J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

PageRank 특징을 활용한 RDP기반 내부전파경로 탐지 및 SHAP를 이용한 설명가능한 시스템☆

RDP-based Lateral Movement Detection using PageRank and Interpretable System using SHAP

윤 지 영[1] 김 동 욱[2] 신 건 윤[2] 김 상 수[3] 한 명 묵[1*]
Jiyoung Yun Dong-Wook Kim Gun-Yoon Shin Sang-Soo Kim Myung-Mook Han

요 약

인터넷이 발달함에 따라 다양하고 복잡한 사이버공격들이 등장하기 시작했다. 공격들을 방어하기 위해 네트워크 외부에서 다양한 방식의 탐지 시스템들이 활용되었으나 내부에서 공격자를 탐지하는 시스템 및 연구는 현저히 드물어 내부에 들어온 공격자를 탐지하지 못해 큰 문제를 야기하기도 했다. 이를 해결하고자 공격자의 움직임을 추적하고 탐지하는 내부전파경로 탐지 시스템에 대한 연구가 등장하기 시작했다. 특히 그중에서도 Remote Desktop Protocol(RDP) 내 특징을 추출해 탐지하는 방식은 간편하면서도 매우 좋은 결과를 나타내었다. 하지만 그럼에도 불구하고 이전 연구들은 각 로그온 된 노드들 자체의 영향 및 관계성을 고려하지 않았으며, 제시된 특징 또한 일부 모델에서는 떨어지는 결과를 제공하기도 했다. 또한 왜 그렇게 판단했는지 판단에 대해 설명하지 못한다는 문제점도 존재했다. 이는 결과적으로 모델의 신뢰성 및 견고성 문제를 야기하게 된다. 이를 해결하기 위해 본 연구에서는 PageRank 특징을 활용한 RDP기반 내부전파경로 탐지 및 SHAP를 이용한 설명가능한 시스템을 제안한다. 페이지랭크 알고리즘과 여러 통계적인 기법을 활용해 여러 모델에서 활용 가능한 특징들을 생성하고 SHAP을 활용해 모델 예측에 대한 설명을 제공한다. 본 연구에서는 이전 연구에 비해 대부분의 모델에서 더 높은 성능을 보여주는 특징을 생성했고 이를 SHAP을 이용해 효과적으로 증명했다.

☞ 주제어 : 내부전파경로 탐지, 페이지랭크 알고리즘, 설명가능한 인공지능, 원격 데스트톱 프로토콜, 특징 추출

ABSTRACT

As the Internet developed, various and complex cyber attacks began to emerge. Various detection systems were used outside the network to defend against attacks, but systems and studies to detect attackers inside were remarkably rare, causing great problems because they could not detect attackers inside. To solve this problem, studies on the lateral movement detection system that tracks and detects the attacker's movements have begun to emerge. Especially, the method of using the Remote Desktop Protocol (RDP) is simple but shows very good results. Nevertheless, previous studies did not consider the effects and relationships of each logon host itself, and the features presented also provided very low results in some models. There was also a problem that the model could not explain why it predicts that way, which resulted in reliability and robustness problems of the model. To address this problem, this study proposes an interpretable RDP-based lateral movement detection system using page rank algorithm and SHAP(Shapley Additive Explanations). Using page rank algorithms and various statistical techniques, we create features that can be used in various models and we provide explanations for model prediction using SHAP. In this study, we generated features that show higher performance in most models than previous studies and explained them using SHAP.

☞ keyword : Lateral Movement, Pagerank Algorithm, Explainable AI, Remote Desktop Protocol, Feature Extraction

1. 서 론

인터넷이 발달함에 따라 다양하고 복잡한 사이버 공격들

1 Department of Software, Gachon University, Sungnam-si, 13120, Korea
2 Department of Computer Engineering, Gachon University, Sungnam-si, 13120, Korea
3 Agency for Defense Development Songpa P.O Box 132, Seoul, 05661 Korea
* Corresponding author (mmhan@gachon.ac.kr)

[Received 5 April 2021, Reviewed 8 April 2021(R2 4 June 2021), Accepted 11 June 2021]
☆ 본 연구는 국방과학연구소 연구용역 지원사업의 연구결과로 수행되었음 (UD200020ED)

이 등장하기 시작했다. 공격들을 방어하기 위해 시스템 외부 및 내부에서 다양한 방식으로 탐지가 진행되었다. 일반적으로 외부 탐지는 침입탐지 시스템, 침입방지 시스템, 방화벽 등 다양한 시스템 및 연구들이 제안되고 있지만 내부 네트워크로 공격자가 들어왔을 경우, 이를 탐지할만한 시스템 및 연구가 많지 않아 외부에 비해 상대적으로 쉽게 공격을 당하고 만다. 이를 해결하기 위해 내부전파 경로 탐지에 대한 연구가 진행되고 있다[1-5]. 내부전파란 공격자가 내부에 들어왔을 때 호스트들을 건너다니면서 중요한 서버에 도착해 정보를 탈취하거나 혹은 서버 자체를 다운시키는 공격 유형으로서 호스트들을 건너다니기 때문에 로그온을 기반으로 탐지가 진행된다. 그렇기 때문에 해당 연구는 크게 기존의 로그온 패턴을 기반으로 탐지하는 방식과 로그온을 시도하는 특징을 기반으로 탐지하는 방식, 두 가지로 나눠지게 된다. 로그온 패턴을 기반으로 탐지하는 방식은 기존의 로그온 패턴에 대한 정보가 필요하기 때문에 현재 공공 데이터로는 존재하지 않으며, 일반적으로 회사나 내부 기관이 데이터를 제공하고 이를 이용해 연구를 진행하는 방식으로 이루어진다[1-3]. 그렇기 때문에 본 연구에서는 공공데이터를 활용하는 로그온 시도 특징을 이용한 연구에 초점을 맞춘다.

