컴퓨터통신망 연구보고서

Machine Learning Models Suitable
for Automation Technologies in 6G networks

박서영
전자정보공학부
20213196

1.1 제안서 요약

6G 네트워크는 5G 네트워크와 달리 지상 네트워크뿐만 아니라 공중, 해상 등을 포함하는 3D 이동통신이 실현된 통신의 편재를 추구하고 있다. 10배~100배 이상의 성능 개선을 목표로 하고 있으며 세부적인 관점에 대한 연구는 어느 정도까지 진행된 상태이지만 전체적인 관점에서의 연구는 완성되지 않았다. 상용화는 2028년쯤 이루어질 것이라고 예상된다. 6G 네트워크를 실행하기 위해 필요한 인에이블링 기술으로는 네트워크 컴퓨팅 융합기술, AI/ML 기반 네트워크 자동화기술, 저지연 서비스 제공 기술 등이 개발되고 있다.

네트워크 컴퓨팅 융합 분야에서는 데이터 처리를 엣지에서 수행하여 사용자와의 거리를 최소화하는 5G 네트워크의 에지 컴퓨팅에 인-네트워크 컴퓨팅 기술을 도입하려는 시도가 계속되고 있다. 6G Usage cases들을 참고했을 때 주로 근거리망 적용이 언급되고 있는 것을 확인할 수 있는데, 이를 통해 6G 네트워크가 초분산 구조를 가지게 될 것이라고 예측하였고, 이때 코어 네트워크 내 이동에 있어서 보안 및 인증이 강화되어야 할 것이라고 제안했다. 또한 대역폭이 넓어짐에 따라 신호를 처리하는 반도체의 성능도 높아져야 한다. 아날로그 신호가 외부 요건에 의해 감쇠가 일어날 것을 감안하여 증폭해서 보내고, 잡음을 수신하더라도 왜곡이 최소화된 상태로 복원을 하는 성능이 향상된 수신기가 필요할 것이고, 그러므로 신호 대비 두배 이상의 빠른 동작 속도를 가지는 트랜지스터가 필요할 것이라고 생각하였다.

제안서는 AI/ML 기반 네트워크 자동화 기술에 대한 내용을 주로 서술하였는데, 이는 네트워크에 있어서 자동화가 저지연, 정확도 등 모든 지표에 가장 큰 영향을 미치는 요소라고 판단했기 때문이다. 이뿐만 아니라 6G는 IoT 및 다양한 스마트 기기들과의 연결을 지원하기 때문에, 자동화 기술은 대규모 디바이스를 관리하고 네트워크의 리소스를 할당하는데 도움을 주고, 민감한 데이터를 다루는 6G 네트워크에서의 보안 관련 문제를 빠르게 탐지하고 대응할 수 있게 한다. 5G 네트워크에서는 각 네트워크 기능을 최적화하고 각 슬라이스의 품질 통계를 제공하여 슬라이스 결정에 핵심적인 역할을 하는 NWDAF(Network Data Analytic Function)을 통해 자동화를 이미 구현하고 있지만, 6G 네트워크는 이 기술을 기반으로 하여 소프트웨어 기반 가상자원의 자동 배치, 트래픽 분석, 네트워크 장애 관리 등을 중심으로 한 망 운용/제어/관리의 자동화를 목표로 하고 있다. 6G 네트워크는 아직 개발 중이며, 상용화가 이루어지지 않았기 때문에 6G 데이터를 사용한 훈련 및 연구는 상대적으로 초기 단계에 있다. ML의 도입 분야 중 강조되는 분야는 '이상징후 탐지' 이다. 이는 앞서 언급한 보안 관련 문제나 네트워크 장애 관리와

