基于机器学习与预训练模型的流量分析方法综述

李学相，郑永利，张怡泽，段鹏松
（郑州大学网络空间安全学院，河南 郑州 450002）

摘　要： 随着互联网的普及与应用程序的多样化，海量网络流量的精细化分类成为优化服务质量和分析用户行为模式的关键。对基于机器学习（ML）和基于预训练模型的网络流量分析方法进行概述，旨在通过多维度对比与分析，推动该领域研究的进一步发展。首先，解析了流量分类的完整流程，涵盖了数据采集、预处理、特征提取过程，分析了数据平衡技术的实践价值，同时介绍了主流公共数据集的数据格式、规模及场景适配性等，从多角度进行对比分析，指出其存在的数据分布、特征冗余与时效性问题。然后，不仅在方法层面总结了传统算法在高维数据处理与实时性上的局限性，还重点通过实验结果对比分析，总结了流量分析领域应用预训练模型技术的趋势，包括基于 Transformer 的预训练模型、与深度学习（DL）的融合模型和轻量化模型在流量分类中的突破性进展。最后，结合动态研究趋势，探讨了未来应用预训练模型存在的机遇和挑战，分析了其在计算成本与隐私保护方面的局限性，提出了未来的研究方向并对研究前景进行展望。

关键词： 流量分析；机器学习；深度学习；预训练模型；特征提取；联邦学习
中图分类号： TP393　　**文献标志码：** A　　**DOI：** 10.19678/j.issn.1000-3428.0252301

Review of Traffic Analysis Methods Based on Machine Learning and Pre-trained Model

LI Xuexiang, ZHENG Yongli, ZHANG Yize, DUAN Pengsong
(School of Cyber Space Security, Zhengzhou University, Zhengzhou 450002, Henan, China)

【Abstract】 With the popularization of the Internet and the diversification of applications, fine-grained classification of massive network traffic has become key to optimizing quality of service and analyzing user behavior patterns. This paper presents an overview of Machine Learning (ML)-based and pretrained model-based network traffic analysis methods to promote further research and development in this field through multidimensional comparison and analysis. First, the complete traffic classification pipeline is deconstructed, covering data acquisition, preprocessing, and feature extraction, and the practical value of data balancing techniques is examined. The data format, scale, and scene suitability of mainstream public datasets are introduced, compared, and analyzed from multiple perspectives, highlighting their data distribution, feature redundancy, and timeliness problems. Second, it summarizes the limitations of traditional algorithms in handling high-dimensional data and meeting real-time requirements, and outlines the trend of applying pretrained models in traffic analytics, through a focused comparative analysis of experimental results. This review includes breakthroughs in Transformer-based pretrained models, their fusion with Deep Learning (DL) models, and advances in lightweight pretrained models for traffic classification. Finally, by considering dynamic research trends, the opportunities and challenges in future applications of pretrained models are discussed, and their limitations in terms of computational cost and privacy protection are analyzed.

【Key words】 traffic analysis; Machine Learning (ML); Deep Learning (DL); pre-trained model; feature extraction; Federated Learning (FL)

0　引言

网络流量分析作为网络管理与安全领域的关键技术，旨在通过解析数据包或数据流的特征，准确识别其对应的应用类型或行为模式[1]。

在海量网络流量中，用户行为模式差异显著，精细化分类能够有效挖掘数据价值[2]。在优化服务质量方面，精细化分类可实现流量的智能调度与资源分配。在云游戏场景中[3]，区分游戏控制指令、高清画面流与语音通信数据，针对性优化传输优先级，能

基金项目： 郑州市协同创新重大专项（20XTZX06013）；中国工程科技发展战略河南研究院战略咨询研究项目（2022HENYB03）；河南省科技攻关项目（232102210050，242102210060）。
作者简介： 李学相，男，教授，主研方向为云计算、物联网安全研究；郑永利、张怡泽，硕士研究生；段鹏松（CCF 专业会员、通信作者），副教授、博士。
收稿日期： 2025-04-08　　**修回日期：** 2025-07-19
E-mail： duanps@zzu.edu.cn

够降低画面卡顿率，显著提升用户体验。在远程手术中，减少主控制器到从控制器以及传感器到控制器的时间延迟、数据包丢失和抖动，提高了手术精度[4]。在用户行为分析领域，精细化分类可助力构建多维度用户画像。在电商场景中[5]，分析用户购物流量中的细粒度特征，能够精准构建用户兴趣图谱，使个性化推荐的点击率提升。在智能汽车场景中[6]，对车载终端流量精细分类，能实现有效的入侵检测系统，应对攻击带来的安全威胁。

尽管流量分析在恶意软件通信检测[7-8]、流量分类和用户行为分析[9-10]等任务中应用广泛并取得了较多的成果。但是，随着网络规模指数级别的增长与加密技术的广泛应用，网络流量分类面临动态协议适配性不足，加密流量特征隐匿以及实时性要求严苛等多重挑战。实现高精度、低延迟且适应复杂网络环境的分类方法成为研究者共同关注的焦点。

流量分析技术在深度学习(DL)与预训练模型技术的推动下，实现了从人工经验驱动到数据智能驱动的范式转变。现有研究技术路线一般分为基于传统机器学习(ML)的方法、基于 DL 的方法和预训练模型方法。基于传统 ML 的方法[11-15]依赖专家经验，特征设计难以应对加密流量的动态伪装。基于 DL 的方法计算成本高和标注数据需求大的缺陷限制了其在边缘场景中的部署。预训练模型方法主要是以 Transformer 架构为代表的各种预训练模型实现预训练范式革新和计算架构升级，提升了跨任务迁移能力。

文献[16]对 ML 算法在异常行为检测任务中的具体应用过程进行综述，分析了使用的方法以及不同方法之间的优缺点。文献[17]系统总结了网络流量指纹识别的方法和数据集，综述了设备识别任务中使用基于规则、基于传统 ML 和基于 DL 的方法。大多研究聚焦传统 ML 或 DL 某一类方法，未建立数据采集、特征提取等各环节间的逻辑关联，难以形成完整的流量分析方法论体系，并且缺乏对新兴预训练模型、联邦学习(FL)等前沿技术的实验数据验证以及系统性探讨。尽管已有研究取得了一定的进展，但仍需要对当前技术的局限性进行深入探索。当前公开数据集研究普遍存在过度的特征提取、未充分利用数据资源以及严重的数据不平衡等问题。

本研究旨在系统梳理流量分类领域的技术演进脉络，通过对比分析 ML 方法与预训练模型方法的优势与局限性，探索融合预训练模型与轻量化技术的新型分类框架的研究现状。首先，对比介绍了传统 ML 和 DL 方法以及当前主流技术和局限性，总结了预训练模型技术的改进模型、融合框架以及轻量化模型，并对预训练模型在数据集上的实验结果进行分析。然后，归纳流量分类流程中数据预处理和特征提取等关键环节的技术挑战与优化策略，解析近年来主流数据集的构建逻辑与适用场景，总结其在特征维度、数据分布与时效性上的不足。最后，对网络流量分析方法的未来研究发展进行了展望。

1 网络流量分析基本流程

图 1 是流量分类的主要框架，数据预处理将原始流量文件进行划分，主要是数据平衡技术的选择，提取出特征向量，划分为训练、验证和测试集。通常使用第一阶段获取的特征和标签进行训练和修正，并得到收敛后的模型。

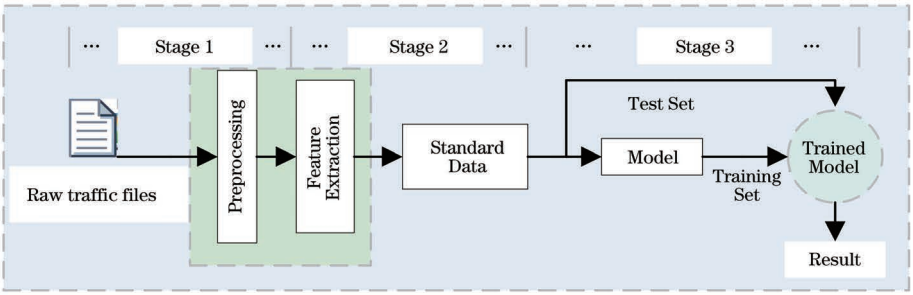

图 1 网络流量分析的基本流程

Fig. 1 Basic flow of network traffic analysis

1.1 数据采集工具

流量应用分类的数据源一般通过局域网进行流量收集。选择 Wireshark 和 Tcpdump 等其他网络监控设备的工具，使用网络协议分析工具获取数据包。

Wireshark 通过设置过滤条件可以截取各种网络数据包。命令行工具 editcap 主要用于编辑和处理捕获的 PCAP 文件。图 2 是 Wireshark 基于虚拟网卡进行网络流量数据采集的基本原理。

