ICSR
Intelligent Cyber Security Research Lab

지능형사이버보안연구실

Explainable AI for Comparative Analysis of Intrusion Detection Models

논문세미나

발표자 정유진

발표일 2025.09.09

CONTENTS

목차

ICSR
Intelligent Cyber Security Research Lab

CONTENTS

ICSR
Intelligent Cyber Security Research Lab

1-1 Overview

■ 제목
- Explainable AI for Comparative Analysis of Intrusion Detection Models

■ 저자
- Pap M. Corea, Yongxin Liu, Jian Wang, Shuteng Niu, Houbing Song

■ 발표 학회
- IEEE MeditCom 2024-WS-05

■ 발표 날짜
- 14 Jun 2024

CONTENTS

02 Introduction

ICSR
Intelligent Cyber Security Research Lab

2-1 배경

- 머신러닝은 침입 탐지에서 새로운 위협을 높은 정확도로 탐지
- 한계
  - 통계적 머신러닝: NIDS 발전에 기여하였으나 과적합의 위험 존재
  - 딥러닝: 대규모 데이터에서 복잡한 패턴 학습, 높은 성능 달성 가능, 데이터 수집 비용이 크고 블랙박스 문제 존재
  - 블랙박스 모델: 내부 동작 원리가 직관적으로 이해되거나 해석되지 않는 모델 (입출력만 알 수 있음)
    - 설명 불투명, 규제 준수, 편향 교정 등의 어려움이 있음
- 실험: UNSW-NB15 데이터 셋을 기반으로 ML 모델을 비교해서 가장 좋은 성능을 가진 모델을 찾기
- UNSW-NB15 데이터 셋
  - UNSW Canberra - Cyber Range Lab
  - Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worm
  - 총 49개의 features
  - 총 2,540,044개의 레코드 수

The UNSW-NB15 Dataset
Intelligent Security Group ISG
UNSW Canberra, Australia
Dr Nour Moustafa
nour.moustafa@UNSW.edu.au

CONTENTS

03 Method

ICSR
Intelligent Cyber Security Research Lab

3-1 데이터 전처리 과정

- 불완전한 레코드 제거
  - 81,173개의 레코드 남음
- Categorical features:
  - one-hot encoding
  - 범주를 0과 1만 있는 벡터로 표현
- Scaling and Normalization:
  - re-scale numerical values -> [0,1]
- Feature Selection:
  - 라벨과의 상관계수가 0.3 보다 작은 feature remove
- Data Synthetic and Model Training:
  - training data - 80%, test data - 20%
- 학습 종료 조건:
  - 정확도 90% or 최근 epoch 대비 정확도 향상 1%미만

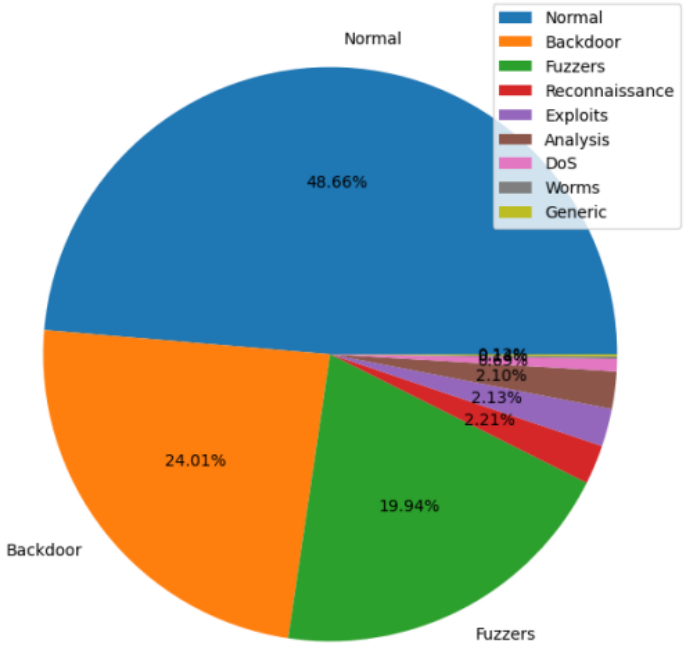

Fig. 1. Distribution of intrusion attack categories after data preprocessing.