관련 있는데, 6G 네트워크는 속도가 빠르고 통과하는 트래픽의 밀도도 높기 때문에 다양한 유형의 트래픽이 네트워크를 통과하게 될 것이다. 따라서 시스템 또는 네트워크에서 비정상적인 활동을 감지하고 식별하는 프로세스인 이상징후 탐지는 자동화 기술에서 필수적이다. 하지만 암호화된 트래픽이 증가하고 침입시도가 증가하고 있고, 현재 기술은 이전에 확인되지 않았던 새로운 정상 네트워크 프레임을 공격 트래픽으로 오판정하는 비율이 높기 때문에 네트워크 트래픽에서 추출된 다양한 카테고리의 데이터셋을 이용하여 비정상 트래픽을 탐지할 때 짧은 훈련 시간을 가지고 있으며 동시에 성능도 우수한 알고리즘과 ML 모델을 찾는 연구는 과거부터 꾸준히 진행되어오고 있다. 이번 보고서에서는 각자 다른 시대에 수행된 선행 연구들에서 가장 효율이 좋다고 판단된 모델에 직접 네트워크 데이터셋을 이용하여 훈련 및 학습을 진행하고 이외에 추가로 더 나은 모델을 찾기 위해 진행한 실험의 과정과 결과를 중심으로 내용을 전개할 것이다.

1.2 네트워크 이상징후 탐지

네트워크 이상징후는 컴퓨터 정보 시스템, 컴퓨터 네트워크, 인프라스트럭처, 개인용 컴퓨터 기기를 대상으로 한 공격인 사이버 공격을 방지하기 위해 평균치와 많은 차이가 있는 값을 탐지하여 검출하는 기술이다. 이러한 사이버 공격은 주로 악의적인 목적을 가진 공격자에 의해 발생하는데, 공격 대상 시스템이 정상적으로 동작하는 것을 방해하거나 데이터를 파손, 파괴하기 때문에 이러한 트래픽을 가려내지 못하면 큰 피해가 발생 할 수 있다. 이러한 네트워크 이상징후에는 첫번째로 네트워크에서 예상하지 못한 대량의 데이터 전송이 감지되는 경우가 있다. 예시로 DDoS 공격은 하나의 에이전트에서 하나의 시스템을 공격하는 것이 아니라 동시에 여러 에이전트를 이용하여 하나의 피해 시스템에 패킷을 과도하게 흘려보내는 DoS 공격이다. 이 경우 공격을 당하는 서버의 서비스는 마비된다. 두번째로 특정 사용자가 평소와 다른 위치에서 로그인을 시도하거나 알 수 없는 장치가 연결되는 경우가 있다. 세번째로는 한 사용자의 계정에서 여러 번의 실패한 로그인 시도가 감지된 경우, 민감한 데이터의 대량 복사 또는 이동 등이 감지되는 데이터 패턴 이상징후가 있다. 이외에도 네트워크에서 일어나는 다양한 비정상적인 활동 또는 행동을 모두 포함한다.

2. 실험 개요

2.1 데이터셋 선정

UNSW-NB15 데이터셋은 9가지 공격 유형과 정상 유형에 대한 42개의 피처로 구성된 트래픽 정보를 포함하고 있다.

smean	dmean	trans_dept	response_	ct_srv_src	ct_state_tt	ct_dst_ltm	ct_src_dpc	ct_dst_spc	ct_dst_src_	is_ftp_logi	ct_ftp_cm	ct_flw_http	ct_src_ltm	ct_srv_dst	is_sm_ips_	attack_cat	label
248	0	0	0	2	2	1	1	1	2	0	0	0	1	2	0	Normal	0
881	0	0	0	2	2	1	1	1	2	0	0	0	1	2	0	Normal	0
534	0	0	0	3	2	1	1	1	3	0	0	0	1	3	0	Normal	0

그림 1: UNSW-NB15 데이터셋 일부

출발, 도착지의 IP/포트 넘버나 TTL 값 이외에 주요 feature 은 다음과 같다.

Dur(레코드 총 지속시간)	공격 패턴이나 정상 행동의 지속 시간
Sbytes	출발지에서 목적지로의 트랜잭션 바이트, 트랜잭션이 전송한 데이터의 양
Sload	출발지 초당 비트수, 출발지에서의 데이터 전송 속도
Spkts	출발지에서 목적지로의 패킷 수
Attack_cat	공격 카테고리의 이름 (정상인 경우 normal 로 표시)

표 1: UNSW-NB15 데이터셋의 주요 Feature

또한 UNSW-NB15 데이터셋은 정상 데이터의 경우 0, 공격 데이터의 경우 1으로 나타내는 binary label을 사용한다. 총 2,540,044개의 데이터로 이루어져 있으며 이중 약 87%가 정상 데이터이고 나머지는 공격 데이터이다. Attack_cat에 표시되는 공격 카테고리는 앞서 설명한 DoS를 포함하여 9가지가 있다.

