Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.31, NO.6, Dec. 2021 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2021.31.6.1279

Intrusion Detection System을 회피하고 Physical Attack을 하기 위한 GAN 기반 적대적 CAN 프레임 생성방법*

김 도 완,[1†] 최 대 선[2‡]
[1,2]숭실대학교 (대학원생, 교수)

GAN Based Adversarial CAN Frame Generation Method for Physical Attack Evading Intrusion Detection System*

Dowan Kim,[1†] Daeseon Choi[2‡]
[1,2]Soongsil University (Graduate student. Professor)

요 약

차량 기술이 성장하면서 운전자의 개입이 필요 없는 자율주행까지 발전하였고, 이에 따라 차량 내부 네트워크인 CAN 보안도 중요해졌다. CAN은 해킹 공격에 취약점을 보이는데, 이러한 공격을 탐지하기 위해 기계학습 기반 IDS가 도입된다. 하지만 기계학습은 높은 정확도에도 불구하고 적대적 예제에 취약한 모습을 보여주었다. 본 논문에서는 IDS를 회피할 수 있도록 feature에 잡음을 추가하고 또한 실제 차량의 physical attack을 위한 feature 선택 및 패킷화를 진행하여 IDS를 회피하고 실제 차량에도 공격할 수 있도록 적대적 CAN frame 생성방법을 제안한다. 모든 feature 변조 실험부터 feature 선택 후 변조 실험, 패킷화 이후 전처리하여 IDS 회피실험을 진행하여 생성한 적대적 CAN frame이 IDS를 얼마나 회피하는지 확인한다.

ABSTRACT

As vehicle technology has grown, autonomous driving that does not require driver intervention has developed. Accordingly, CAN security, an network of in-vehicles, has also become important. CAN shows vulnerabilities in hacking attacks, and machine learning-based IDS is introduced to detect these attacks. However, despite its high accuracy, machine learning showed vulnerability against adversarial examples. In this paper, we propose a adversarial CAN frame generation method to avoid IDS by adding noise to feature and proceeding with feature selection and re-packet for physical attack of the vehicle. We check how well the adversarial CAN frame avoids IDS through experiments for each case that adversarial CAN frame generated by all feature modulation, modulation after feature selection, preprocessing after re-packet.

Keywords: Vehicle Privacy, Evasion Attack, Machine Learning, Adversarial Example

Received(10. 13. 2021), Accepted(10. 29. 2021)
* 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격탐지기술 개발)과 한국연구재단의 지원을 받아 수행된 연구임(No.2021R1A4A1029650, 자율주행 자동차 보안 기초연구실)
† 주저자, dhdp200@soongsil.ac.kr
‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

I. 서 론

자동차의 기술이 빠르게 성장하면서 이에 따라 운전자에 대한 편의성이 증가하여 운전자의 개입이 필요 없는 자율 주행기술의 2단계까지 발전하였다[1]. 현대의 자동차는 차량 내부의 전자 시스템을 제어할 수 있도록 다수의 ECU(Electronic Control Unit)가 포함되어 있는데, ECU는 차량 내부 통신 규격인 CAN(Controller Area Network)[2] 을 사용하여 차량을 제어할 수 있다. 또 자율주행의 핵심기술인 V2X(Vehicle to everything)[3] 기술이 차량과 주변 사물 간의 상호작용되면서 그에 따른 CAN 보안도 중요해졌다. 하지만 CAN은 개발될 당시에 보안 위협이 있을 것이라고 상정하기 어려워 암호학적 보안이 적용되어 있지 않아서 CAN 메시지 해킹 공격에 취약하다는 문제점이 존재한다[4]. 이러한 공격을 탐지하기 위해서 IDS(Intrusion Detection System) 기술을 도입하였고 현재는 머신러닝 및 딥러닝이 발달함에 따라 기계학습 기반 IDS 연구[5, 6]도 활발히 진행되고 있다.

그러나 C. Szegedy[7]는 딥러닝 모델이 높은 분류 정확도를 보임에도 불구하고 데이터에 사람 눈에 띄지 않을 정도의 아주 작은 잡음(Noise)을 추가만 해도 데이터 분류 결과가 달라지는 등 딥러닝 모델이 취약하다는 것을 보여주었다. 이러한 적대적 예제는 사이버 보안 위협 등의 이유로 관심사 연구 중 하나가 되었고, CAN IDS에 대해서 CAN 패킷을 처리한 feature에 아주 작은 잡음을 추가하여 IDS를 회피하는 연구[8]도 진행되었다.

하지만 feature vector 상에서 제약이 없는 실수 변조를 하여 IDS는 회피할 수 있지만, 실제 차량에 주입 시 변조된 잡음은 사라지고 변조로 인해 arbitration ID나 데이터 필드가 다른 값으로 변할 수 있어 공격 자체의 의미가 사라지는 등 physical attack이 불가능한 한계점이 존재한다.

따라서 본 논문에서는 공격의 의미를 사라지지 않도록 전처리를 거친 패킷 feature 중 공격 기능과 상관이 없다고 판단되는 일부를 추출 및 변조하여 공격이 가지는 의미는 없어지지 않고 IDS를 회피하며 최종적으로 실제 차량에 physical attack을 할 수 있는 GAN 기반 적대적 CAN 패킷 프레임 생성방법을 제안한다. 목표는 기계학습 기반 IDS 모델에 의해 잘 탐지되던 공격이 GAN을 통하여 공격이 정상으로 분류되게끔 변조를 가하여 IDS 모델이 공격 탐지를 잘 하지 못하게 만들고 실제 차량에 주입할 수 있도록 CAN 패킷 프레임을 생성방법을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 배경 및 관련 연구에 대하여 설명하고, 3장에서는 CAN 네트워크 패킷에 대한 적대적 예제 생성에 대하여 설명한다. 4장에서는 생성한 적대적 CAN 네트워크 패킷을 이용하여 IDS 회피실험 및 평가를 진행한다. 5장에서는 실제 차량에 대한 Physical attack을 위해 frame 생성방법을 설명하고 마지막 6장에서 고찰 및 향후 연구 계획을 서술한다.

