AI-powered Internet Traffic Classification: Past, Present, and Future

Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, Valerio Persico, Antonio Pescapé

***Abstract*—Traffic classification (TC) is pivotal for network traffic management and security. Over time, TC solutions leveraging Artificial Intelligence (AI) have undergone significant advancements, primarily fueled by Machine Learning (ML). This paper analyzes the history and current state of AI-powered TC on the Internet, highlighting unresolved research questions. Indeed, despite extensive research, key desiderata goals to product-line implementations remain. AI presents untapped potential for addressing the complex and evolving challenges of TC, drawing from successful applications in other domains. We identify novel ML topics and solutions that address unmet TC requirements, shaping a comprehensive research landscape for the TC future. We also discuss the interdependence of TC desiderata and identify obstacles hindering AI-powered next-generation solutions. Overcoming these roadblocks will unlock two intertwined visions for future networks: self-managed and human-centered networks.**

Introduction

While the importance of the Internet for modern life is hard to overstate, yet it is remarkable how a fundamental question like *"what application(s) produced the traffic in this link?"* often lacks an answer, even for network operators. This "visibility" gap impacts resource allocation, planning, security, and stakeholders' grasp of the global critical infrastructure [1].

It comes as no surprise that network **Traffic Classification (TC)**, i.e. inferring the application or service that generated the observed traffic, has been an intensively researched topic since the early days of the global Internet (see Fig. 1 for an overall timeline) [2]. Indeed, TC evolved with Internet usage changes, requiring the adoption of **Artificial Intelligence (AI)**. The latter contributed to research questions that hardly will ever find a definitive answer, due to the *moving target* nature of network traffic and the *arms race* caused by conflicting interests of the diverse set of stakeholders (network operators, over-the-top providers, equipment manufacturers, regulation bodies, citizens), as revealed by debates and regulations on *network neutrality* and *privacy*.

But the wide, diverse, and recently fast-progressing set of themes explored in AI cannot further help TC research unless they are suitably sorted and organized. This prompted us to the following **main contributions**, organized as shown in Fig. 1. First, we summarize the past and current state of AI-powered TC, which nowadays is predominantly focused on **Machine Learning (ML)** and particularly **Deep Learning (DL)**. Second, we discuss open research questions in the practical application of current AI-powered traffic classifiers to the product line [3]. Third, we leverage the AI hype to identify hot research topics valuable for TC. Fourth, we discuss future visions enabled by next-generation AI-powered TC, roadblocks and mitigation strategies.

AI-powered Traffic Classification: past and present

The quest for practical TC has a long history, summarized by the timeline in Fig. 1. Due to its simplicity, **IP/port-based** classification with shared blocklists is still used for security purposes, with recent IoT standards proposing it (as *manufacturer usage description* whitelist). The need for advanced TC arose already in the '90s, as peer-to-peer compromised the reliability of IP addresses and transport ports in inferring the nature of traffic. Anticipating these needs, research turned to application data for inference, starting the **payload-based** classification era (circa 1998). Among payload-based techniques, *Deep Packet Inspection (DPI)* exploited pattern matching for specific sequences of bytes taken as "fingerprint" of the application [2]. DPI resulted to be expert-labor-intensive (for defining signatures) and computationally intensive (during inference): **ML** came to the rescue (circa 2004) [4]. ML allowed automatic mining of common payload subsequences or frequency signatures, leading to *stochastic payload-based* approaches. Yet, the rise of encrypted protocols and privacy concerns prompted for exploring alternative solutions. Also, ML found application with *statistical flow features* and *flow counters* (or their compressed form) serving as input for traffic classifiers. Other approaches considered factors like fan-in and fan-out based on IP or IP-port combinations, as well as mixes of the aforementioned methods (e.g., *rule-based*).

In the dynamic landscape of today's Internet, the effectiveness of existing TC approaches faces unprecedented challenges. Rapidly changing user behaviors (e.g., remote work and videoconferencing during COVID-19 lockdowns) and the nature of the Internet ecosystem contribute to the fast-paced evolution of traffic. This ecosystem encompasses end devices (e.g., smartphones with many apps, IoT devices), service providers (e.g., cloud services and OTT pushing for widespread TLS adoption, including DNS-over-TLS), and developers (e.g., cross-platform frameworks, third-party services, and automatic updates via app markets). To tackle growing complexity, researchers are turning to recent ML advancements inspired by achievements in computer vision and natural language processing. Specifically, **DL** is garnering attention for TC [10, 14]. DL offers a notable advantage with its ability to automatically extract effective features from input data (*end-to-end learning*). This eliminates the need for

The authors are with the DIETI, University of Naples Federico II, Italy.

