SHAP을 통한 AI 투명성 확보 연구*

A Study on Securing Transparency of AI through SHAP

이은규 (Eun-Gyu Lee)(제1 저자) | 호서대학교 대학생 | legleg1216@gmail.com
이태진 (Tae-Jin Lee)(교신저자) | 호서대학교 교수 | kinjecs0@gmail.com

목 차

초 록

클라우드, 5G, IoT와 같은 기술발전으로 네트워크 사용량이 급증하고 있다. 실 환경에서 많은 서비스들이 네트워크에 의존하고 있어 네트워크 보안의 관심이 높아지고 있으며, 이에 따라 관련 연구도 활발하게 진행되고 있다. Intrusion Detection System(IDS) 환경에서의 AI 관련 연구도 진행되고 있다. IDS의 정확도를 높이기 위해 AI의 모델은 더 복잡해지고 불투명성이 증가하고 있다. 모델의 불투명성은 모델이 내린 예측을 적극적으로 수용하는데 방해요소이다. 본 연구는 SHAP을 통해 해석 가능성으로 모델의 투명성을 확보한다. 모델의 투명성을 개선하여 적극적인 사용이 가능하도록 한다. 더불어, 모델의 투명성 확보는 IDS뿐 아니라 금융, 의료분야 등 여러 분야에서도 중요하게 여기고 있어, 여러 분야에 활용될 것을 기대한다.

* **키워드 :** SHAP, 해석 가능성, 투명성, Intrusion Detection System

ABSTRACT

Network usage is rapidly increasing due to technological advances such as cloud, 5G, and IoT. Since many services depend on the network in real environments, the interest in network security is increasing, and accordingly, related research is actively being conducted. Research on AI in the Intrusion Detection System (IDS) environment is also underway. To increase the accuracy of IDS, AI's models are becoming more complex and non-transparency. The non-transparency of the model prevents it from actively accepting predictions made by the model. This study secures the transparency of the model with interpretability through SHAP. Improve the transparency of the model so that it can be actively accepting predictions. In addition, securing transparency in models is important not only in IDS but also in various fields such as finance and medical care, so it is expected to be used in various fields

* **Keywords :** SHAP, Interpretability, Transparency, Intrusion Detection System

* 이 논문은 문화체육관광부 및 한국저작권위원회의 2021년도 저작권기술개발사업의 연구결과로 수행되었음. (No.2019-PF-9500)

• 논문접수일 : 2021년 02월 22일 • 최초심사일 : 2021년 02월 23일 • 게재확정일: 2021년 03월 16일

1. 서론

클라우드, 5G, IoT와 같은 발전으로 네트워크 사용량이 급증하고 있다. 실 환경에서 많은 서비스들이 네트워크에 의존하고 있어 네트워크 보안이 점점 더 중요해지고 있으며, 이에 따라 관련 연구도 활발하게 진행되고 있다. Intrusion Detection System(IDS) 환경에서의 AI 관련 연구도 진행되고 있으며(Kim, Park, & Lee, 2020; Yin et al., 2017), AI의 강력한 학습 알고리즘으로 IDS의 정확도를 높이는데 성과를 보이지만, 정확도가 높아질수록 구조가 복잡해져 모델의 불투명도는 커지고 있다. 모델이 투명할수록 모델을 신뢰하고 적극적으로 사용할 수 있게 되기에 모델의 해석 가능 여부는 중요하다(Marino, Wickramasinghe, & Manic, 2018; Wang et al., 2020). 따라서 모델이 내린 예측을 이해할 수 있는 것이 중요하다.

본 연구에서는 AI로 IDS를 구축하고 XAI(eXplainable Artificial Intelligence) 중 하나인 SHAP을 통해 모델을 해석한다. SHAP 기반 모델 해석을 통해 모델이 공격 유형별 중요하게 여긴 feature들과 실제 공격 유형과의 연관성을 살펴본다.