II. 배경 지식 및 관련 연구

2.1 CAN Data Frame

1986년, 독일의 Bosch사에서 개발된 프로토콜인 CAN(Controller Area Network)은 차량 내부의 ECU(Electronic Control Unit)들이 서로 통신하기 위해서 설계된 표준 통신 규격이다. CAN 네트워크는 ECU가 단일의 CAN 인터페이스를 보유함으로써 차량의 비용과 중량을 줄일 수 있다는 장점이 있어 차량업계에서 신속하게 CAN을 도입하였다. CAN 프로토콜의 특징으로는 메시지의 우선순위에 따라 ID를 할당하고, ID를 이용하여 메시지를 구별한다. CAN은 데이터 프레임, 리모트 프레임, 에러 프레임, 오버로드 프레임 4가지 프레임 타입이 있다. 데이터 프레임은 데이터 전송에 사용하고, 리모트 프레임은 메시지 수신 노드에서 원하는 메시지를 전송할 수 있는 전송 노드에게 전송을 요청할 때 사용된다. 에러 프레임은 메시지 에러가 감지되었을 때 시스템에 알려준다. 오버로드 프레임은 메시지 동기화에 사용된다. 이러한 CAN은 메시지 프레임을 사용하여 데이터 송수신을 한다. CAN 메시지 프레임 구조는 Fig. 1.과 같다. SOF(Start Of Frame)는 메시지의 처음을 지시하고, Arbitration Field는 11비트의 ID와 1비트의 RTR(Remote Transmission Request)로 구성된다. ID는 메시지의 우선순위를 지정하고 RTR은 리모트 프레임인지 데이터 프레임인지 결정하는 데 사용된다. Control Field는 데이터 길이(Bytes 수)를 알려주는 DLC(Data Length Code)가 있다. Data Field는 8 Bytes까지 데이터를 저장할 수 있다. CRC(Cyclic Redundancy Check) Field는 메

Fig. 1. CAN Data structure

시지 에러 유무를 검사하는 데 사용하고 ACK(Acknowledgement) Field는 전송 노드에서 ACK 비트 유무를 확인하고 없을 시 재전송을 하여 버스 작동 영향을 주지 않게 한다. EOF(End Of Frame)는 메시지의 끝을 알려주는 역할을 한다. 본 논문에서는 Arbitration Field 중 ID, Control Field 중 DLC, Data Field의 Data를 사용한다.

2.2 IDS

IDS(intrusion Detection System)는 호스트나 네트워크에서 비정상적인 사용이나 남용 등에 대한 정보를 실시간으로 수집, 분석하여 침입의 징후를 탐지하여 보고하는 시스템을 의미한다[9]. IDS는 오용탐지(Misuse Detection)와 비정상행위탐지(Anomaly Detection)로 나누어지는데, 오용탐지는 이미 알려진 침입행위에 대한 정보를 분석한 비정상적인 패턴과 비교하여 일치하거나 유사한지 인식하여 탐지한다. 오탐률(False positive rate)이 낮지만, 새로운 공격의 패턴은 탐지하기 어렵다는 단점이 있다. 비정상행위탐지는 미리 시스템이나 네트워크에 정상적인 입력패턴을 규정해두고 이 패턴과 비교하여 벗어나게 되면 침입이라고 판단하는 방법이다. 오용탐지와 다르게 새로운 패턴의 공격은 탐지할 수 있지만 오탐률이 높다는 단점이 있다.

M.J Kang[10]은 DNN(Deep Neural Network)[11]을 사용하여 정상 및 공격 패킷의 통계적 특성을 파악하고 feature를 추출하여 공격을 식별하였다. 특히 파라미터를 효율적으로 학습하기 위해 DBN(Deep Belief Network)[12] 알고리즘을 사용하였다. 로지스틱 값 1과 0의 형태로 정상 패킷과 공격 패킷을 분류할 수 있는 확률을 출력한다. Spoofing 공격을 적용한 TPMS(Tire Pressure Monitoring System) 패킷을 사용하여 검증했을 때, 99% 검출비율을 보여준다.

E. Seo[6]는 GAN(Generative Adversarial Network)[13] 기반 IDS를 제안하였다. CAN의 Arbitration ID를 one-hot encoding 방식을 사용하여 이미지로 변환한다. 제안한 모델은 2개의 Discriminator와 Generator로 구성되는데, 첫 번째 Discriminator는 이미 알려진 공격을 탐지하는 용도로 사용되고 두 번째 Discriminator는 새로운 공격을 탐지하기 위해 Generator를 사용하여 가짜 CAN 이미지를 생성한 후 가짜 CAN 이미지를 탐지하도록 한다. 훈련시킨 IDS 모델은 DoS, Fuzzy, RPM, Gear 공격에 대하여 모두 99% 이상의 탐지율을 보여준다.

A.Rehman[14]은 단일 공격과 혼합 공격을 탐지하기 위하여 CNN(Convolutional Neural Network)[15]과 AGRU(Attention based Gated Recurrent Unit)[16]를 조합한 CANintelliIDS를 제안하였다. CANintelliIDS는 CAN의 Data Field의 Data가 CNN에 입력으로 주입되고 AGRU에 입력으로 들어가기 위해 feature를 추출하게 된다. AGRU는 가중치를 조절하기 위하여 Data의 상관관계 측면과 전후 관계 정보를 기반으로 학습한다. 해당 벡터가 공격인지 정상인지 예측하는 컨텍스트 벡터 8을 생성한다. 해당 모델은 2개 계층의 CNN(128, 64)과 3개 계층의

Table 1. Random Forest performance

Attack type	Type	Precision	Recall	F1-score
Flooding	Normal [39,521]	1.0	1.0	1.0
	Attack [38,521]	1.0	1.0	1.0
Fuzzing	Normal [23,620]	0.99	0.99	0.99
	Attack [22,620]	0.99	0.99	0.99
Replay	Normal [11,509]	0.91	0.97	0.94
	Attack [10,509]	0.96	0.90	0.93
Spoofing	Normal [1,988]	1.0	1.0	1.0
	Attack [988]	1.0	1.0	1.0

ARGU(250, 150, 150)으로 구성된다. 이 모델의 성능은 다른 IDS 모델과 비교하여 최소 5.32% 정도 차이가 나는 성능을 보여준다.

