피싱 사이트 탐지를 위한 설명 가능한 다중 뷰 앙상블 모델

홍혜원[1], 박지희[1], 전상훈[2*]
[1]수원대학교 정보보호학과 학생, [2]수원대학교 정보보호학과 교수

Explainable Multi-View Ensemble Model for Phishing Website Detection

Hye-won Hong[1], Ji-hee Park[1], Sanghoon Jeon[2*]
[1]Student, Department of Information Security, The University of Suwon
[2]Professor, Department of Information and Security, The University of Suwon

요 약 본 논문에서는 피싱 사이트의 고도화된 공격 방식에 효과적으로 대응하기 위해서, 다양한 출처의 특징을 그룹화한 멀티-뷰 앙상블 기반 피싱 탐지 모델을 제안한다. 총 113개의 특징을 URL, 도메인, 디렉토리, 파일, 파라미터, 네트워크 등 6개 그룹으로 나누고, 각 그룹별로 LightGBM 모델을 독립적으로 학습하고, 이를 Soft Voting 기법으로 결과를 통합하였다. 제안한 멀티-뷰 앙상블 모델은 실험 결과로 Accuracy 96.0%, F1-score 94.9% 등 모든 주요 지표에서 개별 모델 대비 우수한 탐지 성능을 보였다. 또한 SHAP 기반 설명 가능한 AI(XAI) 기법을 활용해, 특징별 기여도를 정량적 분석 및 시각화를 통해 모델의 해석 가능성을 확인하였다. 그리고, 외부 데이터셋에 대한 일반화 실험에서도 안정적인 탐지 성능을 유지하여 현실 적용 가능성과 확장성을 확인하였다. 본 연구에서 개발된 설명 가능한 다중 뷰 앙상블 모델 개발을 통해 피싱 사이트 탐지 및 예방에 기여하기를 기대한다.

주제어 : 피싱 탐지, 멀티-뷰 학습, 앙상블 모델, XAI, SHAP, LightGBM

Abstract This study proposes a multiview ensemble-based phishing detection model to effectively respond to sophisticated phishing website attacks. A total of 113 features were categorized into six groups: URL, domain, directory, file, parameter, and network. Each group was independently trained using the LightGBM. The results were then integrated using a soft voting mechanism. The proposed model demonstrated superior performance over the individual models, achieving an accuracy of 96.0% and an F1-score of 94.9%. In addition, the SHAP-based explainable AI (XAI) technique was employed to quantitatively analyze and visualize the contribution of each feature group, thereby enhancing model interpretability and reliability. Generalization experiments on external datasets confirmed a stable detection performance, demonstrating the model's applicability and scalability in real-world environments. We expect that the explainable multiview ensemble model developed in this study will contribute to the detection and prevention of phishing sites.

Key Words : Phishing detection, Multi-view learning, Ensemble model, Explainable AI, SHAP, LightGBM

본 연구는 원자력안전위원회의 재원으로 한국원자력안전재단의 지원을 받아 수행한 원자력안전연구사업의 연구결과입니다 (No. RS-2024-00403596).
*교신저자 : 전상훈(shjeon@suwon.ac.kr)
접수일 2025년 07월 12일 수정일 2025년 08월 01일 심사완료일 2025년 08월 12일

www.kci.go.kr

1. 서론

악성 URL(Uniform Resource Locator)을 활용한 사이버 위협이 지속적으로 발생하고 있어 다양한 연구들이 진행되고 있다. 최근 피싱 사이트 공격은 기존 단순 사칭 수준을 넘어, URL 패턴을 정상처럼 꾸미거나, 클릭 이후 특정 조건에서만 악성 행위를 수행하는 등 고도화되고 있다[1]. 한국인터넷진흥원에서 발간한 2022년 상반기 악성코드 은닉사이트 탐지 동향 보고서에 따르면 2021년 하반기 대비 악성코드 유포지가 38% 증가하였고, 전체 악성 URL 탐지 건수도 16% 증가하는 등 공격 규모가 꾸준히 증가하고 있다는 것을 확인할 수 있다[2]. 기존 연구들에서의 URL 기반 탐지 시스템은 URL 문자열의 길이나, 하이픈(-) 존재 여부 같은 정적 패턴에만 의존한다[3]. 그러나 실제 환경에서는 URL 통계 외에도 도메인 속성, 웹 트래픽, SSL 인증 등 다양한 출처의 특징이 함께 존재한다[4]. 이러한 이질적인 특징셋을 모두 반영할 수 있는 탐지 구조가 필요하다.