로그온 특징을 활용하는 연구는 사용자, 송신호스트, 수신호스트를 기반으로 특징을 추출하는 연구인데 최근Remote Desktop Protocol(RDP)를 활용한 연구가 적은 리소스를 활용하면서도 좋은 결과를 나타내고 있다[4]. 하지만 그럼에도 불구하고 세 개의 제한점이 존재한다. 첫 번째로, 각 그래프들의 노드들 즉, 호스트들의 중요성 및 관계성을 다루지 않았다. 내부전파경로는 결과적으로 중요 노드에 침투하는 것을 목적으로 하며, 이를 고려할 때 중요노드 없이 특징을 생성할 경우 유의미한 특징 생성을 놓칠 수 있다. 두 번째로, 제안한 특징들이 일부 모델들에서는 성능저하를 보여 제안 특징이 데이터들을 잘 표현했다고 보기 어렵다. 마지막으로, 이렇게 만들어진 특징들을 활용해 예측 모델을 실행할 때 해당 예측에 대한 근거 및 타당한 설명을 제공하지 못한다는 문제점을 지닌다. 모델에 대한 설명을 제공하지 않을 경우, 모델 자체에 문제가 존재해도 이를 발견하기 어렵고 새로운 인사이트를 창출해도 이를 전달하기 어려워 결과적으로 실제 필드에서 활용하기 어렵다는 문제점이 존재한다[6].

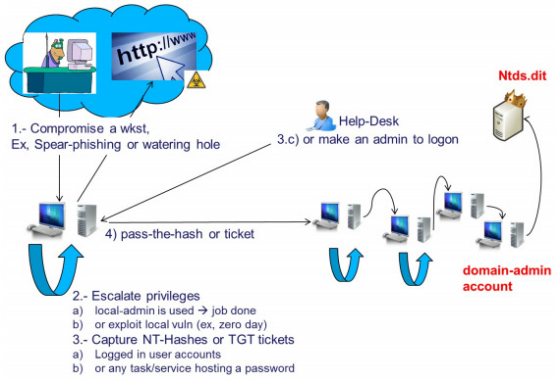
이를 해결하기 위해 본 연구는 페이지랭크(PageRank) 특징을 활용한 RDP기반 내부전파경로 탐지 및 SHAP를 이용한 설명가능한 시스템을 제안한다. 제안 모델은 크게 3단계로 나눠진다. 첫 번째로 호스트 이벤트 데이터들에 전처리를 진행해 RDP 기반의 로그온 이벤트 데이터를 추출한다. 두 번째로 이를 이용해 사용자-수신 호스트와 송신호스트-수신호스트 입장의 그래프를 생성해 페이지랭크 알고리즘을 이용한 각각의 노드 중요도를 추출하고 이를 특징으로 활용한다. 이와 함께 기존의 이벤트에서 수집될 수 있는 기본적인 특징인 로그온 시간차와 로그온 횟수를 이용해 통계적인 처리를 거쳐 특징으로 활용한다. 마지막으로 추출된 특징들을 이용해 다양한 모델에 적용시켜보고 이들의 결과에 SHAP를 적용해 설명을 제공한다. 실험 결과 제안 연구의 특징을 활용할 경우 모든 모델에서 평균 97.50%정도의 F1score를 기록했으며, SHAP로 설명을 생성해 확인해 본 결과, 해당 연구에서 제안한 특징들이 매우 큰 영향을 끼침을 확인했다.

2절에서는 관련연구에 대해 소개하고, 3절에서는 제안모델과 그 프로세스에 대해 설명하며, 4절에서는 실험 및 결과에 대해 언급하고, 5절에서는 결론을 내며 마무리한다.

2. 관련연구

2.1 내부전파경로 탐지 시스템

내부전파란 공격자가 내부 시스템에 들어와 호스트들을 건너가면서, 결과적으로 핵심적인 서버에 들어가 중요한 정보를 탈취하거나 공격하는 행위를 말한다[7]. 그림 1은 내부전파경로의 단계를 보여준다.



(그림 1) 내부전파경로 단계[7]
(Figure 1) Process of lateral movement[7]

그림 1에서 볼 수 있듯이 내부전파는 취약한 서버 및 피싱을 통해 내부로 들어가는 단계, 취약점을 이용해 권한상승하는 단계, 자격증명을 획득하는 단계, 해당 자격증명을 이용해 다른 호스트로 넘어가는 단계 총 4단계로 이루어진다[7]. 이로 인해 일반적으로 내부전파경로 탐지 시스템은 로그온 이벤트를 기반으로 탐지를 진행하며 크게 정상 로그온 패턴을 학습하는 방식과 로그온 시도 특징을 이용해 탐지하는 방식 두 가지 방식으로 이루어진

다. 로그온 패턴을 학습하는 방식은 그래프를 활용한 방식, 패턴을 활용한 방식, 확률을 활용한 방식으로 나눠진다. 그래프를 활용한 방식에서는 가장 먼저 기존 사용자의 로그온 패턴을 가중치를 부여한 그래프로 생성한다. 이후 해당 연결에 PCA(Principal Component Analysis)를 진행해 하나의 벡터값으로 생성하고 이를 다시 재구성(reconstuct)해 재구성오차(reconstruction error)를 계산한다. 해당 오차가 기존의 오차값에 비해 높을 경우, 이를 이상행위 로그온으로 간주하게 된다[1]. 두 번째로 패턴을 활용한 방식에서는 각각 부서의 직원들이 어떻게 로그온을 수행했는지에 대한 데이터를 활용해 패턴을 생성한다. 즉, 각 부서를 기반으로 해당 패턴을 생성하며 이에서 벗어날 경우 이상행위로 분류한다[2]. 마지막으로 확률을 이용한 방식에서는 기존의 Keberos 요청 횟수에 대해 확률을 생성한 후 해당 확률이 갑자기 증가할 경우 이를 이상행위로 판단하기도 한다[3]. 이렇게 정상 패턴을 학습할 경우, 기존의 이상행위에 대해 빠른 판단이 가능하며 직접 필드에서 사용하기도 편하다는 장점이 존재한다. 하지만 정상 패턴을 학습하려면 정상 사용 데이터와 이상행위 데이터가 하나의 기관 내 존재해야하는데 공공으로 해당 데이터를 공개한 경우는 없으며 일반적으로 내부의 데이터 혹은 다른 기관의 데이터를 받아 진행한 연구가 대부분이다. 하지만 공공 데이터를 활용하는 경우가 존재하는데 로그온 시도 특징을 활용한 연구가 대표적이다.