본 논문에서는 머신러닝의 한 종류인 랜덤 포레스트(Random Forest)[17]를 사용하여 IDS를 수행한다. 랜덤 포레스트는 다수의 의사결정 나무(Decision Tree)를 학습하여 의사결정 나무들의 결과를 취합하여 예측하는 앙상블 기법이다. 랜덤 포레스트는 총 164개의 feature를 학습하여 Table 1.에 CAN 네트워크 공격에 대한 성능을 나타낸다.

2.3 CAN 주입 공격의 종류

CAN에 대한 주입 공격 시나리오는 Flooding 공격, Fuzzing 공격, Replay 공격, Spoofing 공격 4가지로 구성된다[18]. Flooding 공격은 우선순위가 가장 높은 CAN 패킷 메시지를 대량으로 주입하여 다른 CAN 패킷 메시지가 작동하지 못하도록 하는 공격이다. Fuzzing 공격은 임의로 선택한 Arbitration ID에 무작위 데이터를 주입하는 공격에 해당한다. Replay 공격은 정상 CAN 패킷 메시지를 일정 시간 동안 추출하고 추출한 메시지를 다시 주입하는 공격이다. Spoofing 공격은 공격자가 특정 arbitration ID에 원하는 공격이 발생할 수 있도록 데이터를 선택적으로 주입하는 공격이다.

2.4 적대적 예제

적대적 예제는 크게 공격 목표에 따른 적대적 예제와 공격자의 지식에 따른 적대적 예제 2가지로 나눌 수 있다.

공격 목표에 따른 적대적 예제는 생성된 적대적 예제에 대한 딥러닝 모델의 결과가 공격자가 의도하는 class로 분류되도록 하는 공격을 표적 공격(Targeted attack)이라고 한다. 무표적 공격(Untargeted attack)은 생성된 적대적 예제에 대한 딥러닝 모델의 결과가 원래 입력 데이터에 대한 class가 아닌 임의의 다른 class로 분류되도록 하는 공격을 의미한다.

공격자의 지식에 따른 적대적 예제 공격은 White-box 공격과 Black-box 공격으로 나누어진다. White-box 공격은 공격자가 기계학습 모델의 구조나 파라미터 등 모든 정보를 아는 상태에서 공격하는 것을 의미하고, Black-box 공격은 공격자가 공격할 대상 모델에 대해 입력값에 대한 출력값(분류 정보)만 알 수 있는 환경에서 공격하는 것을 의미한다.

적대적 공격방법 중 하나인 FGSM(Fast Gradient Sign Method)[19]은 잡음의 크기를 나타내는 ϵ을 설정하여 손실함수의 기울기(Gradient)를 정답 label인 y의 반대되는 방향으로 업데이트하여 오분류를 발생시킬 수 있는 적대적 예제를 생성하게 된다.

$$x_{adv} = x - \epsilon \cdot sign(\nabla L_{F,t}(x)) \quad (1)$$

두 번째로, Carlini Wagner(CW)[20] 공격은 White-box 공격 중에서 100%의 성공률을 보인다. L_p 기반 distance metric을 최소화하여 최적의 적대적 예제를 찾는 방법을 제안하였다. 7가지의 목적함수와 3가지의 L_p 기반 distance metric을 테스트하여 최적화 방안을 찾는다.

세 번째는 네트워크 적대적 공격으로써 Zilong Lin[21]은 IDS를 회피하는 IDSGAN을 제안하였다. NSL-KDD 데이터를 사용하여 GAN을 기반으로 Denial of Service(DoS), User to Root(U2R), Root to Local(R2L) 공격을 IDS 회피하도록 재생성하였다. 머신러닝 및 딥러닝 모델 총 7가지 IDS 모델을 공격하는데, 탐지율(Detection Rate)과 회피증가율(Evasion Increase Rate)을 사용하여 성능을 평가한다. 원본 공격탐지율(Original DR)은 IDS가 CAN 공격 메시지를 얼마나 잘 탐지하는지를 나타내고 적대적 공격탐지율(Adversarial DR)은 IDS가 GAN으로 생성한 적대적 CAN 공격 메시지를 얼마나 잘 탐지하는지를 나타낸다. 회피증가율은 원본 탐지율과 적대적 공격탐지율을 비교함으로써 적대적 CAN 공격 메시지가 IDS를 얼마나 잘 회피하는지 정도를 나타낸다. DoS 공격에 대하여 원본 탐지율은 평균 79% 이상을 나타내지만, 적대적 공격탐지율은 평균 0.97%, 회피증가율이 평균 98.8%를 나타내면서 대부분의 공격이 회피했음을 입증했다. U2R 및 R2L 공격에서는 원본 탐지율이 평균 4.7% 정도인데, 이는 학습 데이터셋에서 U2R 및 R2L의 양이 적어서 탐지율이 낮다고 언급한다. 적대적 공격탐지율은 평균 0.02%, EIR은 평균 99.5%를 보이면서 생성된 적대적 공격이 IDS를 대부분 회피하였음을 확인하였다.

Timestamp	Arbitration ID	DLC	Data	Class	Subclass
1599046456.773016	367	8	00 00 00 00 05 00 DF 10	Normal	Normal
1599046456.773254	368	8	00 0B 50 00 02 7A 0E 44	Normal	Normal
1599046456.773501	479	8	52 00 00 00 00 00 00 00	Normal	Normal
1599046456.773743	130	8	20 7D A8 7E 00 00 0F 60	Normal	Normal
1599046456.773982	140	8	00 7F 00 71 20 00 0F 4E	Normal	Normal
1599046456.774212	251	8	04 03 B4 9C 00 3C 37 7B	Normal	Normal

