JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 8, pp. 2321-2330, Aug. 2026

Check for updates

의료 IoT 환경의 네트워크 데이터셋에서 SHAP 중요도의 통계적 재검증

임 동 성*
오산대학교 컴퓨터소프트웨어과 교수

Statistical Revalidation of SHAP Importance in Network Datasets from IoMT Environments

DongSung Im*

Professor, Department of Computer Software Engineering, Osan University, Osan 18119, Korea

[요 약]

SHAP과 같은 설명 가능 인공지능 기법이 의료사물인터넷(IoMT) 침입탐지에 활용되고 있으나, 이 기법이 제시하는 feature 중요도는 독립적인 통계적 근거로 검증되지 않은 채 쓰이는 경우가 많다. 본 연구는 클래스당 3,200건씩 균형을 이룬 5개 클래스로 구성된 CICIoMT-2024 데이터셋을 대상으로, SHAP 상위 10개 feature 전체에 대해 p-value와 Cohen's d를 함께 산출하고 LIME의 지역 설명과 교차 검증하였다. Rate, Max, Header_Length는 통계적으로 유의했으나 효과크기는 negligible로 나타나, 표본이 클수록 p-value가 무의미한 차이도 과대평가할 수 있음을 보여주었다. IAT는 negligible한 효과크기에도 SHAP·LIME 모두에서 지속적으로 상위로 채택된 반면, rst_count는 두 기법 모두에서 확인된 실질적인 large 효과를 보여, 두 기법의 합의만으로는 신뢰성을 보장할 수 없음을 실증하였다. 이는 IoMT 환경에서 탐지 결과의 판별 근거를 보안 관제 요원이 검증할 수 있는 형태로 제공하여, 설명 가능하고 신뢰할 수 있는 AI 기반 침입탐지 체계 구축에 실질적으로 기여할 수 있다.

[Abstract]

Explainable artificial intelligence techniques such as Shapley additive explanations (SHAP) are increasingly being applied in internet of medical things (IoMT) intrusion detection. However, the reported feature importances are often used without independent statistical validation. This study used the CICIoMT-2024 dataset comprising five balanced classes with 3,200 samples each to jointly compute the p-value and Cohen's d for every top-10 SHAP feature across all classes and cross-check the results against local interpretable model-agnostic explanations (LIME). Rate, Max, and Header_Length were statistically significant, but showed negligible effect sizes, demonstrating that large sample sizes could inflate p-values for practically meaningless differences. IAT was consistently ranked high by both SHAP and LIME, despite its negligible effect size, whereas rst_count showed a genuine large effect, as confirmed by both methods, demonstrating that their agreement alone could not guarantee reliability. This provides security analysts in IoMT environments with a verifiable basis for detection results, contributing to explainable and trustworthy AI-based intrusion detection.

색인어 : 의료 사물인터넷(IoMT), XAI, 침입탐지시스템, SHAP, LIME

Keyword : Internet of Medical Things (IoMT), XAI (eXplainable AI), Intrusion-Detection System, SHAP, LIME

http://dx.doi.org/10.9728/dcs.2026.27.8.2321

Received 25 July 2026; **Revised** 03 August 2026
Accepted 05 August 2026

***Corresponding Author; DongSung Im**

Tel: +82-31-370-2696
E-mail: seaids@osan.ac.kr

Ⅰ. 서 론

의료 사물인터넷(IoMT, Internet of Medical Things)은 인슐린 펌프, 심박동기, 환자 모니터링 장비 등 의료 현장의 각종 기기를 네트워크로 연결하여 진료와 관리의 효율을 높이는 기술로, 디지털 헬스케어 시장의 성장과 함께 그 보급이 빠르게 확대되고 있다[1]. 그러나 이러한 연결성 확대는 곧 공격 표면의 확대를 의미하며, 실제로 독일 뒤셀도르프의 한 대학병원에서 랜섬웨어 공격으로 병원 시스템이 마비되어 환자 치료에 직접적인 피해가 발생한 사례가 보고된 바 있다[2]. 이는 개별 사건에 그치지 않는데, 2023년 미국 보건복지부 보고서에 따르면 500명 이상에게 영향을 미친 대규모 의료 데이터 침해가 그해에만 732건 접수되었으며 이 중 해킹이 전체 사건의 81%를 차지하였다[3]. 이에 따라 IoMT 환경에 특화된 침입탐지 기술의 필요성이 지속적으로 제기되어 왔고, 최근에는 머신러닝 기반 탐지 모델을 적용하려는 연구가 활발히 이루어지고 있다[4].

다만 머신러닝 기반 탐지 모델은 예측의 근거를 스스로 드러내지 않는 블랙박스로 작동한다는 한계를 지니며, 이를 보완하기 위해 SHAP(SHapley Additive exPlanations)과 같은 XAI(eXplainable AI) 기법이 함께 활용되고 있다[5]. SHAP은 각 feature가 예측에 얼마나 기여했는지를 정량적으로 제시하지만, 이 값이 크다는 것은 모델이 해당 feature를 예측 과정에서 많이 참조했다는 의미일 뿐, 그 feature가 실제로 클래스 간에 통계적으로 유의미한 차이를 갖는다는 것을 보장하지는 않는다. 즉 SHAP이 제시하는 중요도와 feature의 실질적 판별력 사이에는 검증되지 않은 간극이 존재할 수 있다.

이러한 간극은 표본 규모가 클수록 더욱 두드러진다. 통계적으로 두 집단을 비교할 때 표준오차는 표본 크기의 제곱근에 반비례하여 줄어들기 때문에, 표본이 클수록 실질적으로 무시할 만한 수준의 차이조차 p-value상 통계적으로 유의하게 산출될 수 있다. 이 경우 p-value에만 의존하여 SHAP 결과의 타당성을 판단하면, 실제로는 판별력이 없는 feature를 중요한 것으로 오인할 위험이 있다. 한편 SHAP의 전역적 한계를 보완하기 위해 개별 샘플 단위의 지역 설명을 제공하는 LIME(Local Interpretable Model-agnostic Explanations)을 함께 적용하는 시도도 있었으나, 두 기법이 동일한 교란변수에 공통으로 의존할 경우 이러한 병행 적용만으로는 문제가 해결되지 않는다는 점이 기존 연구들을 통해서도 확인된다.