로그온 시도 특징을 활용한 연구는 그래프 특징 및 세션 특징을 활용한 연구로 이루어진다. 로그온 그래프를 그리고 그래프 내 특징을 추출한 연구에서는 각각의 사용자를 기준으로 그래프를 그리고 이들의 특징을 추출해 활용했으며 대표적으로 사용자 기준 평균 로그온 이벤트 발생 횟수, 로그온 시간차의 평균 값 등을 활용하였다. 이때 정상 데이터의 개수가 많기 때문에 언더샘플링을 진행했다[5]. 이 외에는 RDP를 활용한 원격 로그온 데이터를 가지고 세션특징을 추가해 탐지를 진행한 연구가 존재한다. 해당 연구의 경우, RDP를 활용한 이벤트만을 추출해 크기를 줄였으며, 이외에 세션을 특징으로 추가했다[4]. 하지만 세션 특징의 경우, 공격데이터에서는 수집할 수 없어 임의적으로 정상데이터에서 샘플링되었기 때문에 공격데이터 자체를 표현하지 못한다는 제한점을 지닌다.

2.2 PageRank Algorithm

페이지랭크 알고리즘이란 구글의 검색엔진에서 활용한 노드 중요도 평가 알고리즘으로 각 노드의 연결 횟수뿐만 아니라 연결된 노드의 중요도도 함께 평가하는 알고리즘이다. 기존의 중요노드는 많은 연결에 의해 결정되었다면, 해당 알고리즘은 연결되어 있는 노드들 자체의 중요도도 함께 판단한다. 각 그래프의 노드들은 검색결과 페이지를 나타내며, 간선들은 각 페이지에서 다른 페이지와의 링크를 나타낸다[8]. 즉, 단순히 많은 링크를 가진 페이지가 아닌, 다른 중요한 페이지들과 많은 링크를 갖는 페이지들을 중요 페이지로 판단하게 된다.

2.3 SHAP(SHapley Additive exPlanations)

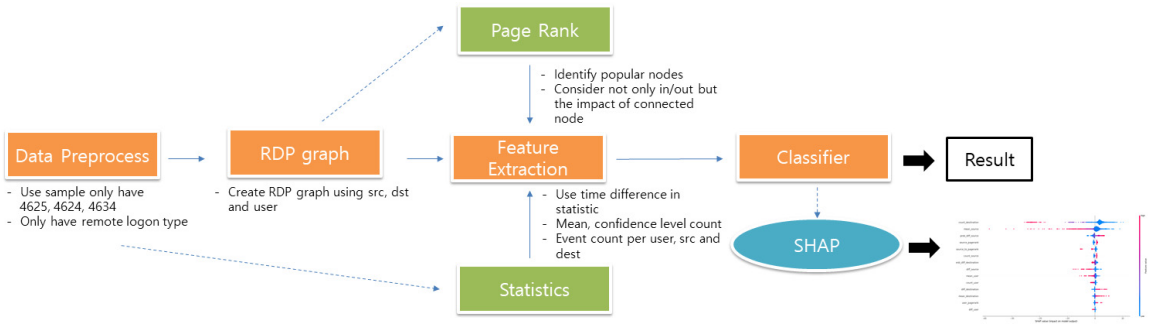
SHAP이란 Shapley value을 활용한 설명모델로서, 기존 게임이론의 Shapley 개념을 적용한 모델이다. 예측 및 판단을 하나의 게임으로, 예측 및 판단에 사용된 특징들은 해당 게임에 참여한 인원으로 치환해 계산된다[9]. SHAP은 기존의 설명모델인 LIME(Local Interpretable Model-Agnostic Explanation)[10]과 다르게 각각의 특징들과 다른 특징집합과의 영향력 비교 및 대조가 가능하다는 장점을 지니는 반면 특징의 수가 많은 경우, 계산이 기하급수적으로 증가해 학습이 상대적으로 오래 걸린다는 단점이 존재한다.

3. 제안연구

제안연구는 내부전파경로탐지로 내부전파 의심노드(의심노드)를 찾아 중요서버로 침입하기 전 공격을 탐지하는 것을 목표로 한다. 제안연구의 전체적인 프레임워크는 그림 2와 같이 이벤트 전처리 단계, 특징 추출 단계, 예측 및 설명제공 단계 총 3 단계로 나눠진다.

3.1 이벤트 전처리 단계

이벤트 전처리 단계에서는 많은 데이터 중 RDP를 활용하는 이벤트들만을 추출하고 정제하는 것을 목적으로 하며 크게 RDP 데이터 추출단계, 공격데이터 주입단계, 정제 단계 총 3단계로 이루어진다. RDP 데이터 추출 단계에서는 모든 호스트의 윈도우 이벤트 중 로그온과 로그온 실패 및 로그오프와 관련된 4624, 4625, 4634 의 이벤트들만을 추출한다. 이때 로그온도 여러 가지 종류가 있는데, 여러 가지 로그온 종류 중 원격로그온과 관련된 종류만을 선택해 추출한다. 공격 주입단계에서는 공격자의 로그온 이벤트들을 추출된 정상이벤트 데이터프레임



(그림 2) 제안연구 프레임워크
(Figure 2) Framework of proposed model

에 시간(time stamp)에 맞춰 추가해 전체적인 로그온 데이터셋을 완성시킨다. 마지막으로 정제단계에서는 사용자 정보, 송신 호스트 정보 등의 특징들이 결측된 데이터들을 찾아 이를 제거한다. 해당 3단계를 진행하면 정제된 RDP 데이터셋을 생성하게 된다.

3.2 특징 추출 단계

특징 추출 단계에서는 앞서 생성된 데이터셋을 활용해 특징들을 생성하고 추출하는 작업을 진행한다. 해당 단계에서는 그래프 생성 및 특징 추출, 통계적 특징 추출 총 2가지 작업이 진행된다.