본 연구에서는 113개의 특징으로 구성된 데이터셋을 사용하여 URL, 도메인, 디렉토리, 파일 경로, 파라미터, 네트워크 등 기능별로 그룹화를 하였다. 각 그룹별 모델을 각각 학습시키고, 예측 결과를 멀티-뷰 앙상블 방식으로 결합하는 모델을 제안한다. 또한, 설명 가능한 인공지능 기법(eXplainable Artificial Intelligence, XAI)인 SHAP(SHapley Additive exPlanations)을 적용하여 각 뷰별·전체 모델의 결정 근거를 시각화하고, 해석력을 확보한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 정리하고, 3장에서는 본 연구의 데이터셋과 특징 그룹화, 멀티-뷰 앙상블 모델 설계 방법을 기술한다. 4장에서는 실험 결과 및 분석을 통해 제안한 시스템의 성능을 검증하며, 5장에서는 결론을 제시한다.

2. 관련 연구

악성 URL 탐지 연구에서는 오랫동안 URL 문자열의 통계적 특징을 중심으로 한 머신러닝 기반 접근이 활발히 이루어져 왔다. 김영준과 이재우[5]의 연구에서는 URL 데이터 분석을 통해 기존 연구에 반영된 URL 어휘적인 특징 이외에도 “URL Days”, “URL Words”, “URL Abnormal” 등 3종 9개 주요 특징을 추가로 제안하였고, Decision Tree, Random Forest, Extra Trees, Gradient Boosting의 네가지 머신러닝 알고리즘을 적용하여 F1-score, 정확도 지표로 성능을 측정하였다. 기존 연구와의 비교 분석에서 평균 0.9% 향상된 성능과 F1-score 및 정확도 지표에서 최고 98.5%가 측정됨에 따라, 해당 주요 특징이 정확도 및 성능 향상에 기여함을 확인하였다. 그러나 이 연구는 URL 기반의 정적 특징만을 분석 대상으로 삼고, 실제 피싱 환경에서 접속 후 발생하는 동적 행위 및 도메인·사이트의 속성 정보는 반영하지 않았다.

김대엽[6]의 연구에서는 4종류의 특징(URL lexical, host, contents, 3rd-party)과 다양한 기계학습 알고리즘(머신러닝, 딥러닝)을 추가로 활용하여 탐지 모델의 해석 범위를 확대하고, LIME과 SHAP 설명 가능 인공지능 기법을 활용하여 사용자 의사결정을 고려한 피싱 사이트 탐지 모델을 제안하였다. 그러나 전문지식이 없는 사용자가 각 특징의 구체적인 의미를 정확히 파악하기 어렵고, 영향력이 낮은 특징을 제거하는 최적화 과정이 부족하다는 한계가 있다.

본 연구에서는 이러한 한계를 보완하기 위해, 기능별 특징 그룹으로 데이터셋을 분할해 각 그룹별로 독립적인 모델을 학습하고, 예측 결과를 결합하여 탐지 성능을 향상시키는 멀티-뷰 앙상블 탐지 구조를 제안한다. 이를 통해 다양한 출처의 특징을 포괄적으로 반영하며, 최종 탐지 결과뿐만 아니라 뷰별 탐지 결과까지 설명 가능 인공지능 기법으로 해석함으로써 사용자 신뢰도와 해석 가능성까지 확보하는 것을 목표로 한다.