이에 본 연구는 클래스당 3,200건으로 균형을 이룬 CICIoMT-2024 데이터셋을 대상으로, SHAP과 LIME 사이에 Cohen's d 효과 크기를 독립적인 통계적 게이트키퍼로 배치하여 두 XAI 기법의 판단이 실제 통계적 근거를 동반하는지를 실증적으로 검증한다. 그 결과 p-value와 Cohen's d의 판정이 서로 어긋나는 사례와, SHAP·LIME이 합의했음에도 그 합의가 통계적으로 뒷받침되지 않는 사례를 확인함으로써, XAI 결과에 대한 통계적 재검증 절차가 왜 필요한지를 보여준다. 이러한 검증 결과는 SHAP과 LIME이 서로 합의했다는 사실만으로 XAI 결과를 신뢰할 수 없으며, Cohen's d와 같은 독립적인 통계적 검증이 반드시 병행되어야 함을 시사한다. 이는 IoMT 환경에서 탐지 결과의 판별 근거를 보안 관제 요원이 직접 검증할 수 있는 형태로 제공함으로써, 설명 가능하고 신뢰할 수 있는 AI 기반 침입탐지 체계를 구축하는 데 실질적으로 기여할 수 있다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 IoMT 침입탐지 및 XAI 관련 선행연구들을 검토하고 본 연구의 필요성을 도출한다. Ⅲ장에서는 CICIoMT-2024 데이터셋의 구성과 전처리 절차를 설명한다. Ⅳ장에서는 ML 평가, 경험적 분포 통계 검증, 설명 일치성 검증으로 이어지는 연구 방법론을 제시한다. Ⅴ장에서는 분류 성능과 SHAP·Cohen's d 재검증 결과를 다루고, Ⅵ장에서는 LIME 분석과 SHAP·Cohen's d와의 일치성을 분석한다. 마지막으로 Ⅶ장에서는 결론과 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

2-1 CICIoMT-2024 및 IoMT 침입탐지 연구

CICIoMT-2024는 IoTFlock 시뮬레이션 환경에서 다양한 IoMT 통신 프로토콜의 공격 트래픽을 수집하여 구축된 데이터셋으로, 33종의 세부 공격 유형과 정상 트래픽을 포함한다[6]. 이 데이터셋을 활용한 선행 연구는 주로 여러 머신러닝 알고리즘의 탐지 성능을 비교하는 데 집중되어 왔으며, 모델이 어떤 근거로 특정 클래스를 예측했는지를 설명하는 해석 가능성 측면은 상대적으로 다루어지지 않았다. 이러한 경향은 IoMT 침입탐지 연구 전반에서도 반복되는데, 초기 연구들은 대체로 분류 정확도 향상에 집중하였을 뿐 모델의 판단 근거를 검증하는 절차를 포함하지 않아, 예측 결과가 실제 공격 패턴을 반영한 것인지 아니면 데이터셋 고유의 우연적 상관관계를 반영한 것인지 구분하기 어렵다는 한계를 가지고 있다.

2-2 SHAP 기반 XAI 침입탐지 연구

이러한 블랙박스 문제에 대응하기 위해 SHAP[5]을 IoMT 침입탐지에 도입하는 연구가 이어져 왔다. SHAP은 협력게임이론의 Shapley 값을 기반으로 각 feature가 예측에 기여한 정도를 전역적으로 산출하는 기법으로, 클래스별로 독립적인 값을 산출하기 때문에 멀티클래스 환경에서도 클래스마다 feature의 기여 크기와 방향을 구분해서 볼 수 있다. Alani et al.[7]은 IoMT 침입탐지에 SHAP을 적용한 초기 연구 중 하나이나 이진 분류에 머물러 공격 유형별 해석에는 이르지 못하였다. Aljuhani et al.[8]은 앙상블·딥러닝 결합 모델에

SHAP 기반 feature 순위화를 적용하여 멀티클래스 탐지와 설명 가능성을 함께 달성하였으나, SHAP이 제시한 feature의 통계적 타당성을 별도로 검증하지는 않았다. Lui et al.[9]은 WUSTL 데이터셋에서 XGBoost와 SHAP을 결합하여 Spoofing 탐지에 기여하는 feature를 분석하였으나, 그 결과의 통계적 판별력에 대한 독립적 검증은 이루어지지 않았다.

SHAP의 전역적 설명은 데이터셋 전체에서 어떤 feature가 중요한지 파악하는 데는 유용하지만 개별 샘플 하나의 예측 근거를 설명하는 데는 한계가 있다. LIME[10]은 설명 대상 샘플 주변에 교란된 합성 이웃을 생성하고 그 국소 영역에서 해석 가능한 대리 모델을 학습시켜, 특정 샘플이 왜 그렇게 분류되었는지를 보여주는 모델 비종속적 기법으로 이 한계를 보완한다. Hosain and Çakmak[11]은 WUSTL-EHMS-2020에서 SHAP과 LIME을 결합하여 이진 분류 환경에서 높은 탐지 정확도를 달성하였으나, SHAP이 제시한 이상값을 통계적 검증 없이 곧바로 공격 신호로 해석하였으며, 이진 분류라는 특성상 개별 공격 유형 단위에서 이러한 해석이 실제로 성립하는지는 확인하지 못하였다. 이는 SHAP과 LIME을 함께 적용하는 것만으로는 두 기법이 공통으로 의존하는 교란변수를 걸러낼 수 없음을 시사한다.

2-3 통계적 Feature 재검증 분석

Cohen's d 효과크기[12]는 두 집단 간 평균 차이를 표본 크기와 무관하게 표준화하여 나타내는 지표로, 전통적으로 feature 선택이나 클래스 불균형 보정의 근거로 활용되어 왔다. 그러나 XAI가 제시한 feature의 신뢰성을 독립적으로 검증하는 수단으로 이를 채택한 연구는 드물다. 앞서 살펴본 선행연구들을 종합하면 공통된 한계가 확인된다. SHAP이 제시한 feature 중요도가 실제로 클래스 간 통계적 판별력을 갖는지 검증하지 않은 채 사용되었으며, SHAP과 LIME을 함께 적용한 경우에도 두 기법의 합의 자체가 신뢰성의 근거로 간주될 뿐 그 합의가 통계적으로 실재하는지는 확인되지 않았다. 본 연구는 이 공백을 메우기 위해 SHAP과 LIME 사이에 Cohen's d를 독립적인 게이트키퍼로 배치하여, 두 XAI 기법의 합의가 실제 통계적 근거를 동반하는지를 균형 표본 조건에서 직접 검증한다. 이러한 게이트키퍼 배치가 필요한 이유는 SHAP과 LIME이 모델이 feature를 얼마나 참조했는지를 측정하는 모델 충실성 지표일 뿐, 그 feature가 실제 데이터에서 클래스를 구분하는 근거를 갖는지는 측정하지 않기 때문이다. 최근 연구에서도 SHAP과 같은 feature importance 순위가 표본 추출에 따른 불안정성을 가지며, 이를 보정할 통계적 검증 절차가 필요하다는 점이 지적된 바 있다[13]. p-value는 이러한 불안정성 속에서 관찰된 차이가 우연인지 여부만을 판단할 뿐 그 차이의 실질적 크기는 알려주지 않으므로, 표본 크기와 무관하게 클래스 간 실제 분리 정도를 표준화하여 나타내는 효과크기가 모델 충실성과 데이터 근거성을 구분하는 독립적 기준으로 요구된다.

