KNOM Review '23-02 Vol.26 No.02
https://doi.org/10.22670/knom.2023.26.2.33

딥러닝 기반의 응용 프로그램 트래픽 분류를 위한 데이터셋 사용 및 전처리 방법

장 윤 성◆, 박 지 태*, 백 의 준*, 김 주 성**, 이 대 국*, 김 명 섭°

Dataset Usage and Pre-processing Method of Deep-learning Based Application Traffic Classification

Yoon-Seong Jang◆, Jee-Tae Park*, Ui-Jun Baek*, Ju-Sung Kim**, Dae-Guk Lee*, Myung-Sup Kim°

요 약

네트워크 트래픽 분류는 네트워크 관리 분야의 핵심 기술로 이를 위해 다양한 공개 데이터셋이 활용되고 있다. 연구 목적에 맞는 데이터셋을 선택하는 것은 중요하며, 데이터셋의 메타데이터 비교 분석 결과는 데이터셋 선택 과정에 도움을 줄 수 있다. 기존 연구에서는 설명, 데이터셋 레벨 그리고 정답지 포함 여부에 대한 비교를 제공하고 있으나 연구 목적에 맞는 데이터셋을 선택하기에 충분하지 않다. 본 논문에서는 데이터셋 선택을 위한 고려사항을 제시하고 데이터셋의 크기, 수집 환경 등 보다 상세한 정보를 제공하고 이를 비교한다. 또한, 다양한 연구에서 사용한 데이터셋 전처리 방법, 분류 방법 그리고 재현성을 데이터셋 별로 분류하여 후속 연구자들의 연구 객관성 확보를 돕고자 한다. 마지막으로, 현재 공개된 데이터셋들의 한계점 및 데이터셋 구축에 대한 도전 과제를 제시한다.

Key Words : Deep Learning, Dataset, Network traffic, Classification, Pre-processing

ABSTRACT

Network traffic classification is a core technology in the field of network management, and various public datasets are being utilized for this purpose. Selecting the appropriate dataset that aligns with the research objectives is crucial, and the comparative analysis of dataset metadata can assist in the dataset selection process. Existing research provides comparisons based on descriptions, dataset levels, and the presence of ground truth labels, but this may not be sufficient for selecting datasets that are tailored to specific research goals. In this paper, we present considerations for dataset selection and provide more detailed information on aspects such as dataset size and collection environment, facilitating their comparison. Additionally, we categorize dataset preprocessing methods, classification techniques, and reproducibility employed in various studies by dataset, aiming to enhance the objectivity of future research endeavors. Finally, we outline the limitations of currently available datasets and pose challenges related to dataset construction.

※ 본 연구는 국토교통부 AI기반 스마트하우징 기술개발사업의 연구비지원(20SHTD-B157018-01)과 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(00235509, ICT융합 공공 서비스●인프라의 암호화 사이버위협에 대한 네트워크 행위기반 보안관제 기술 개발)
◆ First Author : Korea University Department of Computer Convergence Software, brave1094@korea.ac.kr
* Korea University of Department of Computer and Information Science, {pjj5846, pb1069, daekuglee}@korea.ac.kr
** Korea University Department of Computer Convergence Software, jsung0514@korea.ac.kr
° Corresponding Author : Korea University Department of Computer and Information Science, tmskim@korea.ac.kr
논문번호 : KNOM2023-02-05, Received November 17, 2023; Revised December 3, 2023; Accepted December 12, 2023

Ⅰ. 서 론

트래픽 분류는 침입탐지와 같은 네트워크 보안, 성능 최적화, 서비스 품질 관리 등 다양한 네트워크 관리 및 보안 목적을 위해 중요한 역할을 하는 분야이다. 이 분야는 네트워크에서 전송되는 데이터 패킷을 그 특성과 행동에 따라 다른 카테고리 또는 클래스로 분류하는 프로세스를 포함하며, 이러한 분류는 데이터의 특성을 이해하고 관리하기 위해 필수적이다. 컴퓨터 네트워크에서 발생하는 트래픽은 다양한 출발지과 목적지 간에 전송되며, 웹 브라우징, 비디오 스트리밍, 파일 전송, 음성 통화 등 다양한 응용 프로그램 및 서비스에서 발생한다. 트래픽 분류는 이러한 다양한 트래픽 유형을 식별하고 분석함으로써 네트워크 관리자에게 중요한 정보를 제공하고, 보안 위협을 탐지하고 대응하는 데 도움을 준다.

딥러닝은 최근 몇 년 동안 컴퓨터 비전, 자연어 처리, 음성 인식 등 다양한 분야에서 혁신적인 성과를 거두며 주목받고 있는 인공지능 기술 중 하나이다. 이러한 딥러닝 기술은 트래픽 분류 분야에서도 중요한 역할을 하며, 네트워크 관리와 보안에 관련된 여러 측면에서 큰 잠재력을 가지고 있다. 딥러닝을 활용한 트래픽 분류는 기존의 통계적 방법과 비교하여 더 높은 정확도와 효율성을 제공할 수 있다. 딥러닝 모델은 데이터로부터 복잡한 패턴을 스스로 학습하며, 이를 기반으로 트래픽을 분류하는 데 사용된다. 이러한 모델은 대규모 데이터셋과 충분한 학습을 통해 네트워크 트래픽의 다양한 특징과 동작을 이해하고 식별할 수 있다. 또한, 딥러닝 기반 트래픽 분류는 다양한 네트워크 관리 및 보안 시나리오에서 적용된다. 예를 들어, 악성 코드나 DDoS 공격과 같은 보안 위협을 탐지하고 차단하는 데 사용될 수 있으며, 서비스 품질(QoS) 관리나 네트워크 최적화에도 활용된다.

데이터셋은 딥러닝 기반 트래픽 분류 연구에 있어서 핵심적인 구성 요소 중 하나로, 모델 학습 및 평가를 위한 핵심 자원이다. 트래픽 분류를 위한 데이터셋은 실제 네트워크 트래픽을 포함하고 있으며, 다양한 트래픽 클래스 또는 카테고리로 구분되어 있다. 데이터셋은 트래픽 클래스의 다양성과 양적 측면에서 다를 수 있으며, 연구자들은 주로 공개된 데이터셋을 활용하거나 자체적으로 수집하고 구성하기도 한다. 데이터셋의 특성은 연구 결과에 큰 영향을 미치며, 딥러닝 모델을 효과적으로 학습하고 평가하기 위해 데이터셋을 신중하게 선택하고 전처리하는 것이 중요하다. 기존 연구에서 제공하는 데이터셋의 특성에는 데이터셋의 크기나 수집환경과 같은 세부 정보가 포함되지 않았다. 데이터셋의 세부 정보에 따라 적절한 전처리 과정을 선정하여 적용한 뒤 연구에 사용할 데이터셋을 확보해야 하므로, 이러한 데이터셋의 세부 정보의 비교가 요구된다. 또한, 데이터셋의 단점을 보완하는 전처리 기법에 대한 조사는 학습 과정에서의 다양한 문제의 해결책을 제시하므로, 이에 대한 연구가 필요하다.

본 논문은 2013년부터 현재까지 발표된 논문 중 Deep Learning, Network Traffic Classification, Dataset, Pre-Processing을 키워드로 검색된 논문을 조사 및 분석한 것을 토대로 데이터셋 선택을 위한 고려사항을 제시하고, 이에 따라 공개 데이터셋 UNSW-NB15, KDDCup99, NSL-KDD, CIC-IDS-2017, ISCX VPN2016(VPN-nonVPN), CMU-SynTraffic-2022에 대하여 데이터셋의 크기, 수집 환경 등 보다 상세한 정보를 제공하고 이를 비교한다. 또한, 다양한 연구에서 사용한 데이터셋 전처리 방법, 분류 방법 그리고 재현성을 데이터셋 별로 분류하여 후속 연구자들이 고안한 모델을 기존의 연구와 비교할 때 동일한 전처리 과정을 적용시킨 동일한 조건에서의 비교를 지원하고자 한다. 마지막으로, 현재 공개된 데이터셋들의 한계점과, 이를 보완하기 위한 연구의 필요성을 제시하고, 후속 연구자들의 새로운 공개 데이터셋 구축에 대한 도전 과제를 제시한다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 관련 연구에는 최신 연구 논문에서 발표된 공개 데이터셋의 비교에 대해 서술하고, 이를 보완할 방향에 대하여 논의한다. 본론에서는 데이터셋 선택 시 고려사항에 대해 제시하고, 이를 바탕으로 공개 데이터셋의 세부 정보를 비교하였다. 또한, 데이터셋 별로 사용된 트래픽 분류 모델을 조사하여 비교하였다. 마지막으로 데이터셋 별로 사용된 전처리 기법에 대한 비교를 제공하였고, 향후 연구자들의 모델 비교를 위해 논문 별로 전처리 과정과 딥러닝 모델의 소스코드 공개 여부를 비교하여 제공하였다. 논의 사항에서는 향후 연구를 위한 연구 방향성을 제시하고, 결론에서는 논문의 마무리를 맺는다.