3-2 이진 분류 분석

- 정상 트래픽(비침입), 비정상 트래픽(침입)을 구분하는 작업
- Scikit-learn 라이브러리 사용
- Linear Regression, Logistic Regression, Linear SVM, Random Forest, Decision Tree, MLP
- Figure. 2. Correlation을 보여주는 Matrix
- Figure. 3. 그중 쓸만한 feature들을 뽑음 -> ct_state_ttl, sttl 같은 연결 상태/TTL값 중요

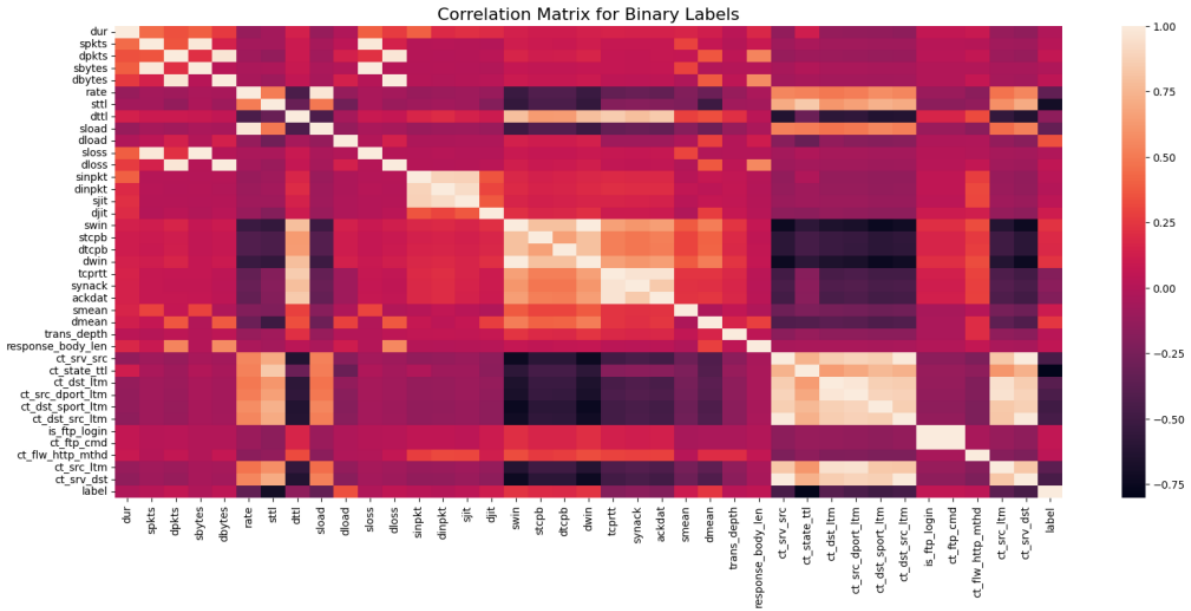

Fig. 2. Feature correlation matrix

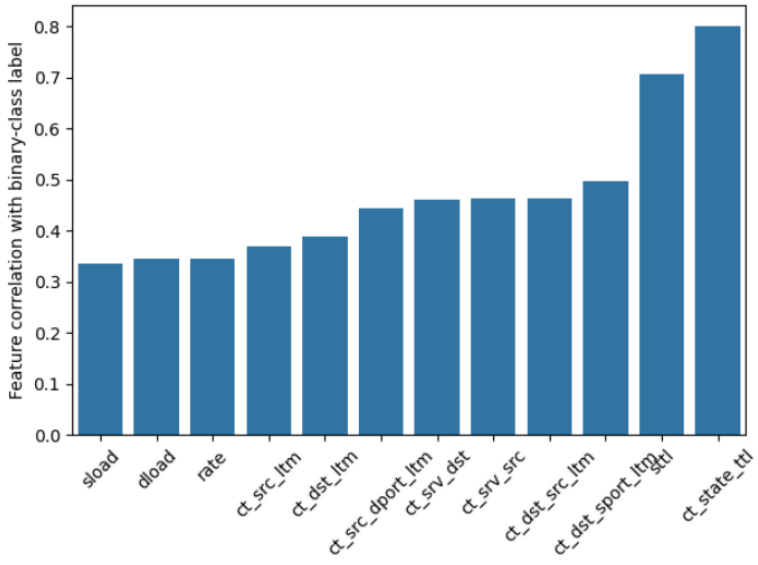

Fig. 3. Selected features for binary classifiers.

3-3 다중 분류 분석

- 동일한 머신러닝, 딥러닝 모델 사용
- 공격 유형별로 분류