Linux 的命令行工具 Tcpdump 适合脚本化和

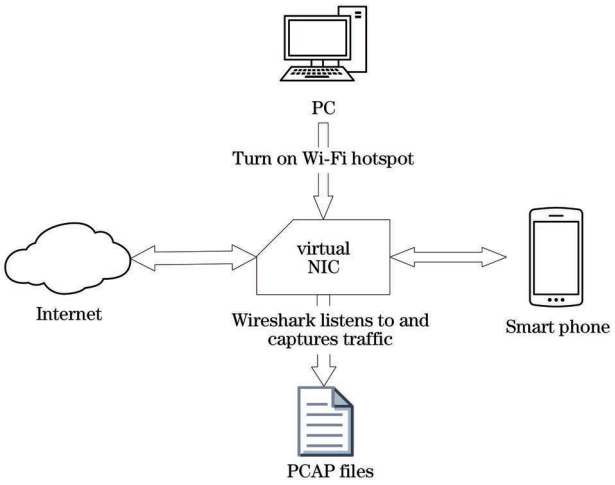

图 2 Wireshark 基于虚拟网卡的数据采集

Fig. 2 Wireshark collects data based on virtual NICs

自动化，特点是小巧快速，支持大规模数据包捕获，也允许使用自定义过滤规则，图 3 是具体的工作流程。

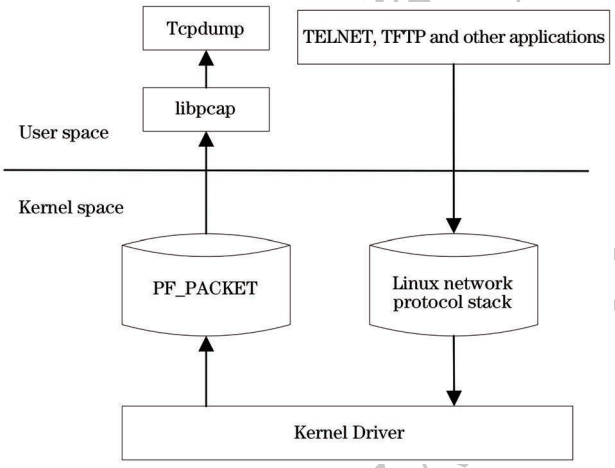

图 3 Linux 中 Tcpdump 工作流程

Fig. 3 Tcpdump workflow in Linux

1.2 数据平衡技术

对基本数据进行预处理，让数据适应模型的需求，从而实现分类任务，提高模型训练的性能和有效性[18]。针对各种处理方式，主要有数据清洗、数据过滤、数据平衡技术[19-22]、IP 掩盖[23]、数值标准化和数值归一化处理[24-25]等步骤。

文献[19]研究了多种解决数据集不平衡的方法。文献[20]使用合成少数群体过度采样技术创建了合成样本，人为地增加了少数群体类别，还将合成少数类过采样技术(SMOTE)与随机欠采样相结合，增强泛化能力，避免过拟合。

文献[21]提出的 PacketCGAN 基于 CGAN，应用流量的类型被作为条件信息添加到输入中，可以控制样本生成的模式。实验结果表明，增强的平衡数据集分类取得了更好的性能。文献[22]引入了双向生成对抗网络(BiGAN)，该方法通过同时学习数据的潜在空间表示，不仅使模型能更有效地捕获数据的本质分布，还促使生成器不断优化，最终生成了质量更高、更逼真的合成样本。

1.3 特征提取

文献[26]根据特征提取粒度分为数据包级、网络流级和会话级 3 种数据处理方式。文献[27]指出流级和会话级特征，更能捕捉长期行为模式，适用于传输层安全性协议(TLS)加密流量的分类，但在动态网络环境下适应性较差。

特征提取主要的技术有 DL 和新兴的预训练模型技术。表 1 是两种技术进行特征提取的对比。基于 DL 的特征提取通过卷积层提取空间特征[28-30]，用长短期记忆(LSTM)网络及其改进模型提取时序特征，充分利用样本数据的依赖关系。基于预训练模型的自动特征提取能够显著提升泛化能力，文献[31]基于 Transformer 架构对大规模未标注流量数据进行预训练，采用预训练-微调的两阶段模型。文献[32]通过 FL 实现分布式特征学习，核心点是边缘-云端协同和差分隐私，并提出 FL 框架 FLCNN。

表 1 预训练模型和传统方法特征提取的比较

Table 1 Comparison of feature extraction between pre-trained model and traditional methods

Dimension	Traditional Feature Engineering	Automatic Extraction of Pre-trained Models
Data dependency	Manual feature selection required	End-to-end learning from raw data
Generalizability	Limited to specific protocols and scenarios	Cross-protocol scenarios apply
Computational cost	Low	High (GPU cluster pre-training)
Privacy	Data needs to be centralized	Support federal learning

2 相关公共数据集

流量分析可选择的数据源非常丰富，这些公共数据集不仅丰富了研究的实验基础，还确保了研究成果的可重复性和可信性。表 2 是对各个数据集进行的对比和总结，在目前流量分析领域使用较多的数据集是 NSL-KDD[33-36]、USTC-TFC 2016[37-38]、ISCX 2016[39]、Moore[40]、MIRAGE-2019[41]、TON-IoT 2018[42]。公共数据集不仅为研究提供了基础数据，还为算法的评估与性能对比提供了保证。

MIRAGE-2019 数据集[41]专注于安卓应用的网络流量特征，为移动流量分类、建模、隐私保护等研究提供了标准化的基准数据。数据集使用 3 种安卓设备，覆盖 40 个应用，覆盖加密与非加密通信，其中 TLS 流量占比高。数据集只分享了前 30 个 IP 数据包的长度，因此不适合做包长序列的模型，适合做以载荷为主要输入的模型。

TON-IoT 2018 数据集[42]模拟真实工业网络环境，首次发布于 2018 年，2021 年更新包含更丰富的攻击类型和异构数据源。数据集支持从原始数据到预处理特征的全流程研究，包含原始数据、处理后数据、训练测试数据、特征描述、攻击标签共 5 个子目录，但恶意样本占比低(8%)，边缘计算场景覆盖不足。

CIC-IDS 2018 数据集[43]支持入侵检测系统研发与评估。每个样本标注为正常或攻击类型，支持监督学习任务，但是部分特征存在高维度相关性，可通过降维提升模型效率。

CICIoV 2024 数据集[44]主要的研究目标是提出一个现实的支持为车联网开发的基准数据集。通过对福特汽车进行攻击获取这一数据集，以 CSV 格式存储数据的字节、标签和类别，便于处理。但是，由于是直接收集没有处理的数据包，在文件中有很多重复的数据包，因此后续数据集需要根据任务进行整理和优化。

UNIBS 数据集[45]是基于意大利都灵理工大学研究小组在其校园网络边界路由器上捕获的流量数据构建的。由于数据集的协议特征比较单一，因此应用比较受限，适用于运营商级别的流量分类。

文献[37]为了应对 KDD 99 和 NSL-KDD 带来的挑战，开发了 USTC-TFC 2016 数据集。该数据集的第一部分从真实网络环境中收集而来，适用于实时流量分析与研究。文献[38]包含数据集的第二部分，是模拟设备收集的涵盖了 8 类常见应用的 10 种正常流量。此外，它还提供了开放源代码和处理工具，供研究人员在分析中使用。

ISCX 2016 数据集[39]包含 Tor 和非 Tor 流量，覆盖了多种应用场景，数据采集使用 Wireshark 和 Tcpdump，帮助研究人员更深入地了解 Tor 流量的特征和行为模式，提高检测此类流量的准确性。但是，数据集依赖固定 Tor 节点 IP 捕获流量，未覆盖 Tor 网络的动态节点更新机制。Tor 节点动态地更换 IP，但数据集标签基于静态黑名单，导致标签滞后，无法检测 2016 年后采用动态路由的 Tor 通信。

表 2 流量分析的公共数据集对比

Table 2 Comparison of public datasets for traffic analysis

Dataset	Time	Data Distribution and Limitations	Application Scenarios
MIRAGE-2019[41]	2019	Coverage of 40 Android Apps, 4 606 PCAP tracking files, IP addresses are private and user privacy is ethical	Mobile application categorization, privacy
TON-IoT 2018[42]	2018	Simulated heterogeneous IoT devices with multi-protocol traffic containing 12 types of attacks; 8% of malicious samples, insufficient coverage of edge computing scenarios	IoT intrusion detection, privacy protection
CIC-IDS 2018[43]	2018	Enterprise-class network topology containing 7 categories of modern attacks, 80+ dimensional traffic characterization; 85% normal traffic, unbalanced attack categories	Encrypted traffic detection, intrusion detection
CICIoV 2024[44]	2024	Direct data collection not processed, large number of duplicate packets in CSV files need to be reorganized	Telematics security, intrusion detection
UNIBS[45]	2016	Operator network traffic, containing multiple attacks; geographic limitations, single protocol characteristics (IPv4-based)	Malicious traffic detection, carrier grade traffic classification
NSL-KDD[33-36]	2009	Classic intrusion detection dataset, improved based on KDD99, contains 4 types of attacks; old attack types, insufficient support for encrypted traffic	Intrusion detection system
USTC-TFC 2016[37-38]	2016	Distinguish between encrypted and non-encrypted traffic, and label 10 types of malicious traffic; encrypted traffic feature extraction relies on port number, which is vulnerable to port masquerade attacks	Malicious traffic detection, encrypted protocol traffic classification
ISCX 2016[39]	2016	Includes Tor, non-Tor and other anonymous traffic, combined with time-series features; Tor node IP dynamic changes, labeling is easily outdated	Network security (intrusion detection, traffic classification)
Moore[40]	2017	Home network scenarios, labeling parameters (such as latency, packet loss rate); single device type (mainly consumer-grade IoT, no industrial devices)	Home network traffic analysis, quality of service monitoring