Fig. 2. Data structure

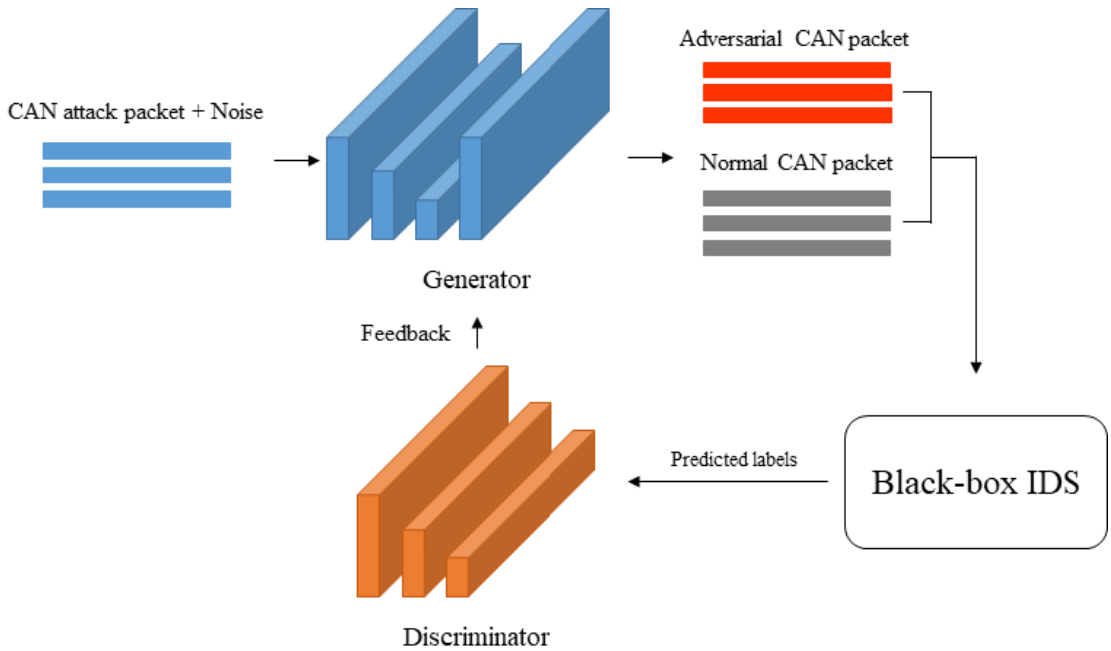

Fig. 3. GAN structure

III. CAN 적대적 예제 생성

본 장에서는 GAN을 기반으로 IDS를 회피할 수 있는 적대적 CAN feature를 생성하는 과정을 설명한다. 적대적 CAN feature가 IDS 모델을 회피하는지 확인하기 위해 전처리를 거친 모든 feature에 잡음(Noise)을 추가하는 시나리오와 생성한 적대적 CAN feature를 실제 차량에 대해 공격하기 위한 physical attack 시나리오로 나누어 설명한다.

3.1 데이터셋 및 데이터 전처리

본 논문에서 사용한 데이터셋은 2020년도 사이버 보안 챌린지 자동차 해킹 공격/방어 부분에서 제공한 예선 데이터를 사용한다[22]. 데이터의 형태는 Fig. 2.와 같다. Timestamp는 CAN 메시지가 로깅되는 시간을 나타낸다. Arbitration ID는 CAN 메시지의 식별 ID이며, DLC는 Data의 크기(Bytes 수)를 의미한다. Data는 CAN 메시지의 데이터가 적재된 필드이고 Class와 Subclass는 각각 해당 CAN 메시지가 정상인지 공격인지와 공격에 해당하면 어떤 공격에 해당하는지를 나타내준다. CAN 네트워크 패킷의 전처리는 one-hot encoding과 Min-Max Scaler를 주로 사용하여 진행한다. Min-Max Scaler는 데이터 스케일링 중 하나로 모든 feature가 0부터 1 사이에 존재하도록 재조정한다.

$$MinMax(x) = \frac{x - min(x)}{max(x) - min(x)} \quad (2)$$

데이터 중 Arbitration ID는 16진수로 이루어져 있고 Data는 8바이트로 이루어져 있어, ID는 one-hot encoding으로 처리하였고, Data는 1바이트를 비트로 변환하였다. 이전 Arbitration ID도 현재 ID와 똑같이 one-hot encoding을 사용하였고, 이전 같은 ID와의 시간 차이, 이전 같은 데이터 간의 시간 차이, 1000개의 패킷 중 같은 ID의 개수, 1000개 패킷 중 같은 데이터 개수 이 4개의 feature는 Min-max scaler를 적용하여 0부터 1 사이의 값으로 변환하여 사용한다. 따라서 사용되는 모든 feature 개수인 164개를 사용하여 모델을 학습 및 테스트한다.

3.2 IDS를 회피하는 CAN 적대적 예제 생성

IDS를 회피하는 CAN 적대적 예제 생성은 2.4절 관련 연구 중 적대적 예제에서 서술한 Zilong Lin의 Wasserstein GAN[23] 구조를 인용한다. GAN의 구조는 Fig. 3.에 나타낸다. Generator는 5층 신경망으로 구성되어 있고 각 층의 출력에는 ReLU 활성화 함수를 사용한다. 마지막 출력층은 입력 feature와 같은 차원의 feature를 출력하도록 설정한다. Discriminator는 5층의 신경망으로 구성되어 있고 각 층의 출력에는 LeakyReLU 활성화 함수를 사용한다. 먼저 CAN 원본 공격 패킷 feature에 0부터 1 사이의 잡음을 더한 값을 Generator에 입력으로 넣어준다, Generator는 적대적 CAN 패킷 feature를 출력하게 되고 정상 CAN 패킷 feature와 함께 Black-box IDS에 주입한다. IDS는 해당 feature가 정상인지 공격인지 분류하면서 feature에 대한 label 값을 출력한다. Discriminator는 Black-box IDS가 출력한 label 값을 기반으로 적대적 CAN feature와 정상 feature가 공격인지 정상인지 판단하고 Generator에 피드백한다. Generator에 대한 손실함수는 다음과 같이 나타낸다.

$$L_G = E_{M \in S_{attack}, N} D(G(M,N)) \quad (3)$$

여기서, M은 CAN 네트워크에서 원본 공격 패킷의 feature, N은 0부터 1 사이의 잡음을 나타낸다. G와 D는 각각 Generator와 Discriminator를 의미한다. Generator는 L_G을 최소화하는 것을 목표로 한다.

Discriminator는 Black-box IDS가 출력한 label과 Discriminator의 출력값으로 손실 값을 측정한다, Discriminator에 대한 손실함수는 다음과 같다.

$$L_D = E_{S \in B_{normal}} D(S) - E_{S \in B_{attack}} D(S) \quad (4)$$

S는 Discriminator에 입력되는 정상 및 적대적 CAN feature를 의미하고, B_{normal}은 Black-box IDS가 예측한 정상 CAN feature, B_{attack}은 Black-box IDS가 예측한 공격, 즉 적대적 CAN feature를 의미한다. Discriminator는 정상 CAN feature와 적대적 CAN feature의 차이를 최소화하는 것이 목표이다.