Ⅲ. 데이터셋 및 전처리

3-1 데이터셋 개요

본 연구는 캐나다 사이버보안연구소가 IoTFlock 시뮬레이션 환경에서 구축한 CICIoMT-2024 데이터셋을 실험에 활용한다. 이 데이터셋은 다양한 IoMT 통신 프로토콜에서 발생하는 33종의 세부 공격 트래픽과 정상 트래픽을 함께 제공한다. 본 연구에서는 서로 다른 통신 계층에 속한 4가지 공격 유형인 ARP_Spoofing, MQTT-DoS-Connect_Flood, Recon-Port_Scan, TCP_IP-DoS-SYN1과 Benign 트래픽을 포함하여 총 5개 클래스에 대해 실험을 진행하였다. 클래스별 표본 수 차이로 인한 통계 검정 왜곡을 막기 위해, 각 클래스에서 동일하게 3,200건씩 추출하여 16,000건 규모의 균형 데이터셋을 구성하였다.

3-2 CICIoMT-2024의 구성 및 특성

각 트래픽 샘플은 CICFlowMeter를 통해 산출된 45개의 플로우 통계 feature로 표현된다[14]. 여기에는 패킷 헤더 크기인 Header_Length, 전송 계층 플래그 발생 횟수인 syn_flag_number, fin_flag_number, ack_count 등과 패킷 크기 분포인 Min, Max, AVG, Std, 패킷 도착 간격인 IAT 등이 포함된다. 또한 프로토콜 유형이나 상위 계층 정보와 무관하게 모든 클래스에 동일한 feature 집합이 적용된다. 이 데이터셋의 가장 큰 특징은 클래스 간 표본 수가 균등하다는 점과, 일반 기업 네트워크가 아닌 의료기기 환경에서 발생한 트래픽 정보만으로 구성된 순수 침입 탐지용 데이터라는 점이다.

3-3 전처리 절차

원본 데이터는 각 공격 유형별로 별도 파일로 제공되며 feature 구성은 모두 동일하다. 우선 5개 클래스에서 각각 3,200건씩 무작위로 추출하여 클래스 간 표본 불균형을 제거하였다. 이어서 클래스 레이블을 정수로 인코딩하고, 값의 범위 차이가 큰 feature가 모델 학습에 왜곡된 영향을 주지 않도록 표준화 스케일링을 적용하였다. 예를 들어 Header_Length는 수십만 단위인 반면 syn_flag_number는 0에서 1 사이 값을 가진다. 전체 표본은 각 클래스의 비율을 유지한 채로 학습용 80퍼센트와 평가용 20퍼센트로 분할하였으며, 그 결과 학습 데이터 12,800건과 클래스당 2,560건, 평가 데이터 3,200건과 클래스당 640건으로 구성되었다.

이렇게 구성된 균형 표본 구조는 Ⅴ장에서 다룰 SHAP 재검증 분석의 전제 조건이 된다.

Ⅳ. 연구 방법론

제안 모델의 전체적인 파이프라인은 그림 1과 같이 크게 세 단계로 구성된다. 먼저 원본 데이터를 전처리하여 모델을 학습하고 평가하는 ML 평가 단계, SHAP이 제시한 feature의 실제 판별력을 통계적으로 재검증하는 경험적 분포 통계 검증 단계, 마지막으로 전역 설명인 SHAP과 지역 설명인 LIME이 서로 일관되는지를 확인하는 설명 일치성 검증 단계이다.

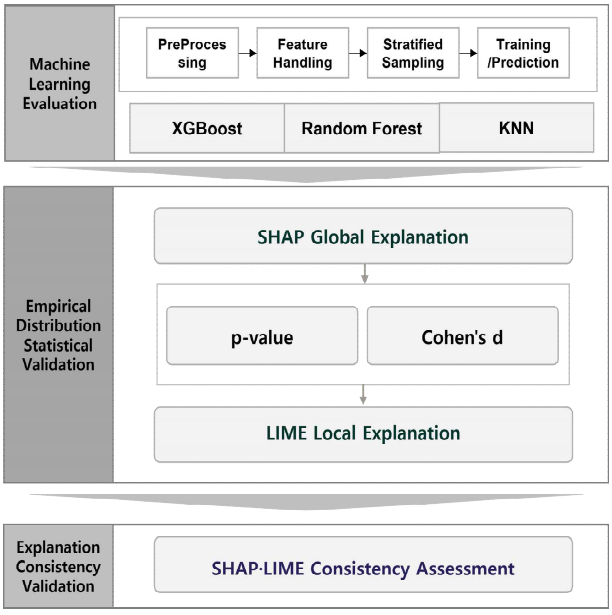

그림 1. 제안하는 통계적 재검증 파이프라인
Fig. 1. Proposed statistical re-validation pipeline

4-1 ML 평가

원본 데이터는 전처리, feature 정제, 클래스 비율을 유지한 표본 추출을 거쳐 학습 및 예측 단계로 이어진다. 전처리 과정에서 클래스 레이블은 정수로 인코딩하고, feature 값은 표준화 스케일링을 적용하여 값의 범위가 큰 feature와 작은 feature 간 척도 차이가 모델 학습에 왜곡된 영향을 주지 않도록 하였다. 이 단계에서는 XGBoost[15], Random Forest[16], KNN[17] 세 모델을 공통으로 적용한다. 순차적 부스팅, 배깅 앙상블, 비모수 거리 기반이라는 서로 다른 계열의 모델을 함께 비교함으로써, 이후 SHAP 및 LIME 분석에 사용할 기준 모델을 분류 성능에 근거하여 선정한다. 분류 성능은 클래스 간 표본 수가 균등한 조건에서 각 클래스를 동등하게 평가하는 Macro F1-score를 주 지표로 사용한다.