3 基于 ML 的网络流量分析

ML 通常分为传统 ML 和 DL。传统 ML 从数据中提取有意义的特征来进行建模，适用于结构化数据，需要人工设计特征。DL 可以自动从数据中学习特征，在图像处理和自然语言处理等任务上表现优异。在大规模数据和复杂问题中，通过多层网络自动学习数据中的复杂模式。

3.1 基于传统 ML 的网络流量分析方法

传统 ML 方法在网络流量分析中有着广泛的应用，表 3 是流量分析领域常用的算法模型。在使用传统 ML 方法时，应特别注意对原始流量数据的处理。

传统 ML 方法从数据中提取有意义的特征来进行建模，图 4 是对传统 ML 的分类，分为有监督 ML 方法和无监督 ML 方法。有监督学习方法常见的算法包括朴素贝叶斯[46-47]、隐马尔可夫模型[48]、k 近邻[49-51]和支持向量机(SVM)[52-53]、逻辑回归等。无监督学习方法常见的算法包括聚类算法(如 k-means)[54-58]、高斯混合模型[15]、关联规则挖掘和主成分分析(PCA)[59]等。有监督学习能够通过标注数据进行模型训练，适合需要明确分类的应用场景，而无监督学习则在标签缺失的情况下发挥重要作用，可以发现数据内在的结构和模式。未来的研究可以结合两者的优势，形成更为综合和高效的分析方法。

3.1.1 有监督学习

有监督 ML 方法的目标是通过已知的输入-输出对来训练一个模型，使其能够对新的输入数据作出准确的预测或分类。

2021，GUARINO 等[48]结合马尔可夫链(MC)和隐马尔可夫模型(HMM)，在不同情况下采用不同的混合阶马尔可夫模型进行流量分析相关的工作。

SVM 本质上是作为二元分类技术而设计的，通常用于模式识别，不注重局部优化，可以通过正确选择的决策函数提供最佳统计分类。2021 年，DONG[52]提出了 SVM 算法 CMSVM，该算法引入每类流量的权重值和主动学习来解决不平衡问题，对于比重较大的网络流量，可以达到 0.94 的真正例(TP)值，评价指标优于其他改进模型。2023 年，RAMRAJ 等[53]在对比了 XGBoost 和随机森林之后得出 SVM 与自动编码器(AE)结合使用时效果更优。

2023 年，CHEN 等[60]对加密流量通过协议类型进行分类后发现 XGBoost 和随机森林更能威胁密集流。2025 年，JISI 等[61]引入了一种具有主动学习成本敏感型 XGBoost(AL-CSXGB)算法，该算法通过不确定性抽样，对 MOORE_SET 中占比 5%的异常流量的检测率从 75%提升至 91%，在 ISCX 数据集中将 VPN 流量误判为正常流量的概率从 12%降至 3.2%，显著提升了 SD-IoT 环境下不平衡流量的分类性能，在实际中很多应用程序中的流量是加密的，该文对数据包进行深入分析，包括对其报头和内容进行分析，从而实现对网络流量特征的分析，这对于未来应用级别的分类任务有很好的启示作用。

表 3 传统 ML 流量分析方法

Table 3 Traditional ML traffic analysis methods

Methodology	Literature	Description and Characterization	Limitations
Supervised	Reference [48]	Different mixed-order Markov models that group applications based on functionality allow for real-time streaming classification	Possible overfitting of specific video protocols due to different functions of default applications
	Reference [52]	Introducing weighting values and active learning for each type of traffic, better results for traffic with a large ratio	Poor results for some specific flows, insufficient generalization
	Reference [53]	SVM is better when used in combination with autoencoders	Fewer types of recognition and limited dataset
	Reference [60]	Active learning cost-sensitive improved XGBoost algorithm with significant improvement in anomalous traffic detection rate	Large training resources and real-time limitations
	Reference [61]	In-depth analysis of packets with XGBoost and random forest algorithms	Decision trees and iterations create resource dependency
Unsupervised	Reference [54]	Improvement of online k-means algorithm from four aspects to improve generalization ability, algorithm parallelism	Performance improvement is not obvious, large computational resources
	Reference [55]	Based on X-means clustering and label propagation algorithm	Sensitive to noise, dependent on feature quality, feature selection method limits classification accuracy
	Reference [57]	Clustering methods construct clusters of major features, then similarity construction, clustering first and then classification	Clustering first has uncertainty
	Reference [58]	k-means performs secondary classification on suspicious servers after classification, and clustering is more accurate after classification	High time cost for subsequent classification
	Reference [59]	Combining information gain and principal component analysis to improve detection rate through data dimensionality reduction	The detection rate of various attack methods is not obvious

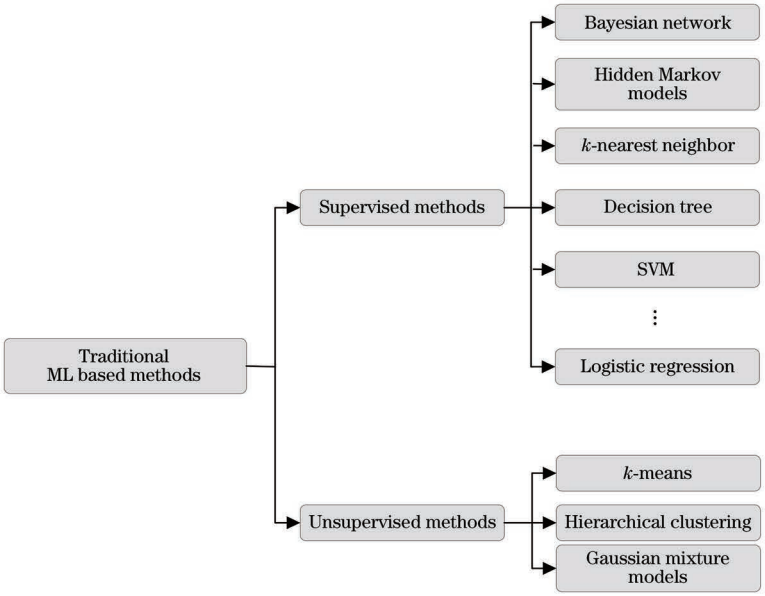

图 4 基于传统 ML 的网络流量分析方法

Fig. 4 Network traffic analysis methods based on traditional ML

3.1.2 无监督学习

传统的 k-means 算法对于样本数较多的数据集，计算成本较大，迭代速度较慢。许家钰[54]从 4 个方面改进在线 k-means 算法：与随机梯度下降相结合，达到小样本更新算法的效果；更新模型参数，选择合适的学习率方向，可以有效提高算法的收敛速度；提高算法的泛化能力，避免过拟合；算法并行化，使得每次迭代不依赖。

文献[55]提出了一种基于 X-means[56]聚类和标签传播算法的新型半监督的网络流量分类方法。该方法的特征选择在标签传播之前应用，标签传播过程在聚类后进行，在数据集上的测试准确率为 0.95。

文献[57]使用聚类框架先将网络会话聚合到几个有主要特征的集群中，减少网络差异，再进行集群之间的相似性图构建。

文献[58]先使用 DL 再用 k-means 分类网络流量数据，进行分类之后输出一个概率值，通过阈值划分新样本，然后使用 k-means 对新数据进行二次分类，以帮助用户识别具体的服务器类型。

文献[59]结合信息增益和 PCA 方法，最大程度地去除冗余信息并提取主要信息，比其他 PCA 的改进方法准确率高 0.3 左右，但是该方法没有提升攻击检测率，只是通过数据降维确定合适的维度。

3.2 基于 DL 的网络流量分析方法

DL 能够适应更高维的数据学习和预测要求，效率更高，可以根据问题自动训练建模，在流量分析方面有很好的应用前景。目前，采用的主要技术模型是卷积神经网络(CNN)、循环神经网络(RNN)及其变体、图神经网络(GNN)、生成对抗网络(GAN)和 AE 以及这几种方法的结合模型。表 4 是几种常用的方法，主流技术可以分为基于图像特征的方法、基于序列特征的方法、基于图特征和 GNN 的方法。

3.2.1 基于图像特征的方法

基于图像特征的方法利用 CNN 通过卷积层自动提取局部特征，适合图像化流量分类，但是容易出现深层网络退化问题[62]，而残差网络(ResNet)模型可以通过残差连接解决深层网络退化问题，提升分类精度。

WANG 等[37]提出了一种基于 CNN 的恶意软件流量分类方法。该方法直接将原始流量数据作为图像输入进行表示学习。实验确定了最佳的流量表示类型为包含所有层的会话表示，平均准确率达到了 99.41%。但是，CNN 模型通常需要大量的计算资源进行训练和推理，在一些资源受限的环境中会限制这个方法的应用。

LOTFOLLAHI 等[63]研究并提出 Deep Packet，该方法结合一维 CNN 提取载荷字节序列的局部空间依赖特征，捕捉协议或应用的独特模式，使用堆叠自动编码器(SAE)对流量进行无监督预训练，生成低维表征，提升分类鲁棒性，在实际复现中，融合 SAE 进行降维容易出现过拟合现象，这可能是模型在训练数据上学习了过多的噪声和细节，导致在测试数据上的性能下降。