Ⅱ. 관련 연구

Zhao 등은 연구에서 기존 트래픽 분류 방법의 성능을 평가하기 위한 기준 목록을 제안하고, 트래픽 분류 모델을 학습시키기 위해 주로 사용되는 데이터셋 및 트래픽 특성에 대한 내용을 서술하였다 [1]. 저자는 해당 논문에서 데이터셋의 이름과, 간단한 설명, 트래픽 분류에 사용되는 단위와 Ground truth의 여부에 대한 비교를 진행하였다. 그러나, 해당 공개 데이터셋 비교에는 데이터셋이 구축된 목적에 대한 표현이 모호하게 서술되었고, 연구의 성과에 영향을 미치는 다양한 세부 정보들에 대한 서술이 없다. 먼저, 해당 논문에서는 데이터셋의 구축 목적이 명확하게 표현되어야 한다. [1]에서 CAIDA 데이터셋의 'for traffic analysis attacks' 설명과, ISCX 데이터셋의 'for anomaly detection' 이와 같은 부분에서는 IDS(Intrusion Detection System)이라는 목적이 공통적으로 존재함에도 불구하고, 다른 방식으로 표현되었다. 이는 어떤 수집 목적을 가지는 데이터셋인지 한 눈에 파악하기 어렵고, 혼동되기 때문에, 명확하게 데이터셋이 구축된 목적을 분류할 필요가 있다. 두 번째로는, 데이터셋의 *크기*에 대한 정보가 없다. 데이터셋의 *크기*는 딥러닝 모델에 다양한 패턴을 학습시킴으로서, 모델의 성능의 향상에 큰 영향을 주는 요소 중 하나이다. 충분한 양의 데이터를 학습시키지 않으면, 딥러닝 모델의 편향학습이나, 과적합의 문제를 불러올 수 있다. 따라서 데이터셋의 *크기*에 대한 정보를 추가함으로서, 연구에 사용할 충분한 *크기*의 데이터인지를 직관적으로 확인할 수 있어야 한다. 세 번째로, 각 데이터셋의 수집된 환경에 대한 정보가 포함되어 있지 않다. 데이터셋이 수집된 환경이 중요한 이유는, 실제 환경에서 수집된 데이터셋의 경우, 실제 환경처럼 다양한 트래픽 패턴을 학습시키고, 네트워크 환경에서 발생하는 노이즈와 현실적인 조건 하에 수집되었기 때문이다. 본 논문에서는 이와 같은 기존 연구의 데이터셋 비교내용의 보완점을 확인하고, 이를 개선하여 본론 3.2에서 제시한다.

Ⅲ. 본론

3.1 데이터셋 선택 시 고려사항

네트워크 트래픽 분류를 위해 많은 대학과 연구기관에서는 공개 데이터셋을 발표하였다. 하지만, 공개 데이터셋들은 레이블의 존재여부나 데이터셋이 수집된 환경과 수집년도, 용량 등 세부 정보에서 많은 차이를 보인다. 본 논문에서는 데이터셋을 선정할 때의 고려사항 총 7가지를 제시함으로써, 향후 연구자들이 연구에 적합한 데이터셋을 사용할 수 있도록 하고자 한다. 본 논문에서 제시하는 데이터셋 선택 시 고려사항은 다음과 같다 :

1) 목적과 사용 사례 : 현재 공개된 데이터셋들은 주로 IDS, QoS의 보장을 그 수집 목적으로 구축되었다. 데이터셋을 사용하는 목적을 명확하게 이해하고, 해당 목적에 적합한 데이터를 선택하는 것은 중요하다. IDS를 위한 데이터셋은 대부분의 경우 클래스의 구성을 Normal, Attack으로 나누어 구성된다. 반면, QoS의 보장을 위한 데이터셋은 대부분의 경우 클래스의 구성을 VoIP, Email, Chat, Streaming과 같은 서비스 종류로 나누어 구성된다. 목적에 적합한 데이터셋을 선택하지 않을 시, 연구 목적과 다른 결과를 취하게 될 수 있으므로, 목적과 사용 사례에 대한 파악이 필요하다.
2) 데이터의 수집년도 : 네트워크 악성 트래픽 공격과 침입 기법들은 시간이 지날수록 그 복잡성과 다양성이 고도화되고 있다. 이러한 관점에서 오래된 데이터셋은 최신의 공격 기법에 대한 클래스를 포함하지 않기 때문에, 비교적 최신의 데이터셋과 함께 사용하거나 최신 데이터를 사용하는 것이나, 최신 데이터와 함께 혼합하여 사용하는 것이 권장된다.
3) 레이블 여부 : 레이블의 유무는 학습된 모델의 성능을 직접적으로 평가할 수 있는 정답지이므로, 포함된 데이터셋을 사용하는 것이 좋다. 네트워크 트래픽 분류에서는 각 트래픽이 정상적인 트래픽인지 공격인지, 혹은 어떠한 응용 트래픽 데이터인지에 대한 정보가 제공되어야 하기 때문에, 레이블의 유무는 필수적이다. 또한, 레이블링 과정에 대한 공개와 오류 검출 가능성에

대한 검토 또한 필요하다. 잘못된 레이블링은 모델이 학습하는 데에 혼동을 발생시켜, 모델의 성능을 저하시키기 때문이다.

4) 데이터셋 크기 : 데이터셋의 크기는 모델의 성능에 영향을 미친다. 이는 충분한 양의 데이터가 존재해야 모델을 훈련시키고 일반화 성능이 우수하기 때문인데, 그렇다고 무조건적으로 다량의 데이터를 사용할 수는 없다. 양질의 데이터를 다량으로 수집하는 것은 많은 비용과 많은 시간이 들기 때문에, 적절한 균형점을 찾아내는 것이 중요하다.
5) 클래스 불균형 : 네트워크 트래픽은 종종 불균형하게 분포가 된다. 예를 들어, KDDCup99 데이터셋에서 DoS 유형의 공격과 U2R 유형의 공격의 비율은 약 74,500배에 이를 정도로 클래스의 불균형이 심하다. 데이터셋이 불균형할 경우 과적합이나 편향 학습의 문제가 발생할 수 있으므로 불균형도에 따라 신중히 데이터셋을 선택할 필요가 있다. 또한, 이러한 불균형 문제를 해결하기 위해 오버샘플링(OverSampling), 언더샘플링 (UnderSampling) 등의 다양한 전처리 기법을 고려할 수 있다 [2].
6) 수집 환경 : 실제 네트워크 환경에서 수집된 데이터셋으로 학습된 모델은 실용가능성 평가에 우수한 결과를 보일 수 있다. 그러나, 데이터의 양과, 익명성 보장 등의 제한사항 때문에 많은 연구에서는 시뮬레이션 된 환경에서 수집된 데이터셋을 사용한다. 이러한 수집 환경에 대해 확인해야 한다.
7) 개인정보 보호 : 네트워크 트래픽은 출발지 IP 주소(Source IP Address), 목적지 IP 주소 (Destination IP Address) 등 다양한 개인 정보를 포함하고 있기 때문에 이에 대한 처리가 필요하다. 많은 데이터셋에서는 해당 부분을 제거하거나, 0으로 대체하는 방식(Zero padding)으로 개인정보를 보호한다. 이러한 데이터셋을 사용하는 것은 연구자의 연구윤리에 해당하므로, 이에 대한 확인이 필요하다.

본 논문에서는 이와 같은 데이터셋 선택을 위한 고려사항을 제시함으로써, 향후 연구자들의 데이터셋 선정과정을 돕고자 한다.