4-2 경험적 분포 통계 검증

ML 평가에서 가장 우수한 성능을 보인 모델에 SHAP TreeExplainer를 적용하여 클래스별 상위 feature를 추출한다. SHAP은 협력 게임 이론의 Shapley value에 기반하여 각 feature가 예측에 기여한 평균적 크기를 산출하는 전역 설명 기법이다. 다만 SHAP 값이 높다는 것은 모델이 해당 feature를 예측 과정에서 많이 참조했다는 의미일 뿐, 그 feature가 실제로 클래스 간 통계적 판별력을 갖는다는 것을 보장하지 않는다.

이에 본 연구는 SHAP이 제시한 feature 각각에 대해 p-value와 Cohen's d를 함께 산출하는 경험적 분포 검증을 수행한다. p-value는 두 그룹 간 관찰된 평균 차이가 우연히 발생했을 가능성을 나타내고, Cohen's d는 그 차이가 표본 크기와 무관하게 실제로 얼마나 큰지를 나타낸다. 표본이 클수록 p-value는 사소한 차이에도 유의하게 산출되는 경향이 있으므로, p-value만으로 feature의 중요성을 판단할 경우 표준오차가 충분히 줄어든 조건에서 실질적 판별력이 없는 feature를 과대평가할 위험이 있다. 따라서 p-value로 통계적 유의성을 먼저 확인하고, Cohen's d로 효과크기가 통상적 기준인 d값 0.2 미만에 미달하는 feature를 negligible로 판정하여 별도로 표시하는 2단계 절차를 사용한다.

4-3 설명 일치성 검증

경험적 분포 통계 검증을 거친 feature를 대상으로 LIME을 적용하여 개별 샘플 단위의 지역 설명을 산출한다. LIME은 특정 샘플 주변에서 분류기를 국소적인 선형 모델로 근사하는 기법으로, LimeTabularExplainer를 사용하여 클래스별 고신뢰 샘플에 대한 상위 feature를 추출한다.

마지막으로 SHAP의 전역 순위와 LIME의 지역 순위를 비교하여 두 설명 기법의 일치 여부를 판정한다. 상위 feature가 겹치면 Match, 일부만 겹치면 Partial Match, 전혀 겹치지 않으면 Mismatch로 분류한다. 특히 Cohen's d로 negligible 판정을 받은 feature가 SHAP과 LIME 양쪽에 공통으로 등장하는 경우는, 통계적 재검증 절차가 없었다면 두 설명 기법 모두에서 신뢰할 만한 근거로 오인될 수 있었던 사례로 다룬다.

Ⅴ. 분류 성능 및 SHAP·Cohen's d 재검증

5-1 분류 성능 비교

전처리를 마친 데이터셋에 대해 XGBoost, Random Forest, KNN 세 모델을 각각 학습하고 평가하였다. Macro F1-score 기준으로 XGBoost가 0.9782로 가장 높은 성능을 보였고, Random Forest가 0.9766으로 근소한 차이로 뒤를 이었으며, KNN은 0.9318로 두 모델 대비 상대적으로 낮은 성능을 나타냈다. 트리 기반 앙상블 모델인 XGBoost와

Random Forest가 거리 기반 모델인 KNN보다 우수한 성능을 보인 결과는, feature 간 비선형 상호작용을 포착하는 능력이 분류 성능에 영향을 미친다는 점을 시사한다. Ⅳ장에서 설명한 선정 기준에 따라 가장 우수한 성능을 보인 XGBoost를 이후 SHAP 및 LIME 분석의 기준 모델로 사용하였다.

세 모델의 클래스별 Precision, Recall, F1-score는 표 1과 같다. 특히 KNN은 전체 클래스에 고르게 낮은 성능을 보이는 것이 아니라 ARP_Spoofing과 Benign 두 클래스에서 F1-score가 각각 0.8505와 0.8436으로 나머지 세 클래스의 0.97 이상과 뚜렷하게 대비된다. 같은 두 클래스에서 XGBoost와 Random Forest는 0.94 이상을 유지하는 것과 비교하면, KNN이 ARP_Spoofing과 Benign 트래픽을 서로 혼동하는 경향이 두드러지며 이는 거리 기반 모델이 비선형 상호작용을 포착하는 능력이 상대적으로 낮다는 앞에서의 설명을 뒷받침하는 구체적 근거가 된다.

표 1. AI 모델에 따른 클래스별 분류 성능
Table 1. Classification performance results by AI model

Model	Class	Precision	Recall	F1-Score
XGBoost	ARP_Spoofing	0.9376	0.9625	0.9499
	Benign	0.9680	0.9453	0.9565
	MQTT-DDoS-Connect_Flood	0.9984	1.0000	0.9992
	Recon-Port_Scan	0.9875	0.9859	0.9867
	TCP_IP-DDoS-SYN1	1.0000	0.9969	0.9984
	Macro avg	0.9783	0.9781	0.9782
Random forest	ARP_Spoofing	0.9331	0.9594	0.9461
	Benign	0.9589	0.9469	0.9528
	MQTT-DDoS-Connect_Flood	1.0000	1.0000	1.0000
	Recon-Port_Scan	0.9921	0.9781	0.9851
	TCP_IP-DDoS-SYN1	1.0000	0.9984	0.9992
	Macro avg	0.9768	0.9766	0.9766
KNN	ARP_Spoofing	0.7975	0.9109	0.8505
	Benign	0.8864	0.8047	0.8436
	MQTT-DDoS-Connect_Flood	1.0000	0.9969	0.9984
	Recon-Port_Scan	0.9934	0.9453	0.9688
	TCP_IP-DDoS-SYN1	0.9969	0.9984	0.9977
	Macro avg	0.9349	0.9313	0.9318

표 1의 클래스별 F1-score를 모델 간 비교가 용이하도록 그림 2에 시각화하였다.

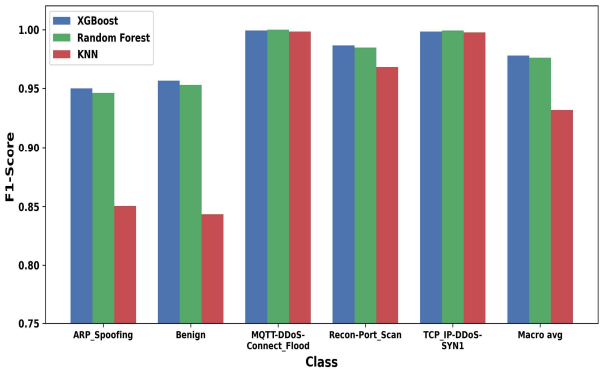

그림 2. 클래스와 모델에 따른 F1-Score 비교
Fig. 2. F1-score comparison by class and model

그림 2에서 다시 확인되는 바와 같이 MQTT-DDoS-Connect_Flood, Recon-Port_Scan, TCP_IP-DDoS-SYN1 세 클래스에서는 세 모델의 F1-score가 모두 0.96 이상으로 큰 차이 없이 밀집되어 있는 반면, ARP_Spoofing과 Benign 두 클래스에서만 KNN의 막대가 뚜렷하게 낮게 나타난다. 이러한 결과는 KNN의 성능 저하가 전체 클래스에 걸친 일반적 현상이 아니라 이 두 클래스에 국한된 문제임을 명확히 보여주며, ARP_Spoofing·Benign 간 feature 공간의 부분적 중첩이라는 해석과 일치한다.