그래프 생성 및 특징 추출 단계에서는 반드시 "사용자=송신 호스트" 가 아니기 때문에 이를 고려해 사용자-수신호스트, 송신호스트-수신호스트 두 종류의 그래프를 생성한다. 이후 각각 노드들의 중요성을 페이지랭크 알고리즘을 활용해 계산한다. 본 연구에서 일반적인 연결횟수가 아닌 페이지랭크 알고리즘을 기준으로 노드의 중요성을 판단한 이유는 해당 알고리즘은 연결된 노드들의 영향력도 같이 고려하기 때문에 더 정확하게 중요성을 파악할 수 있기 때문이다. 해당 알고리즘은 수식 1을 통해 계산되는데 해당 수식에서 T_n 은 n번째 페이지를 의미하며, $C(T_n)$은 T_n의 링크 수를 의미한다. 마지막으로 d는 dumping factor로 사람들이 다른 페이지를 클릭할 확률을 나타내게 된다[8].

$$PR(A) = \frac{(1-d)}{N} + d((\frac{PR(T_1)}{C(T_1)}) + \ldots + (\frac{PR(T_n)}{C(T_n)})) \quad (1)$$

본 연구에서는 사용자-수신호스트, 송신호스트-수신호스트 그래프를 활용하기 때문에 결과적으로 T_n 은 n번째 호스트들을 의미하며, $C(T_n)$은 n번째 호스트와 연결된 호스트 수를 의미하게 된다. Dumping factor의 경우, 일반적으로 사용되는 0.85를 선정한다. 실제로 일반적인 연결횟수를 특징으로 활용했을 때보다, 페이지랭크 값을 활용하였을 때 더 좋은 성능을 나타내며 이를 실험으로 증명한다.

두 번째로 통계적 특징 추출은 호스트를 기준으로 각각의 이벤트 발생횟수, 로그온 시간차, 로그온 시간차의 평균, 로그온 시간차의 신뢰구간 포함횟수를 특징으로 활용한다. 이벤트 발생횟수나 시간차 등의 기본적인 값들만을 특징으로 활용할 경우, 변별력이 떨어지고 각 호스트들만의 패턴을 표현하기 어렵게 때문에 여러 가지 통계기법을 활용한다. 가장 먼저 이벤트 발생 횟수의 경우, 각 로그온 이벤트가 얼마나 발생했는지를 계산하며, 로그온 시간차의 경우, 다음 로그온 시도가 이전 로그온 시도와 얼마만큼의 차이가 나는지를 계산한다. 또한 로그온 시간차 평균의 경우, 앞서구한 시간차를 호스트 기준으로 각각의 평균을 구해 계산한다. 마지막으로 로그온 시간차의 신뢰구간 포함횟수의 경우 가장 먼저 정상 로그온들의 시간차를 호스트 기반으로 분류하고 각 호스트들의 로그온 시간차의 평균과 표준편차를 구해 이들을 이용해 로그온 시간차 가우시안 분포를 생성한다. 이후 정상과 악성을 포함한 각각 호스트들의 시간차가 해당 가우시안 분포의 신뢰구간 안에 들어갈 경우 해당 횟수를 계산한다. 수식 2는 신뢰구간을 구하는 공식을 나타낸다[11]. 해당 공식에서 $\overline{X}$는 호스트의 시간차 평균을, σ는 호스트 시간차의 표준편차를, n은 호스트의 로그온 이벤트 발생횟수를 나타낸다. $z_{\alpha/2}$는 $\alpha/2$에 해당하는 면적을 가진 z 값을 의미하며 일반적인 신뢰구간에서 α는 0.05를 나타내지만 본 연구에선 0.25로 선정한다.

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad (2)$$

결과적으로 해당 수식을 활용해 각 호스트마다의 시간차 신뢰구간을 구하게 되며 해당 구간 안에 들어갈 경우 카운팅이 진행된다. 이렇게 신뢰구간을 활용한 이유는 시간차의 이상치 값이 드물게 존재하기 때문이다. 즉, 시간차가 매우 크거나 매우 작은 값이 아주 가끔씩 존재하며 이들을 반영해 분포를 생성할 경우, 분포가 왜곡될 수 있다는 문제점이 존재한다. 이를 고려해 제안연구에서는 신뢰구간의 바깥 측 즉, 분포에서 거의 발생하지 않을 영역을 제거하고 내부에 해당 시간차 값이 들어오는지를 확인함으로써 좀 더 신뢰성있고 변별력 있는 특징들을 생성하게 된다.

3.3 예측 및 설명제공 단계

예측 및 설명제공 단계에서는 앞서 구해진 특징을 여러 가지 모델을 활용해 예측하고 SHAP를 활용해 설명을 제공하는 작업을 진행한다. 해당 연구에서는 Random forest(RF)[12], Gradient boosting(GB)[13], Decision tree(DT)[14], Gaussian naïve Bayes(GNB)[15], LogitBoost(LB)[16], XGBoost(XGB)[17] 총 6개의 모델을 활용해 평가를 진행했다. SHAP의 경우, 전역적으로 각 모델 판단에 어떤 특징들이 영향을 끼쳤는지 평가하며, 국소적으로 각 인스턴스에 어떤 특징의 어느 값이 영향을 끼쳤는지 설명을 제공한다.

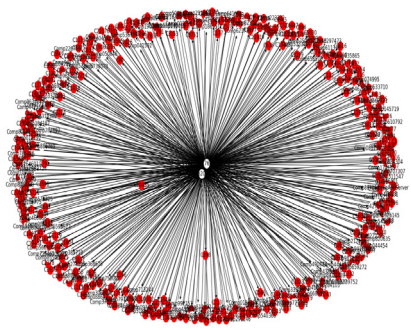
4. 실험 및 성능평가

4.1 데이터셋

실험에서 활용한 데이터셋으로는 LANL(Los Alamos National Laboratory)데이터셋을 활용한다. LANL 데이터셋은 사용자들의 다양한 호스트 이벤트들이 나타나 있을 뿐 아니라 공격 로그온 데이터도 함께 있어, 가장 많이 사용되는 데이터로 본 연구에서도 활용되었다. 정상 데이터셋으로는 LANL의 “Unified Host and Network Data Set”의 Host Event dataset을 활용하며[18], 공격 데이터셋으로는 “Comprehensive, Multi-Source Cyber-Security Events” 의 redteam dataset을 활용한다[19]. Host Event dataset의 경우, 90일간 LANL의 시스템에서 실행된 윈도우 호스트 이벤트 데이터로서 실제 회사와 유사하게 구현하기 위해 Active Directory 서버, 이메일 서버, 스캐닝 시스템 등도 함께 구현되어있다. 해당 데이터셋의 경우 Timestamp, Event ID, Process, Process ID 등 총 21개의 특징들을 활용할 수 있다[18]. Redteam 데이터셋의 경우 총 749개의 로그온 이벤트로 이루어져 있으며, 로그오프 이벤트 및 다른 이벤트 정보는 존재하지 않는다. 사용가능한 특징의 경우 “Timestamp, User, Source host, Destination host, Domain”이 전부이므로 편향되지 않은 정보로 사용될 수 있는 건 Timestamp(시간)만이 유일하다[19]. 본 연구에서는 공격데이터로 749개의 데이터와 정상데이터에서 RDP 기반 원격로그온 데이터 6,756개를 추출해 총 7,505 개의 데이터를 생성했으며 이중 80%는 학습에 20%는 테스트에 활용하였다.