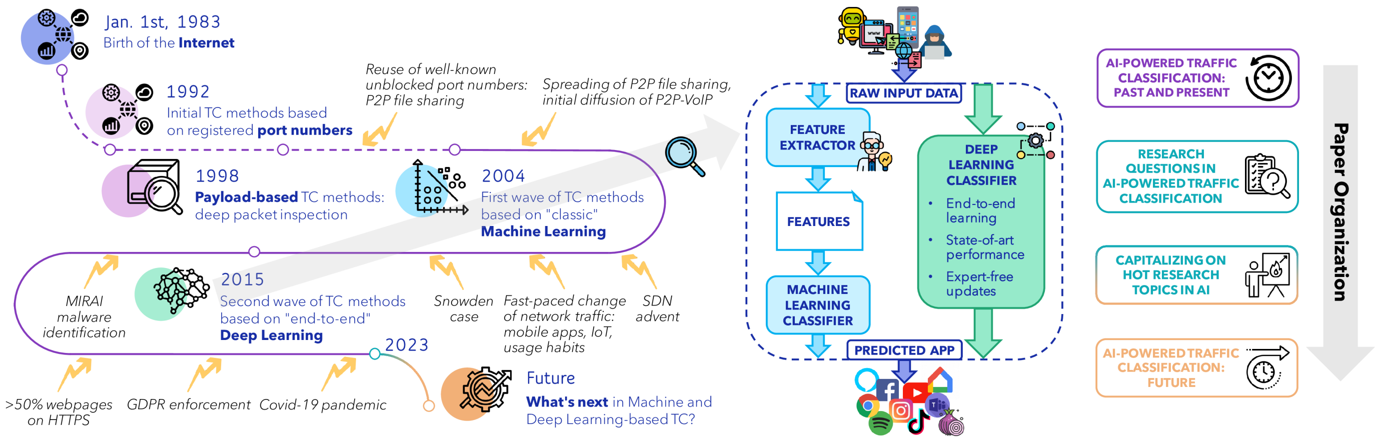

Jan. 1st, 1983
Birth of the Internet
1992
Initial TC methods based on registered port numbers
Reuse of well-known unblocked port numbers: P2P file sharing
Spreading of P2P file sharing, initial diffusion of P2P-VoIP
1998
Payload-based TC methods: deep packet inspection
2004
First wave of TC methods based on "classic" Machine Learning
MIRAI malware identification
2015
Second wave of TC methods based on "end-to-end" Deep Learning
Snowden case
Fast-paced change of network traffic: mobile apps, IoT, usage habits
SDN advent
2023
>50% webpages on HTTPS
GDPR enforcement
Covid-19 pandemic
Future
What's next in Machine and Deep Learning-based TC?
RAW INPUT DATA
FEATURE EXTRACTOR
FEATURES
MACHINE LEARNING CLASSIFIER
DEEP LEARNING CLASSIFIER
End-to-end learning
State-of-art performance
Expert-free updates
PREDICTED APP
AI-POWERED TRAFFIC CLASSIFICATION: PAST AND PRESENT
RESEARCH QUESTIONS IN AI-POWERED TRAFFIC CLASSIFICATION
CAPITALIZING ON HOT RESEARCH TOPICS IN AI
AI-POWERED TRAFFIC CLASSIFICATION: FUTURE
Paper Organization

Fig. 1. TC evolution: from port-based approaches to beyond-DL methods, with a highlight on expert-driven ML vs. end-to-end DL. The dashed line highlights the pre-ML era (before 2004). Paper organization is color-coded to match the timeline (past, present, future).

manual feature engineering by experts. Through joint training of feature extraction and classification, DL models achieve *state-of-art performance* and enable *expert-free updates* [14]. This holds significant promise for tackling the ever-changing and fast-paced nature of traffic to classify.

The research progress in TC is traced in Tab. I, summarizing *IEEE magazine papers* from the past two decades. Particularly, for each work, we report the overall number of citations (from Google Scholar) and we highlight focus on AI approaches or specifically DL. It emerges how the research targets in TC have evolved. Initially, they encompassed generic monitoring methods without AI utilization [1], as well as specialized TC approaches for ad-hoc traffic types like gaming [4], and fine-grained [5] ones. Identifying TC challenges and proposing tackling strategies [2], along with the development of open platforms [6], paved the way for the use of ML models [7], also incorporating transfer-learning [8]. Recently, DL-based TC has seen a surge of interest [9, 10] with the definition of systematic approaches optimizing feature selection [11], DL-based traffic prediction aided by TC [12], and explainable solutions [13].

In the following section, we describe and analyze the main **open research questions** in AI-powered TC.

Research Questions in AI-powered Traffic Classification

Five *research questions* can be identified from the inspection of the literature providing specific AI-based solutions for TC or related surveys [15] in the last two decades. They reflect the limitations of the state-of-the-art, and some arise or are emphasized when resorting to AI. We discuss them in the form of ***the desiderata a TC solution is expected to meet***: *effectiveness, deployability, trustworthiness, robustness, and adaptivity* (see Fig. 2).

Effectiveness: Can it provide me with accurate traffic visibility? Effectiveness is undeniably the primary desideratum in TC. Hence, the existing literature has consolidated the methodological aspects for its evaluation. Regrettably, effectiveness figures are far from requirements even in close-to-real contexts. For instance, up to 99% accuracy is attainable in scenarios with both few classes (< 10 apps or categories with distinct traffic patterns) and the almost complete observation of the traffic aggregate (no routing asymmetries). Still, such performance becomes unrealistic with current methods when

TABLE I
IEEE MAGAZINES ON TC SINCE THE ML ERA (2004–PRESENT, FIRST OCCURRENCES IN 2008). PAPERS ARE LISTED CHRONOLOGICALLY.

Reference	Year	Venue	#Citations	Focus	AI	DL
Kind et al. [1]	2008	Commun. Mag.	26	*Holistic network monitoring approach*: traffic measurement and analysis	○	
But et al. [4]	2008	Commun. Mag.	15	ISP-based system to classify and prioritize *game traffic in real-time* using ML	◐	
Park et al. [5]	2011	Commun. Mag.	34	*Fine-grained TC*: different traffic types within a single application	○	
Dainotti et al. [2]	2012	Netw.	710	*Issues and future directions in TC*: trade-offs in applicability, reliability, and privacy	◐	
De Donato et al. [6]	2014	Netw.	78	*Open platform* for evaluation and combination of TC techniques	◐	
Canovas et al. [7]	2018	Netw.	20	ML-based system for QoE classification of *video traffic*	●	
Grolman et al. [8]	2018	Intell. Syst.	39	User action identification in *mobile-app traffic* via ML with transfer learning	●	
Li et al. [9]	2018	Netw.	59	*DL-based stacked autoencoders for TC* trained via semi-supervised learning	●	✓
Rezaei and Liu [10]	2019	Commun. Mag.	393	Framework for *DL-based TC*: open problems, challenges, and opportunities	●	✓
Shen et al. [11]	2020	Netw.	60	Systematic approach to optimize *feature selection for encrypted TC*	●	
Zhang et al. [12]	2020	Netw.	11	*DL-based traffic classification and prediction scheme* for smart gateways	●	✓
Zhang et al. [13]	2022	Commun. Mag.	12	*Overview of eXplainable AI for networking* with future challenges and directions	●	✓

● focus on / ◐ mention of AI approaches.

the number of classes grows, the classification granularity becomes finer (down to specific apps rather than just categories), or the portion of traffic available for decisions is limited.