3.2 공개 데이터셋

현재 많은 기관과 대학에서는 다양한 네트워크 트래픽 분류를 위한 연구를 지원하기 위해, 공개 데이터셋을 구축하고 발표하였다. 본 논문에서는 3.1에서 제시한 고려사항을 바탕으로 공개 데이터셋에 대한 세부 정보를 제공하며 이는 표 1와 같다. '레이블' 항목은 데이터셋의 레이블 포함 유무를 나타내고, Task와 Class는 데이터셋의 구성요소들을 나타낸다. 불균형도는 데이터셋의 클래스 불균형 정도를 나타내는 것으로 다수 클래스의 샘플 수를 소수 클래스의 샘플 수로 나누어 계산한다. 수식은 다음과 같다.

$$IR(Imbalance\ Ratio) = \frac{N_M}{N_m} \quad (1)$$

수식 1에서 N_m은 소수 클래스의 샘플 수이고, N_M은 다수 클래스의 샘플 수이다. IR이 1보다 크면 불균형이 존재하는 것이고, 높을수록 불균형도가 심하다고 볼 수 있다. 수집환경은 해당 데이터셋이 수집된 환경이 실제 환경일 경우 Real-life로 표기, 시뮬레이션 된 환경일 경우 Benchmark로 표기하였다. 개인정보 고려 여부는 익명성이 보장되었는지 여부를 O, X로 구분하여 작성하였고, 상세설명의 경우 해당 데이터셋을 발표한 홈페이지나 논문에서 데이터셋에 대하여 추가 설명을 서술하거나, 수집 과정과 수집조건에 대한 정보를 공개하는 정도를 O, △, X로 구분하여 작성하였다. CMU-SynTraffic-2022 데이터셋은 Darknet-2020 데이터셋에서 추출한 117,620개의 실제 트래픽 데이터와 그로부터 다양한 기법으로 생성한 합성 트래픽 데이터를 포함하여 총 2,650,467개의 데이터로 구성된 데이터셋이다. 기존의 실제 트래픽 데이터는 연구에 충분한 양의 실제 데이터를 수집하기 어렵다는 점과, 클래스의 불균형으로 인해 학습 모델의 성능 이 저하되는 점, 익명성을 띄는 특성에 의해 제한된 정보와 같은 점 때문에, 이를 보완하기 위해 합성 데이터를 포함하게 되었다. 합성 데이터를 생성하기 위해 사용되는 기법과 간단한 설명은 다음과 같다 :

1) SMOTE(Synthetic Minority Over-sampling TEchnique)는 소수 클래스의 샘플 합성하여 클래스 불균형 문제를 해결하기 위한 기법으로, 딥러닝 모델의 소수 클래스 학습에 도움을 준다.
2) CTGAN(Conditional Tabular Generative Adversarial Network)은 GAN(Generative Adversarial Network)의 한 종류로, 테이블 형태의

표 1. 네트워크 트래픽 분류 공개 데이터셋
Table 1. Public datasets for network traffic classification

Datasets	Label	Year	Size	#.Task	#.Class	Purpose	IR	Environment	Privacy	Detail
UNSW-NB15 [3]	O	2015	99.1GB	2	10	IDS	534.5	Benchmark	X	△
KDDCup99 [4]	O	1999	743MB	2	5	IDS	33,715	Benchmark	O	X
NSL-KDD [5]	O	2009	75MB	5	23	IDS	306	Benchmark	O	O
CICIDS2017 [6]	O	2017	51.1GB	2	14	IDS	214,462	Benchmark	X	O
ISCX VPN 2016 [7]	O	2016	28GB	3	15	QoS	354	Real-life	X	O
USTC-TFC2016 [8]	O	2016	3.71GB	9	21	QoS	17	Real-life	O	△
CMU-SynTraffic-2022 [9]	O	2022	1.2GB	5	40	QoS	2	Synthetic	X	△

데이터를 생성하여, 조건부 확률 분포를 모델링하여, 실제와 유사한 합성 데이터를 생성한다.

3) CopulaGAN(Copula Generative Adversarial Network)는 Copula 모델과 GAN을 결합한 방법으로, 복잡한 데이터들의 상관 관계를 고려하여 합성 데이터를 생성한다.

4) VAE(Variational Autoencoder)는 AutoEncoder 중 한 종류로, 데이터의 특징을 학습하여 새로운 합성 데이터를 사용한다.

Cullen 등은 위와 같은 다양한 기법으로 생성된 합성 데이터를 포함함으로서, 해당 데이터셋은 부족한 실제 데이터에 더하여 대량의 데이터를 포함하게 되었고, 소수 클래스에 대한 보완을 통하여, 데이터 불균형 문제를 개선하였다 [9]. 이처럼 합성 데이터셋을 구축하고 공개 데이터셋으로 사용할 수 있도록 하는 연구가 진행되고 있다.

3.3 데이터셋 별 트래픽 분류 모델

많은 연구에서는 고안한 트래픽 분류 모델의 성능을 평가하기 위해 다양한 트래픽 분류 모델과의 비교를 제시한다. 본 논문에서는 데이터셋 별로 다

표 2. 데이터셋 별 트래픽 분류 모델
Table 2. Traffic classification model for each datasets

	UNSW-NB15	KDD Cup99	NSL-KDD	CIC-IDS2017	ISCX VPN 2016	USTC-TFC2016	Total
CNN	15	4	15	8	9	5	56
LSTM	13	4	13	8	4	2	44
DNN	10	5	4	9	-	-	28
RNN	4	1	5	2	1	-	13
MLP	-	2	4	3	2	-	11
GRU	2	-	2	1	-	1	6
Transformer	1	1	1	2	-	-	5
ANN	-	1	3	1	-	-	5
FNN	1	-	3	-	-	-	4
GAN	1	-	-	-	-	-	1
GCN	-	-	1	-	-	-	1
HNN	-	-	-	-	1	-	1

양한 논문에서 비교에 사용한 딥러닝 모델들에 대해 조사하고 상세한 비교를 제공하여, 후속 연구에 도움이 되고자 한다. 표 2는 최근 연구에서 사용된 기존 트래픽 분류 모델로, 총 사용 빈도수를 기준으로 내림차순으로 정렬하였다. 네트워크 트래픽 데이터는 주로 시계열 특성을 지니게 되므로, 이를 효율적으로 학습할 수 있는 CNN, LSTM, DNN 기법들이 주로 사용되는 것을 확인할 수 있다.

3.4 데이터셋 별 전처리 기법

데이터셋들은 대체로 사용하는 트래픽 분류 모델

표 3. 데이터셋 별 전처리 기법
Table 3. Pre-processing methods for each datasets

Datasets	Data Cleaning	Standardization	Feature Extraction	Splitting
UNSW-NB15	One-hot encoding (11) Label encoding (5) Linear interpolation method (1) SMOTE (2) Early Stopping (2)	Min-Max Normalization (11) Z-score Standardization (5) L2 Normalization (1) Scikit-learn StandardScaler (2)	DAE CNN ReLU함수 LSTM	5-fold (1) 10-fold (1) 3:1:1 (1) 8:2 (3) 2:1 (1) 7:3 (9) 66:31:3 (1) 16h:15h (1)
KDDCup99	One-hot encoding (5) Early Stopping (1) Label encoding (1) Zero padding (1)	Min-Max Normalization (5) Z-score Standardization (3) Scikit-learn StandardScaler (1)	AE CNN	10-fold (1) 6:4(2) 3:2(1) 17:3(1)
NSL-KDD	One-hot encoding (19) SMOTE (2) Label encoding (9) Zero padding (1) t-SNE algorithm (1) ADASYN (1) RapidMiner (1)	Min-max normalization (20) Z-score standardization (6) Log normalization (2) L2 normalization (1) Scikit-learn StandardScaler (1)	AE CNN LSTM Chi-square feature selection	8:2(5) 17:3(11) 3:7(1) 2:8(1) 4-fold (1) 7:1:2(1) By category(1) 66:31:3(1) 7:3(3)
CICIDS2017	One-hot encoding (5) SMOTE (2) Word2vec method (1) Random Oversampling method (2) Label encoding (1)	Min-max normalization (8) Z-score standardization (4) 0.1-0.9 normalization (1) Batch normalization (1)	DAE	6:2:2 1) 4:1 (5) 17:3 (1) 7:3 (2) 4-fold (1) 1:9 (1) 18:1:1 (1)
ISCX VPN2016	CIC-flowmeter (1) 15-s FLP (1) Zero padding (1) nDPI tool (1)	Min-Max normalization (4) Batch normalization (1)	LSTM CNN	60:20:20 (2) 10-fold (1) 5-fold (1) 80:10:10 (1) 70:10:20 (1)