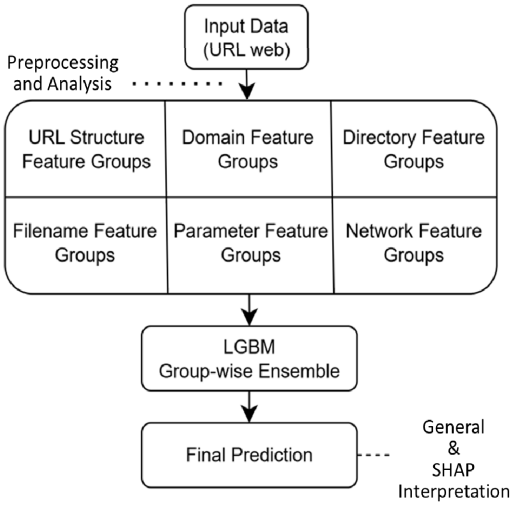
3. 제안 방법

본 논문에서는 기존의 단일 특징셋 기반 피싱 탐지 모델의 한계를 극복하고, 다양한 관점의 데이터 특성을 효과적으로 반영하기 위해 멀티-뷰 앙상블 기반 탐지 모델을 제안한다. 제안하는 방법은 원시 URL 데이터를 다각도로 분석하여, URL 구조, 도메인 속성, 디렉토리 구조, 파일명, 파라미터, 네트워크 등으로 구성된 여러 기능별로 6개의 특징 그룹을 정의한다. 각 특징 그룹별 모델로는 LightGBM(Light Gradient Boosting Machine) 기반의 분류 모델을 독립적으로 학습하였다. LightGBM은 Microsoft에서 개발한 트리 기반의 그래디언트 부스팅 기법으로, leaf-wise 트리 성장 방식을 사용하여 손실 감소가 가장 큰 리프 노드부터 확장함으로써 더 깊고 효율적인 트리 구조를 형성한다. 또한 히스토그램 기반 분

www.kci.go.kr

할 최적화를 통해 연산 속도와 메모리 효율성을 높이며, 결측치 및 범주형 변수 자동 처리, 병렬 학습 및 GPU 가속 지원 등의 기능을 제공함으로써 대규모 데이터셋과 고차원 특징 환경에 최적화되어 있다. 이러한 장점으로 인해 LightGBM은 최근 보안, 이상탐지, 악성코드 탐지, 네트워크 트래픽 분석 등 다양한 분야에서 표준 분류기로 널리 활용되고 있으며, 높은 예측 성능과 실무 적용성이 다수의 연구에서 입증되었다[7]. 이후, 각 모델의 예측 결과(확률값)를 통합하는 소프트 보팅(soft voting) 앙상블 기법을 적용함으로써, 개별 관점의 예측 정보를 효과적으로 결합하고자 한다. 최종적으로, 그룹별 SHAP 값을 활용하여 각 특징 그룹 및 주요 특징의 기여도를 정량적으로 분석하고, 이를 통해 모델의 해석력을 강화한다[8]. 또한, 학습 및 검증 데이터와는 별도의 외부 공개 데이터셋에 대해 동일한 전처리 및 특징 매핑, 이진화 과정을 거친 후 제안된 모델의 일반화 성능을 종합적으로 평가한다. 이러한 설계를 통해 실제 환경에서 발생할 수 있는 다양한 데이터 분포 변화와 공격 시나리오에도 강건하게 대응 가능한 탐지 시스템을 구현하고자 하였으며, 실무적 활용 가능성과 설명 가능성을 동시에 확보하는 것을 목표로 하였다.

[Fig. 1]은 본 연구의 멀티-뷰 앙상블 탐지 모델 전체 구조이다.



[Fig. 1] Overall Architecture of the Multi-view Ensemble-based Phishing Detection Model

3.1 데이터셋 구성 및 전처리

본 연구는 총 13만여 건의 샘플과 113개의 특징을 포함하는 dataset_cybersecurity_michelle.csv(dataset1)[9]을 사용하였다. 네트워크 및 SSL 관련 정보 등 여섯 개의 기능별 특징 그룹으로 분류하였다. 이러한 그룹화는 각 특징이 갖는 정보의 출처와 의미에 기반하여 논리적으로 분리함으로써, 각 뷰 모델이 해당 정보의 특성을 독립적으로 학습할 수 있도록 유도하기 위함이다.

〈Table 1〉은 각 특징 그룹의 구성 정보를 요약한 것이다. 결측치 제거 및 라벨 정렬 과정을 통해 모든 그룹은 동일한 103,758개의 샘플로 구성되어 있으며, 이는 멀티-뷰 학습 시 샘플 간 정합성과 일관된 평가를 보장하기 위함이다. 각 그룹별로 포함된 열(Column) 수 및 라벨을 제외한 순수 특징 수가 상이하다. 학습에는 각 샘플 내에서 해당 그룹에 해당하는 특징만을 추출하여 별도의 학습 및 검증 데이터셋을 구성하였다.