표 1의 수치를 오분류 패턴 차원에서 보다 구체적으로 확인하기 위해, 가장 우수한 성능을 보인 XGBoost의 Confusion Matrix를 그림 3에 제시하였다.

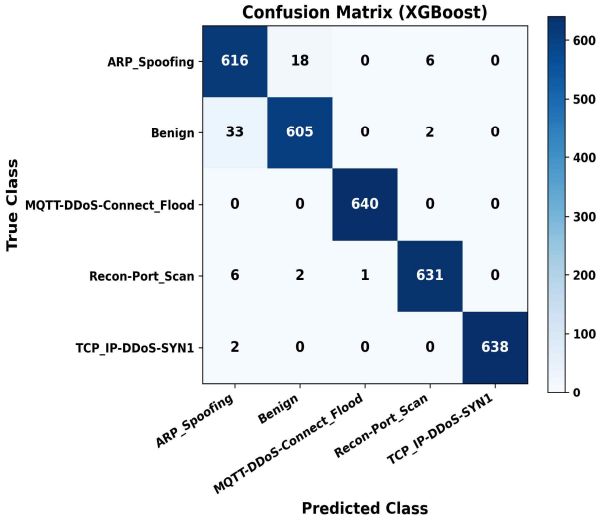

그림 3. XGBoost 모델의 confusion matrix
Fig. 3. Confusion matrix of the XGBoost model

그림 3에서 확인되는 바와 같이, XGBoost의 전체 오분류 70건 가운데 51건이 ARP_Spoofing과 Benign 사이의 상호 혼동에서 발생하여 전체 오류의 73퍼센트를 차지한다. Benign

표 2. p-value 및 Cohen's d 판정 불일치·일치 사례 요약
Table 2. Summary of p-value and Cohen's d agreement and disagreement cases

Class	Feature	p-value	Cohen's d	Verdict
ARP_Spoofing	Rate	0.001	0.082	Significant but negligible
Benign vs All Attacks	Rate	2.45×10^{-9}	−0.118	Significant but negligible
Benign vs All Attacks	Max	7.18×10^{-18}	0.170	Significant but negligible
Benign vs All Attacks	Header_Length	2.37×10^{-17}	−0.168	Significant but negligible
4 classes common	IAT	0.256–0.661	−0.028 to −0.011	Agreement, confounder

640건 중 33건이 ARP_Spoofing으로, ARP_Spoofing 640건 중 18건이 Benign으로 오분류되었으며, 나머지 세 클래스 간에는 산발적으로 1건에서 6건 수준의 오류만 나타났고 MQTT-DDoS-Connect_Flood는 오분류가 전혀 없었다. 이는 앞에서 확인한 KNN의 ARP_Spoofing·Benign 혼동과 동일한 축에서 반복되는 패턴으로, 두 클래스의 feature 공간이 모델 구조와 무관하게 일부 겹친다는 것을 보여준다. 다만 그 겹침의 정도는 모델에 따라 달라, KNN은 이 두 클래스에서 F1-score가 0.84~0.85까지 떨어진 반면 XGBoost는 0.95 이상을 유지하여, XGBoost의 비선형 분할 능력이 이 겹침의 상당 부분을 해소하되 완전히 제거하지는 못함을 시사한다. 또한 오분류 방향에서도 Benign이 ARP_Spoofing으로 오인되는 경우가 그 반대보다 많아, 모델이 정상 트래픽을 공격으로 오탐하는 경향이 상대적으로 두드러진다.

5-2 SHAP 상위 feature의 Cohen's d 재검증

Ⅳ장에서 제시한 절차에 따라 XGBoost 모델에 SHAP TreeExplainer를 적용하여 5개 클래스 각각에 대해 상위 판별 feature를 추출하고, 각 feature에 대해 p-value와 Cohen's d를 함께 산출하였다. 그 결과 다수의 feature가 p-value 기준으로는 유의하지만 Cohen's d 기준으로는 negligible에 해당하는 불일치 사례로 나타났다.

가장 뚜렷한 사례는 Benign 트래픽과 전체 공격 트래픽을 비교한 Rate, Max, Header_Length 세 feature이다. Rate는 p-value가 2.45×10^{-9}으로 극히 낮게 산출되었으나 Cohen's d는 −0.118로 negligible 수준에 그쳤다. Max 역시 p-value가 7.18×10^{-18}에 달할 만큼 통계적으로 유의했지만 Cohen's d는 0.170으로 작은 효과크기 기준인 0.2에도 미달하였다. Header_Length는 p-value 2.37×10^{-17}, Cohen's d −0.168로 같은 양상을 보였다. 세 feature 모두 클래스당 3,200건 조건에서 표준오차가 충분히 줄어든 상태로, 극히 작은 평균 차이조차 통계적 유의성으로 산출된 사례이며, p-value만으로 feature의 중요성을 판단할 경우 실질적 판별력이 없는 feature를 과대평가할 위험을 보여준다.

ARP_Spoofing 클래스에서도 유사한 양상이 확인되었다. Rate feature는 p-value 0.001로 유의하게 산출되었으나 Cohen's d는 0.082에 그쳐 negligible로 판정되었다. 이는 Benign 대 전체공격 비교에서 나타난 Rate의 과대평가 양상이 개별 클래스 단위 비교에서도 재현됨을 보여준다.

반면 IAT는 5개 클래스 전체에서 p-value와 Cohen's d가 모두 일치하는 대조 사례로 확인되었다. ARP_Spoofing, MQTT-DDoS-Connect_Flood, Recon-Port_Scan, TCP_IP-DDoS-SYN1 네 클래스에서 IAT의 Cohen's d는 −0.028에서 −0.011 사이로 모두 negligible에 해당하였고, p-value 또한 0.256에서 0.661 사이로 통계적으로 비유의하였다. 동일하게 표준오차가 충분히 줄어든 표본 조건임에도 유의한 차이로 이어지지 않았다는 점에서, p-value와 Cohen's d가 서로 다른 결론을 내리는 앞선 세 feature와 달리 IAT는 두 지표가 일관되게 negligible을 가리킨다. 이는 CICFlowMeter의 플로우 집계 구조에서 기인하는 구조적 교란변수로 볼 수 있는 근거를 제공한다. 표 2는 앞서 살펴본 재검증 결과를 정리한 것이다.