YANG 等[64]提出了混合方法，该方法研究使用了包含数百万流量样本和数千个应用标签的华为

客户部署数据集，采用一维 CNN 和 XGBoost 模型进行已知流量的分类，在应用程序分类任务中，CNN 作为分类模型实现了 99.41%的平均准确率。该方法结合 XGBoost 模型，但由于仍依赖人工特征工程，因此处理复杂特征数据集时可能存在性能瓶颈。

WANG 等[65]结合 ResNet 的特征提取优势和 SAE 的统计特征编码优势提出 CENTIME 模型实现分类。该方法去掉了池化层，更有效地利用原始流量信息。在加密流量分类方面，该方法达到了 0.99 的 Fl 值，尽管解决了梯度消失问题，但是对标注数据还是有一定的依赖性。

魏明军等[66]提出 AQVAE-RGSNet 方法，该方法使用 ResNet 与改进的分段残差神经网络对输入的样本数据进行联合学习并预测其攻击类型。在保证最佳训练效果的前提下，该方法能够更有效地识别车辆网络中的攻击类别，对于具体协议和应用程序等细粒度识别还存在局限性。

王光明等[8]进行图像转换，改进深度卷积生成对抗网络（DCGAN）结构，解决了模型对抗概念漂移能力差等问题。该方法应用堆叠去噪自动编码器（SDAE）进行特征降维，Tor 流量的在线分类准确率可达 95.7%。在实际流量监管中，流量特征更加复杂，该方法针对连续多标签的流量分割和识别需要继续探索。

GAN 一般结合 CNN、LSTM 和 AE 等方法进行流量分析。WANG 等[21]提出的 PacketCGAN 使用条件生成对抗网络（CGAN）[67]，并将应用程序类型的输入作为条件，从而实现数据均衡；使用多层感知机（MLP）、CNN 和 SAE 进行流量分析改进。文献[68]通过引入条件约束和新的距离度量来构建 GraphCWGAN-GP 模型，采用经典的微型网络卷积核 NIN-CNN[69]作为分类器。

WANG 等[32]提出结合 FL 和 DL 的 FLCNN，还设计了一种基于边缘设备的数据采集方法，结合边缘计算解决数据孤岛问题。该模型通过使用半监督模型来实现数据采集和标记，从而降低数据依赖性并提高准确性，在分类上有很好的效果，是一个有潜力的挖掘模型。

3.2.2 基于序列特征的方法

基于序列特征的方法通常是利用 RNN、LSTM 或门控循环单元（GRU）将流量数据视为时间序列或文本序列，能够有效处理变长序列数据，适应不同长度的输入。传统 RNN 在处理长序列数据时容易遭遇梯度消失和梯度爆炸的问题，LSTM 和 GRU 引入门控机制，提高了对长序列的建模能力。对于多维流量特征的融合和表示，合理设计有效的输入结构和注意力机制是当前研究的热点。

SHAMELI 等[70]的主要目标是检测 5G 软件定义网络环境中的安全漏洞。该方法使用优化的 GAN 模型进行预测攻击，用 GRU 进行训练分类，与现有入侵检测方法相比准确率显著提高，并且检测时间缩短。

为了自适应地为网络选择超参数，ZENDEHDEL 等[71]提出基于 CNN 增强 IoT 入侵检测的 MDNET 神经网络。该网络通过模糊粗糙集理论删除受噪声影响严重的要素，能够提高整体效率，与现有模型相比，有较好的分类准确性和稳定性。

LIANG 等[72]将 CGAN、CNN 和双向 LSTM（BiLSTM）相结合，该方法具有较强的鲁棒性。MO 等[73]结合 CNN、TCN 和 GRU 的优势，对网络流量进行分类，其中，TCN 用于进一步提取数据的高频和低频信息，GRU 提取长期依赖关系，该方法使用手动方式确定模型参数，这虽然有一定的局限性，但可以引入优化算法来选择更好的模型参数。

陈思宇等[74]提出一种基于注意力机制的 CNN 和双向 GRU（BiGRU）的加密流量分类和应用识别方法 CNN-AttBiGRU，该方法对流量所使用的应用程序分类的准确率达 99.8%，能够提升跨协议场景的泛化能力。

3.2.3 基于 GNN 的方法

基于 GNN 的方法通过构建流量关系图，能够直观地表示流量中各个实体之间的相互关系。图结构使得模型能够捕获节点间的依赖性，从而更好地理解和分析流量的整体行为。图卷积网络（GCN）[75]通过局部谱滤波或空间域卷积实现图结构数据的特征提取，通过分层卷积捕捉局部与全局图模式，避免传统 CNN 和 RNN 的局限性，但是也容易出现过平滑和动态图适应性差的问题。

FENG 等[76]针对 Android 恶意软件提出混合特征融合和图嵌入技术（Node2Vec）的 HGDetector 方法。当网络流量特征有效捕获应用的行为时，混合特征的融合可以将恶意软件检测准确率提高约 4%。

XU 等[77]提出基于 AE 和深度 GCN 的 ADGCN 方法，并将其用于小样本学习。在图形构建过程中，该方法需要从较长流量训练样本中学习抽象特征，因此训练依赖于较长流量，对流量数据集要求高。

LIU 等[78]提出的 TB-Graph 模型使用关系图注意力（RGAT）网络。从突发图节点和边缘关系中提取潜在的行为特征，从而有效捕捉流量中的依赖

性和上下文信息。TB-Graph 在两个公共数据集上的细粒度加密恶意流量分类任务中优于各种先进的方法,在多维特征融合和上下文理解等方面表现优异,能够提高准确率和减少误分类。该方法对输入数据的微小扰动敏感,恶意用户可能通过对抗样本攻击来干扰流量分类的效果,具有安全隐患。

ZHAO 等[79]提出基于 CNN 的不平衡流量检测模型 LGSMOTE-IDS,将基于线性图的加权距离 SMOTE 引入入侵检测系统。LGSMOTE-IDS 是一个将 GNN 与过采样算法集成的框架,以解决网络入侵检测系统(NIDS)数据集中的类不平衡问题。

ZHANG 等[80]提出 TFE-GNN 模型,该模型构

表 4 基于 DL 的流量分析方法

Table 4 DL based traffic analysis methods

Methodology	Architecture	Description and Characterization	Limitations
Image	1D-CNN[37]	Strong local feature extraction capability, suitable for application classification	Lack of long sequence modeling capability, relying on large-scale labeled data
	CNN, SAE[63]	Combination of spatial features and deep feature dimensionality reduction	High complexity of model fusion
	1D-CNN, XGBoost[64]	Validated in Huawei's commercial dataset, close to real-world application scenarios	Relying on manual feature engineering, XGBoost may be a performance bottleneck
	ResNet, AutoEncoder[65]	Residual structure mitigates gradient vanishing self-encoder supports unsupervised pre-training	Still has some dependence on labeled data
	ResNet, WGAN[66]	Learning on input sample data and predicting its attack type	Fine-grained recognition still has limitations
	SDAE, DCGAN[8]	Multi-dimensional fusion for data enhancement to improve data quality and SDAE for feature dimensionality reduction	Traffic segmentation and recognition for consecutive multiple labels needs to be improved
	CNN, SAE, CGAN[21]	Advantages of CGAN generate specified samples and condition the inputs of application types for data equalization	Problems of training instability and pattern collapse
	NIN-CNN, CGAN[68]	Introduces conditional constraints and a new distance metric to construct, classification models use miniature network convolution kernels	Fewer recognizable classes
	FLCNN[32]	Protects data privacy and supports cross-organizational co-training	Sensitive to non-independently and identically distributed data, performance may be degraded
Sequence	GAN、GRU[70]	For 5G security vulnerabilities, accuracy is significantly improved and detection time is shortened	More favorable to network security
	CNN[71]	Remove features that are seriously affected by noise to improve efficiency	Not suitable for dynamic traffic analysis scenarios, can consider gray set theory
	CGAN, CNN, BiLSTM[72]	Extract key features based on spatial and temporal information, robustness	Accuracy needs to be improved
	CNN, TCN, GRU[73]	TCN further extracts high-frequency and low-frequency information of data, and GRU extracts long-term dependencies	Manual way to determine model parameters, with some limitations
	CNN, BiGRU, Attention[74]	Fusion of local features, screening of key frames and long time series modeling	Large number of model parameters, poor real-time performance, only suitable for cloud deployment
Graph	Node2Vec[76]	Supplementary function call graph features are effective in improving results when features are insufficient	Not as good as graph representation models in representing overall graph structure relationships
	ADGCN, AE[77]	Traffic classification on small sample datasets, learning abstract features	Training relies on long traffic flows, demanding traffic datasets
	FA-NET, RGAT[78]	Extracts potential behavioral features from bursty graph nodes and edge relationships, effective for malicious flows	Sensitive to small perturbations in input data, with security risks
	GNN, SMOTE[79]	A framework that integrates GNN with oversampling algorithms to address class imbalance in Network Intrusion Detection System (NIDS) datasets	Improvement is needed for extreme minority data identification
	GNN[80]	Embedded in the header and load bytes of packets, stronger feature representation	Inability to adapt to dynamic new network conditions in time