2. 관련 연구

2.1 IDS를 위한 분류 모델 비교 분석

NSL-KDD는 KDD'99에서 단점을 보완하고 정제된 버전이다. 중복 데이터를 줄여 데이터 편향을 줄이고 다른 연구들과 결과를 비교하기에 적절하다(Canadian Institute for Cybersecurity, 2009). 하지만, NSL-KDD는 metadata 형식의 전처리를 통해 사용되는 Dataset으로 실험 환경에서 정확도가 높을지라도, 네트워크 트래픽은 추세에 따라 변하기 때문에 지속적인 재학습 없이 실 환경에서의 정확도는 감소할 수 있다. 머신 러닝을 활용해 NSL-KDD dataset을 분류한 연구에서는, NSL-KDD Dataset을 여러 Classifier를 이용해 결과를 도출한 후 비교를 진행하였다. XGBoost의 정확도가 98.7%로 가장 높았고 Decision Table과 Logistic이 그 뒤를 이었다(Dhaliwal, Nahid, & Abbas, 2018). NSL-KDD Dataset에는 label이 크게 Normal, DoS, Probe, R2L, U2R로 나뉜다. 전체적으로 NSL-KDD에 대해서 Normal과 나머지 공격 유형들을 Anomaly로 묶어 분류한 Binary classification이 Multi classification보다 정확도가 높았다. Multi classification에서는 R2L와 U2R에 대한 FAR(False Alarm Rate)가 높아 다른 공격 유형보다 정확도가 매우 낮게 측정되었다.

2.2 XGBoost

여러 개의 Decision Tree를 조합해서 사용하는 Ensemble 알고리즘으로 Boosting 기법을 이용하여 구현한 알고리즘인 Gradient Boost가 대표적이며, Gradient Boost 알고리즘을 병렬 학습을 지원되도록 구현한 라이브러리가 XGBoost이다. 모델의 파라미터 설정을 통해 과적합 규제가 가능하고 자체 내장된 교차 검증 알고리즘이 있다. 또한, 병렬 학습으로 메모리 자원을 최적으로 사용하여 속도가 매우 빠르다.

XGBoost에는 다양한 파라미터가 있으며, 파라미터를 사용하여 특정 작업을 수행할 수 있다. <표 1>의 XGBoost는 NSL-Dataset을 Binary classification 진행할 때 파라미터 설정을 변경해가며 결과를 비교한 결과 가장 높은 점수가 나온 파라미터를 나타낸다(Dhaliwal, Nahid, & Abbas, 2018).

<표 1> XGBoost Parameters

Parameter	Description	setting
learning_rate	Step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and learning_rate shrinks the feature weights to make the boosting process more conservative.	0.1
max_depth	Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.	5
subsample	Subsample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees. and this will prevent overfitting.	0.9
colsample_bytree	subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed.	0.8
min_child_weight	Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than min_child_weight, then the building process will give up further partitioning.	1
seed	Random number seed.	0
objective	Specify the learning task and the corresponding learning objective.	binary:logistic

2.3 SHAP(SHapley Additive exPlanations)

SHAP은 LIME과 Shapley value를 연결한 이론이다. LIME은 Data에 변형을 주며 블랙박스 모델의 예측에 어떤 영향이 있는지를 테스트하여 가중치를 계산하는 이론이다(Molnar, Christoph,

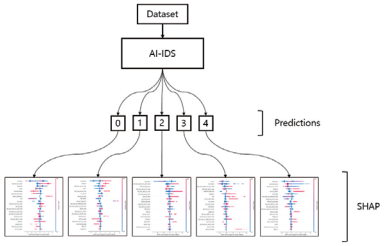
2019). Shapley value는 coalitional game theory를 기반의 이론으로 feature의 기여도를 나타내는 지표이다. feature value에 대한 Shapley value는 다음 방식으로 측정된다. feature로 가능한 모든 coalitions를 만들고 feature value가 입력되었을 때 변화된 기여도의 평균을 계산한 값이다. 하지만 feature의 종류가 늘어날수록 가능한 coalition의 수는 기하급수적으로 증가하여, 몇 가지 sample에 대해서만 기여도를 계산하여 사용한다. 이 두 이론을 연결한 SHAP은 예측에 대한 각 feature의 기여도를 계산하여 모델의 예측을 설명하는 것이다. SHAP value가 클수록 예측에 긍정적인 영향을 미쳤다고 해석할 수 있다.