Fuzzers	소프트웨어에 무작위 또는 임의의 입력을 주입하여 프로그램이 예상치 못한 동작을 하도록 하는 테스트 방법. 시스템에서 예상치 못한 동작이나 취약점을 찾아내기 위해 사용됨.
Analysis	시스템의 구조,동작 또는 코드를 분석하여 취약점을 찾는 과정이나 기술
Backdoor	정상적인 인증 또는 암호화를 우회하는 접근 액세스
DoS	정상적인 트래픽을 처리할 수 없을 때까지 대상 시스템을 요청으로 압도하거나 폭주시켜 추가 사용자에 대한 서비스 거부를 초래
Exploits	보안의 취약점을 이용하여 공격하는 방식
Generic	특정 카테고리에 속하지 않는 일반적인 공격 형태
Reconnaisssance	포트를 확인하거나 네트워크 패킷을 캡처하는 등 서비스를 파악하여 대상 네트워크에 대한 정보를 습득하는 단계
Shellcode	사용자 명령어 라인의 해석기 Shell로 제어를 넘기고 공격당한 프로그램의 권한으로 시스템의 다른 프로그램에 접근
worms	스스로를 복제하며, 호스트 컴퓨터를 공격하고 네트워크를 통해 유포되는 악성 코드가 포함된 프로그램

표 2: UNSW-NB15 데이터셋의 attack_cat feature 종류

ackdat	smean	dmean	trans_dept	response_	ct_srv_src	ct_state_tt	ct_dst_ltm	ct_src_dpc	ct_dst_spc	ct_dst_src_	is_ftp_logi	ct_ftp_cm	ct_flw_http	ct_src_ltm	ct_srv_dst	is_sm_ips_	attack_cat	label
0.224843	0.089581	0.135262	103	83	0	0	4	1	1	1	1	2	0	0	0	1	2	0	Fuzzers	1
0.209077	0.112794	0.096283	84	469	1	2306	1	1	2	1	1	1	0	0	1	1	1	0	Exploits	1
0.141729	0.066968	0.074761	96	273	1	895	2	1	1	1	1	1	0	0	1	1	1	0	DoS	1
0.301089	0.219112	0.081977	98	91	0	0	3	1	1	1	1	2	0	0	0	1	2	0	Fuzzers	1

그림 2: attack_cat feature

UNSW-NB15 데이터는 다른 네트워크 데이터들에 비해 비교적 최근 제작되었기 때문에

새로운 유형의 공격 데이터들을 많이 가지고 있을 것이라고 판단하여 네트워크 자동화를 위한 ML 모델을 훈련시키기 위한 데이터셋으로 사용하기로 하였다.

2.2 6G 환경을 고려한 전처리

UNSW-NB15 데이터셋은 학습데이터의 class간 불균형이 존재하여서 down sampling을 진행하여야 한다. 다운 샘플링은 주로 다수 클래스의 샘플을 제거하거나 소수 클래스의 샘플을 복제함으로써 데이터셋의 균형을 맞추는 과정이다. 또한 5G에 비해 6G에서는 주파수 대역이 더욱 높아지고, 더 많은 주파수 범위를 사용할 것으로 예상된다고 앞서 설명했다. 이는 세분화된 셀룰러 네트워크를 유발할 수 있는데 이로 인해 공격 유형 데이터를 수집하는 것이 더욱 어려워질 것이라고 예상된다. 각 특화망은 독립적으로 운영되며, 데이터 공유 및 통합이 복잡해질 것이기 때문이다. 따라서 다운 샘플링 과정에서 공격 데이터 개수를 줄이는 과정을 필수적으로 진행하여야 한다. 사용 기기의 성능 제약으로 약 2백만개의 데이터를 모두 훈련에 사용할 수는 없기 때문에 82333개의 데이터만을 훈련 데이터로 사용하고 175341개의 데이터를 학습 데이터로 사용하였다. 각 공격 유형 별로 데이터의 개수를 출력하였을 때의 결과는 다음과 같았다.

```
30 worms_count = df[df['attack_cat'].str.lower() == 'worms'].shape[0]
31 print(f"The number of data points with 'attack_cat' being 'worms': {worms_count}")
```

```
The number of data points with 'attack_cat' being 'normal': 37000
The number of data points with 'attack_cat' being 'fuzzers': 6062
The number of data points with 'attack_cat' being 'analysis': 677
The number of data points with 'attack_cat' being 'backdoor': 583
The number of data points with 'attack_cat' being 'dos': 4089
The number of data points with 'attack_cat' being 'exploits': 11132
The number of data points with 'attack_cat' being 'generic': 18871
The number of data points with 'attack_cat' being 'Reconnaissance': 3496
The number of data points with 'attack_cat' being 'Shellcode': 378
The number of data points with 'attack_cat' being 'worms': 44
```