建字节级流量图，捕捉细粒度依赖关系，包含双嵌入层、基于 GNN 的流量图编码器以及交叉门控的特征融合机制，能够分别嵌入数据包的头部和载荷字节，然后将它们融合以获得更强的特征表示。在两个真实数据集上，TFE-GNN 在细粒度加密流量分类任务中的表现超过了多种先进的方法。TFE-GNN 依赖于静态构建的字节级流量图，而在动态的网络环境中，流量模式和拓扑结构可能随时变化，模型可能无法及时适应新的网络条件。

3.2.4 从 ML 到预训练模型的发展

传统 ML 可以解决传统端口和有效载荷难题，但存在专业知识依赖。DL 自动特征学习能够准确分类和识别网络流量数据中的异常行为或潜在威胁，但其依赖大规模标注数据，泛化能力有限。在跨协议场景中，模型需针对每种协议单独训练，导致资源消耗大和部署成本高。

当前 ML 流量分析技术仍有许多局限性。首先，在资源受限条件下，如何在网络边缘设备和弱计算能力的终端上实现分类功能。其次，在训练和推理阶段，计算资源等成本与分类性能如何平衡。最后，如何防范数据集的敏感数据泄露问题。

相较而言，预训练模型（如 Transformer 架构和 BERT 模型）通过引入预训练-微调范式和自注意力机制，实现了显著的技术突破，并且能够做到隐私合规。表 5 展示了 DL 与预训练模型之间在特征提取、数据效率和跨协议场景中应用的差异。

表 5 基于 DL 和预训练模型的网络流量分析技术对比

Table 5 Comparison of DL and pre-trained model based network traffic analysis techniques

Dimension	Deep Learning	Pre-trained Models
Feature extraction	Local features	Global dependencies
Data samples	10^4-level	10^2-level
Cross-protocol	Protocol-specific classification with poor cross-protocol adaptation	Unknown protocol recognition for cross-domain migration

预训练模型具备了少样本学习的能力，在跨领域迁移方面表现突出。此外，自注意力机制使得模型能够捕获长序列依赖，适用于复杂的流量模式，在流量分析领域有很大的研究意义。

4 基于预训练模型的网络流量分析方法

当前基于预训练模型的流量分析方法主要分为 3 类，包括预训练模型优化、多模态特征融合以及轻量化部署，表 6 是对基于预训练模型的流量分析技术的整理总结。

4.1 预训练模型优化及实验分析

通过实施从通用到领域自适应的预训练模型优化，利用大规模无标注数据学习通用流量表征，从而缓解了标注数据稀缺对模型性能的影响问题。Transformer 通过自注意力机制建模长序列依赖，基于 Transformer 的 BERT 预训练和微调模型显著提升了加密流量的多任务分类性能。

PERT[81] 是基于 Transformer 的载荷编码表示，处理加密流量的载荷字节序列，其被应用于载荷分析，通过自注意力机制学习载荷的全局上下文表示，引入加密后仍保留的部分统计特征进行分析，但是实际加密场景中这些特征的有效性有限。

LIN 等[31] 提出一种新的流量分类模型 ET-BERT，该模型利用 Transformer 实现加密流量双向编码。在加密流量分类任务中，ET-BERT 通过微调实现跨协议特征迁移，是从大规模未标记数据中预训练的数据包级模型，具有很强的泛化能力。实验结果表明，在通用加密应用分类、加密恶意软件分类、VPN 加密流量分类、Tor 加密应用分类和 TLS1.3 加密应用分类等任务中，ET-BERT 的性能达到了国际一流水平。

YU 等[82] 提出了一种结合 BERT 模型和流量包头部信息的加密流量分类方法。该方法通过 BERT 对数据包头部序列进行上下文建模，提取深层语义特征，有效捕捉头部字段的序列依赖关系，对加密流量中的协议模式识别能力较强，但仅依赖头部信息可能忽略载荷中的潜在特征（如时序或长度信息）。

表 7 是预训练模型在 ISCX 数据集上按照服务类型分类的实验结果。实验结果表明，BERT 和头部信息的精确率高于其他方法，将头部字段信息用于流量分析的结果显著优于流级和元数据。对比流级别和数据包级别的分类证明了在数据包级别上进行细粒度流量分析更为重要。当前研究结果证明了流量特征注入和特征层次深化的有效性，提供了未来模型优化的发展方向。

表 6 预训练模型流量分析方法

Table 6 Pre-trained model flow analysis methods

Methodology	Architecture	Description and Characterization	Limitations
Pre-trained	PERT[81]	Fine-grained feature extraction, suitable for closed environment deep analysis	High computational complexity, limited feature effectiveness in real-world scenarios
	ET-BERT[31]	Strong generalization ability, context-aware, suitable for small sample or large-scale cross-platform classification	High pre-training cost, insufficient dynamic adaptability, poor real-time performance
	BERT, Head[82]	Low computational overhead, privacy compliance, suitable for real-time monitoring and resource-constrained environments	Incomplete information, context sensitive
Combining DL	BERT, CNN[83]	Robust, complementary features, suitable for high-precision offline classification	Complex models, high training cost
	BERT, BiLSTM[84]	Bidirectional time-series modeling, complex pattern adaptation, suitable for VPN/Tor traffic analysis	High latency, difficult to deploy edge devices
	HGNN, LLM[85]	Gain contextual, interpretable and actionable threat insights, better experimental results	Effectiveness on other data remains to be evaluated
	GNN, BERT[86]	Self-attentive mechanism fuses interaction patterns and load information multimodal approach	Ability to analyze encrypted traffic needs to be improved
Lightweight	Distilled BERT[87]	Lightweight for edge deployments and resource-constrained real-time scenarios	Loss of accuracy and performance degradation
	ViT, Meta-Learning[88-89]	Global dependency modeling with dynamic update support for graphical traffic analysis	Timing information may be lost

表 7 预训练模型在 ISCX 数据集上的实验结果对比

Table 7 Comparison of experimental results of pre-trained model on ISCX dataset

Model	Precision	Recall	F1 value
PERT	0.932 7	0.932 2	0.932 3
ET-BERT-flow	0.975 6	0.973 1	0.973 3
ET-BERT-packet	0.989 1	0.989 0	0.989 0
BERT Header	0.992 6	0.992 4	0.992 4

4.2 预训练模型与 DL 的融合

多模态特征融合通过融合 DL 的 CNN 和 LSTM 等模型，实现全局与局部协同建模。融合模型可以利用头部序列、载荷统计以及时序模式等流量多维度特征，提升分类鲁棒性。

SHI 等[83]提出的 BFCN 结合 BERT 和 CNN。BERT 提取数据包头部和载荷的全局语义特征，CNN 捕获局部空间特征，全局与局部特征提升了对加密流量细微模式的敏感性。在精确率、召回率、F1 值 3 个指标中，BFCN 对比 ET-BERT 分别提高了 0.22%、0.21%和 0.21%，但是 BFCN 的主要目的是对加密流量进行分类，它没有考虑恶意流量的特征。

MA 等[84]提出的 Bi-ETC 是 BERT 和 BiLSTM 的融合。该方法利用 BERT 提取单包特征，使用 BiLSTM 建模流量会话的双向时序依赖。Bi-ETC 在 ISCX-VPN 数据集上的 F1 值和准确率分别为 99.43%和 99.7%。尽管双向时序建模可以显著提升视频流与即时通信等的长会话流量的分类效果，但由于 BiLSTM 的序列处理效率较低并且模型参数量大，难以满足高吞吐场景，边缘设备部署困难。

FARRUKH 等[85]提出集成大语言模型(LLM)和异构神经网络的入侵检测框架 XG-NID。该框架集成 LLM，可以获得上下文、可解释的威胁洞察，通过实验中的定量比较分析，在多分类任务中取得 97%的 F1 值，优于基线和先进模型。但是，该框架在其他数据上的有效性还有待进一步评估。

LU 等[86]提出一种结合 GNN 和 BERT 的多模态分类方法。该方法通过自注意力机制融合交互模式和负载信息，对去中心化应用程序进行流量分类。在构建的数据集和公共数据集上进行了测试，在准确率和鲁棒性方面均优于单模方法和其他先进技术。

4.3 预训练模型轻量化

轻量化部署模型是从云端到边缘的落地实践。通过模型压缩和硬件加速等技术，实现预训练模型在资源受限场景中的实时推理。预训练模型从轻量化应用到流量分类领域的主要模型压缩技术有替换融合模型为轻量化的模块、蒸馏 BERT 模型以及快速微调模型等。轻量化可能牺牲模型表达能力，对小众流量分类性能下降显著，所以在轻量化过程中，需要平衡模型性能和参数的量级。

NetKD[87]是一种基于知识蒸馏的轻量化模型，NetKD 将大型预训练语言模型(如 BERT)压缩为轻量级学生模型，减少资源消耗。蒸馏 BERT 到轻量级 LSTM，模型体积减少了 80%。在保持较高分类精度的同时大幅降低计算和存储开销，适合边缘设备或实时系统，但蒸馏过程可能导致特征信息损失，对小型流量类型的分类性能下降明显。

