 비교한 결과를 나타낸다. Raw Rego 정책은 단순 임계값 기반 조건만 포함되어 있어 의미 해석이 어려운 반면, SHAP 기반 정책은 각 규칙이 어떤 공격 유형과 관련되는지를 제시한다. 이는 정책 수준에서



(그림 2) SHAP 기반 상위 15개 규칙의 피처 중요도 및 공격 유형 특성 분포

<표 4> 설명가능성 및 정책 복잡도 지표

지표	값
SHAP-Rule Overlap@15	**0.987±0.027**
Rule Feature Set Jaccard	0.743±0.066
Number of Rules	**33.0±3.5**
Avg. Conditions Per Rule	1.142±0.106

의 설명 가능성을 제공함으로써, 보안 운영자가 규칙의 의도를 이해하고 활용할 수 있음을 의미한다. 그림 2는 SHAP 기반으로 추출된 상위 규칙들의 중요도를 나타낸다. 각 규칙은 특정 피처 조건과 공격 유형 간의 관계를 반영하며, 높은 confidence 값을 통해 해당 규칙이 공격 탐지에 효과적으로 기여함을 확인할 수 있다. 이는 제안 기법이 데이터 기반으로 의미 있는 정책을 생성함을 뒷받침한다. 표 4는 설명가능성 및 정책 복잡도 지표를 나타낸다. SHAP 기반 피처 중요도와 규칙 간 정렬도(overlap@15)는 0.987로 매우 높은 값을 보였으며, 이는 생성된 규칙이 모델의 의사결정 구조에 잘 반영하고 있음을 의미한다. 또한 규칙 집합 간 안정성(Jaccard)는 0.743으로 나타나, 서로 다른 실험에서도 유사한 규칙이 반복적으로 생성됨을 확인하였다. 이는 제안 기법이 데이터에 대해 일관된 정책을 도출함을 시사한다. 정책 복잡도 측면에서는 평균 33개의 규칙과 규칙당 약 1.14개의 조건으로 구성되어, 단일 조건 기반의 단순한 구조를 유지하였다. 이는 생성된 규칙이 높은 해석 가능성과 함께 실제 보안 정책으로 적용 가능함을 의미한다.

5. 결론 및 향후 연구

본 논문에서는 기계학습 기반 침입 탐지 시스템의 블랙박스 문제와 규칙 기반 시스템의 낮은 일반화 성능을 해결하기 위해, SHAP 기반 설명가능 정책 생성과 하이브리드 탐지를 결합한 프레임워크를 제안하였다. 제안 기법은 SHAP을 활용하여 핵심 피처를 식별하고, 오탐률 제약 하에서 임계값 기반 규칙을 생성함으로써 해석 가능한 정책을 도출하였다. 또한, 생성된 규칙과 기계학습 모델을 결합한 하이브리드 구조를 통해 상호 보완적인 탐지를 수행하였다. 실험 결과, 제안 방법은 이진 분류에서 0.983의 높은 공격 탐지율을 달성하였으며, 다중 분류에서도 0.942의 정확도를 보였다. SHAP 기반 규칙 생성은 모델의 의사결정을 효과적으로 반영하며 높은 정렬도와 안정성을 보였고, 단순한 구조를 유지하면서도 실제 보안 정책으로 활용 가능한 수준의 해석 가능성을 제공하였다. 향후 연구에서는 탐지 성능 향상을 위해 다중 조건 규칙 확장, 시계열 기반 피처 활용, 그리고 정책 생성 과정에서의 적응적 피처 선택 기법을 적용할 예정이다. 또한, 실시간 환경에서의 정책 적용 및 성능 검증을 통해 실제 보안 시스템으로의 확장 가능성을 검토할 계획이다.

Acknowledgement

본 결과물은 농림축산식품부의 재원으로 농림식품기술기획평가원의 농식품과학기술융합형연구인력양성사업의 지원을 받아 연구되었음(RS-2024-00397026).(34%) 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2026-RS-2023-00256629)(33%)이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구임(IITP-2026-RS-2022-00156287)(33%)

참고문헌

[1] Z. A. El Houda, B. Brik and S. -M. Senouci, "A Novel IoT-Based Explainable Deep Learning Framework for Intrusion Detection Systems," in IEEE Internet of Things Magazine, vol. 5, no. 2, pp. 20-23, June 2022

[2] G. Kim, Y. Kim, E. Lee, H. Jang and K. Kim, "Edge-Based Policy Caching for Low Latency Security Enforcement in Hybrid Clouds," 2025 25th Asia-Pacific Network Operations and Management Symposium (APNOMS), Kaohsiung, Taiwan, 2025, pp. 1-6

[3] M. N. Sarwar, M. S. Arman, T. Bhuiyan and F. B. Rafiq, "Optimizing Intrusion Detection with Hybrid Deep Learning Models and Data Balancing Techniques," 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), Houston, TX, USA, 2025, pp. 1-6

[4] G. Kim, Y. Kim, T. Kim and K. Kim, "HCAPO: Transformer-Based Adaptive Policy Orchestration for Hybrid-Cloud Security," 2026 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Tokyo, Japan, 2026, pp. 1-6

[5] D. Manivannan, "Explainable AI-Enabled Intrusion Detection Systems for Computer Networks," 2025 IEEE 12th International Conference on Cyber Security and Cloud Computing (CSCloud), New York City, NY, USA, 2025, pp. 1-6