개요
제안하는 모델은 딥러닝 ANN과 스택 앙상블 기법을 이용하여 현재 네트워크에 존재하는 침입 유형을 탐지한다. 스택형 앙상블 사용 이유는 스택형 앙상블을 사용하여 여러 딥 러닝 모델을 함께 결합하여 보다 효율적이고 강력한 침입 탐지 메커니즘을 얻기 위함이다.

1. 서론
교환된 정보의 기밀성과 무결성을 보호하기 위해 사용된 네트워크는 보안이 필요하므로 네 트워크 보안이 실행되었다. 네트워크 침입 탐지 시스템(NIDS, Network Intrusion Detection System)는 네트워크에 침입을 감지하고 관련 당사자에게 공격 사실을 알린다.

머신러닝의 문제점
NIDS에 사용할 수 있는 전통적인 머신러닝 알고리즘은 매우 우수한 정확도를 보여주지만 비용이 많이 들고 과도한 양의 데이터가 필요하다.

이 논문은 NIDS 모델을 구현하기 위한 딥러닝 기법에 초점을 둔다.
- 미가공 데이터를 사용하는 특성 공학(feature engineering) 기법을 적용해 데이터를 전처리
- 인공신경망(ANN, Artificial Neural Network)을 사용하여 다양한 모델 세트를 교육하고 추가 처리를 위한 최적의 모델을 선택한다. 최고의 모델은 정확도, 내부 구조 분포 및 학습 시간을 기준으로 선택
- ANN 모델을 사용한 스택 앙상블 기법을 통해 여러 모델을 함께 결합하면 효율적이고 일관성 있는 모델이 된다.

2. 관련 연구
관련 논문
• 데이터 셋의 문제점
KDD99 및 NSL-KDD와 같은 이러한 데이터 셋 중 일부는 매우 오래되고 작다. 모델을 설계 하고 이 데이터셋을 탐지에 사용 하는 것은 현대 기술에 적합하지 않다. 이러한 데이터셋 기능과 침입 공격은 도전적인 현대 보안 위협에 충분하지 않으며 매 순간 새로운 보안 문제가 발생한다.

본 논문
• 데이터셋
NF-UQ-NIDS-V2

• 사용 모델
앙상블 학습 머신러닝과 딥러닝 을 함께 사용하는 것은 비용이 많이 들며 다른 기술은 결과를 얻기 위한 접근 방식이 다르다. 우리 모델은 ANN(인 공 신경망) 사용에 중점을 둔다. ANN의 다양한 모델 세트는 최종 분류 결과를 얻기 위해 설계, 훈련 및 결합한다.

3. 방법론

모델의 기본 구조

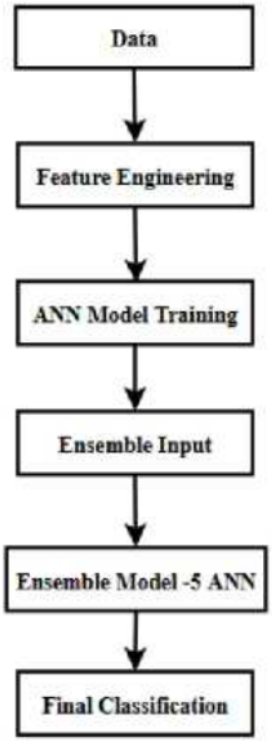

Fig. 1: *The proposed model*

모델의 다양한 변형이 훈련 데이터로 훈련되고 훈련 후 모델을 결합하여 앙상블 Model-5 모델에 대한 완전한 입력이 생성된다.

A. 피쳐 엔지니어링

- 대치: null과 무한대값을 0으로 대치
- Feature
- 이상치 처리: Z-score 기반 기술 사용
  - u = mean(vi) + 3 * std(vi) : 상한선
  - l = mean(vi) - 3 * std(vi) : 하한선
- 스케일링

$$z = \frac{X - \mu}{\sigma}$$

σ: 표준편차. z: 표준화 μ: 평균

- Feature Encoder(Label Encoder)

B. 딥러닝 모델

모든 ANN 모델 중에서 가장 좋은 4개의 ANN 모델 선택
모든 모델은 하나의 입력층, 하나의 출력층 및 4개의 은닉층으로 학습된다. 하이퍼파라미터 튜닝도 수행한다.

- ANN
  - 모델 구조에는 Feed-Forward ANN(전방향 인공신경망)이 사용된다.
  - 은닉층에는 Relu을 사용하고 출력층에는 다중 클래스 분류를 위한 softmax 활성화 함수를 사용

- 과적합 처리
  - 드롭아웃 레이어의 범위는 0.1 ~ 0.5입니다.
  - 데이터가 부족하지 않으므로 과소적합 가능성은 고려하지 않는다.
  - 각 은닉층에는 2^5, 2^6,2^7,2^8 노드가 있을 수 있지만 첫 번째 은닉층에는 모든 모델에 대한 최소 2^6 노드가 있어야 한다.