- Occlusion sensitivity: 특정 feature를 가렸을 때 모델 성능이 얼마나 떨어지는가
- Figure. 4. 탐지 성능이 확 떨어지는 중요한 feature들을 뽑아냄.
- ctd, tcprtt, ackdat 등 세션 응답 시간, 윈도우, ACK 관련 feature 중요
- 공격 유형별 네트워크 동작 패턴이 다르기 때문
  - DoS - 응답 없음, Fuzzers - 비정상 핸드셰이크

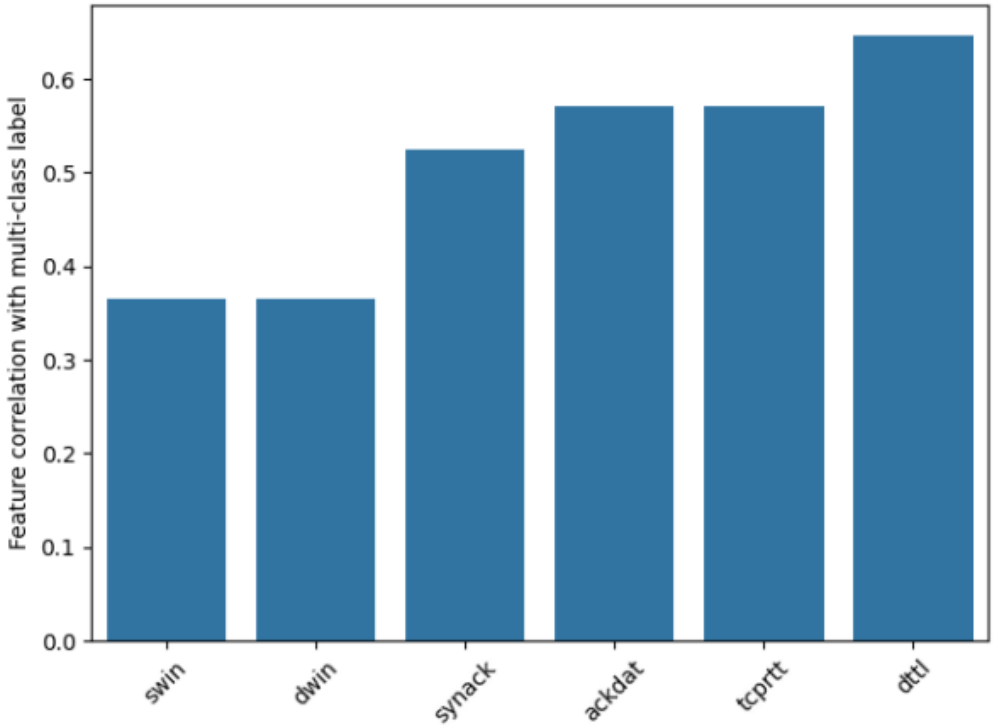

Fig. 4. Selected features for multi-class classifiers

CONTENTS

04 Experiment

ICSR
Intelligent Cyber Security Research Lab

4-1 이진 분류기 분석

- 대부분의 이진 분류 모델 -> 3개 이하의 주요 feature 의존
- Decision Tree: 단일 feature 제거에 취약
- MLP, Random Forest: 다양한 feature 활용