표 2의 수치가 실제 분포에서도 확인되는지 살펴보기 위해, 판정이 어긋난 대표 사례인 Max와 판정이 일치한 대표 사례인 IAT를 대상으로 Benign과 전체공격의 분포를 그림 4에 비교하였다. 두 feature 모두 원본 값이 극단적으로 치우친 분포를 가지므로 로그 변환값을 기준으로 시각화하였다.

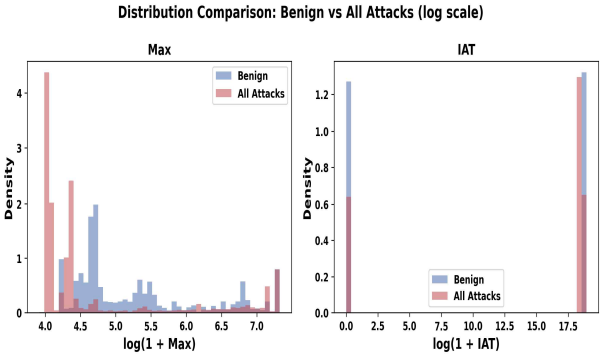

그림 4. Max 및 IAT의 Benign 대 전체공격 분포 비교
Fig. 4. Distribution comparison of Max and IAT between Benign and all attacks

그림 4를 보면 Max는 낮은 구간에서 두 그룹의 봉우리 위치가 다소 다르게 나타나지만 전체적으로는 넓은 구간에 걸쳐 상당 부분 겹쳐 있어, p-value상 유의함에도 실질적 효과

크기는 negligible이라는 판정과 부합한다. 반면 IAT는 두 그룹의 분포가 거의 동일한 두 지점에 집중되어 있어, p-value와 Cohen's d가 함께 negligible로 일치하는 양상을 시각적으로도 뒷받침한다. 다만 그림은 로그 변환된 값을 기준으로 하며, 표 2의 Cohen's d는 원본 스케일 값으로 산출되었다는 점을 참고할 필요가 있다.

따라서 위의 재검증 결과는 클래스당 3,200건 규모의 균형 데이터셋에서 p-value 단독 사용이 SHAP 결과의 신뢰성 해석에 왜곡을 줄 수 있음을 실증적으로 보여준다. 표본 크기가 표준오차를 충분히 줄이는 조건에서는 negligible한 효과 크기도 통계적으로 유의하게 산출될 수 있으므로, Cohen's d를 병행하지 않으면 이러한 왜곡을 걸러낼 수 없다. 5-1절의 Confusion Matrix에서 확인된 ARP_Spoofing과 Benign 간의 혼동 역시, 두 클래스의 feature 분포가 통계적으로 완전히 분리되지 않는 지점이 존재함을 시사한다는 점에서 본 절의 재검증 결과와 같은 방향을 가리킨다.

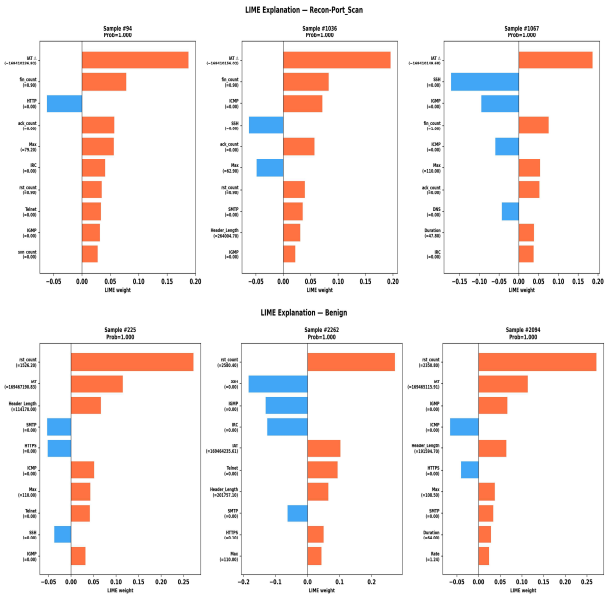

그림 5. Recon-Port_Scan(IAT)과 Benign(rst_count)의 LIME 설명 비교

Fig. 5. LIME explanation comparison: Recon-Port_Scan (IAT) and Benign (rst_count)

Ⅵ. LIME 분석 및 SHAP과의 일치성 분석

6-1 클래스별 LIME 상위 feature

각 클래스에서 예측 확률이 1.000인 고신뢰 샘플 3건씩을 선정하여 LimeTabularExplainer를 적용한 결과, 다음과 같은 공통 패턴이 관찰되었다. ARP_Spoofing에서는 IAT가 세 샘플 모두에서 1순위 또는 2순위로 나타났으며 부호는 일관되게 음수였다. Header_Length와 Max는 세 샘플 모두에서 양의 방향으로 반복 등장하였다. Recon-Port_Scan에서는 IAT가 세 샘플 모두에서 압도적인 1순위로 나타났으며, LIME weight가 나머지 feature보다 두 배 이상 컸다. TCP_IP-DDoS-SYN1에서는 Max가 세 샘플 모두에서 가장 큰 양의 기여를 보였고, IAT는 2순위 또는 그에 준하는 순위로 반복 등장하였다. MQTT-DDoS-Connect_Flood에서는 IAT가 세 샘플 중 두 건에서 1순위로 나타났다. Benign에서는 rst_count가 세 샘플 모두에서 1순위로 나타났으며, IAT는 2순위권에서 반복 등장하였다.

이 중 SHAP과 LIME이 가장 뚜렷하게 갈리는 두 사례를 시각적으로 비교하기 위해, IAT가 압도적 1순위로 나타난 Recon-Port_Scan과 rst_count가 1순위로 나타난 Benign의 LIME 설명을 그림 5에 제시하였다.

그림 5와 같이 Recon-Port_Scan의 세 샘플 모두 IAT가 나머지 feature보다 두 배 이상 큰 LIME weight로 최상위를 차지하는 반면, Benign의 세 샘플 모두에서는 rst_count가 동일한 양상으로 최상위를 차지한다. 두 feature 모두 LIME 단계에서는 동등하게 지배적인 근거로 나타나지만, 그 배후의 통계적 실체는 정반대이다.