학습을 반복하면서 최종적으로 Generator는 정상 feature와 비슷한 적대적 공격 feature를 생성하게 된다.

3.3 Feature 기반 적대적 CAN 패킷 생성

Feature 기반 적대적 CAN 패킷 생성은 전처리를 거친 모든 CAN feature에 0부터 1 사이의 잡음(Noise)을 추가하여 Generator에 입력 데이터로 들어가게 한다. Generator는 164개의 feature를 생성하고 Black-box IDS에 입력 데이터로 주게 된다. Black-box IDS는 해당 패킷이 정상인지 공격인지 분류하고 분류한 label 값을 출력한다. Black-box IDS가 출력한 label을 기반으로 Discriminator는 Black-box IDS와 마찬가지로 정상인지 공격인지 판단하고 Generator에 피드백을 주게 된다. 학습을 진행하면서 최종적으로 Generator는 Black-box IDS를 회피할 수 있는 CAN feature를 생성하게 된다.

하지만 변조된 CAN feature는 0부터 1 사이의 실수 값을 가져서 IDS는 회피할 수 있지만, 실제 차량을 대상으로 공격은 불가능하다. 변조한 feature를 다시 패킷으로 복원해야 하는데, 실수 상태인 feature가 정수로 변환되면서 가지고 있던 잡음이 사라져 변조한 의미가 없어지거나 예를 들어, 원래 Replay였던 공격이 변조되면서 Replay 공격이 가

지는 추출된 패킷을 다시 흘려보내는 공격이 아닌 아예 다른 새로운 공격이 될 가능성이 있어 공격의 의미가 달라질 수 있고, 변조해도 해당 공격이 발현되지 않을 수 있다는 한계점이 존재한다.

3.4 공격 기능을 고려한 feature 추출

모든 feature를 변조하면 공격의 기능이 남아 있지 않을 수 있다는 한계점 때문에 랜덤 포레스트의 Feature importance를 사용한다.

Feature importance는 랜덤 포레스트 모델이 가장 중요하다고 판단되는 feature의 중요도를 추출해주는 기능이다. 각 공격의 feature에서 공격 기능과 상관없다고 판단되는 feature에 0부터 1 사이의 잡음을 추가하여 Generator에 입력한다. 4가지 공격에 대한 feature importance는 Fig. 4-7.에 나타낸다.

Flooding 공격은 우선순위가 높은 ID를 대량으로 전송하는 공격이기 때문에 랜덤 포레스트는 현재 Arbitration ID와 통계 feature를 중요하다고 판단한다. Flooding 공격이 가지는 의미를 사라지지 않도록 현재 ID를 제외한 현재 데이터와 통계에 변조를 가하도록 설정한다.

Fuzzing 공격은 임의로 선택한 ID에 무작위 데이터를 주입하는 공격이라서 통계와 현재 데이터 feature를 중요하다고 판단한다. 따라서 선택된 ID 부분을 제외하고 현재 데이터 및 통계를 변조하도록 설정한다.

Replay 공격은 흐르고 있던 정상 CAN 패킷을 일정 시간 동안 추출하고 다시 주입하는 공격이라서 통계를 가장 중요시 본다. Replay 공격은 ID나 데이터가 변조되어 달라진다면 Replay 공격의 의미가 사라질 수 있다는 한계점으로 인해 통계만 변조하도록 설정한다.

Spoofing 공격은 공격자가 임의로 선택한 ID에 원하는 공격이 발생할 수 있도록 데이터를 조작하여 주입하는 공격이라서 통계, 데이터, ID까지 다양하게 퍼져있다. Spoofing 공격도 Fuzzing 공격과 마찬가지로 ID를 제외하고 데이터와 통계를 변조하도록 설정한다. 하지만 데이터 전체를 변조하면 공격자가 원하는 공격의 기능이 발생하지 않을 수 있다. 따라서 Spoofing 공격은 데이터의 8 Bytes 중에서 공격 기능을 가지는 부분을 제외한 나머지 부분을 선택하여 변조하도록 추가로 설정한다.

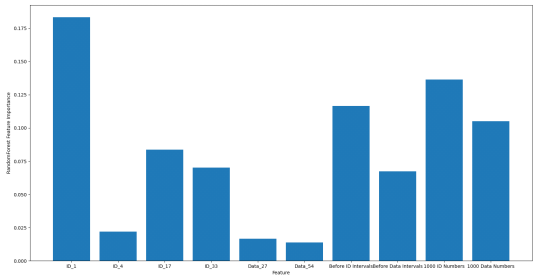

Fig. 4. Flooding attack feature importance

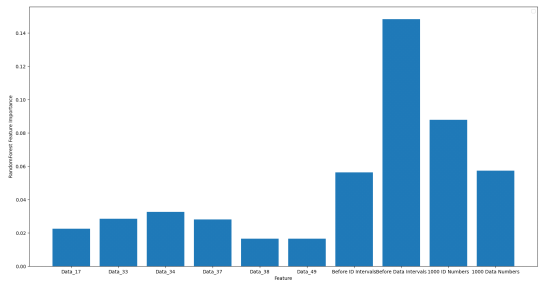

Fig. 5. Fuzzing attack feature importance

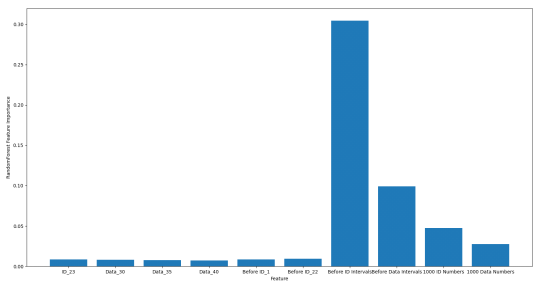

Fig. 6. Replay attack feature importance

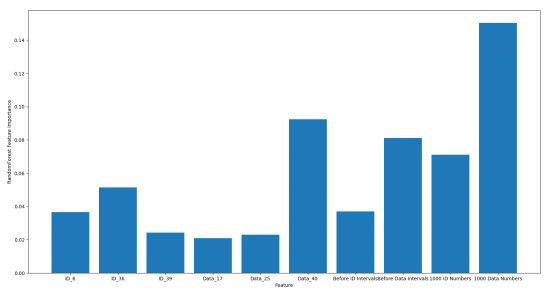

Fig. 7. Spoofing attack feature importance

3.5 Physical attack을 위한 패킷화 적용

선택한 feature를 차량에 주입할 수 있는 패킷 원본 형태로 복원할 수 있도록 복원 방법을 제시한다. 생성한 feature는 실수 상태의 잡음을 포함한 데이터다. 이 데이터를 그대로 차량에 주입하게 되면 잡음이 사라져 변조한 의미가 사라지거나 공격 기능이 바뀌거나 아무런 공격이 발생하지 않을 수 있는 등 문제점이 있을 수 있다. 따라서 Generator에서 출력된 적대적 CAN feature를 패킷화 후 다시 feature로 바꾸는 방법을 선택한다. 패킷화를 하는 방법은 Fig. 8.에 나타낸다.