Deployability: Can I deploy it with my network assets and constraints? A number of physical constraints (time, memory, processing resources, energy) impact the training or inference phase: these limit the practical applicability of TC solutions. TC models must suit the (network) devices where they operate, whose nature varies widely, encompassing legacy routers, specialized hardware components, AI-focused hardware accelerators (e.g., Huawei Ascend or Google Tensor Processing Units), and middleboxes. While virtualized edge/cloud tiers (e.g., Amazon Web Services or Google Cloud platforms) can alleviate resource limitations, they may not be ideal when classification outcomes must drive fast and local real-time decisions (e.g., traffic-flow scheduling, attack blocking). Collection and labeling of traffic samples can impose further constraints. Cross-admin-boundaries transfer of these data could be restricted by intellectual property or privacy policies and concerns. For instance, the *EU GDPR* expressly focuses on healthcare and children-compliant apps, and even IP addresses are categorized as personally-identifiable info.

Trustworthiness: Why should I trust it and to what extent? ML-based TC solutions can deliver highly accurate output but they are often treated as black boxes. Their automated construction solely driven by traffic data creates complex inner mechanisms that are challenging for human understanding. This clashes with the need for operators to understand *why* a model follows a specific decision process, assess its reliability, predict its behavior, and make improvements. As all ML methods are based on data, the quality of traffic datasets also impacts the reliability of the outcome. This is specifically critical for TC, as input data have an extremely complex structure (as opposed to pixel colors or letter sequences), with different and varying semantics of the observed sample, given by the stack of network protocols. Understanding which subset of the dataset (specific applications/services/protocols) or input (specific protocol fields or parts of the payload stream) impacts the model creation and its performance is essential. This provides the necessary insight to trust (or reject, improve, fix) the TC solution. Regulatory issues further complicate this aspect. Emerging regulations and standards, such as the *EU AI Act*, prohibit the deployment of AI systems with potential impacts on individuals' lives unless they ensure either technical transparency or explainability.

Robustness: Will it keep working *when* the network context changes? ML solutions are trained in *specific* network contexts, encompassing a mix of applications (and their versions), user population (and their habits), and devices (and their settings). However, it is crucial that these traffic classifiers can be effectively applied—without significant impact on the aforementioned desiderata—in *other contexts* or when the *same operating context is subject to changes*. Furthermore, the nature of network traffic can change due to adversarial behaviors, such as network attacks or the deliberate use of crafted inputs to evade malware identification. While some of these factors can be easily documented and tracked, others can be completely hidden. Hence, model robustness should be routinely re-evaluated by assessing potential degradation between design/training and deployment scenarios.

Adaptivity: If its visibility of traffic drops, can I fix it, and at what cost? When a TC model is transferred to a different context, there are two scenarios to consider: (*i*) model performance may deteriorate below the expected level due to the lack of robustness (e.g., due to concept drift); (*ii*) the new context may necessitate a different classification task, such as accommodating new apps/services or different traffic fingerprints. In such cases, adaptation becomes necessary. Ideally, this adaptation should provide the same performance in terms of effectiveness, trustworthiness, and deployability achieved by designing a model ex-novo in the new context (with a minor drop with respect to the original model, in the worst case). Also, it is imperative that such a process incurs limited costs, including the collection of new datasets and retraining procedures. The degree to which these properties are fulfilled determines the level of adaptivity.

Assessing Desiderata: Evaluation Setups. Evaluation setups are crucial for quantitative analysis of research questions (Fig. 2). Tailored setups assess TC desiderata using either a broad set of *evaluation metrics and tools* or adopting some of them in a *comparative scenario*. Regarding *effectiveness*, metrics range from concise ones like accuracy (measured flow- or byte-wise [2]) to per-app breakdown and soft-output capitalization [14]. *Trustworthiness* is measured through calibration and interpretability tools, assessing to what extent a traffic classifier can be trusted and which flow/packet portions mainly contribute to its outcomes. *Deployability* setups group metrics for assessing training requirements, run-time, memory occupation, time-to-insight, and throughput. Robustness and adaptivity instead inherit metrics from other desiderata but used in a differential fashion. Regarding *robustness*, open-world, cross-dataset, and adversarial setups assess how well a traffic classifier responds to unseen apps/services, deployment in different conditions (e.g., different vantage points), and packet mutation attacks, respectively. Differently, *adaptivity* is evaluated via incremental learning setups, to measure how well (and at what cost) an existing traffic classifier can accommodate refreshed app patterns, additional apps, or further network visibility tasks (e.g., QoS classes). The evaluation compares it to an ideal classifier built from scratch with improved knowledge.

TC Desiderata Interplay. It is important to note that these desiderata are *interdependent*, thus necessitating joint evaluation and possibly trade-offs to find the sweet spot. Figure 2 illustrates their interplay, highlighting potential *positive* and *negative impacts* among them (measured via the evaluation setup previously discussed). Effectiveness as a main goal is easily affected by any additional requirement, limiting viable solutions and incurring negative impacts from all the other desiderata. For instance, incorporating trustworthiness may lead to an explainable but less accurate traffic classifier (e.g., packet features simpler to understand but less discerning).