6-2 SHAP·Cohen's d·LIME 일치성 분류

표 3은 Ⅴ장의 Cohen's d 재검증 결과를 기준으로 SHAP과 LIME의 상위 feature를 교차 정리한 것이다.

ARP_Spoofing의 Rate는 SHAP에서 3순위로 제시되었으나 세 샘플의 LIME 상위 10개 feature 어디에도 등장하지 않았다. SHAP과 LIME이 서로 다른 판단을 내린 Mismatch이며, Cohen's d 재검증에서도 negligible로 확인되어 LIME이 통계적으로 무의미한 feature를 결과적으로 걸러낸 사례

표 3. SHAP·Cohen's d·LIME 교차 검증 요약

Table 3. Cross-validation summary of SHAP, Cohen's d, and LIME

Class	Feature	SHAP Rank	Cohen's d Verdict	LIME Appearance	Classification
ARP_Spoofing	Rate	3	Negligible (confounder)	Not in top 10 (0/3 samples)	Mismatch
ARP_Spoofing / Recon / MQTT / TCP_IP	IAT	2, 3, 7, 2	Negligible (confounder, 4-class)	Top 1-2 (majority of samples)	Match
Benign vs All Attacks	Max, Header_Length	7, 9 (Benign)	Negligible (confounder)	Recurring mid-rank	Partial Match
Benign	rst_count	1	Large, d=1.43 (PASSED)	Top 1 (3/3 samples)	Match
Benign / MQTT / Recon / TCP_IP	Number	5, 4, 2, 7	Negligible (confounder, 4-class)	Not in top 10	Mismatch

이다. IAT는 다섯 클래스 SHAP top10에서 예외 없이 등장하고, LIME에서도 네 클래스 전부에서 예외 없이 상위 1~2순위를 차지하였다. SHAP과 LIME이 일관되게 합의하는 Match이다. 그러나 Cohen's d 재검증 결과는 negligible로, 두 XAI 기법이 사이좋게 합의한 신호가 실제로는 통계적 근거가 없는 교란변수였음을 보여준다. Max와 Header_Length는 Benign 대 전체공격 비교에서 SHAP 중위권으로 제시되었고, LIME에서도 세 샘플 중 일부 또는 중간 순위로만 등장하여 완전히 일치하지도 완전히 상충하지도 않는 Partial Match로 분류하였다. Benign의 rst_count는 SHAP 1위, LIME에서도 세 샘플 모두 1위로 나타나 명확한 Match이다. 이 경우 Cohen's d 재검증에서도 d=1.43의 large 효과로 PASSED되어, SHAP과 LIME의 합의가 통계적으로도 실제였음이 확인되었다. Number는 Benign, MQTT-DDoS-Connect_Flood, Recon-Port_Scan, TCP_IP-DDoS-SYN1 네 클래스의 SHAP top10에 반복적으로 등장하였으나 LIME 상위 10개 목록 어디에도 나타나지 않아 Mismatch로 분류하였다. Cohen's d 재검증에서도 negligible로 확인되어, LIME이 이 교란변수를 자연스럽게 걸러냈다는 판단이 통계적으로도 뒷받침된다.

앞에서 확인된 다섯 건의 feature-클래스 조합을 SHAP·LIME 일치성 기준으로 집계하면 표 4와 같다.

표 4. SHAP·LIME 일치성 분류 집계
Table 4. Summary counts of SHAP-LIME agreement classification

Classification	Count	Features
Match	2	IAT, rst_count
Partial Match	1	Max, Header_Length
Mismatch	2	Rate, Number

총 5건 중 Match와 Mismatch가 각각 2건으로 가장 많고 Partial Match가 1건으로 나타났다. 다만 이 집계에서 특히 주목할 점은 단순한 건수 비율이 아니라 Match로 분류된 두 사례의 성격이 정반대라는 것이다. IAT의 Match는 Cohen's d negligible 판정과 결합되어 통계적 근거가 없는 합의였던 반면, rst_count의 Match는 large 효과크기로 뒷받침되는 실질적 합의였다. 즉 SHAP·LIME 일치성 분류만으로는 이 두 Match 사례를 구별할 수 없으며, 이것이 다음 절에서 다루는 핵심 논점이다.

6-3 IAT와 rst_count로 본 SHAP·LIME 일치성의 이중적 함의

본 연구에서 가장 주목할 지점은 IAT와 Number가 Cohen's d 기준으로는 동일하게 negligible한 교란변수이면서도 SHAP·LIME 일치성 차원에서는 정반대의 결과를 보인다는 것이다. Number는 SHAP만 상위로 꼽고 LIME은 채택하지 않아 Mismatch로 끝났지만, IAT는 SHAP과 LIME 모두 상위로 채택해 Match로 수렴하였다. 만약 Cohen's d 재검증 없이 SHAP·LIME 일치 여부만으로 신뢰도를 판단했다면, 두 기법이 합의했다는 이유로 IAT를 rst_count와 동등하게 신뢰할 만한 근거로 오인했을 것이다. 그러나 실제로 rst_count의 합의는 large 효과 크기로 뒷받침되는 실질적 근거였고, IAT의 합의는 negligible 효과 크기 뒤에 가려진 통계적 허상이었다. 이는 SHAP과 LIME이 서로 합의한다는 사실 자체가 신뢰성의 근거가 될 수 없으며, Cohen's d와 같은 독립적인 통계적 검증이 반드시 병행되어야 함을 보여주는 본 연구의 가장 핵심적인 근거이다.

Ⅶ. 결 론

본 연구는 의료용 IoMT 네트워크 데이터셋인 CICIoMT-2024를 대상으로, XAI 기법인 SHAP이 제시하는 feature 중요도를 통계적으로 재검증해야 할 필요성을 실증적으로 제기하였다. SHAP은 모델이 예측 과정에서 어떤 feature를 얼마나 참조했는지는 보여주지만, 그 feature가 실제로 클래스 간 통계적 판별력을 갖는지는 보장하지 않는다. 이 간극은 표본이 클수록 두드러지는데, 클래스당 3,200건으로 구성된 균형 데이터셋에서는 표준오차가 충분히 줄어들어 negligible 수준의 사소한 효과 크기조차 p-value상 유의하게 산출될 수 있는 조건이 형성되기 때문이다.