4.2 실험과정

전처리를 통해 생성된 7,505개의 데이터를 활용해 RDP 그래프를 생성했으며 그림 3은 송신 호스트-수신 호스트 그래프 중 송신 호스트의 페이지랭크 값을 기준으로 높은 페이지랭크 값을 가진 상위 1,000개 노드의 연결을 나타낸다. 붉은 노드는 의심노드, 하얀 노드는 정상노드를 나타내며 1,000개 중 의심노드 701개, 정상노드 299개로 이루어져있다.



(그림 3) source_pagerank 상위 1,000 노드 그래프
(Figure 3) Top 1,000 source_pagerank node in graph

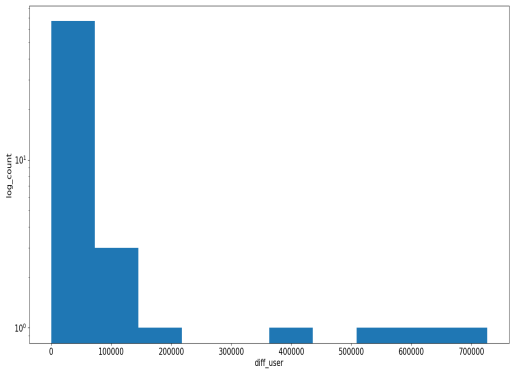
그림 3에서 볼 수 있듯이 공격자는 수신 호스트에 비해 송신 호스트가 한정적이기 때문에, 동일한 송신 호스트를 많이 사용하게 되고 이로 인해 공격에서 사용된 의심노드의 송신자 페이지랭크 값이 상대적으로 높음을 알 수 있다. 다음으로 통계적인 특징 중 로그온 시간차에 대해 보면 그림 4와 같이 대부분의 시간이 0근처로 모여 있으며 매우 큰 시간차를 갖는 사건은 드물게 일어남을 확인할 수 있다. 이를 가우시안 분포로 생성하면 그림 5처럼 나타나며 그림에서 볼 수 있듯이 매우 넓은 분포를 갖

(표 1) 사용된 특징집합
(Table 1) Used feature set

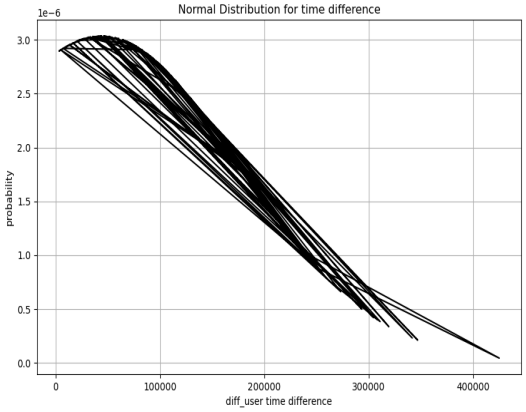
Feature Name	Description
diff_user	사용자 별 다음 로그온 이벤트와의 시간 차
diff_source	송신 컴퓨터 별 다음 로그온 이벤트와의 시간 차
diff_destination	수신 컴퓨터 별 다음 로그온 이벤트와의 시간 차
mean_diff_user	사용자 별 시간 차의 평균 값
mean_diff_source	송신 컴퓨터 별 시간 차의 평균 값
mean_diff_destination	수신 컴퓨터 별 시간 차의 평균 값
prob_diff_source*	다른 정상 송신 컴퓨터 로그온 신뢰구간 내 존재 횟수
prob_diff_destination*	다른 정상 수신 컴퓨터 로그온 신뢰구간 내 존재 횟수
count_user	사용자 로그온 횟수
count_source*	송신 컴퓨터 별 로그온 횟수
count_destination*	수신 컴퓨터 별 로그온 횟수
user_pagerank*	사용자의 페이지랭크 값
source_pagerank*	송신 컴퓨터 페이지랭크 값
source_to_pagerank*	송신 컴퓨터를 수신 컴퓨터와 연결했을 때 페이지 랭크 값

* : 본 연구에서 제안한 특징들을 의미함

고 있음을 확인할 수 있다. 그림 5의 확률분포를 보면 시간차가 커질수록 발생할 확률은 빠르게 0에 가까워진진다. 이를 반영해 확률 값이 아닌 신뢰구간을 생성해 포함횟수를 계산하게 된다. 마지막으로 앞서 한 설명을 토대로 총 7개의 특징을 추출했으며 기존의 존재했던 특징들을 추가해 총 14개의 특징을 활용하였다. 표 1은 사용된 특징들과 그에 대한 설명을 나타낸다.



(그림 4) 로그온 시간차의 히스토그램
(Figure 4) Histogram of logon time difference

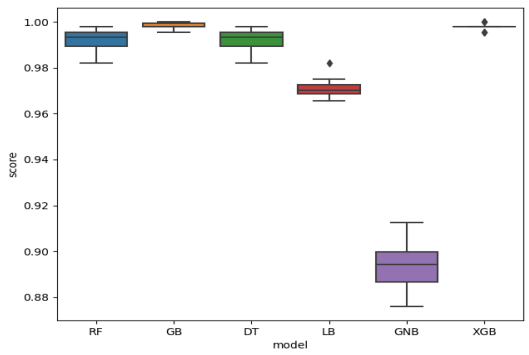


(그림 5) 로그온 시간차 정규분포
(Figure 5) Gaussian distribution of logon time difference

4.3 성능평가

성능평가에서는 불균형 데이터셋인 것을 고려해 정확도가 아닌 F1score를 사용하며[20] 평가 모델로는 RF, GB, DT, GNB, LB, XGB를 활용한다. 실험은10-fold cross validation을 진행하였으며 결과는 그림 6과 같다.