모든 ANN모델은 입력층과 출력층에 대해 도일한 수의 노드를 갖는다.
표-1는 기본 모델의 각 모델의 은닉층에 대한 노드 수

TABLE I: ANN structure of four Base Model

	Hidden Layer Node distribution			
	layer-1	layer-2	layer-3	Layer-4
Model-1	64	128	128	64
Model-2	64	128	256	128
Model-3	256	64	64	256
Model-4	128	64	32	128

밀도가 높은 은닉층 사이에 드롭아웃을 사용

C. 스택 앙상블

- Model-5 ANN을 모델링할 때 하나의 은닉층만 사용하는 것이 효율적이다.
- 입력층의 경우 기본 ANN 모델의 출력에서 입력이 생성된다.
- 은닉층에서 변환을 거친 후 최종 분류 출력이 분류된다.
- 출력층에는 softmax 활성화 기능이 사용된다.

4. 실험

데이터 세트: 80:20
사용 언어 및 라이브러리: 파이썬, 텐서 보드, 파이썬 시각화 라이브러리
하드웨어: 100GB 메모리 공간, 13GB RAM, 2.2Ghz 2CPU
데이터 세트 분포: 양성 흐름 33.12%, 공격 66.88%
평가 매트릭스: 정확도, 정밀도, Recall, F1-Score

5. 결과

4개의 기본 모델 및 Model-5 ANN 결과

Model	Trainable parameter	Accuracy	Batch size	Patience
Model-1	37,077	98.19%	1024	8
Model-2	79,573	98.30%	512	8
Model-3	53141	98.30%	1024	8
Model-4	22,517	98.25%	512	16
Model-5	6,805	98.40%	1	Epoch-32

Early stopping을 사용하여 과적합 조절
기본 모델의 배치 크기는 2^10 및 2^9 이다.

첫 번째 은닉층과 출력층에 대한 편향, 가중치, bin이 있는 4가지 ANN 기반 모델(특정 시간 단계 =10이라고 가정)

	First Hidden layer			Output layer		
	bias	weight	bin	bias	weight	bin
Model-1	0.00166	-0.5 to 0.5	9.48	0.334	-8 to 4	1.53
Model-2	-0.0774	-1 to 1	6.06	-0.415	-4 to 1.5	1.58
Model-3	0.0991	-1.5 to 1.5	1.96	-0.553	-6 to 2	1.40
Model-4	-0.142	-4 to 4	67.0	0.0339	-3 to 1	2.89

TABLE VI: 전체 데이터 세트에서 각 클래스에 대해 테스트된 샘플 수를 포함하는 클래스별 네트워크 침입 정밀도, 재 현율 및 F1 점수

Attack name	Precision	Recall	F1-Score	Support
Analysis	0.21	0.72	0.33	36
Backdoor	1.00	0.88	0.94	234
Benign	1.00	1.00	1.00	335354
Bot	1.00	1.00	1.00	1970
Brute Force	1.00	0.97	0.98	1630
DDoS	0.99	0.99	0.99	290346
DoS	0.98	0.98	0.98	238243
Exploits	0.80	0.77	0.78	435
Fuzzers	0.48	0.85	0.61	274
Generic	0.99	0.65	0.78	203
Infilteration	0.92	0.75	0.83	1534
Reconnaissance	0.92	0.87	0.89	34825
Shellcode	0.01	0.00	0.00	23
Theft	0.38	0.26	0.31	23
Worms	0.00	0.00	0.00	3
injection	0.88	0.77	0.82	9193
mitm	0.79	0.32	0.46	102
password	0.92	0.92	0.92	15315
ransomware	0.88	0.86	0.87	44
scanning	0.96	1.00	0.98	50355
xss	0.93	0.97	0.95	33031

이 표는 모델의 전반적인 효율성을 나타내며 여기서 Support은 테스트를 위해 수행된 각 공격 유형에 대해 임의로 선 택된 샘플 수를 나타내며 Worm은 샘플 수가 3개에 불과하기 때문에 모든 평가 매트릭스(matrices)에서 가장 낮다. 반면에 bot은 정밀도와 재현율이 1이며, 이는 모델이 bot을 100% 정확하게 식별하고 있음을 의미한다.

6. 향후 작업

스택 앙상블은 여러 모델을 함께 연결할 수 있으므로 모델 효율성을 높이고 다양한 유형의 모델을 사용하면 더 강력한 탐지 모델을 얻을 수 있습니다.