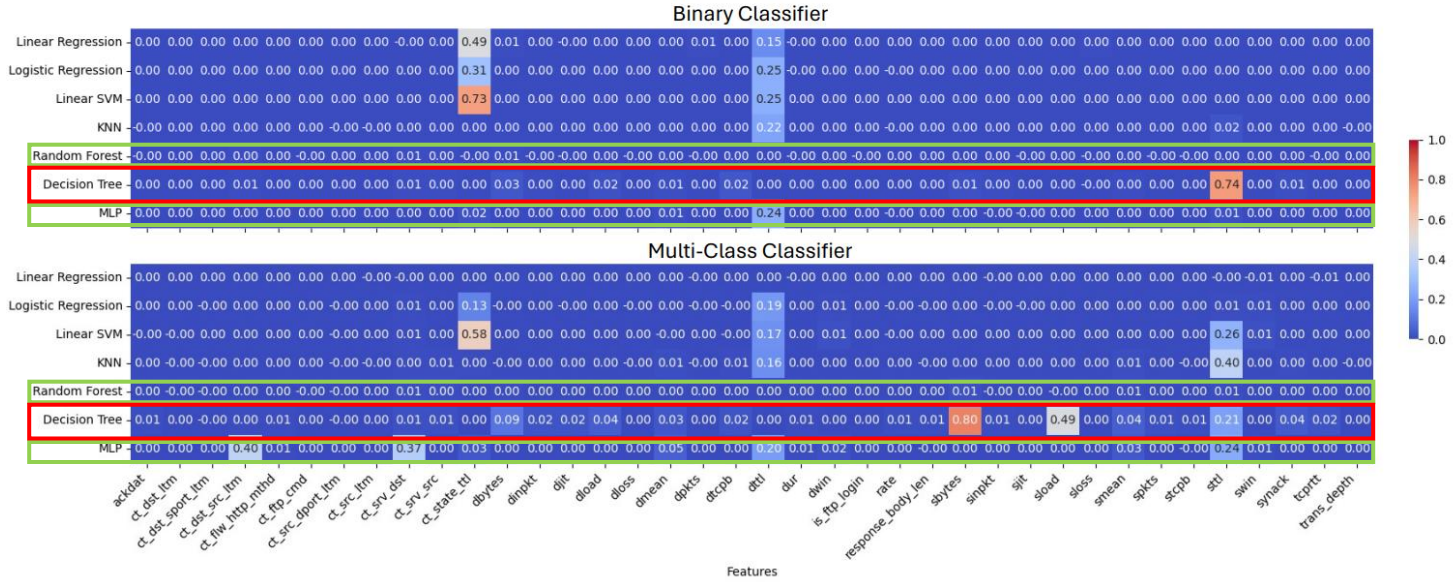

Fig. 5. Feature sensitivity of intrusion detection model classifiers trained with complete features.