학습 및 검증 데이터는 Phishing 컬럼을 기준으로 라벨 비율이 유지되도록 stratified 방식으로 8:2 비율로 분할하였다. 원시 데이터에는 결측치 및 이상치가 거의 존재하지 않았으며, 간단한 정제 과정을 거쳐 모델 학습에 바로 사용하였다.

또한, 모델의 일반화 성능 및 데이터 분포 불일치 상황을 검증하기 위해, dataset1의 수치형 특징을 모두 이진화(binarization)한 dataset2를 별도로 생성하였다. 이진화는 각 특징의 값이 0이면 0, 0을 초과하면 1로 변환하거나, 분포의 중앙값 또는 도메인 지식을 기반으로 설정한 임계값을 기준으로 0과 1을 할당하는 방식으로 수행하였다. 생성된 dataset2는 학습에 사용되지 않으며, 학습된 멀티-뷰 앙상블 모델의 외부 테스트용으로만 활용하였다.

〈Table 1〉 Feature Group-wise Dataset Composition for Phishing Detection

Feature Group	Number of Rows	Number of Columns (including Label)	Number of Features (excluding Label)
URL	103,758	20	19
Domain	103,758	22	21
Directory	103,758	19	18
File	103,758	19	18
Params	103,758	21	20
Network	103,758	16	15

www.kci.go.kr

3.2 멀티-뷰 앙상블 설계

하나의 데이터셋은 동일한 샘플들로 구성되어 있으나, 각 샘플은 여러 종류의 특징을 가진다. 이 특징들을 각 기능별로 그룹화한 후, 그룹별로 독립적인 모델로 학습시키는 방식이 멀티-뷰 접근법이다[10]. 이는 상호 보완적인 여러 관점의 특징을 결합함으로써 단일 관점보다 학습 성능과 일반화 능력을 향상시킬 수 있음을 보여준다[11]. 이후 각 모델의 결과를 결합하여 최종 판단하는 구조인 멀티-뷰 앙상블 모델을 제안한다.

각 특징 그룹별로 LightGBM 기반의 분류 모델을 독립적으로 학습하였다. 하이퍼파라미터는 Optuna 기반 Bayesian Optimization을 통해 최적화하였으며, 교차검증(3-fold StratifiedKFold)과 조기 종료를 적용해 과적합을 방지하였다. 일반적으로 5-fold 또는 10-fold cross-validation이 널리 사용되지만, 본 연구에서는 특징 그룹 및 모델 조합이 많고, 실험 반복에 소요되는 시간과 리소스를 고려하여 3-fold StratifiedKFold를 채택하였다. 이는 실험 효율성과 일반화 성능 간의 균형을 고려한 결정이며, 각 fold별로 라벨 분포가 유지되도록 stratified 방식을 적용하였다.

학습된 그룹별 모델은 검증 데이터셋에 대해 확률 예측값을 생성하였고, 멀티-뷰 앙상블 모델을 활용하여 예측 확률을 단순 평균(soft voting) 하여 최종 예측을 수행하였다. 이러한 방식은 다양한 관점의 정보를 균형 있게 반영하여 모델의 과적합 위험을 감소시키고, 일반화 성능을 향상시킨다. 마지막으로, 평가 지표는 Accuracy, Precision, Recall, Specificity, F1-score, ROC AUC, PR AUC, Matthews Correlation Coefficient (MCC)를 사용하여 다각도로 모델 성능을 분석하였다.

3.3 환경 설정

본 연구는 Google Colab Pro 환경에서 실험을 수행하였다. 로컬 환경은 Windows 64비트 운영체제 기반의 시스템에서, 13세대 Intel Core i7-1360P 프로세서 (2.20GHz)와 32GB RAM을 사용하여 실험을 진행하였다.

4. 실험 결과 및 분석

제안한 멀티-뷰 앙상블 기반 피싱 탐지 모델의 성능을 다양한 관점에서 평가하였다. 개별 특징 그룹별 모델, 멀티-뷰 앙상블 모델, 이진화 테스트 모델, 일반화 성능을 비교하였다. 실험 결과는 〈Table 2〉에 요약되어 있다.

4.1 개별 특징 그룹 모델 성능 분석

〈Table 2〉의 URL, DOMAIN, DIRECTORY, FILE, PARAMS, NETWORK는 각 특징 그룹별로 학습된 개별 모델의 성능을 보여준다.