해석이 쉬운 모델의 경우 구조가 단순하거나 정확도가 떨어지는 경우가 많은 반면, 해석이 어려운 모델의 구조는 매우 복잡하고 높은 정확도를 보인다. 여러 분야에서 강력한 학습 알고리즘을 사용함과 동시에 해석 가능성을 확보하기 위해 XAI(eXplainable Artificial Intelligence)를 모델에 적용하는 연구들이 진행되고 있다. 생물학 분야에서는 생물학적 연령과 관련된 연구에서 SHAP을 사용하였고(Wood et al., 2019). 인지 특성의 가변성을 예측하는 연구에서 인지의 신경해부학적 기초를 SHAP을 통해 해석한다(Azevedo et al., 2019). 국내 시중은행들은 비대면 대출 심사에 AI를 도입하였고, AI 기반의 상담 챗봇, 보험금 지급에도 AI를 활용하고 있다. 이렇듯 민감한 결정을 AI가 도입되어 해결하려 하고 있으며 AI가 내리는 것에는 근거가 충분히 마련되어야 실현 가능하다. AI의 예측 근거를 확보하여 명확한 근거를 제시할 수 있어야 하기에 모델의 투명성 확보가 중요하다.

3. 제안 방법

본 연구에서는 SHAP을 통해 IDS의 투명성을 높이는 방법을 제안한다. SHAP으로 내린 모델의 해석을 기반으로 IDS 예측의 근거와 실제 공격 유형과의 연관성을 살펴본다.

제안 방법의 구조는 <그림 1>과 같다. 먼저, AI로 IDS 모델을 구축하고 생성된 모델을 이용해 각 공격 유형별로 SHAP value를 산출한다. 각 공격 유형별 SHAP value를 plot으로 시각화하여 해당 공격으로 판단하는데 주요하게 작용한 feature들을 확인한다. 더불어 feature들과 공격 유형간의 관계 해석을 통해 AI model의 예측을 해석한다.



<그림 1> 제안 방법 구조

3.1 AI-IDS Model

Multi classification을 위해 속도와 정확도 측면에서 우수했던 XGBoost를 사용한다. SHAP의 Explainer의 종류는 TreeExplainer, DeepExplainer, GradientExplainer, LinearExplainer, KernelExplainer 5가지가 있다. KernelExplainer는 Model-agnostic으로 모든 모델에 적용 가능하지만 다른 유형의 알고리즘보다 느리다. 이에 본 논문에서는 정확하고 속도가 빠른 TreeExplainer을 사용한다.

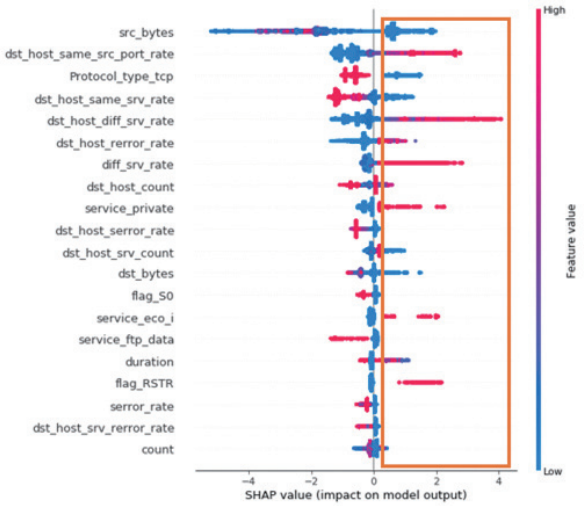
3.2 SHAP Local 해석

SHAP은 <그림 2>와 같은 plot을 제공한다. Data 하나를 모델이 예측할 때 사용된 feature별 SHAP value를 직관적으로 보여준다. 해당 Data가 예측하는데 어느 feature가 얼마나 positive or negative effect가 있었는지 영향도가 큰 순서대로 확인 가능하다.

<그림 2> SHAP Plot for Local interpretation

3.3 SHAP Global 해석

SHAP은 <그림 3>과 같은 plot을 제공한다. <그림 3>는 실험에 사용한 NSL-KDD Train Dataset 중 Probe 공격 유형에 대한 SHAP summary plot이다. 모델이 Data를 분류하는데 기여도가 높은 feature 순서로 내림차순 정렬된다. 표시된 오른쪽 영역은 SHAP value가 0보다 큰 지점이다. 해당 영역에 있는 feature value는 Probe로 판단하는데 Positive effect가 있다고 해석된다.