그림 6에서 확인할 수 있듯이, 제안 연구의 특징을 활



(그림 6) 10-fold cross validation F1score 결과
(Figure 6) Result of 10-fold cross validation in F1score

용할 경우, GNB를 제외한 대부분의 모델에서 큰 차이 없이 준수한 성능을 보여줌을 확인할 수 있다. 특히 대부분의 모델에서 큰 분산없이 평균 근처에서 존재함으로써 본 연구에서 제안한 특징의 일반성(generalization)도 확인할 수 있다. 표 2는 제안연구와 기존 연구를 비교한 결과로서 표에서 확인할 수 있듯 제안 연구가 LB를 제외한 대부분의 모델에서 더 좋은 성능을 보여줌을 확인할 수 있다.

마지막으로 본 연구에서 제안한 특징들의 유효성을 확인한다. 본 연구에서는 노드의 중요도를 위해 페이지랭크 알고리즘을 활용했으며 통계적 특징으로 시간차 신뢰구간 포함횟수를 활용했다. 가장 먼저 표 3을 통해 페이지랭크 알고리즘이 일반적인 연결횟수에 비해 더 정확한 중요도를 제공해준다는 것을 증명한다. 표 3은 페이지랭크 값과 연결 횟수만을 기준으로 한 성능평가 결과를 나타낸다. 표에서 볼 수 있듯이 페이지랭크 값을 사용해 중요도를 표현했을 때 더 정확한 탐지 결과를 추출할 수 있다. 두 번째로 신뢰구간 포함횟수만을 이용해 진행한 성능평가의 결과를 표 4에 나타낸다. 표 4를 통해 신뢰구간 포함횟수만을 사용했을 때도 준수한 성능을 보여줌을 확인할 수 있다.

대부분의 결과에서 GNB만이 유독 낮은 성능을 보였는데 이는 GNB의 기본이 되는 Gaussian, 정규분포의 영향으로 볼 수 있다. 즉, GNB는 모든 특징들이 정규분포를 따른다고 가정하고 확률밀도를 계산해 분류를 진행하

(표 2) 제안연구와 이전연구 성능비교
(Table 2) Comparison of performance in proposed model and previous model

F1score	Random Forest	Gradient Boosting	Decision Tree	Gaussian NB	Logit Boost	XGBoost
제안연구	0.99553	1.0	0.98876	0.88429	0.98190	1.0
이전연구	0.978	0.994	0.962	0.8466	0.997	-

(표 3) 페이지랭크와 연결횟수를 활용한 성능평가
(Table 3) Comparison of performance in pagerank algorithm and the number of links

F1score	Random Forest	Gradient Boosting	Decision Tree	Gaussian NB	Logit Boost	XGBoost
페이지랭크	0.99103	1.0	0.99777	0.47555	0.98190	1.0
연결횟수	0.45578	0.45578	0.45578	0.32378	0.45578	0.45578

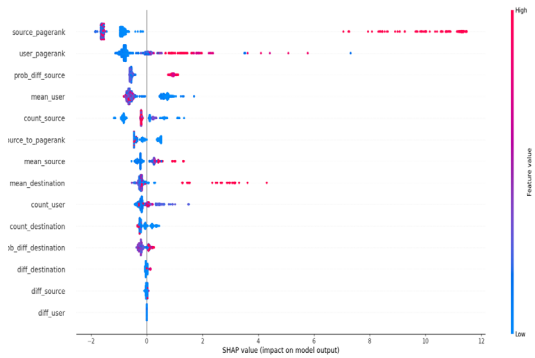
(표 4) 신뢰구간 포함횟수를 활용한 성능평가
(Table 4) Comparison of performance in the "prod_diff_source"(confidence level included)

F1score	Random Forest	Gradient Boosting	Decision Tree	Gaussian NB	Logit Boost	XGBoost
신뢰구간 포함횟수	0.98876	0.99777	0.99328	0.55172	0.98190	0.99777

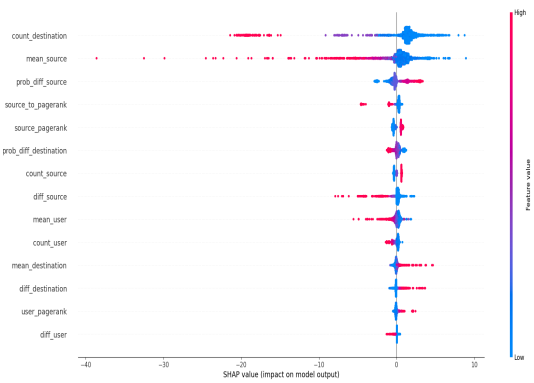
기 때문에 정규분포를 따르지 않는 특징에 의해 성능이 낮게 나올 수 있다. 이와 함께 RF, DT 및 XGB처럼 트리 기반의 분류기는 매우 좋은 성능을 보여줬는데 이는 일반적으로 불균형데이터들이 트리계열에서 좋은 성능을 보인다는 지식과 일치한다.

4.4 SHAP을 활용한 설명

마지막으로 SHAP을 활용해 설명을 제공한다. 설명은 크게 중요특징들과 그들의 분포값을 보여주는 모델기준의 설명과 각 인스턴스 하나의 예측에 대해 특징과 그 범위를 제공하는 인스턴스기준의 설명으로 나눠지며 먼저 모델 기준의 설명부터 나타낸다. 그림 7은 가장 좋은 성능을 기록한 XGB에 대한 설명을 나타내며, 그림 8은 가장 저조한 성능을 기록한 GNB에 대한 설명을 나타낸다.