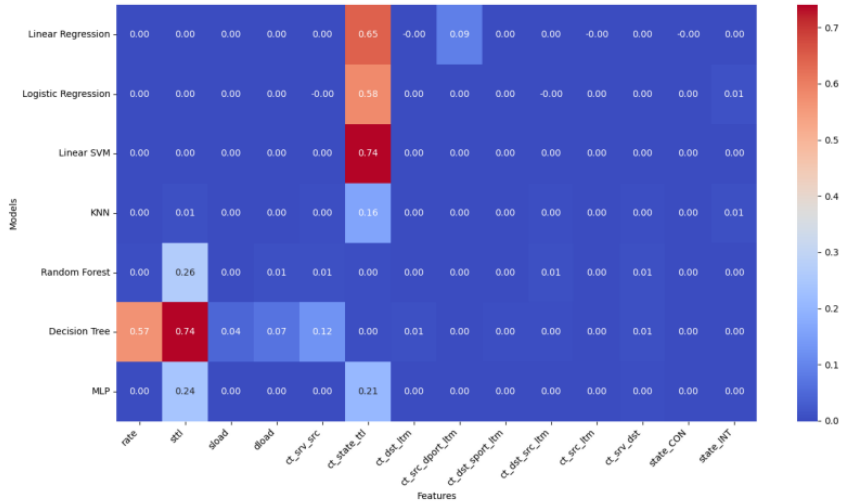

Fig. 6. Feature sensitivity of binary intrusion detection model classifiers trained with selected features.

4-1 이진 분류기 분석

- 상위 2개 feature만 마스킹
- Random Forest, KNN만 성능 저하가 거의 없다

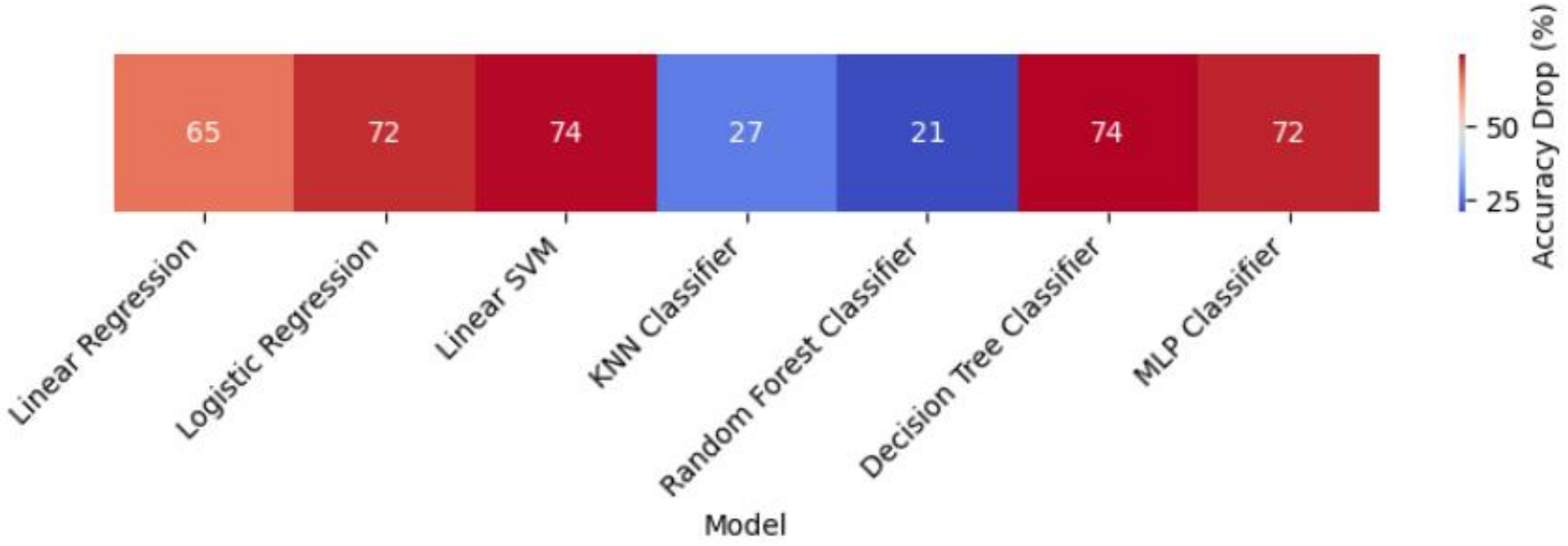

Fig. 7. Accuracy degradation after masking the Top-2 features.