<그림 3> SHAP Plot for Global interpretation

4. 실험

본 연구에서는 실험에서 사용된 Dataset, IDS 모델 학습 과정, SHAP을 통한 모델 해석을 설명한다. 이 실험의 목표는 모델 해석을 통해 투명성을 확보하고, 모델이 내린 예측에 대한 설명을 제공할 수 있음을 확인한다.

4.1 Dataset

제안 모델의 실험을 위해 NSL-KDD Dataset이 사용되었다. NSL-KDD Dataset은 IDS 실험에 널리 사용되던 KDD'99의 단점을 보완하고 정제된 버전으로, 중복 레코드를 제거하여 빈번한 데이터에 의해 모델이 편향되지 않는다. 더불어 다양한 침입 탐지 방법을 비교하는데 도움이 된다. NSL-KDD의 Train Dataset과 Test Dataset은 <표 2>와 같이 구성되어 있다.

<표 2> NSL-KDD Dataset

Dataset	Total No. of Instance					
	Instance	Normal	Dos	Probe	U2R	R2L
Train	125,973	67,343	45,927	11,656	52	995
Test	22,544	9711	7460	2421	67	2885

NSL-KDD Dataset에는 Normal, DoS, Probe, R2L, U2R과 같이 크게 5가지의 공격 유형으로 분류한다. Nomal은 class 0, DoS는 class 1, Probe는 class 2, R2L은 class 3, U2R은 class 4로 분류하였다. 각각의 공격에 대한 상세 설명과 포함된 공격 유형은 <표 3>과 같다.

<표 3> 공격 유형

Major Categories	Subcategories
Denial of Service(Dos)	네트워크 서비스에 대한 정상적인 액세스를 방해하는 공격으로, 주요 목적은 메모리 자원을 완전히 잡아먹어 네트워크 자원이 정상적인 네트워크 요청을 처리할 수 없게 하여 사용자의 서비스 액세스를 거부시키는 공격 유형 * Ping of Deth, LAND, Neptune, Backscatter, Smurf, Teardrop
Probing	공격 대상을 분석하기 위해 스캔하고 대상 정보를 수집하는 행위로, 공격자는 정확하고 효율적인 공격을 하기 위해 개체에서 알려진 취약점을 찾기 위해 정보를 검색하고 사용함 * Ipsweeping, nmap, Portsweeping, Satan
Remote to Local(R2L)	외부 네트워크에서 로컬 리소스에 불법적으로 액세스하는 공격 유형 * FTP-write, Password guessing, Imap, Multi-hop, phf, spy, Warezclient, warezmaster
User to Root(U2R)	일반적인 사용자 권한을 관리자 권한으로 불법 승격시키는 공격 유형 * Buffer Overflow, Loadmodule, Perl, Rootkit

4.2 Data Preprocessing

NSL-KDD Dataset에는 41가지 feature가 있다. feature 유형은 크게 Binary, Symbolic, Continuous 3가지로 나뉜다. Binary features는 Land, logged_in, root_shell, su_attempted, Is_hot_login, Is_guest_login 총 6가지로 별다른 전처리 과정 없이 그대로 사용하고, Symbolic features는 Protocol_type, Service, Flag로 feature value의 종류는 <표 4>와 같다. Symbloic feature의 unique한 값은 One Hot Encoder를 거쳐 Protocol Type은 3가지, Service는 70가지, Flag는 11가지의 feature로 변환된다. Flag 종류에 대한 설명은 <표 5>와 같다. Continuous features는 그 외 32가지로 Min-Max Scaling으로 전처리 과정을 거쳐 사용된다. 전처리 후 41가지 feature는 122개의 feature로 변환된다.

<표 4> Symbolic features의 One Hot Encoder

Protocol Type	Service					Flag
TCP	aol	exec	klogin	pop_2	telnet	OTH
UDP	auth	finger	kshell	pop_3	tftp_u	REJ
ICMP	bgp	ftp	ldap	printer	tim_i	RSTO
	courier	ftp_date	link	private	time	RSTOS0
	csnet_ns	gopher	login	red_i	whois	RSTR
	ctf	garvest	name	mtp	smtp	S0
	daytime	hostnames	netbios_dgm	remote_job	urh_i	S1
	discard	http	netbios_ns	uucp	urp_i	S2
	domain	http_2784	netbios_ssn	uucp_path	rje	S3
	domain_u	http_443	netstat	sql_net	shell	SF
	echo	http_8001	nnsp	ssh	vmnet	SH
	eco_i	imap4	nntp	sunrpc	X11	
	ecr_i	IRC	ntp_u	supdup	Z39_50	
	efs	iso_tsap	pm_dump	systat	other	