视觉 Transformer(ViT)[88]是将图像分割为一个个的块(Patch)序列，利用自注意力机制建模全局依赖，不仅考虑 Patch 之间的信息，还考虑每个 Patch 的内部信息。基于视觉 Transformer 与元学习的实时分类框架 ConViTML[89]利用元学习的快速微调，满足边缘设备动态需求。该方法将流量数据转换为字节矩阵的图像形式，利用 ViT 提取空间特征，并结合元学习实现快速模型自适应。元学习支持动态环境下的快速模型更新，但图像转换过程会丢失部分时序信息，未来如何保证动态更新的同时平衡时序信息的获取也是一个方向。

5 未来展望

5.1 基于 FL 的隐私保护

FL 为流量分析中的数据隐私保护和跨机构协作提供了新思路。当前研究(FLCNN[32]、FedAvg[90]、FL[91])已初步验证了 FL 在边缘设备间协同训练轻量化模型的可行性，但仍面临模型异构性、非独立同分布(Non-IID)数据适配性不足等挑战。

现有的解决方案是在边缘设备上完成流量特征工程，仅上传模型参数，避免包含敏感隐私的原始流量泄露。文献[92]提出 FLIC 框架，在智能手机端对 HTTP/TLS 流量进行本地训练，利用联邦服务器聚合模型更新，在 non-IID 数据下，应用分类准确率达 88%，且未泄露用户隐私。

未来可探索联邦聚合策略，形成“联邦大模型”范式，结合差分隐私[93]与同态加密技术，在确保数据不可逆推的同时提升全局模型泛化能力。在华为 5G 案例中[94]，non-IID 数据下平均绝对误差(MAE)较 FedAvg 降低 18%，设备掉线率从 25%降至 9%。在梯度聚合时添加拉普拉斯噪声进行差分隐私增强。同态加密是对传输的模型参数进行加密处理，确保中间结果不可解密。

5.2 未知流量识别

由于当前新型加密协议和自定义协议的快速涌现，因此未知流量识别成为流量分析的核心难点。现有方法通过预训练大模型捕捉流量上下文语义，但面对动态协议伪装和零样本场景仍显不足。当前研究人员在加密环境下的恶意流量检测领域已取得了一定的进展，但这些研究大多聚焦于特定类型的恶意攻击场景[95]，主要针对 TLS 加密的恶意软件流量或者加密的拒绝服务(DoS)攻击流量等。

未来可结合元学习构建自适应特征空间，使模型通过少量标注样本快速适应未知协议特征，同时引入主动学习技术，通过不确定性采样筛选高价值样本进行人工标注，降低标注成本。此外，需探索多模态表征学习，融合流量头部、载荷字节和行为时序特征来构建复合判别模型。针对加密流量的隐蔽性，可结合 GNN 挖掘流量实体间的拓扑关联，捕捉协议交互的潜在模式，提升对未知威胁的早期预警能力。

5.3 轻量化与边缘计算

轻量化部署是流量分析技术落地的关键瓶颈。现有模型压缩方法，如知识蒸馏，虽能降低参数量，但易导致细粒度分类性能下降。

针对计算成本的问题，VICENZI 等[96]提出结合动态剪枝与硬件虚拟层的轻量化方案，实验能耗降低了 137%。XU 等[97]提出 FastTraffic 框架，针对加密流量的轻量化分类问题，在 ISCX-VPN 数据集上实现 95.2%的分类准确率，模型参数量较传统 CNN 减少 90%。

未来可探索动态剪枝技术、硬件层和算法层面的优化。研究方向还应关注数据预处理与模型轻量化的协同优化，通过数据预处理[30,98-99]使数据更加适用于轻量化模型。在算法层面，复数神经网络[100]通过扩展特征维度提升模型容量与鲁棒性，为轻量化提供新路径。此外，深度模型本身的轻量化[101]也是一个方向。

6 结束语

网络流量分析作为网络管理与安全的核心技术，其方法演进始终与人工智能技术的突破紧密相连。本文系统梳理了从传统 ML、DL 到预训练模型的网络流量分析技术的发展脉络，揭示了不同方法在特征提取、计算效率与跨协议适应性等方面的内在关联与局限。尽管已有显著进展，但仍面临着诸多挑战。下一步将从如何平衡模型复杂度和泛化能力、识别未知流量以及隐私保护等方面进行研究和探讨，进一步优化网络流量分析技术。

参考文献

[1] LIN X J, XIONG G, GOU G P, et al. Respond to change

with constancy: instruction-tuning with LLM for non-I. I. D. network traffic classification [J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 5758-5773.
[2] SEWAK M, SAHAY S K, RATHORE H. Deep reinforcement learning in the advanced cybersecurity threat detection and protection [J]. Information Systems Frontiers, 2023, 25(2): 589-611.
[3] CARVALHO M, SOARES D, MACEDO D F. QoE estimation across different cloud gaming services using transfer learning[J]. IEEE Transactions on Network and Service Management, 2024, 21(6): 5935-5946.
[4] KUMAR R, KUMAR R, NIGAM M J. Alleviation of delay in tele-surgical operations using Markov approach-based Smith predictor [J]. International Journal of Business Analytics, 2022, 9(3): 1-14.
[5] ZHAO P H, DING Z J, WANG M M, et al. Behavior analysis for electronic commerce trading systems: a survey [J]. IEEE Access, 2019, 7: 108703-108728.
[6] HUAN S, ZHANG X Y, SHANG W L, et al. T-shaped CAN feature integration with lightweight deep learning model for in-vehicle network intrusion detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(12): 21183-21196.
[7] 梁松林, 林伟, 王珏, 等. 面向后渗透攻击行为的网络恶意流量检测研究[J]. 计算机工程, 2024, 50(5): 128-138.
LIANG S L, LIN W, WANG J, et al. Research on network malicious traffic detection for post-exploitation attack behavior[J]. Computer Engineering, 2024, 50(5): 128-138. (in Chinese)
[8] 王光明, 李冬青, 蒋从锋. 不平衡数据集下的数据中心网络流量异常检测[J]. 计算机工程, 2025, 51(8): 227-237.
WANG G M, LI D Q, JIANG C F. Network traffic anomaly detection for data centers in imbalanced datasets [J]. Computer Engineering, 2025, 51(8): 227-237. (in Chinese)
[9] PARK J T, SHIN C Y, BAEK U J, et al. User behavior detection using multi-modal signatures of encrypted network traffic[J]. IEEE Access, 2023, 11: 97353-97372.
[10] ANAMURO C V, BLANC A, LAGRANGE X. Statistical analysis and characterization of signaling and user traffic of a commercial multi-band LTE system[J]. Telecommunication Systems, 2024, 87(2): 437-453.
[11] DAINOTTI A, PESCAPE A, CLAFFY K C. Issues and future directions in traffic classification[J]. IEEE Network, 2012, 26(1): 35-40.
[12] SUN G L, XUE Y B, DONG Y F, et al. An novel hybrid method for effectively classifying encrypted traffic [C]// Proceedings of the IEEE Global Telecommunications Conference. Washington D. C., USA: IEEE Press, 2011: 1-5.
[13] VELAN P, ČERMÁK M, ČELEDA P, et al. A survey of methods for encrypted traffic classification and analysis[J]. International Journal of Network Management, 2015, 25(5): 355-374.
[14] ARNDT D J, ZINCIR-HEYWOOD A N. A Comparison of three machine learning techniques for encrypted network traffic analysis[C]//Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA). Washington D. C., USA: IEEE Press, 2011: 107-114.
[15] YAO Z J, GE J G, WU Y L, et al. Encrypted traffic classification based on Gaussian mixture models and Hidden Markov Models [J]. Journal of Network and Computer Applications, 2020, 166: 102711.
[16] WU S G, WANG H Y, WANG Y, et al. Technology analysis of network anomalous behavior detection based on machine learning[C]//Proceedings of the 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). Washington D. C., USA: IEEE Press, 2022: 730-737.
[17] SHENG C, ZHOU W, HAN Q L, et al. Network traffic fingerprinting for IIoT device identification: a survey[J]. IEEE Transactions on Industrial Informatics, 2025, 21(5): 3541-3554.
[18] YAN X G, HE L K, XU Y F, et al. High-speed encrypted traffic classification by using payload features[J]. Digital Communications and Networks, 2025, 11(2): 412-423.
[19] VU L, VAN TRA D, NGUYEN Q U. Learning from imbalanced data for encrypted traffic identification problem[C]// Proceedings of the 7th Symposium on Information and Communication Technology. New York, USA: ACM Press, 2016: 147-152.
[20] SARANYA N, HALDORAI A. Efficient intrusion detection system data preprocessing using deep sparse autoencoder with differential evolution[J]. IET Information Security, 2024(1): 9937803.
[21] WANG P, LI S H, YE F, et al. PacketCGAN: exploratory study of class imbalance for encrypted traffic classification using CGAN [C] // Proceedings of the 2020 IEEE International Conference on Communications (ICC). Washington D.C., USA: IEEE Press, 2020: 1-7.
[22] 张震, 周一成, 田鸿朋. 基于空间特征和生成对抗网络的网络入侵检测[J]. 郑州大学学报(工学版), 2024, 45(6): 40-47.
ZHANG Z, ZHOU Y C, TIAN H P. Network intrusion detection based on spatial features and generative adversarial networks[J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(6): 40-47. (in Chinese)
[23] 孙文茜. 基于流量特征的网络流量分类算法研究[D]. 南京: 南京信息工程大学, 2024.
SUN W Q. Research on network traffic classification algorithm based on traffic characteristics [D]. Nanjing: Nanjing University of Information Science & Technology, 2024. (in Chinese)
[24] BROWNLEE J. How to use power transforms for machine learning [EB/OL]. [2025-03-07]. https://machinelearningmastery.com/power-transforms-with-scikit-learn/.
[25] WEISBERG S. Yeo-Johnson power transformations [EB/OL]. [2025-03-07]. https://www.geeksforgeeks.org/machine-learning/yeo-johnson-transform/.
[26] 梅汉涛, 程光, 朱怡霖, 等. Tor 被动流量分析综述[J]. 软件学报, 2025, 36(1): 253-288.
MEI H T, CHENG G, ZHU Y L, et al. Survey on Tor passive traffic analysis [J]. Journal of Software, 2025, 36(1): 253-288. (in Chinese)
[27] 刘赟, 张位, 周阳. 基于连续上下行数据传输特征的 TLS 加密流量分类方法[J]. 通信技术, 2024, 57(9): 955-964.
LIU Y, ZHANG W, ZHOU Y. Classification of TLS encrypted traffic based on continuous forward and backward data transmission features [J]. Communications Technology, 2024, 57(9): 955-964. (in Chinese)
[28] DO N Q, SELAMAT A, LIM K C, et al. An improved ensemble deep learning model based on CNN for malicious website detection [C] // Proceedings of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Berlin, Germany: Springer, 2022: 497-504.
[29] 张光华, 王子昱, 蔡明伟. 基于不平衡数据的物联网异常流量检测[J]. 信息安全研究, 2024, 10(11): 1012-1019.
ZHANG G H, WANG Z Y, CAI M W. Abnormal traffic detection in the Internet of Things based on imbalanced data [J]. Journal of Information Security Research, 2024, 10(11): 1012-1019. (in Chinese)