4-2 다중 분류기 분석

- 다중 분류기 -> 더 많은 featur를 활용
- 하지만 feature selection -> 대부분 모델에서 TTL 관련 feature가 중요 - 재등장
- Random Forest: 단일 feature 마스킹에 거의 영향을 받지 X
- Decision Tree: 이진 분류보단 덜한 성능 하락
- Figure. 8. 각 모델이 어떤 feature에 민감한가

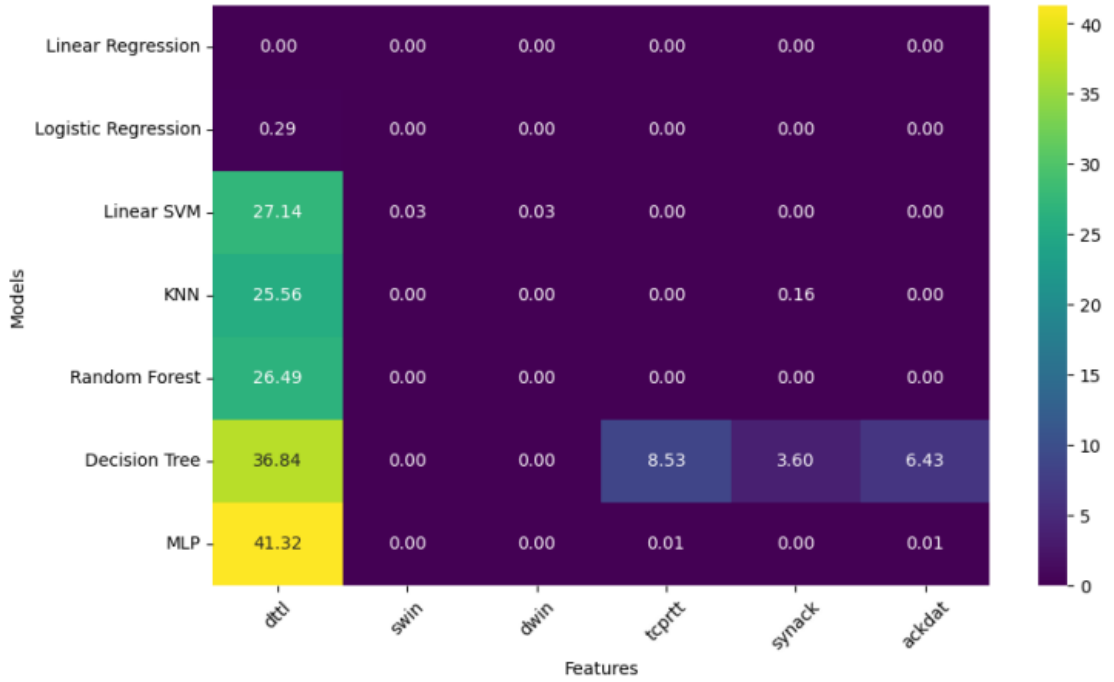

Fig. 8. Feature sensitivity w.r.t. classification accuracy of multi-class intrusion detection model.

4-2 다중 분류기 분석

- Figure. 9. 상위 feature(TTL 관련 등)를 제거 -> 모델 다시 학습 -> 어떤 feature에 민감한가
  - Decision Tree가 유난히 특정 feature에 의존
- Figure. 10. 가장 중요한 feature 2개 마스킹 -> 모델 정확도가 얼마나 떨어지는가
  - Logistic Regression, Linear SVM, MLP -> 정확도 41% 하락
  - KNN, Decision Tree -> 정확도 37~39% 하락
  - Random Forest -> 26% 하락