<표 5> Flag feature 설명

Flag	Descrption
OTH	No SYN seen, just midstream traffic (a “partial connection” that was not later closed)
REJ	Connection attempt rejected
RSTO	Connection established, originator aborted (sent a RST)
RSTOS0	Originator sent a SYN followed by a RST, we never saw a SYN-ACK from the responder
RSTR	Responder sent a RST
S0	Connection attempt seen, no reply
S1	Connection established, not terminated
S2	Connection established and close attempt by originator seen (but no reply from responder)
S3	Connection established and close attempt by responder seen (but no reply from originator)
SF	Normal establishment and termination. Note that this is the same symbol as for state S1. You can tell the two apart because for S1 there will not be any byte counts in the summary, while for SF there will be
SH	Originator sent a SYN followed by a FIN, we never saw a SYN ACK from the responder (hence the connection was “half” open)

4.3 IDS Model

NSL-KDD Dataset을 가지고 여러 Classifier를 비교한 연구에서 XGBoost의 정확도가 가장 우수하여 XGBoost를 사용하여 실험한다. XGBoost에서 사용한 파라미터는 <표 6>과 같다.

<표 6> XGBoost Parameters

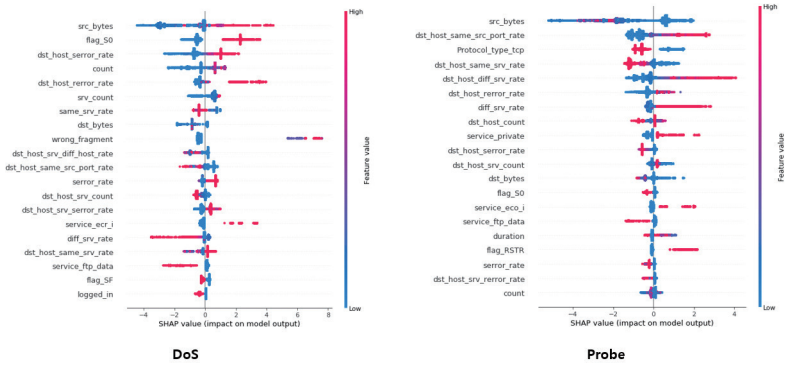
Parameter	setting	Parameter	setting
learning_rate	0.3	min_child_weight	1
max_depth	3	seed	0
subsample	1	objective	multi:softprob
colsample_bytree	1	num_class	5

4.4 Global 해석 결과

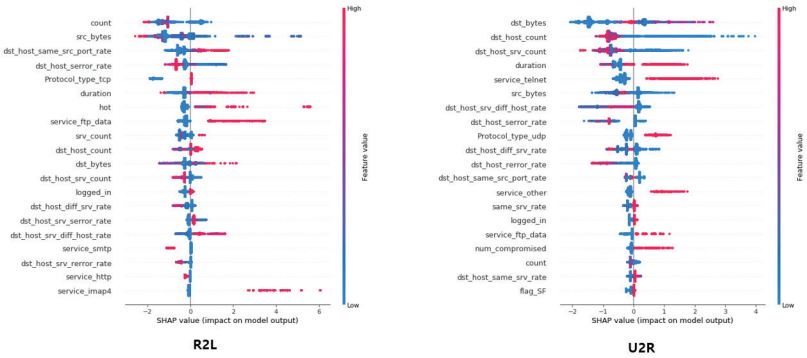
<그림 4>, <그림 5>은 Train Dataset으로 학습시킨 AI에서 추출한 공격 유형별 SHAP summary plot이다. 명확한 해석이 가능한 그래프는 feature value에 다른 SHAP value는 넓은 영역에 분산되어 있지 않고 특정 구간에 군집되어 있어야 한다. feature value가 낮을 때 SHAP value가 높다면 반대로 feature value가 높을 때 SHAP value는 낮게 나타나야 명확한 해석이 가능하다.