본 연구는 5개 클래스의 SHAP 상위 10개 feature 전체, 총 50개 feature-클래스 조합에 대해 p-value와 Cohen's d를 함께 산출하였으며, Cohen's d 0.2 미만을 negligible로 판정하였다. 그 결과 Rate, Max, Header_Length는 negligible한 효과크기에도 p-value상 유의하게 산출되어 SHAP의 과대평가 위험을 확인하였다. 반면 rst_count는 두 지표 모두 뚜렷한 판별력을 가리켜, 재검증이 항상 SHAP을 기각하는 것은 아님을 보여주었다.

LIME을 통한 지역 설명 단계에서는 더 중요한 발견이 확인되었다. IAT는 4개 클래스의 SHAP과 LIME 모두에서 일관되게 상위로 채택되어 두 XAI 기법이 서로 합의하였으나, Cohen's d는 이 합의가 negligible한 효과 크기 뒤에 가려진 통계적 근거가 없는 것임을 보여주었다. 반면 같은 성격의 교란변수인 Number는 LIME 단계에서 자연스럽게 걸러져 Mismatch로 나타났다. 이는 SHAP과 LIME이 서로 합의한다는 사실 자체가 신뢰성의 근거가 될 수 없으며, 두 기법의 합의가 rst_count처럼 실제 효과 크기로 뒷받침되는 경우와 IAT처럼 그렇지 않은 경우를 두 기법만으로는 구별할 수 없다는 것을 보여준다. 이것이 Cohen's d를 SHAP과 LIME 사이의 독립적인 게이트키퍼로 병행해야 하는 근거이며, 본 연구가 제시하는 핵심 결론이다. 따라서 SHAP과 LIME이 합의했다는 사실만으로 그 설명을 신뢰하기보다, Cohen's d로 그 합의의 통계적 실체를 먼저 확인하는 절차가 XAI 기반 침입 탐지 결과의 신뢰성을 담보하는 데 필요하다. 이는 IoMT 환경에서 탐지 결과의 판별 근거를 보안 관제 요원이 검증할 수

있는 형태로 제공함으로써, 설명 가능하고 신뢰할 수 있는 AI 기반 침입탐지 체계를 구축하는 데 실질적으로 기여할 수 있을 것이다.

본 연구는 CICIoMT-2024 단일 데이터셋을 기반으로 수행되어 다양한 IoMT 환경에 대한 일반화에는 한계가 있다. 따라서 향후 실 환경에서 수집된 다양한 IoMT 데이터셋들을 추가하여, 본 연구 결과의 일반화 가능성을 검증할 것이다.

감사의 글

본 연구는 2026년 오산대학교 경기도 지역혁신중심 대학 지원체계(RISE) 구축·지원 사업 연구지원비에 의하여 수행된 연구임.

참고문헌

[1] Precedence Research. Internet of Medical Things Market Size and Growth [Internet]. Available: https://www.precedenceresearch.com/internet-of-medical-things-market.

[2] P. H. O'Neill. A Patient Has Died After Ransomware Hackers Hit a German Hospital [Internet]. Available: https://www.technologyreview.com/2020/09/18/1008582/a-patient-has-died-after-ransomware-hackers-hit-a-german-hospital/.

[3] U.S. Department of Health and Human Services Office for Civil Rights. Annual Report to Congress on HIPAA Privacy, Security, and Breach Notification Rule Compliance for Calendar Year 2023 [Internet]. Available: https://www.hhs.gov/sites/default/files/breach-report-to-congress-2023.pdf.

[4] A. Si-Ahmed, M. A. Al-Garadi, and N. Boustia, "Survey of Machine Learning Based Intrusion Detection Methods for Internet of Medical Things," *Applied Soft Computing*, Vol. 140, 110227, 2023. https://doi.org/10.1016/j.asoc.2023.110227

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30, pp. 4765-4774, 2017.

[6] S. Dadkhah, E. C. P. Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. A. Ghorbani, "CICIoMT2024: A Benchmark Dataset for Multi-Protocol Security Assessment in IoMT," *Internet of Things*, Vol. 28, 101351, December 2024. https://doi.org/10.1016/j.iot.2024.101351

[7] M. M. Alani, A. Mashatan, and A. Miri, "Explainable Ensemble-Based Detection of Cyber Attacks on Internet of Medical Things," in *Proceedings of the IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC/PiCom/CBDCom/CyberSciTech)*, Abu Dhabi, pp. 609-614, 2023. https://doi.org/10.1109/DASC/PiCom/CBDCom/Cy59711.2023.10361448

[8] A. Aljuhani, A. Alamri, P. Kumar, and A. Jolfaei, "An Intelligent and Explainable SaaS-Based Intrusion Detection System for Resource-Constrained IoMT," *IEEE Internet of Things Journal*, Vol. 11, No. 15, pp. 25454-25463, 2024. https://doi.org/10.1109/JIOT.2023.3327024

[9] P. H. Lui, L. P. Siqueira, J. F. Kazienko, V. E. Quincozes, S. E. Quincozes, and D. Welfer, "On the Performance of Cyber-Biomedical Features for Intrusion Detection in Healthcare 5.0," in *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2025)*, Porto Alegre, Brazil, pp. 389-400, 2025. https://doi.org/10.5753/sbcas.2025.7182

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference Knowledge Discovery and Data Mining (KDD)*, San Francisco: CA, pp. 1135-1144, 2016. https://doi.org/10.1145/2939672.2939778

[11] Y. Hosain and M. Çakmak, "XAI-XGBoost: An Innovative Explainable Intrusion Detection Approach for Securing Internet of Medical Things Systems," *Scientific Reports*, Vol. 15, 22278, 2025. https://doi.org/10.1038/s41598-025-07790-0

[12] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[13] J. Goldwasser and G. Hooker, "Statistical Significance of Feature Importance Rankings," in *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI),* Rio de Janeiro, Brazil, pp. 1476-1496, 2025.

[14] A. H. Lashkari, G. Draper Gil, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of Tor Traffic Using Time Based Features," in *Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP)*, Porto, Portugal, pp. 253-262, 2017.

[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco: CA, pp. 785-794, 2016. https://doi.org/10.1145/2939672.2939785

[16] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[17] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers: A Tutorial," *ACM Computing Surveys*, Vol. 54, No. 6, pp. 1-25, 2021. https://doi.org/10.1145/3459665

임동성(DongSung Im)

2015년 : 전남대학교 대학원 (이학석사)
2018년 : 전남대학교 대학원
(이학박사-정보보안)

1995년～2002년: LG전자
2002년～2011년: LG엔시스
2011년～2021년: 안랩
2022년～현 재: 오산대학교 컴퓨터소프트웨어과 교수
※관심분야 : IoMT, AI 보안, 정보보호관리체계, 네트워크 보안 등