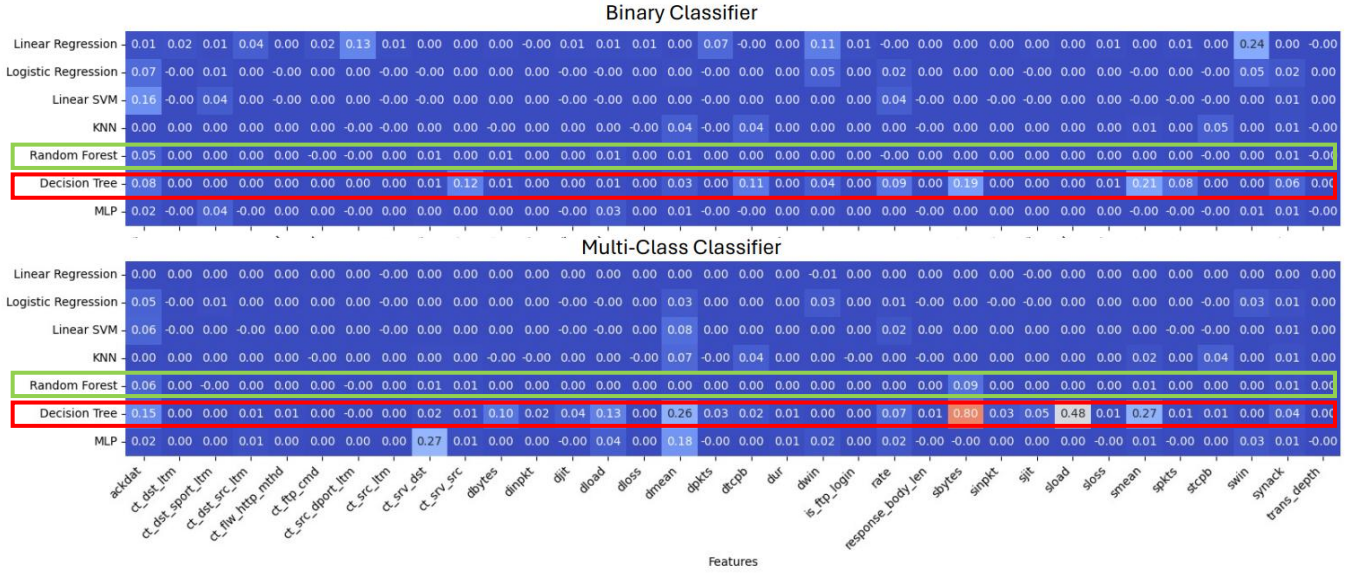

Fig. 9. Feature sensitivity of intrusion detection model classifiers trained without the top features.

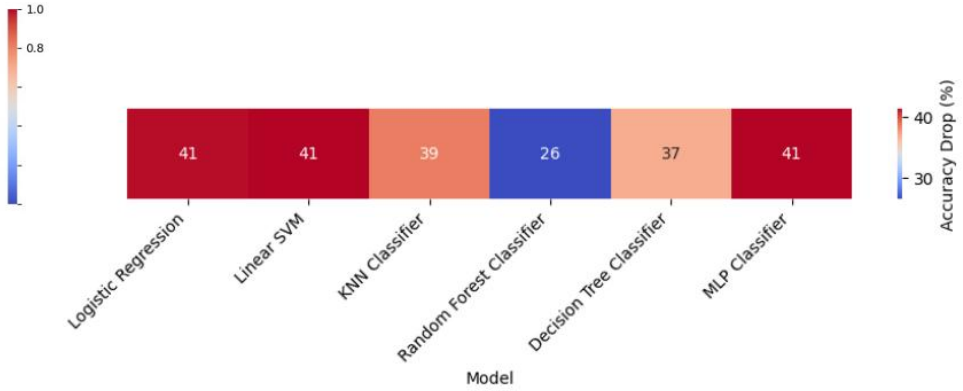

Fig. 10. Accuracy degradation after masking the Top-2 features

4-3 모델 오버헤드 비교

■ Fig. 11. 모델 별 훈련 시간과 Overhead(추론 시간) 비교한 막대그래프
- 각 모델이 training과 inferenc에 얼마나 시간이 걸리는지