에 따라 혹은 데이터셋의 특징이나 제공된 데이터의 형식에 따라 적절한 전처리 과정을 거쳐 연구에 사용할 데이터셋으로 구성된다. 본 논문에서는 전처리의 과정을 정제, 표준화, 특징 추출, 분할의 4단계로 구분하여, 데이터셋이 각 단계별로 어떠한 전처리 과정을 거쳐 연구 데이터셋으로 확보되는지를 비교하여 표 3으로 제시한다. 데이터 정제 단계는 데이터 형식 변환, 누락값 처리, 데이터의 불균형 처리, 과적합 방지, 데이터 크기의 조정, 차원 축소와 같은 전처리 기법을 포함하고, 표준화 단계는 정규화 기법을 포함한다. 특징 추출 단계에서는 주로 연구에서 고안한 모델을 사용하는 경우가 많고, 그와 비교하기 위한 모델은 사용 빈도수가 높은 모델만을 제공한다. 데이터 분할은 Train:Test 혹은 Train:Validation:Test와 같은 분할이나, k-fold 기법 분할을 포함한다. 본 논문에서는 다음과 같은 단계별로 데이터셋 전처리 기법을 구분하였고, 이를 통해, 해당 데이터셋을 사용한 기존의 트래픽 모델과의 비교를 할 때, 동일한 전처리 과정을 거쳐 객관적인 비교가 가능하도록 하고자 한다.

전체적으로 높은 빈도를 보이는 전처리 기법이 5가지가 존재하였는데, 데이터 정제 단계의 One-hot encoding, Label Encoding, SMOTE 기법과, 표준화 단계의 Min-Max Normalization, Z-score standardization 기법이 이에 해당된다. 데이터 정제 단계에서는 데이터의 형식을 범주형 데이터에서 숫자 형태로 변환하는 One-hot Encoding, Label Encoding 방법이 가장 많이 사용되었는데, 이는 대부분의 딥러닝 모델에서 수치형 데이터로 입력을 받기 때문에 이러한 데이터 입력 포맷에 적합하도록 데이터 변환 기법을 전처리 과정에 포함하는 것으로 보인다. 소수 클래스 데이터를 합성하는 SMOTE 기법과 소수 클래스 데이터에 더 많은 가중치를 부여하는 방식의 ADASYN 기법과 같은 오버샘플링(Oversampling) 기법 또한 높은 사용 빈도를 보인다. Wang 등은 오버샘플링기법의 중요성을 서술하면서, 기존 SMOTE기법을 사용하고 소수 클래스 샘플이 주변의 과반수 클래스 샘플과 겹칠 가능성이 존재하다는 점을 해결하기 위해, ENN 알고리즘을 활용하여 노이즈 샘플을 제거, 다운 샘플링 작업을 동시에 사용하는 방법론을 제시하였다 [10]. 정규화 단계에서는 트래픽 데이터의 모든 값을 0(최솟값)-1(최댓값) 범위 내로 변환시키는 Min-Max Normal-ization 기법과, 평균이 0이고 표준편차가 1인 형태로 변환시키는 Z-score Standardization 기법을 사용하여, 입력되는 데이터의 특징들에 대해 민감한 kNN이나 Gradient Descent과 같은 모델에 사용할 수 있도록, 데이터를 표준화 시키는 것이다. 데이터 분할에서는 주로 Train:Test 비율로서 7:3의 비율이 비교적 높은 사용빈도를 보이고, Train:Validation :Test 비율로서 3:1:1의 비율이 가장 높은 빈도를 보이는 비율이다. K-fold 분할 기법도 많은 사용빈도를 보였는데, 그 중 10-fold기법이 가장 많았다. 클래스마다 다른 비율을 가지도록 하는 경우도 있었는데, 이러한 분할 방법도 고려해볼 수 있다.

3.5 데이터셋 별 재현성 확인

많은 연구에서는 다양한 트래픽 분류 모델을 사용하여 고안한 모델을 비교한다. 하지만, 데이터셋의 전처리 과정과 전처리 후의 데이터셋 변화, 사용 딥러닝 모델의 소스코드와 알고리즘에 대해 명시되지 않으면, 후속 연구에서 해당 연구와의 연구 성과를 객관적으로 비교할 수 없다. 그러므로, 본 논문에서는 각 데이터셋을 사용한 논문 별 전처리 과정과 모델의 소스코드의 공개 여부를 표 4를 통하여 제공한다 [11-82]. 표에서 #는 논문의 인용번호, PP는 전처리과정의 공개 여부, MA는 모델의 소스코드 공개 여부를 의미한다. 전처리 과정의 공개 여부는 논문에서 제공하는 전처리 과정에 따라 3단계로 나누어 구분하였다. O는 전처리 과정에 대한 수식이나, 활용 기법에 대해 자세히 설명하고, 전처리 과정을 거친 후의 데이터셋의 변화를 자세히 설명하는 것이고, △는 전처리 과정이나 전처리 후의 데이터셋 변화에 대해 서술되었으나, 그 표현이 모호한 경우이다. X는 전처리 과정에 대한 언급이 되지 않은 경우이다. 이와 동일하게, 모델의 소스코드 공개 여부 또한 3단계로 나누어 구분하였다. O는 모델의 소스코드나 알고리즘, 각 파라미터에 대해 서술한 논문이고, △는 서술되었으나, 그 표현이 모호한 경우이다. 마지막으로, X는 모델을 공개하지 않은 경우이다. 표 4에서 확인할 수 있듯이, 대다수의 연구에서는, 사용한 데이터셋의 전처리 과정과 전처리 전후의 데이터셋 변화, 그리고 연구 모델의 소스코드에 대하여 자세하게 서술하지 않고 있다. 하지만, 후속 연구자들이 연구결과를 기존 연구와 비교하기 위해서는, 동일한 조건 하에 객관적인 비교가 가능하기 때문에, 이에 대한 공개가 필요한 상황이다. 예시로서 백의준 등은 ISCX VPN 2016 데이터셋을

표 4. 데이터셋 별 재현성
Table 4. Reproducibility for each datasets

Datasets	#	PP	MA	#	PP	MA	#	PP	MA
UNSW-NB15	[11]	X	X	[12]	△	X	[13]	△	△
	[14]	X	△	[15]	O	△	[16]	△	X
	[17]	O	X	[18]	△	O	[19]	O	△
	[20]	△	O	[21]	△	△	[22]	△	X
	[23]	△	△	[24]	X	△	[25]	△	X
	[26]	△	X	[27]	X	△	[28]	△	△
	[29]	X	O	[30]	△	O	[31]	X	△
	[32]	X	△	[33]	X	△	[34]	O	X
	[35]	△	△	[36]	△	△	[37]	X	O
	[38]	△	O						
KDDCup99	[39]	△	△	[40]	△	X			
NSL-KDD	[10]	△	△	[41]	O	△	[42]	O	△
	[43]	O	X	[44]	△	X	[45]	△	O
	[46]	△	△	[47]	O	X	[48]	X	X
	[49]	△	△	[50]	△	X	[51]	△	X
	[52]	O	△	[53]	△	X	[54]	O	△
	[55]	△	△	[56]	O	△	[57]	△	X
	[58]	O	△						
CICIDS2017	[59]	O	X	[60]	X	X	[61]	△	X
	[62]	△	X	[63]	△	X	[64]	X	△
	[65]	X	O	[66]	X	X	[67]	X	X
	[68]	O	X						
ISCX VPN2016	[69]	△	X	[70]	△	X	[71]	△	X
	[72]	△	O	[73]	△	△	[74]	△	X
	[75]	△	△	[76]	X	X	[77]	△	△
	[78]	△	△						
USTC-TFC2016	[79]	△	O	[80]	O	X	[2]	△	X
	[81]	△	X	[82]	△	△			

사용한 연구에서 네트워크 연결 또는 통신 중에 발생한 것으로 추정되는 프로토콜을 사용하는 플로우와. 3-way 핸드셰이크가 보존되지 않은 TCP 프로토콜을 사용하는 플로우를 제거하면서, 해당 전처리 과정의 필요성과, 전후의 데이터를 자세하게 공개하였다 [83]. 또한, Asgharzadeh 등은 연구에 사용된 BMECapSA 모델의 수도코드를 자세하게 공개하면서 서술하였다 [45]. 이처럼 전처리 과정과 전후 데이터셋 변화, 분류 모델의 소스코드 공개는 후속 연구자들의 객관적인 비교를 가능하게 한다.