NETWORK 그룹은 Accuracy 0.9525, F1-score 0.9408 등 전반적으로 높은 성능을 보이며, 개별 모델 중 가장 뛰어난 성능을 기록하였다. 이는 네트워크 및 도메인 환경 기반 특징이 피싱 탐지에 매우 유의미함을 시사한다. URL과 DIRECTORY 그룹 또한 F1-score이 각각 0.8808, 0.8582로 높은 성능을 나타내어 전반적으로 균형 잡힌 성능을 보인다. FILE 그룹에서 Recall은 0.9593으로 매우 높지만, Precision이 0.7412로 낮아 오탐 가능성이 존재한다. DOMAIN 그룹은 전 지표가 낮으며, 탐지력과 정밀도 모두 부족하여 단독 사용엔 부적절하다. PARAMS 그룹은 Precision이 0.9526으로 매

〈Table 2〉 Performance Comparison of Feature Group-based and Binarized Models for Phishing Detection

Model / Dataset	Accuracy	Precision	Recall	Specificity	F1-score	ROC AUC	PR AUC
URL	0.9040	0.8793	0.8824	0.9186	0.8808	0.9690	0.9574
DOMAIN	0.7419	0.7016	0.6230	0.8218	0.6600	0.7927	0.7348
DIRECTORY	0.8867	0.8639	0.8526	0.9097	0.8582	0.9517	0.9386
FILE	0.8489	0.7412	0.9593	0.7747	0.8362	0.9132	0.8699
PARAMS	0.6817	0.9526	0.2194	0.9926	0.3567	0.6092	0.7455
NETWORK	0.9525	0.9420	0.9397	0.9611	0.9408	0.9868	0.9822
Multi-view Ensemble	0.9600	0.9609	0.9387	0.9743	0.9497	0.9923	0.9901
dataset1(Binarized)	0.8451	0.7312	0.9723	0.7596	0.8347	0.9075	0.8626
dataset2(Generalized)	0.8353	0.7328	0.8605	0.8209	0.7915	0.8277	0.7681

www.kci.go.kr

우 높지만, Recall이 0.2194로 탐지 성능은 낮았다.

4.2 멀티-뷰 앙상블 모델 성능 분석

개별 모델의 예측값을 Soft Voting 방식으로 통합한 멀티-뷰 앙상블 모델은 모든 주요 지표에서 최고 성능을 기록하며, 단일 모델을 압도하였다. 특히 Recall과 Precision이 균형 있게 높아, 실질적인 환경에서의 오탐, 누락 위험을 최소화한다. MCC와 AUC 계열 지표에서도 우수한 일반화 및 예측력을 확인하였다. 결과적으로, 멀티-뷰 앙상블은 개별 모델의 장점을 통합하여 가장 이성적인 성능을 보여준다.

4.3 이진화 모델의 학습 및 일반화 성능 분석

모델의 실질적인 적용 가능성과 일반화 능력을 검증하기 위해, 연속형 수치 특징으로 구성된 dataset1을 이진화 처리한 후 모델을 학습시켰다. 이후 학습 성능과 별도 도메인의 데이터셋에 대한 일반화 성능을 각각 비교하였다. Train에서 Recall이 0.9723으로 매우 높지만, Precision 0.7312, Specificity 0.7596으로 낮아 오탐률이 높은 경향이 있다.

일반화 실험은 web-page-phishing.csv(dataset2)[12]을 사용하였으며, 해당 데이터셋은 URL 기반 특징만 포함하고 있어 URL 단일 모델로 평가하였다. 결과적으로 Recall이 다소 낮아졌지만, Precision과 Specificity는 개선이 되었고, F1-scores는 비교적 안정적인 성능을 유지하였다. 이는 이진화된 모델이 타 도메인의 데이터셋에서도 일정 수준 이상의 탐지 성능을 유지하고 있음을 보여주며, 일반화 가능성이 충분함을 의미한다.

4.4 XAI 기반 특징 중요도 해석

모델 해석 가능성 확보를 위해, 각 특징 그룹별로 SHAP 값을 활용하여 특징 중요도를 분석하였다. SHAP는 예측 결과에 대한 각 입력 특징의 기여도를 정량적으로 산출하는 기법으로, 모델의 설명 가능성과 투명성을 확보하여 사용자의 해석 가능성과 신뢰도를 향상 시켜준다[13]. 이 방법은 여러 연구에서 모델 내부의 의사결정 과정을 설명하기 위한 정량적 XAI 도구로 자리잡았다[14]. 본 연구에서는 SHAP값의 평균 절대값을 기준으로 상위 10개 특징을 추출하고 분석하였다. 각 그룹별 분석 결과는 [Fig. 2]에 시각화되어 있으며, 다음과 같은 시사점을 도출할 수 있다.