그림 3: 각 유형 별 데이터 개수

Worms의 경우 그 수가 44개 밖에 되지 않았으며 generic의 경우 개수가 20000개에 가까운 것을 확인할 수 있었다. 이러한 불균형한 클래스 분포에 의해 이 경우 모델이 해당 클래스를 올바르게 학습하기 어려울 수 있거나 과적합 될 가능성이 있으며 worms 값을 가진 데이터에 대한 정확한 평가가 어려워질 수 있다. 이런 경우 다수 클래스의 데이터를 감소시키거나, 소수 클래스에 높은 가중치를 부여하거나, 혹은 다수 클래스와 소수 클래스의 특성을 결합하여 새로운 데이터를 생성할 수 있는데, 이 중 가장 성능이 높게 나오는 방식으로 실험을 진행하기로 했다. 또한 feature 중에 attack_cat이나 프로토콜 유형은 문자열로 되어있어 훈련을 하기 전 label encoder을 이용하여 이를 숫자 형식으로 변환해주었다. 아래 사진에서 proto 와 attack_cat이 숫자로 변환된 것을 확인 할

수 있다.

	id	dur	proto
0	1	0.000011	117
1	2	0.000008	117
2	3	0.000005	117
3	4	0.000006	117
4	5	0.000010	117

	attack_cat	label
0	6	0
1	6	0
2	6	0
3	6	0
4	6	0

[5 rows x 45 columns]

```
for col in categorical_columns:
    df[col] = label_encoder.fit_transform(df[col])
```

그림 4: label encoder 코드 및 실행 결과

2.3 ML 모델 선택과 성능 비교

6G 네트워크의 이상징후 탐지를 위해 머신러닝 모델을 선택할 때는 다음과 같은 다양한 특성을 고려해야한다. 먼저 6G 네트워크는 초고속 데이터 전송을 제공해야한다. 이로 인해 데이터의 양과 속도가 빠르게 증가하므로 머신러닝 모델은 대량의 데이터를 처리하고 실시간으로 분석할 수 있는 능력이 필요하다. 또한 6G는 초저지연 연결을 제공하며, 이로 인해 머신러닝 모델은 빠른 응답 시간과 실시간 이상징후 감지 능력을 갖추어야한다. 더불어 대규모의 사물 인터넷(IoT) 장치를 지원할 것으로 예상되므로 다양한 IoT 센서 데이터를 처리하고 IoT 장치에서의 이상징후를 식별할 수 있어야 한다. 이번 실험에서는 오토인코더, NB, Random Forest, LGBM, Extra Tree 모델을 사용하여 성능과 시간을 확인하였다.

2.3.1 오토인코더

6G 네트워크 데이터는 대부분 다양한 특성들을 갖고 있다. 고차원 데이터의 경우 모델의 복잡성이 증가할 수 있어, 입력층보다 적은 수의 뉴런을 가진 은닉층을 중간에 넣어 차원을 줄이는 오토인코더를 사용한다면 시간에 따라 변화하는 네트워크 흐름에 빠르게 대응할 수 있을 것이라고 예측했다.

```
Epoch 50/50
2059/2059 [==============================] - 7s 3ms/step - loss: 0.6720 - val_loss: 0.6540
515/515 [==============================] - 2s 3ms/step
테스트 세트의 평균 제곱 오차: 0.6539832949638367
```

그림 5: 오토인코더를 사용한 결과

여기서 epoch는 전체 훈련 데이터셋이 모델을 한번 통과했을 때 1씩 증가하고, 모델의 성능 향상을 위해 50번 훈련을 반복했다. 이때 오차가 0에 가까울수록 성능이 좋다고 판단하는데 오토인코더 모델로 네트워크 데이터셋을 훈련한 결과 시간이 너무 오래 걸리고 오차가 지나치게 크다고 판단하게 되었다.