Ⅳ. 논의 사항

현재 다양한 공개 데이터셋들이 발표되고 사용되고 있으나, 아직 많은 개선점과 연구과제가 존재한다. 이는 다음과 같다 :

1) 데이터셋의 태스크와 클래스를 명확하게 구분할 기준이 없다. 예를 들어, 네이버 메일에 관련된 트래픽의 경우, 1번 데이터셋에서는 웹 관련 트래픽으로 레이블링되고, 2번 데이터셋에서는 메일 관련 트래픽으로 레이블링된다면, 모델의 학습과정에서 혼동이 발생하여 모델의 일반화 성능을 저하시킬 수 있다. 그러므로, 이에 대한 명확한 기준이 존재해야 한다.

2) 데이터셋의 불균형도에 대한 연구가 필요하다. 불균형 데이터는 모델의 편향 학습을 야기하고, 일반화 성능을 저하시킨다. 따라서 편향 학습을 지양하기 위해, 균형적인 데이터를 학습해야 한다. 그러나, 실제 환경에서 네트워크 트래픽 데이터는 스트리밍과, FTP와 같은 서비스에 대해서는 Chat과 같은 서비스보다 더 넓은 대역의 트래픽을 발생시킨다. 그러므로, 매우 불균형한 네트워크 트래픽을 가진다. 실제 환경에 적용시킬 수 있는 네트워크 트래픽 분류 모델을 구축하기 위해서는, 이 두 가지 관점의 중점에 위치하는 적절한 불균형도에 대한 연구가 필요하다.
3) 실제 환경의 네트워크 트래픽 분류 모델을 구축하기 위해서는 실제 환경에서 발생하는 노이즈와 조건 속에서 학습하는 모델이 더 좋은 성능이 낼 수 있다. 그러나 실제 환경에서 수집한 데이터셋은 수집비용과 시간 그리고 레이블링의 어려움 때문에 현재 많은 연구에서는 가상 환경에서 수집된 벤치마크 데이터셋을 사용하고 있다. 이러한 실제 환경의 데이터셋에 대한 대안으로, 실제 환경에서 수집된 데이터셋을 바탕으로 SMOTE 기법 등 다양한 합성 기법을 통하여 합성 데이터셋을 구축하여 사용하는 방법이 연구 중이다. [24: 논문 147]에서는 CMU-SynTraffic-2022 합성 데이터셋을 소개한다. 이와 같은 데이터셋의 시뮬레이션 된 수집 환경에 대한 보완이 요구된다.
4) 각 연구에서는 데이터셋 사용 시 각 전처리 단계와 분할 비율을 포함하여, 전후의 데이터셋 변화, 딥러닝 모델의 소스코드 공개가 필요하다. 본론에서 언급하였듯, 연구 모델의 비교를 위해서는 객관적인 비교가 가능하도록 재현성이 확보되어야 한다. 따라서 연구에 대한 객관성을 위해 더 상세한 전처리 과정과 모델 소스코드가 필요하다.
5) 데이터셋의 클래스 최신화가 필요하다. 최근 Similarweb 사이트에서 공개한 응용 프로그램 사용 순위를 확인해보면, 배달의 민족과 같은 배달 응용 프로그램과, 비교적 최근 급속도로 발전하는 웨어러블(Wearable) 기기 응용 트래픽과 같은 새로운 클래스에 대한 도입이 필요하다. 현재까지의 데이터셋은 기존의 네트워크 트래픽에 대한 데이터만 포함하고 있었지만, 응용 프로그램의 다양화와 발전으로 인해, 트래픽 클래스의 최신화 또한 연구가 필요하다.

Ⅴ. 결론

본 논문에서는 네트워크 트래픽 분류 분야의 최신 연구 동향에 대한 검토를 제공한다. 먼저, 네트워크 트래픽 분류 모델을 학습시키기 위한 데이터셋 선정 시 고려사항을 제시하고, 이를 바탕으로 공개 데이터셋의 세부정보를 비교하였다. 그 후, 데이터셋별로 고안된 연구의 비교 목적으로 사용된 딥러닝 모델을 비교하였다. 또한, 데이터셋 별 전처리 기법을 데이터 정제, 표준화, 특징 추출, 분할의 4단계로 분류하여 동일한 조건 하에 비교하기를 돕고자 하였고, 데이터셋 별로 각 논문에서 전처리 과정과 그에 따른 데이터셋의 변화, 딥러닝 모델의 소스코드나 알고리즘 공개 여부에 대한 확인을 통해 객관적인 비교가 가능한 연구에 대해 제시하였다. 이를 통하여 후속 연구자들의 데이터셋 선정과 적절한 전처리 과정의 선택을 지원하고자 한다. 또한, 기존 연구와의 객관적인 비교를 위한 재현성을 제시함으로서, 후속 연구에서 고안될 모델의 비교를 돕고자 한다. 향후 연구과제로서는 태스크와 클래스를 명확하게 구분할 기준과 데이터셋 불균형도에 대한 연구를 제시하고, 향후 연구에서 모델의 재현을 통한 비교를 위하여 전처리과정과 모델의 소스코드에 대한 공개를 촉구하는 바이다. 마지막으로, 실제 환경의 네트워크 트래픽 데이터셋 수집의 어려움을 보완하고, 시대의 기술 발전에 따른 최신 트래픽 클래스를 추가하는 새로운 데이터셋의 발전을 미래 연구방향으로 제시한다.

References

[1] Jingjing Zhao, Xuyang Jing, Zheng Yan, Witold Pedrycz, “Network traffic classification for data fusion: A survey”, Information Fusion, Volume 72, 2021, Pages 22-47, ISSN 1566-2535, (https://doi.org/10.1016/j.inffus.2021.02.009.)

[2] DONG, Shi; XIA, Yuanjun; PENG, Tao, “Traffic identification model based on generative adversarial deep convolutional network”, Annals of Telecommunications, 77, 573-587, 23 August 2021.(https://doi.org/10.1007/s12243-021-00876-6)

[3] N. Moustafa and J. Slay, ″UNSW-NB15: a

comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 2015, pp. 1-6, (https://doi.org/10.1109/MilCIS.2015.7348942)

[4] KDDCup99, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (Accessed 14 December 2023).

[5] RUIZHE ZHAO, February 2, 2022, "NSL-KDD", IEEE Dataport, (https://dx.doi.org/10.21227/8rpg-qt98.)

[6] Sharafaldin, Iman & Habibi Lashkari, Arash & Ghorbani, Ali, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", 108-116, January 2018. (https://doi.org/10.5220/0006639801080116)

[7] Gerard Drapper Gil, Arash Habibi Lashkari, Mohammad Mamun, Ali A. Ghorbani, "Characterization of Encrypted and VPN Traffic Using Time-Related Features", In Proceedings of the 2nd International Conference on Information Systems Security and Privacy(ICISSP 2016) , pages 407-414, Rome, Italy, February 2016. (https://doi.org/10.5220/0005740704070414)

[8] USTC-TFC2016, https://github.com/yungshenglu/USTC-TFC2016 (Accessed 14 December 2023)

[9] D. Cullen, J. Halladay, N. Briner, R. Basnet, J. Bergen and T. Doleck, "Evaluation of Synthetic Data Generation Techniques in the Domain of Anonymous Traffic Classification", in IEEE Access, vol. 10, pp. 129612-129625, 2022, (https://doi.org/10.1109/ACCESS.2022.3228507.

[10] Shiyu Wang, Wenxiang Xu, Yiwen Liu, "Res-TranBiLSTM: An intelligent approach for intrusion detection in the Internet of Things, Computer Networks", Volume 235, 109982, ISSN 1389-1286, 2023. (https://doi.org/10.1016/j.comnet.2023.109982.)