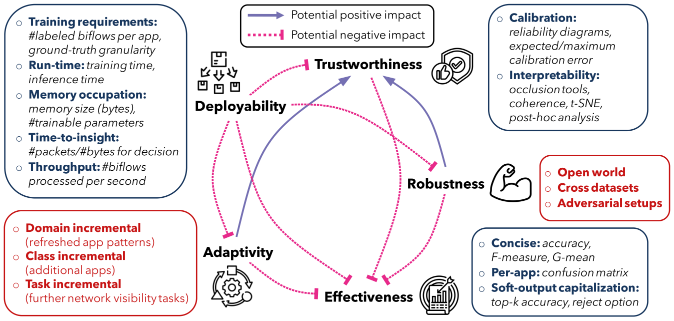

Potential positive impact
Potential negative impact
Training requirements: #labeled biflows per app, ground-truth granularity
Run-time: training time, inference time
Memory occupation: memory size (bytes), #trainable parameters
Time-to-insight: #packets/#bytes for decision
Throughput: #biflows processed per second
Deployability
Trustworthiness
Calibration: reliability diagrams, expected/maximum calibration error
Interpretability: occlusion tools, coherence, t-SNE, post-hoc analysis
Robustness
Open world
Cross datasets
Adversarial setups
Domain incremental (refreshed app patterns)
Class incremental (additional apps)
Task incremental (further network visibility tasks)
Adaptivity
Effectiveness
Concise: accuracy, F-measure, G-mean
Per-app: confusion matrix
Soft-output capitalization: top-k accuracy, reject option

Fig. 2. TC desiderata interplay. Positive and negative (trade-offs) potential impacts are shown between them, as measured by relevant evaluation setups. The setups grouping metrics or tools are in blue, while those based on comparative scenarios are in red. Specific evaluation metrics and tools are reported in *italic* font.

Conversely, deployability introduces physical, logistic, and cost constraints that negatively impact all other desiderata. For instance, when a classifier is specialized for deployment to a local vantage point, any network context change may affect its robustness and product-line constraints may hamper its interpretability. Robustness and adaptivity can also cater to trustworthiness, thanks to a more general traffic representation provided by the model. These interactions call for a *comprehensive quantitative assessment* of AI-powered TC solutions as a first necessary step towards a theoretical framework.

CAPITALIZING ON HOT RESEARCH TOPICS IN AI

Based on research questions, we explore *AI topics* that *can* address unsatisfied or improvable requirements in TC. Leveraging established AI reports like Gartner, we identify promising areas for TC research based on our expertise via the methodological steps in Fig. 3 (top). These areas encompass several recent *AI solutions* and established ones that have however untapped potential for modern TC operations. Figure 3 illustrates the maturity and research interest of these AI solutions in TC using a *hype cycle* representation (left). Also, the grid (right) illustrates how each AI topic is focused on (or could improve) each of the five desiderata.

ModelOps. *ModelOps* is an emerging trend in industry innovation that offers a comprehensive approach to harness the practical benefits of AI. It encompasses the automation of the ML lifecycle, also known as MLOps, utilizing platforms such as `MLflow` and `Neptune`. These enable tracking and management of data and processes, ensuring seamless storage and deployment on multiple infrastructures. Lifecycle automation is essential for the end-to-end development of traffic classifiers, starting from raw data collection to their auditable deployment and maintenance across different network points.

Automated managed processes include data preprocessing, model optimization, maintenance, and update. Model optimization can leverage the *AutoML* umbrella to devise node-focused traffic classifiers at scale with no human expert in the loop, for instance via `Auto-Sklearn` or `AutoKeras` Python libraries. Conversely, model maintenance and update can capitalize *continual learning*, which addresses updating AI models with novel knowledge when new data stream in, such as recognizing traffic from emerging apps.

TRiSM. AI *(T)rust, (Ri)sk, & (S)ecurity (M)anagement* is aimed at providing AI tools with model governance, trustworthiness, fairness, reliability, efficacy, security, and data protection. This includes solutions for model transparency, adversarial attack resistance, and privacy.