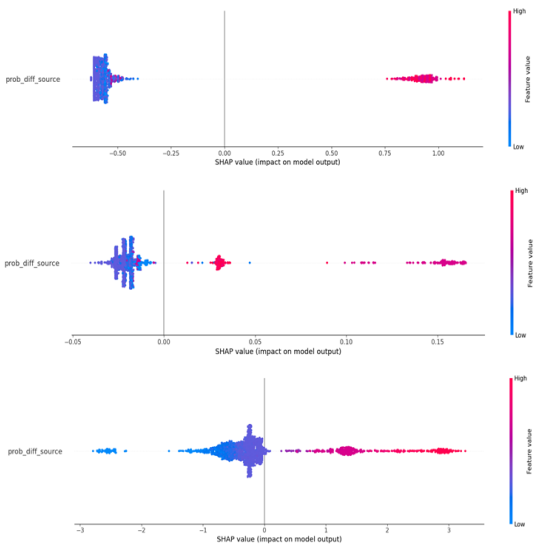


(그림 7) XGBoost의 SHAP summary plot
(Figure 7) SHAP summary plot in XGBoost



(그림 8) 가우시안 나이브 베이즈 SHAP summary plot
(Figure 8) SHAP summary plot in GNB

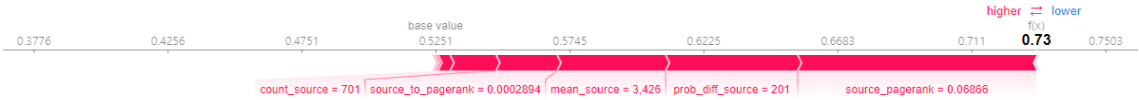
붉은색으로 표시된 포인트는 해당 특징의 높은 값을 의미하며, 푸른색은 낮은 값을 의미한다. XGB를 보게 되면 source_pagerank, user_pagerank, prob_diff_source, mean_user, count_source 가 상위 5개의 특징을 차지했으며 이 중 제안 연구가 추출한 특징이 4개를 차지한다. 가장 낮은 성능을 보여준 GNB에서도 제안한 특징들이 상위 5개 중 3개를 차지함을 확인할 수 있다. 실제로 페이지랭크의 경우 그 값이 클수록, 악성행위 판단에 긍정적인 영향을 끼치며, 특히 “prob_diff_source”에 의해 정상 로그온 시간차 신뢰구간에 더 많이 포함될수록 의심노드일 경우가 높다는 사실을 발견했다. 기존에 일반적인 상식에서는 정상 분포의 신뢰구간을 활용했기 때문에 정상호스트가 더 많이 포함될 것이라고 예측했으나, 의심노드가 더 많이 포함된 것을 확인할 수 있다. 실제로 이런 경향은 대부분의 모델에서 발생했으며 그림 9에서 이에 대해 확인할 수 있다. 그림 9는 위에서부터 XGB, RF, GNB에서 신뢰구간 포함횟수가 의심노드 라고 판단하는데 끼친 영향도를 표현한 것이다.



(그림 9) 모델마다 “prod_diff_source” 영향도
(Figure 9) Impact of “prod_diff_source” in model

이는, 정상적인 로그인 시도 시간 내 더 많이 포함될수록 의심노드일 경우가 높다는 것으로 정상적인 사람들은 자신들이 일반적으로 원하는 시간에 시도하는 반면에 악성행위자들은 일반적인 로그온 시간차와 유

(그림 10) 정상 인스턴스에 대한 특징 영향도 평가
(Figure 10) Feature impact on normal instance



(그림 11) compromised 인스턴스에 대한 특징 영향도 평가
(Figure 11) Feature impact on compromised instance

사하게 행동한다는 것을 확인할 수 있는 대목이다.

다음은 인스턴스 기반의 설명제공에 대해 다룬다. 앞서 XGB와 GNB에 대해 설명을 제공했기 때문에 뒤의 인스턴스 설명은 랜덤포레스트를 기반으로 진행한다. 가장 먼저 그림 10은 정상 인스턴스를 그림 11은 의심노드 인스턴스에 대한 설명을 나타낸다. 해당 그림에서 푸른색은 판단에 부정적인 영향, 붉은색은 긍정적인 영향을 나타내며, base value는 각각의 샘플들의 결과를 클래스 기반으로 확인했을 때 값으로 모든 훈련데이터들의 아웃풋은 클래스 0을 기준으로 0.711의 평균값을 가지며, 클래스 1에 대해 0.5241의 평균값을 가짐을 나타낸다. 이와 함께 f(x)는 해당 샘플의 클래스에 대한 값을 나타낸다. 즉 그림 11을 해석해보면, 훈련 데이터는 클래스 1(의심노드)을 기준으로 평균 0.5251의 결과값을 가지지만 해당 샘플은 붉은색과 푸른색의 특징에 의해 0.73의 값을 가진다고 해석가능하다. 그림 10 에서 보면 mean_user가 판단에 부정적인 영향을 끼쳤으나, source_pagerank, prod_diff_source 등의 특징에 의해 긍정적인 영향을 받음을 확인할 수 있다. 그림 11의 경우는 부정적인 영향을 받은 특징은 없으며, 앞선 그림 10과 유사한 특징들에 의해 긍정적인 영향을 받음을 확인할 수 있다. 즉, 이를 통해 source_pagerank, prod_diff_source, mean_source 등의 값들이 클 경우 의심노드인 경향을 띈다는 것을 확인할 수 있다.

5. 결 론

여러 가지의 실험을 거쳐 제안 연구에서 제시한 특징들이 판단에 큰 영향을 끼치고 이들이 어떤 경향을 보이는지 SHAP 이라는 설명제공 모델을 통해 검증했다. 가장 먼저 특징 측면에서, 기존에 일반적으로 사용되던 연결횟수 기반의 노드 중요도를 페이지랭크 알고리즘을 이용해 더 높은 활용도를 가질 수 있도록 만들었으며, 신뢰구간을 활용해 적은 기본특징에 대해서도 데이터를 대표할 수 있는 정제된 특징집합을 생성했다. 결과적으로 본 연구는 이전연구보다 기존 데이터에 변별력을 띄는 데이터셋을 생성할 수 있었다. 두 번째로 본 연구에서는 SHAP을 이용해 설명을 제공했다. 설명을 제공함으로써 정상과 의심노드간의 차이를 구분할 수 있었으며, 본 연구에서 제안한 특징들이 중요한 영향을 끼친다는 것을 확인할 수 있었다. 본 연구는 앞으로 설명에 초점을 맞춰 모델을 발전시켜나갈 생각이다. 이번에 사용된 SHAP는 매우 유명하고 모든 모델에서 사용될 수 있는 훌륭한 설명모델이지만, 모델 자체에서 설명을 제공하는 것이 아니기 때문에 설명과 결과까지의 프로세스가 조금씩 다를 수 있다[21]. 그렇기 때문에 다음연구에서는 모델 자체가 설명성을 가지는 ad-hoc 형태의 설명 모델을 만들어 좀 더 정확한 설명을 제공할 생각이다.