[11] Dutta, Vibekananda & Pawlicki, Marek & Kozik, Rafał & Choraś, Michał, "Unsupervised network traffic anomaly detection with deep autoencoders", Logic Journal of IGPL, p 30, February 2022. (https://doi.org/10.1093/jigpal/jzac002)

[12] Zhaoxu Ding, Guoqiang Zhong, Xianping Qin, Qingyang Li, Zhenlin Fan, Zhaoyang Deng, Xiao Ling, Wei Xiang, "MF-Net: Multi-frequency intrusion detection network for Internet traffic data", Pattern Recognition, Volume 146, 109999, ISSN 0031-3203, 2024 (https://doi.org/10.1016/j.patcog.2023.109999.)

[13] Feng, Y., Wang, C. Network Anomaly "Early Warning through Generalized Network Temperature and Deep Learning". J Netw Syst Manage 31, 38, 2023. (https://doi.org/10.1007/s10922-023-09727-2)

[14] Prabhakaran, V., Kulandasamy, A. "mLBOA-DML: modified butterfly optimized deep metric learning for enhancing accuracy in intrusion detection system". J Reliable Intell Environ 9, 333–347 , 2023. (https://doi.org/10.1007/s40860-022-00197-y)

[15] Acharya, Toya & Annamalai, Annamalai & Chouikha, Mohamed, "Efficacy of CNN-Bidirectional LSTM Hybrid Model for Network-Based Anomaly Detection", 348-353, May 2023. (https://doi.org/10.1109/ISCAIE57739.2023.10165088)

[16] Hailong Xie, Chenxian Hao, Jie Li, Min Li, Peng Luo, Jinpeng Zhu, "Anomaly Detection For Time Series Data Based on Multi-granularity Neighbor Residual Network", International Journal of Cognitive Computing in Engineering, Volume 3, Pages 180-187, ISSN 2666-3074, June 2022. (https://doi.org/10.1016/j.ijcce.2022.10.001.)

[17] Bowen, B., Chennamaneni, A., Goulart, A. et al, "BLoCNet: a hybrid, dataset-independent intrusion detection system using deep learning", Int. J. Inf. Secur. 22, 893–917, March 2023. (https://doi.org/10.1007/s10207-023-00663-5)

[18] Ghani, H. & Virdee, Bal & Salekzamankhani, Shahram, "A Deep Learning Approach for Network Intrusion Detection Using a Small Features Vector", Journal of Cybersecurity and

Privacy, 3, 451-463, August 2023. (https://doi.org/10.3390/ jcp3030023)

[19] Lal, Jaya & Ayoub, Shahnawaz & Prashant, Dr & Subbian, Prabagar & Vijay, Dr & Tiwari, Mohit. "Hybrid Deep Learning based Attack Detection and Classification Model on IoT Environment". July 2023. (https://doi.org/10.1109/ICECCT56650.2023.10179678)

[20] Y. Yue, X. Chen, Z. Han, X. Zeng and Y. Zhu, "Contrastive Learning Enha -nced Intrusion Detection", in IEEE Transactions on Network and Service Management, vol. 19, no. 4, pp. 4232-4247, December 2022, (https://doi.org/10.1109/TNSM.2022.3218843.)

[21] T.V. Ramana, M. Thirunavukkarasan, Amin Salih Mohammed, Ganesh Gopal Devarajan, Senthil Murugan Nagarajan, "Ambient intelligence approach: Internet of Things based decision performance analysis for intrusion detection", Computer Communications, Volume 195, Pages 315-322, ISSN 0140-3664, 2022. (https://doi.org/10.1016/j.comcom.2022.09.007)

[22] J. Du, K. Yang, Y. Hu and L. Jiang, "NIDS-CNNLSTM: Network Intrusion Detection Classification Model Based on Deep Learning", in IEEE Access, vol. 11, pp. 24808-24821, 2023. (https://doi.org/10.1109/ACCESS.2023.3254915)

[23] Sakthi, K. & Kumar, Palanichamy, "A novel attention-based feature learning and optimal deep learning approach for network intrusion detection". Journal of Intelligent & Fuzzy Systems, 45, 1-18, July 2023. (https://doi.org/10.3233/ JIFS-231758.)

[24] Ponmalar, A. & Dhanakoti, V, "Hybrid Whale Tabu algorithm optimized convolutional neural network architecture for intrusion detection in big data", Concurrency and Computation: Practice and Experience, 34, May 2022. (https://doi.org/10.1002/ cpe.7038.)

[25] Amma, N.G.B. "A vector convolutional deep autonomous learning classifier for detection of cyber attacks". Cluster Comput 25, 3447–3458, January2022. (https://doi.org/10.1007/s10586-022-03577-4)

[26] Binbusayyis, A., Vaiyapuri, T., "Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM", Appl Intell 51, 7094–7108, January 2021. (https://doi.org/10.1007/s10489-021-02205-9)

[27] Mijalkovic, J.; Spognardi, "A. Reducing the False Negative Rate in Deep Learning Based Network Intrusion Detection Systems". Algorithms, 15, 258, 26 July 2022. (https://doi.org/10.3390 /a15080258)

[28] Giuseppina Andresini, Annalisa Appice, Luca De Rose, Donato Malerba, "GAN augmentation to deal with imbalance in imaging-based intrusion detection", Future Generation Computer Systems, Volume 123, Pages 108-127, ISSN 0167-739X, October 2021. (https://doi.org/10.1016/j.future. 2021.04.017)

[29] P Rajesh Kanna, P Santhi, Unified "Deep Learning approach for Efficient Intrusion Detection System using Integrated Spatial–Temporal Features", Knowledge-Based Systems, Volume 226, 107132, ISSN 0950-7051, 17 August 2021. (https://doi.org/10.1016/j.knosys.2021.107132.)

[30] Sreelatha, G., Babu, A.V. & Midhunchakkaravarthy, D. "Improved security in cloud using sandpiper and extended equilibrium deep transfer learning based intrusion detection". Cluster Comput 25, 3129–3144, 28 January 2022. (https://doi.org/10.1007 /s10586-021-03516-9)

[31] P. Rajesh Kanna, P. Santhi, Hybrid "Intrusion Detection using MapReduce based Black Widow Optimized Convolutional Long Short-Term Memory Neural Networks", Expert Systems with Applications, Volume 194, 116545, ISSN 0957-4174, 15 May 2022. (https://doi.org/10.1016/j.eswa.2022.116545.)

[32] Abdulla, Shubair & Alashoor, Ahmed, "An Artificial Deep Neural Network for the Binary Classification of Network Traffic", International Journal of Advanced Computer Science and Applications, 11, January 2020. (https://doi.org/10.14569/IJACSA.2020.0110150

.)

[33] Keserwani, Pankaj & Govil, Mahesh & Shubhakar, Emmanuel, "An Optimal Intrusion Detection System using GWO-CSA-DSAE Model", Cyber-Physical Systems, 7, 1-24, August 2020.(https://doi.org/10.1080/23335777.2020. 1811383.)

[34] Manuel Lopez-Martin, Belen Carro, Antonio Sanchez-Esguevillas, Jaime Lloret, "Shallow neural network with kernel approximation for prediction problems in highly demanding data networks", Expert Systems with Applications, Volume 124, Pages 196-208, ISSN 0957-4174, 15 June 2019. (https://doi.org/10.1016/j.eswa.2019.01.063.)

[35] Vaiyapuri, Thavavel & Binbusayyis, Adel., "Enhanced Deep Autoencoder Based Feature Representation Learning for Intelligent Intrusion Detection System", Computers, Materials & Continua, 68, 3271-3288, January 2021. (10.32604/cmc.2021.017665)

[36] Sethi, K., Sai Rupesh, E., Kumar, R. et al, "A context-aware robust intrusion detection system: a reinforcement learning-based approach", Int. J. Inf. Secur. 19, 657–678, 03 December 2019. (https://doi.org/10.1007/s10207-019-00482-7)

[37] H. He, X. Sun, H. He, G. Zhao, L. He and J. Ren, "A Novel Multimodal-Sequential Approach Based on Multi-View Features for Network Intrusion Detection", in IEEE Access, vol. 7, pp. 183207-183221, 12 December 2019. (https://doi.org/10.1109/ ACCESS. 2019.2959131.)

[38] Ibor, A.E., Oladeji, F.A., Okunoye, O.B. et al, "Conceptualisation of Cyberattack prediction with deep learning", Cybersecur 3, 14, 17 June 2020. (https://doi.org/10.1186/s42400-020-00053-7)

[39] G. Muhammad, M. S. Hossain and S. Garg, "Stacked Autoencoder-Based Intrusion Detection System to Combat Financial Fraudulent", in IEEE Internet of Things Journal, vol. 10, no. 3, pp. 2071-2078, 1 Feb.1, 2023, (https://doi.org/10.1109/JIOT.2020.3041184.)