ID는 one-hot encoding 방식을 사용하였으므로 Generator 이후 ID를 나타내는 48차원에서 각 16차원 중 가장 큰 자리는 1, 나머지는 0으로 복원하였다. 데이터는 0과 1만 가지는 정수 형태 데이터이므로 실수 상태의 ID와 데이터를 수식 (5)을 사용하여 복원한다. 통계는 전처리 과정에서 Min-Max scaler를 적용했기 때문에 생성한 적대적 feature를 수식 (6)을 사용하여 복원한다. 복원한 feature는 원본 패킷의 형태로 변경하고, 복원한 통계 feature는 다시 수식 (2)를 적용하여 Black-box IDS에 입력으로 들어가게 한다.

$$0 < x \leq 1, \quad \lceil x \rceil \tag{5}$$

$$MinMaxInverse(x) = inverse(\frac{x - min(x)}{max(x) - min(x)}) \tag{6}$$

복원한 통계 feature가 패킷화를 할 때, 통계값이 일치하지 않는 현상이 발생할 수 있다. 따라서 공격 ID와 같은 정상 ID에 대한 통계의 평균을 구하여 Generator가 생성한 통계가 정상 feature 통계의 평균으로 수렴하도록 하여 통계가 일치하지 않는 현상을 발생하지 않도록 한다.

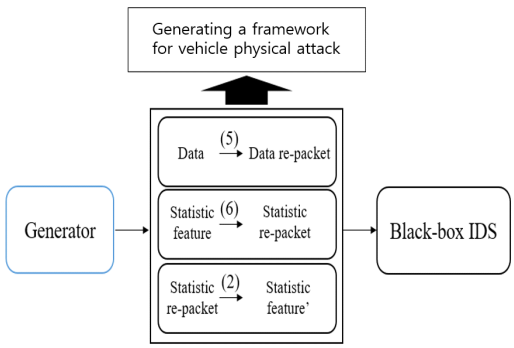

Fig. 8. Re-packet method

II. GAN 기반 적대적 CAN 패킷 생성 실험

4.1 실험 평가

실험결과의 평가방법은 2.4절 관련 연구 중 적대적 예제 부분에서 서술한 Zilong Lin 연구에서 사용한 탐지율(Detection Rate, DR)과 회피증가율(Evasion Increase Rate, EIR)을 사용하여 평가한다.

$$DR = \frac{Num.of\ correctly\ detected\ attacks}{Num.of\ all\ the\ attacks} \tag{7}$$

$$EIR = 1 - \frac{Adversarial\ detection\ rate}{Original\ detection\ rate} \tag{8}$$

탐지율은 Black-box IDS가 공격 메시지 프레임이라고 예측한 모든 공격 메시지 중 진짜 공격 메시지에 대한 비율을 의미한다. 여기서 원본 공격탐지율(Original DR)은 IDS가 원본 공격 CAN 메시지와 정상 CAN 메시지를 얼마나 잘 분류하는지 알려주고 적대적 공격탐지율(Adversarial DR)은 적대적 CAN 메시지와 정상 CAN 메시지를 분류함으로써 IDS가 적대적 CAN 메시지를 얼마나 잘 탐지하는지에 대한 것으로, 수치가 낮을수록 적대적 CAN 메시지를 탐지하지 못하는 것이다. 회피증가율은 원본 공격탐지율과 적대적 공격탐지율을 비교함으로써 적대적 CAN 메시지가 IDS를 얼마나 회피를 하는지에 대한 비율을 나타낸다. 이는 수치가 높을수록 적대적 CAN 메시지가 IDS를 잘 회피한다는 것을 나타낸다.

4.2 실험 설정

데이터에서 정상 패킷이 각 공격 패킷에 비해 더 많은 데이터가 존재하므로 Black-box IDS의 학습 및 테스트 데이터의 정상:공격 비율 1:1로 설정하여 학습 및 테스트를 진행한다. IDS로 사용한 랜덤 포레스트의 트리 개수는 100개로 설정하였고 Generator와 Discriminator는 epoch 25, 배치

사이즈를 256으로 설정하여 실험에 사용하였다.

4.3 실험결과

3.3부터 3.5까지 설명한 방법으로 생성한 적대적 CAN의 공격 성능을 확인한다. 3.3 방법으로 생성한 적대적 CAN의 Black-box IDS 회피 성능은 Table 2.에 나타낸다. 모든 feature에 잡음을 추가한 적대적 CAN feature는 모든 공격에 대한 IDS의 adversarial DR이 모두 0%, EIR이 100%인 것을 확인할 수 있다. 이는 모든 적대적 CAN feature가 IDS를 회피한다는 것이다. 하지만 앞에서 설명했듯이 실제 차량에 대해서는 공격이 실패할 수 있다.

3.4 방법으로 생성한 적대적 CAN의 Black-box IDS 회피 성능은 Table 3.에 공격 성공/전체 공격 개수로 나타낸다. Feature Importance를 이용하여 추출한 feature에만 잡음을 추가한 공격은 현재 데이터만 변조하거나 현재 데이터 및 통계를 변조하는 시나리오가 공격이 성공하는 것을 확인할 수 있다.