URL 그룹은 qty_slash_url(2.64), length_url(1.85), qty_dot_url(0.75) 등의 특징이 높은 중요도를 보였다. 이는 URL의 구조적 깊이와 전체 길이가 피싱 여부 판단에 영향을 미친다는 것을 의미하며, 피싱 사이트가 일반적으로 길고 복잡한 URL 구조를 사용하는 경향과 일치한다[15]. 이러한 결과는 URL의 형식적 복잡성이 피싱 탐지의 강력한 신호로 작용할 수 있음을 시사한다.

DOMAIN 그룹에서는 qty_dot_domain(0.74), domin_length(0.39), qty_vowels_domain(0.34) 등이 주요 특징으로 보여졌다. 이는 도메인 내 문자 수, 형태, 모음 비율 등이 피싱 사이트 도메인 네이밍 특성과 연관됨을 의미한다.

DIRECTORY 그룹에서는 qyt_dot_directory(1.69), directory_length(0.99), qty_slash_directory(0.70) 등의 특징이 중요한 역할을 하였다. 디렉토리의 구조적 복잡성과 깊이가 탐지에 유의미하게 작용함을 나타낸다.

File 그룹에서는 qty_dot_file의 SHAP 값이 2.53으로 압도적으로 높았으며, 그 외 file_length, qty_hypen_file 등도 주요 특징으로 보여졌다. 이는 파일명에 포함된 특수기호와 확장자 관련 정보가 피싱 URL 식별에 핵심적인 역할을 한다는 것을 보여준다.

PARAMS 그룹에서는 qty_dot_params(0.69), params_length(0.67) 등 파라미터의 문자 수, 복잡성이 중요한 영향을 미쳤다. 파라미터 내의 특수문자 빈도도 일부 영향을 주는 것으로 나타났으며, 이는 파라미터에 삽입된 인젝션 또는 리디렉션 유사 형태의 공격 가능성과 관련이 있다.

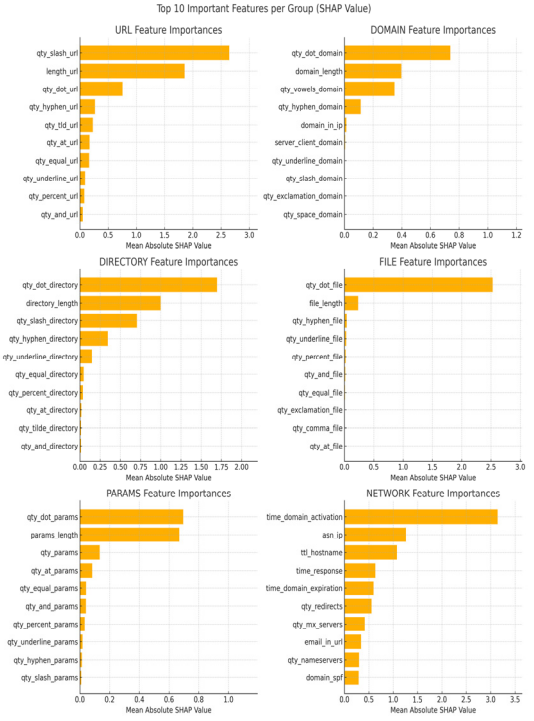
NETWORK 그룹은 time_domain_activation(3.14)이 전체 그룹 중 가장 높은 SHAP 값을 기록하였으며, asn_ip, ttl_hostname, time_response, time_domain_expiration 등 네트워크 지연 및 인증 관련 정보가 높은 중요도를 보였다. 이는 피싱 도메인이 일반적으로 운영 안정성이 낮고 갱신 주기가 짧은 비정상적인 특성을 가지며, 이러한 정보가 피싱 탐지에 핵심적으로 작용함을 보여준다.