[30] 赵广龙. 基于深度学习的轻量化网络流量分类方法研究[D]. 哈尔滨: 黑龙江大学, 2023.
ZHAO G L. Research on lightweight network traffic classification method based on deep learning[D]. Harbin: Helongjiang University, 2023. (in Chinese)

[31] LIN X J, XIONG G, GOU G P, et al. ET-BERT: a contextualized datagram representation with pre-training Transformers for encrypted traffic classification[C]// Proceedings of the ACM Web Conference 2022. New York, USA: ACM Press, 2022: 633-642.

[32] WANG Z X, LI Z Y, FU M Y, et al. Network traffic classification based on federated semi-supervised learning[J]. Journal of Systems Architecture, 2024, 149: 103091.

[33] REVATHI S, MALATHI A. A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection[J]. International Journal of Engineering Research & Technology, 2013, 2(12): 1848-1853.

[34] TAVALLAEE M, BAGHERI E, LU W, et al. A detailed analysis of the KDD CUP 99 data set[C]//Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications. Washington D.C., USA: IEEE Press, 2009: 1-6.

[35] CREECH G, HU J K. Generation of a new IDS test dataset: time to retire the KDD collection[C]//Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC). Washington D.C., USA: IEEE Press, 2013: 4487-4492.

[36] HADDADI F, ZINCIR-HEYWOOD A N. Data confirmation for botnet traffic analysis[C]//Proceedings of the Foundations and Practice of Security. Berlin, Germany: Springer, 2015: 329-336.

[37] WANG W, ZHU M, ZENG X W, et al. Malware traffic classification using convolutional neural network for representation learning[C]//Proceedings of the International Conference on Information Networking (ICOIN). Washington D.C., USA: IEEE Press, 2017: 712-717.

[38] CELIK Z B, WALLS R J, MCDANIEL P, et al. Malware traffic detection using tamper resistant features[C]// Proceedings of the 2015 IEEE Military Communications Conference. Washington D.C., USA: IEEE Press, 2015: 330-335.

[39] LASHKARI A H, GIL G D, MAMUN M S I, et al. Characterization of tor traffic using time based features[C]// Proceedings of the 3rd International Conference on Information Systems Security and Privacy. Porto, Portugal: Science and Technology Publications, 2017: 253-262.

[40] Moore dataset[EB/OL]. [2025-03-07]. https://gitcode.com/open-source-toolkit/8c98d.

[41] MIRAGE-2019 dataset[EB/OL]. [2025-03-07]. https://www.traffic.comics.unina.it/mirage/mirage-2019.html.

[42] The TON-IoT datasets[EB/OL]. [2025-03-07]. https://research.unsw.edu.au/projects/toniot-datasets.

[43] CSE-CIC-IDS 2018 on AWS[EB/OL]. [2025-03-07]. https://www.unb.ca/cic/datasets/ids-2018.html.

[44] NETO E C P, TASLIMASA H, DADKHAH S, et al. CICIoV 2024: advancing realistic IDS approaches against DoS and spoofing attack in IoV CAN bus[J]. Internet of Things, 2024, 26: 101209.

[45] Archived news[EB/OL]. [2025-03-07]. https://www.ing.unibs.it/ntw/tools/traces/index.php.

[46] AULD T, MOORE A W, GULL S F. Bayesian neural networks for Internet traffic classification[J]. IEEE Transactions on Neural Networks, 2007, 18(1): 223-239.

[47] TANG B, HE H B, BAGGENSTOSS P M, et al. A Bayesian classification approach using class-specific features for text categorization[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(6): 1602-1606.

[48] GUARINO I, NASCITA A, ACETO G, et al. Mobile network traffic prediction using high order Markov chains trained at multiple granularity[C]//Proceedings of the IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI). Washington D.C., USA: IEEE Press, 2021: 394-399.

[49] NGUYEN T T T, ARMITAGE G. A survey of techniques for Internet traffic classification using machine learning[J]. IEEE Communications Surveys & Tutorials, 2008, 10(4): 56-76.

[50] SHI Y, BISWAS S. Protocol-independent identification of encrypted video traffic sources using traffic analysis[C]// Proceedings of the IEEE International Conference on Communications (ICC). Washington D.C., USA: IEEE Press, 2016: 1-6.

[51] RAN D B, DVIR A, PELE O, et al. I know what you saw last minute—encrypted HTTP adaptive video streaming title classification[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(12): 3039-3049.

[52] DONG S. Multi class SVM algorithm with active learning for network traffic classification[J]. Expert Systems with Applications, 2021, 176: 114885.

[53] RAMRAJ S, USHA G. Unsupervised feature learning methodology for tree based classifier and SVM to classify encrypted traffic[J]. International Journal of Advanced Computer Science and Applications, 2023, 14(2): 1-7.

[54] 许家钰. 基于 k-means 算法的 WiFi 用户行为分析系统设计与实现[D]. 北京: 北京邮电大学, 2019.
XU J Y. Design and implementation of WiFi user behavior analysis system based on k-means algorithm[D]. Beijing: Beijing University of Posts and Telecommunications, 2019. (in Chinese)

[55] NOORBEHBAHANI F, MANSOORI S. A new semi-supervised method for network traffic classification based on X-means clustering and label propagation[C]//Proceedings of the 8th International Conference on Computer and Knowledge Engineering (ICCKE). Washington D.C., USA: IEEE Press, 2018: 120-125.

[56] PELLEG D, MOORE A W. X-means: extending k-means with efficient estimation of the number of clusters[C]// Proceedings of the International Conference on Machine Learning. Washington D.C., USA: IEEE Press, 2000: 1-10.

[57] DU Y X, HE M S, WANG X J. A clustering-based approach for classifying data streams using graph matching[J]. Journal of Big Data, 2025, 12(1): 37.

[58] LIU J G, ZHANG P Y, SUN Y M, et al. Network traffic classification method of power system based on DNN and k-means[C]//Proceedings of the International Symposium on Artificial Intelligence and Robotics. Singapore: Springer, 2022: 303-317.

[59] 王旭仁, 马慧珍, 冯安然, 等. 基于信息增益与主成分分析的网络入侵检测方法[J]. 计算机工程, 2019, 45(6): 175-180.
WANG X R, MA H Z, FENG A R, et al. Network intrusion detection method based on information gain and principal components analysis[J]. Computer Engineering, 2019, 45(6): 175-180. (in Chinese)

[60] CHEN L, WANG Q J, SONG Y Q, et al. Security is readily to interpret: quantitative feature analysis for botnet encrypted malicious traffic[C]//Proceedings of the IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). Washington D.C., USA: IEEE Press, 2023: 753-758.