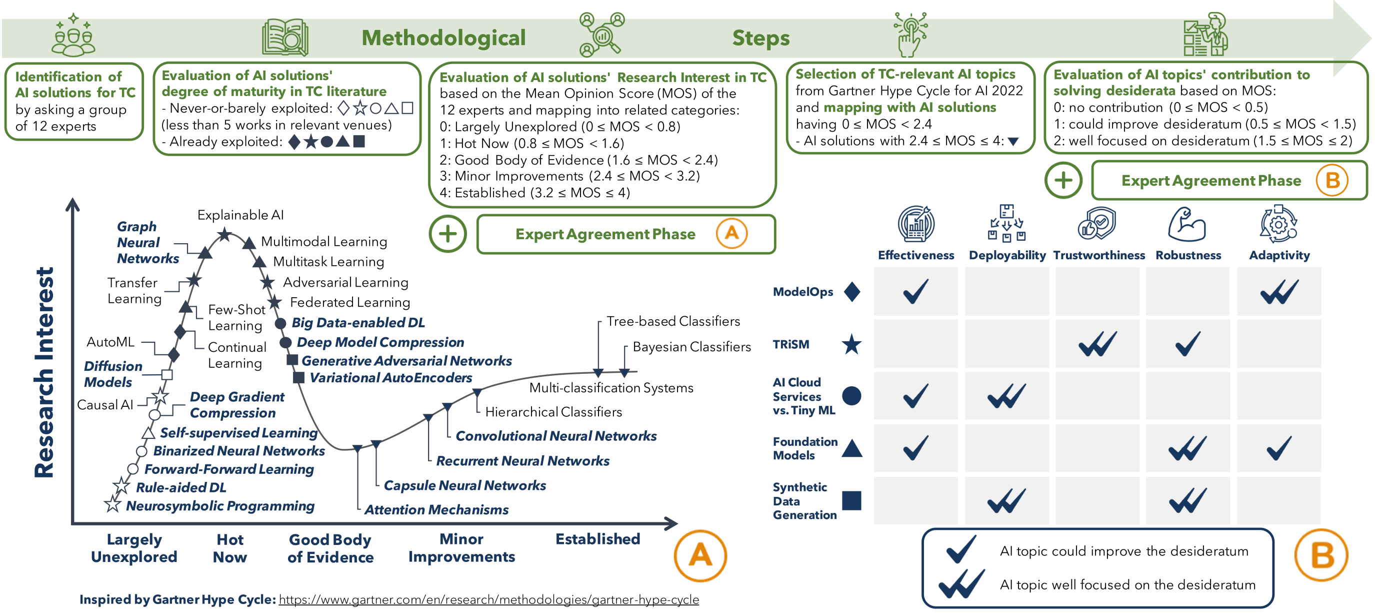
To ensure transparency of black-box AI-based traffic classifiers, *eXplainable AI (XAI)* tools are crucial, providing post-hoc explanations (e.g., `LIME` and `SHAP`) or using novel explainable-by-design architectures. In the latter context, *Causal AI* techniques build traffic classifiers that can infer cause-effect relations and predict the outcome of parameter changes such as degraded network status. With the same aim, *rule-aided DL* can integrate a-priori human information (e.g., protocol-originated) into solely black-box TC decisions. This is useful in cases with minimal data supervision and provides decision consistency with format/network constraints. *Neurosymbolic programming* combines DL for feature extraction with symbolic program synthesis to enhance traffic classifiers. It generates human-readable code (e.g., in `Scallop` language) systematically incorporating a-priori knowledge.

Conversely, *adversarial learning* addresses AI model vulnerabilities against attacks like poisoning and evasion. The idea is that ad-hoc altering/forging network traffic (traffic morphing) can trick the model into misclassifying a given app or giving away traffic-sensitive information. The aim is to make TC tools robust in such adversarial network environments.

Privacy in training data-driven traffic classifiers can be achieved through various strategies. *Transfer learning* splits the training process into two stages. Pre-training uses a large dataset to create a rich model, while fine-tuning uses a specialized dataset to enhance and adapt it. Assigning these tasks to two distinct actors, transfer learning enables asymmetric knowledge sharing (e.g., between big and small network operators) and eliminates the need to share traffic data. *Federated learning* decouples learning from dataset storage, allowing distributed training with periodic updates to/from a central server via dedicated frameworks such as `OpenFL` and `TensorFlow Federated`. Edge devices perform local training operations without sharing traffic data, improving privacy protection, and accommodating business-sensitive constraints.

Cloud AI Services vs. Tiny ML. When considering learning paradigms and technologies to support AI-powered TC solutions, conflicting trends emerge. On one hand, cloud datacenters provide virtually unlimited computing resources, facilitating centralized training. On the other hand, practical limitations push for decentralized learning on edge devices, even on the same vantage point capturing the traffic.

Network traffic data align with the characteristics of Big Data in volume, variability, and velocity. Leveraging abundant computing power can significantly accelerate the training process in *Big Data-enabled DL*. Techniques such as data-parallel schemes, where multiple CPUs handle subsets of training data, and careful data layout design, contribute to this acceleration, with ad-hoc libraries (`Apache Spark MLlib`) and frameworks (`Horovod`) adopted to this aim.


Methodological
Steps
Identification of AI solutions for TC by asking a group of 12 experts
Evaluation of AI solutions' degree of maturity in TC literature
- Never-or-barely exploited: ◇☆○△□
(less than 5 works in relevant venues)
- Already exploited: ◆★●▲■
Evaluation of AI solutions' Research Interest in TC based on the Mean Opinion Score (MOS) of the 12 experts and mapping into related categories:
0: Largely Unexplored (0 ≤ MOS < 0.8)
1: Hot Now (0.8 ≤ MOS < 1.6)
2: Good Body of Evidence (1.6 ≤ MOS < 2.4)
3: Minor Improvements (2.4 ≤ MOS < 3.2)
4: Established (3.2 ≤ MOS ≤ 4)
Selection of TC-relevant AI topics from Gartner Hype Cycle for AI 2022 and mapping with AI solutions having 0 ≤ MOS < 2.4
- AI solutions with 2.4 ≤ MOS ≤ 4: ▼
Evaluation of AI topics' contribution to solving desiderata based on MOS:
0: no contribution (0 ≤ MOS < 0.5)
1: could improve desideratum (0.5 ≤ MOS < 1.5)
2: well focused on desideratum (1.5 ≤ MOS ≤ 2)
+ Expert Agreement Phase A
+ Expert Agreement Phase B
Research Interest
Explainable AI
Graph Neural Networks
Multimodal Learning
Multitask Learning
Transfer Learning
Adversarial Learning
Federated Learning
Few-Shot Learning
Big Data-enabled DL
AutoML
Continual Learning
Deep Model Compression
Generative Adversarial Networks
Diffusion Models
Variational AutoEncoders
Causal AI
Deep Gradient Compression
Self-supervised Learning
Binarized Neural Networks
Forward-Forward Learning
Rule-aided DL
Neurosymbolic Programming
Tree-based Classifiers
Bayesian Classifiers
Multi-classification Systems
Hierarchical Classifiers
Convolutional Neural Networks
Recurrent Neural Networks
Capsule Neural Networks
Attention Mechanisms
Largely Unexplored
Hot Now
Good Body of Evidence
Minor Improvements
Established
A
Inspired by Gartner Hype Cycle: https://www.gartner.com/en/research/methodologies/gartner-hype-cycle
B