SHAP 기반 분석을 통해 각 특징 그룹이 탐지 결과에 미치는 영향을 정량적으로 시각화함으로써, 단일 특징 기반 모델보다 멀티-뷰 앙상블 모델의 필요성과 효과를 수치적으로 입증하였다. 특히 URL 및 NETWORK 그룹의 특징은 전반적으로 높은 SHAP 값을 보여, 피싱 탐지에서 가장 핵심적인 역할을 수행함을 확인할 수 있었다.

기존의 URL 기반 피싱 탐지 연구에서는 URL 길이, 특수문자 빈도 등의 표면적 특성에 의존하여 탐지 성능 향상에 집중했지만, 모델의 의사결정 구조를 체계적이고

www.kci.go.kr

정량적으로 해석하려는 시도는 부족하였다[15]. 침입 탐지 분야에서도 설명 가능한 인공지능(XAI) 기법이 일부 적용되었으나, 주로 시각화 중심의 정성적 해석에 머무르며, 특징 기여도에 대한 정량 분석은 제한적이었다[13]. 이에 반해 본 연구는 SHAP 값을 활용하여 여섯 개의 독립된 특징 그룹별 기여도를 수치적으로 분석함으로써, 모델 예측의 해석 가능성과 투명성을 실질적으로 확보하였고, 이는 기존 연구들 대비 설명력과 실무 적용성 측면에서 차별적인 강점을 지닌다.



[Fig. 2] SHAP-based Feature Importance Visualization for Each Multi-view Group

4.5 종합 평가

개별 특징 모델은 특징의 특성에 따라 일부 그룹에서 우수한 성능을 보였지만, 그룹 간 성능 편차가 존재하였다. 이는 단일 특징에 의존할 경우 피싱 사이트의 복합적인 특성을 충분히 반영하기 어렵다는 한계를 보여준다.

이에 반해, 그룹별 모델을 결합한 멀티-뷰 앙상블 모델은 개별 모델 대비 모든 지표에서 성능이 향상되었으며, 실질적인 환경에 적용하기에 가장 적합한 구조임을 확인하였다. 또한, 이진화 모델과 일반화 실험에서도 안정적인 성능을 유지하였으며, 이는 제안된 모델이 실제 환경에서 새로운 데이터에 적용되었을 때 과도한 성능 저하 없이 일반화가 가능함을 보여준다.

마지막으로, SHAP 분석을 통한 XAI 해석 결과는 모델이 단순히 수치적 성능에 의존하지 않고, 논리적 판단 기반에서 설명 가능하고 신뢰할 수 있는 예측을 수행함을 입증하였다.

5. 결론

본 논문에서는 실제 환경에서 발생할 수 있는 다양한 데이터 분포 변화와 공격 시나리오에도 강건하게 대응 가능한 탐지 시스템을 구현하고자 하였으며, 실무적 활용 가능성과 설명 가능성을 동시에 확보하는 것을 목표로 하였다.

피싱 사이트 탐지의 성능과 해석 가능성을 동시에 향상 시키기 위해, 113개의 특징으로 이루어진 데이터셋을 각 특징의 특성에 따라 6개의 그룹으로 나눈 후, 각 그룹별로 개별 LGBM 모델을 학습한 후 Soft Voting 기반의 멀티-뷰 앙상블 구조를 제안하였다. 실험 과정에서는 개별 특징 모델과 앙상블 모델의 성능을 비교하였다.

실험 결과, 개별 특징 그룹 모델에서는 각 그룹별로 성능 편차가 존재하였고 멀티-뷰 앙상블 모델은 모든 지표에서 가장 우수한 성능을 보였으며, Precision과 Recall의 균형도 우수하여 실전 적용 가능성이 높음을 확인하였다. 또한, SHAP 기반의 특징 중요도 분석을 통해 각 그룹에서 탐지 결과에 크게 기여한 주요 특징들을 도출함으로써, 모델 해석 가능성도 확보하였다. 추가로, 연속형 특징을 이진화한 모델을 별도로 학습시켜 외부 도메인 기반의 dataset2에 대해 일반화 실험을 수행한 결과, 안정적인 성능을 유지함으로써 제안된 모델의 일반화 가능성과 확장성 또한 확인할 수 있었다.