2.3.2 NB 모델

6G 네트워크 데이터는 이진 분류가 필요한 모델로써, 선형모델과 유사하지만 선형 분류기보다 훈련속도가 빠른 NB 모델이 훈련에 유리할 수도 있다고 생각하여 NB 모델을 훈련 모델 후보로 뽑았다.

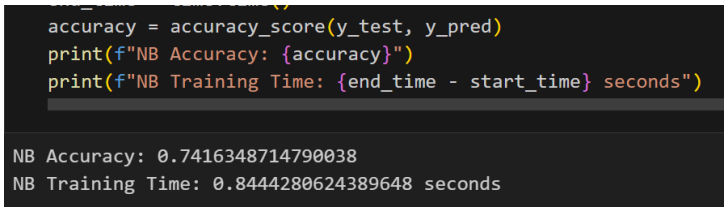

그림 6: NB 모델을 사용한 결과

NB 모델 훈련 및 학습 결과 이론대로 훈련 시간은 매우 빨랐지만 정확도가 기대하는 정도에 미치지 못하는 것을 확인할 수 있었다.

2.3.3 Random Forest

여러 모델을 같이 이용하여 학습을 진행하는 모델을 앙상블 모델이라고 한다. Random Forest는 앙상블 학습의 한 종류로, 여러 개의 결정 트리(Decision Tree)를 사용하여 데이터를 학습하고 예측하는 모델이다. 6G 데이터셋은 feature이 많다는 특징이 있는데, Random Forest는 이러한 다양한 특성의 중요도를 측정할 수 있으며 여러 개의 결정 트리를 결합하므로 단일 결정 트리보다 높은 성능을 제공할 수 있다는 점에서 적합하다고 추측하였다. 특히, 데이터의 복잡한 패턴과 관계를 잘 학습할 수 있기 때문에 학습 성능을 측정해보기로 하였다. 모든 feature에 대해 훈련을 진행한 경우 96퍼센트의 낮지 않은 정확도와 만족스러운 속도를 보여주었고, 중요한 feature만 골라서 훈련을 진행한 경우 시간은 감소하지만 정확도가 현저히 낮아지는 것을 확인할 수 있었다.

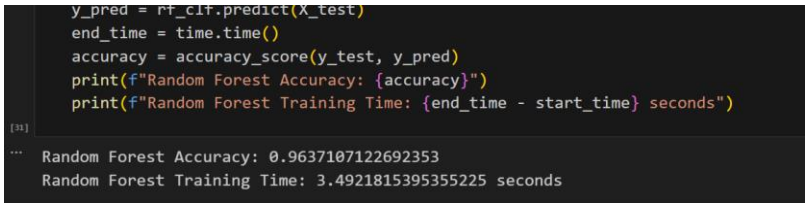

그림 7 : 모든 feature으로 random forest를 실행한 결과

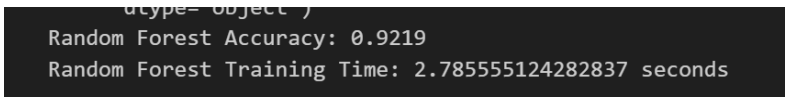

그림 8: Random Forest를 주요 feature로만 훈련한 결과

2.3.4 LGBM

LGBM은 대용량 데이터셋에서 뛰어난 성능을 보인다. 메모리를 효율적으로 활용하며 앞

서 문자열 클래스를 숫자형으로 바꿔준 전처리를 하지 않아도 카테고리컬 특성을 직접 처리할 수 있는 능력을 가지고 있다. 그런데 본 실험에서는 약 69퍼센트 정도의 낮은 정확성을 보여주었다.

```
[LightGBM] [Warning] Stopped training because there are no more leaves that meet the split requirements
[LightGBM] [Warning] boosting is set=gbdt, boosting_type=gbdt will be ignored. Current value: boosting=gbdt
Accuracy: 0.6806
LGBM Training Time: 4.357361316680908 seconds
```