■ Linear Regression/Logistic Regression
- 시간이 거의 안들어가지만 성능이 제한적

■ Linear SVM
- Training 시간이 상대적으로 큼 - 실시간 탐지에 부담

■ KNN
- Training 시간 거의 X (메모리에 저장), inference 시간 큼

■ Random Forest
- 둘 다 적당함

■ Decision tree
- 시간이 적고 빠르지만 특정 feature 의존 문제

■ MLP
- Overhead가 일단 너무 큼

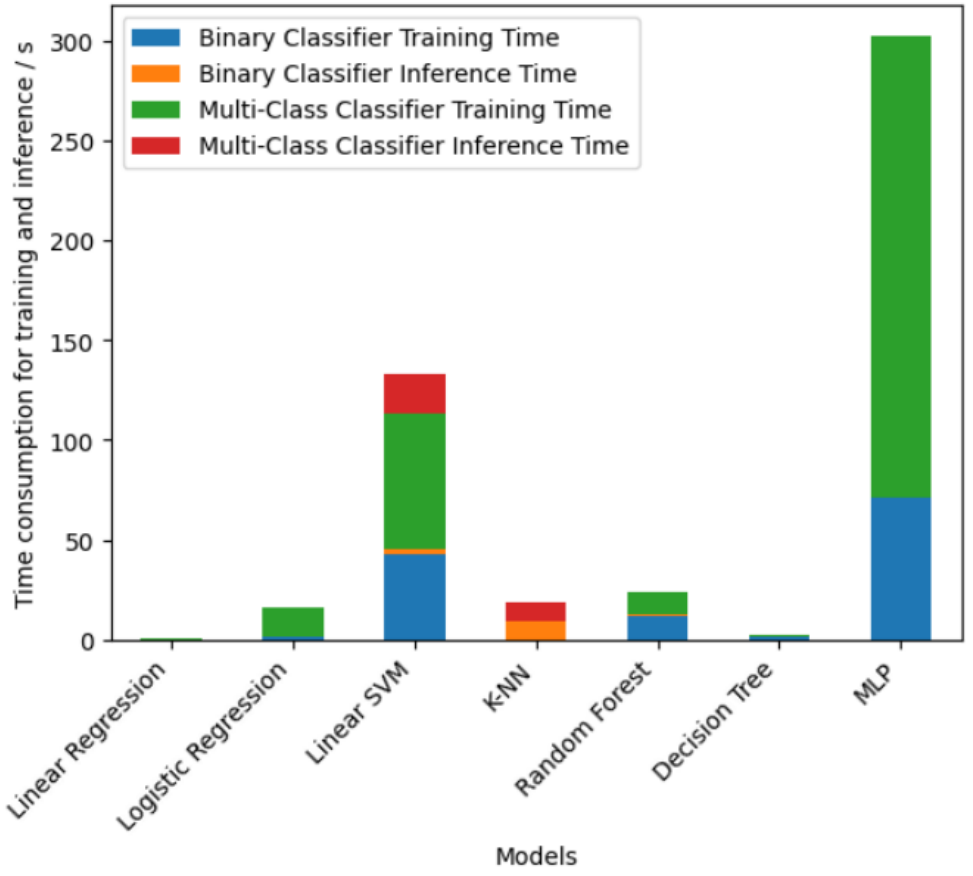

Fig. 11. Comparison of model overhead.

05 Conclusion

ICSR
Intelligent Cyber Security Research Lab

5-1 결론

- UNSW-NB15 데이터셋을 대상으로 머신러닝의 feature 중요도를 비교하기 위해 Occlusion Sensitivity 사용
- NIDS를 설계할 때 Random Forest가 가장 좋다
  - 특정 feature에 과도하게 의존 X
  - 성능 + 효율
    - MLP 같은 신경망보다 -> 비슷한 성능, 학습시간 짧음
  - 강인(robust)
    - 단일 feature 제거에도 크게 흔들리지 X