3.5 방법으로 생성한 적대적 CAN의 Black-box IDS 회피 성능은 Table 4.에 공격 성공/전체 공격 개수로 나타낸다. 3.4 방법으로 추출한 feature에

Table 2. Feature based adversarial attack performance

Attack type	Original DR	Adversarial DR	EIR
Flooding	1.0	0.0	1.0
Fuzzing	0.99	0.0	1.0
Replay	0.97	0.0	1.0
Spoofing	1.0	0.0	1.0

Table 3. Feature importance based adversarial attack performance

Feature	Flooding	Fuzzing	Replay	Spoofing
Data	256/256	256/256	-	256/256
Statistic feature	0/256	3/256	8/256	0/256
Data and Statistic feature	256/256	256/256	-	256/256

Table 4. Re-packet and scale based adversarial attack performance

Feature	Flooding	Fuzzing	Replay	Spoofing
Data	256/256	256/256	-	256/256
Statistic feature	0/256	2/256	24/256	0/256
Data and Statistic feature	256/256	256/256	-	256/256

잡음을 추가하고 Generator로 변조한 값을 재복원 후 다시 scale 적용한다. 사실상 제약을 추가했음에도 불구하고 공격 성능이 떨어지지 않는 결과를 나타낸다.

III. 패킷 프레임 생성

생성한 데이터를 실제 차량에 주입할 경우, 공격이 성공하는 것은 현재 확인이 불가능하다는 한계점이 존재하지만, 추후 연구를 위한 실시간으로 적대적 CAN 메시지를 생성하여 IDS를 회피하고 실제 차량에서 공격의 기능까지 발현 가능할 수 있도록 프레임 워크를 제시한다. 3장에서 사용한 방법으로 생성한 적대적 CAN feature를 원본 데이터 형태로 변환한다. Data만 변환하여 생성한 적대적 CAN 메시지 프레임과 Data 및 통계를 변환하여 생성한 적대적 CAN 메시지 프레임을 Fig. 9.에 나타낸다. Data만 변환한 적대적 CAN 메시지 프레임은 원본 데이터 프레임 중 공격 데이터 프레임의 Data 필드만 변환된 모습을 확인할 수 있다. Data 및 통계를 변환하여 생성한 적대적 CAN 메시지 프레임은 원본 데이터 프레임과 비교했을 때, 공격 데이터의 순서가 통계대로 변경된 모습을 확인할 수 있다. 생성

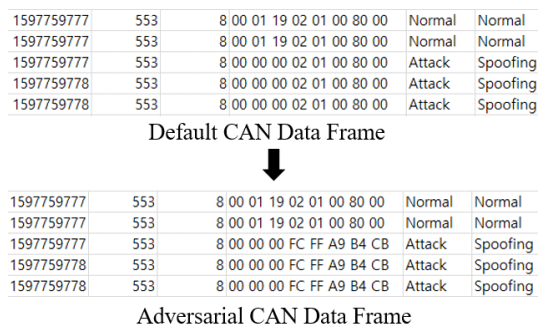

Fig. 9. Adversarial CAN Data Frame

Table 5. Re-packet and preprocessing based adversarial attack performance

Feature	State	Flooding	Fuzzing	Spoofing
Data	Before reconstruction feature	256/256	256/256	256/256
	Reconstruction	256/256	256/256	256/256
Data and Statistic feature	Before reconstruction feature	256/256	256/256	256/256
	Reconstruction	226/256	256/256	253/256

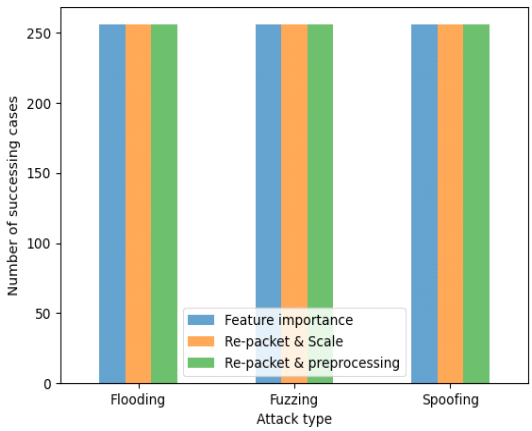

Fig. 10. Adversarial attack performance that modulates only data features

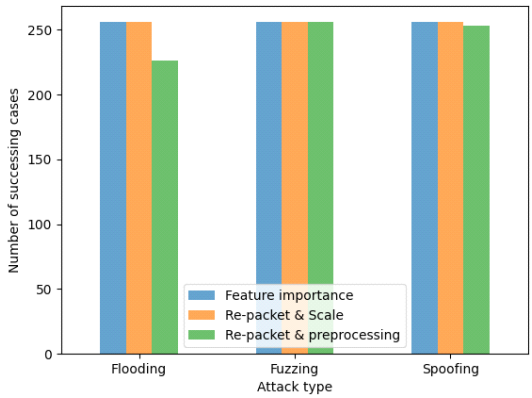

Fig. 11. Adversarial attack performance that modulates data and statistic features

한 적대적 CAN 메시지 프레임을 다시 전처리하여 IDS 모델에 입력하였을 때 결과를 Table 5.에 표시하고 Fig. 10.과 Fig. 11.에 3가지 공격에 대한 데이터 및 데이터와 통계 feature를 변조하여 공격한 실험 3가지 성능을 그래프로 표시한다. 그래프에서 3가지 공격 모두 모든 실험에서 약 88% 이상의 공격 성공률을 보여주는 것을 확인할 수 있다.

IV. 결 론

본 논문은 차량의 Physical attack을 하기 위해 GAN 기반 적대적 CAN frame 생성방법을 제안하였다. 실제 차량에 Physical attack을 할 수 있도록 모든 CAN feature 기반 적대적 CAN feature 생성방법부터 feature importance를 이용한 중요 feature 추출 후 feature 변조 방법, re-packet을 위한 패킷화 적용, packet frame 생성까지 순서대로 실험을 진행하였다.

결과적으로, 처음으로 시행한 모든 feature에 잡음을 추가하여 변조한 실험은 4가지 공격 모두 IDS를 회피하는 결과를 나타내었다. 하지만 차량에 주입할 수 없는 feature의 형태이고 주입할 수 있는 형태로 변환하는 과정에서 GAN으로 생성한 잡음이 사라져 각 공격이 가지는 의미도 사라질 수 있기 때문에 공격의 의미를 없애지 않는 feature만 변조하는 실험을 진행하였다. 그 결과, 통계 feature만 변조하는 공격을 제외한 데이터만 변조, 데이터 및 통계 feature 변조가 IDS를 회피하는 결과를 확인할 수 있었다. 이후 변조한 feature를 다시 패킷화 처리를 진행하고 다시 전처리한 후 IDS에 주입하여 공격이 성공하는지 확인하였다. Flooding, Fuzzing, Spoofing 3가지 공격이 약 88% 이상의 공격 성공률을 보여주는 것을 확인할 수 있었다.