그림 9: LGBM 훈련 결과

2.3.5 Extra Tree

Extra tree는 랜덤 포레스트와 비슷한 앙상블 학습 방법 중 하나로, 랜덤 포레스트와 달리 각 트리의 분할을 결정할 때 가장 좋은 임계값을 찾지 않고 랜덤하게 결정하는 특징이 있다. 앞선 random forest 모델 훈련 결과가 긍정적이었기 때문에 Extra tree 또한 성능이 좋을 것이라고 예측하였다. Extra tree의 파라미터를 기본값으로 설정하였을 때 정확도는 약 97퍼센트였고, 시간은 3.12초 정도였다. 그리고 최적의 parameter을 찾는 과정인 grid search 를 수행하고 도출된 결과로 파라미터를 설정하였는데 정확도는 99퍼센트로 우수한 수준이었고, 훈련 시간 또한 1.45초로 매우 짧았다.

```
accuracy = accuracy_score(y_test, y_pred)
print(f'Extra Trees Accuracy: {accuracy:.4f}')
print(f'Training Time: {training_time:.2f} seconds')

Extra Trees Accuracy: 0.9740
Training Time: 3.12 seconds
```

그림 10: Default Parameter으로 학습한 결과

```
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f'Extra Trees Accuracy: {accuracy:.4f}')
print(f'Training Time: {training_time:.2f} seconds')

Extra Trees Accuracy: 0.9985
Training Time: 1.45 seconds
```

그림 11: Best Parameter으로 학습한 결과

Extra Tree는 일반적으로 메모리 사용량이 적다고 알려져 있고 훈련시간도 빠른 편이며 많은 다양한 데이터셋에 대해 가장 높은 성능을 자랑한다.

3. 결론 및 한계점

본 보고서는 아직 개발중인 6G 네트워크 환경에 적합한 조건을 갖춘 머신러닝 모델을 찾는 것을 목표로 작성되었고, USNW 검증 결과 ExtraTree 모델이 가장 최적의 성능과 조건을 가지고 있다고 제안하였다.

6G 네트워크에 있어 AI를 이용한 자동화 기술은 성능지표를 실현하기 위해서 필수적이다. 아직 상용화 단계에 이르지 못하였고 개발이 한창 진행중이기 때문에 현재 상황에서 절대적으로 적합한 모델을 찾는 것은 어렵다. 또한 6G 네트워크에서 적합한 머신러닝

모델을 찾기 위한 선행 연구들 사이의 의견이 일치되지 않고, 본 보고서 또한 선행 연구들에서 긍정적으로 평가된 머신러닝 모델의 성능이 가장 낮게 나오는 등 큰 차이가 있는 경우도 존재했다. 연구 환경 및 상황, 그리고 아직 존재하지 않는 6G 네트워크 데이터셋을 구현하는 방법에 따라서 결과가 다르게 도출되기 때문에 연구 과정에 한계가 있다.

4. 참고 문헌

박승균.(2021).무선 네트워크 침입탐지를 위해 개선된 CNN 분석.융복합지식학회논문지,9(3),147-154.

서승수. "네트워크 트래픽 이상징후 탐지율 향상을 위한 자기지도학습 기반의 오토인코더 최적화 연구." 국내석사학위논문 서울대학교, 2020. 서울

봉기정. "6G 환경의 AI 기반 보안 모델 및 데이터 생성기법 검증." 국내석사학위논문 USTETRI School, 2023. 대전

예충일, 김근영, 김영진, 김용선, 김진경, 남상우, 이우용, 장성철, 고영조.(2020).6G 비전, 서비스 및 기술 동향.한국통신학회지(정보와통신),37(2),11-22.

전효진. (2023). 6G 이동통신 연구개발 동향분석 및 발전방안 연구 -미래 국방사업 적용 방안-. 한국산학기술학회 논문지, 24(7), 336-342, 10.5762/KAIS.2023.24.7.336

이태양,and 이종혁. "Beyond 5G 및 6G 보안 국제표준화 동향 분석." 韓國通信學會論文誌 48.4 (2023): 423-434

나중찬. "실시간 트래픽 분석에 의한 예측모형 기반 네트워크 이상징후 탐지 기법." 국내박사학위논문 忠南大學校, 2004. 대전