	Effectiveness	Deployability	Trustworthiness	Robustness	Adaptivity
ModelOps ◆	✓				✓✓
TRiSM ★			✓✓	✓	
AI Cloud Services vs. Tiny ML ●	✓	✓✓			
Foundation Models ▲	✓			✓✓	✓
Synthetic Data Generation ■		✓✓		✓✓	

✓ AI topic could improve the desideratum
✓✓ AI topic well focused on the desideratum

Fig. 3. Hot research topics in AI for TC: *methodological steps* for devising the *Hype Cycle* (top) showing the *degree of research interest* of noteworthy solutions versus their *level of maturity* (left) and *matching with TC desiderata* they contribute to fulfill (right). Each marker symbolizes the AI topic to which the AI solution belongs (center). We report the AI solutions specific to DL in ***bold italic** and blue color.*

Conversely, *tiny ML* focuses on ML technologies and applications (hardware, software, and processes) to be run on devices with low memory, computing, and bandwidth resources, possibly battery-operated (which are ubiquitous in networks). Therefore, it can help perform training directly on edge devices (enabling transfer learning and federated learning) by simplifying training algorithms, coping with memory limitations, and reducing exchanged information. Several strategies in this respect can be implemented, such as the more energy-efficient *forward-forward learning* instead of classic backpropagation, *online learning* techniques as opposed to batch learning, and *deep gradient compression* to significantly reduce the communication bandwidth in distributed training. Besides training, tiny ML aims to simplify the inference stage by reducing the complexity of ML models for deployment on network interface cards. This can be achieved through approaches like *binarized neural networks* (`Larq` Python library being a popular implementation) or *deep model compression* (involving knowledge distillation, pruning, and quantization). Energy-efficient hardware (e.g., neuromorphic chips such as `IBM TrueNorth`) can also improve the environmental impact of DL solutions by disabling inactive neurons.

Unified Representation Learning / Foundation Models. A *foundation model* is a large AI model trained on a vast quantity of unlabeled data at scale, that can be adapted to a wide range of downstream tasks. Recent trends in AI rely on the novel concept of *self-supervised learning* to distill unified representations from unlabeled data via suitably defined pretext tasks, echoing successful applications in natural language processing (`ChatGPT`) and computer vision (`Dall-E`). Self-supervised learning hugely benefits from structured and *multimodal-like* inputs (as those related to network traffic aggregates) and provides benefits in terms of out-of-distribution detection (which maps into open-world TC). Effective design hinges on aligning the input with the model, with *graph neural networks* (built using `PyG` or `DGL` libraries) having the potential to capture complex interactions of network data at different levels of abstraction (e.g., flows, sequences of flows).

Downstream tasks encompass diverse network visibility perspectives, including the recognition of services or apps at different granularity, or app-conditional traffic prediction. *Multi-task learning* can handle these diverse problems with a single solution. Also, the challenges associated with collecting an extensive ground truth for pursued TC tasks can be addressed using *few-shot learning* techniques. These techniques leverage foundational models while enabling fine-tuned adaptation to the specific task at hand with minimal supervision.

Synthetic Data Generation. Generative modeling is an unsupervised learning task that involves automatically discovering the regularities or patterns in input data for generating new samples virtually indistinguishable from the original. The generation of synthetic traffic mitigates the poor availability of public network-traffic datasets due to difficulties in collection. When combined with few-shot training, generative approaches mitigate data scarcity issues even in scenarios characterized by the emergence of new apps/services or zero-day network attacks. Also, generative approaches can lower the barrier to the sharing of datasets by providing privacy-preserving tools.

Various DL networks have been proposed for synthetic traffic generation. *Variational methods* learn data-generating distributions via variational inference. *Generative adversarial networks* put two networks against each other, with the generator producing synthetic samples indistinguishable from real ones. Recently, latent and conditional *diffusion models* have found successful application in the generation of images given their textual description (e.g., `Stable Diffusion` and

TABLE II
TAXONOMY OF DATASETS EMPLOYED FOR ENCRYPTED-TRAFFIC CLASSIFICATION* RELEASED IN THE LAST DECADE (2013–2023, FIRST OCCURRENCES IN 2016). DATASETS ARE LISTED CHRONOLOGICALLY BY RELEASE YEAR.

Dataset	Release Year	Traffic Nature	🧍	Label Space	Raw Data	Capture Span
ISCXVPN2016	2016	🖥	●	2 encapsulation types / 7 traffic types / 15 apps	✓	03/15 – 06/15
ISCXTor2016	2016	🕵	●	8 traffic types / 18 apps	✓	07/15 – 02/16
Anon17	2017	🕵	●	3 anonymity tools / 8 traffic types / 21 apps		2014 – 2017
MTD	2018	📱	●	12 apps		10/16 – 03/17
QUIC	2018	🖥	○	5 QUIC services		03/18
UNSWIoT	2018	🤖	●	28 devices	✓	10/16 – 04/17
MIRAGE-2019	2019	📱	●	40 apps		05/17 – 05/19
MIRAGE-VIDEO	2020	📱	●	4 video categories / 8 apps		06/19 – 03/20
Orange'20	2020	📱	●	8 traffic types		11/07/19
UTMobileNetTraffic2021	2021	📱	◐	16 apps / 31 user activities		03/18 – 04/18
MIRAGE-COVID-CCMA-2022	2022	📱	●	9 apps / 3 user activities		04/21 – 12/21
CICIoT2022	2022	🤖	◐	3 device types / 40 devices	✓	09/21 – 12/21
AppClassNet	2022	📱🖥	●	500 apps		N/A
CESNET-QUIC22	2022	📱🖥	●	7 traffic types / 102 services		11/22

Traffic Nature: 📱 = Mobile Apps, 🕵 = Anonymity Tools, 🖥 = Desktop, 🤖 = IoT.
🧍 = Human-generated: whether it is completely/partially generated by real human experimenters, as opposed to bots or scripts.
Raw Data: PCAP files are available.
*The datasets related to network-anomaly detection or attack classification have been excluded due to their specific focus on network security.