다만, 각 특징 그룹별 모델의 성능 편차가 존재하고, 멀티-뷰 앙상블 방식에서도 그룹별 기여도에 따라 가중치를 차등 적용하는 전략은 반영되지 못하였다. 또한, 모델 학습 및 이진화 과정에서의 하이퍼파라미터 탐색 자동화나 성능 최적화 절차가 미흡한 한계가 있다. 향후 후속 연구에서는 가중치 기반 동적 앙상블 기법, 자동화된 특징 처리 및 하이퍼파라미터 튜닝 기법, 그리고 실제 서비스 환경에서의 적용 최적화 방안을 중심으로 모델의 성능과 실용성을 더욱 고도화할 예정이다.

www.kci.go.kr

REFERENCES

[1] S. Marchal, J. Francois, R. State, and T. Engel, "PhishStorm: Detecting phishing with streaming analytics," IEEE Trans. Netw. Serv. Manag., Vol. 11, No. 4, pp. 458-471, 2014.

[2] Korea Internet & Security Agency, "Detection trend report of malicious code hidden sites in the first half of 2022," KISA Report, 2022.

[3] C.-R. Han, S.-H. Yun, M.-J. Han, and I.-G. Lee, "A machine learning-based malicious URL detection method," J. Korea Inst. Inf. Secur. Cryptol., Vol. 32, No. 3, pp. 555-564, 2022.

[4] H. Doshi, Phishing websites features, M.S. thesis, Dept. of Computer Science, San Jose State Univ., pp. 1-45, 2012. [Online]. Available: https://core.ac.uk/reader/30732240

[5] Y. Kim and J. Lee, "Development of a malicious URL machine learning detection model reflecting the main feature of URLs," J. Korea Inst. Inf. Commun. Eng., Vol. 26, No. 12, pp. 1786-1793, 2022.

[6] D.-Y. Kim, "Interpretable phishing website detection model based on XAI (eXplainable Artificial Intelligence) considering user decision," J. Korea Multimedia Soc., Vol. 26, No. 8, pp. 1013-1026, 2023.

[7] Y. Zhang, Y. Lu, and X. Ma, "A novel phishing website detection model based on LightGBM and domain-name features," Sensors, Vol. 23, No. 8, p. 3981, 2023.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., Vol. 30, pp. 4765-4774, 2017.

[9] Michelle V. P., Dataset Phishing Domain Detection || CyberSecurity [Internet], Kaggle. Available: https://www.kaggle.com/datasets/michellevp/dataset-phishing-domain-detection-cybersecurity

[10] Z. Yu, Z. Dong, C. Yu, K. Yang, Z. Fan, and C. L. P. Chen, "A review on multi-view learning," Front. Comput. Sci., Vol. 19, No. 7, p. 197334, 2025.

[11] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," CoRR, abs/1304.5634, 2013.

[12] Shashwat, Web page Phishing Detection Dataset [Internet], Kaggle. Available: https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset

[13] Il-Ok Jung, Woo-Bin Choi, and Soo-Chul Kim, "Enhancement of intrusion detection reliability using explainable artificial intelligence (XAI)," J. Converg. Secur., Vol. 22, No. 3, pp. 101-109, 2022.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., Vol. 30, pp. 4765-4774, 2017.

[15] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in Proc. 15th ACM SIGKDD, pp. 1245-1254, 2009.

홍 혜 원(Hye-Won Hong) [준회원]

- 2023년 3월 ~ 현재 : 수원대학교 정보보호학과 학사

〈관심분야〉
보안 컨설팅, AI 보안, 금융 보안

박 지 희(Ji-Hee Park) [준회원]

- 2023년 3월 ~ 현재 : 수원대학교 정보보호학과 학사

〈관심분야〉
정보보호

전 상 훈(Sanghoon Jeon) [정회원]

- 2012년 2월 : 경북대학교 IT대학 심화 전자공학 공학사
- 2014년 2월 : 대구경북과학기술원 정보통신융합공학전공 공학석사
- 2020년 8월 : 대구경북과학기술원 정보통신융합전공 공학박사
- 2020년 3월 ~ 2020 8월 : 한양대학교 산학협력단 선임연구원
- 2020년 9월 ~ 2022 9월 : 한양대학교 의과대학 응급의학과 포닥연구원
- 2022년 10월 ~ 2023 9월 : 한양대학교 의과대학 응급의학과 연구조교수
- 2023년 10월 ~ 현재 : 수원대학교 지능형SW융합대학 정보보호학과 조교수

〈관심분야〉
웨어러블컴퓨팅, 의료인공지능, CPS보안

www.kci.go.kr