본 논문의 방법으로 생성한 적대적 CAN Frame은 IDS를 회피할 수 있고 뿐만 아니라 실제 차량에 주입하여 공격할 수 있도록 해주는 framework의 첫 사례를 보여준다. 현재 한계점을 극복하기 위해서 향후 연구로 GAN 기반 적대적 CAN frame을 실시간으로 생성할 수 있도록 모델 및 feature 패킷화의 경량화와 생성한 적대적 CAN 메시지를 실제 차량에 흐르는 CAN 메시지대로 실시간으로 주입할 수 있는 기술 연구를 수행하고, 이후 공격한 차량에서 메시지를 다시 추출하여 IDS 모델을 회피하는지 확인하는 실험 진행할 예정이다.

References

[1] "Taxonomy and definitions for term related to driving automation systems for on-road motor vehicles," SAE International in United States, Apr. 2021.

[2] R.B. GmbH, "CAN specification version 2.0," Sep. 1991.

[3] Jian Wang, Yameng Shao, Yuming Ge, and Rundong Yu, "A survey of vehicle to everything(V2X) testing," Sensors 2019, vol. 19, no. 2, Jan. 2019.

[4] S.H. Kim and Y.S. Kim, "Control area network security technology trends," The Magazine of the IEIE, 42(8), pp. 46-53, 2015.

[5] H.M. Song, Jiyoung Woo, and H.K Kim, "In-vehicle network intrusion detection using deep convolutional neural network," Vehicular Communications 21, vol. 21, pp. 1-13, Jan. 2020.

[6] Eunbi Seo, H.M. Song, and H.K. Kim, "GIDS: GAN based Intrusion detection system for in-vehicle network," 16th Annual Conference on Privacy, Security and Trust(PST), vol. 1, pp. 1-6, Aug. 2018.

[7] Christian Szegedy, Wojeciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, I.J. Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," arXiv preprint, Dec. 2013.

[8] Dowan Kim and Daeseon Choi, "GAN based adversarial CAN packet generation method for evading IDS," Conference on Information Security and Cryptography Summer 2021, pp. 74-77, Jun. 2021.

[9] R.S. Pressman and B.R. Maxim, Software engineering a practitioner's approach, ISBN-10: 1259872971, 3rd Ed. McGraw Hill, Sep. 2019.

[10] M.J. Kang and J.W. Kang, "Intrusion detection system using deep neural network for in-vehicle network security," PloS one, vol. 11, no. 6, Jun. 2016.

[11] Geoffrey Hinton, Li Deng, Dong Yu, G.E. Dahl, et. al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," IEEE Signal Processing Magazine, vol 29, no. 6, pp. 82-97, Nov. 2012.

[12] G.E. Hinton, Simon Osindero, and Y.W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527-1554, Jul. 2006.

[13] I.J. Goodfellow, J.P. Abadie, Mehdi Mirza, et. al., "Generative adversarial nets," Advances in Neural Information Processing Systems 2014, vol. 2, pp. 2672-2680, Jun. 2014.

[14] A.R. Javed, S.U. Rehman, M.U. Khan, and Mamoun Alazab, "CANintelliIDS: detecting in-vehicle intrusion attacks on a controller area network using CNN and attention-Based GRU," 2021 3rd International Cyber Resilience Conference(CRC), vol. 8, no. 2, pp. 1456-1466, Jan. 2021.

[15] Alex Krizhevsky, Ilya Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems 2012, vol. 1, pp. 1097-1105, Dec. 2012.

[16] Zhewen Niu, Zeyuan Yu, Wenhu Tang, Qinghua Wu, and Marek Reformat, "Wind power forecasting using attention-based gated recurrent unit network," Energy, vol. 196, Apr.

2020.
[17] T.K. Ho, "Random decision forest," Proceeding of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278-282, Aug. 1995.
[18] Hyunjae Kang, B.I. Kwak, Y.H. Lee, Hwejae Lee, and H.K. Kim, "Car hacking and defense competition on in-vehicle network," Workshop on Automotive and Autonomous Vehicle Security, pp. 1-6, Feb. 2021.
[19] I.J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint, Dec. 2014.
[20] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," 2017 IEEE Symposium Security and Privacy, vol. 1, pp. 39-57, May. 2017.
[21] Zilong Lin, Yong Shi, and Zhi Xue, "IDSGAN: generative adversarial networks for attack generation against intrusion detection," arXiv preprint, Sep. 2018.
[22] K-CyberSecurityChallenge2020, "K-cyber security challenge," http://datachallenge.kr/challenge20/car/rules/, Aug. 2021.
[23] Martin Arjovsky, Soumith Chintala, and Leon Bottou, "Wasserstein generative adversarial networks," Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 214-223, 2017.

〈저자소개〉

김 도 완 (Dowan Kim) 학생회원
2019년 2월: 공주대학교 의료정보학과 학사
2019년 3월~2020년 8월: 공주대학교 대학원 의료정보학과 석사과정
2020년 8월~현재: 숭실대학교 대학원 융합소프트웨어학과 석사과정
〈관심분야〉 정보보호, 인공지능 보안, 차량 보안

최 대 선 (Daeseon Choi) 종신회원
1995년 2월: 동국대학교 컴퓨터공학과 졸업
1997년 2월: 포항공과대학교 컴퓨터공학과 석사
2009년 1월: 한국과학기술원 전산학과 박사
1997년 1월~1999년 6월: 현대정보기술 선임
1999년 7월~2015년 8월: 한국전자통신연구원 인증기술연구실 실장/책임연구원
2020년 9월~2020년 8월: 공주대학교 의료정보학과 부교수
2020년 9월~현재: 숭실대학교 소프트웨어학부 교수
2016년~현재: 정보보호학회 차세대인증연구회장
〈관심분야〉 인증, 개인정보보호, 차량 보안, 의료정보보안, 머신러닝, 인공지능 보안