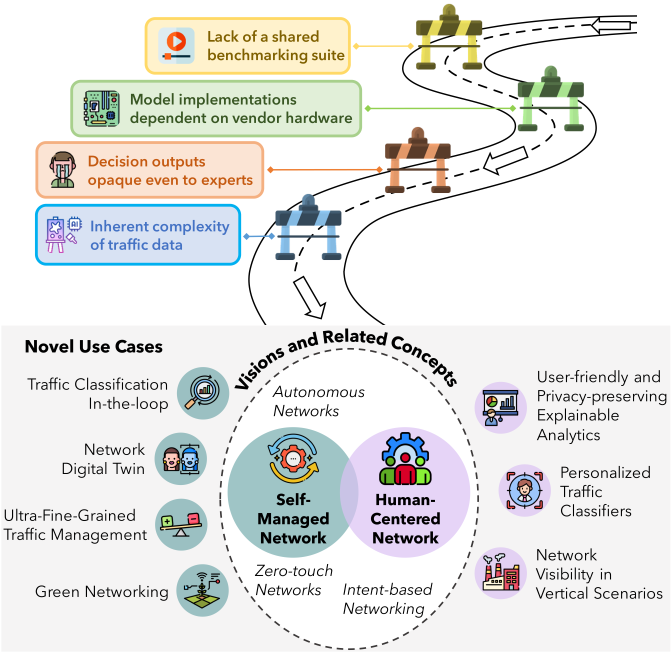

Lack of a shared benchmarking suite
Model implementations dependent on vendor hardware
Decision outputs opaque even to experts
Inherent complexity of traffic data
Novel Use Cases
Visions and Related Concepts
Traffic Classification In-the-loop
Network Digital Twin
Ultra-Fine-Grained Traffic Management
Green Networking
Autonomous Networks
Self-Managed Network
Human-Centered Network
Zero-touch Networks
Intent-based Networking
User-friendly and Privacy-preserving Explainable Analytics
Personalized Traffic Classifiers
Network Visibility in Vertical Scenarios

Fig. 4. AI-powered TC future: roadblocks (top) and novel use cases (bottom).

`Google Imagen`) but are yet to be explored in the TC context.

AI-POWERED TRAFFIC CLASSIFICATION: FUTURE

The future of TC is tied to AI, but what is the ultimate goal? Two intertwined visions emerge (with AI-powered TC as a core enabler): **self-managed network** and **human-centered network** (see Fig. 4). These aim at automation, while keeping goals and supervisory control firmly in human hands. The *self-managed network* aims to exclude direct human intervention by creating a *proactive system* that monitors and reconfigures itself. The *human-centered network* seeks a *human-in-command* approach through an abstracted interface for defining goals and policies. **AI is crucial for materializing these ideas** and specifically addressing traffic monitoring, policing, and human understanding: hereafter we highlight how the AI research topics empower novel TC **use cases**.

AI-powered TC for the Self-Managed Network. The ModelOps paradigm supports the fully-automated management of network infrastructures by leveraging TC outcomes (*TC-in-the-loop*). This approach capitalizes on the opportunities offered by software-defined networking (e.g., via `OpenFlow` and `P4`) and enables reinforcement learning. The *network digital twin* plays a crucial role by collecting network environment information, simulating TC-informed decisions, and predicting their impact before putting them in play. Improving effectiveness through foundation models and the ModelOps lifecycle is key to advancing QoS management alongside emerging 6G technologies. This advancement facilitates *ultra-fine-grained traffic management*, supporting applications such as augmented/virtual reality metaverse, vehicle-to-vehicle, and unmanned-vehicle communications. Additionally, cloud AI services combined with tiny ML can support a wide range of traffic monitors aligned with *green networking* principles. Their joint use simplifies TC and learning phases for diverse network devices. Indeed, in the latter case, the training process can be lightweight at the edge or delegated to the cloud.

AI-powered TC for the Human-Centered Network. The assurance of TRiSM techniques in TC is the necessary stepping stone toward the design of *user-friendly* and *privacy-preserving* dashboards providing *explainable analytics* that will widen the adoption of TC onto more diverse contexts, also including personal networks. Foundation models are particularly beneficial for creating fine-grained *personalized traffic classifiers*. This is especially relevant in "Small Office Home Office" environments where the bring-your-own-device policy is common. Personalized classifiers enhance transparency and accountability in operations, leading to improved efficiency and fairness in communication resource management. Finally, network visibility emerges as a valuable source of information also in *vertical scenarios* where network services are not the primary focus, like controlling critical infrastructures or managing physical assets in Industry 4.0.

Roadblocks and Mitigations. Various **roadblocks** impede full AI technology utilization in realizing these visions (see Fig. 4); we discuss possible **mitigation** strategies.

Reproducibility and thorough evaluations pose significant TC research challenges. A *shared benchmarking suite* (similar to `Kaggle` or `GitHub`) housing open-source implementations, datasets, and metrics, could expedite progress. Initiatives like the ITU "AI/ML in 5G Challenge" are the first steps in this direction. Standardized methods for generating, collecting, preprocessing, labeling and sharing traffic data are crucial but often lacking due to diverse and privacy-sensitive collection contexts. Data availability has long been an issue [2], especially for ML and DL. New privacy-preserving solutions and incentive systems (e.g., Blockchain-based NFTs) can help. Table II categorizes public datasets for encrypted-traffic classification, released from 2016 onwards to accommodate growing encrypted protocol use, particularly TLS. Older datasets no longer represent the current traffic landscape. We flag datasets providing *raw traffic data* for DL end-to-end learning.