[40] B. Yan and G. Han, "Effective Feature Extraction via Stacked Sparse Autoencoder to Improve Intrusion Detection System", in IEEE Access, vol. 6, pp. 41238-41248, 23 July 2018. (https://doi.org/10.1109/ACCESS.2018.2858277.)

[41] Almahadin, Ghayth & Aoudni, Yassine & Shabaz, Dr. Mohammad & Agrawal, Anurag & Yasmin, Ghazaala & Alomari, Esraa & Al-Khafaji, Hamza Mohammed Ridha & Dansana, Debabrata & Maaliw III, Renato., "VANET Network Traffic Anomaly Detection Using GRU-Based Deep Learning Model", IEEE Transactions on Consumer Electronics, PP, October 2023. (https://doi.org/10.1109/TCE. 2023.3326384)

[42] H. Xu, L. Sun, G. Fan, W. Li and G. Kuang, "A Hierarchical Intrusion Detection Model Combining Multiple Deep Learning Models With Attention Mechanism", in IEEE Access, vol. 11, pp. 66212-66226, 28 June 2023.(https://doi.org/10.1109/ACCESS.2023.3290613.)

[43] Çavuşoğlu, Ü., Akgun, D. & Hizal, S. "A Novel Cyber Security Model Using Deep Transfer Learning." Arab J Sci Eng, 24 July 2023. (https://doi.org/10.1007/s13369-023-08092-1)

[44] Lv, Xiang & Han, Dezhi & Li, Dun & Xiao, Lijun & Chang, Chin-Chen, "Network abnormal traffic detection method based on fusion of chord similarity and multiple loss encoder", EURASIP Journal on Wireless Communications and Networking, October 2022. (https://doi.org/10.1186 /s13638-022-02180-w.)

[45] Hossein Asgharzadeh, Ali Ghaffari, Mohammad Masdari, Farhad Soleimanian, harehchopogh,"Anomaly -based intrusion detection system in the Internet of Things using a convolutional neural network and multi-objective enhanced Capuchin Search Algorithm", Journal of Parallel and Distributed Computing, Volume 175, 2023, Pages 1-21, ISSN 0743-7315, January 9 2023.

(https://doi.org/10 .1016/j.jpdc.2022.12.009.)

[46] Li, Xue Jun, Maode Ma, and Yihan Sun. 2023. ″An Adaptive Deep Learning Neural Network Model to Enhance Machine-Learning-Based Classifiers for Intrusion Detection in Smart Grids″ Algorithms 16, no. 6, 288, 2 June 2023. (https://doi.org/10.3390 /a16060288)

[47] Azzaoui, H., Boukhamla, A.Z.E., Arroyo, D. et al. ″Developing new deep-learning model to enhance network intrusion classification″. Evolving Systems 13(2022), pp 17–25, 19 January 2021. (https://doi.org/10. 1007/s12530-020-09364-z)

[48] A.H. Nasreen Fathima, S.P. Syed Ibrahim, ″Multi-stage deep investigation pipeline on detecting malign network traffic″, Materials Today: Proceedings, Volume 62, Part 7, (2022), Pages 4726-4731, ISSN 2214-7853, 30 March 2022.(https:// doi.org/10 .1016/j.matpr.2022.03.211.)

[49] MD Moizuddin, M. Victor Jose, ″A bio-inspired hybrid deep learning model for network intrusion detection, Knowledge-Based Systems″, Volume 238, 2022, 107894, ISSN 0950-7051, December 11 2021. (https://doi.org/10. 1016/j.knosys.2021.107894.)

[50] Qigang Liu, Deming Wang, Yuhang Jia, Su yuan Luo, Chongren Wang, ″A multi-task based deep learning approach for intrusion detection, Knowledge-Based Systems″, Volume 238, 2022, 107852, ISSN 0950-7051, 10 December 2021 (https://doi.org/10. 1016/j.knosys2021.107852.)

[51] Alshammri, Ghalib & Samha, Amani & Hemdan, Ezz El-Din & Amoon, Mohammed & El-Shafai, Walid, ″An Efficient Intrusion Detection Framework in Software-Defined Networking for Cybersecurity Applications, Computers″, Materials and Continua. 72. pp3529–3548, March 2022. (https://doi.org/10.32604/cmc. 2022.025262)

[52] Imrana, Yakubu, Yanping Xiang, Liaqat Ali, Zaharawu Abdul-Rauf, Yu-Chen Hu, Seifedine Kadry, and Sangsoon Lim, ″χ2-BidLSTM: A Feature Driven Intrusion Detection System Based on χ2 Statistical Model and Bidirectional LSTM″ Sensors 22, no. 5, 4 March 2022 (https://doi.org/10. 3390/s22052018)

[53] Wang, Zhihao & Jiang, Dingde & Liuwei, Huo & Yang, Wei, ″An efficient network intrusion detection approach based on deep learning″. Wireless Networks. pp 1-14, July 2021.(https:// doi.org/10.1007/s11276-021-02698-9)

[54] Kalaivani, K., and M. Chinnadurai. ″A Hybrid Deep Learning Intrusion Detection Model for Fog Computing Environment.″ Intelligent Automation &Soft Computing 30.1, 16 April 2021. (https://doi.org/10.32604/iasc.2021.017515)

[55] Judy Simon, N. Kapileswar, Phani Kumar Polasi, M. Aarthi Elaveini, Hybrid intrusion detection system for wireless IoT networks using deep learning algorithm, Computers and Electrical Engineering, Volume 102, 2022, 108190, ISSN 0045-7906, 1 July 2022. (https://doi.org/10.1016/j. compeleceng.2022.108190)

[56] Muhuri, Pramita & Chatterjee, Prosenjit & Yuan, Xiaohong & Roy, Kaushik & Esterline, Albert. ″Using a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to Classify Network Attacks″. Information. 11. 243. May 2020 (https://doi.org/ 10.3390/info11050243)

[57] J. Hassannataj Joloudari, M. Haderbadi, A. Mashmool, M. Ghasemigol, S. S. Band and A. Mosavi, ″Early Detection of the Advanced Persistent Threat Attack Using Performance Analysis of Deep Learning″, in IEEE Access, vol. 8, pp. 186125-186137, 2020, (https://doi.org/ 10.1109/ACCESS.2020.3029202.

[58] J. Hassannataj Joloudari, M. Haderbadi, A. Mashmool, M. Ghasemigol, S. S. Band and A. Mosavi, ″Early Detection of the Advanced Persistent Threat Attack Using Performance Analysis of Deep Learning,″ in IEEE Access, vol. 8, pp. 186125-186137, 06 October 202. (https://doi.org/10.1109/ACCESS.2020.3029202)

[59] Alrayes FS, Zakariah M, Driss M, Boulila W, Deep Neural Decision Forest "(DNDF): A Novel Approach for Enhancing Intrusion Detection Systems in Network Traffic Analysis". Sensors, 2023, 23(20), 8362, 10 October 2023. (https://doi.org/10.3390/s23208362)

[60] M. J. Hashemi, E. Keller and S. Tizpaz-Niari, "Detecting Unseen Anomalies in Network Systems by Leveraging Neural Networks," in IEEE Transactions on Network and Service Management, vol. 20, no. 3, pp. 2515-2528, Sept. 2023, 09 November 2022. (https://doi.org/10.1109/TNSM. 2022.3220775.)

[61] Nan Wei, Lihua Yin, Xiaoming Zhou, Chuhong Ruan, Yibo Wei, Xi Luo, Youyi Chang, Zhao Li, "A feature enhancement-based model for the malicious traffic detection with small-scale imbalanced dataset", Information Sciences, Volume 647, 2023, 119512, ISSN 0020-0255, 11 August 2023. (https://doi.org/10. 1016/j. ins2023.119512.)