[61] JISI C, ROH B H, ALI J. An effective scheme for classifying imbalanced traffic in SD-IoT, leveraging

XGBoost and active learning[J]. Computer Networks, 2025, 257: 110939.
[62] 耿丽丽, 牛保宁. 基于通道相似度熵的卷积神经网络裁剪[J]. 计算机工程, 2024, 50(7): 133-143.
GENG L L, NIU B N. Convolutional neural network pruning based on channel similarity entropy[J]. Computer Engineering, 2024, 50(7): 133-143. (in Chinese)
[63] LOTFOLLAHI M, SIAVOSHANI M J, ZADE R S H, et al. Deep Packet: a novel approach for encrypted traffic classification using deep learning[J]. Soft Computing, 2020, 24(3): 1999-2012.
[64] YANG L X, FINAMORE A, JUN F, et al. Deep learning and zero-day traffic classification: lessons learned from a commercial-grade dataset[J]. IEEE Transactions on Network and Service Management, 2021, 18(4): 4103-4118.
[65] WANG M N, ZHENG K F, NING X Y, et al. CENTIME: a direct comprehensive traffic features extraction for encrypted traffic classification[C]//Proceedings of the IEEE 6th International Conference on Computer and Communication Systems (ICCCS). Washington D. C., USA: IEEE Press, 2021: 490-498.
[66] 魏明军, 李凤, 刘亚志, 等. 基于改进 WGAN-GP 和 ResNet 的车联网入侵检测方法[J]. 郑州大学学报(工学版), 2024, 45(4): 30-37.
WEI M J, LI F, LIU Y Z, et al. An intrusion detection method for Internet of vehicles based on improved WGAN-GP and ResNet[J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(4): 30-37. (in Chinese)
[67] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. [2025-03-07]. https://arxiv.org/abs/1411.1784.
[68] ZHAI J T, LIN P, CUI Y F, et al. GraphCWGAN-GP: a novel data augmenting approach for imbalanced encrypted traffic classification[J]. Computer Modeling in Engineering & Sciences, 2023, 136(2): 2069-2092.
[69] LIN M, CHEN Q, YAN S C. Network in network[EB/OL]. [2025-03-07]. https://arxiv.org/abs/1312.4400.
[70] SHAMELI R, RAJKUMAR S. High-speed threat detection in 5G SDN with particle swarm optimizer integrated GRU-driven generative adversarial network[J]. Scientific Reports, 2025, 15: 10025.
[71] ZENDEHDEL M, DEHAKITOROGHI A, HAMIDZADEH J. MDNET: a novel neural network based on CNN and fuzzy rough set with adaptive parameters for intrusion detection in the Internet of Things[J]. International Journal of Engineering, 2025, 38(12): 2965-2993.
[72] LIANG X Y, XING H Y, HOU T H. Network intrusion detection method based on CGAN and CNN-BiLSTM[C]// Proceedings of the IEEE 16th International Conference on Electronic Measurement & Instruments (ICEMI). Washington D.C., USA: IEEE Press, 2023: 396-400.
[73] MO L N, QI X G, LIU L F. Network traffic grant classification based on 1DCNN-TCN-GRU hybrid model[J]. Applied Intelligence, 2024, 54(6): 4834-4847.
[74] 陈思宇, 马海龙, 张建华, 等. 基于自注意力的 CNN 和 Bi-GRU 加密流量分类[J]. 计算机科学, 2023, 50(8): 396-402.
CHEN S Y, MA H L, ZHANG J H. Encrypted traffic classification of CNN and Bi-GRU based on self-attention[J]. Computer Science, 2023, 50(8): 396-402. (in Chinese)
[75] BHATTI U A, TANG H, WU G L, et al. Deep learning with graph convolutional networks: an overview and latest applications in computational intelligence[J]. International Journal of Intelligent Systems, 2023, 2023(1): 8342104.
[76] FENG J Y, SHEN L M, CHEN Z, et al. HGDetector: a hybrid Android malware detection method using network traffic and Function call graph[J]. Alexandria Engineering Journal, 2025, 114: 30-45.
[77] XU S W, HAN J J, LIU Y L, et al. Few-shot traffic classification based on autoencoder and deep graph convolutional networks[J]. Scientific Reports, 2025, 15: 8995.
[78] LIU M, YANG Q C, WANG W Q, et al. TB-Graph: enhancing encrypted malicious traffic classification through relational graph attention networks[J]. Computers, Materials & Continua, 2025, 82(2): 2985-3004.
[79] ZHAO G Y, LI L W, HE H D, et al. LGSMOTE-IDS: line graph based weighted-distance SMOTE for imbalanced network traffic detection[J]. Expert Systems with Applications, 2025, 281: 127645.
[80] ZHANG H Z, YU L, XIAO X, et al. TFE-GNN: a temporal fusion encoder using graph neural networks for fine-grained encrypted traffic classification[C]//Proceedings of the ACM Web Conference 2023. New York, USA: ACM Press, 2023: 2066-2075.
[81] HE H Y, YANG Z G, CHEN X N. PERT: payload encoding representation from Transformer for encrypted traffic classification[C]// Proceedings of the ITU Kaleidoscope: Industry-Driven Digital Transformation. Washington D.C., USA: IEEE Press, 2020: 1-8.
[82] YU J, CHOI Y, KOO K, et al. A novel approach for application classification with encrypted traffic using BERT and packet headers[J]. Computer Networks, 2024, 254: 110747.
[83] SHI Z L, LUKTARHAN N, SONG Y Y, et al. BFCN: a novel classification method of encrypted traffic based on BERT and CNN[J]. Electronics, 2023, 12(3): 516.
[84] MA X T, LIU T, HU N, et al. Bi-ETC: a bidirectional encrypted traffic classification model based on BERT and BiLSTM[C]// Proceedings of the 8th International Conference on Data Science in Cyberspace (DSC). Washington D.C., USA: IEEE Press, 2024: 197-204.
[85] FARRUKH Y A, WALI S, KHAN I, et al. XG-NID: dual-modality network intrusion detection using a heterogeneous graph neural network and large language model[J]. Expert Systems with Applications, 2025, 287: 128089.
[86] LU H Y, ZHANG R, KONG T. Analyzing decentralized applications traffic: a multimodal approach based on GNN and BERT[C]//Proceedings of the International Conference on Information Security and Cryptology. Singapore: Springer, 2025: 235-254.
[87] MA J J, LI X G, LUO H, et al. NetKD: towards resource-efficient encrypted traffic classification using knowledge distillation for language models[C]//Proceedings of the 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). Washington D.C., USA: IEEE Press, 2024: 3011-3016.
[88] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. [2025-03-07]. https://arxiv.org/abs/2010.11929.
[89] YANG L, GUO S T, LIU D F, et al. ConViTML: a convolutional vision Transformer-based meta-learning framework for real-time edge network traffic classification[J]. IEEE Transactions on Network and Service Management, 2024, 21(3): 3344-3357.
[90] MAJEED U, HASSAN S S, HONG C S. Cross-silo model-based secure federated transfer learning for flow-based traffic classification[C]//Proceedings of the International Conference on Information Networking (ICOIN). Washington D.C., USA: IEEE Press, 2021: 588-593.

[91] PEKAR A, MAKARA L A, BICZOK G. Incremental federated learning for traffic flow classification in heterogeneous data scenarios[J]. Neural Computing and Applications, 2024, 36(32): 20401-20424.
[92] MUN H, LEE Y. Internet traffic classification with federated learning[J]. Electronics, 2021, 10(1): 27.
[93] MAO W J, YU B, ZHANG C, et al. FedKT: Federated learning with knowledge transfer for non-IID data[J]. Pattern Recognition, 2025, 159: 111143.
[94] JIANG W W, MU J B, HAN H Y, et al. Federated learning-based mobile traffic prediction in satellite-terrestrial integrated networks[J]. Software: Practice and Experience, 2025, 55(4): 613-628.
[95] 唐政治, 曾学文, 陈君, 等. 基于机器学习的网络流量分析综述[J]. 网络新媒体技术, 2020, 9(5): 1-8.
TANG Z Z, ZENG X W, CHEN J, et al. A review of network traffic analysis based on machine learning[J]. Network New Media Technology, 2020, 9(5): 1-8. (in Chinese)
[96] VICENZI J C, KOROL G, JORDAN M G, et al. Exploiting virtual layers and pruning for FPGA-based adaptive traffic classification[C]//Proceedings of the 27th Euromicro Conference on Digital System Design (DSD). Washington D.C., USA: IEEE Press, 2024: 194-201.
[97] XU Y W, CAO J, SONG K H, et al. FastTraffic: a lightweight method for encrypted traffic fast classification[J]. Computer Networks, 2023, 235: 109965.
[98] 张琬茜. 面向异构设备的高效网络流量分类技术的研究[D]. 大连: 大连理工大学, 2020.
ZHANG W X. The efficient network traffic classification technologies for heterogeneous devices[D]. Dalian: Dalian University of Technology, 2020. (in Chinese)
[99] 张磊. 基于深度学习的物联网恶意流量识别技术研究[D]. 济南: 齐鲁工业大学, 2024.
ZHANG L. Research on malicious traffic identification technology in Internet of Things based on deep learning[D]. Jinan: Qilu University of Technology, 2024. (in Chinese)
[100] IZADI M, SAFAYANI M, MIRZAEI A. Knowledge distillation on spatial-temporal graph convolutional network for traffic prediction[J]. International Journal of Computers and Applications, 2025, 47(1): 45-56.
[101] 陈梓延, 王晓龙, 何迪, 等. 基于改进 YOLOv8 的轻量化车辆检测网络[J]. 计算机工程, 2025, 51(5): 314-325.
CHEN Z Y, WANG X L, HE D, et al. Lightweight vehicle detection network based on improved YOLOv8[J]. Computer Engineering, 2025, 51(5): 314-325. (in Chinese)

文字编辑 陆燕菲
栏目编辑 宋 圆