Concerning deployability, although TC solutions based on tiny ML address scalability concerns, they might rely on *specific network-device hardware implementations*, thus becoming vendor-dependent. Hence, innovative AI-powered TC solutions suited for existing open hardware (e.g., `Corundum FPGA NIC`) and/or legacy systems are highly needed. A similar reasoning applies to novel open hardware architectures specifically designed for AI-powered TC.

Interpreting DL-based TC solutions is a challenge due to *non-expert-friendly explanations* [13]. Bridging the gap between network-input attribution and actionable interpretability remains a challenge. Unlike mature fields like computer vision, current solutions struggle to provide intuitive answers to questions like "why these protocol fields or traffic-flow portions have driven this decision?", especially in presence of encryption. Tools that provide higher-level interpretability, mapping traffic portions to their "network semantics", are becoming essential. Addressing sensitive fields like cleartext payload must consider *privacy* and data protection to make explanations effective while safeguarding personal information.

The *complex and multimodal nature of traffic data* poses challenges in defining effective foundation models (e.g., avoiding unwanted dependence on specific facets of traffic inputs) and generating diverse (while realistic) synthetic traffic. Generating traces from user prompts, akin to image generation in computer vision, is currently unfeasible. The exploitation of multi-view representations (e.g., traffic in graph and time-series forms) and domain knowledge (e.g., expert rules encoded into AI techniques) can address this challenge.

Wrap-up. With our research experience in AI-powered TC, we offer an up-to-date and comprehensive understanding of this field. The convergence of novel tools, approaches, and evolving perspectives on the human role in technology has given rise to compelling future visions. We encourage TC researchers to overcome the *effectiveness-only* viewpoint and dismantle the roadblocks hindering progress. By doing so, we can transform these envisioned futures into tangible reality, also overcoming resistance to AI transition due to concerns regarding hardware/software investments, maintenance, and return on investment.

References

[1] A. Kind *et al.*, "Advanced network monitoring brings life to the awareness plane," *IEEE Commun. Mag.*, vol. 46, no. 10, pp. 140–146, 2008.

[2] A. Dainotti *et al.*, "Issues and future directions in traffic classification," *IEEE Netw.*, vol. 26, no. 1, pp. 35–40, 2012.

[3] A. Lavin *et al.*, "Technology readiness levels for machine learning systems," *Nature Communications*, vol. 13, no. 1, p. 6039, 2022.

[4] J. But *et al.*, "Outsourcing automated QoS control of home routers for a better online game experience," *IEEE Commun. Mag.*, vol. 46, no. 12, pp. 64–70, 2008.

[5] B. Park *et al.*, "Toward fine-grained traffic classification," *IEEE Commun. Mag.*, vol. 49, no. 7, pp. 104–111, 2011.

[6] W. De Donato *et al.*, "Traffic identification engine: an open platform for traffic classification," *IEEE Netw.*, vol. 28, no. 2, pp. 56–64, 2014.

[7] A. Canovas *et al.*, "Multimedia Data Flow Traffic Classification Using Intelligent Models Based on Traffic Patterns," *IEEE Netw.*, vol. 32, no. 6, pp. 100–107, 2018.

[8] E. Grolman *et al.*, "Transfer learning for user action identification in mobile apps via encrypted traffic analysis," *IEEE Intell. Syst.*, vol. 33, no. 2, pp. 40–53, 2018.

[9] P. Li *et al.*, "An Improved Stacked Auto-Encoder for Network Traffic Flow Classification," *IEEE Netw.*, vol. 32, no. 6, pp. 22–27, 2018.

[10] S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 76–81, 2019.

[11] M. Shen *et al.*, "Optimizing Feature Selection for Efficient Encrypted Traffic Classification: A Systematic Approach," *IEEE Netw.*, vol. 34, no. 4, pp. 20–27, 2020.

[12] J. Zhang *et al.*, "Intelligent and application-aware network traffic prediction in smart access gateways," *IEEE Netw.*, vol. 34, no. 3, pp. 264–269, 2020.

[13] T. Zhang *et al.*, "Interpreting AI for Networking: Where We Are and Where We Are Going," *IEEE Commun. Mag.*, vol. 60, no. 2, pp. 25–31, 2022.

[14] G. Aceto *et al.*, "Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges," *IEEE Trans. Netw. Serv. Manag.*, vol. 16, no. 2, pp. 445–458, 2019.

[15] M. S. Sheikh and Y. Peng, "Procedures, criteria, and machine learning techniques for network traffic classification: a survey," *IEEE Access*, vol. 10, pp. 61 135–61 158, 2022.

Giuseppe Aceto (giuseppe.aceto@unina.it) is an Associate Professor at the University of Napoli Federico II. His research concerns network performance, traffic analysis, Internet censorship, and ICTs applied to health.

Domenico Ciuonzo [SM] (domenico.ciuonzo@unina.it) is a Tenure-Track Professor at the University of Napoli Federico II. His research concerns data fusion, network analytics, IoT, signal processing, and AI.

Antonio Montieri (antonio.montieri@unina.it) is an Assistant Professor at the University of Napoli Federico II. His research concerns network measurements, traffic classification, modeling and prediction, and AI for networks.

Valerio Persico (valerio.persico@unina.it) is a Tenure-Track Professor at the University of Napoli Federico II. His work concerns network measurements, traffic analysis, cloud-network monitoring, and Internet topology discovery.

Antonio Pescapé [SM] (pescape@unina.it) is a Full Professor at the University of Napoli Federico II. His work focuses on measurement, monitoring, and analysis of the Internet.

Open Access funding provided by Università degli Studi di Napoli Federico II within the CRUI CARE agreement.