[62] Elnakib, O., Shaaban, E., Mahmoud, M. et al. "EIDM: deep learning model for IoT intrusion detection systems". J Supercomput 79 (2023), 13241–13261, 22 March 2023. (https://doi.org/10. 1007/s11227-023-05197-0)

[63] Li, Yimin & Han, Dezhi & Cui, Mingming & Yuan, Fan & Zhou, Yachao, "RESNETCNN: An abnormal network traffic flows detection model", Computer Science and Information Systems, 20, pp. 4-4, January 2023. (https://doi.org/10.2298/CSIS221124004L)

[64] He, Junpeng & Luo, Lei & Xiao, Kun & Fang, Xiyu & Li, Yun, "Generate qualified adversarial attacks and foster enhanced models based on generative adversarial networks", Intelligent Data Analysis, 26, 1359-1377, September 2022. (https://doi.org/10.3233/IDA-216134.)

[65] Y. Hou, S. G. Teo, Z. Chen, M. Wu, C. -K. Kwoh and T. Truong-Huu, "Handling Labeled Data Insufficiency: Semi-supervised Learning with Self-Training Mixup Decision Tree for Classification of Network Attacking Traffic," in IEEE Transactions on Dependable and Secure Computing, pp. 1-14, 01 August 2022. (https://doi.org/ 10.1109/TDSC.2022.3195534.)

[66] T. Ye, G. Li, I. Ahmad, C. Zhang, X. Lin and J. Li, "FLAG: Few-Shot Latent Dirichlet Generative Learning for Semantic-Aware Traffic Detection," in IEEE Transactions on Network and Service Management, vol. 19, no. 1, pp. 73-88, March 2022.(https://doi.org/ 10.1109/TNSM.2021.3131266.)

[67] Mo, Chen & Xiaojuan, Wang & Mingshu, He & Lei, Jin & Javeed, Khalid & Wang, Xiaojun, "Network Traffic Classification Model Based on Metric Learning", Computers, Materials & Continua, 64, pp. 941-959, January 2020. ·(https://doi.org/10. 32604/cmc.2020.09802)

[68] Sun, Pengfei & Liu, Pengju & Li, Qi & Liu, Chenxi & Lu, Xiangling & Hao, Ruochen & Chen, Jinpeng, "DL-IDS: Extracting features using CNN-LSTM hybrid network for intrusion detection system", Security and Communication Networks, 2020, pp. 1-11, (https:// doi.org/10.1155/2020/8890306)

[69] Ul Islam, Faiz & Liu, Guangjie & Liu, Weiwei & Haq, Qazi Mazhar Ul, "A deep learning based framework to identify and characterise heterogeneous secure network traffic", IET Information Security, 17, 294-308, October 2022. (https://doi.org/10.1049/ise2.12095)

[70] Y. Yang et al., "A Network Traffic Classification Method Based on Dual-Mode Feature Extraction and Hybrid Neural Networks", in IEEE Transactions on Network and Service Management, vol. 20, no. 4, pp. 4073-4084, March 27. 2023. (https://doi.org/10.1109/TNSM.2023.3262246)

[71] B. Mareri, G. Owusu Boateng, R. Ou, G. Sun, Y. Pang and G. Liu, "MANTA: Multi-Lane Capsule Network Assisted Traffic Classification for 5G Network Slicing," in IEEE Wireless Communications Letters, vol. 11, no. 9, pp. 1905-1909, 27 June 2022, (https://doi.org/10.1109/LWC.2022.3186529)

[72] Wenting Wei, Huaxi Gu, Wenshuai Deng, Zhe Xiao, Xinming Ren, "ABL-TC: A lightweight

design for network traffic classification empowered by deep learning", Neurocomputing, Volume 489, Pages 333-344, ISSN 0925-2312, 18 March 2022. (https://doi.org/10.1016/j. neucom.2022.03.007.)

[73] A. Telikani, A. H. Gandomi, K. -K. R. Choo and J. Shen, "A Cost-Sensitive Deep Learning-Based Approach for Network Traffic Classification," in IEEE Transactions on Network and Service Management, vol. 19, no. 1, pp. 661-670, September 13 2021, (https://doi.org/10.1109/TNSM.2021.3112283.)

[74] P. Wang, F. Ye, X. Chen and Y. Qian, "Datanet: Deep Learning Based Encrypted Network Traffic Classification in SDN Home Gateway," in IEEE Access, vol. 6, pp. 55380-55391, September 27 2018. (https://doi.org/10.1109/ACCESS.2018.2872430.)

[75] Nikolić, Nedeljko & Tomovic, Slavica & Radusinovic, Igor, "A Comparative Study of Deep Learning and Decision Tree Based Ensemble Learning Algorithms for Network Traffic Identification", Telfor Journal, 14, 61-66, December 2022. (https://doi.org/10.5937/telfor2202061N)

[76] Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, Antonio Pescapé, DISTILLER: Encrypted traffic classification via multimodal multitask deep learning, Journal of Network and Computer Applications, Volumes 183–184, 102985, ISSN 1084-8045, 20 January 2021. (https://doi.org/10.1016/j.jnca.2021.102985)

[77] Liu, Xinlei, "Identification of Encrypted Traffic Using Advanced Mathematical Modeling and Computational Intelligence", Mathematical Problems in Engineering, 2022, pp. 1-10, August 2022. (https://doi.org/10.1155/2022/1419804)

[78] Izadi, S., Ahmadi, M. &Rajabzadeh, A, "Network Traffic Classification Using Deep Learning Networks and Bayesian Data Fusion", J Netw Syst Manage 30, 25 (2022), pp. 1–21, January 20 2022. (https://doi.org/10.1007/s10922-021-09639-z)

[79] Zhang, Weijie, Lanping Zhang, Xixi Zhang, Yu Wang, Pengfei Liu, and Guan Gui, "Intelligent Unsupervised Network Traffic Classification Method Using Adversarial Training and Deep Clustering for Secure Internet of Things", Future Internet 15, no. 9, 298, 1 September 2023. (https://doi.org/10.3390/fi15090298)

[80] Li, Daoquan & Dong, Xueqing & Gao, Jie & Hu, Keyong, "Abnormal Traffic Detection Based on Attention and Big Step Convolution", IEEE Access, p.1, January 2023. (https://doi.org/10.1109/ACCESS.2023.3289200)

[81] Guan, J., Cai, J., Bai, H. et al, "Deep transfer learning-based network traffic classification for scarce dataset in 5G IoT systems", Int. J. Mach. Learn. &Cyber, 12, 3351–3365, August 19 2021.(https://doi.org/10.1007/s13042-021-01415-4)

[82] H. Yao, P. Gao, J. Wang, P. Zhang, C. Jiang and Z. Han, "Capsule Network Assisted IoT Traffic Classification Mechanism for Smart Cities", in IEEE Internet of Things Journal, vol. 6 (2019), no. 5, pp. 7515-7525, February 24 2019. (https://doi.org/10.1109/JIOT.2019.2901348)

[83] U. -J. Baek, M. -S. Lee, J. -T. Park, J. -W. Choi, C. -Y. Shin and M. -S. Kim, "Preprocessing and Analysis of an Open Dataset in Application Traffic Classification," 2023 24st Asia-Pacific Network Operations and Management Symposium (APNOMS), Sejong, Korea, Republic of, 2023, pp. 227-230. September 25 2023.

장 윤 성 (Yoon-Seong Jang)

2018 고려대학교, 컴퓨터융합 소프트웨어학과 학사

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

박 지 태 (Jee-Tae Park)

2017년 : 고려대학교 컴퓨터정보학과 학사

2017년 ~ 현재 : 고려대학교 컴퓨터정보학과 석박사 통합 과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

백 의 준 (Ui-Jun Baek)

2018년 : 고려대학교 컴퓨터정보학과 학사

2018년 ~ 현재 : 고려대학교 컴퓨터정보학과 석박사 통합 과정

<관심분야> 블록체인 거래 모니터링, 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

김 주 성 (Ju-Sung Kim)

2018 고려대학교, 컴퓨터융합 소프트웨어학과 학사

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

이 대 국(Dae-Guk Lee)

2004년 : 영국 런던 대학 UCL 컴퓨터 과학 학사

2018년 : 고려대학교 컴퓨터정보학과 석사

2018년 ~ 현재 : 고려대학교 컴퓨터정보학과 박사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

김 명 섭(Myung-Sup Kim)

1998년 : 포항공과대학교 전자계산학과 학사

2000년 : 포항공과대학교 전자계산학과 석사

2004년 : 포항공과대학교 전자계산학과 박사

2006년 : Dept. of ECS, Univ of Toronto Canada

2006년 ~ 현재 : 고려대학교 컴퓨터정보학과 교수

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크