참고문헌(Reference)

[1] BA Powell, "Detecting malicious logins as graph anomalies.", Journal of Information Security and Applications, Vol. 54, No. 102557, 2020. https://doi.org/10.1016/j.jisa.2020.102557.

[2] H Siadati, and M Nasir, "Detecting structurally anomalous logins within enterprise networks.", Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security., pp. 1273-1284, 2017.

https://doi.org/10.1145/3133956.3134003

[3] Q Liu, J W Stokes, R Mead, T Burrell, I Hellen, J Lambert *et al*, "Latte: Large-scale lateral movement detection." MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM) pp. 1-6, 2O18 https://doi.org/10.1109/MILCOM.2018.8599748

[4] T Bai, H Bian, M. A. Salahuddin, A. A. Daya, N Limam, R Boutaba, "RDP-based Lateral Movement detection using Machine Learning", Computer Communications, Vol. 165, pp. 9-19., 2021. https://doi.org/10.1016/j.comcom.2020.10.013

[5] G Kaiafas, G Varisteas, S Lagraa, R State, and C. D. Nguyen et al, "Detecting malicious authentication events trustfully", NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, Taipei, pp. 1-6, 2018. https://doi.org/10.1109/NOMS.2018.8406295.

[6] M. Hind, "Explaining explainable AI." XRDS: Crossroads, The ACM Magazine for Students, Vol.25 No.3 pp. 16-19, 2019, https://doi.org/10.1145/3313096

[7] M Soria-Machado, D Abolins, C Boldea, and K Socha, "Detecting Lateral Movements in Windows Infrastructure", CERT-EU Security Whitepaper 17‐002, 2017. https://media.cert.europa.eu/static/WhitePapers/CERT-EU_SWP_17-002_Lateral_Movements.pdf

[8] L Page, S Brin, R Motwani, and T Winograd, "The PageRank citation ranking: Bringing order to the web", Stanford InfoLab., 1999. http://ilpubs.stanford.edu:8090/422/

[9] S Lundberg, and SI Lee, "A Unified Approach to Interpreting Model Predictions", arXiv preprint arXiv:1705.07874, 2017. https://arxiv.org/abs/1705.07874v2

[10] MT 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. https://doi.org/10.1145/2939672.2939778

[11] L Schmetterer, "Introduction to mathematical statistics", Springer Science & Business Media, Vol. 202., 2012. https://doi.org/10.1007/978-3-642-65542-5

[12] L Breiman, "Random forests", Machine learning Vol. 45, No. 1, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[13] JH Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", The Annals of Statistics, Vol. 29, No. 5, pp. 1189-1232, 2001. https://www.jstor.org/stable/2699986

[14] JR Quinlan, "Induction of decision trees", Machine learning, Vol. 1, No. 1, pp. 81-106, 1986. https://doi.org/10.1007/BF00116251

[15] DJ Hand, and K Yu, "Idiot's Bayes—not so stupid after all?", International statistical review, Vol. 69, No. 3, pp. 385-398, 2001, https://doi.org/10.1111/j.1751-5823.2001.tb00465.x

[16] J Friedman, T Hastie, and R Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)", Annals of statistics, Vol. 28, No.2, pp. 337-407, 2000. https://doi.org/10.1214/aos/1016218223

[17] T Chen and C Guestrin, "Xgboost: A scalable tree boosting system", Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794, 2016. https://doi.org/10.1145/2939672.2939785

[18] MJM Turcotte, AD Kent, and C Hash, "Unified Host and Network Data Set", Data Science for Cyber-Security, pp. 1-22, 2018. https://doi.org/10.1142/9781786345646_001

[19] AD Kent, "Comprehensive, multi-source cyber-security events data set", No. LA-UR-15-23810. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2015. https://doi.org/10.17021/1179829

[20] H He, and Y Ma, "Imbalanced Learning: Foundations, Algorithms, and Applications", 2013. https://doi.org/10.1002/9781118646106

[21] C Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence Vol. 1, No. 5, pp. 206-215, 2019. https://doi.org/10.1038/s42256-019-0048-x

◐ 저 자 소 개 ◑

윤 지 영(Jiyoung Yun)
2020년 가천대학교 컴퓨터공학과(공학사)
2020년~현재 가천대학교 일반대학원 소프트웨어학과 석사과정
관심분야 : XAI, Machine Learning, Statistics, Intrusion Detection
E-mail : apfhd9043@naver.com

김 동 욱 (Dong-Wook Kim)
2015년 가천대학교 컴퓨터공학과(공학사)
2017년 가천대학교 일반대학원 컴퓨터공학과(공학석사)
2017년~현재 가천대학교 컴퓨터공학과 박사과정
관심분야 : Cyber Security, Data Mining, Artificial intelligence
E-mail : kog7306@naver.com

신 건 윤(Gun-Yoon Shin)
2017년 가천대학교 인터랙티브 미디어 융합학과 학사
2018년 가천대학교 일반대학원 컴퓨터공학과(공학석사)
2018년~현재 가천대학교 컴퓨터공학과 박사과정
관심분야 : 기계 학습, 악성코드 분석, 공격자 식별, 저자 분석, 인공지능
E-mail : tlsrjsdbs@gmail.com

김 상 수(Sang-Soo Kim)
1997년 경북대학교 전자공학과(공학사)
2003년 경북대학교 컴퓨터공학과(공학석사)
2003년~현재 국방과학연구소 연구원
관심분야 : Cyber Security, Cyber Situation Awareness, Artificial Intelligence
E-mail : wisdory@naver.com

한 명 묵(Myung-Mook Han)
1980년 연세대학교 공과대학(공학사)
1987년 뉴욕공과대학교 대학원 컴퓨터공학과(공학석사)
1997년 오사카시립대학교 대학원 정보공학부(이학박사)
1998년~현재 가천대학교 소프트웨어학과 교수
관심분야 : 정보보호, 알고리즘, 데이터 마이닝, 기계 학습
E-mail : mmhan@gachon.ac.kr