dst_host_diff_srv_rate를 해석한 내용은 다음과 같다. dst_host_diff_srv_rate는 동일한 목적지에 대한 연결 사이에 다른 서비스들을 사용한 비율을 의미한다. dst_host_diff_srv_rate의 feature value가 높을수록 SHAP value가 높고, feature value가 낮을수록 SHAP value가 낮은 것이 확인된다. 즉 동일한 목적지에 대한 연결 사이에 다른 서비스를 사용한 비율이 높을수록 Probe로 판단하는 Positive effect가 있고 반대로, 비율이 낮을수록 Negative effect가 있다고 해석된다. Probe 공격 중 nmap과 같은 경우 정보 수집을 위해 여러 서비스로 접근을 시도하는 경우 dst_host_diff_srv_rate의 feature value가 높게 측정될 것이다. 결론적으로 dst_host_diff_srv_rate는 Probe 공격 유형과 관련성 있는 feature이며, feature value에 따른 SHAP value가 구분이 명확해 모델이 data에 해당 feature를 적절하게 사용하고 있다고 해석된다. 반대로 R2L의 src_bytes는 summary plot에서 별다른 특징을 해석할 수 없다. 같은 feature value임에도 SHAP value가 일관성이 없는 것을 볼 수 있다. src_bytes는 R2L을 예측하는데 있어 noise로 작용했을 가능성이 있다고 볼 수 있다.

위와 같이 SHAP을 통해 공격 유형별 feature 해석을 제공할 수 있다. <그림 2>와 같은 IDS가 예측한 Test Data의 local 해석을 Train Dataset에서 나온 global 해석과 대조하여 IDS의 예측에 근거를 제공할 수 있다.



<그림 4> DoS & Probe summary plot



<그림 5> R2L & U2R summary plot

5. 결론

네트워크에 의존하는 서비스가 증가하고 네트워크 사용량이 급증하고 있는 만큼 네트워크 보안도 중요해지고 있다. IDS와 AI의 접목은 활발히 연구되고 있는 분야이다. 하지만 정확한 예측을 위해 AI의 모델은 점점 더 구조가 복잡해지고 불투명성의 문제를 안고 있다. 모델의 불투명성은 모델이 내린 예측을 신뢰하고 적극적으로 사용하는데 방해가 된다.

본 연구에서는 AI로 구축한 IDS를 SHAP을 통해 설명하는 방법에 대한 연구를 진행했다. SHAP은 IDS가 내린 예측의 이유를 설명하고 이해하는데 기여한다. IDS의 투명성을 확보하여 예측을 신뢰하고 적극적으로 사용할 수 있는 근거를 제공한다. SHAP은 어느 모델에도 적용 가능하여 보안 분야뿐 아닌 판단 근거를 중요시하는 금융, 의료분야에도 적용 가능할 것으로 기대한다.

참고문헌

Azevedo, T., Passamonti, L., Lió, P., & Toschi, N. (Eds.). (2019). A Machine Learning Tool for Interpreting Differences in Cognition Using Brain Features. Berlin: Springer, Cham.

Canadian Institute for Cybersecurity (2009). NSL-KDD Dataset. Retrieved February 03, 2021, from https://www.unb.ca/cic/datasets/nsl.html

Dhaliwal, S. S., Nahid, A. A., & Abbas, R. (2018). Effective Intrusion Detection System Using XGBoost. Information, 9(7), 149. doi:10.3390/info9070149

Kim, A., Park, M., & Lee, D. H. (2020). AI-IDS: Application of Deep Learning to Real-Time Web Intrusion Detection. IEEE Access, 8, 70245-70261. doi: 10.1109/ACCESS.2020.2986882.

Marino, D. L., Wickramasinghe, C. S., & Manic, M. (2018). An Adversarial Approach for Explainable AI in Intrusion Detection Systems. IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society., Washington, DC, USA. doi:10.1109/IECON.2018.8591457

Molnar, Christoph (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable. Retrieved February 10, 2021, from https://christophm.github.io/interpretable-ml-book/

Wang, M., Zheng, K., Yang, Y., & Wang, X. (2020). An Explainable Machine Learning Framework for Intrusion Detection Systems. IEEE Access, 8, 73127-73141. doi:10.1109/ACCESS.2020.2988359

Wood, T. R., Kelly, C., Roberts, M., & Walsh, B. (2019). An interpretable machine learning model of biological age. F1000Research, 8(17), 17. doi:10.1109/ACCESS.2017.2762418

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. IEEE Access, 5, 21954-21961. doi:10.1109/ACCESS.2017.2762418